

ESTIMATIVA DE DESEMPENHO ACADÊMICO A PARTIR DE MÉTODOS ESTATÍSTICOS PREDITIVOS: UM ESTUDO DE CASO COM ALUNOS DO 1º ANO DO ENSINO MÉDIO DO INSTITUTO FEDERAL DO CEARÁ, CAMPUS CRATO

Robson G. Fechine Feitosa¹, Guilherme A. R. M. Esmeraldo², Cícero Carlos F. de Oliveira³

RESUMO

O processo de avaliação escolar, ainda hoje, é tema de muita polêmica. Contudo, o processo de avaliação tem como base a análise quantitativa para mensurar o nível de aprendizagem do aluno, e decidir se o mesmo está ou não apto a cursar as séries seguintes. Diante desse contexto, o presente trabalho analisa o histórico de notas dos alunos do 1º ano do ensino médio, do IFCE *Campus* Crato, do ano de 2006 a 2010, através do método de regressão linear múltipla, a fim de descobrir qual a influência das notas dos dois primeiros bimestres em relação à média final do aluno.

PALAVRAS-CHAVE: Modelagem de Desempenho Acadêmico. Regressão Linear Múltipla. Correlação Linear.

AN ESTIMATE OF ACADEMIC PERFORMANCE FROM PREDICTIVE STATISTICAL METHODS: A CASE STUDY WITH 1st YEAR HIGH SCHOOL STUDENTS FROM THE FEDERAL INSTITUTE OF CEARÁ, CRATO CAMPUS

ABSTRACT

The academic assessment process, even now, is a very controversial subject. However, it is based on quantitative analysis in order to measure the student's learning level, and decide whether he/she is able to attend the following school years or not. Facing this context, the present study examined the history of grades of students in the 1st year of high school, from the IFCE Crato Campus, between 2006 and 2010, by using the method of multiple linear regression analysis, in order to find out the influence the first and second quarters grades have in relation to the student's final average.

KEYWORDS: *Academic Performance Modeling, Multiple Linear Regression, Linear Correlation.*

¹Mestre pelo Programa de Mestrado Acadêmico em Ciência da Computação da Universidade Estadual do Ceará. Professor do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE, *Campus* Crato. E-mail: robsonfeitosa@ifce.edu.br

²Doutor em Ciência da Computação pela Universidade Federal de Pernambuco. Professor de Ensino Médio, Técnico e Tecnológico do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - *Campus* Crato. E-mail: guilhermealvaro@ifce.edu.br

³Mestre em Biometria e Estatística Aplicada pela UFRPE. Professor Efetivo de Estatística do IFCE - *Campus* Crato. E-mail: ciceroCarlos@ifce.edu.br.

1 INTRODUÇÃO

A missão de qualquer instituição de ensino é contribuir com excelência na formação dos seus alunos. Contudo, muitos percalços surgem durante o processo de ensino-aprendizagem. Visando melhorar tal processo, diversos profissionais da área educacional têm demandado esforço e tempo na análise de fatores que dificultam à aprendizagem dos alunos, e, com isso, propõem melhorias a mesma.

A avaliação do desempenho acadêmico é tema de muita polêmica. Várias são as iniciativas para melhorar o processo de avaliação. Exemplos clássicos desse tipo de iniciativa são encontrados em Cronbach e Webb (1975), Burstein, Linn e Capell (1978) e Longford (1985). Entretanto, a avaliação quantitativa, por disciplina, ainda é usada como parâmetro base para o cálculo do rendimento escolar do aluno nas mais diversas instituições do Brasil. Para os alunos do ensino médio, a avaliação ocorre a cada bimestre, possibilitando ao aluno recuperar, ou manter seu desempenho no decorrer de um ano. Após esse prazo é realizada uma média entre as avaliações bimestrais; e, caso esse valor esteja dentro de um limite mínimo, o aluno é considerado apto a cursar a série seguinte; caso contrário, o mesmo é retido na mesma série.

Sabe-se que são muitos os fatores que corroboram à retenção do aluno. Pedagogos enumeram uma grande quantidade de fatores qualitativos, como: déficit de atenção, hiperatividade, problemas domésticos, *bullying*, dentre outros. Também são levados em consideração fatores quantitativos, como: quantidade de disciplinas cursadas, horas de estudo após as aulas, notas das avaliações. Tanto as características quantitativas, quanto as qualitativas são muito relevantes para a análise do desempenho do aluno, Lenning (1982) agrupa alguns destes fatores. Mas, avaliar os aspectos qualitativos é uma tarefa que demanda muito tempo, recursos humanos e financeiros, e o resultado pode chegar tarde demais, ou seja, depois que o aluno ficou retido.

Sabendo da importância da avaliação quantitativa (das notas das avaliações por disciplina) para a análise do rendimento escolar, o presente trabalho visa utilizar regressão linear múltipla para estabelecer um comportamento no desempenho final a partir das notas das avaliações dos dois primeiros bimestres.

Com isso, busca-se avaliar a probabilidade do aluno não atingir uma nota satisfatória, com tempo hábil o suficiente para a tomada de ações preventivas, visando o apoio à recuperação do aluno. Por exemplo: o encaminhamento ao reforço do conteúdo com os monitores, ou mesmo, aulas de recuperação paralela com os professores das disciplinas, grupos de estudo e leitura, dentre outros. Como resultado das medidas citadas, espera-se a diminuição do percentual de alunos retidos.

Dessa forma, o presente trabalho está organizado do seguinte modo: na seção de Materiais e Métodos descreve-se como foi feita a coleta dos dados, bem como os métodos

estatísticos utilizados. Na seção seguinte, são descritos os procedimentos de formulação dos modelos lineares múltiplos para as disciplinas do estudo de caso. Posteriormente, apresentam-se as discussões e análises resultados obtidos com a abordagem. Finalmente, em seguida, apresentam-se conclusões e trabalhos futuros.

2 MATERIAL E MÉTODOS

2.1 Especificação formal do modelo

Os dados utilizados no presente trabalho foram coletados na Secretaria de Educação Profissional e Tecnológica do IFCE, *Campus* Crato. Para o intervalo do ano de 2006 a 2010, de todos os alunos matriculados no 1º ano do Ensino Médio, foram selecionados todos os registros (exceto os registros de alunos transferidos e desistentes) de notas nos 4 (quatro) bimestres referentes às disciplinas de Português, Matemática e Biologia, bem como o resultado final, média resultante do cálculo abaixo:

$$md = \frac{(n1 + n2 + n3 + (2.n4))}{5} \quad (1)$$

Regressão linear múltipla é um método utilizado para modelar um relacionamento linear entre uma variável dependente e duas ou mais variáveis independentes. Assim, como nos modelos de regressão linear simples, para se obter um modelo bem ajustado aos dados do problema, devem ser consideradas as suposições de normalidade, homocedasticidade e independência dos erros.

Neste trabalho, o método utilizado para modelagem incluiu três etapas: formulação e ajustes dos modelos, bem como inferência. A etapa de formulação consiste em escolher as variáveis do modelo: variáveis independentes (também chamadas de sistemáticas ou preditoras) e variável dependente (ou resposta). O ajuste busca estimar os parâmetros lineares dos modelos e determinar as funções das estimativas desses parâmetros. Um método bastante utilizado, e que também é utilizado neste trabalho, é o método de mínimos quadrados, o qual conduz a um procedimento de estimação bastante simples.

A última etapa, inferência, verifica a adequação do modelo e realiza a análise de discrepâncias locais, que quando significantes, podem acarretar na escolha de outro modelo, ou em aceitar a existência de dados discrepantes. Para a etapa de inferência, neste trabalho utilizou-se análise gráfica residual, por consistir de um método mais flexível, em relação aos testes de hipóteses, e permitir um tipo de análise mais subjetiva. Utilizaram-se os gráficos: (1) qq-plot e (2) histograma, para verificar a suposição de normalidade; (3) de dispersão entre resíduos e valores

ajustados, para avaliar a suposição de homocedasticidade; e (4) de ordem de coleta dos resíduos, para avaliar a independência entre erros.

Para efeito de simplificação foram escolhidas 3(três) disciplinas para a geração dos modelos, Português, Matemática e Biologia. Para cada disciplina foi gerado um modelo de regressão múltipla resultante da análise nos dados dos registros, assumindo apenas as notas dos dois primeiros bimestres, para inferir a média final. De posse desse modelo é possível estimar a média final do aluno, um semestre antes da retenção. As ferramentas utilizadas no apoio à análise, testes e geração dos gráficos estatísticos foram: R (versão 2.12) (Gentleman e Ihaka, 2003) e BrOffice.org Calc (versão 3.3.1) (BrOffice.org, 2013).

3 PROCEDIMENTO DE FOMULAÇÃO DOS MODELOS LINEARES MÚLTIPLOS PARA AS DISCIPLINAS DO ESTUDO DE CASO

Esta seção apresenta os resultados obtidos a partir da aplicação do método apresentado na seção anterior ao estudo de caso. A subseção a seguir apresenta os resultados das etapas de formulação e ajuste para os modelos selecionados para as disciplinas de Português, Matemática e Biologia.

3.1 Etapas de formulação e ajuste dos modelos para as disciplinas do estudo de caso

Para cada disciplina foi gerado um modelo de regressão múltipla resultante da análise nos dados dos registros, assumindo apenas as notas dos dois primeiros bimestres (n_1 e n_2), como variáveis independentes, para inferir a média final (md), modelada como variável resposta. De posse desse modelo, busca-se estimar a média final do aluno, dois semestres antes da retenção.

Para verificar a correlação linear entre as variáveis, foi calculado o coeficiente de correlação de Pearson (Tabela 1). Os resultados mostraram que existe correlação moderada positiva para as disciplinas de Português e Matemática e fraca positiva para Biologia. Para os dados da disciplina de Português, tem-se:

Figura 1 – Cálculo dos coeficientes e intercepto do modelo de regressão múltipla para a disciplina de Português.

```
Residuals:
      Min       1Q   Median       3Q      Max
-1.50985 -0.32816 -0.02505  0.29995  1.63839

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.32135     0.12031   27.61  <2e-16 ***
n1             0.24021     0.01745   13.76  <2e-16 ***
n2             0.28392     0.01611   17.63  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4974 on 611 degrees of freedom
Multiple R-squared:  0.5622,    Adjusted R-squared:  0.5608
F-statistic: 392.3 on 2 and 611 DF,  p-value: < 2.2e-16
```

Fonte: Produzida pelo autor

Como resultado do modelo observa-se o modelo de regressão múltipla 2, ilustrada a seguir:

$$md = 3.321 + 0,24.n1 + 0,284.n2 \quad (2)$$

Para os dados da Disciplina de Matemática, tem-se:

Figura 2 – Cálculo dos coeficientes e intercepto do modelo de regressão múltipla para a disciplina de Matemática.

```
Residuals:
      Min       1Q   Median       3Q      Max
-2.84093 -0.42214 -0.02536  0.50475  1.89808

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.24440     0.13647   16.45  <2e-16 ***
n1             0.28530     0.01877   15.20  <2e-16 ***
n2             0.38078     0.02071   18.38  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7225 on 603 degrees of freedom
Multiple R-squared:  0.6457,    Adjusted R-squared:  0.6445
F-statistic: 549.5 on 2 and 603 DF,  p-value: < 2.2e-16
```

Fonte: Produzida pelos autores

Como resultado do modelo observa-se o modelo de regressão múltipla 3, ilustrada a seguir:

$$md = 2.224 + 0,285.n1 + 0,381.n2 \quad (3)$$

Para os dados da Disciplina de Biologia, tem-se:

Figura 3 – Cálculo dos coeficientes e intercepto do modelo de regressão múltipla para a disciplina de Biologia.

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.47692 -0.42367  0.00134  0.45080  1.85080

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.61092    0.15097   23.92  <2e-16 ***
n1             0.24945    0.01806   13.81  <2e-16 ***
n2             0.21331    0.01582   13.48  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7241 on 648 degrees of freedom
Multiple R-squared:  0.4388,    Adjusted R-squared:  0.4371
F-statistic: 253.4 on 2 and 648 DF,  p-value: < 2.2e-16
```

Fonte: Produzida pelos autores.

Como resultado do modelo observa-se o modelo de regressão múltipla 4, ilustrada a seguir:

$$md = 3.611 + 0,249.n1 + 0,213.n2 \quad (4)$$

Tabela 1 – Resumo da Relação linear entre as variáveis dos modelos para as disciplinas de Português, Matemática e Biologia.

Disciplina	Coefficiente de correlação de Pearson
Português	56,08%
Matemática	64,45%
Biologia	43,71%

Ou seja, utilizando a técnica de mínimos quadrados foi possível estabelecer os parâmetros dos três modelos, como pode ser visto na Tabela 2.

Tabela 2 – Resumo dos Modelos lineares múltiplos para as disciplinas de Português, Matemática e Biologia.

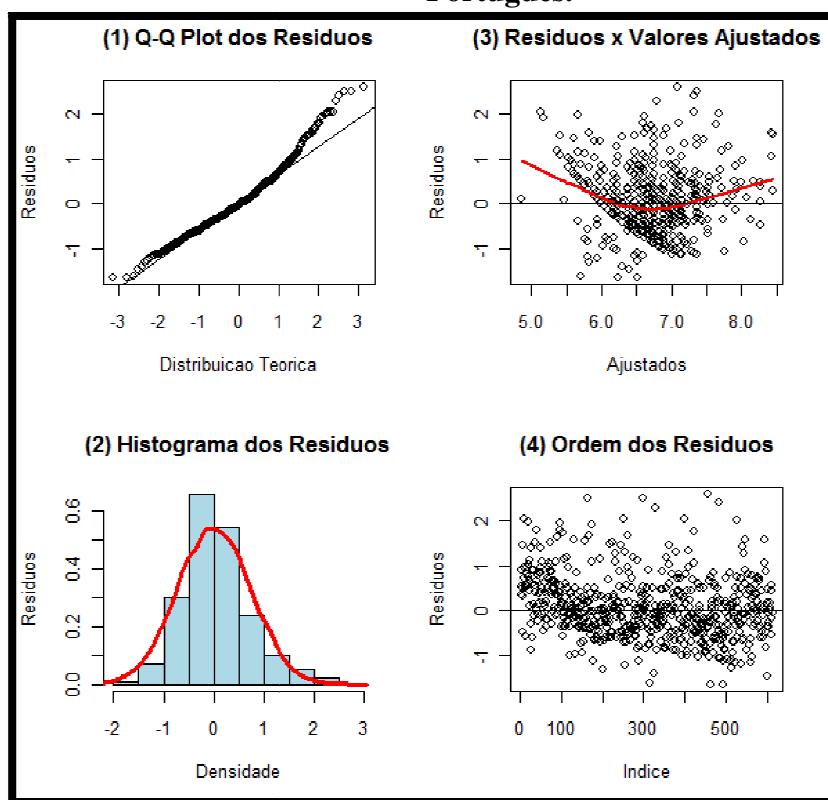
Disciplina	Modelo Linear Múltiplo
Português	$md = 3,321 + 0,24 \cdot n_1 + 0,284 \cdot n_2$
Matemática	$md = 2,224 + 0,285 \cdot n_1 + 0,381 \cdot n_2$
Biologia	$md = 3,611 + 0,249 \cdot n_1 + 0,213 \cdot n_2$

Uma vez selecionados os modelos para os dados das três disciplinas, a etapa seguinte, que será descrita na próxima subseção, consiste da análise residual.

3.2 Análise residual

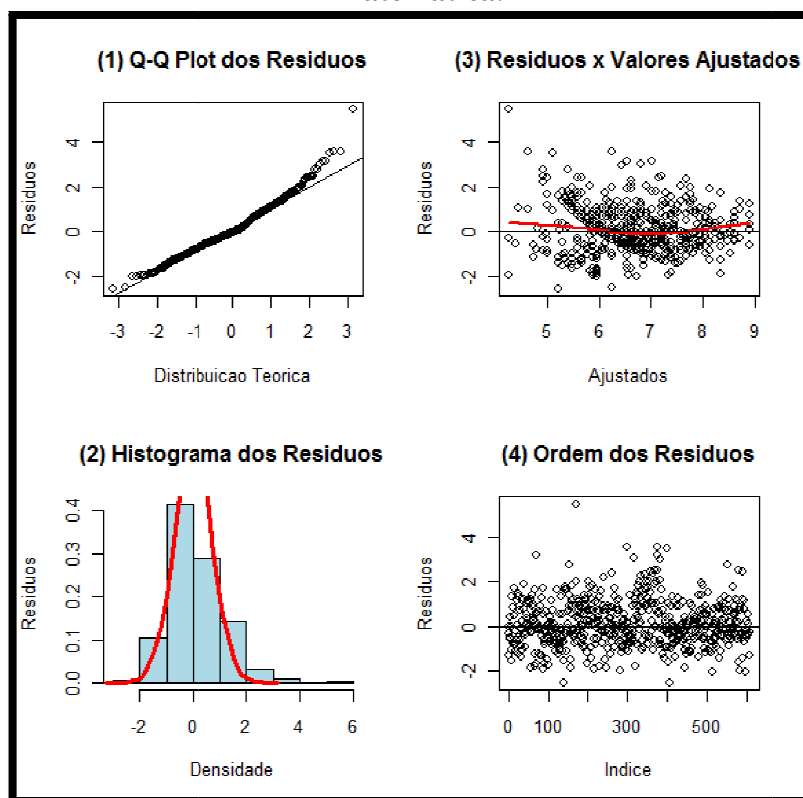
Para verificar as suposições de normalidade, homocedasticidade e independência sobre os erros dos modelos selecionados, bem como o grau de dispersão dos resíduos foram gerados os gráficos 1,2,3 e 4 apresentados nas Figuras 4, 5 e 6.

Figura 4 – Gráficos para análise residual para o modelo encontrado para a disciplina de Português.



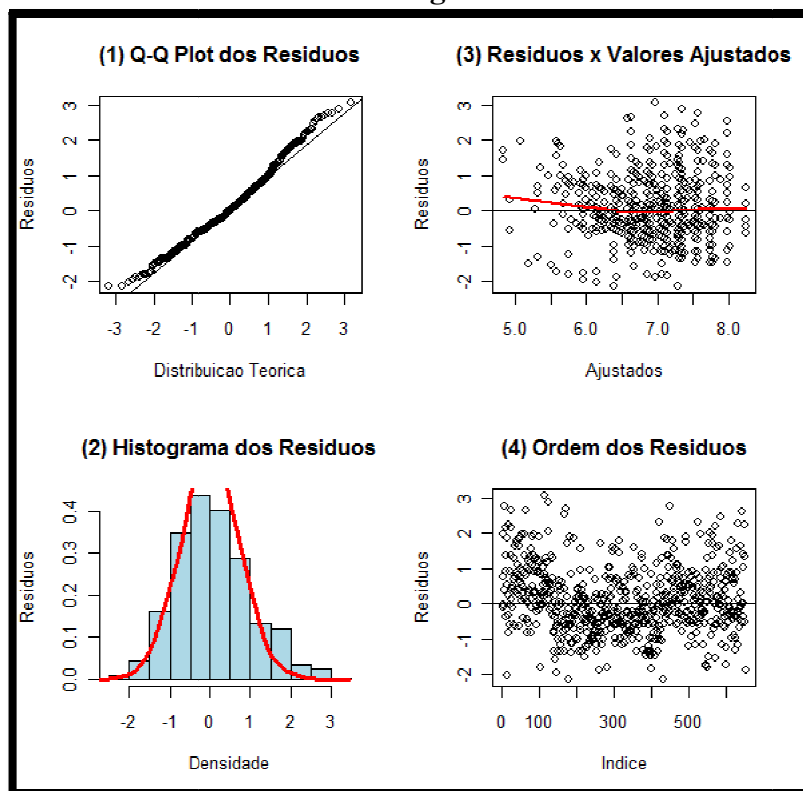
Fonte: Produzida pelos autores

Figura 5 – Gráficos para análise residual para o modelo encontrado para a disciplina de Matemática.



Fonte: Produzida pelos autores

Figura 6 – Gráficos para análise residual para o modelo encontrado para a disciplina de Biologia.



Fonte: Produzida pelos autores

Analizando os gráficos qq-plot (1) e histogramas (2), contidos nas Figuras 4, 5 e 6, percebe-se que os três modelos selecionados (Tabela 2) apresentaram erros normalizados.

A análise dos gráficos de dispersão, apresentados nas Figuras 4, 5 e 6, mostra que ocorreu heterocedasticidade, devido ao padrão de dispersão variável entre os pontos nos gráficos. Porém, apesar deste comportamento, essa característica pode ser relaxada, uma vez que a dispersão é bastante sensível.

Por fim, verifica-se que não existe relação de dependência entre os erros, como pode ser visto nos gráficos das Figuras 4, 5 e 6, dada as distribuições aleatórias dos pontos nos respectivos gráficos.

4 DISCUSSÃO DOS RESULTADOS

A análise das influências das notas dos dois primeiros bimestres no resultado final, expressadas através dos coeficientes de correlação de Pearson, expostos na Tabela 1, mostrou que as referidas notas são mais significativas e menos significativas, no desempenho final para as disciplinas de Matemática e Biologia, respectivamente.

Ainda desta análise, pode-se concluir que os alunos que obtiveram bons índices nas duas primeiras avaliações tendem a ter bons desempenhos finais nas disciplinas de Matemática e Português. Já para a disciplina de Biologia, a influência das duas notas é semelhante à média aritmética de todas as notas. Em outras palavras, o peso das notas dos dois primeiros bimestres, para Biologia, seria igual ao de uma média aritmética.

Com os modelos lineares múltiplos selecionados, pode-se estabelecer um comportamento geral para o desempenho acadêmico nas três disciplinas. Utilizando esses modelos e considerando as notas dos dois primeiros bimestres para um aluno, é possível ter uma estimativa do seu desempenho final. Por exemplo, um aluno que obteve as notas $n_1 = 5$ e $n_2 = 6$ na disciplina de Matemática, considerando o modelo linear múltiplo da Tabela 2 seu desempenho final tende a ser 5,9 aproximadamente. Neste caso, como as demais notas (n_3 e n_4) possuem menor influência no resultado final, o aluno deverá ter um desempenho acima de média no restante da disciplina para não ser retido.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou um estudo sobre as influências dos resultados das Avaliações dos dois primeiros bimestres no resultado acadêmico final para três disciplinas (Português, Matemática e Biologia) do ensino médio do Instituto Federal do Ceará - *Campus Crato*.

O estudo consistiu na seleção de modelos lineares múltiplos para caracterizar essas influências. Os modelos foram formulados mediante o histórico acadêmico de todos os alunos que cursaram essas disciplinas entre os anos de 2006 e 2010.

A análise dos resultados mostrou que as notas dos dois primeiros bimestres possuem maior influência na média final para as disciplinas de Português e Matemática. Além disso, a partir dos modelos lineares múltiplos selecionados, pode-se estimar o desempenho final.

Como trabalhos futuros, objetiva-se realizar o mesmo estudo utilizando o modelo de inferência preditivo bayesiano (*Madureira et al, 2009*). Com isso, os modelos selecionados serão uma ferramenta de apoio estratégico ao setor pedagógico na busca por ações que minimizem o percentual de retenção.

REFERÊNCIAS

- BROffice 2013. Disponível em: <<http://www.broffice.org/>>. Acesso em: 05 mar. 2013.
- BURSTEIN, L.; LINN, R. L.; CAPELL, F. J. Analysing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics*, v. 3, n. 4, p. 347-383, 1978.
- CRONBACH, L. J.; WEBB, N. Between class and within class effects in a reported aptitude X treatment interaction: a reanalysis of a study by G. L. Anderson. *Journal of Educational Psycology*, v.67, n.6, p. 717-724, 1975.
- GENTLEMAN, R; IHAKA R. *The R project for statistical computing*. 2003. Disponível em: <<http://www.r-project.org/>>. Acesso em: 05 mar. 2013.
- LENNING, O. T. Variable-selection and measurement concerns. In: E. Pascarella. (Ed.) *Studying Student Attrition. New Directions for Institutional Research*, 36. San Francisco: Jossey-Bass, 1982.
- LONGFORD, N. T. Mixed linear models and application to schools effectiveness. *Computational statistics quaterly*, n.2, p. 109-117, 1985.
- MADUREIRA, A.P. et al. Inferência baysiana na predição de valores genéticos do peso aos 365 dias de bovinos de corte. *Archivos de Zootecnia*, Espanha, v. 58, n. 222, p. 265-275, 2009.

Trabalho recebido em 11/10/2012.

Trabalho aceito em 05/03/2013.