Universidade Aberta

Análise Multinível: Aplicação em Avaliação de Desempenho Escolar

Emerson Andrade Pires

Mestrado em Estatística, Matemática e Computação:

Especialidade – Estatística Computacional

Março de 2020

Universidade Aberta

Análise Multinível: Aplicação em Avaliação de Desempenho Escolar

Emerson Andrade Pires

Mestrado em Estatística, Matemática e Computação:

Especialidade – Estatística Computacional

Dissertação de mestrado orientada pela Professora Doutora Maria do Rosário Olaia Duarte Ramos

Março de 2020

RESUMO

A natureza, a quantidade de variáveis envolvidas no processo de ensino aprendizagem, a forma como estas variáveis estão agrupadas: variáveis do aluno, da sala, do professor, da escola, da família, etc. nos impõe que qualquer estudo aplicado nesta área não deve descurar esta hierarquia organizacional, pois se tal acontecer, as conclusões advenientes correm o risco de serem desprovidas de rigor científico.

Nesta vertente, aparecem os modelos de regressão multiníveis que estatisticamente trazem esta mais valia que é a possibilidade de analisar dados que possuem uma organização hierárquica, o que na educação se verifica de uma forma bem óbvia, visto termos alunos (com caraterísticas próprias) agrupados em turmas (um outro nível hierárquico) que por sua vez estão agrupadas em escolas, que também podem pertencer a agrupamentos escolares. Apesar desta evidente organização hierárquica, há uma regra estatística que nos irá permitir avançar ou não com o ajustamento dos dados a um modelo de regressão multinível: o Coeficiente de Correlação Intraclasse. Caso não seja possível o ajuste dos dados pelo modelo multinível devemos pesquisar outros procedimentos estatísticos que melhor se adequem aos dados.

Neste trabalho, procuramos modelar os dados de avaliação de desempenho dos alunos das escolas secundárias da cidade do Porto Novo – Santo Antão – Cabo Verde dos anos letivos 2016 a 2019, tendo como variáveis dependentes as classificações finais nas disciplinas de Matemática e Português. Os dados não suportaram estatisticamente um modelo multinível, pelo que decidimos recorrer a regressão múltipla, considerando a Nota Final à Matemática e à Português como variáveis respostas. O estudo foi conduzido com o apoio do software IBM SPSS (*Satistical Package for Social Science*) Versão 25.

Palavras-Chave: Desempenho escolar, modelos multiníveis ou hierárquicos, modelos de regressão múltipla.

ABSTRAT

The nature, the quantity of variables involved in the learning process, the way these variables are grouped: the students, the classroom, the teacher, the school, the family, etc requires us that any study applied in this area should not neglect this organizational hierarchy, for if it happens, the conclusions are likely to be devoid of scientific rigour.

In this aspect, the multilevel regression models that statistically bring this added value appear, which is the possibility of analysing data that have a hierarchical organization, which in education is verified in a very obvious way, since we have students (with their own characteristics), grouped in classes (another hierarchical level), which in turn are grouped in classes, which also belong to school groupings. Despite this clear hierarchical organization, there is a statistical rule that will allow us to proceed or not with the adjustment of data to a multilevel regression model: the Intraclass Correlation Coefficient. In case it is not possible the adjustment of the data by the multilevel model we must research other procedures that better fit the data.

In this work, we try to model the performance evaluation data of the students of the secondary schools in Porto Novo City - Santo Antão - Cape Verde from the school year 2016 to 2019, having as dependent variables the final classifications of the Mathematics and Portuguese subjects. The data did not statistically support a multilevel model, so we decided to use multiple regression, considering the Final Grade to Mathematics and Portuguese as the final response variable. The study was conducted with the support of IMB SPSS (Statistical Package for Social Science) Version 25.

Key words: School performance, Multilevel or Hierarchical Models, Multiple Regression Models.

AGRADECIMENTOS

Um agradecimento especial à Professora Doutora Maria do Rosário Duarte Olaia Ramos pela disponibilidade demonstrada em acompanhar-me nesta caminhada, pelas orientações e pela boa vontade que demonstrou ao longo deste trabalho.

Um agradecimento também especial à Delegação Escolar da cidade do Porto Novo, na pessoa da Delegada Escolar que disponibilizou a base de dados, ferramenta importante neste estudo.

Aos meus filhos, Ézio Marónio e Aldson Muriel, por terem sido um dos focos neste percurso, pelas horas que este trabalho lhes impediram de estarem mais pertos.

A minha esposa e mãe dos meus filhos, Osvaldina Ramos, um muito obrigado pela força e palavras de motivação, pelos gestos que sempre demonstrou.

Aos meus colegas de profissão, Professores Camilo Rodrigues, Eneida Carvalho e Maria de Fátima Reis pela disponibilidade que demonstraram em ajudar neste trabalho.

Às pessoas que moldaram o que hoje sou: minhas mães Maria das Dores, Margarida Gomes (Falecida) e Filomena Delgado e o meu pai António Pires (Falecido).

ÍNDICE

RUÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

SIMBOLOGIAS E NOTAÇÕES

ANOVA – Análise de Variância

CIC – Coeficiente de Correlação Intraclasse

EBI – Ensino Básico Integrado

ESASP – Escola Secundária António Silva Pinto

ETJV – Escola Técnica João Varela

NFMat – Nota Final à Matemática

NFPort – Nota Final à Português

SPSS – Statistical Package for Social Science

INTRODUÇÃO

Se em determinados estudos científicos, cujo objetivo é inferir sobre informações observadas de um dado fenómeno, conseguirmos definir no agrupamento destes dados uma certa hierarquia organizacional, estaremos perante a um contexto que nos impele a levar tal caraterística em conta na hora de formular alguma ilação sobre os dados. Ignorar esta caraterística, ou seja, se não levarmos em conta esta forma da disposição dos dados, em hierarquias, isto pode acarretar que as conclusões sobre os dados sejam desprovidas de rigor científico.

Informações com esta particularidade na sua estrutura exigem da parte do investigador uma abordagem diferenciada e adequada. Neste sentido, os modelos de regressão multinível (modelos de regressão hierárquicos) aparecem como a solução no tratamento dos mesmos, pois nesta disposição hierárquica é possível encontrar informações cujas caraterísticas se identificam num determinado grupo, o que não exclui a possibilidade destes se agruparem em outros grupos, ou seja, há uma variabilidade de dados num nível que pode ser explicado pela variabilidade de dados num outro nível.

Na educação nem sempre é possível isolar um determinado fenómeno para estudá-lo sem levar em conta o meio que o rodeia, portanto faz mais sentido, aceitar a hierarquia das observações onde determinada variável num nível pode ser afetada por uma outra num outro nível. Para Bergamo (2002), se o objetivo é estudar não apenas um determinado evento, mas estudar tudo o que o envolve, é possível ajustar um modelo que leve em conta toda a variabilidade entre os experimentos e incorpore os diferentes aspetos de cada um deles. Neste sentido, os modelos multiníveis representam uma extensão do modelo de regressão clássica quando as variáveis são dispostas em vários níveis de agregação. Esta técnica é um tipo de análise de regressão que, simultaneamente tem em consideração múltiplos níveis de agregação.

Apesar da necessidade que o investigador possa achar sobre o ajuste dos dados a um modelo de regressão multinível, este deve ser tomado com bases científicas para que possíveis conclusões sejam também cientificamente validadas. Nesta perspetiva, o

Coeficiente de Correlação Intraclasse é a estatística que fornece ao investigar tal suporte, pois é ela que avalia a homogeneidade ou não das informações entre as classes. Num contexto educacional, querendo analisar a necessidade de ajustar os dados do rendimento escolar, por exemplo na disciplina de Matemática a um modelo de regressão multinível, é este coeficiente que nos fornece a magnitude da influência da escola sobre tal rendimento. Caso houver significância estatística neste parâmetro ou coeficiente, estaremos perante a validade na prossecução pelo ajuste dos dados a um modelo multinível, ou seja, há na escola determinadas variáveis que exercem influência nas avaliações dos alunos e, caso contrário, não havendo significância o Coeficiente de Correlação Intraclasse ou a estatística for próximo de zero, verifica-se a homogeneidade no rendimento entre as escolas e a recomendação pelo ajuste dos dados ao modelo de regressão multinível não é válida, cabendo ao investigador optar por outras possíveis alternativas, tais como os modelos de regressão linear.

Um dos pilares de desenvolvimento de qualquer sociedade é a educação. Mas para que este desenvolvimento seja efetivo, há que criar políticas educacionais adequadas e ajustadas para dar resposta as demandas que vão surgindo no dia a dia. Nesta ótica, há que ter consciência sobre a necessidade de periodicamente avaliar as políticas educacionais, e uma das formas de fazer esta avaliação é, na nossa opinião, estudar as informações referentes ao sistema educativo numa visão abrangente.

Sendo uma primeira aventura em trabalhos desta natureza, aplicados a dados do sistema nacional de educação Cabo-verdiano, pretendemos analisar a possibilidade de um possível ajustamento das informações educacionais das duas escolas secundárias da cidade do Porto Novo – ilha de Santo Antão a um modelo de regressão multinível, tentando perceber até que ponto o contexto da organização hierárquica pode influenciar a classificação desses alunos nas disciplinas de Matemática e Português, aplicados a dados dos anos letivos 2016–2017 a 2018–2019, excetuando os alunos da área de Humanística que, no 3º ciclo, não possuem a disciplina de Matemática no programa escolar.

Este trabalho é uma tentativa de definir uma relação entre o resultado da avaliação de desempenho dos alunos com algumas variáveis envolvidas no processo e, assim tentar ajudar na procura de respostas a inquietações enquanto professor e agente do sistema educativo, usando informações dos alunos respeitantes aos anos letivos 2016 a 2019 nas duas escolas secundárias da cidade do Porto Novo, ilha de Santo Antão, Cabo Verde. Para tal, estudaremos a possibilidade de ajustar os dados a um modelo de regressão multinível com as variáveis respostas Nota Final à Matemática e à Português e as explicativas sendo: sexo, idade, escola que frequenta, distância da casa à escola, nível de instrução do encarregado de educação, número de reprovações no sistema de ensino e repetente no ano que estuda.

No que concerne à estrutura, esta tese possui três capítulos, sendo que no Capítulo I iniciamos com um enquadramento teórico sobre os modelos de regressão linear simples e múltipla, ou modelos com somente um nível. Ainda no mesmo capítulo introduzimos uma revisão teórica sobre os modelos de regressão multinível, focando na construção de um modelo com dois níveis e apresentando exemplos de aplicações práticas da regressão multinível, visto que o campo da aplicabilidade dos métodos de regressão multinível é vasto. O Capítulo II é dedicado a parte inferencial teórica iniciando num primeiro momento com uma abordagem dos métodos de estimação para os modelos de regressão multinível, seguido dos testes de hipóteses para os efeitos fixos e os componentes da variância dos modelos de regressão multiníveis bem como de outros testes que se figuraram, para nós, como sendo relevantes neste trabalho. No Capítulo III que é essencialmente prático, usando o software estatístico o SPSS Versão 25.0, iniciamos com algumas considerações sobre a nossa base de dados, seguido de uma apresentação da análise exploratória dos mesmos, de possíveis relações entre as possíveis variáveis respostas e explicativas e terminando com o ajuste dos dados a modelos de regressão multinível e linear múltipla.

CAPÍTULO I: MODELOS DE REGRESSÃO

1.1. Modelos de Regressão Linear ou modelos com um nível

Quantas vezes, e de uma forma informal e intuitiva, vimo-nos perante a necessidade de relacionar duas variáveis, indagando sobre qual a influência que possivelmente uma exercerá sobre a outra. Ora, querendo analisar somente um acontecimento (que seria para nós a variável resposta), descurando os possíveis fatores envolventes, recomenda-se a aplicação dum modelo de regressão linear simples.

Portanto, podemos dizer que a regressão linear, (se possuir somente uma variável explicativa, trata-se de um modelo linear simples, caso contrário, será um modelo linear múltiplo) surge como uma ferramenta estatística para dar resposta à necessidade de se relacionar um conjunto de observações, consideradas variáveis explicativas ou covariáveis com um evento, para nós a variável resposta.

O modelo de regressão linear, de acordo com Castro (2015, *apud* Matos, 1995), pode ter dois objetivos diferentes: explicativo ou preditivo. O explicativo passa por demonstrar uma relação matemática de aparente causa-efeito entre uma variável resposta e uma ou mais variáveis explicativas, e o preditivo pretende estimar o comportamento da variável resposta por meio de uma combinação linear de parâmetros que dependem do valor das variáveis explicativas, de forma a futuramente, a partir da observação destas prever o valor da variável resposta, sem que seja necessário medi-la.

1.1.1. Método e descrição

No modelo de regressão linear simples, a relação entre a variável resposta e a variável explicativa, geralmente é definida como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{com} \quad i = 1,2,\ldots,n$$

em que:

y_i representa a variável resposta observada no *i*-ésimo indivíduo;

β_0 ordenada na origem, o valor esperado da variável resposta y_i para $x_i = 0$;

β_1 a mudança esperada em y_i quando x_i aumenta uma unidade;

x_i variável explicativa do *i*–ésimo;

ε_i é o erro associado ao *i*-ésimo indivíduo $\left(\varepsilon_i \sim N\left(0,\sigma^2\right)\right)$ e $\varepsilon_i^{'}$s independentes, com as suposições:

$E\left(\varepsilon_i\right) = 0$;

$Var\left(\varepsilon_i\right) = \sigma^2$;

$Cov\left(\varepsilon_i, \varepsilon_j\right) = 0$ para $i \neq j$

Decidindo por não incluir no nosso modelo nenhuma variável explicativa, estaremos perante um modelo considerado nulo, definido como:

$$y_i = \beta_0 + \varepsilon_i \quad \text{com} \quad i = 1,2,\ldots,n$$

Neste modelo, a variável resposta é definida exclusivamente pela sua média β_0, ou constante do modelo.

Caso não for nenhum dos casos anteriormente propostos, ou seja, se o modelo de regressão possuir mais que uma variável explicativa, esta deverá ser escrita como:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon_i$$

em que:

cada uma das k variáveis explicativas, são observadas n vezes;

$Y = \left[y_1,\ldots,y_n\right]^T$ é o vetor da variável resposta

β_0 é a ordenada na origem para cada variável resposta;

β_i é o valor para o parâmetro de regressão associado à *i*-ésima variável explicativa;

$X_i = \left[x_{1i},\ldots,x_{ni}\right]^T$, com $i = 1,2,\ldots,k$ o vetor de valores da *i*-ésima variável explicativa para os n indivíduos;

$\varepsilon = \left[\varepsilon_1,\ldots,\varepsilon_n\right]^T$ é o vetor dos erros associados aos n indivíduos, com $\varepsilon_i \sim N\left(0,\sigma^2\right)$;

Sob as mesmas condições é possível levar a cabo a realização de vários experimentos e para cada um deles obteremos sua respetiva reta de regressão para explicar a variabilidade do experimento. Nesta ótica, o experimento que apresentar maior inclinação, nos leva a concluir que a variável explicativa é mais forte tornando-a mais preditivo em relação à outra e, o com menor ordenada na origem é o que proporciona um pior resultado na origem.

Devido a impossibilidade de, na maioria das vezes, termos acesso ao universo das informações em qualquer trabalho estatístico, recorrendo para tal a amostras representativas, temos a necessidade de trabalharmos com estimativas dos parâmetros para um modelo de regressão. No tópico seguinte, este tema será abordado bem como a necessidade de se produzir boas estimativas para os parâmetros.

1.1.2. Pressupostos e Estimação

Em qualquer ajuste de um modelo de regressão, o primeiro passo consiste na estimação, pois o real valor de um determinado parâmetro nem sempre é conhecido.

O ajuste de um modelo linear conduz a uma procura de uma combinação de elementos da amostra, que é mais fácil nestes modelos comparativamente aos modelos não

lineares, denominado por estimador e pretende que a sua utilização no modelo permita estimar corretamente a variável resposta (Castro, 2015).

Dado que há uma dependência linear entre os parâmetros de um modelo de regressão linear, então existe a possibilidade desta relação poder ser descrita através de várias retas. Portanto, a questão que o investigador se depara é o de escolher a melhor reta que se ajusta aos dados observados, por forma a minimizar o erro inerente a qualquer processo de estimação.

Este método, que procura definir a melhor reta que se ajusta aos dados minimizando o erro, é denominado por método dos mínimos quadrados, pois trata-se de escolher uma reta entre todas as possíveis retas, mas que minimize a diferença entre os pontos observados e esta mesma reta, ou seja, entre o observado e o estimado.

A independência entre as observações e a ausência de multicolinearidade entre as variáveis explicativas são as duas premissas que norteiam os modelos de regressão linear, sendo que o primeiro é apontado por muitos autores como sendo o ponto fraco dos modelos, pois ignora o meio envolvente. Este acarreta outro problema se pensarmos na possibilidade de, com evidências estatísticas, rejeitarmos a premissa de independências das observações, visto que tal condição influencia na subestimação dos erros padrão dos coeficientes de regressão.

A variabilidade de um experimento pode ser explicada por um conjunto de retas de regressão que podem ser originadas da realização de um experimento várias vezes sob as mesmas condições. Quando o objetivo é estudar não apenas um determinado evento, mas tudo o que o envolve, é necessário ajustar um modelo que leve em conta toda a variabilidade entre os experimentos e incorpore os diferentes aspetos de cada um deles. Considerar os dados de acordo com uma estrutura hierárquica leva em conta tal tipo de análise (Bergamo, 2002).

1.2. Modelos de Regressão Multinível

Uma das desvantagens que podemos apontar aos modelos de regressão tradicionais é o fato de que perante um estudo que comporta em si dados agrupados em hierarquias serem incapazes de levar em conta tal hierarquização das observações, o que pode levar o investigador a falsas conclusões. Neste sentido os modelos multiníveis aparecem como a solução para suplantar tal limitação, pois além de levar em conta a forma de agrupamento das observações, permite que em cada nível seja definida uma equação que serão posteriormente combinadas num único modelo que melhor se ajuste aos dados.

Pela essência dos modelos multiníveis, podemos distinguir duas partes: a fixa e a aleatória. A parte fixa é aquela que é comum a todos os grupos e a aleatória é a que expõe a especificidade de cada grupo sendo estimada pela variabilidade nos diferentes níveis. Nos modelos de regressão multinível, os coeficientes do primeiro nível são tratados como aleatórios no segundo nível.

1.2.1. Modelos de regressão multinível com dois níveis

A Figura 1 mostra-nos uma ilustração de um modelo multinível com dois níveis, de onde se constata que as observações são classificadas de acordo com os níveis, ou seja, unidades de nível 1 e unidades de nível 2, ocorrendo, portanto, n_j unidades de nível 1 para cada unidade j $\left(j=1,2,\ldots,J\right)$ do nível 2.

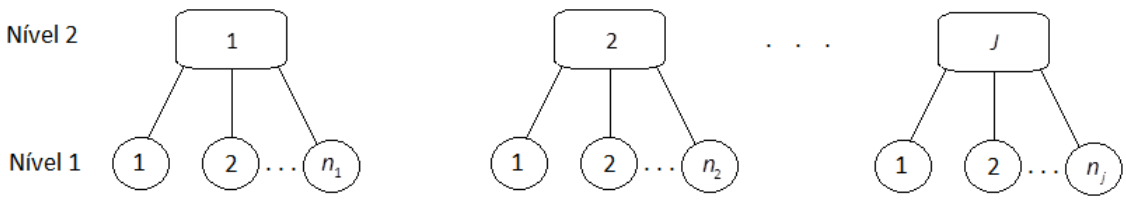


Figura 1: Ilustração de um modelo multinível com dois níveis

Nesta visão desenvolve-se os modelos para o nível 1 separadamente quando se leva em conta uma possível variação da ordenada na origem e as inclinações, para cada unidade

j do nível 2. Assim, e de acordo com Cruz (2010), os modelos no nível 1 estão relacionados através de um modelo de nível 2, no qual os coeficientes de regressão do nível 1 se "incorporam" num 2º nível de variáveis explicativas, e assim sucessivamente para os diferentes níveis. Ainda, e segundo Pinheiro (2005), este modelo permite que diferentes níveis sejam especificados em modelos separados e depois combinados num único modelo.

Por norma, num modelo de regressão deste tipo, onde há n_j unidades do nível 1, para cada uma das j unidades do nível 2, com $j = 1,2,\ldots,J$, o modelo será, no nível 1, representado na forma:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \quad \text{com} \quad i = 1,2,\ldots,n_j \text{ e } j = 1,2,\ldots,J$$

em que:

y_{ij} é a variável resposta do *i*-ésimo indivíduo do nível 1, do *j*-ésimo grupo;

x_{ij} é a variável explicativa referente ao *i*-ésimo indivíduo do nível 1, agrupadas para o *j*-ésimo grupo do nível 2;

β_{0j} é a ordenada na origem para o *j*-ésimo grupo;

β_{1j} é a inclinação associada à variável explicativa x_{ij}, da *i*–ésima unidade do nível 1 para o *j*-ésimo grupo;

ε_{ij} é o erro aleatório associado à *i*–ésima unidade do nível 1, do *j*–ésimo grupo do nível 2, com as suposições:

$\varepsilon_{ij} \sim N\left(0,\sigma^2\right)$;

ε_{ij} são independentes.

Sabendo que quando $x_{ij} = 0$, a ordenada na origem β_{0j} é o valor esperado da variável resposta no y_{ij} do *j*–ésimo grupo, e que teremos para o nível 2 J modelos, em que para cada um há aleatoriamente diferentes ordenadas na origem aleatórios (β_{0j}) e inclinações (β_{1j}) para $j = 1,2,\ldots,J$, poderemos então modelar esses coeficientes como:

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

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

onde

γ_{00} é o valor esperado das ordenadas nas origens nas unidades do nível 2;

γ_{10} é o valor esperado das inclinações na população do nível 2;

u_{0j} é o efeito aleatório na ordenada na origem β_{0j} da *j*–ésima unidade do nível 2;

u_{1j} é o efeito aleatório na inclinação β_{1j} da *j*–ésima unidade do nível 2;

com as suposições:

$u_{0j} \sim N(0, \tau_{00})$;

$u_{1j} \sim N(0, \tau_{11})$;

todos os u_{0j} e u_{1j} são independentes entre si e também de cada ε_{ij} correspondente. Logo podemos afirmar que:

$$\beta_{0j} \sim N(\gamma_{00}, \tau_{00}) \text{ e } \beta_{1j} \sim N(\gamma_{10}, \tau_{11})$$

τ_{00} é a variância populacional das ordenadas nas origens;

τ_{11} é a variância populacional das inclinações;

τ_{01} é a covariância entre β_{0j} e β_{1j}.

Poderemos ter a necessidade de incluir variáveis explicativas no nível 2, com o fito de melhor explicarmos a variabilidade nesse mesmo nível. Para isso, consideremos a variável w_j, e as equações:

$$\beta_{0j} = \gamma_{00} + u_{0j} \text{ e } \beta_{1j} = \gamma_{10} + u_{1j}$$

que combinadas, poderão ser reescritas como

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

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

em que:

β_{0j} é a ordenada na origem para a *j*-ésima unidade do nível 2;

β_{1j} é a inclinação para a *j*-ésima unidade do nível 2;

γ_{00} é o valor esperado da ordenada na origem para w_j;

γ_{10} é o valor esperado das inclinações w_j;

γ_{01} é o coeficiente de regressão associado à variável explicativa w_j do nível 2 relativo à ordenada na origem;

γ_{11} é o coeficiente associado à variável explicativa w_j do nível 2, referente a inclinação do nível 1;

u_{0j} é o efeito aleatório da *j*-ésima unidade do nível 2 sobre a ordenada na origem para w_j;

u_{1j} é o efeito aleatório da *j*-ésima unidade do nível 2 sobre a inclinação para w_j;

τ_{00} é a variância populacional das ordenadas na origem corrigida pela varável w_j;

τ_{11} é a variância populacional das inclinações corrigida pela varável w_j;

τ_{01} é a covariância entre β_{0j} e β_{1j}, com as suposições:

$u_{0j} \sim N(0, \tau_{00})$;

$u_{1j} \sim N(0, \tau_{11})$;

todos os v_{0j} e v_{1j} são independentes entre si e também de cada ε_{ij} correspondente;

$Co(u_{0j}, u_{1j}) = \tau_{01}$.

Sendo que τ_{00}, τ_{11} e τ_{01} são componentes da variância, podemos relacionar as equações

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j} \quad \text{e} \quad \beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

com

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$$

e obter um modelo, abarcando variáveis explicativas de nível 1 (x_{ij}) e nível 2 (w_j), escrita como

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + u_{0j} + (\gamma_{10} + \gamma_{11} w_j + u_{1j}) x_{ij} + \varepsilon_{ij}$$

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij} + u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}$$

Além das variáveis explicativas de nível 1 e 2, este modelo envolve ainda:

um termo entre os níveis $w_j x_{ij}$;

um termo complexo $u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}$, ou seja, um erro complexo;

1.2.2. Construção do modelo de regressão com dois níveis

A construção de um modelo final de regressão multinível é dividida em várias etapas, sendo que subjacente a cada etapa está a possibilidade de obtermos vários outros modelos. Neste trabalho, utilizamos o método proposto por Hox, que consiste em elaborar um modelo de regressão multinível em cinco passos (Laros & Marciano, 2008 *apud* Hox, 2002).

1.2.1. Modelo Nulo

Sendo o primeiro passo na construção de um modelo de regressão multinível, é denominado também de modelo vazio ou modelo ANOVA com efeitos aleatórios, visto não envolver variáveis explicativas em nenhum dos níveis mas sim efeitos aleatórios, ou seja, o efeito dos grupos u_{0j} são interpretados como aleatórios.

Com toda a variação nos dados devida à variabilidade dos resíduos não explicada pelo modelo, temos que a equação do nível 1 dada por:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

onde y_{ij} representa o desempenho do aluno i na escola j

e a de nível 2 definida como

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

Portanto a equação do modelo nulo, que modela somente a ordenada na origem é escrita como:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}$$

onde

β_{0j} é a média da variável resposta na *j*-ésima escola (parâmetro a ser estimado)

γ_{00} é a ordenada na origem da regressão

u_{0j} é o resíduo de nível 2, ou seja, o efeito aleatório associado ao nível 2 (estima-se a variância dos resíduos u_{0j})

$u_{ij} \sim N(0, \tau_{00})$, u_{ij} independentes

ε_{ij} é o resíduo de nível 1, ou seja, o efeito aleatório associado ao nível 1 (estima-se a variância dos resíduos ε_{ij})

$\varepsilon_{ij} \sim N(0, \sigma^2)$, ε_{ij} independentes

Segundo Castro (2015, *apud* Fernandes, 2005), este modelo é composto por duas partes: a parte fixa e a parte aleatória. A parte fixa passa por modelar o efeito da média total sobre as observações e a parte aleatória decompõe a variância total de y pelos níveis.

A decomposição da variância da variável resposta nos dois níveis é a primeira consideração que se deve ter numa análise de regressão multinível. Este não pretende explicar nenhuma variabilidade, mas sim tão-somente decompor a variância em duas componentes independentes: σ^2_{u0} (variância dos resíduos σ_{u0} do nível 2) e (variância dos resíduos $\sigma_{\varepsilon 0}$ do nível 1), a partir da qual é possível estimar o coeficiente de correlação intraclasse (CIC), dado pela equação:

$$\rho = \frac{\sigma^2_{\text{entre grupos}}}{\sigma^2_{\text{entre grupos}} + \sigma^2_{\text{dentro dos grupos}}} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{\varepsilon 0}}$$

Relacionado com o nosso estudo, o CIC define a percentagem da variância total explicada pelo fator escola, ou seja, fornece-nos o grau de comparação entre alunos de escolas diferentes e alunos da mesma escola.

No caso de $\rho = 1$, a conclusão que se pode tirar é a de que a variabilidade nas classificações dos alunos deve-se à diferença entre as instituições escolares, descurando as caraterísticas individuais de cada aluno. No outro extremo, na possibilidade de $\rho = 0$,

não se justifica o estudo usando a regressão multinível, visto que toda a variabilidade nas classificações situa-se no nível 1, ou seja, deve-se única e exclusivamente às caraterísticas individuais dos alunos. De acordo com Cruz (2010, *apud* Merlo, 2005) caso um valor pequeno ocorrer para a CIC, isto não impede a existência de associações significativas entre as variáveis no nível 1 e nível 2.

Contrariamente ao modelo nulo, os modelos que abordaremos de seguida conhecidos como modelos de regressão de componentes da variância, são elaborados mediante a introdução das variáveis explicativas tanto de nível 1 como de nível 2.

1.2.2. Modelos de regressão de componentes da variância

Uma vez findada a construção do modelo vazio, inicia-se a construção do modelo com as variáveis explicativas. Diferentemente do modelo nulo, onde se estima o efeito do fator previamente estabelecido (escola) desconsiderando variáveis explicativas dos níveis, nesta fase há a incorporação dos mesmos nos diferentes níveis.

Partindo da assunção que a ordenada na origem varia entre os níveis mas considerando os coeficientes de regressão como fixos, esses modelos são considerados como componentes da variância pelo fato de decomporem a variância da ordenada na origem em, para cada nível hierárquico, diferentes componentes de variância.

Modelo ANCOVA *one–way* com efeitos aleatórios

Considerando a ordenada na origem como aleatória, a introdução das variáveis explicativas de nível 1 consideradas fixas (não variando de grupo para grupo) é feita nesta fase.

Dado que se pretende estimar a contribuição de cada variável explicativa sobre o modelo, e caso fosse somente no nível 1, a equação seria assim definida, com:

uma variável explicativa x_{ij},

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + \varepsilon_{ij}$$

k variáveis explicativas x_{ij},

$$y_{ij} = \gamma_{00} + \gamma_{k0} x_{kij} + u_{0j} + \varepsilon_{ij}$$

Neste modelo, é possível estimar a proporção da variância explicada no nível 1 pelo modelo. De acordo com Laros & Marciano (2008, *apud* Hox, 2002), tal relação é dada por:

$$R_1^2 = \frac{\sigma^2_{\varepsilon|M_0} - \sigma^2_{\varepsilon|M_{COMP}}}{\sigma^2_{\varepsilon|M_0}}$$

onde

$\sigma^2_{\varepsilon|M_0}$ é a variância residual do nível 1 para o modelo nulo (vazio);

$\sigma^2_{\varepsilon|M_{COMP}}$ é a variância residual do nível 1 neste modelo considerado de comparação;

Modelo de regressão de médias como respostas

Neste modelo pretende-se estudar qual o peso que as variáveis explicativas de nível 2 possuem sobre a proporção da variância no nível 1, sendo para tal acrescidas variáveis explicativas do nível 2.

Seja w_j uma variável explicativa do nível 2, sabemos que a equação do nível 1 é dada por:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

A inclusão de w_j no nível 2, permite-nos escrever neste nível como:

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

Portanto com a inclusão de uma variável explicativa, a equação deste modelo será definido por:

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + u_{0j} + \varepsilon_{ij}$$

onde

γ_{00} é a ordenada na origem média dos grupos quando $w_j = 0$;

γ_{01} é a diferença média entre os grupos;

u_{0j} é o efeito aleatório do *j*-ésimo grupo sobre a ordenada na origem quando $w_j = 0$;

ε_{ij} é o resíduo de nível 1, que segue uma distribuição normal com média zero e variância σ^2

A equação acima está definida para somente uma variável explicativa no nível 2 e sem inclusão de variáveis explicativas do nível 1.

Caso considerássemos para cada nível:

uma variável explicativa, teríamos a equação:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} w_j + u_{0j} + \varepsilon_{ij}$$

q variáveis explicativas w_{qj} do nível 2 e k variáveis explicativas x_{ij} do nível 1, a equação final para os modelos de componentes da variância seria assim:

$$y_{ij} = \gamma_{00} + \gamma_{k0} x_{kij} + \gamma_{0q} w_{qj} + u_{0j} + \varepsilon_{ij}$$

À semelhança do modelo *one-way* com efeitos aleatórios, também é possível definir a proporção da variância explicada no nível 2, por este modelo, usando a igualdade:

$$R_2^2 = \frac{\sigma^2_{u0|M_0} - \sigma^2_{u0|M_{COMP}}}{\sigma^2_{u0|M_0}}$$

sendo

$\sigma^2_{\varepsilon|M_0}$ é a variância residual do nível 2 para o modelo nulo (vazio);

$\sigma^2_{\varepsilon|M_{COMP}}$ é a variância residual do nível 2 neste modelo;

1.2.3. Modelo de regressão com coeficientes aleatórios

Caso considerarmos tanto a ordenada na origem como o coeficiente de inclinação como aleatórios (variando de grupo para grupo), estaremos perante um modelo de regressão com coeficientes aleatórios, ou seja, neste modelo pretendemos investigar se os coeficientes de regressão das variáveis explicativas de nível 1 possuem componentes significativos de variância (diferente de zero) entre os grupos de nível 2.

Considerando a variável explicativa x_{ij} do nível 1, então podemos escrever neste nível a equação dada como:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$

com as equações do nível 2, escritas como:

$$\beta_{0j} = \gamma_{00} + u_{0j} \text{ e } \beta_{1j} = \gamma_{10} + u_{1j}$$

sendo

β_{0j} é a ordenada na origem média;

β_{1j} é o coeficiente de inclinação;

γ_{00} é a ordenada média das unidades de nível 2;

u_{0j} é o efeito aleatório associado ao grupo *j*

$u_{0j} \sim N(0, \tau_{00})$ e $u_{0j}^{'} s$ são independentes dos $\varepsilon_{0j}^{'} s$;

γ_{10} é a inclinação média dos grupos;

u_{1j} é o efeito do *j*-ésimo grupo sobre o coeficiente de inclinação;

Da combinação das três equações acima, obtemos o modelo final de regressão com coeficientes aleatórios assim definido:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} = \gamma_{00} + u_{0j} + \left(\gamma_{10} + u_{1j}\right) x_{ij} + \varepsilon_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{1j} x_{ij} + u_{0j} + \varepsilon_{ij}$$

Na equação definida acima, identificamos que a variável resposta é definida em função da:

Parte fixa: $\gamma_{00} + \gamma_{10} x_{ij}$

Parte aleatória: $u_{1j} x_{ij} + u_{0j} + \varepsilon_{ij}$

com as seguintes componentes

u_{0j} é o efeito aleatório do *j*-ésimo grupo sobre a média;

u_{1j} é o efeito aleatório do *j*-ésimo grupo sobre o coeficiente de inclinação;

ε_{ij} é o resíduo aleatório de nível 1

Neste modelo, podemos ainda representar a dispersão dos efeitos aleatórios usando uma matriz de variância e covariância:

$$\text{var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = T$$

com

$\text{var}\left(u_{0j}\right) = \tau_{00}$ a variância incondicional nas ordenadas da origem do nível 1;

$\text{var}\left(u_{1j}\right) = \tau_{11}$ a variância incondicional nas inclinações do nível 1;

$\text{cov}\left(u_{0j}, u_{1j}\right) = \tau_{01}$ a covariância incondicional entre nas ordenadas da origem e as inclinações do nível 1;

Caso considerássemos somente uma variável explicativa em cada nível, equação assumiria o seguinte formato:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} w_j + u_{1j} x_{ij} + u_{0j} + \varepsilon_{ij}$$

e caso fosse k variáveis explicativas x_{kij} do nível 1 e q variáveis explicativas w_{qj} do nível 2, então ela seria redefinida como:

$$y_{ij} = \gamma_{00} + \gamma_{k0} x_{kij} + \gamma_{0q} w_{qj} + u_{kj} x_{kij} + u_{0j} + \varepsilon_{ij}$$

com u_{kj} os resíduos de nível 2 dos coeficientes das k variáveis explicativas x_{kij} de nível 1, onde:

$\gamma_{00} + \gamma_{k0} x_{kij} + \gamma_{0q} w_{qj}$ é a parte fixa ou determinística do modelo

$u_{kj} x_{kij} + u_{0j} + \varepsilon_{ij}$ é a parte aleatória ou estocástica do modelo

1.2.4. Modelo Final com as interações entre níveis

Neste modelo final é adicionado as interações entre níveis, entre as variáveis explicativas de nível 2 com as de nível 1, mas que se revelaram variâncias significativas no modelo de regressão com coeficientes aleatórios.

Com o aparecimento destas interações, o modelo será escrito na sua forma final, dada por:

$$y_{ij} = \gamma_{00} + \gamma_{k0} x_{kij} + \gamma_{0q} w_{qj} + \gamma_{kq} x_{kij} w_{qj} + u_{kj} x_{kij} + u_{0j} + \varepsilon_{0j}$$

Os termos $x_{kij} w_{qj}$ são termos de interação que aparecem no modelo como consequência da variação da inclinação β_{1j} da regressão modelada, considerando a variável ao nível do individuo (x) com a variável ao nível do grupo (w). Uma vez que o termo do erro aleatório u_{1j} é multiplicado pela variável explicativa x_{ij}, o erro total resultante será diferente para diferentes valores de x_{ij}, uma situação que em regressão múltipla comum é chamada de heteroscedástica (Pinheiro, 2005 *apud* Hox, 1995).

1.3. Vantagens em usar modelos de regressão multinível

Segundo Coelho (2017, *apud* Goldstein, 1986), para estudar fenómenos relativos a esse tipo de dados, uma possibilidade é considerar modelos de regressão multinível, que permitem a incorporação de elementos aleatórios associados a cada um dos níveis. Uma diferença marcante nesses modelos em relação aos modelos de regressão linear múltipla é o fato de que leva em consideração a possível correlação existente entre os dados de um mesmo grupo, como nos diferentes níveis de hierarquia.

Em qualquer processo, independentemente de qual área estamos a debruçar, há sempre prós e contras, sendo que cabe ao investigador adotar estratégias para minimizar os erros e assim produzir conclusões mais adequadas à realidade. Tratando-se dos modelos de regressão multinível, as vantagens são imensuravelmente maiores que as desvantagens, tendo sempre em questão que a consequência desta comparação vai depender da forma como estes modelos são utilizados.

- A primeira vantagem que ressalta é a qualidade das estimativas dos coeficientes de regressão e de variação em relação aos modelos tradicionais, pois estes modelos são mais estruturados e flexíveis pelo fato de utilizarem um maior nível de informação da amostra;
- Dado que os modelos de regressão multinível não descuram a hierarquização das observações, incorporam uma maior fiabilidade dos intervalos de confiança, testes de hipóteses, erros padrão;
- Ora um dos pressupostos nos modelos de regressão tradicionais é a independência entre as observações, mas uma vez que a disposição dos dados está em hierarquias, acarretando em si uma possível correlação entre as observações num mesmo grupo como também nos diferentes níveis de hierarquia, esta independência nos modelos de regressão multinível não se verifica;
- Os modelos de regressão multinível permitem uma melhor avaliação da variabilidade das observações nos diferentes níveis, e ajustar um modelo que melhor se adeque à realidade do problema de acordo com a proporção explicada

por cada um deles, ou seja, segundo Castro (2015), possibilita que a variabilidade da variável resposta seja explicada pelas covariáveis incluídas nos diferentes níveis e quantificada de forma que a proporção de variabilidade explicada em cada nível possa ser diretamente comparada, salvaguardando a introdução de erros nos estudos;

- Uma vez que incorpora variáveis explicativas em cada nível, é possível estabelecer para cada uma sua respetiva equação. Este permite análises individuais em cada grupo, o que possibilita mais rigor no estudo da variabilidade entre nível e intra nível, derivado da decomposição nos diversos níveis da variância do erro.

Estudar informações que por natureza estão agrupadas em hierarquias sem levar em conta tal caraterística, nos leva a certas insuficiências processuais, que por sua vez nos levam a conclusões desajustadas da realidade. De acordo com Cruz (2010), se considerarmos um exemplo de aplicação na área da educação (na qual se consideram alunos aninhados em escolas), temos diversos aspetos a ter em conta na utilização da modelação multinível, tais como:

- Heterogeneidade das retas de regressão

  Espera-se que o desempenho médio seja diferente entre as escolas (cada escola terá a sua reta de regressão, distinta uma das outras). Dado que em princípio, cada escola está sujeita a um conjunto variado de fatores que contribuem para explicar as diferenças encontradas, existe um declive diferente nas respetivas retas de regressão. Esta variabilidade não deve ser ignorada, qualquer que seja o tipo de análise.
- Ausência de independência nas observações

  Por norma, indivíduos pertencentes a um mesmo contexto tendem a ser mais semelhantes no seu comportamento do que os que pertencem a contextos diferentes. Os alunos de uma mesma escola tendem a ser similares, em razão do processo de seleção por esta empregue, do ambiente e da história comuns que os alunos compartilham por frequentar a mesma escola. Assim, ao lidar com variáveis em diferentes níveis, o modelo de regressão tradicional pode não ser o

mais adequado, pois não tem em consideração a correlação entre indivíduos associados a um mesmo nível de agregação. Quanto maior for essa correlação maior é a inadequação do modelo de regressão tradicional, ou seja, quanto maior essa dependência, mais a análise multinível se torna necessária.

- Agregação

  Esta inconveniência pode levar a uma perda de informação substancial útil, podendo ser analisada em duas vertentes:

  Caso 1: os dados são agrupados ao nível das escolas (ignorando a variação inter-individual dos alunos). Neste caso, existe uma perda substancial de informação útil, pois a informação acerca dos alunos não é tida em conta na análise (Cruz, 2010 *apud* Hox, 2002; Leeuw, 2005).

  Caso 2: apenas são considerados os dados ao nível das diferenças entre sujeitos (como ocorre em estudos de regressão linear simples ou múltipla), ignorando os efeitos da variação encontrada ao nível das próprias escolas. Neste e citando Cruz (2010, *apud* Hox, 2002; Leeuw, 2005), consideram-se dependências nos dados que, na realidade não existem, pois, os resultados dos alunos são os mesmos, considerando ou não a escola desagregada. Neste caso, não é possível estudar a forma como variam as relações entre as variáveis através dos alunos (Cruz, 2010, *apud* Nezlek, 2001), e por outro lado, as inferências podem ser erróneas, por se considerar que os dados desagregados são independentes entre si (Cruz, 2010, *apud* Fox & Glas, 2002; Hox, 2002).

  Segundo Cruz (2010), é necessário ter em atenção que a forma como os dados são tratados pode induzir-nos a erro, pois tradicionalmente, nas investigações quantitativas, por exemplo na área da educação, analisavam-se conjuntamente as variáveis respeitantes aos alunos e as variáveis respeitantes à turma. Neste caso, havia duas alternativas, ambas erróneas. Por um lado, a falácia atomística (*apud* Hox, 1995) que consiste na recolha de dados de cada sujeito de forma independente e individual e, em seguida, agrupavam-se de forma a tirar conclusões do grupo a que pertenciam, e por outro lado, a falácia ecológica (*apud*. Hill & Rowe, 1996; Hox, 1998; Goldstein, 2003) que consiste em atribuir

incorretamente as caraterísticas do contexto aos sujeitos, visto que se considerava que a unidade de análise fosse a turma ou a escola, as variáveis respeitantes aos alunos incluíam-se nos dados da escola.

1.4. Exemplos de aplicação de modelos de regressão multiníveis

1.4.1. Aplicado à Geografia

Em Janeiro de 2011, Chasco & Lopez, ambas professoras da Universidade Autónoma de Madrid, publicaram o resultado de um trabalho científico denominado ***Modelos Multiníveis: uma aplicação ao modelo de convergência beta***, que tinha por objetivo determinar em que medida os países descentralizados na Europa (a partir de uma perspetiva política e económica), no período entre 1992 a 2006, albergaram um maior crescimento económico nas suas regiões correspondentes do que países com um estado unitário clássico, testando o impacto da descentralização regional sobre o crescimento do rendimento regional.

Segundo os autores, mesmo reconhecendo a importância dos modelos multiníveis na aplicação em ciências sociais, médicas e biológicas, notou-se até a data da publicação do estudo uma certa ignorância da aplicação dos modelos hierárquicos aplicado à ciência regional, principalmente nos casos de modelos de crescimento e de convergência, que pressupõem-se levar em conta não só os fatores regionais, mas também os efeitos nacionais (políticas económicas, legislação, instituições, religião, etc). Para fundamentar a necessidade e importância do estudo, os autores usaram conclusões de Darbi (2005), onde defendia que os países com descentralização regional promovem a inovação e o crescimento económico, porque a descentralização de serviços como a educação ou os cuidados de saúde promovem o crescimento e, por outro lado esta descentralização regional deveria conduzir a igualização fiscal dentro dos países em relação à prestação de serviços públicos mais eficientes.

Utilizaram o método da máxima verosimilhança restrita para estimar os parâmetros num modelo de ajustamento de regressão com dois níveis, sendo a amostra composta por 233 regiões em 20 países da União Europeia. Para o estudo consideraram, tendo como referência o ano de 1991, variáveis explicativas: PIB, percentagem de empregados, formas de governos regionais, desfasamento espacial do PIB, desfasamento espacial da percentagem de empregados, direção este-oeste, direção norte-sul e regimes espaciais e, como variável resposta o PIB médio no período de 1991 a 2006.

Os autores confirmaram a necessidade da análise dos dados pelo ajustamento de um modelo multinível, pois concluíram, pelo coeficiente de correlação intrapaís, que quase 90% da variabilidade total do rendimento do PIB pode ser atribuída às diferenças entre os países. Segundo os autores, não se revelou tão evidente um impacto positivo da descentralização sobre o desenvolvimento económico dos países, pois se analisassem este impacto numa perspetiva como todo, encontrariam indícios não estatisticamente significativos a favor da descentralização e, caso considerassem o agrupamento dos países em centrais e periféricos concluiriam que a descentralização seria importante somente no centro (Finlândia, Suécia e Reino Unido).

1.4.2. Aplicado à Saúde

Fausto *et al*, publicaram em 2008, um exemplo de aplicação dos modelos de regressão multinível denominado de *O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados* que tinha por objetivo demonstrar a aplicação do modelo linear de efeitos mistos na análise de dados de crescimento de crianças. Considera-se que um estudo seja longitudinal quando há uma sequência temporal de duas ou mais observações em cada indivíduo e caso estas observações forem feitas em tempos diferentes, estaremos perante dados desbalanceados.

Na ótica dos autores do trabalho, caso se decidisse pela aplicação do modelo de regressão simples em observações desta natureza, acarretaria que as inferências fossem

menos confiáveis, ou seja, a confiabilidade nas estimativas dos erros padrão dos coeficientes estariam comprometidos. Portanto, devemos encarar a estrutura hierárquica nos dados longitudinais, visto que as medidas repetidas são agrupadas dentro do indivíduo, o que nos leva a supor o pressuposto de independência entre as observações dos indivíduos e que as respeitantes ao individuo possuem a caraterística de dependência com erros correlacionados.

Nesta via de pensamento, o modelo linear de efeitos mistos ou modelo de efeitos aleatórios (que assume que o padrão de crescimento tem a mesma forma funcional para todos os indivíduos, mas que estes podem apresentar comportamento longitudinal diferentes) é o modelo que melhor se adequa aos dados longitudinais desbalanceados em estrutura hierárquica, visto serem estes modelos que permitem que os coeficientes da regressão variem entre indivíduos, possuindo um componente intraindividual (mudança longitudinal descrita pelo modelo de regressão) e outro interindividual (variação na ordenada na origem e inclinação individual), figurando-se este modelo como o mais adequado para observações em que a variação interindividual seja maior que a intraindividual, como é o caso das curvas de crescimento.

A amostra do estudo foi constituída por 139 crianças nascidas de mulheres soropositivas, com 97 lactentes sororrevertores que foram observadas, em relação a altura, 907 vezes e 42 lactentes vivendo com HIV/AIDS, que contribuíram com 411 medidas de altura. Para o modelo, foi considerado as covariáveis: grupo (HIV_+ = 1; sororrevertor = 0), sexo (feminino = 0; masculino = 1) e idade (em meses).

Usando os métodos de máxima verosimilhança completo e restrito na estimação dos parâmetros com β_0 a ordenada na origem; β_1 o efeito associado ao sexo; β_2 o efeito atribuído à infeção pelo HIV; β_3 e β_4 os efeitos atribuídos à idade e ε_{ij} o erro que define a quantidade que a variável resposta desvia da trajetória quadrática, consideraram-se num primeiro momento o modelo de efeitos aleatórios apenas na ordenada na origem definida como $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{3ij}^2 + u_{0j} + \varepsilon_{ij}$ em que y_{ij} é a variável resposta (altura da criança i no momento j). De acordo com os autores, dado que na

análise exploratória verificou-se um comportamento não linear no crescimento da altura, tiveram a necessidade de ajustar o modelo de efeitos aleatórios com um termo quadrático $\left(x_{3ij}^2\right)$ na variável idade. Num segundo momento deu-se a inclusão da inclinação aleatória $\left(u_{3i}x_{3ij}\right)$ no modelo, cujo efeito de u_{3ij} é permitir que a altura de cada criança pudesse deferir, no momento j, da média geral de crescimento.

Com $\beta_1 = 1{,}44$, $\beta_3 = 2{,}98$ e $\beta_4 = -0{,}08$, as variáveis que revelaram estatisticamente significativas para o modelo com apenas a ordenada na origem aleatória foram sexo, a idade e o polinomial da idade. Usando a análise univariada e tendo em conta a forma como foi codificada a variável sexo, concluiu-se que, em média, os meninos são 1,44 mais altos que as meninas. Dos valores de β_3 e β_4, respetivamente, depreenderam que há um aumento da altura com a idade e que a velocidade de crescimento da altura diminui com o aumento da idade. Todas as variáveis foram significativamente associadas com a altura, quando usaram a análise multivariada usando o modelo de efeitos mistos. Observaram pelo valor da ordenada na origem que as crianças nascidas com mães infetadas com HIV era de 48,9 cm, sendo que ao relacionarem este efeito com o sexo viram que este aplicava as crianças do sexo feminino e que para os meninos a média era de 50,45 cm. Porém, após terem somado o efeito da ordenada na origem, do sexo e da variável grupo concluíram: (1) a altura média das meninas não infetadas foi de 48,9 cm e as infetadas 46,8; (2) a altura média dos meninos foram, respetivamente, de 50,45 cm e 47,93, para os não infetados e os infetados.

Os autores do estudo decidiram que o melhor ajuste dos dados ao modelo foi o usando a máxima verosimilhança restrita, pois este apresentou ligeiro aumento na estimativa do desvio padrão da ordenada aleatória.

Aquando da inclusão de efeitos aleatórios na variável idade verificaram uma modificação no valor da estimativa da média para a variável grupo (de $-2{,}52$ para $-1{,}99$) e na variável sexo (de 1,55 para 1,20) o que sugere a existência de interação entre o grupo, sexo e o polinomial da idade. Portanto a inclusão desse efeito aleatório na inclinação da curva de crescimento melhorou o ajuste do modelo (com $p-\text{value} < 0{,}0001$).

Tendo em conta o modelo final, estimaram as equações que descrevem o crescimento (em cm) das crianças e as que estimam a velocidade do crescimento, definidas por:

$$\text{altura}_{\text{crianças sororrevertores}} = 49 + (0{,}58 \times \text{sexo}) + (2{,}98 \times \text{idade}) - (0{,}08 \times \text{idade}^2) + \\ + (0{,}26 \times \text{sexo} \times \text{idade}) - (0{,}009 \times \text{sexo} \times \text{idade}^2)$$

$$\text{altura}_{\text{crianças HIV+}} = 49 + (0{,}58 \times \text{sexo}) - 0{,}88 + (2{,}98 \times \text{idade}) - (0{,}08 \times \text{idade}^2) + \\ + (0{,}26 \times \text{sexo} \times \text{idade}) - (0{,}009 \times \text{sexo} \times \text{idade}^2) - (0{,}46 \times \text{idade}) + \\ + (0{,}02 \times \text{idade}^2)$$

$$\text{velocidade de crescimento}_{\text{meninas sorerrevertores}} = 2{,}98 - (0{,}08 \times \text{idade})$$

$$\text{velocidade de crescimento}_{\text{meninas HIV+}} = 2{,}98 - (0{,}08 \times \text{idade}) - 0{,}46 + (0{,}02 \times \text{idade})$$

$$\text{velocidade de crescimento}_{\text{meninos sorerrevertores}} = 2{,}98 - (0{,}08 \times \text{idade}) + 0{,}26 - (0{,}009 \times \text{idade})$$

$$\text{velocidade de crescimento}_{\text{meninos HIV+}} = 2{,}98 - (0{,}08 \times \text{idade}) + 0{,}26 - (0{,}009 \times \text{idade}) - 0{,}46 + \\ + (0{,}02 \times \text{idade})$$

Segundo os autores, o modelo linear de efeitos mistos revelou-se adequado neste estudo longitudinal com dados desbalanceados e que em relação às curvas de crescimento, permitiu não apenas descrever o comportamento longitudinal da variável resposta, mas também estimar as velocidades de crescimento. Recomendam-se a aplicação deste modelo noutras áreas da saúde, nomeadamente: avaliações de respostas imunológica e de testes laboratoriais em indivíduos infetados ou doentes, avaliação das infeções respiratórias e diarreicas e resposta da fase aguda em crianças acompanhadas por um período de tempo.

1.4.3. Aplicado à Educação

Denominado de NELS: 88 (*National Education Longitudinal Study*) é uma pesquisa sob a responsabilidade do departamento político dos Estados Unidos da América encarregue das políticas educativas e que possuía como principal objetivo identificar o efeito de

variáveis socioeconómicas e escolares sobre o desempenho escolar. Aos 21.580 alunos da oitava série do ensino fundamental (correspondente ao 9º Ano no sistema de ensino Cabo-verdiano), distribuídos por 1.003 escolas foram, além da aplicação de um teste de matemática, colocadas diversas questões relacionadas com a vivência no seio familiar e meio envolvente, relacionamento no meio escolar.

Em 2008, Laros & Marciano usaram o NELS: 88 com o objetivo de mostrar a importância da aplicabilidade dos modelos de regressão multiníveis em dados agrupados por hierarquias em detrimento dos modelos de regressão tradicionais publicando o artigo com o nome *Análise Multinível aplicada aos dados do NELS:88*. Primeiramente retratam a inconveniência em usar a regressão múltipla, defendendo que a violação da independência das observações afigura-se como o maior problema em tratar dados nas ciências sociais e humanas. Segundo os mesmos autores, quando existe uma estrutura hierárquica na população de interesse, a análise multinível é a opção metodologicamente correta para estabelecer as relações entre as variáveis.

Considerando como variável resposta a nota y_{ij} obtida no teste de matemática, propuseram ilustrar um modelo de regressão com dois níveis definido por $y_{ij} = \gamma_{00} + \gamma_{p0}x_{pij} + \gamma_{0q}z_{qj} + \gamma_{pq}z_{qj}x_{pij} + u_{pj}x_{pij} + u_{0j} + \varepsilon_{ij}$, sendo no nível 1 considerado os alunos e a escola definido como nível 2.

Não considerando nenhuma variável explicativa nos dois níveis, o modelo nulo confirmou-lhes a necessidade de se usar a análise através dos modelos multiníveis em detrimento dos modelos tradicionais, pois o modelo mostrou que 26% da variância das classificações no teste de matemática era atribuída ao nível da escola, com a média geral dos alunos no teste de matemática igual a 50,80 pontos. Ainda este modelo mostrou que, com significância comprovada usando o teste de Wald, dentro das escolas havia variabilidade nas classificações ficando em 76,62.

O efeito de todas as variáveis do nível do aluno (nível socioeconómico do aluno, etnia, horas usadas com os trabalhos de casa, escolaridade dos pais e género) que foram introduzidas no primeiro modelo de componentes da variância revelaram-se

significativas. Ainda é de realçar a diminuição de todos os parâmetros em comparação com o modelo nulo: a variância do nível da escola (26,56 para 7,32) em relação ao modelo nulo, fato que os autores defenderam dever-se a desigualdade entre a proporção das variáveis em cada escola; a variância ao nível do aluno também teve uma diminuição para 66,07; o coeficiente de correlação intraclasse também sofreu uma diminuição para 0,10 e a variância explicada no nível do aluno por este modelo a ser definido em 0,138.

A tendência para a descida dos parâmetros estimados manteve-se com a introdução de variáveis do nível da escola. Com a variância explicada, por este modelo, no nível da escola de 0,724; é de realçar que este modelo apresentou como sendo a mais ajustada aos dados do que o modelo nulo, pois a medida do grau de desajuste do modelo (usada para comparar modelos), diminui do modelo nulo (156,966) para o modelo com as variáveis inseridas (152,872). Ainda o modelo com as variáveis dos dois níveis revelou-se mais ajustada em relação ao modelo onde somente as variáveis do nível do aluno eram consideradas com o valor de teste dada, respetivamente, por $\chi^2 = 818{,}80$ e $\chi^2 = 77{,}00$.

Das variáveis de nível do aluno anteriormente inseridas a que apresentou coeficiente aleatório não significativo no modelo dos coeficientes aleatórios foi a escolaridade dos pais. Das que foram significativas, a que registou maior influência foi o género enquanto que, com menos influência o modelo apresenta o nível socioeconómico do aluno. É de realçar também uma melhoria no ajuste deste modelo aos dados comparando com o anterior com o valor de teste $\chi^2 = 7{,}21$.

O último modelo apresentado pelos autores do estudo foi o das interações entre as variáveis do nível do aluno e da escola, de onde se destaca que a única interação com coeficiente significativa foi entre o nível socioeconómico do aluno com etnia, levando os autores a concluírem que o efeito do nível socioeconómico do aluno é diferente entre escolas de etnia predominantemente branca e asiática em comparação com as outras etnias.

Como conclusão, os autores demonstram, pelo valor de χ^2, dada pela comparação entre o Modelo Nulo e o das Interações

$\left(\frac{Deviance_{\text{Modeno Nulo}} - Deviance_{\text{Modelo das interações}}}{\text{Parâmetros}_{\text{Modelo das interações}} - \text{Parâmetros}_{\text{Modelo Nulo}}} \approx 189{,}62 \right)$, que o modelo final melhor se ajustou aos dados do que o inicial. Neste sentido, segundo os mesmos a base de dados, revelou-se adequada a utilização dos modelos de regressão multiníveis.

CAPÍTULO II: INFERÊNCIA ESTATÍSTICA

2.1. Métodos de estimação de parâmetros

Na maior parte das vezes, nem os parâmetros e nem a distribuição da população em estudo são do conhecimento do investigador. Neste sentido a importância de se escolher um melhor estimador de seus parâmetros e de também se estimar a função de distribuição das observações é de extrema importância para se chegar a conclusões concisas.

Um modelo estatístico é uma representação simplificada da realidade, sendo composto por equações que descrevem as relações entre determinadas quantidades aleatórias. Essas equações contêm um conjunto de parâmetros aleatórios, associados aos parâmetros fixos nos diferentes níveis, que são estimados para que o modelo se ajuste o melhor possível à realidade (Castro, 2015).

De acordo com Souza (2015, *apud* Bryk & Raudenbush, 2002), existem três tipos de parâmetros que podem ser estimados num modelo linear hierárquico com dois níveis, são eles: efeitos fixos, coeficientes aleatórios do nível 1 e componentes de variância e covariância.

Havendo outros métodos para a estimação de parâmetros, os métodos de máxima verosimilhança completa e restrita são os mais usados na estimação de parâmetros nos modelos de regressão multinível.

Sendo a função de máxima verosimilhança uma medida relativa de probabilidade de ocorrência de uma amostra específica, o método que incide sobre tal função conduz-nos à maximização desta função na referida amostra levando-nos à produção de estimadores dos parâmetros.

Para que possamos obter estimadores pelo método de máxima verosimilhança é necessário que conheçamos a distribuição da variável em estudo. De acordo com Pinheiro (2005), o método de máxima verosimilhança adota como estimativas dos parâmetros, os valores que maximizam a probabilidade (variável discreta) ou a densidade de probabilidade (variável contínua).

2.1.1. Método da máxima verosimilhança completa

Neste método, os coeficientes de regressão bem como os componentes da variância são incluídos na função de verosimilhança. Segundo Osio (2013, a*pud* Hox, 2010), o método de máxima verosimilhança é um dos mais utilizados para se obter as estimativas dos coeficientes nos modelos multiníveis, pois tem a vantagem de produzir estimativas que são assintoticamente eficientes (para um tamanho de amostra grande, o estimador de máxima verosimilhança é aproximadamente não enviesado com variância mínima) e consistentes (podem ser bastante próximos do verdadeiro valor do parâmetro com alta probabilidade, se muitos dados forem recolhidos).

Neste método não se regista enviesamento nas estimativas dos parâmetros fixos, mas apresenta estimadores enviesados na estimação de parâmetros aleatórios. Segundo Castro (2015), este enviesamento deve-se à perda de graus de liberdade da estimação que resulta da estimação dos parâmetros fixos.

2.1.2. Método da máxima verosimilhança restrita

Na função de máxima verosimilhança restrita os componentes da variância (que são baseados nos resíduos) são incluídos na função e na segunda etapa são estimados os coeficientes da regressão. Afigura-se como solução para o problema que o método de máxima verosimilhança completa acarreta em si, que é o das estimativas enviesadas. Neste quesito, Castro (2015, *apud* Torman, 2011), é de opinião que aquando da estimação dos efeitos aleatórios, o método de máxima verosimilhança restrita retira o viés introduzido pela perda de graus de liberdade que a estimação dos efeitos fixos introduz nos dados. Ainda e de acordo com Castro (2015), este método além de ponderar o ajuste do número de graus de liberdade, é o mais indicado no estudo de dados não equilibrados.

Portanto, este método leva em conta o número de graus de liberdade usado nas estimativas dos efeitos fixos, quando se estima os componentes da variância e

covariância e, de acordo com Souza (2015), ao invés de aperfeiçoar diretamente a verosimilhança das observações, ele aperfeiçoa o integral da verosimilhança dos resíduos.

Independente de qual for o processo de ajuste que estejamos a tentar elaborar, a importância dos testes de hipóteses no mesmo é extrema, visto serem eles responsáveis pela determinação da significância do modelo bem como das estimativas inerentes ao processo. De seguida abordamos os testes de hipóteses usados em modelos de regressão multinível bem como alguns outros testes que revelaram serem importantes tendo em conta a natureza do nosso estudo.

2.2. Testes de hipóteses

2.2.1. Teste de hipótese para os efeitos fixos

É importante que a influência das variáveis explicativas sobre a resposta seja analisada após a estimativa dos parâmetros do modelo de regressão multinível, ou seja, é necessário que a significância de cada parâmetro estimado seja analisada (analisando para tal a significância estatística de cada um dos parâmetros estimados).

Para a análise da significância estatística dos parâmetros fixos do modelo, recomenda-se a aplicação do teste de Wald, com as hipóteses assim definidas:

$$H_0 : \gamma_k = 0$$

versus

$$H_1 : \gamma_k \neq 0$$

onde γ_k é um dos elementos do vetor dos parâmetros fixos.

Sob a hipótese nula, a estatística de teste que é obtida considerando o estimador de máxima verosimilhança restrita, é dada por:

$$W = \frac{\hat{\gamma}_k}{\sqrt{\text{var}\left(\hat{\gamma}_k\right)}}$$

que segue uma distribuição *t* de Student para dados balanceados e para alguns não balanceados.

2.2.2. Teste de hipótese para os componentes da variância

A aplicação do teste de Wald pode ser aplicada para testar a significância dos parâmetros de variância e assim avaliar quais os efeitos aleatórios a serem incluídos no modelo multinível.

Usando o estimador de máxima verosimilhança completa e restrita para o erro padrão, a estatística de teste que é unilateral, é definida como:

$$W = \frac{\hat{\tau}_k}{\sqrt{\text{var}\left(\hat{\tau}_k\right)}}$$

com τ_k é um dos elementos do vetor dos componentes da variância.

A estatística de teste segue assimptoticamente uma distribuição Qui-Quadrado, com os graus de liberdade dados pela diferença entre os parâmetros estimados no modelo especificado na hipótese alternativa e nula.

As hipóteses são definidas como:

$$H_0 : \tau_{uij}^2 = 0$$

versus

$$H_1 : \tau_{uij}^2 \neq 0$$

No processo inferencial estatístico, impõem-se a tomada da decisão entre usar a aplicação de testes paramétricos ou não, sendo que tal é suportada pelas estatísticas fornecidas pelos testes K-S e Levene que analisam, respetivamente, os dois pressupostos para a aplicação dos testes paramétricos: (1) que se verifique a distribuição normal sob a variável dependente e (2) que não haja heterogeneidade nas variâncias caso estejamos a comparar mais que dois grupos. De seguida abordamos os testes K-S e Levene.

2.2.3. Teste *k–S* para testar se uma distribuição amostral é normal

Quando a dimensão da amostra a ser testada for superior a 50, recomenda-se a aplicação deste teste com vista a estudar a normalidade de uma variável em estudo com parâmetros μ e σ .

As hipóteses de estudos podem ser assim definidas:

H_0 : a variável x em estudo possui distribuição normal com parâmetros μ e σ, ou seja, $x \sim N(\mu,\sigma)$;

versus

H_1 : a variável x em estudo não possui distribuição normal com parâmetros μ e σ, ou seja, $x \not\sim N(\mu,\sigma)$.

Cálculo da estatística de teste

Consiste nos seguintes passos (Maroco, 2003, *apud* Steel & Torrie, 1980):

Ordenar, por ordem crescente, as observações da variável em estudo x;

Calcular a frequência acumulada de cada observação;

Calcular o valor absoluto da diferença entre a frequência acumulada de cada uma das observações e a frequência acumulada que essa observação teria se a sua distribuição de probabilidade fosse normal;

Calcular o valor absoluto da diferença entre a frequência acumulada de cada uma das observações e a frequência acumulada da observação anterior.

Terminado este processo, a estatística de teste é dada por:

$$D_{\text{calc}} = \max\left\{\max\left(\left|F\left(x_i\right) - F_0\left(x_i\right)\right|\right); \max\left(\left|F\left(x_{i-1}\right) - F_0\left(x_i\right)\right|\right)\right\}$$

Se $D_{\text{calc}} > D_{\text{tabelado}\,(\alpha)}$, então rejeita-se H_0 ao nível de significância α. Devemos rejeitar H_0 ao nível de significância α, $p-\text{value} < \alpha$. Este valor, $p-\text{value}$, é calculado usando a correção de Lilleforts (Maroco, 2003 *apud* Lilleforts, 1967), pois por norma a inferência é feita com parâmetros estimados (μ e σ) a partir de uma amostra e dos parâmetros populacionais μ e σ.

2.2.4. Teste de Levene para testar a homogeneidade de vareiâncias

Consistindo basicamente na modificação das observações iniciais para que sobre eles seja aplicado o teste ANOVA, o teste de Levene é uma estatística inferencial que tem por objetivo avaliar a igualdade de variâncias em relação a uma variável, mas quando é determinada para pelo menos dois grupos.

A transformação inicial proposta pelo autor apenas utilizava a média de z, devendo ser usada quando a variável $x \sim N\left(\mu,\sigma\right)$, é definida como:

$$z_{ij} = \left|x_{ij} - \bar{x}_i\right|$$

com

$i = 1,2,\ldots,k$ amostras;

$j = 1,2,\ldots,n_i$ é a dimensão de cada uma das amostras;

z_{ij} são as observações transformadas;

x_{ij} representa os dados iniciais;

$\bar{x}_i$ é a média da amostra i das observações iniciais;

Cálculo da estatística de teste

A estatística do teste é definida como (Maroco, 2003 *apud* Levene, 1960):

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} n_i \left(z_i - \bar{z}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(z_{ij} - \bar{z}_i\right)^2}$$

onde

$N = n_1 + n_2 + \cdots + n_k$;

$\bar{z}_i$ é a média de z_i na amostra i ;

$\bar{z}$ é a média da amostra z_i na amostra global;

Como já referido, a fórmula de Levene definida acima somente contemplava a possibilidade das observações atenderem ao quesito de normalidade, portanto usando somente a média. De acordo com Maroco (2003, *apud* Brown & Forsythe, 1974), caso existirem evidências estatísticas que a variável não obedece à distribuição normal, o teste foi expandido para usar a mediana ou ainda a média aparada, atribuindo a partir deste momento uma robustez e potência necessária aos casos onde ocorrem desvios de normalidade.

Com esta expansão, z deve ser calculado como:

$$z_{ij} = \left| x_{ij} - x_i \right|$$

Onde x_i é a mediana da amostra i. Se $p - \text{value} < \alpha$ devemos rejeitar H_0 ao nível de significância α.

Se $W \geq f_{(1-\alpha;(k-1,n-k))}$, então rejeita-se H_0 ao nível de significância α, com as hipóteses assim definidas:

H_0 : as variâncias são iguais para as k populações; o seja, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$;

versus

H_1 : existe pelo menos duas populações i e j com $i \neq j$ onde as variâncias diferem, ou seja, $\exists_{i \neq j} : \sigma_i^2 \neq \sigma_j^2$.

Uma outra possibilidade que este teste nos fornece é de aplicarmos o teste ANOVA sobre os dados transformados $z_{ij} = \left| x_{ij} - \bar{x}_i \right|$, rejeitando H_0 caso a estatística F for significativa.

2.2.5. Teste de independência Qui–Quadrado

Dado que não depende nem da média e nem do desvio padrão, o teste de independência ou associação enquadra na classe dos testes não paramétricos. Possuí como princípio básico a comparação das frequências observadas e esperadas dos dados, nos diferentes grupos de uma variável aleatória, pois se para cada grupo a diferença entre a frequência esperada e a observada for próxima de zero, podemos depreender que há um comportamento similar entre ambas, ou seja, estão associadas.

Para que possamos aplicar o teste Qui-Quadrado com rigor, devemos verificar os seguintes pressupostos: (1) dimensão tem que ser maior que 20; (2) que todas as

frequências esperadas sejam maiores que 1; (3) que pelo menos 80% das frequências esperadas sejam superiores ou iguais a 5 (Maroco, 2003), podendo optar pela agregação de categorias contíguas caso não se verificar (2).

As hipóteses podem ser assim definidas:

> H_0 : as frequências esperadas não são diferentes das observadas, ou seja, as variáveis não estão associadas, isto é, são independentes.

versus

> H_1 : as frequências esperadas e as observadas diferem, ou seja, as variáveis não são independentes.

Cálculo da estatística de teste

Sob H_0, a estatística de teste χ^2, que melhora a sua aproximação quanto maior for a dimensão da amostra, é dada por:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$$

onde

O_{ij} frequência observada na célula $\left(i,j\right)$, $i=1,2,\ldots,k$ e $j=1,2,\ldots,p$;

$E_{ij}=\frac{n_i \times n_j}{n}$, n_i é o total da linha i e n_j o total da linha j;

Caso $\chi^2_{\text{calculado}} > \chi^2_{\left((k-1)(p-1),1-\alpha\right)}$, então ao nível de significância rejeita-se H_0.

2.2.6. Teste Mann–Witney para comparar duas amostras independentes

Sendo a alternativa não paramétrica para o teste paramétrico *t* de Student, o teste de Mann–Witney compara de igualdade de parâmetros de localização de duas amostras independentes (de dimensão n_1 e n_2), ou seja, averigua se dois grupos de dados foram ou não extraídos de uma mesma população, com as hipóteses definidas como:

$$H_0 : \vartheta = k$$

versus

$$H_1 : \vartheta \neq k$$

Cálculo da estatística de teste

Baseando na ordem das observações, começa-se primeiro por definir $D_i = \left| x_i - k \right|$, seguido da ordenação por ordem crescente e eliminando os $D_i = 0$. No caso de ocorrerem empates, atribui-se a média das observações empatadas.

Ela é baseada na soma das ordens das observações dada por:

$$T = \sum_{i=1}^{n} R(i)$$

Se $n_1 \leq 10$ e $n_2 \leq 10$, a estatística do teste é:

$$U = \min(U_1, U_2)$$

com

$$U_1 = n_1 n_2 - \frac{n_1 (n_1 + 1)}{2} - T_1$$

e

$$U_2 = n_1 n_2 - \frac{n_2\left(n_2+1\right)}{2} - T_2$$

Se $U_{\text{calculado}} < U_{\text{crítico}}$ (consultado na tabela de Mann–Witney), rejeitamos H_0 ao nível de significância α

Se $n_1 > 10$ e $n_2 > 10$

E partindo do pressuposto que a medida que os valores das dimensões nos dois grupos aumentam, a distribuição tende para a normal, os valores críticos são tabelados pela distribuição normal reduzida. Se a probabilidade associada ao valor z observado for inferior a α, então rejeita-se H_0 (Oliveira, 2004), com:

$$Z = \frac{T - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2\left(n+1\right)}{12}}} \approx N\left(0,1\right)$$

Caso ocorrerem empates entre duas ou mais observações de ambas as amostras, então devemos efetuar a correção do valor de T, da seguinte forma

$$Z = \frac{T - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{\left(n_1+n_2\right)\left(n_1+n_2-1\right)}\left(\frac{n^3-n}{12} - \sum_{k=1}^{N_{emp}} \frac{e_k^3 - n_k}{12}\right)}} \approx N\left(0,1\right)$$

com

N_{emp} número de empates;

e_k número de observações empatadas para um dado posto. Caso $p-\text{value} < \alpha$, rejeita-se H_0 ao nível de significância α.

2.2.7. Teste kruskal–Wallis: comparar três ou mais grupos independentes

Constituindo opção não paramétrica à ANOVA *one–way* (análise de variância simples), é um teste muito útil para verificar se k grupos aleatórios independentes provêm de populações diferentes.

Admitindo que existe independência não só entre elementos do mesmo grupo, mas sim entre elementos de grupos diferentes, o teste apresenta as hipóteses assim formuladas:

H_0 : as k populações partilham a mesma distribuição

versus

H_1 : pelo menos duas das k populações diferem na distribuição

Cálculo da estatística de teste

Conjuntamente são ordenados todas as observações dos k grupos, atribuindo-lhes a respetiva ordem e, calcular a soma das ordens $\left(R_i\right)$ para cada grupo;

Os valores dos quantis encontram-se tabelados, caso verificar que os $n_i < 5$ e $k = 3$. Neste caso, rejeita-se H_0 ao nível de significância α, se $T > t_{(\text{crítico},\,\alpha)}$.

Caso houver empates, a estatística de teste dada por:

$$T = \frac{12}{n\left(n+1\right)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\left(n+1\right)$$

onde

k é o número de grupos;

n_i é o número de elementos do grupo i;

R_i é a soma dos postos no grupo i;

n é o número total de elementos em todas os grupos combinados;

Se os tamanhos dos k grupos não forem muito pequenos, utiliza-se a distribuição Qui-Quadrado com $gl = k - 1$.

Caso houver empates, será atribuída às observações empatadas a média aritmética dos postos empatados. Dado que o valor da estatística de teste é afetado pelos empates, uma correção sobre o mesmo, consistindo em:

$$T_{\text{Corrigido}} = \frac{\dfrac{12}{n(n+1)}\displaystyle\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)}{1 - \dfrac{\displaystyle\sum_{j=1}^{Aemp}\left(t_j^3 - t_j\right)}{n^3 - n}}$$

onde

$Aemp$ número de amostras com diferentes ordens de empates;

t_j é o número de empates na amostra j.

Com base nas evidências estatísticas, se $T > \chi^2_{k-1}$, então rejeita-se H_0 ao nível de significância α. Caso $p - \text{value} < \alpha$, rejeita-se H_0.

No processo anterior, a rejeição de H_0 só nos permite concluir a existência de diferenças entre grupos, não nos fornecendo evidências estatísticas para decidir em que par de grupos tal diferença ocorre. Neste caso, impõe-nos a necessidade de efetuar o teste de comparações múltiplas de Kruskal–Wallis que baseia em efetuar $C(k,2)$ testes com as hipóteses assim definidas:

H_0 : para qualquer $i \neq j$, as distribuições nos dois grupos é a mesma

versus

H_1 : para qualquer $i \neq j$, as distribuições nos dois grupos difere

Rejeita-se H_0 se:

$$\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right| > t_{\left(n-k;1-\frac{\alpha}{2}\right)} s \sqrt{\frac{n-T-1}{n-k}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

onde

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{n=1}^{k}\sum_{j=1}^{n_i} R\left(x_{ij}\right)^2 - \frac{n\left(n+1\right)^2}{4}\right]}$$

Caso $p-\text{value} < \alpha$, rejeita-se H_0. Neste caso o $p-\text{value}$ é corrigido pelo correção de Bonferroni.

CAPÍTULO III: ANÁLISE DE DADOS DE DESEMPENHO ESCOLAR NAS ESCOLAS SECUNDÁRIAS DA CIDADE DO PORTO NOVO – ILHA DE SANTO ANTÃO

3.1. Considerações sobre o sistema de ensino e a base de dados escolares

A cidade do Porto Novo, ilha de Santo Antão possui à disposição da comunidade educativa, duas instituições de ensino secundário, a Escola Secundária António Silva Pinto e a Escola Técnica João Varela, este último cujo patrono tenha sido o renomado Escritor e Neurocientista Cabo-verdiano que viveu entre 1937 e 2007.

O sistema de Ensino em cabo Verde, tem sido ao longo dos tempos sujeito a alterações com vista a dar respostas as exigências constantes que a educação exige. Em 2010 foi aprovada a revisão da Lei de Bases do Sistema Educativo publicado em 1990 (Lei n.º 103 / III / 1990 de 29 de Dezembro) e revista pela lei n.º 113 / V / 1999. Esta lei tinha por pressuposto atualizar o sistema de ensino cabo-verdiano às demandas exigidas pela época, com vista a adequar o sistema aos desafios do desenvolvimento do país e das perspetivas do futuro, num quadro estrutural mais amplo.

Este decreto preconizava o alargamento da escolaridade obrigatória para oito anos, assentes na universalidade de acesso, observância dos parâmetros de qualidade, equidade e da sustentabilidade do sistema de ensino. Ainda o mesmo diploma previa a necessidade de gradativamente alargar a escolaridade obrigatória até o 12º Ano.

Na altura, o modelo de Ensino Básico compreendia três ciclos sequenciais, sendo o primeiro de quatro anos e o segundo e o terceiro de dois anos cada. Assim o Ensino Secundário passaria a ser de quatro anos, compreendidos de dois ciclos de dois anos cada, sendo que o 1º Ciclo, 9º e 10º ano de escolaridades, constituía um ciclo de consolidação do ensino básico e de orientação vocacional. No 2º Ciclo (11º e 12º ano de escolaridades) havia a possibilidade de se optar por uma via geral ou uma via técnica profissionalizante.

Ainda neste processo de reforma do Ensino Básico e Secundário Cabo-verdiano e com o intuito de adequá-lo com novas políticas educativas, foi publicado em 07 de Dezembro de 2018 o decreto legislativo n.º 13, que além de outros assumia-se a gratuidade do Ensino Básico Obrigatório (EBO) sob a modalidade de ensino público gratuito de 8 anos. Após a publicação deste diploma, deixamos de ter o EBI (Ensino Básico Integrado) passando a

figurar o EBO (Ensino Básico Obrigatório) que compreende dois ciclos de aprendizagens sequências, de quatro anos cada. O Ensino Secundário passou a ter um ciclo único de quatro anos, do 9º ao 12º ano de escolaridade, estruturando-se em duas vias alternativas, via geral e via técnica, podendo os alunos optar por uma delas.

As alterações que o decreto lei, anteriormente referenciado, trouxe não foram levadas em conta neste trabalho, visto que a nossa base de dados de estudo compreende os anos letivos 2016 – 2017 a 2018 – 2019, logo antes da publicação.

Em qualquer estudo, a recolha de informações acarreta consigo um risco de que nem sempre o resultado final é o pretendido. Neste sentido, inicialmente a nossa base de estudos foi alvo de uma filtragem com o intuito de aprimorá-lo, eliminando os casos que foram mal preenchidos ou que possuíam informações relevantes em falta. Neste sentido, cerca de 7,83% dos casos na ESASP foram excluídos enquanto que, na ETJV ficou pelos 5,21%. No cômputo geral esta perda de informação ronda os 6%, passando o nosso universo de 4996 alunos para 4662 alunos nos três anos letivos (Tabela 1).

Tabela 1: Primeira filtragem da base de dados

	ESASP					ETJV					GERAL				
	Válidos		Excluídos		Total	Válidos		Excluídos		Total	Válidos		Excluídos		Total
	N	Perc.	N	Perc.		N	Perc.	N	Perc.		N	Perc.	N	Perc.	
2016 - 2017	317	91,88	28	8,12	345	1283	94,55	74	5,45	1357	1600	94,01	102	5,99	1702
2017 - 2018	305	92,71	24	7,29	329	1234	96,26	48	3,74	1282	1539	95,53	72	4,47	1611
2018 - 2019	308	91,94	27	8,06	335	1215	93,61	83	6,39	1298	1523	93,26	110	6,74	1633
Total	930	92,17	79	7,83	1009	3732	94,79	205	5,21	3937	4662	94,26	284	5,74	4946

Esta base de dados, que é preenchida sempre no final de cada ano letivo pelo professor/diretor de cada turma, abarca informações pessoais dos alunos bem como a avaliação dos mesmos em todas as disciplinas estudadas. Dessas, as consideradas relevantes para nós foram:

Variáveis respostas

- Classificação Final à Matemática – variável resposta contínua que traduz a classificação anual do aluno na disciplina, medida na escala de 0 a 20 valores;
- Classificação Final à Português – variável resposta contínua que traduz a classificação anual do aluno na disciplina, medida na escala de 0 a 20 valores;

Variáveis explicativas

- Sexo do aluno – variável explicativa dicotómica, (Masculino e Feminino);
- Idade do aluno – variável explicativa contínua;
- Ciclo de estudos – variável explicativa categórica, sendo:
    - 1º ciclo – corresponde ao 7º e 8º Anos de escolaridade;
    - 2º ciclo – corresponde ao 9º e 10º Anos de escolaridade;
    - 3º ciclo – corresponde ao 11º e 12º Anos de escolaridade;
- Escola que frequenta – variável explicativa dicotómica, com:
    - ESASP – Escola Secundária António Silva Pinto;
    - ETJV – Escola Técnica João Varela;
- Ano letivo – variável explicativa categórica, sendo:
    - Ano letivo 2016 – 2017;
    - Ano letivo 2017 – 2018;
    - Ano letivo 2018 – 2019;
- Distância da casa à escola – variável explicativa categórica, com as distâncias assim classificadas:
    -]0 – 1] quilómetro;
    -]1 – 3] quilómetros;
    -]3 – 6] quilómetros;
    -]6 – …[quilómetros;
- Nível de instrução do encarregado de educação – variável explicativa categórica, assim classificadas:
    - Sem instrução
    - 1º ciclo do Ensino Básico Integrado, traduzindo no 1º e 2º Anos de escolaridade;
    - 2º ciclo do Ensino Básico Integrado, traduzindo no 3º e 4º Anos de escolaridade;
    - 3º ciclo do Ensino Básico Integrado, traduzindo no 5º e 6º Anos de escolaridade;
    - 1º ciclo do Ensino Secundário;

    - 2º ciclo do Ensino Secundário;
    - 3º ciclo do Ensino Secundário;
    - Curso médio profissionalizante;
    - Curso Superior sem Licenciatura;
    - Licenciatura;
    - Mestrado / Pós-Graduação;
    - Outras
- Repetente no ano de estudos – variável explicativa dicotómica (Repetente ou Não repetente);
- Número de reprovações que possui no sistema de ensino secundário – variável explicativa quantitativa;
- Resultado Final no ano – variável explicativa dicotómica (Aprovado ou não);

Terminadas as considerações sobre a nossa base de dados, passamos de seguida a apresentação da análise exploratória dos mesmos, para que possamos melhor compreender as caraterísticas das variáveis envolventes no estudo, usando o software estatístico SPSS versão 25.0.

3.2. Análise exploratória

A Tabela 2 mostra a distribuição dos alunos que estudaram o ensino secundário nos anos letivos 2016–2017 a 2018–2019, cuja área de estudos possui simultaneamente as disciplinas de Matemática e Português. O ano letivo que absorveu mais alunos foi o de 2016–2017 com 1600 alunos, sendo que dos três anos letivos a que menos teve alunos foi o de 2018–2019, com 1523 alunos.

Tabela 2: Distribuição dos alunos pelos anos letivos

		Frequência	Percentagem	Percentagem acumulativa
Válido	16-17	1600	34,3	34,3
	17-18	1539	33,0	67,3
	18-19	1523	32,7	100,0
	Total	4662	100,0	

De seguida, as Tabelas 3, 4 e 5 permitem-nos analisar as informações referentes aos anos letivos, sexo dos alunos e ciclo de estudos nas duas instituições escolares (como as variáveis ***Ano letivo***, ***Sexo*** e ***Ciclo de Estudos*** estão distribuídas pelas escolas).

Pela análise da Tabela 3, constatamos que dos 4662 alunos, 930 frequentaram a ESASP (escola secundária) que é uma instituição de ensino secundário situada no interior da cidade, enquanto que 3732 foram alunos da ETJV. Dos alunos que estudaram nas escolas secundárias da cidade do Porto Novo em 2016–2017, 19,8% o fizeram na ESASP, o que no universo dos alunos que frequentaram a instituição nos três anos letivos, representa 6,8%. A menor percentagem de alunos que estudaram na ETJV nos três anos letivos ocorreu em 2018–2019, representando 26,1% de todos os alunos dessa instituição nos três anos considerados neste estudo.

Tabela 3: Frequências absolutas e relativas para as variáveis Ano letivo e Escola que frequenta

			Escola que frequenta		
			ESASP	ETJV	Total
Ano letivo	16-17	Contagem	317	1283	1600
		% em Ano letivo	19,8%	80,2%	100,0%
		% em Escola que frequenta	34,1%	34,4%	34,3%
		% do Total	6,8%	27,5%	34,3%
	17-18	Contagem	305	1234	1539
		% em Ano letivo	19,8%	80,2%	100,0%
		% em Escola que frequenta	32,8%	33,1%	33,0%
		% do Total	6,5%	26,5%	33,0%
	18-19	Contagem	308	1215	1523
		% em Ano letivo	20,2%	79,8%	100,0%
		% em Escola que frequenta	33,1%	32,6%	32,7%
		% do Total	6,6%	26,1%	32,7%
Total		Contagem	930	3732	4662
		% em Ano letivo	19,9%	80,1%	100,0%
		% em Escola que frequenta	100,0%	100,0%	100,0%
		% do Total	19,9%	80,1%	100,0%

A Tabela 4 indica-nos que a maior parte dos alunos, 2445, são do sexo feminino. Destes 79,3% frequentaram a ETJV, o que representam cerca 52% dos alunos da escola no período em referência. Em ambas as instituições de ensino secundário, a prevalência de alunos do sexo feminino é maior, sendo que esta diferença é de 8,6% e 4%, para a ESASP e a ETJV, respetivamente.

Tabela 4: Frequências absolutas e relativas para as variáveis Sexo e Escola que frequenta

			Escola que frequenta		
			ESASP	ETJV	Total
Sexo do aluno	Feminino	Contagem	505	1940	2445
		% em Sexo do aluno	20,7%	79,3%	100,0%
		% em Escola que frequenta	54,3%	52,0%	52,4%
		% do Total	10,8%	41,6%	52,4%
	Masculino	Contagem	425	1792	2217
		% em Sexo do aluno	19,2%	80,8%	100,0%
		% em Escola que frequenta	45,7%	48,0%	47,6%
		% do Total	9,1%	38,4%	47,6%
Total		Contagem	930	3732	4662
		% em Sexo do aluno	19,9%	80,1%	100,0%
		% em Escola que frequenta	100,0%	100,0%	100,0%
		% do Total	19,9%	80,1%	100,0%

Analisando os dados pelo Ciclo de estudos (Tabela 5), notamos que aproximadamente 43% dos alunos frequentaram o 1º Ciclo de estudos (7º e 8º Anos de escolaridade). Destes 79,4% o fizeram na ETJV contra os 20,6% que frequentaram a ESASP. Naturalmente, e dado a condições físicas e geográficas das escolas, somente 3,1% do total dos alunos estudaram o 3º ciclo de estudos na ESASP, contrastando com 17,4% do total de todos os alunos que estudaram este último ciclo de estudos como alunos da ETJV. Esta discrepância de números justifica-se pois, além das condições físicas distintas, a ETJV é uma escola do meio urbano e de agrupar as valências gerais e técnico profissionais que são lecionadas no ciclo em questão.

Tabela 5: Frequências absolutas e relativas para as variáveis Ciclo de estudos e Escola que frequenta

			Escola que frequenta		
			ESASP	ETJV	Total
Ciclo de estudos	1º Ciclo	Contagem	412	1586	1998
		% em Ciclo de estudos	20,6%	79,4%	100,0%
		% em Escola que frequenta	44,3%	42,5%	42,9%
		% do Total	8,8%	34,0%	42,9%
	2º Ciclo	Contagem	372	1336	1708
		% em Ciclo de estudos	21,8%	78,2%	100,0%
		% em Escola que frequenta	40,0%	35,8%	36,6%
		% do Total	8,0%	28,7%	36,6%
	3º Ciclo	Contagem	146	810	956
		% em Ciclo de estudos	15,3%	84,7%	100,0%
		% em Escola que frequenta	15,7%	21,7%	20,5%
		% do Total	3,1%	17,4%	20,5%
Total		Contagem	930	3732	4662
		% em Ciclo de estudos	19,9%	80,1%	100,0%
		% em Escola que frequenta	100%	100,0%	100,0%
		% do Total	19,9%	80,1%	100,0%

A tabela acima, nos apresenta que o 3º Ciclo foi frequentado por 956 alunos, o que no todo representa cerca de 20,5% dos alunos, distribuídos por 453 no 11º ano e 503 no 12º Ano. Destes, analisando a Tabela 6, 146 o fizeram na ESASP exclusivamente na via geral pois esta instituição dedica-se especificamente a esta vertente e 810 foram alunos da ETJV. Sendo a ETJV, uma escola onde funciona as duas vias de ensino, 392 estudaram a via geral contra 418 alunos que fizeram seus percursos como estudantes do 3º Ciclo da Via Técnica.

Tabela 6: Frequências absolutas e relativas para as variáveis Via de estudos no 3º ciclo, Anos de estudos no 3º ciclo e Escola que frequenta

			Ano de estudos no 3º Ciclo			Escola que frequenta		
			11º Ano	12º Ano	Total	ESASP	ETJV	Total
Via de estudos no 3º Ciclo	Via Geral	Contagem	255	283	538	146	392	538
		% do Total	5,5%	6,1%	11,5%	3,1%	8,4%	11,5%
	Via Técnica	Contagem	198	220	418	0	418	418
		% do Total	4,2%	4,7%	9,0%	0,0%	9,0%	9,0%
Total		Contagem	453	503	956	146	810	956
		% do Total	9,7%	10,8%	20,5%	3,1%	17,4%	20,5%

As duas escolas encontram-se a extremos quando se compara a distância que os alunos fazem de suas casas à escola. Pela análise da Figura 2, nota-se que 17% dos alunos da ESASP residem a menos de um quilómetro da instituição, enquanto que esta percentagem sobe para 44% quando se considera a ETJV. Já em relação a maior distância, 7% dos alunos da ETJV percorrem diariamente mais que seis quilómetros de suas casas à escola, sendo que esta distância é percorrida por 33% dos alunos da ESASP.

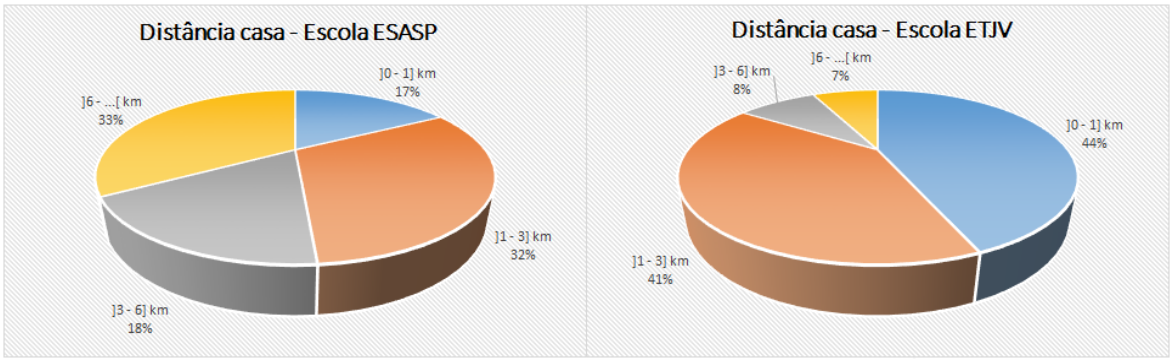


Figura 2: Comparação entre as distâncias casa - escola

A tabela seguinte nos mostra a relação entre o nível de instrução dos Encarregados de Educação em relação às escolas. Nota-se que mais de metade (62,3%) dos encarregados de educação dos alunos da ESASP possuem o Ensino Básico como nível de instrução. A percentagem de encarregados de educação dos alunos da ETJV que possuem este nível de ensino como habilitação literária é de aproximadamente 50%. O ensino secundário foi frequentado por 28,1% dos encarregados de educação de alunos da ETJV sendo que, para a ESASP, esta fatia desce para 16,8%. Menos de 2% dos encarregados de educação da

ESASP frequentaram pelo menos um curso médio, percentagem que sobe neste nível de instrução para os encarregados de educação da ETJV.

Tabela 7: Frequências absolutas e relativas para as variáveis Escola e Nível de instrução do encarregado de educação

			Escola que frequenta		
			ESASP	ETJV	Total
Instrução do encarregado de educação	Sem instrução	Contagem	179	384	563
		% em Escola que frequenta	19,2%	10,3%	12%
	1º Ciclo - EB	Contagem	185	327	512
		% em Escola que frequenta	19,9%	8,8%	11%
	2º Ciclo - EB	Contagem	293	921	1214
		% em Escola que frequenta	31,5%	24,7%	26%
	3º Ciclo - EB	Contagem	101	633	734
		% em Escola que frequenta	10,9%	17,0%	16%
	1º Ciclo - ES	Contagem	32	272	304
		% em Escola que frequenta	3,4%	7,3%	6,5%
	2º Ciclo - ES	Contagem	71	354	425
		% em Escola que frequenta	7,6%	9,5%	9,1%
	3º Ciclo - ES	Contagem	54	420	474
		% em Escola que frequenta	5,8%	11,3%	10%
	Curso Médio	Contagem	7	109	116
		% em Escola que frequenta	0,8%	2,9%	2,5%
	Curso Superior sem Licenciatura	Contagem	0	34	34
		% em Escola que frequenta	0,0%	0,9%	0,7%
	Licenciatura	Contagem	8	214	222
		% em Escola que frequenta	0,9%	5,7%	4,8%
	Mestrado / Pós Graduação	Contagem	0	9	9
		% em Escola que frequenta	0,0%	0,2%	0,2%
	Outras	Contagem	0	55	55
		% em Escola que frequenta	0,0%	1,5%	1,2%
Total		Contagem	930	3732	4662
		% em Escola que frequenta	100,0%	100%	100%

A Figura 3 apresenta-nos que 814 dos alunos são repentes no ano que estuda, o que representa cerca de 17,5%. Regista-se uma percentagem de 18,7% dos alunos que têm uma reprovação no sistema de ensino secundário. Querendo avaliar os alunos que

possuem duas reprovações, nota-se através da figura seguinte que são aproximadamente 8,4%, descendo para cerca de 2% os que apresentam 3 reprovações. Já os que possuem pelo menos 4 reprovações representam menos de 1%.

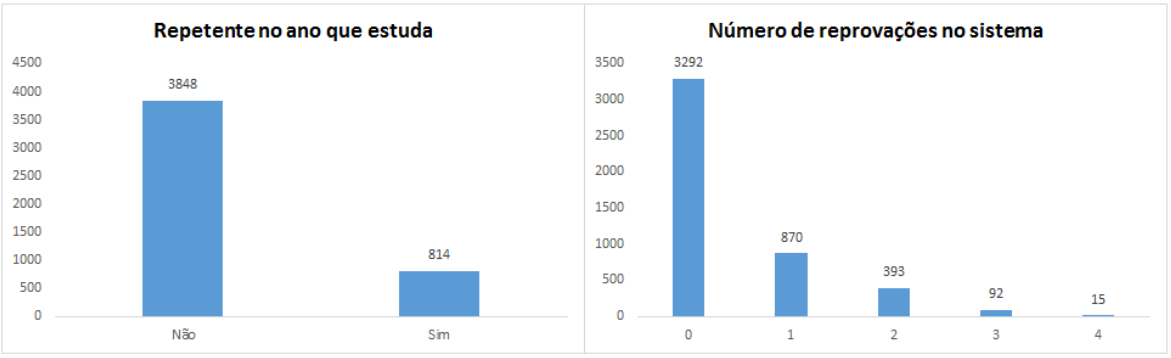


Figura 3: Repetente no ano que estuda e número de reprovações no sistema

De seguida, a Figura 4 e a Tabela 8, apresentam-nos as estatísticas descritivas das variáveis respostas Nota Final a Português e a Matemática, e da variável explicativa Idade. A média da idade dos alunos situa-se nos 14,95 anos sendo que idade mínima registada foi de 12 anos e a máxima de 23 anos. Em relação às classificações nas disciplinas Matemática e Português, ambas registaram a nota máxima permitida, diferenciando na classificação mínima, com respetivamente, 4 e 5 valores para Matemática e Português. A Figura 4, nos mostra que os dados apresentam uma simetria mais acentuada na disciplina de Matemática contrariando com a da disciplina de Português, o que pode ser confirmado pelos coeficientes de simetria fornecidos pela Tabela 8. Pela visualização dos histogramas (Figura 4) e dos coeficientes de achatamento (Tabela 8), concluímos que ambas as disciplinas apresentam curvas afiladas, sendo que este achatamento é mais realçado na de Português.

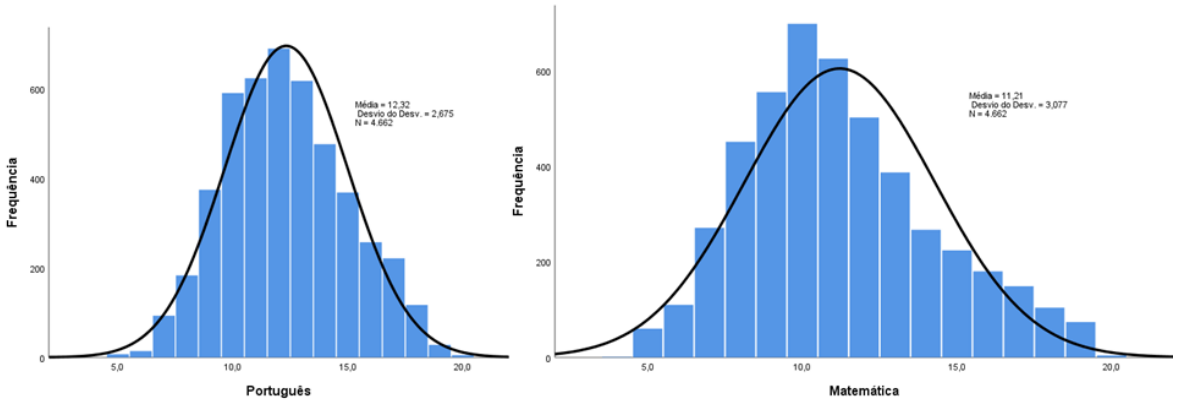


Figura 4: Histogramas para as notas finais em Matemática e Português

Tabela 8: Estatística descritiva para as variáveis Matemática, Português e Idade

	Mínimo	Máximo	Média	Desvio	Assimetria		Curtose	
	Estatística	Estatística	Estatística	Estatística	Estatística	Erro	Estatística	Erro
Nota Final a Português	5,0	20,0	12,319	2,6748	,219	,036	-,416	,07
Nota Final a Matemática	4,0	20,0	11,206	3,0770	,496	,036	-,203	,07
Idade do aluno	12	23	14,95	2,054	,514	,036	-,081	,07

Como já mencionado anteriormente, as notas mínimas registadas nas disciplinas de Matemática e Português foram, respetivamente, 4 e 5 valores, comprovadas pelas Tabelas 9 e 10, que permitem-nos visualizar com exatidão em que ano letivo foram registadas. Para a disciplina de Matemática a nota mínima foi registada no ano letivo 2017-2018, enquanto que para português, tal fato aconteceu nos outros dois anos letivos. Já a nota máxima, foi respetivamente, para Português e Matemática, nos anos letivos 2016-2017 e 2018-2019.

Tabela 9: Estatística descritiva para variável Matemática em relação aos anos letivos

	Nota Final a Matemática								
	Ano letivo								
	16-17			17-18			18-19		
	Média	Mínimo	Máximo	Média	Mínimo	Máximo	Média	Mínimo	Máximo
Estatística	11,369	5,0	19,0	11,04	4,0	20,0	11,20	5,0	20,0
Erro	,0744			,0803			,0795		

Tabela 10: Estatística descritiva para variável Português em relação aos anos letivos

	Nota Final a Português								
	Ano letivo								
	16-17			17-18			18-19		
	Média	Mínimo	Máximo	Média	Mínimo	Máximo	Média	Mínimo	Máximo
Estatística	12,38	5,0	20,0	12,26	6,0	20,0	12,315	5,0	19,0
Erro	,0657			,0669			,0711		

A Figura 5, fornece-nos uma perceção de como ocorreram as classificações nas disciplinas de Matemática e Português, consideradas como variáveis resposta, no período referência do estudo. Nota-se que os alunos com aproveitamento negativo à Matemática representam mais do dobro dos que obtiveram o mesmo desempenho à Português. De uma forma geral, a disciplina de Português apresenta melhores resultados de aproveitamento em relação à Matemática.

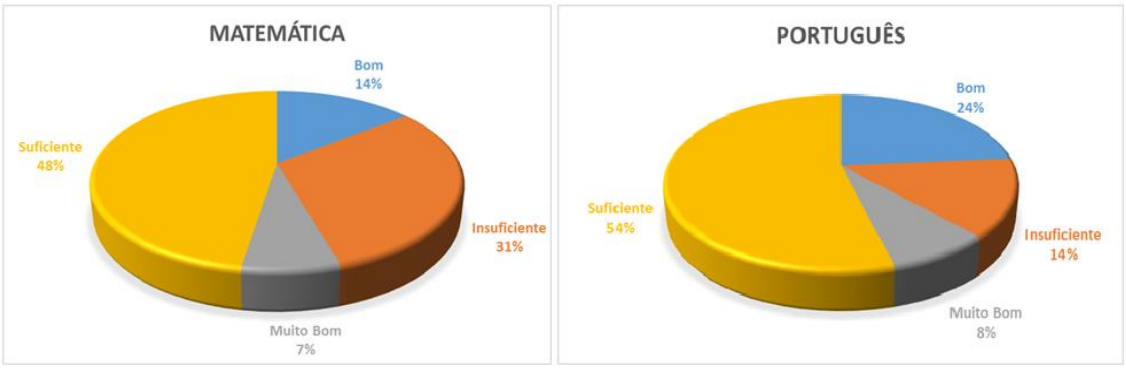


Figura 5: Classificação Qualitativa nas disciplinas de Matemática e Português

Feita a análise exploratória dos dados e para que possamos compreender melhor a influência das variáveis explicativas sobre a resposta, o tópico seguinte apresenta-nos algumas estatísticas inferenciais sobre os dados.

3.3. Importância das possíveis variáveis explicativas sobre as respostas

A influência que as variáveis explicativas exercem sobre as variáveis respostas pode diferir tendo em conta as caraterísticas de cada uma. É neste sentido, e com a intenção de

analisar a forma como as variáveis explicativas influenciam as variáveis respostas, que exporemos uma análise inferencial sobre os dados envolvidos no estudo.

Como já referenciado anteriormente, o sistema de ensino Cabo-verdiano possui uma particularidade no terceiro ciclo do ensino secundário, que é a possibilidade dos alunos escolherem entre o ramo profissional e o geral, portanto acarretando que os programas curriculares sejam diferentes. Neste sentido, tentando indagar sobre a relevância de trabalharmos com todos os alunos do sistema de ensino (Via Técnica e Via Geral) ou não, apresentamos de seguida evidências estatísticas para comparar se as médias desses alunos (terceiro ciclo do ensino secundário) diferem. Para tal, recomenda-se primeiramente que se estude a normalidade e a homogeneidade dos dados com vista a decidir entre a opção de um teste paramétrico ou não.

Para o teste K–S de normalidade, definiremos as hipóteses como:

H_0: a distribuição dos dados das variáveis respostas aproxima-se da normal

versus

H_1: a distribuição dos dados das variáveis respostas não se aproxima da normal

Para o teste de homogeneidade das variâncias, as hipóteses serão assim definidas:

H_0: não existem diferenças significativas entre as variâncias das classificações dos grupos em cada variável resposta.

versus

H_1: existe pelo menos dois grupos em cada variável resposta com diferenças entre as variâncias.

A Tabela 11 apresenta-nos o resumo descritivo das variáveis respostas referentes as duas de vias de ensino no 3º ciclo. As Tabelas 12 e 13, nos levam a concluir que a um nível de significância de 5%, temos evidências estatísticas $(p-\text{value}<0{,}001)$ para não aceitar a condição de normalidade das observações nas classificações das disciplinas de

Matemática e Português nas duas vias de ensino. Tendo registado $p-\text{value}=0,868>0,05$, concluímos que com base na média temos evidências estatísticas para aceitar a homogeneidade das variâncias somente na variável Nota Final a Português, visto que para a variável resposta Nota Final a Matemática o teste de Levene forneceu-nos um $p-\text{value}<0,001$.

Tabela 11: Estatística descritiva para as variáveis respostas referentes ao 3º Ciclo

	Via de estudos no 3º Ciclo			Estatística	Erro
Nota Final a Português	Via Geral	Média		13,05	,113
		95% Intervalo de Confiança para Média	Limite inferior	12,82	
			Limite superior	13,27	
		Erro Desvio		2,623	
	Via Técnica	Média		12,07	,131
		95% Intervalo de Confiança para Média	Limite inferior	11,81	
			Limite superior	12,33	
		Erro Desvio		2,688	
Nota Final a Matemática	Via Geral	Média		11,90	,141
		95% Intervalo de Confiança para Média	Limite inferior	11,62	
			Limite superior	12,18	
		Erro Desvio		3,276	
	Via Técnica	Média		11,34	,130
		95% Intervalo de Confiança para Média	Limite inferior	11,08	
			Limite superior	11,59	
		Erro Desvio		2,661	

Tabela 12: Teste de normalidade K-S das variáveis respostas referentes às vias de ensino

	Via de estudos no 3º Ciclo	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estatística	df	Sig.	Estatística	df	Sig.
Nota Final a Português	Via Geral	,123	538	,000	,968	538	,000
	Via Técnica	,092	418	,000	,983	418	,000
Nota Final a Matemática	Via Geral	,112	538	,000	,977	538	,000
	Via Técnica	,139	418	,000	,962	418	,000

a. Correlação de Significância de Lilliefors

Tabela 13: Teste de homogeneidade de Levene das variáveis respostas referentes às vias de ensino

		Estatística de Levene	df1	df2	Sig.
Nota Final a Português	Com base em média	,027	1	954	,868
	Com base em mediana	,021	1	954	,885
	Com base em mediana e com df ajustado	,021	1	950,7	,885
	Com base em média aparada	,062	1	954	,803
Nota Final a Matemática	Com base em média	21,799	1	954	,000
	Com base em mediana	21,897	1	954	,000
	Com base em mediana e com df ajustado	21,897	1	918,0	,000
	Com base em média aparada	22,711	1	954	,000

Tendo somente aceitado a condição a homogeneidade das variâncias na variável resposta Nota Final a Português e de termos rejeitado a suposição de normalidade dos dados nas duas variáveis respostas, a opção dos testes não paramétricos afigura-se como o mais adequado. Para tal, usaremos a alternativa não paramétrica ao teste *t* independente que é o teste Mann–Witney com as hipóteses assim formuladas:

> H_0: a distribuição das notas finais nas variáveis respostas é a mesma nas duas vias de ensino no 3º Ciclo

versus

> H_1: a distribuição das notas finais nas variáveis respostas não é a mesma nas duas vias de ensino no 3º Ciclo

Tabela 14: Postos e soma das classificações dos dados nas variáveis respostas

	Via de estudos no 3º Ciclo	N	Posto Médio	Soma de Classificações
Nota Final a Português	Via Geral	538	517,69	278519,00
	Via Técnica	418	428,06	178927,00
	Total	956		
Nota Final a Matemática	Via Geral	538	501,09	269586,00
	Via Técnica	418	449,43	187860,00
	Total	956		

Tabela 15: Estatística do teste Mann–Witney com Via de Estudos no 3º Ciclo como variável de agrupamento

	Nota Final a Português	Nota Final a Matemática
U de Mann-Whitney	91356,000	100289,000
Wilcoxon W	178927,000	187860,000
Z	-5,010	-2,887
Significância Sig. (bilateral)	,000	,004

Dado que o teste é não paramétrico, a Tabela 14 apresenta-nos relação dos postos das observações, pois em vez das observações absolutas são usados os postos na elaboração do teste. O *output* fornecido pela Tabela 15 nos leva a, em ambos os casos $p-\text{value}<0{,}05$, rejeitar a hipótese nula, ou seja, há diferenças significativas nas medianas das observações referentes às variáveis respostas Matemática e Português.

Não tendo evidências estatísticas para aceitar que a distribuição nas notas finais referentes às disciplinas de Matemática e Português nas duas vias de ensino no terceiro ciclo do ensino secundário são iguais, portanto reconhecendo que dado a natureza de cada ciclo com suas peculiaridades há fatores diferentes que podem interferir neste processo, decidimos restringir a nossa base de dados de estudos somente a alunos da Via Geral. Neste sentido, excluídos os 418 alunos da Via Técnica, a dimensão da base de dados integra agora 4244 alunos desde o 7º ao 12º Ano de escolaridade, com 930 e 3314 alunos distribuídos, respetivamente, pela ESASP e ETJV (Tabela 16).

Tabela 16: Segunda filtragem da base de dados

	ESASP					ETJV					GERAL				
	Válidos		Excluídos		Total	Válidos		Excluídos		Total	Válidos		Excluídos		Total
	N	Perc.	N	Perc.		N	Perc.	N	Perc.		N	Perc.	N	Perc.	
2016 - 2017	317	100,00		0,00	317	1117	87,06	166	12,94	1283	1434	89,63	166	10,38	1600
2017 - 2018	305	100,00		0,00	305	1100	89,14	134	10,86	1234	1405	91,29	134	8,71	1539
2018 - 2019	308	100,00		0,00	308	1097	90,29	118	9,71	1215	1405	92,25	118	7,75	1523
Total	930	100,00		0,00	930	3314	88,80	418	11,20	3732	4244	91,03	418	8,97	4662

As inferências estatísticas anteriormente feitas incidiram somente sobre o 3º Ciclo do ensino secundário tendo como grupos as duas vias de ensino, técnica e geral. Doravante e dado que a Via Técnica foi excluída da nossa base, as inferências a serem feitas terão a base de dados como referência.

Pretendendo analisar a existência de associação entre as variáveis respostas e algumas das variáveis explicativas, o teste de independência Qui–Quadrado apresenta-se como o ideal, pois nos fornece evidências estatísticas para aceitar ou não essa associação. O teste determina se existe uma associação entre duas variáveis categóricas, comparando as frequências observadas e as esperadas, visto que quanto maior for a associação entre duas variáveis, maior é a expetativa de que as frequências esperadas difiram das observadas.

Para isto, houve a necessidade de se criar duas variáveis categóricas (*Class_Port: "Classificação Final à Português"* e *Class_Mat: "Classificação Final à Matemática"*), com base nos resultados finais nas disciplinas de Matemática e Português, codificados da seguinte forma:

- Insuficiente: $[0; 10[$ valores;
- Suficiente: $[10; 14[$ valores;
- Bom: $[14; 17[$ valores;
- Muito Bom: $[17; 20]$ valores.

As Tabelas 17 a 24, nos mostram que os pressupostos necessários para aplicarmos o teste foram respeitados e, portanto, estamos em condições de testar a associação entre as variáveis explicativas e as respostas Nota Final à Português e à Matemática.

Tabela 17: Tabela bivariada com as variáveis Português, Sexo e Ciclo de estudos

			Sexo do aluno			Ciclo de estudos			
			Feminino	Masculino	Total	1º Ciclo	2º Ciclo	3º Ciclo	Total
Classificação Final a Português	Insuficiente	Contagem	184	414	598	394	175	29	598
		Contagem Esperada	324,4	273,6	598,0	281,5	240,7	75,8	598,0
		Resíduos ajustados	-12,4	12,4		9,9	-5,9	-6,2	
	Suficiente	Contagem	1183	1125	2308	1072	942	294	2308
		Contagem Esperada	1251,9	1056,1	2308,0	1086,6	928,9	292,6	2308,0
		Resíduos ajustados	-4,3	4,3		-,9	,8	,1	
	Bom	Contagem	656	331	987	397	444	146	987
		Contagem Esperada	535,4	451,6	987,0	464,7	397,2	125,1	987,0
		Resíduos ajustados	8,8	-8,8		-4,9	3,5	2,3	
	Muito Bom	Contagem	279	72	351	135	147	69	351
		Contagem Esperada	190,4	160,6	351,0	165,2	141,3	44,5	351,0
		Resíduos ajustados	9,9	-9,9		-3,4	,7	4,1	
Total		Contagem	2302	1942	4244	1998	1708	538	4244
		Contagem Esperada	2302,0	1942,0	4244,0	1998,0	1708,0	538,0	4244,0

Tabela 18: Tabela bivariada com as variáveis Português, Escola que frequenta e Distância casa à escola

			Escola que frequenta			Distância da escola à casa				
			ESASP	ETJV	Total	]0 - 1] km	]1 - 3] km	]3 - 6] km	]6 - ...[km	Total
Classificação Final a Português	Insuficiente	Contagem	127	471	598	225	197	76	100	598
		Contagem Esperada	131,0	467,0	598,0	232,9	222,5	61,2	81,4	598,0
		Resíduos ajustados	-,4	,4		-,7	-2,3	2,2	2,4	
	Suficiente	Contagem	536	1772	2308	891	841	244	332	2308
		Contagem Esperada	505,8	1802,2	2308,0	898,9	858,7	236,0	314,3	2308,0
		Resíduos ajustados	2,3	-2,3		-,5	-1,1	,8	1,6	
	Bom	Contagem	195	792	987	390	411	83	103	987
		Contagem Esperada	216,3	770,7	987,0	384,4	367,2	100,9	134,4	987,0
		Resíduos ajustados	-1,9	1,9		,4	3,3	-2,2	-3,3	
	Muito Bom	Contagem	72	279	351	147	130	31	43	351
		Contagem Esperada	76,9	274,1	351,0	136,7	130,6	35,9	47,8	351,0
		Resíduos ajustados	-,7	,7		1,2	-,1	-,9	-,8	
Total		Contagem	930	3314	4244	1653	1579	434	578	4244
		Contagem Esperada	930,0	3314,0	4244,0	1653,0	1579,0	434,0	578,0	4244,0

Tabela 19: Tabela bivariada com as variáveis Português, Repetente no ano que estuda e Número de reprovações no sistema

			Repetente no ano em que estuda			Número de reprovações no sistema secundário					
			Não	Sim	Total	Nenhuma	Uma vez	Duas vezes	Três vezes	Mais que Três vezes	Total
Classificação Final a Português	Insuficiente	Contagem	442	156	598	333	168	84	13	0	598
		Contagem Esperada	500,6	97,4	598,0	424,3	113	49,7	10,0	,8	598,0
		Resíduos ajustados	-7,0	7,0		-8,9	6,2	5,5	1,0	-1,0	
	Suficiente	Contagem	1865	443	2308	1535	499	219	49	6	2308
		Contagem Esperada	1932,2	375,8	2308,0	1637,5	437	192,0	38,6	3,3	2308,0
		Resíduos ajustados	-5,6	5,6		-7,0	4,9	3,0	2,5	2,2	
	Bom	Contagem	910	77	987	821	111	47	8	0	987
		Contagem Esperada	826,3	160,7	987,0	700,2	187	82,1	16,5	1,4	987,0
		Resíduos ajustados	8,2	-8,2		9,7	-7,0	-4,6	-2,4	-1,3	
	Muito Bom	Contagem	336	15	351	322	25	3	1	0	351
		Contagem Esperada	293,9	57,1	351,0	249,0	66,4	29,2	5,9	,5	351,0
		Resíduos ajustados	6,4	-6,4		9,0	-5,9	-5,3	-2,1	-,7	
Total		Contagem	3553	691	4244	3011	803	353	71	6	4244
		Contagem Esperada	3553,0	691,0	4244,0	3011,0	803	353,0	71,0	6,0	4244,0

Tabela 20: Tabela bivariada com as variáveis Português e Nível de instrução do encarregado de educação

			Instrução do encarregado de educação												
			Sem instrução	1º Ciclo - EB	2º Ciclo - EB	3º Ciclo - EB	1º Ciclo - ES	2º Ciclo - ES	3º Ciclo - ES	Curso Médio	Curso Superior sem Licenciatura	Licenciatura	Mestrado / Pós Graduação	Outras	Total
Classificação Final a Português	Insuficiente	Contagem	93	81	182	95	26	51	34	12	0	19	0	5	598
		Contagem Esperada	71,9	69,2	154,3	94,4	39,6	53,4	62,7	14,4	4,4	27,5	1,1	5,2	598,0
		Resíduos ajustados	2,9	1,6	2,8	,1	-2,4	-,4	-4,1	-,7	-2,3	-1,8	-1,1	-,1	
	Suficiente	Contagem	301	313	591	387	155	189	229	43	14	64	2	20	2308
		Contagem Esperada	277,4	267,0	595,5	364,4	152,8	206,1	242,0	55,5	16,9	106,0	4,4	20,1	2308,0
		Resíduos ajustados	2,2	4,4	-,3	1,9	,3	-1,8	-1,3	-2,5	-1,0	-6,2	-1,7	,0	
	Bom	Contagem	100	78	239	148	74	98	134	25	10	64	5	12	987
		Contagem Esperada	118,6	114,2	254,7	155,8	65,4	88,1	103,5	23,7	7,2	45,3	1,9	8,6	987,0
		Resíduos ajustados	-2,1	-4,1	-1,3	-,8	1,3	1,3	3,6	,3	1,2	3,2	2,6	1,3	
	Muito Bom	Contagem	16	19	83	40	26	41	48	22	7	48	1	0	351
		Contagem Esperada	42,2	40,6	90,6	55,4	23,2	31,3	36,8	8,4	2,6	16,1	,7	3,1	351,0
		Resíduos ajustados	-4,5	-3,8	-1,0	-2,4	,6	1,9	2,0	4,9	2,9	8,5	,4	-1,8	
Total		Contagem	510	491	1095	670	281	379	445	102	31	195	8	37	4244
		Contagem Esperada	510,0	491,0	1095,0	670,0	281,0	379,0	445,0	102,0	31,0	195,0	8,0	37,0	4244,0

Tabela 21: Tabela bivariada com as variáveis Matemática, Sexo e Ciclos de estudos

			Sexo do aluno			Ciclo de estudos			
			Feminino	Masculino	Total	1º Ciclo	2º Ciclo	3º Ciclo	Total
Classificação Final a Matemática	Insuficiente	Contagem	649	702	1351	715	516	120	1351
		Contagem Esperada	732,8	618,2	1351,0	636,0	543,7	171,3	1351,0
		Resíduos ajustados	-5,5	5,5		5,2	-1,9	-5,1	
	Suficiente	Contagem	1069	903	1972	893	831	248	1972
		Contagem Esperada	1069,6	902,4	1972,0	928,4	793,6	250,0	1972,0
		Resíduos ajustados	,0	,0		-2,2	2,3	-,2	
	Bom	Contagem	362	244	606	271	223	112	606
		Contagem Esperada	328,7	277,3	606,0	285,3	243,9	76,8	606,0
		Resíduos ajustados	2,9	-2,9		-1,3	-1,9	4,6	
	Muito Bom	Contagem	222	93	315	119	138	58	315
		Contagem Esperada	170,9	144,1	315,0	148,3	126,8	39,9	315,0
		Resíduos ajustados	6,0	-6,0		-3,4	1,3	3,2	
Total		Contagem	2302	1942	4244	1998	1708	538	4244
		Contagem Esperada	2302,0	1942,0	4244,0	1998,0	1708,0	538,0	4244,0

Tabela 22: Tabela bivariada com as variáveis Matemática, Escola que frequenta e Distância da casa à escola

			Escola que frequenta			Distância da escola à casa				
			ESASP	ETJV	Total	]0 - 1] km	]1 - 3] km	]3 - 6] km	]6 - ...[km	Total
Classificação Final a Matemática	Insuficiente	Contagem	270	1081	1351	522	474	135	220	1351
		Contagem Esperada	296,0	1055,0	1351,0	526,2	502,6	138,2	184,0	1351,0
		Resíduos ajustados	-2,1	2,1		-,3	-2,0	-,3	3,5	
	Suficiente	Contagem	443	1529	1972	746	753	220	253	1972
		Contagem Esperada	432,1	1539,9	1972,0	768,1	733,7	201,7	268,6	1972,0
		Resíduos ajustados	,8	-,8		-1,4	1,2	1,9	-1,4	
	Bom	Contagem	157	449	606	235	242	55	74	606
		Contagem Esperada	132,8	473,2	606,0	236,0	225,5	62,0	82,5	606,0
		Resíduos ajustados	2,6	-2,6		-,1	1,5	-1,0	-1,1	
	Muito Bom	Contagem	60	255	315	150	110	24	31	315
		Contagem Esperada	69,0	246,0	315,0	122,7	117,2	32,2	42,9	315,0
		Resíduos ajustados	-1,3	1,3		3,3	-,9	-1,6	-2,0	
Total		Contagem	930	3314	4244	1653	1579	434	578	4244
		Contagem Esperada	930,0	3314,0	4244,0	1653,0	1579,0	434,0	578,0	4244,0

Tabela 23: Tabela bivariada com as variáveis Matemática, Repetente no ano que estuda e Número de reprovações no sistema

			Repetente no ano em que estuda			Número de reprovações no sistema secundário					
			Não	Sim	Total	Nenhuma	Uma vez	Duas vezes	Três vezes	Mais que Três vezes	Total
Classificação Final a Matemática	Insuficiente	Contagem	1061	290	1351	793	339	177	38	4	1351
		Contagem Esperada	1131,0	220,0	1351,0	958,5	256	112,4	22,6	1,9	1351,0
		Resíduos ajustados	-6,3	6,3		-12,0	7,0	7,7	4,0	1,8	
	Suficiente	Contagem	1655	317	1972	1439	361	143	28	1	1972
		Contagem Esperada	1650,9	321,1	1972,0	1399,1	373	164,0	33,0	2,8	1972,0
		Resíduos ajustados	,3	-,3		2,7	-1,0	-2,3	-1,2	-1,5	
	Bom	Contagem	539	67	606	496	79	27	3	1	606
		Contagem Esperada	507,3	98,7	606,0	429,9	115	50,4	10,1	,9	606,0
		Resíduos ajustados	3,8	-3,8		6,4	-4,0	-3,7	-2,4	,2	
	Muito Bom	Contagem	298	17	315	283	24	6	2	0	315
		Contagem Esperada	263,7	51,3	315,0	223,5	59,6	26,2	5,3	,4	315,0
		Resíduos ajustados	5,4	-5,4		7,7	-5,3	-4,3	-1,5	-,7	
Total		Contagem	3553	691	4244	3011	803	353	71	6	4244
		Contagem Esperada	3553,0	691,0	4244,0	3011,0	803	353,0	71,0	6,0	4244,0

Tabela 24: Tabela bivariada com as variáveis Matemática e Nível de instrução do encarregado de educação

		Instrução do encarregado de educação												
		Sem instrução	1º Ciclo - EB	2º Ciclo - EB	3º Ciclo - EB	1º Ciclo - ES	2º Ciclo - ES	3º Ciclo - ES	Curso Médio	Curso Superior sem Licenciatura	Licenciatura	Mestrado / Pós Graduação	Outras	Total
Classificação Final a Matemática	Contagem	174	200	386	218	79	113	104	24	7	32	1	13	1351
	Contagem Esperada	162,3	156,3	348,6	213,3	89,5	120,6	141,7	32,5	9,9	62,1	2,5	11,8	1351
	Resíduos ajustados	1,2	4,5	2,8	,4	-1,4	-,9	-4,1	-1,8	-1,1	-4,7	-1,2	,4	
	Contagem	244	216	515	328	146	167	209	42	12	70	2	21	1972
	Contagem Esperada	237,0	228,1	508,8	311,3	130,6	176,1	206,8	47,4	14,4	90,6	3,7	17,2	1972
	Resíduos ajustados	,7	-1,2	,4	1,4	1,9	-1,0	,2	-1,1	-,9	-3,0	-1,2	1,3	
	Contagem	73	49	134	82	35	71	92	20	7	39	3	1	606
	Contagem Esperada	72,8	70,1	156,4	95,7	40,1	54,1	63,5	14,6	4,4	27,8	1,1	5,3	606,0
	Resíduos ajustados	,0	-2,9	-2,2	-1,6	-,9	2,6	4,1	1,6	1,3	2,3	1,9	-2,0	
	Contagem	19	26	60	42	21	28	40	16	5	54	2	2	315
	Contagem Esperada	37,9	36,4	81,3	49,7	20,9	28,1	33,0	7,6	2,3	14,5	,6	2,7	315,0
	Resíduos ajustados	-3,4	-1,9	-2,8	-1,2	,0	,0	1,3	3,2	1,9	11,1	1,9	-,5	
Total	Contagem	510	491	1095	670	281	379	445	102	31	195	8	37	4244
	Contagem Esperada	510,0	491,0	1095,0	670,0	281,0	379,0	445,0	102,0	31,0	195,0	8,0	37,0	4244

Estando em condições para testar a existência ou não de associação entre as variáveis respostas Nota Final a Português e a Matemática e as variáveis explicativas, então definiremos as nossas hipóteses como:

H_0: as variáveis não estão associadas, ou seja, elas são independentes, ou ainda, não existem diferenças entre as amostras relativamente à distribuição nas classes da variável resposta;

versus

H_1: as variáveis estão associadas, ou seja, são dependentes, ou ainda, existem diferenças significativas entre os grupos ou populações de onde foram extraídas as amostras.

Tabela 25: Teste Qui-Quadrado de associação entre a variável Matemática e as variáveis explicativas

		Valor	gl	Significância Assintótica (Bilateral)	Significância aproximada
Sexo	Qui-Quadrado de Pearson	61,765	3	,000	
	Razão de verossimilhança	63,035	3	,000	
	V de Cramer	,121			,000
Ciclo de estudos	Qui-Quadrado de Pearson	63,258	6	,000	
	Razão de verossimilhança	62,136	6	,000	
	V de Cramer	,086			,000
Escola que frequenta	Qui-Quadrado de Pearson	10,447	3	,015	
	Razão de verossimilhança	10,337	3	,016	
	V de Cramer	,050			,015
Distância da casa à escola	Qui-Quadrado de Pearson	27,296	9	,001	
	Razão de verossimilhança	27,079	9	,001	
	V de Cramer	,046			,001
Instrução do Encarregado de Educação	Qui-Quadrado de Pearson	248,373	33	,000	
	Razão de verossimilhança	205,284	33	,000	
	V de Cramer	,140			,000
Repetente no ano que estuda	Qui-Quadrado de Pearson	66,216	3	,000	
	Razão de verossimilhança	73,452	3	,000	
	V de Cramer	,125			,000
Número de reprovações no sistema	Qui-Quadrado de Pearson	204,171	12	,000	
	Razão de verossimilhança	214,325	12	,000	
	V de Cramer	,127			,000

Tabela 26: Teste Qui-Quadrado de associação entre a variável Português e as variáveis explicativas

		Valor	gl	Significância Assintótica (Bilateral)	Significância aproximada
Sexo	Qui-Quadrado de Pearson	290,566	3	,000	
	Razão de verossimilhança	301,079	3	,000	
	V de Cramer	,262			,000
Ciclo de estudos	Qui-Quadrado de Pearson	130,250	6	,000	
	Razão de verossimilhança	134,654	6	,000	
	V de Cramer	,124			,000
Escola que frequenta	Qui-Quadrado de Pearson	5,560	3	,135	
	Razão de verossimilhança	6,602	3	,133	
	V de Cramer	,036			,135
Distância da casa à escola	Qui-Quadrado de Pearson	30,479	9	,000	
	Razão de verossimilhança	30,725	9	,000	
	V de Cramer	,049			,000
Instrução do Encarregado de Educação	Qui-Quadrado de Pearson	251,917	33	,000	
	Razão de verossimilhança	240,626	33	,000	
	V de Cramer	,141			,000
Repetente no ano que estuda	Qui-Quadrado de Pearson	145,746	3	,000	
	Razão de verossimilhança	162,936	3	,000	
	V de Cramer	,185			,000
Número de reprovações no sistema	Qui-Quadrado de Pearson	243,347	12	,000	
	Razão de verossimilhança	270,610	12	,000	
	V de Cramer	,138			,000

As Tabelas 25 e 26 apresentam-nos evidências estatísticas que nos permitem analisar, através do teste Qui–Quadrado, a associação entre as variáveis explicativas Sexo, Ciclo de estudos, Escola que frequenta, Distância da escola à casa e Instrução do encarregado de educação e as variáveis respostas Nota Final à Matemática e à Português. Estatisticamente o teste nos forneceu evidências para somente rejeitar a associação entre a variável explicativa Escola que frequenta e a variável resposta Português, com $p-\text{value}=0{,}135>0{,}05$, ou seja, a variável escola não influencia significativamente a classificação dos alunos na disciplina de Português. Através do coeficiente de Cramer, podemos avaliar a força das associações das variáveis explicativas com as respostas, e neste sentido, as variáveis explicativas que revelaram estarem mais associadas com a

Classificação Final a Português foram Sexo, Repetente no ano que estuda e Instrução do encarregado de educação com uma associação de 26,2%, 18,5% e 14,1%, respetivamente (Tabela 26). Em relação a variável resposta Classificação Final a Matemática, nota-se uma associação das variáveis Instrução do encarregado de educação com 14,0%, Número de reprovações no sistema com 12,7% e Repetente no ano que estuda com 12,5% (Tabela 25).

Aconselha-se a aplicação de testes paramétricos quando se pretende analisar a influência de uma certa variável explicativa sobre uma resposta. Nesta ótica, inicialmente pretendemos aplicar o teste Análise de Variância (ANOVA) com o objetivo de estudar a importância das variáveis explicativas sobre as respostas.

Sendo a ANOVA um teste paramétrico, devemos garantir, para sua aplicação os seguintes pressupostos: (1) que a variável dependente possua distribuição normal; (2) que as variâncias populacionais sejam homogéneas em cada grupo. Para tal, testaremos a normalidade dos dados das variáveis respostas usando o Teste K–S e a homogeneidade de variâncias será verificada com o Teste de Levene.

Relembremos que a Tabela 12, nos levou a concluir a não normalidade dos dados, mas somente em relação ao 3º Ciclo. Generalizar essa suposição de não normalidade para o todo seria uma conclusão errónea, por isso que de seguida apresentamos para a totalidade da nossa base de dados.

Tabela 27: Teste K–S de Normalidade dos dados

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	df	Sig.	Estatística	df	Sig.
Nota Final a Português	,100	4244	,000	,979	4244	,000
Nota Final a Matemática	,120	4244	,000	,967	4244	,000

a. Correlação de Significância de Lilliefors

Tabela 28: Teste de Levene das variáveis respostas e as explicativas

		Sexo	Ciclo de estudos	Escola que frequante	Distância casa-escola	Instrução do encarregado de educação	Repetente no ano que estuda	Número de reprovações no sistema
Nota Final a Português	Com base na média	,013	,011	,492	,342	,000	,000	,000
	Com base na mediana	,008	,019	,513	,509	,000	,000	,000
Nota Final a Matemática	Com base na média	,000	,064	,938	,001	,000	,000	,000
	Com base na mediana	,000	,035	,815	,013	,000	,000	,000

A Tabela 27 (estatística referente a teste de normalidade) fornece-nos evidências estatísticas (para ambas as variáveis respostas temos $p-\text{value}<0,001$) que nos levam a não aceitar a suposição de normalidade das classificações das disciplinas consideradas como variáveis respostas. Mesmo não aceitando a normalidade dos dados, poderíamos aplicar o teste de Levene (Tabela 28, estatística referente ao teste de homogeneidade das variâncias), pois o mesmo é robusto a não normalidade dos dados. Neste caso, e usando o teste baseado na média, concluímos que há evidências estatísticas, somente nas variáveis Ciclo de Estudos e Escola que frequenta, para não rejeitarmos a suposição de homogeneidade de variâncias quando a analisamos em relação a variável resposta Classificação Final a Matemática. Dever-se-á tirar a mesma conclusão sobre a homogeneidade referente à Classificação Final à Português considerando as variáveis explicativas Escola que frequenta e Distância da casa à escola. Podemos notar que, dado a não aceitação da normalidade dos dados, poderíamos ter feito a análise anterior considerando como referência à mediana, o que nos levaria a mesma conclusão. Este facto, nos colocou perante a impossibilidade de aplicar os testes paramétricos nomeadamente o teste *t* de Student para a comparação de duas médias populacionais e a comparação de médias de mais do que duas populações (Análise de Variância).

Dado que não aceitamos a condição homogeneidade das variâncias, e nem a normalidade, aplicaremos os testes não paramétricos Mann–Whitney e Kruskal–Wallis que se figuram, respetivamente, como alternativa aos testes paramétricos teste *t* de Student para a comparação de duas médias populacionais e ANOVA.

As hipóteses para o teste Mann–Whitney são definidas como:

H_0: a média das classificações nas variáveis respostas é a mesma para os dois grupos, ou seja, $\mu_{\text{Grupo 1}} = \mu_{\text{Grupo 2}}$.

versus

H_1: a média das classificações nas variáveis respostas não é a mesma para os dois grupos, ou seja, $\mu_{\text{Grupo 1}} \neq \mu_{\text{Grupo 2}}$.

Tabela 29: Teste Mann–Witney das variáveis respostas segundo Sexo, Escola e Repetente no ano que estuda

			Nota Final a Português	Nota Final a Matemática
Variável de Agrupamento	Sexo	U de Mann-Whitney	1515618,000	1913086,500
		Wilcoxon W	3402271,000	3799739,500
		Significância bilateral	,000	,000
	Escola que frequenta	U de Mann-Whitney	1480535,000	1482913,500
		Wilcoxon W	1913450,000	6975868,500
		Significância bilateral	,065	,077
	Repetente no ano que estuda	U de Mann-Whitney	839403,000	989319,000
		Wilcoxon W	1078489,000	1228405,000
		Significância bilateral	,000	,000

Pela análise da Tabela 29, o teste de Mann–Whitney nos forneceu evidências estatísticas para concluir que as variáveis explicativas Sexo e Repetente no ano que estuda têm efeito sobre as classificações das variáveis respostas, em ambos os casos, com $p-\text{value} < 0{,}05$. Registando $p-\text{value} > 0{,}05$, não aceitamos que a variável Escola teve influência significativa nas classificações referentes às variáveis respostas.

Uma vez que as outras variáveis explicativas são compostas por três ou mais grupos independentes, aplicaremos então o teste não paramétrico Kruskal–Wallis, com as hipóteses assim definidas:

H_0: a distribuição da nota final nas variáveis respostas é a mesma entre as categorias das variáveis explicativas.

H_1: existe pelo menos dois grupos onde as classificações nas variáveis respostas são diferentes.

Tabela 30: Teste Kruskal–Wallis das variáveis respostas segundo Ciclo de estudos, Distância casa-escola, Instrução do encarregado de educação e Número de reprovações

			Nota Final a Português	Nota Final a Matemática
Variável de Agrupamento	Ciclo de Estudos	H de Kruskal-Wallis	118,442	41,421
		gl	2	2
		Significância	,000	,000
	Distância da casa à escola	H de Kruskal-Wallis	26,264	15,053
		gl	3	3
		Significância	,000	,002
	Instrução do Encarregado de Educação	H de Kruskal-Wallis	217,566	154,983
		gl	11	11
		Significância	,000	,000
	Número de Reprovações no Sistema	H de Kruskal-Wallis	278,493	221,015
		gl	4	4
		Significância	,000	,000

Pela Tabela 30 que apresenta-nos estatísticas do teste Kruskal–Wallis, concluímos que há evidências $(p-\text{value}<0,05)$ para rejeitarmos a hipótese nula, ou seja, que existe pelo menos dois grupos em cada variável explicativa constante na tabela acima onde as classificações médias nas variáveis respostas são diferentes, ou por outras palavras existe influência dos grupos em cada variável explicativa sobre as classificações médias nas disciplinas de Matemática e Português.

Concluindo quais as variáveis explicativas que fazem efeito sobre as respostas, torna-se necessário investigar onde de facto tais diferenças ocorrem, sendo que um dos processos para estudar tais diferenças seria efetuar vários testes Mann–Whitney (em termos paramétricos seria o mesmo que efetuar vários testes *t* de Student após realizar uma ANOVA). Este acarretaria um problema, pois a realização de vários testes conduz-nos a

um aumento do erro do tipo I (rejeitar a hipótese nula sendo ela verdadeira). Sendo assim, sugere-se que o valor de *p* seja retificado, na realização das Mann–Whitney, pela correção de Bonferroni. Esta correção consiste em dividir o valor de p pelo número de comparações dois a dois a fazer em cada grupo. No nosso exemplo, e usando a variável Distância da casa à escola, o número de comparações dois a dois é dado por $C_2^4 = 6$, portanto o valor de p corrigido seria de $p = \frac{0,05}{6}$.

A conclusão da rejeição da hipótese nula, não nos permite saber com exatidão quais os grupos onde tais diferenças se verificam. Neste sentido, tratando-se do teste não paramétrico de Kruskal–Wallis existe a possibilidade de fazermos múltiplas comparações dos grupos dois a dois e assim analisar estatisticamente quais os grupos onde de facto as diferenças se verificam.

As tabelas 31 a 42 nos mostram além das estatísticas descritivas dos grupos de cada variável explicativa, também as comparações dos grupos dois a dois, usando a correção de Bonferroni, para que possamos identificar as diferenças nos diversos grupos. Pela análise das Tabelas 31 e 33, podemos verificar que, à primeira vista, a diferença entre as médias nos diferentes ciclos parecem serem significativas para ambas as variáveis respostas, perceção esta confirmada pelas evidências estatísticas fornecidas pelas Tabelas 32 e 34.

Tabela 31: Estatísticas descritivas dos grupos da variável Ciclo de estudos em relação a Português

Nota Final a Português					
Ciclo de estudos	Média	N	Erro Desvio	Mínimo	Máximo
1º Ciclo	11,89	1998	2,709	5	20
2º Ciclo	12,66	1708	2,550	5	20
3º Ciclo	13,05	538	2,623	7	20
Total	12,34	4244	2,673	5	20

Tabela 32: Teste de comparação múltipla referente a Português e Ciclo de Estudos com *p* ajustado pela correção de Bonferroni

Amostra1-Amostra2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Ajust.
1º Ciclo-2º Ciclo	-361,897	40,126	-9,019	,000	,000
1º Ciclo-3º Ciclo	-508,948	59,142	-8,606	,000	,000
2º Ciclo-3º Ciclo	-147,051	60,198	-2,443	,015	,044

Tabela 33: Estatísticas descritivas dos grupos da variável Ciclo de estudos em relação a Matemática

Nota Final a Matemática

Ciclo de estudos	Média	N	Erro Desvio	Mínimo	Máximo
1º Ciclo	10,95	1998	3,055	4	19
2º Ciclo	11,26	1708	3,096	5	20
3º Ciclo	11,90	538	3,276	5	20
Total	11,19	4244	3,115	4	20

Tabela 34: Teste de comparação múltipla referente a Matemática e Ciclo de Estudos com *p* ajustado pela correção de Bonferroni

Amostra1-Amostra2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Ajust.
1º Ciclo-2º Ciclo	-110,662	40,168	-2,755	,006	,018
1º Ciclo-3º Ciclo	-378,096	59,205	-6,386	,000	,000
2º Ciclo-3º Ciclo	-267,434	60,261	-4,438	,000	,000

A comparação do efeito das diferenças entre as classificações nos grupos da variável Distância da casa à escola sobre as duas variáveis respostas revelou-se estatisticamente significativa para os pares]0 - 1[km -]6 - …[km e]1 - 3[km -]6 - …[km (Tabelas 36 e 38). Além das antes referidas, podemos constatar que existem diferenças nas classificações médias na variável resposta Português entre os grupos]1 - 3[km -]3 - 6[km (Tabela 36).

Tabela 35: Estatísticas descritivas dos grupos da variável Distância da escola à casa em relação a Português

Nota Final a Português					
Distância da escola à casa	Média	N	Erro Desvio	Mínimo	Máximo
]0 - 1] km	12,43	1653	2,673	5	20
]1 - 3] km	12,49	1579	2,679	6	20
]3 - 6] km	12,05	434	2,611	5	20
]6 - ...[km	11,93	578	2,648	6	19
Total	12,34	4244	2,673	5	20

Tabela 36: Teste de comparação múltipla referente a Português e Distância da casa à escola com *p* ajustado pela correção de Bonferroni

Amostra1-Amostra2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Ajust.
]6 - ...[km-]3 - 6] km	54,081	77,338	,699	,484	1,000
]6 - ...[km-]0 - 1] km	225,695	58,838	3,836	,000	,001
]6 - ...[km-]1 - 3] km	260,957	59,194	4,408	,000	,000
]3 - 6] km-]0 - 1] km	171,614	65,674	2,613	,009	,054
]3 - 6] km-]1 - 3] km	206,876	65,993	3,135	,002	,010
]0 - 1] km-]1 - 3] km	-35,262	42,847	-,823	,411	1,000

Tabela 37: Estatísticas descritivas dos grupos da variável Distância da escola à casa em relação à Matemática

Nota Final a Matemática					
Distância da escola à casa	Média	N	Erro Desvio	Mínimo	Máximo
]0 - 1] km	11,29	1653	3,233	5	20
]1 - 3] km	11,28	1579	3,086	4	20
]3 - 6] km	11,02	434	2,862	5	19
]6 - ...[km	10,79	578	3,000	5	20
Total	11,19	4244	3,115	4	20

Tabela 38: Teste de comparação múltipla referente a Matemática e Distância da casa à escola com *p* ajustado pela correção de Bonferroni

Amostra1-Amostra2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Ajust.
]6 - ...[km-]3 - 6] km	97,442	77,420	1,259	,208	1,000
]6 - ...[km-]0 - 1] km	186,509	58,901	3,167	,002	,009
]6 - ...[km-]1 - 3] km	214,759	59,257	3,624	,000	,002
]3 - 6] km-]0 - 1] km	89,066	65,743	1,355	,175	1,000
]3 - 6] km-]1 - 3] km	117,317	66,063	1,776	,076	,455
]0 - 1] km-]1 - 3] km	-28,250	42,892	-,659	,510	1,000

As Tabelas 39 e 41 realçam-nos que a média dos alunos, nas disciplinas de matemática e português, que nunca reprovaram no sistema de ensino secundário tende a ser significativamente maior que os demais. De fato, o teste de comparação múltipla de Kruskal–Wallis nos mostra (Tabelas 40 e 42) que somente quando se compara, para ambas as variáveis respostas, o grupo Nenhuma com Uma vez, Duas vezes e Três vezes teremos evidências estatísticas para aceitar o efeito das diferenças em relação às variáveis respostas Matemática e Português.

Tabela 39: Estatísticas descritivas dos grupos da variável Número de reprovações no sistema em relação à Português

Nota Final a Português					
Número de reprovações no sistema secundário	Média	N	Erro Desvio	Mínimo	Máximo
Nenhuma	12,78	3011	2,701	5	20
Uma vez	11,38	803	2,344	5	18
Duas vezes	11,12	353	2,159	5	17
Três vezes	11,01	71	2,135	7	18
Mais que Três vezes	11,33	6	1,211	10	13
Total	12,34	4244	2,673	5	20

Tabela 40: Teste de comparação múltipla referente à Português e Número de Reprovações no sistema com *p* ajustado pela correção de Bonferroni

Amostra1-Amostra2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Ajust.
Três vezes-Duas vezes	58,520	158,371	,370	,712	1,000
Três vezes-Mais que Três vezes	-133,858	517,667	-,259	,796	1,000
Três vezes-Uma vez	184,404	150,758	1,223	,221	1,000
Três vezes-Nenhuma	820,085	146,198	5,609	,000	,000
Duas vezes-Mais que Três vezes	-75,338	501,296	-,150	,881	1,000
Duas vezes-Uma vez	125,884	77,758	1,619	,105	1,000
Duas vezes-Nenhuma	761,565	68,501	11,118	,000	,000
Mais que Três vezes-Uma vez	50,546	498,943	,101	,919	1,000
Mais que Três vezes-Nenhuma	686,227	497,585	1,379	,168	1,000
Uma vez-Nenhuma	635,681	48,360	13,145	,000	,000

Tabela 41: Estatísticas descritivas dos grupos da variável Número de reprovações no sistema em relação à Matemática

Nota Final a Matemática					
Número de reprovações no sistema secundário	Média	N	Erro Desvio	Mínimo	Máximo
Nenhuma	11,64	3011	3,157	4	20
Uma vez	10,26	803	2,779	5	19
Duas vezes	9,87	353	2,578	5	18
Três vezes	9,61	71	2,638	5	19
Mais que Três vezes	9,50	6	2,950	7	15
Total	11,19	4244	3,115	4	20

Tabela 42: Teste de comparação múltipla referente à Matemática e Número de Reprovações no sistema com *p* ajustado pela correção de Bonferroni

Amostra1-Amostra2	Estatística de Teste	Std. Erro	Erro Estatística de Teste	Sig.	Sig. Ajust.
Mais que Três vezes-Três vezes	104,275	518,215	,201	,841	1,000
Mais que Três vezes-Duas vezes	241,167	501,827	,481	,631	1,000
Mais que Três vezes-Uma vez	413,095	499,471	,827	,408	1,000
Mais que Três vezes-Nenhuma	945,118	498,111	1,897	,058	,578
Três vezes-Duas vezes	136,892	158,539	,863	,388	1,000
Três vezes-Uma vez	308,820	150,917	2,046	,041	,407
Três vezes-Nenhuma	840,844	146,353	5,745	,000	,000
Duas vezes-Uma vez	171,928	77,840	2,209	,027	,272
Duas vezes-Nenhuma	703,951	68,573	10,266	,000	,000
Uma vez-Nenhuma	532,024	48,411	10,990	,000	,000

Devido ao número elevado de comparações a serem realizados nos grupos da variável explicativa Nível de Instrução do Encarregado de Educação, as tabelas do teste de comparação múltiplas para as duas variáveis respostas estão disponíveis como Anexo 1.

Mesmo sendo a ANOVA um teste robusto quando as variâncias não são homogéneas desde que as amostras dos grupos sejam de iguais dimensões ou praticamente, há a possibilidade de homogeneizar as variâncias usando, para tal, transformações matemáticas. Segundo Maroco (2003), apesar das transformações terem sido desenvolvidas para homogeneizar, em muitos casos, conduzem também à normalização da variável sob estudo.

A tabela 43 mostra-nos o panorama do teste de homogeneidade das variáveis respostas antes e após sofrerem transformações. Após a transformação das variáveis não ocorreu uma mudança significativa na homogeneização das variáveis, e aliado a este facto as

Tabelas 44 a 46 nos mostram que a opção dos testes não paramétricos se revelou como a mais correta neste caso, visto que a dimensão dos grupos em cada variável explicativa não é igual e nem similar,

Tabela 43: Teste de Levene com as variáveis sem e após sofrerem transformações

	Sexo		Ciclo de Estudos		Escola que frequenta		Distância da casa à escola		Instrução do Encarregado de Educação		Repetente no ano que estuda		Número de reprovações no sistema	
	Mat	Port	Mat	Port	Mat	Port	Mat	Port	Mat	Port	Mat	Port	Mat	Port
Sem transformar	0,000	0,014	0,977	0,005	0,585	0,470	0,000	0,402	0,000	0,000	0,000	0,000	0,000	0,000
ln(y)	0,019	0,008	0,012	0,000	0,877	0,529	0,019	0,934	0,059	0,000	0,000	0,000	0,166	0,014
$\frac{1}{\sqrt{y}}$	0,037	0,000	0,000	0,000	0,809	0,495	0,019	0,540	0,378	0,000	0,000	0,072	0,269	0,502
$\sqrt{y}$	0,000	0,867	0,227	0,000	0,615	0,423	0,005	0,913	0,000	0,000	0,000	0,000	0,000	0,000
$\log_{10}(y)$	0,019	0,008	0,012	0,000	0,877	0,529	0,019	0,934	0,059	0,000	0,000	0,000	0,166	0,014
$\frac{1}{y}$	0,327	0,000	0,001	0,000	0,818	0,518	0,029	0,592	0,334	0,000	0,000	0,030	0,400	0,381

Tabela 44: Frequência das variáveis Sexo, Ciclo de escola e Escola que frequenta

		Frequência	Percentagem	Percentagem acumulativa
Sexo	Feminino	2302	54,2	54,2
	Masculino	1942	45,8	100,0
	Total	4244	100,0	
Ciclo de estudos	1º Ciclo	1998	47,1	47,1
	2º Ciclo	1708	40,2	87,3
	3º Ciclo	538	12,7	100,0
	Total	4244	100,0	
Escola que frequenta	ESASP	930	21,9	21,9
	ETJV	3314	78,1	100,0
	Total	4244	100,0	

Tabela 45: Frequência das variáveis Distância casa à escola e Instrução do encarregado de educação

		Frequência	Percentagem	Percentagem acumulativa
Distância da escola à casa	]0 - 1] km	1653	38,9	38,9
	]1 - 3] km	1579	37,2	76,2
	]3 - 6] km	434	10,2	86,4
	]6 - ...[km	578	13,6	100,0
	Total	4244	100,0	
Instrução do encarregado de educação	Sem instrução	510	12,0	12,0
	1º Ciclo - EB	491	11,6	23,6
	2º Ciclo - EB	1095	25,8	49,4
	3º Ciclo - EB	670	15,8	65,2
	1º Ciclo - ES	281	6,6	71,8
	2º Ciclo - ES	379	8,9	80,7
	3º Ciclo - ES	445	10,5	91,2
	Curso Médio	102	2,4	93,6
	Curso Superior sem Licenciatura	31	,7	94,3
	Licenciatura	195	4,6	98,9
	Mestrado / Pós Graduação	8	,2	99,1
	Outras	37	,9	100,0
	Total	4244	100,0	

Tabela 46: Frequência das variáveis Repetente no ano que estuda e Número de reprovações no sistema

		Frequência	Percentagem	Percentagem acumulativa
Repetente no ano que estuda	Não	3553	83,7	83,7
	Sim	691	16,3	100,0
	Total	4244	100,0	
Número de reprovações no sistema	Nenhuma	3011	70,9	70,9
	Uma vez	803	18,9	89,9
	Duas vezes	353	8,3	98,2
	Três vezes	71	1,7	99,9
	Mais que Três vezes	6	,1	100,0
	Total	4244	100,0	

3.4. Construção dos Modelos de Regressão: Modelos Multiníveis

Tendo a Nota Final à Matemática e a Nota Final à Português como variáveis respostas e a Escola como fator, o objetivo primordial nesta primeira etapa é o de realizar uma análise de variância com coeficientes aleatórios, por forma a concluirmos, com evidências estatísticas, sobre a variação observada na classificação das variáveis respostas explicada pelo efeito da escola, ou seja, é este modelo (nulo, dada pela equação $y_{ik} = \gamma_{00} + u_{ok} + e_{ik}$) que nos fornece a variância entre as médias das duas escolas e a variância entre as médias dos alunos na mesma escola, respetivamente, variância de nível 2 e 1.

A Tabela 47 fornece-nos a estimativa do único parâmetro fixo no modelo (ordenada na origem, ou a média das classificações das duas instituições de ensino) para as duas variáveis respostas, que são respetivamente, $\gamma_{00} = 11{,}2165$ e $\gamma_{00} = 12{,}3186$, para Matemática e Português. Para um nível de significância de 5%, ambos os parâmetros revelaram-se estatisticamente significativos, com $p-\text{values}$ inferiores a 0,05. Estes valores representam a média dos alunos nos anos letivos de referência para este trabalho sem considerar nenhuma outra variável explicativa, sendo tão somente uma confirmação dos valores constantes na Tabela 8.

Tabela 47: Estimativa dos efeitos fixos no Modelo Nulo para as duas variáveis respostas considerada a amostra como um todo

						95,0% Intervalo de Confiança	
Parâmetros	Estimativa	Erro	gl	t	Sig.	Limite inferior	Limite superior
Interceto [Mat]	11,2165	,0789	,734	142,145	,016	8,7711	13,6619
Interceto [Port]	12,3186	,0750	,805	164,345	,010	10,6063	14,0310

As estatísticas apresentadas na tabela acima e na abaixo foram obtidas tendo como referência a amostra como um todo, ou seja, todos os alunos desde o 7º ao 12º ano de escolaridade. Uma vez que nos interessa indagar sobre o quanto as variáveis respostas variam dentro dos níveis, é possível pela análise da Tabela 48 concluir que, para a variável

Nota Final à Matemática a variância dos resíduos no nível dos alunos é de $\sigma_\varepsilon = 9{,}7006$, enquanto que, a variância do fator escola é de $\sigma_u = 0{,}0068$. Para a variável Nota Final à Português, a variância do fator escola difere da outra variável resposta somente na quarta casa decimal, sendo $\sigma_u = 0{,}0069$, e a variabilidade dentro das escolas (resíduo de nível 1) desce para $\sigma_\varepsilon = 7{,}1401$.

Com a estimação destes parâmetros acima é possível definir, para cada uma das variáveis respostas, o Coeficiente de Correlação Intraclasse, sendo estes de, $\rho_{\text{Mat}} = \dfrac{0{,}0068}{0{,}0068 + 9{,}7006} = 0{,}0007$ e $\rho_{\text{Port}} = \dfrac{0{,}0069}{0{,}0069 + 7{,}1401} = 0{,}0010$, respetivamente para a Nota Final à Matemática e à Português. Ora estes coeficientes, nos colocam perante a ponderação em usar o modelo multinível ajustado aos dados, visto serem ambos próximos de zero permitindo-nos concluir que toda a variabilidade nas classificações situa-se no nível 1 (nível dos alunos), conclusão comprovada pelas evidências estatísticas dada pelos valores de significância do teste de Wald, ($p-\text{value} = 0{,}721 > 0{,}05$ e $p-\text{value} = 0{,}679 > 0{,}05$, respetivamente, para a variável dependente Nota Final à Matemática e à Português). Uma vez que, para ambas as variáveis, tivemos evidências estatísticas para rejeitar a hipótese de que a variância populacional do fator Escola é nula, abordaremos o ajuste dos dados a um outro tipo de modelo que não o multinível.

Tabela 48: Estimativas dos parâmetros de covariâncias do Modelo Nulo para as variáveis respostas considerada a amostra como um todo

Parâmetro		Estimativa	Erro	Wald Z	Sig.	95,0% Intervalo de Confiança	
						Limite inferior	Limite superior
Resíduo [Mat]		9,7006	,21	46,054	,000	9,29643	10,1224
Interceto [Mat: Fator = Escola]	Variância	,0068	,0191	,257	,721	,00003	1,64749
Resíduo [Port]		7,1404	,1550	46,054	,000	6,84265	7,45060
Interceto [Port: Fator = Escola]	Variância	,0069	,0168	,414	,679	,00006	,78801

Dado que, pela análise dos dados como um todo não será possível um ajuste dos mesmos usando os modelos multiníveis, pesquisaremos de seguida a possibilidade de fazer tal ajuste considerando os dados desagregados, ou seja, analisando os três Ciclos de estudos separados. A Tabela 49 nos mostra que a estimativa da ordenada na origem é maior, nos três ciclos, para variável resposta Nota Final à Português. Nesta disciplina a maior estimativa regista-se no 3º Ciclo com $\gamma_{00} = 12{,}7392$, contrastando com a estimativa da ordenada na origem para a variável resposta Nota Final à Matemática de $\gamma_{00} = 11{,}9032$.

Tabela 49: Estimativa dos efeitos fixos no Modelo Nulo para as duas variáveis respostas considerada a amostra por ciclos de estudo

Parâmetros	Estimativa	Erro	gl	t	Sig.	95,0% Intervalo de Confiança	
						Limite inferior	Limite superior
Interceto [Mat 1º Ciclo]	11,1180	,3151	,993	32,286	,018	7,04733	15,18874
Interceto [Mat 2º Ciclo]	11,1938	,1634	,892	68,501	,014	8,41912	13,96847
Interceto [Mat 3º Ciclo]	11,9032	,1415	536	84,122	,000	11,62521	12,18113
Interceto [Port 1º Ciclo]	12,0504	,2976	,995	40,488	,016	8,21952	15,88129
Interceto [Port 2º Ciclo]	12,4836	,3260	,997	38,292	,017	8,30732	16,65981
Interceto [Port 3º Ciclo]	12,7392	,6944	,999	18,345	,035	3,89772	21,58064

A desagregação da nossa amostra em três partes distintas, correspondendo aos três ciclos de estudo, não acarretou na possibilidade de haver um ajuste dos dados a um modelo de regressão multinível, pois de acordo com a Tabela 50, a influência do fator Escola deu estatisticamente não significativo para as variáveis respostas em todos os ciclos de estudo. Destes, destacamos o CIC de $\rho = 0{,}1252$ para a variável resposta Nota Final à Português, mas que a estatística do teste de Wald deu como não significativo com um $p - \text{value} = 0{,}494$, e a redundância do efeito da variabilidade do fator Escola sobre as classificações da variável resposta Nota Final à Matemática.

Tabela 50: Estimativas dos parâmetros de covariâncias do Modelo Nulo para as variáveis respostas considerada a amostra por ciclos de estudo

Parâmetro		Estimativa	Erro	CIC	Wald Z	Sig.	95,0% Intervalo de Confiança	
							Limite inferior	Limite superior
Resíduo [Mat 1º Ciclo]		9,2736	,2936		31,591	,000	8,71572	9,86715
Interceto [Mat 1º Ciclo: Fator = Escola]	Variância	,18472	,2813	,0195	,657	,511	,00934	3,65337
Resíduo [Mat 2º Ciclo]		9,5745	,3278		29,206	,000	8,95307	10,23908
Interceto [Mat 2º Ciclo: Fator = Escola]	Variância	,0385	,0777	,0040	,495	,620	,00074	2,01211
Resíduo [Mat 3º Ciclo]		10,7518	,6568		16,371	,000	9,53863	12,11927
Interceto [Mat 3º Ciclo: Fator = Escola]	Variância	,0000	,0000					
Resíduo [Port 1º Ciclo]		7,2852	,2306		31,591	,000	6,84691	7,75145
Interceto [Port 1º Ciclo: Fator = Escola]	Variância	,1663	,2509	,0223	,663	,508	,00864	3,20064
Resíduo [Port 2º Ciclo]		6,4341	,2203		29,206	,000	6,01645	6,88066
Interceto [Port 2º Ciclo: Fator = Escola]	Variância	,2017	,3009	,0304	,670	,503	,01084	3,75367
Resíduo [Port 3º Ciclo]		6,5244	,3989		16,355	,000	5,78750	7,35500
Interceto [Port 3º Ciclo: Fator = Escola]	Variância	,9338	1,364	,1252	,685	,494	,05330	16,35971

Portanto, e dado que para ambas as variáveis respostas o valor do CIC (coeficiente de correlação intraclasse) pode ser considerado como irrisório e estatisticamente não significativo, ocorrendo num caso a redundância, concluímos que o efeito da Escola nas classificações obtidas nas disciplinas de Matemática e Português é praticamente inexistente. Sendo assim, tentaremos recorrer a outras técnicas e procedimentos mais adequados às caraterísticas dos dados e, uma vez que a nossa base de dados possui mais que uma variável considerada como explicativa, passaremos a modelar as informações usando a Regressão Linear Múltipla.

3.5. Construção dos Modelos de Regressão: Modelos Lineares

Nesta etapa do trabalho, e dado que a inclusão deste procedimento estatístico se deveu à inadequação do modelo multinível sobre os dados, decidimos enveredar pelo caminho da Regressão Múltipla mantendo como variáveis respostas as notas finais à Matemática e à Português.

Primeiramente apresentaremos uma análise da correlação entre as variáveis quantitativas (Nota Final à Matemática e à Português, Idade e Número de reprovações no

sistema de ensino secundário). A opção pela análise das correlações usando o coeficiente de Spearman é dada pela Tabela 51, pois um dos pré-requisitos, a normalidade dos dados, foi violada (para ambas temos $p-\text{value}<0{,}001$) e, portanto, a análise da Tabela 52, nos mostra uma associação significativa entre todas as variáveis, sendo que ela é positiva entre as variáveis Nota Final à Matemática e Nota Final à Português e entre Número de reprovações e Idade, e negativa entre as variáveis respostas e Idade e Número de reprovações no sistema de ensino secundário. A relação positiva entre as classificações nas duas disciplinas sugere que alunos com boas notas à Matemática tendem também a ter boas notas à Português e vice-versa, enquanto que a medida que aumenta o número de reprovações no sistema de ensino os alunos tendem a ter piores notas nas disciplinas de Matemática e Português.

Tabela 51: Teste de normalidade K-S das variáveis quantitativas

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	df	Sig.	Estatística	df	Sig.
Nota Final a Matemática	,120	4244	,000	,967	4244	,000
Nota Final a Português	,100	4244	,000	,979	4244	,000
Idade do aluno	,136	4244	,000	,944	4244	,000
Número de reprovações no sistema secundário	,424	4244	,000	,616	4244	,000

a. Correlação de Significância de Lilliefors

Tabela 52: Correlação entre as variáveis quantitativas

			Nota Final a Português	Nota Final a Matemática	Idade do aluno	Número de reprovações no sistema secundário
rô de Spearman	Nota Final a Português	Coeficiente de Correlação	1,000	,612	-,099	-,256
		Sig. (2 extremidades)	.	,000	,000	,000
		N	4244	4244	4244	4244
	Nota Final a Matemática	Coeficiente de Correlação	,612	1,000	-,135	-,228
		Sig. (2 extremidades)	,000	.	,000	,000
		N	4244	4244	4244	4244
	Idade do aluno	Coeficiente de Correlação	-,099	-,135	1,000	,383
		Sig. (2 extremidades)	,000	,000	.	,000
		N	4244	4244	4244	4244
	Número de reprovações no sistema secundário	Coeficiente de Correlação	-,256	-,228	,383	1,000
		Sig. (2 extremidades)	,000	,000	,000	.
		N	4244	4244	4244	4244

A correlação positiva entre as classificações finais nas disciplinas não é a mesma para os sexos, pois a Tabela 53 apresenta-nos diferentes coeficientes de correlação (feminino $\rho_S = 0{,}656$ e masculino $\rho_S = 0{,}552$), ou seja, a relação positiva verifica-se mais forte em alunos do sexo feminino. A mesma conclusão deverá ser feita, quando se compara o efeito da correlação tendo como referência a variável escola (Anexo 2), onde se regista que o coeficiente positivo é maior para a ETJV, com $\rho_S = 0{,}622\,1$ contra o da ESASP com $\rho_S = 0{,}584$. A Tabela 54 nos exibe um contraste no valor do coeficiente de correlação em relação aos alunos que são ou não reprovados no ano que estuda, visto que mesmo sendo positivo esta relação entre as classificações nas duas disciplinas, para os alunos que não são repetentes no ano que estuda o coeficiente de correlação é $\rho_S = 0{,}639$, ficando o coeficiente em $\rho_S = 0{,}391$ para os alunos que são reprovados. Nota-se, pelo Anexo 3, que a medida que aumenta o nível de instrução do encarregado de educação a relação positiva entre as notas finais à Matemática e à Português tende a aumentar, verificando nos dois extremos coeficientes de correlação de e 0,556 e 0,751 para, respetivamente, alunos cujos encarregados não possuem instrução e os que possuem o ensino superior.

Tabela 53: Correlação entre as variáveis Nota Final à Matemática e à Português em relação ao sexo

	Sexo do aluno			Nota Final a Português	Nota Final a Matemática
rô de Spearman	Feminino	Nota Final a Português	Coeficiente de Correlação	1,000	,656
			Sig. (2 extremidades)	.	,000
			N	2302	2302
		Nota Final a Matemática	Coeficiente de Correlação	,656	1,000
			Sig. (2 extremidades)	,000	.
			N	2302	2302
	Masculino	Nota Final a Português	Coeficiente de Correlação	1,000	,552
			Sig. (2 extremidades)	.	,000
			N	1942	1942
		Nota Final a Matemática	Coeficiente de Correlação	,552	1,000
			Sig. (2 extremidades)	,000	.
			N	1942	1942

Tabela 54: Correlação entre as variáveis Nota Final à Matemática e à Português em relação a reprovado no ano que estuda

	Repetente no ano em que estuda			Nota Final a Português	Nota Final a Matemática
rô de Spearman	Não	Nota Final a Português	Coeficiente de Correlação	1,000	,639
			Sig. (2 extremidades)	.	,000
			N	3553	3553
		Nota Final a Matemática	Coeficiente de Correlação	,639	1,000
			Sig. (2 extremidades)	,000	.
			N	3553	3553
	Sim	Nota Final a Português	Coeficiente de Correlação	1,000	,391
			Sig. (2 extremidades)	.	,000
			N	691	691
		Nota Final a Matemática	Coeficiente de Correlação	,391	1,000
			Sig. (2 extremidades)	,000	.
			N	691	691

Sendo uma extensão da correlação (o grau de relação entre duas variáveis), pretendemos com a regressão além da correlação, prever uma variável (variável dependente) através de uma outra (variável independente) e tentar encontrar uma condição matemática que descreva tal relação.

Como definido anteriormente o modelo de regressão linear é definida por $y_j = \beta_0 + \sum_{i=1}^{p} \beta_i x_{ij} + \varepsilon_j$, sendo que quando:

$p = 1$, trata-se de um modelo de regressão linear simples;

$p > 1$, o modelo de regressão designa-se como linear múltipla.

β_0 a ordenada na origem, o seja o valor de y_j quando $x_j = 0$;

β_1 a variação que espera no valor de y_j quando x_j aumenta uma unidade;

x_j a nota final do aluno na disciplina de Português ou Matemática;

ε_j é o resíduo associado ao aluno j, ou seja, é a diferença entre o valor previsto e o valor observado $\left(y_j - y_j\right)$;

De seguida passaremos a ajustar os dados aos dois modelos de regressão múltipla, tendo como (1) variáveis respostas: as Notas Finais à Matemática e à Português e (2) variáveis explicativas: Idade, Sexo, Número de reprovações no sistema, Reprovado no ano que estuda, Escola que frequenta, Ano letivo, Ciclo de estudo, Distância da casa à escola e Nível de instrução do encarregado de educação. O método usado no ajuste dos modelos será o passo a passo (*Stepwise*), que de acordo com Field (2009), cada vez que uma variável explicativa for adicionada no modelo, um teste de remoção é feito sobre a variável explicativa menos útil. De acordo com Maroco (2003), a vantagem deste método, é que permite a remoção de uma variável cuja importância no modelo é reduzida pela adição de novas variáveis.

Teremos a necessidade de, no ajuste dos modelos de regressão linear múltipla, levarmos em consideração variáveis independentes qualitativas. O problema que se coloca quando estas variáveis são incluídas no modelo é, a necessidade de termos em mente que na maior parte das vezes a forma como foram codificadas (as variáveis qualitativas sofreram um processo de discretização) não deverá ser levada em conta na elaboração do modelo. Ilustrando este fato, na nossa base de dados, a atribuição dos valores da variável Distância da Escola à Casa (discretizada em 1 – [0 – 1[km; 2 – [1 – 3[km; 3 – [3 – 6[km; 4 – [6 – …[km) da mesma forma que foi codificada não é aconselhável, primeiro porque a variável não é discreta e depois porque indicaria uma relação de magnitude não presente na mesma. Neste sentido, é possível sim incluí-las no nosso modelo de regressão múltipla, recorrendo para tal às variáveis auxiliadoras indicadoras, conhecidas por variáveis "*dummy*".

Uma variável *dummy* assume apenas dois valores, 0 ou 1, que traduz na presença ou não de determinada caraterística. Desta forma, pode-se dizer que representa um estado, ou por outras palavras, uma variável *dummy* representa algo que não possui valores numéricos e, caso possuir, estes valores não possuem significado numérico.

De uma forma geral, se uma variável tiver n estados, devemos trabalhar com $n-1$ variáveis *dummys*. Em relação a nossa variável Distância da escola à casa, dado que a mesma possui 4 estados, poderíamos definir 3 variáveis *dummys*, da seguinte forma:

Tabela 55: Variáveis auxiliadores indicadoras "*dummy*" para distância da escola à casa

	$D_{[0 - 1[\ km}$	$D_{[1 - 3[\ km}$	$D_{[3 - 6[\ km}$
[0 - 1[km	1	0	0
[1 - 3[km	0	1	0
[3 - 6[km	0	0	1
[6 - ...[km	0	0	0

Portanto, nesta etapa do ajuste dos dados ao modelo de regressão múltipla, incluímos as variáveis qualitativas, Sexo, Ciclo de estudo, Escola, Ano letivo, Distância da escola à Casa, Instrução do encarregado de educação e Repetente no ano que estuda e, dado ao número de categorias presentes na variável Instrução do encarregado de educação

decidimos combinar algumas categorias (Tabela 56) por forma a maximizar a associação com a variável dependente e diminuir o número de variáveis *dummys*.

Tabela 56: Combinação das categorias na variável Instrução do encarregado de educação

Variável Original	Categoria Original	Categoria Nova	Variável Recodificada
Instrução do Encarregado de Educação	Sem Instrução	Sem Instrução	Instrução do Encarregado de Educação Combinada
	1º Ciclo do EBI 2º Ciclo do EBI 3º Ciclo do EBI	Ensino Básico Integrado	
	1º Ciclo do ES 2º Ciclo do ES 3º Ciclo do ES	Ensino Secundário	
	Curso Médio Curso Superior sem Licenciatura Outras	Ensino Médio	
	Licenciatura Mestrado / Pós Graduação	Ensino Superior	

A título de ilustração de um exemplo de um modelo de regressão linear múltipla com variável qualitativa Sexo (0 – Feminino, 1 – Masculino) como uma das variáveis independentes, e cujos coeficientes constam na Tabela 57, o modelo final seria assim definido:

$$y_j = \beta_0 + \beta_1 \times \text{NFPort}_j + \beta_2 \times \text{D}_{\text{sexo}} + \varepsilon_j = 1{,}340 + 0{,}784 \times \text{NFPort}_j + 0{,}370 \times \text{D}_{\text{sexo}} + \varepsilon_j$$

Especificamente, para os estados (sexo masculino e feminino), as equações deveriam assumir as seguintes escritas:

$$\text{Masculino: } y_j = 1{,}340 + 0{,}784 \times \text{NFPort}_j + 0{,}370 + \varepsilon_j$$

e

$$\text{Feminino: } y_j = 1{,}340 + 0{,}784 \times \text{NFPort}_j + \varepsilon_j$$

Tabela 57: Ilustração de um exemplo de regressão linear múltipla com "*dummy*" sexo

Modelo		Coeficientes não padronizados		Coeficientes padronizados		
		B	Erro	Beta	t	Sig.
2	(Constante)	1,340	,190		7,070	,000
	Nota Final a Português	,784	,014	,673	55,851	,000
	Sexo do aluno	,370	,075	,059	4,911	,000

De acordo com a Tabela 58, a primeira variável explicativa inserida no modelo, por isso denominado de Modelo 1, foi o Número de reprovações no ensino secundário. Assemelha-se ao modelo de regressão linear simples, tendo como variável dependente a Nota Final à Matemática e pela tabela, notamos que o Modelo 1 explica 4,6% da variabilidade das notas na disciplina de Matemática. A cada inserção de uma variável no modelo de regressão, é elaborado uma ANOVA da mudança da regressão (com a hipótese nula definida por H_0: o ajuste dos dados aos modelos são iguais) com o objetivo de testar a significância do modelo após a inserção da variável. Nesta ótica, a inserção da variável *dummy* Encarregados de educação que possuem o Ensino Superior como instrução (Modelo 2), revelou-se estatisticamente significativa com o valor das estatísticas da Mudança de $R^2 = 0,020$ e $p-\text{value} < 0,001$.

Portanto, o resumo da qualidade do modelo de regressão linear múltipla, dado pelo Modelo 10, nos fornece o coeficiente do $R^2_{\text{ajustado}} = 0,141$ que pode ser traduzido na percentagem da variação nas classificações na disciplina de Matemática que é explicada pela variação nas variáveis independentes que foram estatisticamente significativas para o modelo. Podemos, ainda na tabela, concluir pelas estatísticas da mudança que o Modelo 10 ajustou-se melhor aos dados, ou seja, que o modelo incluído a variável *dummy* Encarregados de edução sem instrução escolar melhorou, com $R^2_{\text{da mudança}} = 0,001$ e $p-\text{value} = 0,011 < 0,05$, a previsão do modelo. Ainda na tabela é possível analisar um dos pressupostos do modelo de regressão linear que é a independência dos resíduos, com a estatística de teste proposto por Durbin – Watson (Maroco, 2003) e que, segundo o mesmo autor, não rejeitamos a hipótese nula (os resíduos são independentes) caso, no

SPSS, a estatística de teste for igual a $2.0(\pm 0.2)$, portanto, podemos no limite da fronteira aceitar o pré-requisito da independência dos resíduos, a estatística de Durbin – Watson de 1,792.

Tabela 58: Resumo dos modelos de regressão linear múltipla com Nota Final à Matemática como resposta

					Estatísticas de mudança					
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Mudança de R quadrado	Mudança F	df1	df2	Sig. Mudança F	Durbin-Watson
1	,215[a]	,046	,046	3,043	,046	205,280	1	4242	,000	
2	,256[b]	,066	,065	3,012	,020	88,707	1	4241	,000	
3	,278[c]	,077	,077	2,993	,012	53,789	1	4240	,000	
4	,290[d]	,084	,083	2,983	,007	30,723	1	4239	,000	
5	,328[e]	,108	,107	2,944	,024	112,299	1	4238	,000	
6	,357[f]	,128	,126	2,911	,020	97,009	1	4237	,000	
7	,365[g]	,133	,131	2,903	,005	25,556	1	4236	,000	
8	,369[h]	,136	,135	2,897	,004	17,493	1	4235	,000	
9	,374[i]	,140	,138	2,892	,003	16,981	1	4234	,000	
10	,376[j]	,141	,139	2,890	,001	6,505	1	4233	,011	1,792

a. Preditores: (Constante), Número de reprovações no sistema secundário

b. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior

c. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno

d. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo

e. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno

f. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo

g. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta

h. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018

i. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

j. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado, Encarregados de Educação sem instrução escolar

A ideia que está por detrás da ANOVA da regressão linear múltipla é um teste de comparação entre os modelos ajustados em cada etapa com o modelo sem variável explicativa, sendo as hipóteses assim definidas:

H_0: o ajuste dos dados ao modelo com variáveis independentes é o mesmo que um modelo sem variáveis independentes;

versus

H_1: o ajuste dos dados ao modelo com variáveis independentes não é o mesmo que um modelo sem variáveis independentes;

Nesta ótica, a Tabela 59 nos mostra que o ajuste dos dados revelou-se estatisticamente sempre significativo aos modelos (Anexo 7) para variáveis não excluídas, pelo método *Stepwise*. Ainda podemos ver que variáveis como: Repetente no ano que estuda, Ano letivo 2016–2017, Distância da casa à escola e Encarregados de educação que possuem o ensino secundário como nível de instrução foram excluídas do modelo final (Modelo 10).

Tabela 59: ANOVA da regressão linear múltipla[1] com Nota Final à Matemática como resposta

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	1900,323	1	1900,323	205,280	,000[b]
	Resíduo	39269,241	4242	9,257		
	Total	41169,565	4243			
9	Regressão	5759,986	9	639,998	76,526	,000[j]
	Resíduo	35409,579	4234	8,363		
	Total	41169,565	4243			
10	Regressão	5814,320	10	581,432	69,613	,000[k]
	Resíduo	35355,245	4233	8,352		
	Total	41169,565	4243			

b. Preditores: (Constante), Número de reprovações no sistema secundário

j. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

k. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado, Encarregados de Educação sem instrução escolar

A Tabela 60 nos fornece os coeficientes em duas unidades de medidas, os padronizados cujo objetivo é usá-lo na comparação dos diferentes coeficientes e os não padronizados que estão na unidade de medida original dos dados e que, nesta fase é feita para cada coeficiente um teste *t* para indagar sobre a sua significância no modelo. Neste modelo final, as estimativas para as variáveis quantitativas foram, a ordenada na origem de $\beta_0 = 23{,}674$, os coeficientes de inclinações para o número de reprovações no sistema de

[1] Tabela apresentada de uma forma resumida. Consultar Anexo 7 para a tabela completa

ensino secundário $\left(\beta_1 = -,252\right)$ e idade $\left(\beta_5 = -0,618\right)$ que representam os efeitos que se espera na variável dependente quando o número de reprovações ou a idade aumenta unidade.

Um dos pressupostos da regressão linear múltipla é a ausência de multicolinearidade (associação entre as variáveis independentes). De acordo com Maroco (2003), se as variáveis independentes estão fortemente correlacionadas entre si, a análise do modelo de regressão ajustado pode ser extremamente confusa e desprovida de significado.

Podemos também, através da Tabela 60, decidir sobre a multicolinearidade entre as variáveis independentes usando para a estatística da Tolerância ou VIF (Fator de Inflação da Variância). De acordo com Maroco (2003), valores de $VIF = \frac{1}{1 - R_i^2}$ acima de 5 (*apud* Montgomery & Peck, 1982) ou mesmo acima de 10 (*apud* Myers, 1986) indicam problemas com a estimação de β_i devido a problemas de multicolinearidade nas variáveis independentes. Caso a estatística da Tolerância $T = \frac{1}{VIF} > 0,1$, aceitamos a inexistência de multicolinearidade, portanto, o pressuposto da não multicolinearidade entre as variáveis independentes se verifica.

Tabela 60: Coeficientes do modelo de regressão linear múltipla com NFMat como resposta

Modelo		Coeficientes não padronizados		Coeficientes padronizados			Estatísticas de colinearidade	
		B	Erro	Beta	t	Sig.	Tolerância	VIF
10	(Constante)	23,674	,744		31,814	,000		
	Número de reprovações no sistema secundário	-,252	,073	-,059	-3,452	,001	,703	1,423
	Encarregados de Educação que possuem como instrução o Ensino Superior	1,582	,219	,108	7,221	,000	,900	1,111
	Sexo do aluno	-,522	,091	-,083	-5,745	,000	,961	1,040
	Alunos do 1º Ciclo	-3,244	,220	-,520	-14,768	,000	,164	6,110
	Idade do aluno	-,618	,042	-,375	-14,609	,000	,308	3,251
	Alunos do 2º Ciclo	-1,661	,167	-,262	-9,961	,000	,294	3,398
	Escola que frequenta	-,630	,111	-,084	-5,684	,000	,936	1,068
	Ano letivo 2017 - 2018	-,401	,094	-,061	-4,250	,000	,996	1,004
	Encarregados de Educação que possuem como instrução o Ensino Básico Integrado	-,500	,104	-,080	-4,830	,000	,738	1,356
	Encarregados de Educação sem instrução escolar	-,396	,155	-,041	-2,551	,011	,773	1,294

De seguida apresentamos as Figuras 6 a 7 para analisarmos os pressupostos da regressão. Um outro pré-requisito na análise de regressão múltipla é que os resíduos ou os erros possuem distribuição normal com média nula e variância constante, ou seja, $\varepsilon_j \sim N(0,\sigma)$. Este pressuposto é validado pelas Figuras 6 e 7, que são respetivamente o Gráfico de normalidade dos resíduos padronizados e o PP normal da regressão. Quando se analisa a normalidade usando o PP Plot, esperamos numa perspetiva de adequação perfeita que os pontos se sobrepõem na totalidade à reta, e nesta ótica, podemos ver uma sobreposição dos resíduos à reta de regressão (Figura 7). Por outro lado, a Figura 6 mostra-nos que a curva da distribuição normal se adequa não na perfeição ao histograma dos resíduos, mas de uma forma bem admissível para aceitarmos a normalidade dos resíduos.

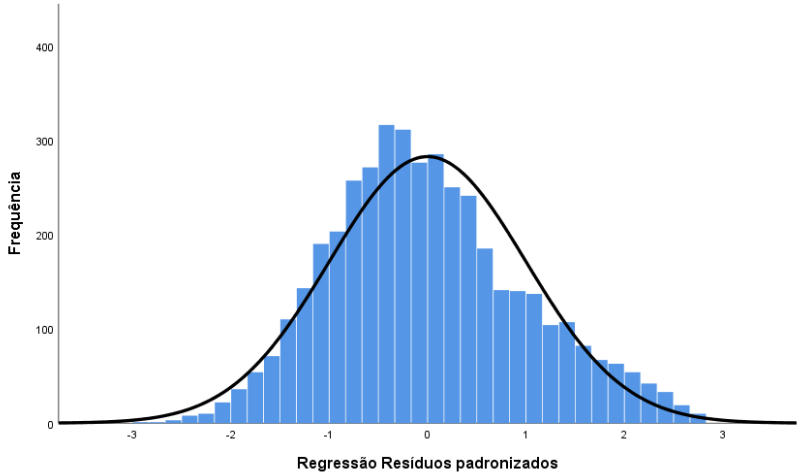


Figura 6: Gráfico de normalidade da regressão linear múltipla dos resíduos padronizados com NFMat como resposta

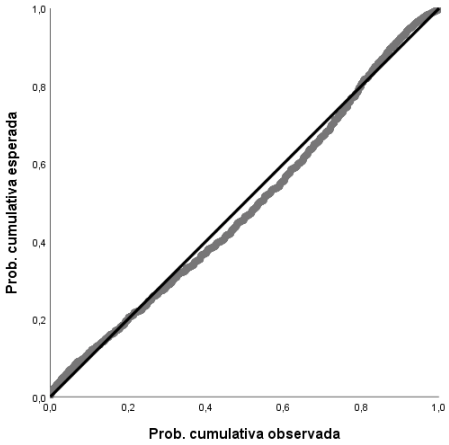


Figura 7: Gráfico P-P Plot Normal de regressão linear múltipla dos resíduos normalizados com NFMat como resposta

Ainda na esfera dos pré-requisitos, devemos analisar a homocedasticidade (variação constante dos resíduos em relação à reta de regressão) dos resíduos, que pode ser feita recorrendo à análise do gráfico de dispersão da regressão entre os valores preditos padronizados e os resíduos padronizados. Espera-se que a distribuição dos pontos (cada resíduo) esteja de uma forma aleatória numa perspetiva de se formarem uma "figura retangular", sendo que caso de não ocorrer a homocedasticidade dos resíduos, esta disposição estaria num formato aproximadamente cónico.

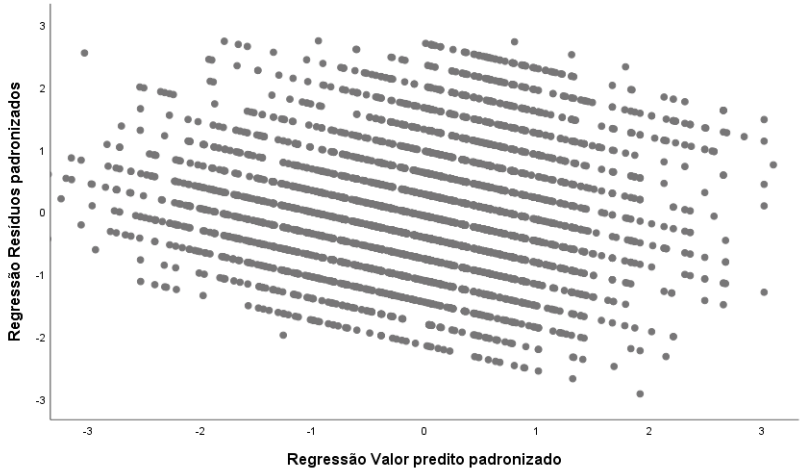


Figura 8: Gráfico de dispersão da regressão linear múltipla entre os valores preditos padronizados e os resíduos padronizados com NFMat como resposta

Verificada a ocorrência dos pressupostos da análise de regressão múltipla entre a variável dependente (Nota Final à Matemática) e as variáveis independentes que foram significativas para o modelo, a relação matemática que define o nosso modelo fica assim definida:

$$\begin{aligned}\text{NFMat}_j = {} & \beta_0 + \beta_1 \times \text{NRep} + \beta_2 \times Dummy_{\text{Inst\_ESup}} + \beta_3 \times Dummy_{\text{Sexo}} + \beta_4 \times Dummy_{\text{Ciclo1}} + \beta_5 \times \text{Idade} \\ & + \beta_6 \times Dummy_{\text{Ciclo2}} + \beta_7 \times Dummy_{\text{Escola}} + \beta_8 \times Dummy_{\text{Anolet17-18}} + \beta_9 \times Dummy_{\text{Inst\_EBI}} + \\ & + \beta_{10} \times Dummy_{\text{Sem\_Inst}} + \varepsilon_j\end{aligned}$$

Ou seja,

$$\begin{aligned}\text{NFMat} = {} & 23{,}674 - 0{,}252 \times \text{NRep} + 1{,}582 \times D_{\text{Inst\_ESup}} - 0{,}522 \times D_{\text{Sexo}} - 3{,}244 \times D_{\text{Ciclo1}} - \\ & - 0{,}618 \times \text{Idade} - 1{,}661 \times D_{\text{Ciclo2}} - 0{,}630 \times D_{\text{Escola}} - 0{,}401 \times D_{\text{Anolet17-18}} - \\ & - 0{,}500 \times D_{\text{Inst\_EBI}} - 0{,}396 \times D_{\text{Sem\_Inst}} + \text{erro}\end{aligned}$$

Após ter findado o modelo de regressão linear múltipla tendo como variável resposta Nota Final à Matemática, passaremos de seguida ao ajuste dos dados definida agora a Nota Final à Português como variável dependente. Sabemos que a presença de *outliers* podem influenciar na estimativa dos coeficientes do nosso modelo de regressão linear múltipla. Mesmo sendo a presença de *outliers* insignificante para o nosso estudo (menos de 1% $\left(7/4244\right)$), decidimos eliminá-los, apresentando como critérios de comparação as estatísticas dos coeficientes sem remover os *outliers* em Anexos.

As Figuras 9 e 10, permitem-nos concluir que a validação do pressuposto da normalidade dos resíduos se verifica assim como a homocedasticidade (Figura 11).

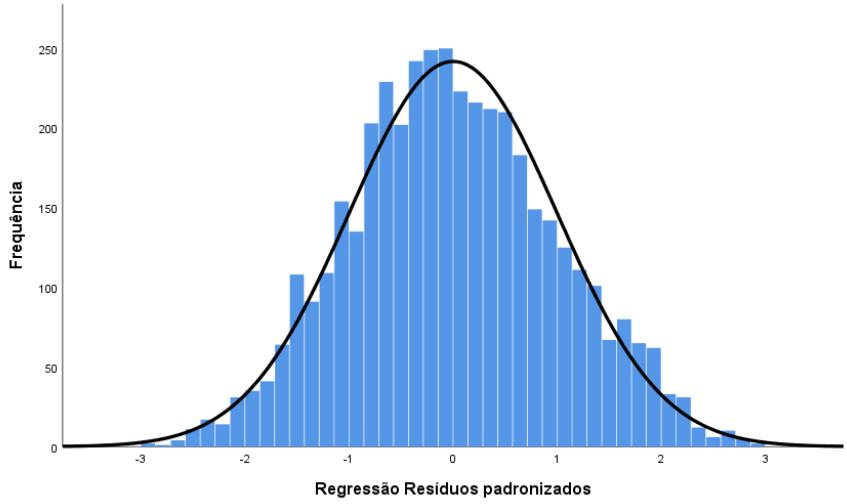


Figura 9: Gráfico de normalidade da regressão linear múltipla dos resíduos padronizados com NFPort como resposta

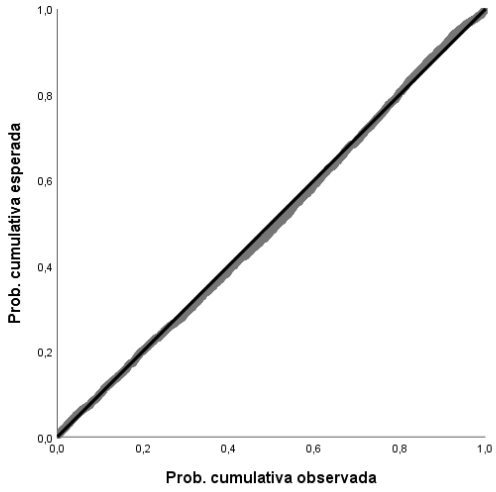


Figura 10: Gráfico P-P Plot Normal de regressão linear múltipla dos resíduos normalizados com NFPort como resposta

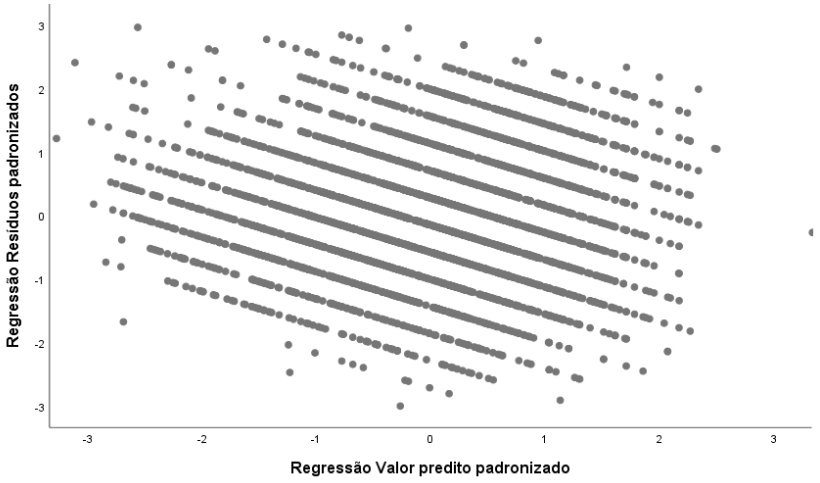


Figura 11: Gráfico de dispersão da regressão linear múltipla entre os valores preditos padronizados e os resíduos padronizados com NFPort como resposta

Querendo concluir, se para este modelo teremos problemas com a independência dos resíduos, notamos através da estatística de Durbin-Watson (Tabela 61), que a validação do pressuposto de independência dos resíduos pode ser validada pelo fato da estatística de Darbin-Watson estar aproximadamente entre 1,8 e 2,2. Ainda e na tabela seguinte, não devemos aceitar a hipótese de que o Modelo 15 ajuste melhor ao ajustamento dos dados comparativamente ao Modelo 14, pois o teste ANOVA para a estatística desta mudança nos forneceu, para um nível de significação de 5%, $R^2_{\text{da mudança}} < 0,001$ e $p-\text{value} = 0,158 > 0,05$. Portando, o nosso modelo, sem as variáveis *dummys* Ano letivo 2016–2017 e Alunos que distam entre 3 à 6 km da escola (Anexo 9), explica 23,1% da variação das classificações finais na disciplina de Português ($R^2 = 0,231$ e $p-\text{value} < 0,001$). De realçar que a remoção dos *outliers* melhorou a estimativa dos coeficientes R^2 e a estatística de Durbin–Watson (Anexo 10).

Tabela 61: Resumo dos modelos de regressão linear múltipla com Nota Final à Português como resposta

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
					Mudança de R quadrado	Mudança F	df1	df2	Sig. Mudança F	
1	,286[a]	,082	,081	2,552	,082	376,315	1	4235	,000	
2	,354[b]	,125	,125	2,492	,043	210,459	1	4234	,000	
3	,387[c]	,150	,149	2,456	,025	122,885	1	4233	,000	
4	,428[d]	,183	,183	2,408	,034	173,711	1	4232	,000	
5	,448[e]	,201	,200	2,383	,017	91,268	1	4231	,000	
6	,458[f]	,210	,209	2,369	,009	49,064	1	4230	,000	
7	,466[g]	,217	,215	2,359	,007	38,192	1	4229	,000	
8	,468[h]	,219	,218	2,355	,002	13,215	1	4228	,000	
9	,470[i]	,221	,219	2,353	,002	9,458	1	4227	,002	
10	,471[j]	,222	,220	2,352	,001	5,647	1	4226	,018	
11	,472[k]	,223	,221	2,351	,001	4,888	1	4225	,027	
12	,473[l]	,224	,222	2,349	,001	5,835	1	4224	,016	
13	,474[m]	,225	,223	2,348	,001	5,174	1	4223	,023	
14	,481[n]	,231	,229	2,339	,006	34,246	1	4222	,000	
15	,480[o]	,231	,228	2,339	,000	1,990	1	4222	,158	1,796

a. Preditores: (Constante), Sexo do aluno

b. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário

c. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo

d. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno

e. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo

f. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior

g. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário

h. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda

i. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018

j. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola

k. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta

l. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola

m. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar

n. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

o. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

Pela análise da Tabela 62 (as ANOVAS para os modelos 1 a 11 foram excluídas na apresentação), que nos fornece a ANOVA da regressão múltipla, podemos confirmar que apesar do Modelo 15 ter sido significativo, a qualidade deste ajuste não é superior que a do Modelo 14.

Tabela 62: ANOVA da regressão linear múltipla com Nota Final à Português como resposta

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
12	Regressão	6728,841	12	560,737	101,597	,000[m]
	Resíduo	23313,307	4224	5,519		
	Total	30042,148	4236			
13	Regressão	6757,371	13	519,798	94,272	,000[n]
	Resíduo	23284,777	4223	5,514		
	Total	30042,148	4236			
14	Regressão	6944,722	14	496,052	90,674	,000[o]
	Resíduo	23097,426	4222	5,471		
	Total	30042,148	4236			
15	Regressão	6933,837	13	533,372	97,473	,000[p]
	Resíduo	23108,311	4223	5,472		
	Total	30042,148	4236			

m. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola

n. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar

o. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

p. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

A Tabela 63 nos mostra as estimativas dos coeficientes significativos para o nosso modelo.

Tabela 63: Coeficientes do modelo de regressão linear múltipla com NFPort como resposta

Modelo		Coeficientes não padronizados		Coeficientes padronizados			Estatísticas de colinearidade	
		B	Erro	Beta	t	Sig.	Tolerância	VIF
14	(Constante)	23,107	,633		36,527	,000		
	Sexo do aluno	-1,240	,074	-,232	-16,824	,000	,958	1,044
	Número de reprovações no sistema secundário	-,167	,066	-,045	-2,526	,012	,564	1,774
	Alunos do 1º Ciclo	-2,822	,179	-,529	-15,732	,000	,161	6,211
	Idade do aluno	-,497	,035	-,352	-14,395	,000	,304	3,291
	Alunos do 2º Ciclo	-1,088	,136	-,200	-8,009	,000	,291	3,438
	Encarregados de Educação que possuem como instrução o Ensino Superior	,345	,244	,028	1,411	,038	,477	2,098
	Encarregados de Educação que possuem como instrução o Ensino Secundário	-,528	,194	-,087	-2,717	,007	,177	5,639
	Repetente no ano em que estuda	-,411	,118	-,057	-3,496	,000	,685	1,461
	Ano letivo 2017 - 2018	-,225	,077	-,040	-2,932	,003	,993	1,007
	Alunos que distam a entre 1 e 3 km da escola	,327	,099	,059	3,318	,001	,569	1,759
	Escola que frequenta	-,330	,095	-,051	-3,457	,001	,830	1,204
	Alunos que distam a menos de 1 km da escola	,204	,101	,037	2,023	,043	,537	1,864
	Encarregados de Educação sem instrução escolar	-1,301	,212	-,159	-6,153	,000	,273	3,668
	Encarregados de Educação que possuem como instrução o Ensino Básico Integrado	-1,109	,190	-,208	-5,852	,000	,144	6,926

Neste sentido, a equação final fica assim definida:

$$\begin{aligned}\text{NFPort} = {} & 23{,}107 - 1{,}240 \times D_{\text{Sexo}} - 0{,}167 \times \text{NRep} - 2{,}822 \times D_{\text{Ciclo1}} - 0{,}497 \times \text{Idade} + \\ & + 0{,}345 \times D_{\text{Inst\_ESup}} - 0{,}528 \times D_{\text{Inst\_ESec}} - 0{,}411 \times D_{\text{Repetente}} - 0{,}225 \times D_{\text{Anolet17-18}} + \\ & + 0{,}327 \times D_{\text{Dist\_1-3km}} - 0{,}330 \times D_{\text{Escola}} + 0{,}204 \times D_{\text{Dist\_inf\_1km}} - 1{,}301 \times D_{\text{Sem\_Inst}} - \\ & - 1{,}109 \times D_{\text{Inst\_EBI}} + \text{erro}\end{aligned}$$

CONCLUSÃO E TRABALHOS FUTUROS

Em qualquer processo, seja ele pessoal, profissional ou outro, é natural que sempre que uma etapa termina façamos balanços, pesando os prós e contras no sentido de quiçá encontrarmos as respostas necessárias para melhorar o processo lá onde estiver positivo e corrigir erros que naturalmente possam surgir.

Infelizmente e de uma forma corrente, em toda a vivência como docente, as conclusões sobre a avaliação de desempenho dos alunos têm sido única e exclusivamente baseadas em: qual a taxa de aproveitamento? Mas como será possível minimizar todo um processo de ensino aprendizagem ao facto de ter obtido nota inferior ou não a 10 valores (dependendo da escala que se usa), sendo que este valor foi convencionalmente estipulado.

Não é efetivamente correto tirar conclusões sobre o fim de um processo se não soubermos quais os caminhos que levaram para tal. Há que trilhar um caminho com passos firmes, há que entender que avaliar todo o processo de uma comunidade educativa sem analisar todas as variáveis intervenientes neste processo não é a atitude mais correta. Há que criar mecanismos necessários para entender o cerne da questão e analisar a possibilidade de abordar o processo como sendo um processo e não o fim.

A nossa ideia de partida para este trabalho, era de ilustrar a necessidade de se aplicar o modelo de regressão multinível aplicada a dados educacionais. Esta nossa presunção deve-se ao fato da forma como os alunos, possuindo suas caraterísticas intrínsecas, estão sujeitos a outras variáveis da sala de aula, dos professores, da escola, meio envolvente, etc., e que estas, de uma forma conjugada ou não, podem influenciar o desempenho escolar dos discentes.

Este estudo provou-nos a necessidade de no início de um trabalho como este, mesmo com a definição dos itens necessários para a sua prossecução, não limitarmos o nosso fio condutor na tentativa de prever a resposta ideal que almejávamos. As informações são as informações e elas falam por si, cabendo a nós somente a capacidade de ler o que elas

nos permitirem ler, de escrever sobre somente o que elas permitirem escrever, e de concluir somente o que elas permitirem concluir.

A nossa primeira etapa na abordagem das informações referentes as classificações dos alunos das duas escolas secundárias da cidade do Porto Novo nas disciplinas de Matemática e Português, nos mostrou que a nossa vontade de ajustar os dados a um modelo de regressão multinível tendo como fator a escola não seria viável (estatisticamente não significativo), pois o modelo ANOVA com coeficientes aleatórios (nulo ou vazio) da regressão multinível forneceu-nos um Coeficiente de Correlação Intraclasse de 0,07% $\left(p-\text{value}=0,721\right)$ e 0,1% $\left(p-\text{value}=0,679\right)$, respetivamente, para a Nota Final à Matemática e à Português. Não sendo uma regra, que para estudos diferentes sobre uma mesma base de dados, as conclusões podem ser generalizadas, a verificação da impossibilidade do ajuste dos dados a um modelo de regressão multinível usando como variáveis respostas a Nota Final à Matemática e à Português veio confirmar, pelas evidências estatísticas fornecidas pelos modelos nulos, o que a Tabela 29 nos havia indicado, ou seja, que a escola não possuía efeito estatisticamente significativo sobre as classificações finais nas duas disciplinas.

Revelando impossível tal fato descrito acima, optamos pela desagregação da nossa amostra em três partes: 1º ciclo – composto pelos alunos do 7º e 8º ano de escolaridade; 2º ciclo – composto pelos alunos do 9º e 10º anos de escolaridade e pelo 3º ciclo – com os alunos do 11º e 12º ano de escolaridade. Novamente esta decisão não trouxe a tão almejada possibilidade de ajustar os dados dos ciclos de estudos a um modelo de regressão multinível usando a escola como fator, pois para: 1º ciclo – ($\rho_{\text{Mat}}=0,0195$, $p-\text{value}_{\text{Mat}}=0,511$, $\rho_{\text{Port}}=0,0223$, $p-\text{value}_{\text{Port}}=0,663$); 2º ciclo – ($\rho_{\text{Mat}}=0,0040$, $p-\text{value}_{\text{Mat}}=0,620$, $\rho_{\text{Port}}=0,0304$, $p-\text{value}_{\text{Port}}=0,503$); 3º ciclo – (verificou-se um problema de redundância no cálculo do Coeficiente de Correlação Intraclasse para a variável Nota Final à Matemática, $\rho_{\text{Port}}=0,1252$, $p-\text{value}_{\text{Port}}=0,494$).

Não sendo estatisticamente viável o ajuste dos dados a um modelo de regressão multinível, e uma vez que na nossa base de dados havia mais que uma possível variável

explicativa, optamos pelo modelo de regressão linear múltipla usando o método passo a passo (*Stepwise*) na inserção das variáveis no modelo, mantendo as duas variáveis respostas definidas anteriormente.

O coeficiente de correlação de Spearman comprovou a relação positiva entre as classificações nas disciplinas de Matemática e Português, corroborando mais a necessidade de se continuar a investir nestas duas disciplinas, mas de uma forma agregada. Por outro lado, as evidências estatísticas nos mostraram a existência de uma correlação moderada negativa entre o número de reprovações e as classificações finais nas duas disciplinas.

Um dos pontos fracos que podemos apontar aos nossos modelos, é a de que percentagem da variabilidade das notas nas disciplinas de Matemática e Português que cada um consegue explicar é relativamente baixa, sendo respetivamente de, 14,1% e 23,1%, para a Nota Final à Matemática e à Português. Destacamos as variáveis globais como o Número de reprovações no sistema, a idade e as variáveis *dummys* relacionadas com o nível de instrução dos encarregados de educação. A medida que a idade ou o número de reprovações aumenta, verifica-se um efeito negativo nas classificações das disciplinas de Matemática e Português, sendo que o efeito da idade é maior em ambas (de $-0{,}618$ para a disciplina de Matemática e $-0{,}497$ para a de Português). A presença das variáveis *dummys* Encarregados de educação sem nível de instrução, com o EBI influenciam também de uma forma negativa as classificações finais das duas disciplinas definidas como variáveis respostas. Por outo lado, a presença da variável *dummy* Encarregados de educação com o Ensino Superior como o nível de instrução tende a aumentar as notas finais nas disciplinas de Matemática e Português (1,502 e 0,345, respetivamente, para a Nota Final à Matemática e à Português). Mas dado ao desequilíbrio na distribuição das categorias, visto que a percentagem de encarregados de educação que possuem como instrução o nível superior é de somente 5%, enquanto que os com o EBI e os sem nível de instrução são, cumulativamente, cerca de 65%, o modelo não consegue aferir de uma forma mais exata o peso da mesma. Uma possibilidade seria

a de trabalhar com amostras mais equilibradas nas categorias, mas esta eventualidade levaria a uma perda de informação substancial.

Pensamos que em futuros estudos desta natureza se deva investir mais em outras variáveis, ou seja, que haja na base de dados mais informações sobre os professores, sobre as escolas, sobre as famílias, por forma que a capacidade explicativa de um determinado modelo seja melhorada.

No nosso estudo, a modelo de regressão linear múltipla, para cada variável resposta ficou definida pela relação assim definida:

- Nota Final à Matemática como variável resposta:

$$\begin{aligned}\text{NFMat} = {} & 23{,}674 - 0{,}252 \times \text{NRep} + 1{,}582 \times D_{\text{Inst\_ESup}} - 0{,}522 \times D_{\text{Sexo}} - 3{,}244 \times D_{\text{Ciclo1}} - \\ & - 0{,}618 \times \text{Idade} - 1{,}661 \times D_{\text{Ciclo2}} - 0{,}630 \times D_{\text{Escola}} - 0{,}401 \times D_{\text{Anolet17-18}} - \\ & - 0{,}500 \times D_{\text{Inst\_EBI}} - 0{,}396 \times D_{\text{Sem\_Inst}}\end{aligned}$$

- Nota Final à Português como resposta:

$$\begin{aligned}\text{NFPort} = {} & 23{,}107 - 1{,}240 \times D_{\text{Sexo}} - 0{,}167 \times \text{NRep} - 2{,}822 \times D_{\text{Ciclo1}} - 0{,}497 \times \text{Idade} + \\ & + 0{,}345 \times D_{\text{Inst\_ESup}} - 0{,}528 \times D_{\text{Inst\_ESec}} - 0{,}411 \times D_{\text{Repetente}} - 0{,}225 \times D_{\text{Anolet17-18}} + \\ & + 0{,}327 \times D_{\text{Dist\_1-3km}} - 0{,}330 \times D_{\text{Escola}} + 0{,}204 \times D_{\text{Dist\_inf\_1km}} - 1{,}301 \times D_{\text{Sem\_Inst}} - \\ & - 1{,}109 \times D_{\text{Inst\_EBI}}\end{aligned}$$

BIBLIOGRAFIA

Aguinis, H, Gottfredson, K, & Culpepper, A. (2013). **Best–Practice Recommendations for Estimating Cross–Level Interaction Effects Using Multilevel Modeling**. Journal of Management, 39(6). Disponível em https://doi.org/10.1177/0149206313478188 1490–1528. [01 de Fevereiro de 2020].

Albright, J. & Maronova, D. M (2015). **Estimating Multilevel Models using SPSS, Stata, SAS and R**. Disponível em https://hdl.handle.net/2022/19737 [12 de Dezembro de 2019].

Bergamo, G. C. (2002). **Aplicação de Modelos Multiníveis na Análise de Dados de Medidas Repetidas no Tempo**. Tese de Mestrado, Escola Superior de Agricultura Luiz de Queiroz: Universidade de São Paulo, Piracicaba. doi.10.11606/D.11.2002.tde-08012003-083811. Disponível em www.teses.usp.br [14 de Junho de 2019].

Bono, Roser & Arnau, Jaume & Balluerka, Nekane. (2007). **Using linear mixed models in longitudinal studies: Application of SAS PROC MIXED.** REMA, ISSN 1135-6855, Vol. 12, Nº. 2, 2007, pags. 15-31. 12. Disponível em https://www.researchgate.net/publication/28183327 [06 de Janeiro de 2020].

Chasco, C. & Lopez, A. M. (2011). **Modelos Multiníveis: uma aplicação ao modelo de convergência beta**. Disponível em https://www.researchgate.net/publication/256706776 [10 de Outubro de 2019].

Castro, C. S. (2015). **Aplicação de Modelos Multiníveis para o Estudo da Sinistralidade no Retalho Alimentar**. Tese de Mestrado: Faculdade de Ciências da Universidade do Porto. Disponível em https://dhll.handle.net/10216/82347 [16 de Junho de 2019].

Cerqueira, R. I. (2016). **Fatores de complicações após transplantes renais: Análise Estatística Multivariada**. Tese de Mestrado: Universidade Aberta. Disponível em http://hdl.handle.net/10400.2/6332 [27 de Outubro de 2019].

Coelho, F. H. (2017). **Seleção de Modelos Multiníveis para Dados de Avaliação Educacional**. Tese de Mestrado: Universidade Federal de São Carlos. Disponível em

https://repositorio.ufscar.br/bitstream/handle/ufscar/9429/DissFRC.pdf?sequence=1&isAllowed=y [12 de Fevereiro de 2019].

Cruz, C. C. M. F. (2010). **Modelos Multi-nível: Fundamentos e Aplicações**. Tese de Mestrado: Universidade Aberta. Disponível em http://hdl.handle.net/10400.2/1729 [20 de Setembro de 2019].

Dupont, E. & Martensen, H. (Eds) (2007). **Multilevel modeling and time series analyses in traffic research – Methodology**. Deliverable D7.4 of the EU FP6 project SafetyNet. Disponível em https://trimis.ec.europa.eu/sites/default/files/project/documents/20130131_135329_47313_D7.4.pdf [10 de Janeiro de 2020].

Fausto, M. & Carneiro, M. & Antunes, C. & Pinto, J. & Colosimo, E. (2008). **O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados**. Cadernos De Saude Publica - CAD SAUDE PUBLICA. 24. 10.1590/S0102-311X2008000300005. Disponível em https://www.researchgate.net/publication/250026818 [08 de Novembro de 2019]

Field, A. (2009). **Descobrindo a Estatística usando o SPSS**. Porto Alegre: Artmed. Disponível em https://docero.com.br/doc/e18vnn- [15 de Maio de 2019].

Gelmam, A. Hill, J. (2007). **Data Analyses Using Regression and Multilevel / Hierarchical Models**. Cambridge University Press.

Goldstein, H. (1999). **Multilevel Statistical Models**. London: Institute of Education. Disponível em https://stats.idre.ucla.edu/wp-content/uploads/2016/02/goldstein-1.pdf [31 de Dezembro de 2019].

Laros, Jacob & Marciano, Joao. (2008). **Análise multinível aplicada aos dados do NELS:88. Estudos em Avaliação Educacional**. 19. 263-278. 10.18222/eae194020082079. Disponível em https://www.scielosp.org/article/csp/2008.v24n3/513-524/ [12 de Outubro de 2019].

Maia, J. et al (2003). **Modelação hierárquica ou multinível. Uma metodologia estatística e um instrumento útil de pensamento na investigação em Ciências do Desporto**. Disponível em http://hdl.handle.net/10198/3175 [08 de Janeiro de 2020].

Maroco, J. (2003). **Análise Estatística com utilização do SPSS**. Edições Sílabo Lda: 2ª Edição. Lisboa.

Osio, M. M. G. (2013). **Análise de Modelos de Regressão Multiníveis Simétricos**. Tese de Mestrado, Instituto de Ciências Matemáticas e de Computação: Universidade de São Paulo, São Carlos. doi.10.11606/D.55.2013.tde-05072013-161440. Disponível em www.teses.usp.br [27 de Julho de 2019].

O'Dwyer, L. M. & Parker, C. E. (2014). **A primer for analysing nested data: multilevel modeling in SPSS using an exemple from a REL study.** Disponível em https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2015046.pdf [08 de Dezembro de 2019].

Oliveira, T. C. A. (2004). **Estatística Aplicada***.* Universidade Aberta, Lisboa.

Pereira, A. (2008). **SPSS – Guia Prático de Utilização**. Edições Sílabo Lda: 7ª Edição. Lisboa.

Pinheiro, S. M. C. (2005). **Modelo Linear Hierárquico: um Método Alternativo para Análise de Desempenho Escolar**. Tese de Mestrado: Universidade Federal de Pernambuco. Disponível em https://repositorio.ufpe.br/handle/123456789/6476 [17 de Julho de 2019].

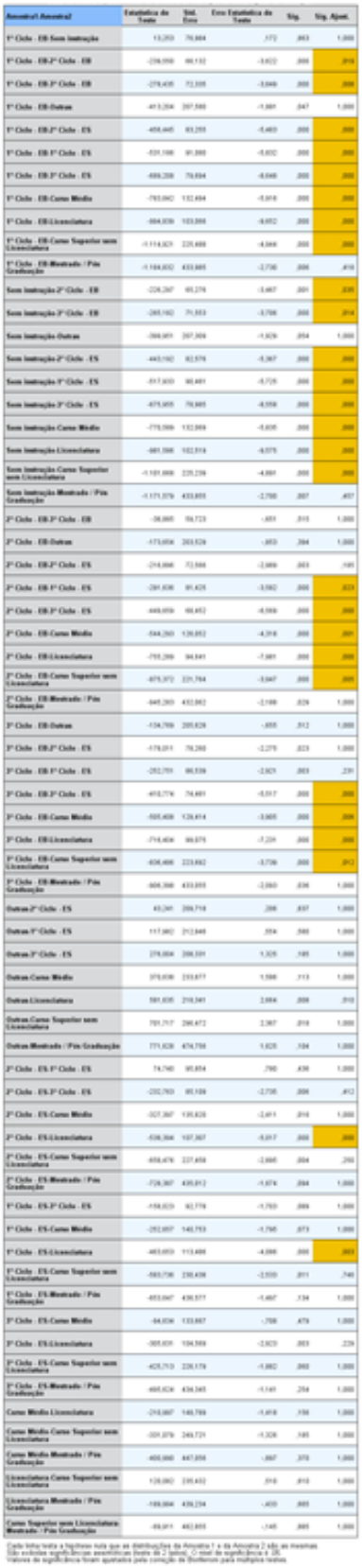
Reis, E., et al. (2008). **Estatística Aplicada Vol. 2**. Edições Sílabo Lda: 4ª Edição. Lisboa

Reis, E. (2001). **Estatística Multivariada Aplicada**. Edições Sílabo Lda: 2ª Edição. Lisboa.

Veech, J. (2012). **Significance testing in ecological null models. Theoretical Ecology**. 5. 10.1007/s12080-012-0159-z. https://www.researchgate.net/publication/257769520 [10 de Dezembro de 2019].

ANEXOS

Anexo 1: Estatística para o teste de comparação múltipla referente as variáveis respostas e Instrução do Encarregado de Educação com *p* ajustado pela correção de Bonferroni



Anexo 2: Correlação entre as variáveis Notas Finais à Matemática e à Português em relação à escola

	Escola que frequenta			Nota Final a Português	Nota Final a Matemática
rô de Spearman	ESASP	Nota Final a Português	Coeficiente de Correlação	1,000	,584
			Sig. (2 extremidades)	.	,000
			N	930	930
		Nota Final a Matemática	Coeficiente de Correlação	,584	1,000
			Sig. (2 extremidades)	,000	.
			N	930	930
	ETJV	Nota Final a Português	Coeficiente de Correlação	1,000	,622
			Sig. (2 extremidades)	.	,000
			N	3314	3314
		Nota Final a Matemática	Coeficiente de Correlação	,622	1,000
			Sig. (2 extremidades)	,000	.
			N	3314	3314

Anexo 3: Correlação entre as variáveis Notas Finais à Matemática e à Português em relação à Instrução do encarregado de educação

Instrução do encarregado de educação			Nota Final a Português	Nota Final a Matemática
Sem Instrução	Nota Final a Português	Coeficiente de Correlação	1,000	,556
		Sig. (2 extremidades)	.	,000
		N	510	510
	Nota Final a Matemática	Coeficiente de Correlação	,556	1,000
		Sig. (2 extremidades)	,000	.
		N	510	510
Ensino Básico Integrado	Nota Final a Português	Coeficiente de Correlação	1,000	,571
		Sig. (2 extremidades)	.	,000
		N	2256	2256
	Nota Final a Matemática	Coeficiente de Correlação	,571	1,000
		Sig. (2 extremidades)	,000	.
		N	2256	2256
Ensino Secundário	Nota Final a Português	Coeficiente de Correlação	1,000	,640
		Sig. (2 extremidades)	.	,000
		N	1105	1105
	Nota Final a Matemática	Coeficiente de Correlação	,640	1,000
		Sig. (2 extremidades)	,000	.
		N	1105	1105
Ensino Médio	Nota Final a Português	Coeficiente de Correlação	1,000	,709
		Sig. (2 extremidades)	.	,000
		N	170	170
	Nota Final a Matemática	Coeficiente de Correlação	,709	1,000
		Sig. (2 extremidades)	,000	.
		N	170	170
Ensino superior	Nota Final a Português	Coeficiente de Correlação	1,000	,751
		Sig. (2 extremidades)	.	,000
		N	203	203
	Nota Final a Matemática	Coeficiente de Correlação	,751	1,000
		Sig. (2 extremidades)	,000	.
		N	203	203

Anexo 4: *Dummys* para a variável Nível de Instrução do Encarregado de Educação

	$Dummy_{EBI}$	$Dummy_{ES}$	$Dummy_{EM}$	$Dummy_{ESup}$
Sem Instrução	0	0	0	0
Ensino Básico Integrado	1	0	0	0
Ensino Secundário	0	1	0	0
Ensino Médio	0	0	1	0
Ensino Superior	0	0	0	1

Anexo 5: Dummys para a variável Ano Letivo

	$Dummy_{Alet1617}$	$Dummy_{Alet1718}$
Ano letivo 2016 – 2017	1	0
Ano letivo 2017 – 2016	0	1
Ano letivo 2018 – 2019	0	0

Anexo 6: Dummys para a variável Ciclo de Estudo

	$Dummy_{Ciclo1}$	$Dummy_{Ciclo2}$
1º Ciclo	1	0
2º Ciclo	0	1
3º Ciclo	0	0

Anexo 7: ANOVA da regressão linear múltipla com Nota Final à Matemática como resposta

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
2	Regressão	2704,873	2	1352,436	149,116	,000[c]
	Resíduo	38464,692	4241	9,070		
	Total	41169,565	4243			
3	Regressão	3186,729	3	1062,243	118,578	,000[d]
	Resíduo	37982,836	4240	8,958		
	Total	41169,565	4243			
4	Regressão	3460,036	4	865,009	97,237	,000[e]
	Resíduo	37709,528	4239	8,896		
	Total	41169,565	4243			
5	Regressão	4433,477	5	886,695	102,292	,000[f]
	Resíduo	36736,087	4238	8,668		
	Total	41169,565	4243			
6	Regressão	5255,748	6	875,958	103,343	,000[g]
	Resíduo	35913,817	4237	8,476		
	Total	41169,565	4243			
7	Regressão	5471,122	7	781,589	92,744	,000[h]
	Resíduo	35698,443	4236	8,427		
	Total	41169,565	4243			
8	Regressão	5617,974	8	702,247	83,654	,000[i]
	Resíduo	35551,590	4235	8,395		
	Total	41169,565	4243			
10	Regressão	5814,320	10	581,432	69,613	,000[k]
	Resíduo	35355,245	4233	8,352		
	Total	41169,565	4243			

c. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior

d. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno

e. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo

f. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno

g. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo

h. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta

i. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018

k. Preditores: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado, Encarregados de Educação sem instrução escolar

Anexo 8: Variáveis excluídas no modelo final da regressão linear múltipla com Nota Final à Matemática como resposta

Modelo		Beta In	t	Sig.	Correlação parcial	Estatísticas de colinearidade		
						Tolerância	VIF	Tolerância mínima
10	Repetente no ano em que estuda	,029[k]	1,696	,090	,026	,686	1,457	,162
	Ano letivo 2016 - 2017	,010[k]	,588	,557	,009	,738	1,356	,163
	Alunos que distam a menos de 1 km da escola	,012[k]	,785	,432	,012	,934	1,071	,164
	Alunos que distam a entre 1 e 3 km da escola	,018[k]	1,269	,205	,019	,987	1,013	,163
	Alunos que distam a entre 3 e 6 km da escola	-,006[k]	-,442	,659	-,007	,979	1,021	,163
	Encarregados de Educação que possuem como instrução o Ensino Secundário	-,035[k]	-1,033	,302	-,016	,179	5,588	,146

k. Preditores no Modelo: (Constante), Número de reprovações no sistema secundário, Encarregados de Educação que possuem como instrução o Ensino Superior, Sexo do aluno, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Escola que frequenta, Ano letivo 2017 - 2018, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado, Encarregados de Educação sem instrução escolar

Anexo 9: Variáveis excluídas no modelo final da regressão linear múltipla com Nota Final à Português como resposta

Modelo		Beta In	t	Sig.	Correlação parcial	Estatísticas de colinearidade		
						Tolerância	VIF	Tolerância mínima
14	Ano letivo 2016 - 2017	,004[o]	,256	,798	,004	,721	1,387	,144
	Alunos que distam a entre 3 e 6 km da escola	,009[o]	,518	,604	,008	,629	1,590	,144
15	Ano letivo 2016 - 2017	,004[p]	,262	,794	,004	,721	1,387	,160
	Alunos que distam a entre 3 e 6 km da escola	,009[p]	,528	,598	,008	,629	1,590	,161
	Encarregados de Educação que possuem como instrução o Ensino Superior	,028[p]	1,411	,158	,022	,477	2,098	,144

o. Preditores no Modelo: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

p. Preditores no Modelo: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Alunos que distam a menos de 1 km da escola, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

Anexo 10: Resumo dos modelos de regressão linear múltipla com Nota Final à Português como resposta sem remover Outliers

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
					Mudança de R quadrado	Mudança F	df1	df2	Sig. Mudança F	
1	,284[a]	,081	,081	2,563	,081	372,899	1	4242	,000	
2	,349[b]	,122	,122	2,505	,041	199,576	1	4241	,000	
3	,383[c]	,147	,146	2,470	,025	121,908	1	4240	,000	
4	,423[d]	,179	,178	2,423	,032	167,637	1	4239	,000	
5	,443[e]	,196	,195	2,398	,017	88,615	1	4238	,000	
6	,452[f]	,205	,203	2,385	,009	46,262	1	4237	,000	
7	,460[g]	,211	,210	2,376	,007	35,567	1	4236	,000	
8	,462[h]	,214	,212	2,372	,003	13,750	1	4235	,000	
9	,464[i]	,216	,214	2,369	,002	10,225	1	4234	,001	
10	,466[j]	,217	,215	2,368	,001	5,807	1	4233	,016	
11	,467[k]	,218	,216	2,367	,001	5,685	1	4232	,017	
12	,468[l]	,219	,217	2,365	,001	5,465	1	4231	,019	
13	,474[m]	,225	,222	2,357	,006	33,055	1	4230	,000	
14	,474[n]	,224	,222	2,357	,000	2,062	1	4230	,151	1,789

a. Preditores: (Constante), Sexo do aluno

b. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário

c. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo

d. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno

e. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo

f. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior

g. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário

h. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda

i. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018

j. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola

k. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta

l. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Encarregados de Educação sem instrução escolar

m. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Superior, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado

n. Preditores: (Constante), Sexo do aluno, Número de reprovações no sistema secundário, Alunos do 1º Ciclo, Idade do aluno, Alunos do 2º Ciclo, Encarregados de Educação que possuem como instrução o Ensino Secundário, Repetente no ano em que estuda, Ano letivo 2017 - 2018, Alunos que distam a entre 1 e 3 km da escola, Escola que frequenta, Encarregados de Educação sem instrução escolar, Encarregados de Educação que possuem como instrução o Ensino Básico Integrado