

Classificação dos Exames Nacionais 2011/2012 de Matemática A

Abordagem Bayesiana

Pedro Martins
Ricardo Cotrim

Lisboa, 8 Julho 2015

- 1 Motivação e Objetivos
- 2 Limitações
- 3 Impacto dos Exames Nacionais
- 4 Fonte dos dados
- 5 Procedimentos
- 6 Caracterização da amostra
- 7 Abordagem Bayesiana vs. Clássica
- 8 Modelação
- 9 Resultados
- 10 Comentários e trabalhos futuros

Motivação e Objectivos

Motivação

1. Familiarização com uma nova metodologia e software Bayesianos;
2. Criação de um modelo, ainda em fase preliminar, que funcione como suporte de um estudo aprofundado dos exames de 12º ano

Objetivos

1. Elaborar um modelo que identifique as variáveis que contribuem significativamente para explicar a variável de interesse: “classificação dos exames de matemática A”.
2. Comparar os resultados das abordagens Clássica e Bayesiana

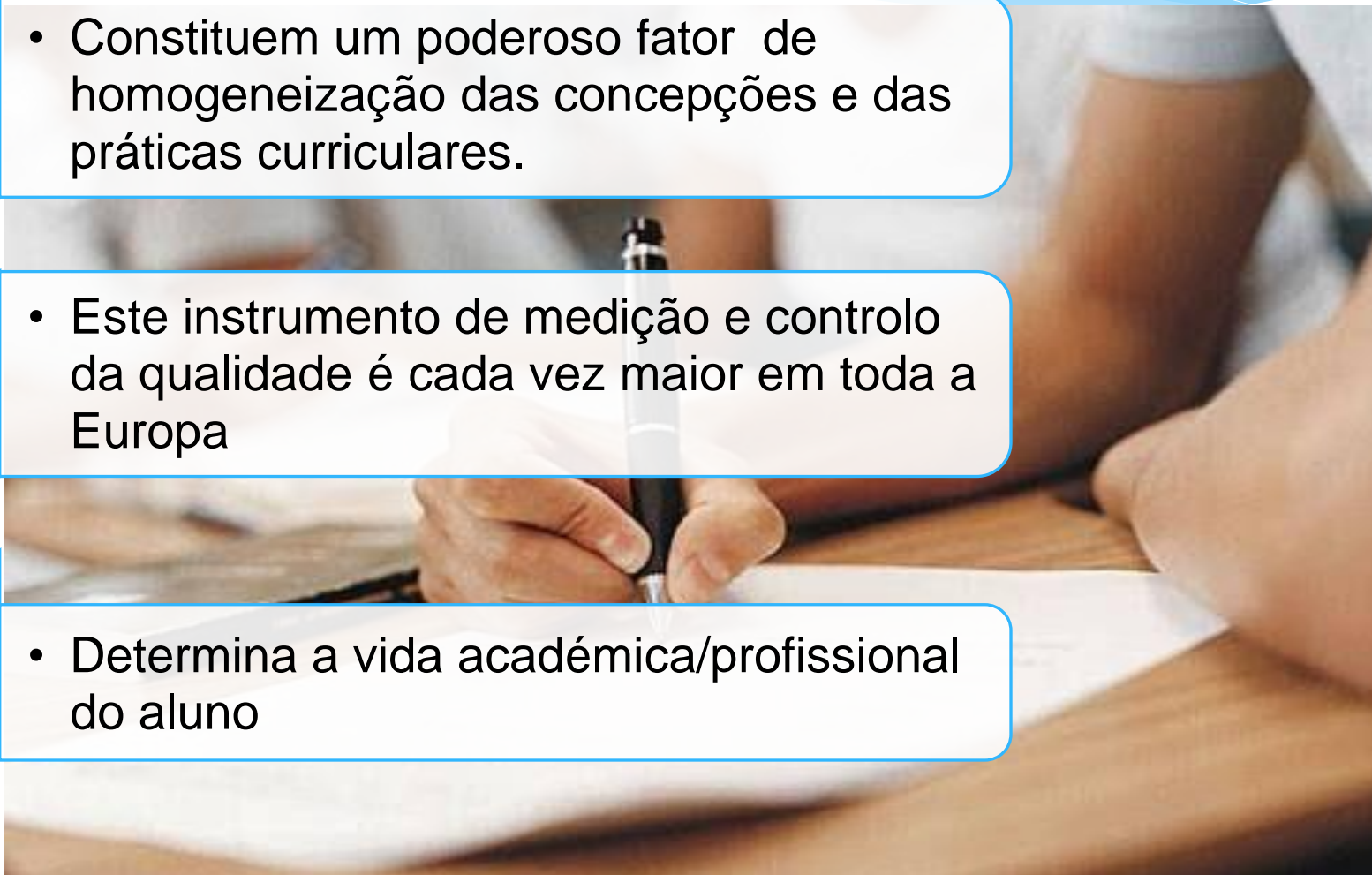
Limitações

- ❖ Este estudo resulta de um trabalho para a unidade curricular de estatística bayesiana e foi o nosso primeiro contacto com a metodologia e software bayesianos (OpenBugs, livre utilização);
- ❖ A dimensão da base de dados (MISI) é grande (mais de 22.000.000 registos), dificultando a extração de dados e a verificação de incongruências;
- ❖ Algumas variáveis foram agregadas de modo a simplificar a construção do modelo;
- ❖ Os dados analisados foram obtidos através de uma amostragem aleatória simples;
- ❖ Criação de uma variável proxy da dimensão da turma baseada na nota de 3º período



Impacto dos Exames Nacionais

- Constituem um poderoso fator de homogeneização das concepções e das práticas curriculares.
- Este instrumento de medição e controlo da qualidade é cada vez maior em toda a Europa
- Determina a vida académica/profissional do aluno



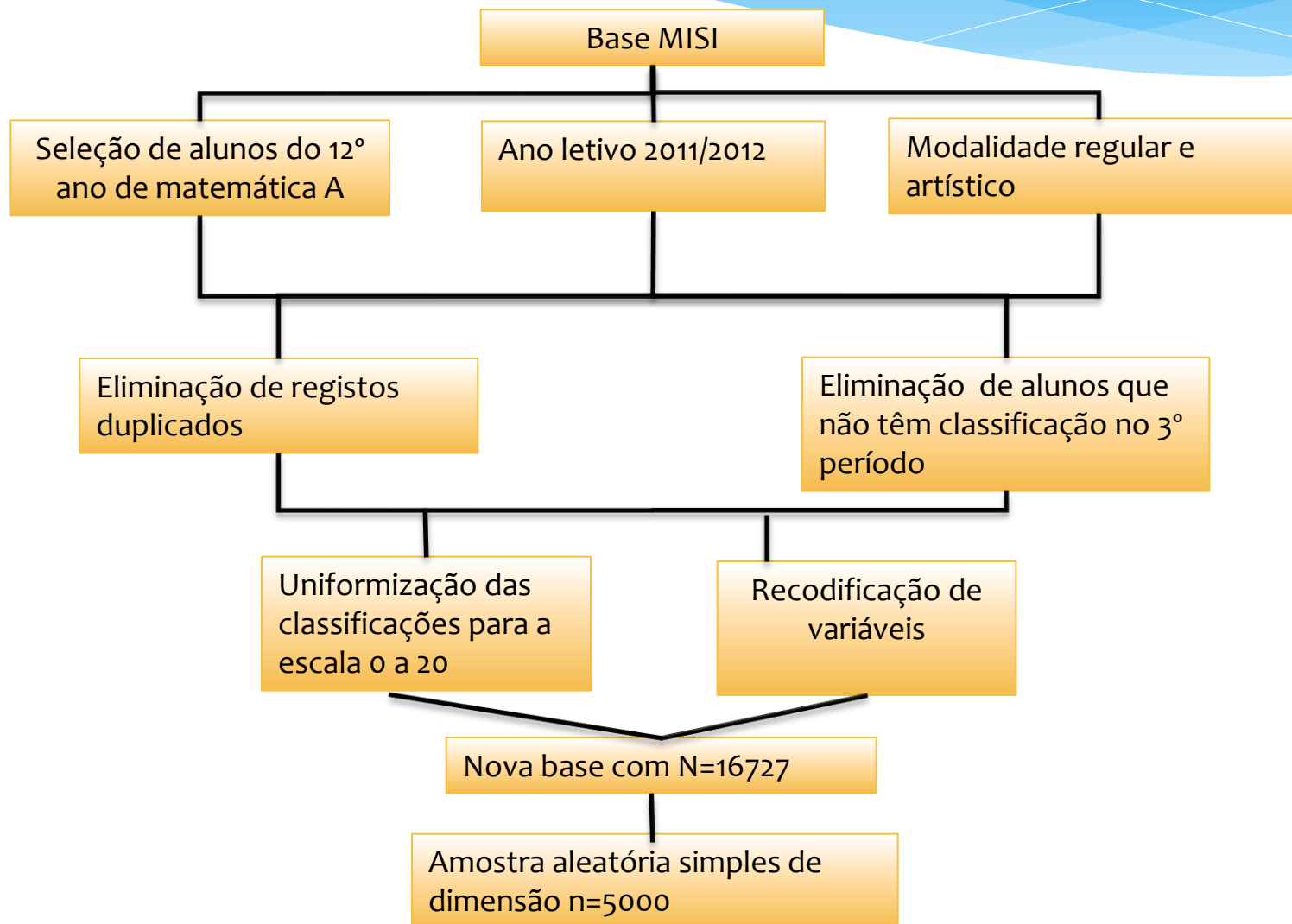
Fonte de dados: MISI

O sistema MISI (Equipa de Missão para o Sistema de Informação) tem como objetivo:

- ✓ Centralizar a recolha de informação da educação pré-escolar e do ensino básico e secundário;
- ✓ Facultar aos organismos centrais do MEC a informação necessária para a prossecução das suas atribuições;
- ✓ Servir como base de suporte à produção de estatísticas da educação e à tomada de decisão.



Procedimentos



Caracterização da amostra

Variáveis explicativas:

- Sexo
- NUTSII
- Número de matrículas
- Encarregado de educação (E.E.)
- Formação académica do E.E.
- Beneficiário da ASE
- Curso
- Internet em casa
- Dimensão da turma
- Idade do aluno

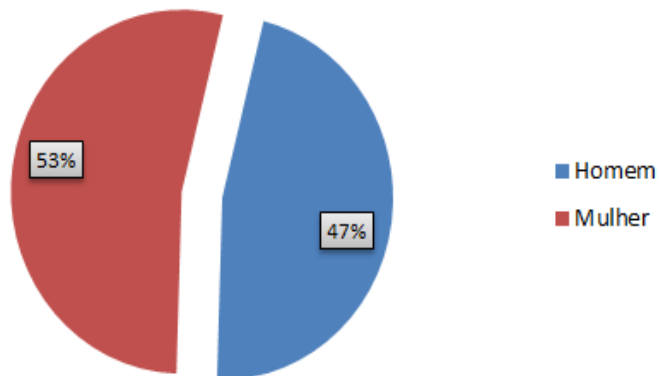
Variável de interesse:

Classificação de exame

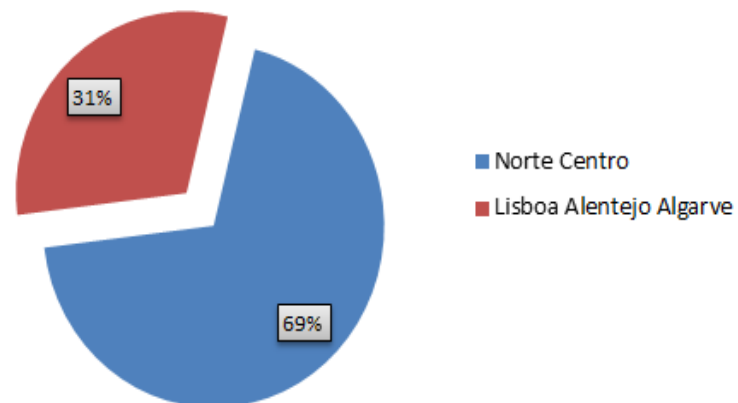
	n	%
Total	5000	100%
Sexo		
Homem	2330	47%
Mulher	2670	53%
NUTS II		
Norte Centro	3459	69%
Lisboa Alentejo Algarve	1541	31%
Número de matrículas		
Uma	4553	91%
Duas ou mais (repetente)	447	9%
Encarregado de educação		
Pai	951	19%
Mãe	3.525	71%
O próprio	430	9%
Outro	94	2%
Formação académica do encarregado de educação		
Sem habilitações	72	1%
Básico	2.395	48%
Secundário	1.198	24%
Graduado	1.335	27%
Beneficiário da ASE (Apoio escolar: alimentar, material e transportes)		
Não beneficia (famílias mais ricas)	4058	81%
Escalão B	603	12%
Escalão A (famílias mais pobres)	339	7%
Curso		
Ciências	4447	89%
Economia	478	10%
Humanidades e Artes	75	2%
Internet em casa		
Não	898	18%
Sim	4.102	82%
	Média	Desvio padrão
Dimensão da Turma	23	4,60
Classificação do exame	11,05	4,42
Idade	17,30	0,67

Caracterização da amostra

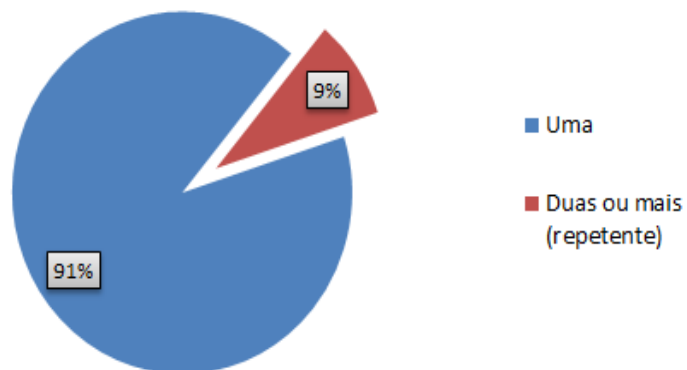
Alunos por sexo



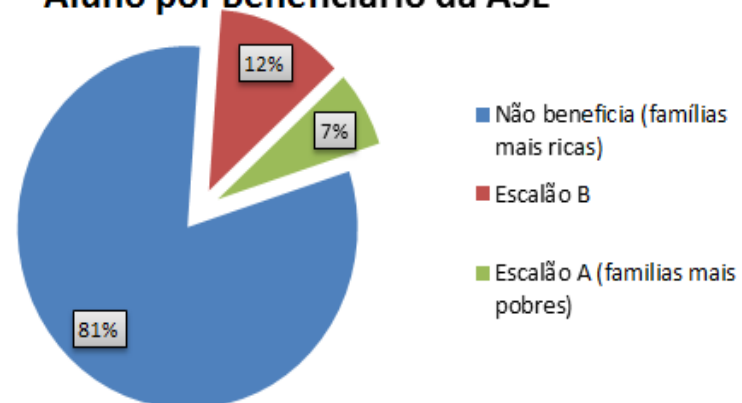
Alunos por NUTSII



Alunos por número de matrículas

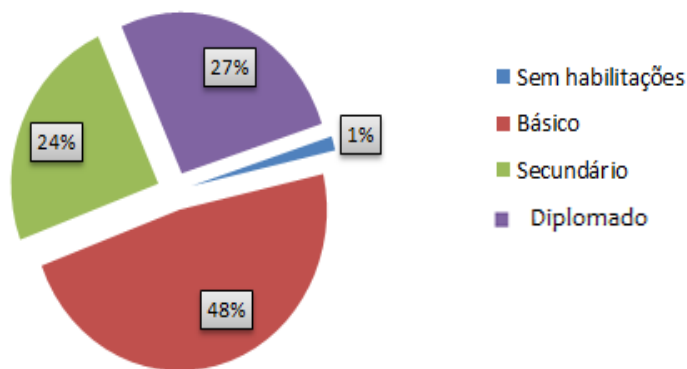


Aluno por Beneficiário da ASE

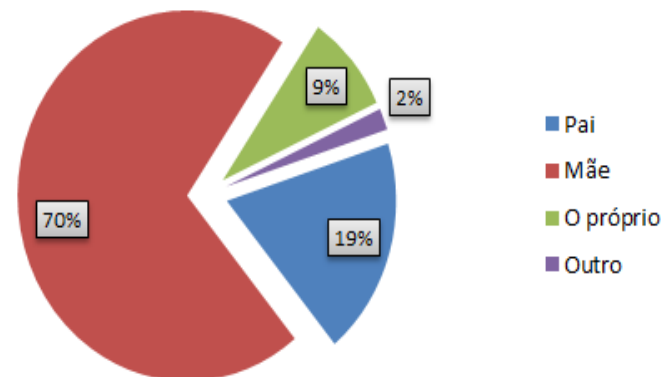


Caracterização da amostra

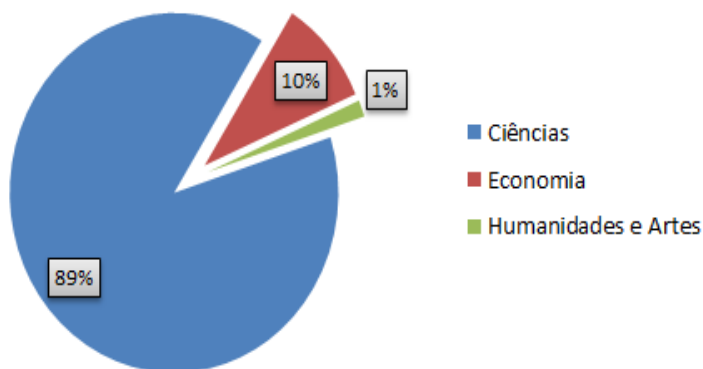
**Encarregado de educação por
Formação académica**



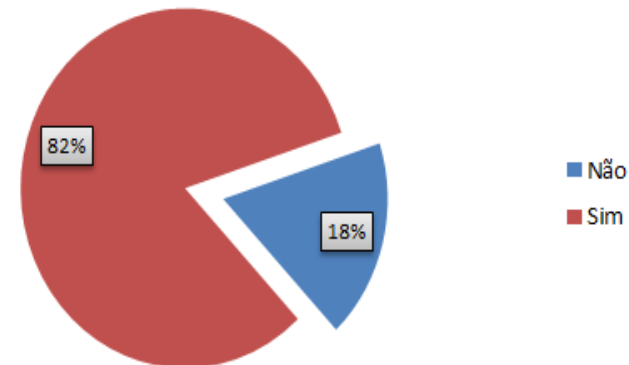
Encarregado de educação



Alunos por curso

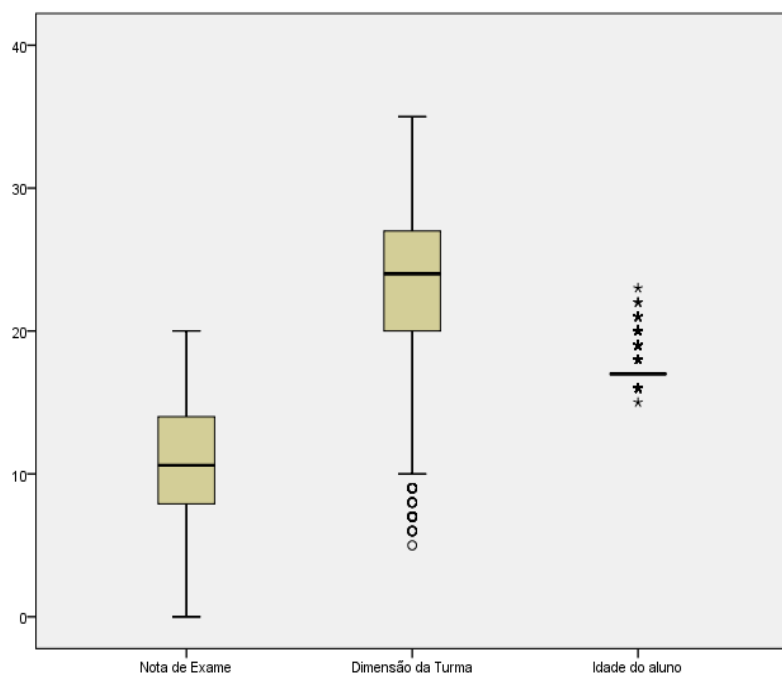


Alunos com internet em casa

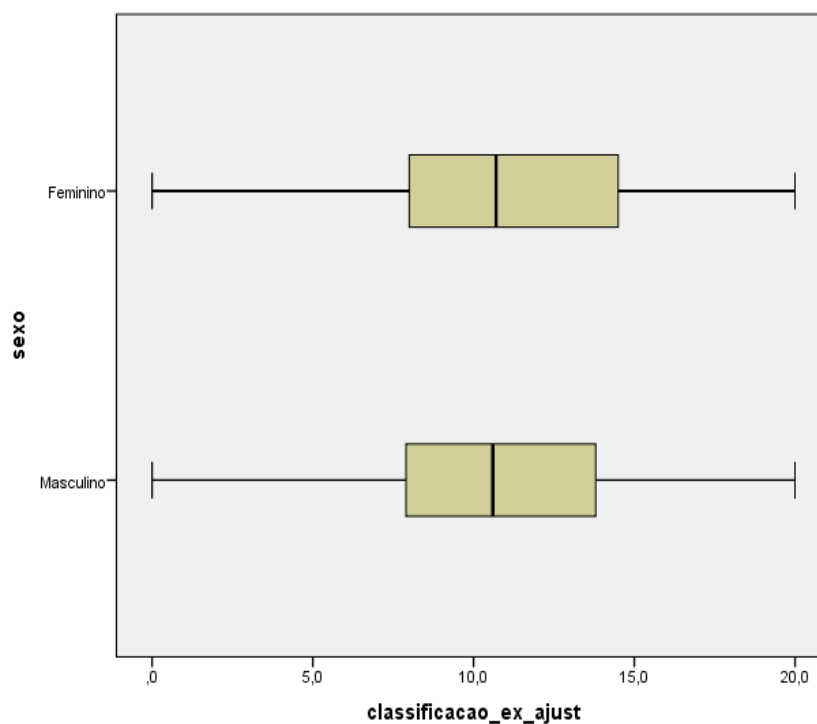


Caracterização da amostra - Boxplot

Classificação de exame, dimensão das turmas e idade dos alunos

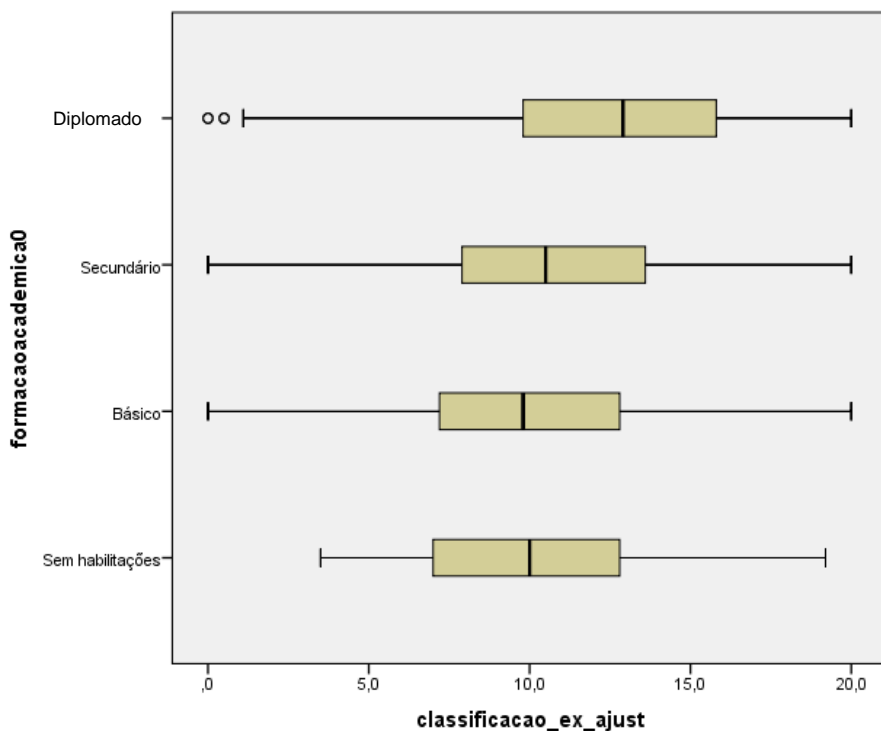


Classificação de exame por sexo

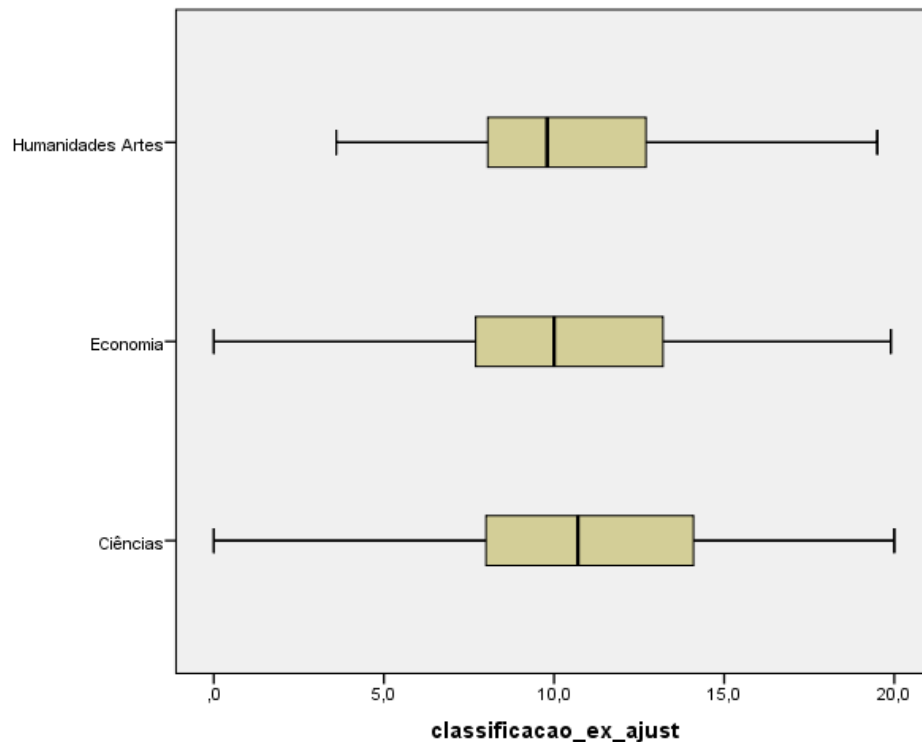


Caracterização da amostra - Boxplot

Classificação de exame por formação académica do encarregado de educação

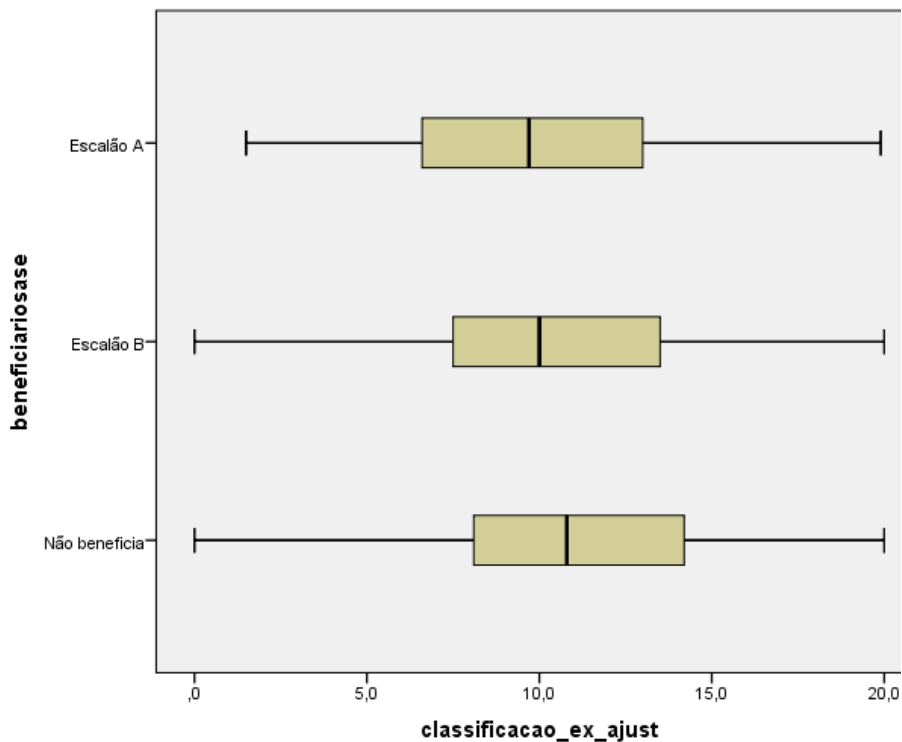


Classificação de exame por curso

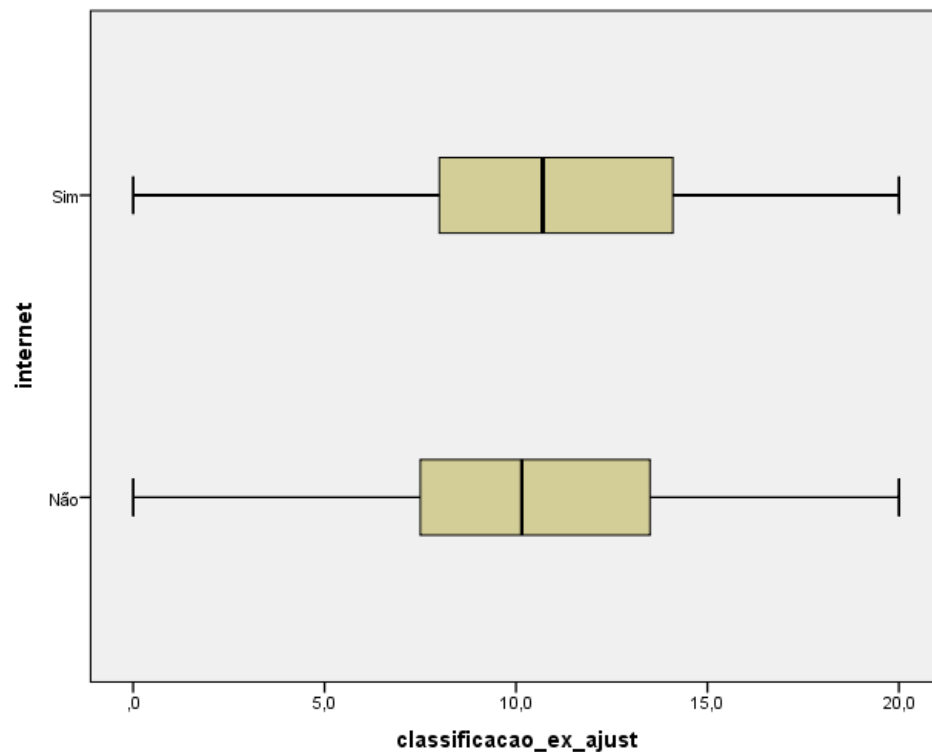


Caracterização da amostra - Boxplot

Classificação de exame
por beneficiários ASE



Classificação de exame
por Internet em casa



Thomas Bayes



Esta nova abordagem apenas teve o *boom* em 1984!



Teorema de Bayes

Teorema de Bayes:

Seja X uma v.a. com f.d.p $f(\cdot | \theta)$, a qual pertencente à família de modelos

$$\mathcal{F} = \{f(x | \theta) : \theta \in \Theta\}, x \in X$$

sendo Θ o espaço-parâmetro.

O teorema de Bayes no caso contínuo

Suponhamos que se observa $X = x$. O teorema de Bayes é dado por:

$$p(\theta | x) = \frac{f(x | \theta)p(\theta)}{\int_{\Theta} f(x | \theta)p(\theta)d\theta}, \theta \in \Theta,$$

É habitual apresentar-se o teorema de Bayes num formato não normalizado expresso por:

$$p(\theta | \mathbf{x}) \propto L(\theta | \mathbf{x})p(\theta)$$

sendo $L(\theta | \mathbf{x})$ a verosimilhança de θ dada por:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n f(x_i | \theta).$$

Abordagem Clássica Vs Bayesiana

Seja (X_1, X_2, \dots, X_n) uma amostra aleatória proveniente de uma população com f.m.p. ou f.d.p. $f(x|\theta)$.
Tem-se:

Estatística Clássica	Estatística Bayesiana
Os dados x são aleatórios e θ é um vetor de parâmetros fixos quase sempre de valores desconhecidos.	Os dados x são fixos e θ é um vetor de parâmetros aleatórios cuja incerteza é formalizada através da distribuição à priori $p(\theta)$.
Geralmente estimamos o valor de θ por máxima verosimilhança.	Através do Teorema de Bayes, obtemos a distribuição à posteriori de θ (considerar a média à posteriori como estimativa pontual).
Faz-se inferência estatística por exemplo construindo intervalos de confiança ou fazendo testes de hipóteses.	Faz-se inferência através das Regiões de Credibilidade e fazem-se testes de hipótese utilizando a distribuição à posteriori.
Não utiliza a informação de experiências anteriores.	Permite utilizar informação resultante de experiências anteriores e pode ser usada de forma sequencial.

Modelo Bayesiano: distribuições à priori não informativas

- Consideramos que não dispomos de qualquer informação sobre os parâmetros de interesse.

“let the data speak for themselves”

Ronald Fisher or John Tukey

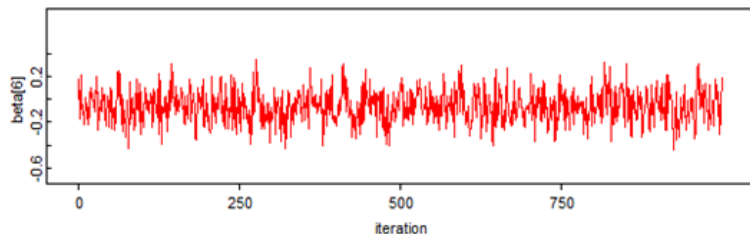
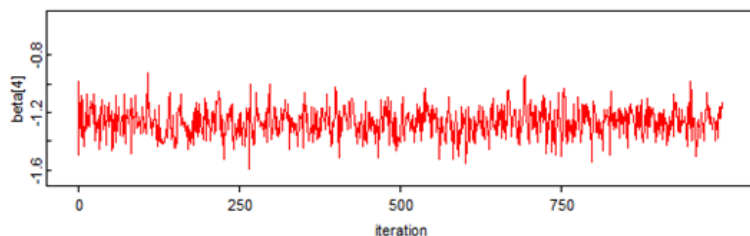
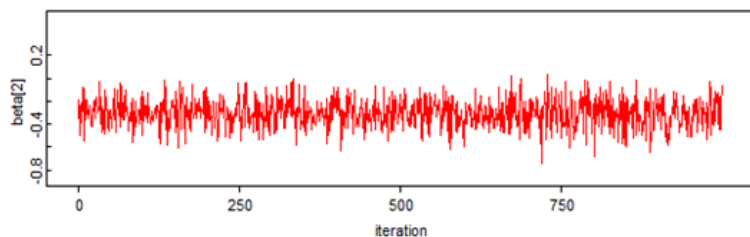
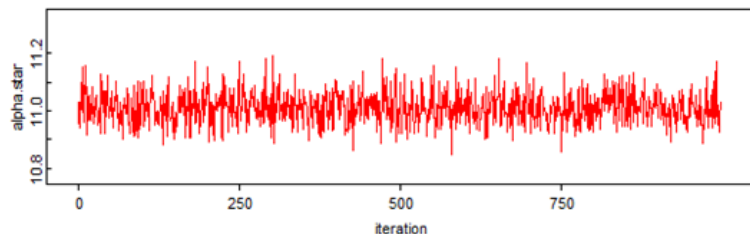
- Foram criadas variáveis dummy para cada uma das variáveis categóricas.
- Obtemos o modelo de regressão bayesiana (os coeficientes β são aleatórios):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_{16} x_{16} + \varepsilon$$

- Para fazer inferência Bayesiana foi utilizado o software OpenBugs.

Modelo Bayesiano: nº de iterações para os coeficientes de regressão

Update de 1.000 iterações:



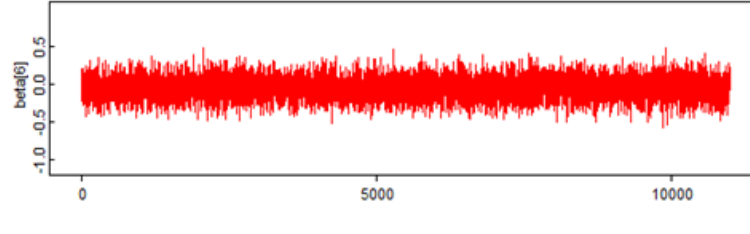
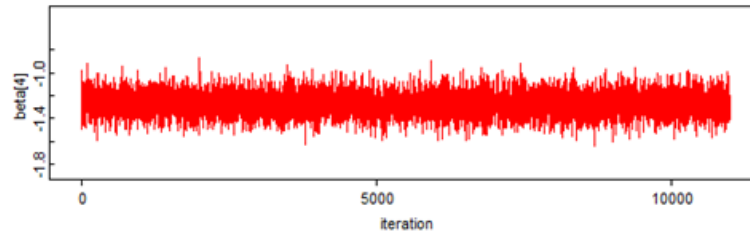
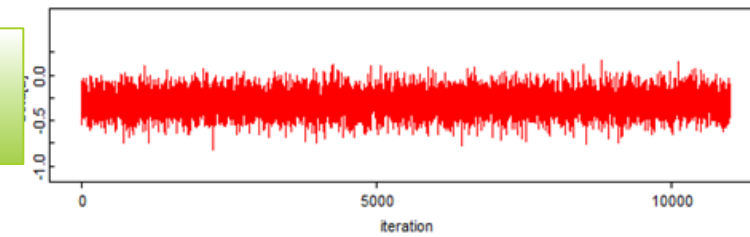
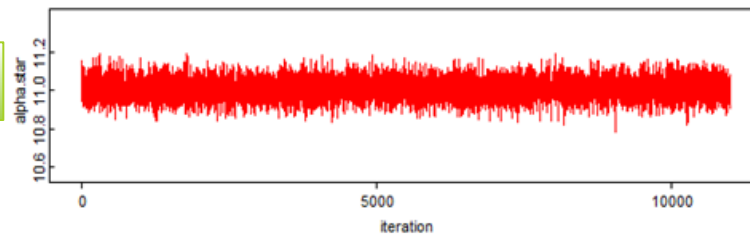
Alpha start

Beta 2
(NUTS: Lisboa/
Alentejo/Algarve)

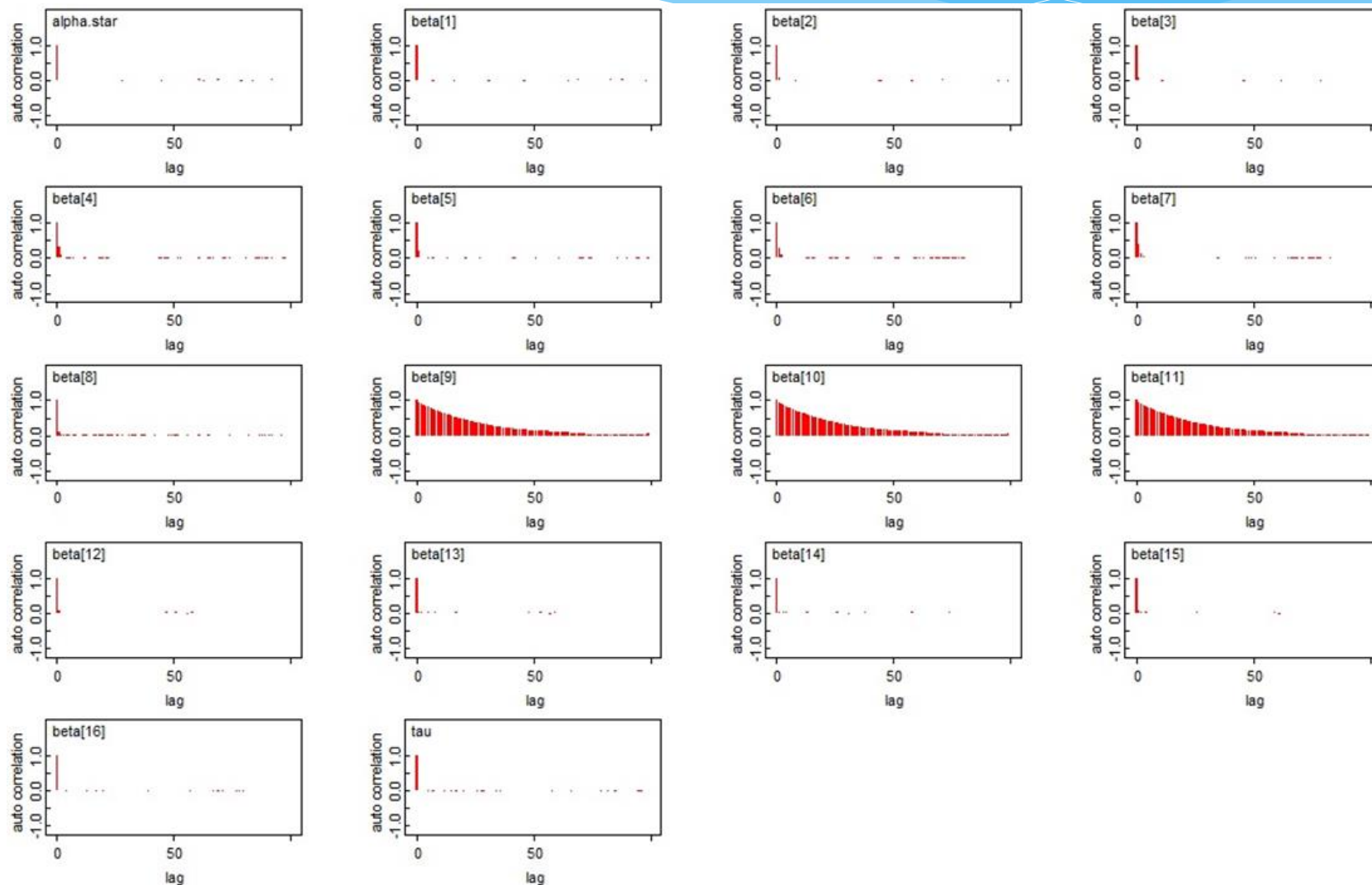
Beta 4
(Idade)

Beta 6
(Enc. Educação:
Mãe)

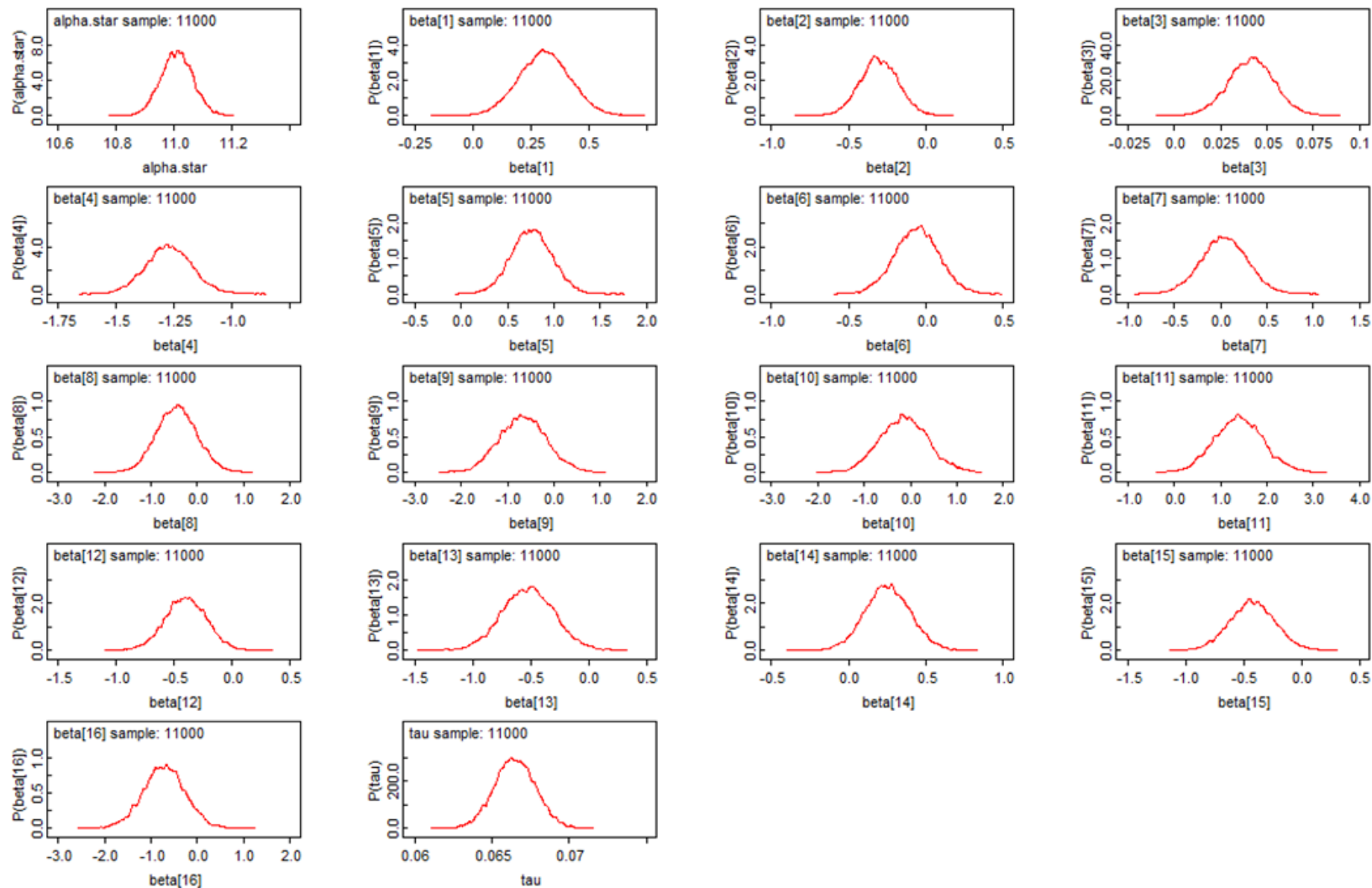
Update de mais 10.000 iterações:



Modelo Bayesiano: auto-correlações



Modelo Bayesiano: estimativas de kernel para as densidades à posteriori



Modelo Bayesiano: output

Variável/Constante	Média	Desvio Padrão	quantil de 2,5%	quantil de 95%	Significância
alpha.star	11,0100	0,0547	10,9000	11,1200	significativo
Sexo(1=Mulher)	0,3098	0,1109	0,0937	0,5288	significativo
NUTSII0_2 (1= LAA)	-0,3030	0,1235	-0,5447	-0,0613	significativo
dimturma	0,0416	0,0125	0,0169	0,0661	significativo
idade	-1,2730	0,0986	-1,4680	-1,0790	significativo
numeromatriculas	0,7597	0,2180	0,3339	1,1910	significativo
encarregadoedu0_2 (1=Mãe)	-0,0587	0,1420	-0,3411	0,2222	não significativo
encarregadoedu0_3(1=Próprio)	0,0383	0,2487	-0,4538	0,5207	não significativo
encarregadoedu0_4(1=Outro)	-0,4398	0,4264	-1,2510	0,4104	não significativo
formacaoacademica0_2(1=Básico)	-0,6639	0,4961	-1,6100	0,3432	não significativo
formacaoacademica0_3(1=Secundário)	-0,1381	0,5045	-1,1010	0,8891	não significativo
formacaoacademica0_4(1=Graduado)	1,3930	0,5087	0,4292	2,4130	significativo
beneficiariosase_2 (1=Escalão B)	-0,3995	0,1769	-0,7471	-0,0528	significativo
beneficiariosase_3 (1=Escalão A)	-0,5221	0,2255	-0,9669	-0,0779	significativo
Internet (1= tem internet em casa)	0,2525	0,1452	-0,0347	0,5399	não significativo
curso0_2 (1=Economia)	-0,4260	0,1932	-0,8037	-0,0453	significativo
curso0_3 (1=Artes e Humanidades)	-0,7147	0,4571	-1,6330	0,1763	não significativo
tau	0,0665	0,0013	0,0639	0,0691	significativo

Abordagem Clássica vs. Bayesiana

Estatística Clássica Estatística Bayesiana

Sexo

Homem		
Mulher	0.310*** (0.110)	0.310 Sig (0.111)

NUTS II

Norte Centro		
Lisboa Alentejo Algarve	-0.301** (0.123)	-0.303 Sig (-0.124)

Dimensão da Turma

0.042*** (0.012)	0.042 Sig (0.012)
---------------------	----------------------

Idade

-1.270*** (0.098)	-1.273 Sig (0.100)
----------------------	-----------------------

Número de matrículas

Uma		
Duas ou mais (repetente)	-0.765*** (0.216)	-0.760 Sig (0.218)

Encarregado de educação

Pai		
Mãe	-0.058 (0.143)	-0.059 (0.142)
O próprio	0.038 (0.249)	0.038 (0.249)
Outro	-0.438 (0.424)	-0.44 (0.426)

Observações	5000
R ²	0.1245

*** significância a 99%

** significância a 95%,

* significância a 90%

Entre os parenteses apresentamos o desvio padrão.

Abordagem Clássica vs. Bayesiana

	Estatística Clássica	Estatística Bayesiana
Formação académica do encarregado de educação		
Sem habilitações		
Básico	-0.662 (0.473)	-0.664 (0.496)
Secundário	-0.137 (0.481)	-0.138 (0.505)
Diplomado	1.396*** (0.484)	1.393 Sig (0.509)
Beneficiário da ASE (Apoio escolar: alimentar, material e transportes)		
Não beneficia (famílias mais ricas)		
Escalão B	-0.399** (0.176)	-0.400 Sig (0.177)
Escalão A (famílias pobres)	-0.521** (0.226)	-0.522 Sig (0.226)
Internet em casa		
Não		
Sim	0.251* (0.145)	0.253 (0.145)
Curso		
Ciências		
Economia	-0.424** (0.193)	-0.426 Sig (0.193)
Humanidades e Artes	-0.716 (0.455)	-0.715 (0.457)

Observações	5000
R ²	0.1245

*** significância a 99%

** significância a 95%,

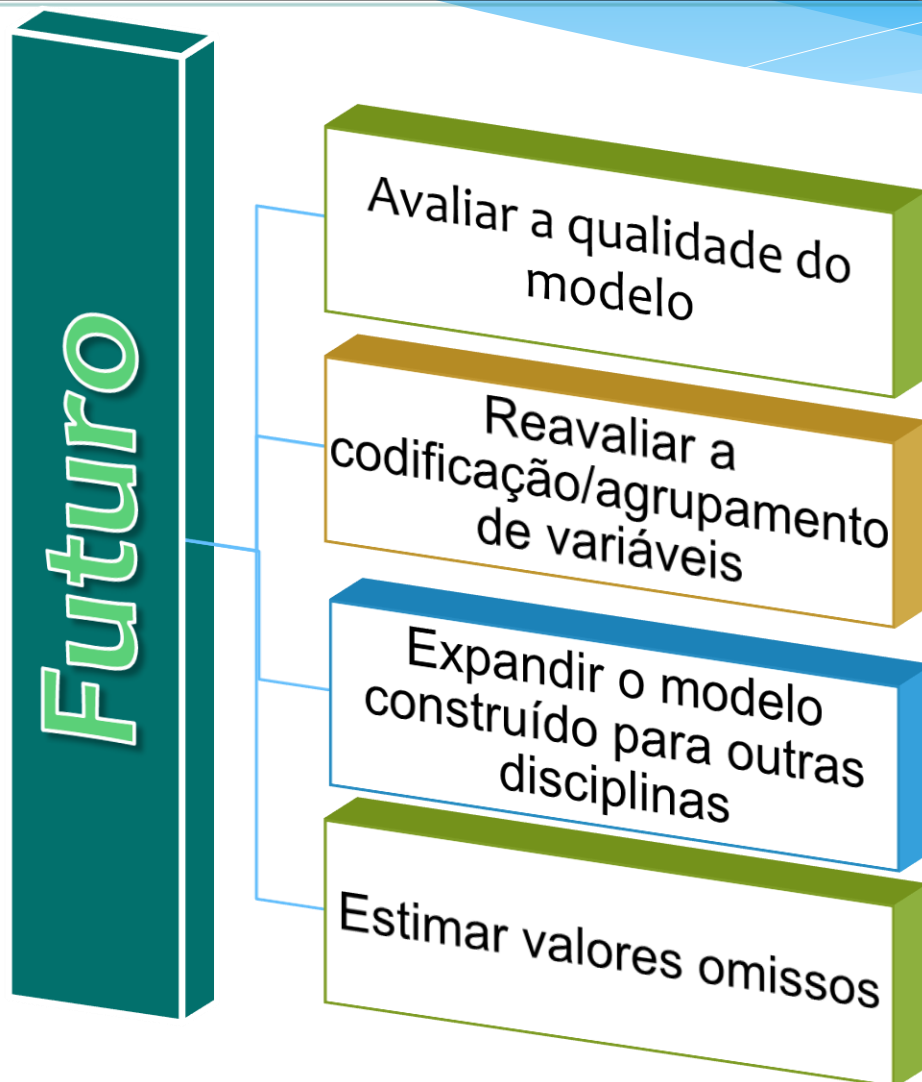
* significância a 90%

Entre os parenteses apresentamos o desvio padrão.

Conclusões

Variáveis	Classe de Referência	Impacto da variável
sexo(1= Mulher)	Homem	↑
NUTSII0_2 (1= LAA)	Norte e Centro	↓
dimturma	n.a. (variável quantitativa)	↑
idade	n.a. (variável quantitativa)	↓
numeromatriculas	n.a. (variável quantitativa)	↓
Formacaoacademica (1=Diplomado)	Sem habilitações	↑
beneficiariosase_2 (1=Escalão B)	Não beneficia	↓
beneficiariosase_3(1=Escalão A)	Não beneficia	↓
curso0_2 (1=Economia)	Ciências	↓

Comentários e trabalhos futuros



- “The Bugs Book A Practical Introduction to Bayesian Analysis”, Lunn David e outros.
- “Abordagem Bayesiana para modelar dados com excesso de zeros- aplicação à Parasitologia”, João Filipe Azevedo dos Santos.
- Handouts das aulas de Estatística Bayesiana, Patrícia Cortés de Zea Bermudez.

DGEEC/MEC

pedro.martins@dgeec.mec.pt

ricardo.santos@dgeec.mec.pt

Obrigado!