Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Estatística

Aplicação de um modelo Beta Auto-Regressivo a indicadores educacionais do Brasil

LU TAI YI – 200755016

Juiz de Fora

2011

Lu Tai Yi

200755016

Aplicação de um modelo Beta Auto-Regressivo a indicadores educacionais do Brasil

Monografia apresentada ao Departamento da Estatística da Universidade Federal de Juiz de Fora como requisito necessário para obtenção do grau de Bacharel em Estatística.

Orientador : Prof. Dr. Tufi Machado Soares

Juiz de Fora

2011

Lu Tai Yi

200755016

Aplicação de um modelo Beta Auto-Regressivo a indicadores educacionais do Brasil

BANCA EXAMINADORA

Tufi Machado Soares

Pós-Doutor em Estatística – Universidade Federal de Rio de Janeiro

---

Clécio da Silva Ferreira

Doutor em Estatística – Universidade de São Paulo

Joaquim Henriques Vianna Neto

Mestre em Estatística – Universidade do Rio de Janeiro

“Aprender e praticar o que se aprendeu frequentemente, não é uma alegria? Ter amigos que vêm de longe, não é uma felicidade? Ficar imperturbável quando não compreendidos pelos outros, não é um homem nobre?”

Confúcio

“Eu me examino três vezes por dia: Se não sou leal àqueles em nome dos quais estou agindo? Se não sou de confiança ao lidar com meus amigos? Se não me pratico o que me foi pregado?”

Zeng Zi

Agradecimentos

A Deus, por toda a sua condução e proteção no trajeto da minha vida.

Aos meus pais, pelas suas ajudas e compreensões nos momentos difíceis, pelos conselhos no caminho da minha vida.

Aos meus irmãos, pelos bons momentos que passamos juntos.

Aos meus amigos, Carol, Jarbas, Laura, Marcos, Priscila, Samuel, Thales, Thiago e Victor, pelo companheirismo durante todos esses anos de curso.

À Sarah, Raquel,Leiliane e Manoel pela amizade, por me apoiarem, por me ajudarem e pelas suas companhias em todos momentos de dificuldades e alegria do ICE.

Aos professores do Departamento da Estatística pelos ensinamentos.

Ao meu orientador, Professor Tufi, pela excelente orientação e supervisão neste trabalho.

Ao Neimar, Alexandre, Raquel Morani e Douglas, pela ajuda nesse trabalho.

Resumo

Pretende-se propor um modelo previsor através da técnica de regressão beta para algumas séries educacionais, tais como o número e o percentual de concluintes do Ensino Fundamental e Médio. Análise com dados da Pesquisa Nacional por Amostra de Domicílios – PNAD – realizada pelo Instituto Brasileiro de Geografia e Estatística – IBGE – do ano 1976 até 2009, demonstram a eficácia e a eficiência do modelo.

Palavras-Chave: Regressão beta, PNAD, IBGE

Sumário

Gráficos e Figuras

1 - Introdução

A educação sempre foi e sempre será uma área de interesse de todos, pois ela influencia diretamente no desenvolvimento futuro do país. Neste trabalho, foram utilizados os dados da Pesquisa Nacional por Amostra de Domicílios – PNAD – realizada pelo Instituto Brasileiro de Geografia e Estatística – IBGE – do ano 1976 até 2009, a fim de fazer um estudo das séries históricas sobre escolaridade do brasileiro e, a partir dela, produzir com precisão para os próximos anos.

O levantamento da PNAD iniciou-se no segundo trimestre de 1967 e os resultados, então, eram apresentados trimestralmente, o que permaneceu até o primeiro trimestre de 1970. A partir de 1971, passaram a ser realizados com periodicidade anual, ocorrendo sempre no último trimestre. Nos anos 1980, 1991 e 2000 foram interrompidas para a realização do censo, e em 1994, por razões excepcionais, não foi realizado o levantamento. Os dados referentes a estes anos faltantes serão imputados pela interpolação.

A escolha da PNAD 1976 como a primeira série de análise ocorre por falta de disponibilidade das bases nos anos anteriores a 76. Não foram encontrados no *site* do IBGE, os dados das PNADs anteriores a 1976. Lembrando que a PNAD não é realizada nos anos em que há realização de Censo e a sua seleção é feita por meio de uma amostra probabilística de domicílios em três estágios. (PNAD 2006)

A pesquisa feita pelo IBGE tem como finalidade a produção de informações básicas para o estudo do desenvolvimento socioeconômico do País. Este é um sistema de pesquisa de amostra por domicílio que investiga várias características socioeconômicas. De caráter permanente nas pesquisas são coletadas variáveis referentes às características gerais da população, a educação, trabalho, rendimento e habitação, e outras que variam com periodicidade são características da migração, fecundidade, nupcialidade, saúde, nutrição e outros temas que são incluídos no sistema de acordo com as necessidades de informação para o País. (PNAD 2006)

Amostras obtidas na PNAD foram usadas a técnica da seleção probabilística de domicílios, cuja seleção da amostra é feita em três estágios, onde as unidades primárias são os municípios, as secundárias são os setores censitários e as terciárias são os domicílios (domicílios particulares e unidades de habitação em domicílios coletivos). (PNAD 2006)

Uma das características marcantes da PNAD é a sua especificidade, isto é, em cada ano de pesquisa trata-se de um foco temático diferente. Por exemplo, em 1981, o foco era na área de saúde, 1982 na educação e assim por diante, variando o seu tema ano a ano de acordo com as necessidades de informação do país.

A maioria das bases digitais das PNADs vieram divididas em 2 arquivos eletrônicos, um para domicílio e outro para as pessoas que moram nos domicílios selecionados. A primeira contém as informações dos bens e da casa, tais como o número de eletrônicos, saneamento básico, tipo de construção e entre outros. Já

na segunda, são apresentadas as características gerais do morador, como por exemplo, grau de instrução, gênero, idade, situação de emprego e entre outros.

Conjuntamente com o Censo, a PNAD se constitui o segundo maior trabalho realizado pelo IBGE. Este não faz a estimação da população, mas seus pesos projetam para a população brasileira estimada. A base é constituída por mais de 200 variáveis e cerca de 300 mil casos para cada ano de pesquisa. Considerando também que a técnica de armazenagem dos registros é diferente em cada década, foi feito um grande esforço para importar os dados e uniformizá-los.

O objetivo deste trabalho é, baseando-se nos dados da PNAD, propor um modelo previsor através da técnica de regressão beta para algumas séries educacionais, tais como o percentual de concluintes do Ensino Fundamental e Médio. A partir das informações obtidas de séries históricas, analisam-se os dados para demonstrar a eficácia e a eficiência do modelo, além de realizar previsões para os próximos anos.

2 - Fundamentação Teórica

2.1 - Plano Amostral

A técnica da seleção da amostra utilizada no PNAD é a Amostragem por Conglomerado em três estágios.

A ideia de Amostragem por Conglomerados é dividir a população em conglomerados considerando cada um como uma unidade amostral, que será selecionado posteriormente por Amostragem Aleatória Simples. A partir disso, faz-se uma pesquisa mais abrangente nos grupos selecionados aleatoriamente. Isto é aplicado em amostra de um estágio.

A unidade amostral contém mais de um elemento populacional, isto é, várias informações. Por exemplo: considerando a região como unidade amostral, apresenta também outras informações adicionais como estados, municípios, domicílios etc. Portanto, o princípio da técnica de amostra por conglomerados é tentar maximizar e reduzir as diferenças dentro do grupo e inter-grupo, respectivamente.

A vantagem desta técnica é a facilidade para aplicação e o custo baixo, mas o problema que este apresenta é a alta correlação dentro do mesmo conglomerado e, isto, influencia na sua eficiência.

Para resolver a questão de semelhança dentro do mesmo conglomerado faz-se o sub-sorteio dentro dos grupos selecionados. Este método foi usado na

PNAD, onde se faz a seleção em três estágios. As unidades primárias são os municípios, as secundárias são os setores censitários e as terciárias são unidades domiciliares (domicílios particulares e unidades de habitação em domicílios coletivos).

As unidades primárias (no caso, os municípios) foram classificadas em duas categorias: auto-representativas (probabilidade de pertencer à amostra é igual 1) e não auto-representativas. Os municípios que pertencem à segunda passaram por um processo de estratificação e, foram feitas seleções com reposição em cada estrato e, com probabilidade proporcional à população residente obtida no Censo de 2000.

Nas unidades secundárias (no caso, os setores censitários) foram feitas seleções em cada município da amostra, também usando a probabilidade proporcional e com reposição, sendo utilizado o número de unidades domiciliares existentes por ocasião do Censo de 2000 como medida de tamanho.

No último estágio foram selecionados, com probabilidades iguais em cada setor censitário da amostra, os domicílios particulares e as unidades de habitação em domicílios coletivos para investigação das características dos moradores e da habitação. (PNAD 2006)

2.2 - Peso Amostral

Quando se trabalha com pesquisas de amostragem complexas, como no caso da PNAD, para não ter estimativas com baixa eficiência, é necessário a

aplicação dos pesos nas unidades amostradas. Estes são usados para cada unidade amostrada.

O peso da amostra utiliza estimadores de razão. Na PNAD usa-se a projeção populacional residente em cada Unidade de Federação com o tipo de área (região metropolitana e não metropolitana de divulgação da pesquisa). As projeções populacionais consideram a evolução populacional ocorrida entre os Censos subsequentes, sob hipóteses de crescimento associadas a taxas de fecundidade, mortalidade e migração. (PNAD 2006)

2.3 - Coorte

A palavra coorte vem do latim cujo significado inicial era parte de uma legião romana. Hoje, o termo significa grupo de pessoas com característica em comum que avança em conjunto em relação a um intervalo de tempo. Exemplo: coorte de crianças nascidas em 2006; coorte de pessoas que concluíram Ensino Fundamental em 2001, entre outros.

No presente trabalho, foram feitos estudos das "pseudocoortes" das pessoas nascidas em 1961 até as nascidas em 1994. Para acompanhamento e ilustração da análise das séries históricas, foram utilizadas as nascidas em 1961, 1966, 1971, 1976, 1981, 1986 e 1990.

O motivo por qual foram denominadas de "pseudocoortes" é porque não são coortes verdadeiras, e sim, uma estimação da população feita pelo IBGE.

Uma das dificuldades dos estudos nas séries históricas foi a impossibilidade de estudar e acompanhar as coortes, pois são apenas estimações

feitas pelo IBGE. No entanto, este problema será corrigido a partir desse ano, pois, no PNAD 2011 haverá uma mudança estrutural, e este passa a acompanhar as coortes verdadeiras.

2.4 - Série Temporal

Série temporal é o conjunto de observações igualmente espaçado de uma variável observada em um período de tempo, denominado de janela, e pode ser classificada em discreta ou contínua, determinística ou estocástica, multivariada ou univariada. (Souza 1996)

- Discretas: são os casos em que o conjunto das observações é finito ou infinito numerável.

-Contínuas: são casos em que o conjunto de observações for infinito não numerável.

-Determinísticas: são casos em que uma função matemática exata pode ser usada para a previsão dos valores futuros.

-Estocásticas: são casos que a previsão da série só pode ser estabelecida por uma função densidade de probabilidade, isto é, um termo probabilístico, é composto por variável aleatória.

-Multivariada ou univariada: são casos em que a série é representada por vetor de ordem rx1, ou seja, quando o instante da série é composto por uma ou várias condições. Ex: Na meteorologia, pode haver três condições na série ao mesmo tempo: a temperatura, índice de umidade e a índice pluviométrico.

Como foi mencionada anteriormente, não foram realizadas as PNADs nos anos 1980, 1991, 1994 e 2000, por isso, para não apresentar valores perdidos na análise é preciso fazer uma imputação de dados. Para imputar os dados, foi utilizado o método da "Interpolação Cúbica".

Série temporal pode ser considerada como um processo estocástico, pois, são conjuntos de variáveis aleatórias do elemento x medidos em relação ao tempo.

Há diversos métodos de previsão para séries temporais, e a maioria dos métodos aqui considerados usam as observações de um tempo passado para traçar um padrão de comportamento da série histórica. A série pode ser decomposta em partes: sazonais, tendência cíclica, tendência e aleatória. (para maior detalhe, ver Souza, 1996; Morettin, 1987; Neto, 2003)

- Sazonal refere-se ao movimento semelhante de uma série apresentada durante um mesmo período de tempo dos anos sucessivos.

- Tendência Cíclica refere-se ao movimento oscilante que ocorre ao longo do tempo em torno da linha de tendência.

- Tendência refere-se a um comportamento dominante que a série apresenta durante um período de tempo.

- Aleatória ou também chamada de erro aleatório refere-se à variação da série. É o movimento que apresenta grande instabilidade, isto é, uma oscilação intensa.

A seguir vemos alguns exemplos da série sazonal, cíclica, tendência e aleatória.

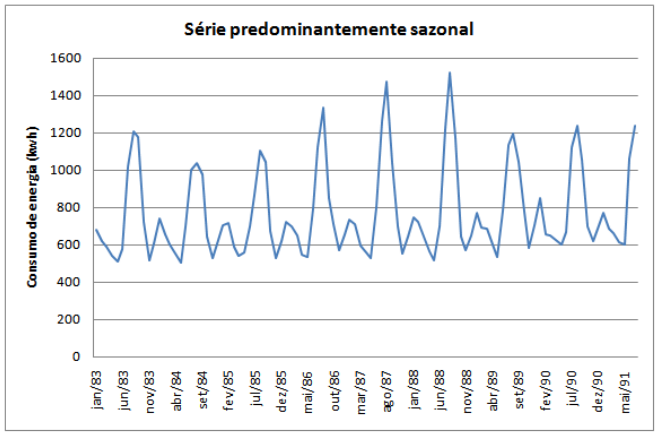


Gráfico 2.4.1 – Série predominantemente sazonal

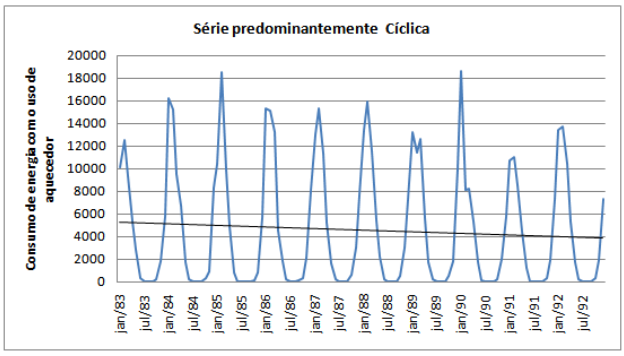


Gráfico 2.4.2 – Série predominantemente cíclica

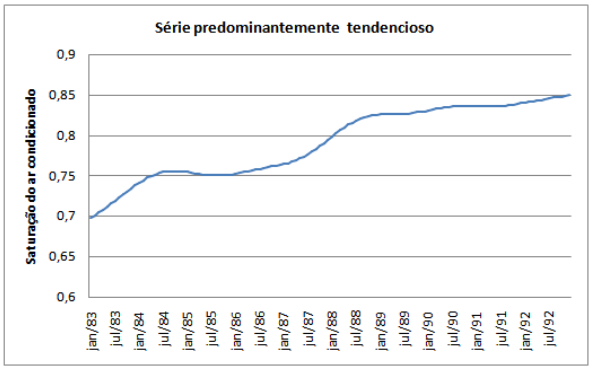


Gráfico 2.4.3 – Série predominantemente tendenciosa

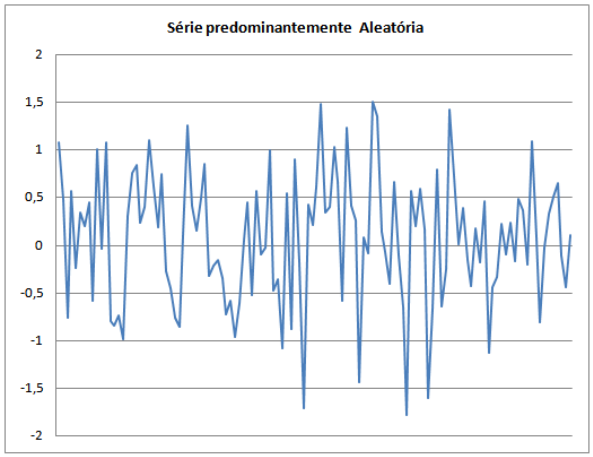


Gráfico 2.4.4 – Série predominantemente aleatório

Alguns dos principais métodos de previsão são Média Móvel, Amortecimento Exponencial e o modelo de Box e Jenkins.

- A média móvel usa a média das últimas N observações para fazer a previsão. Uma das dificuldades é a escolha de N. Quando menor for o N, a previsão tende a oscilar mais. Este método considera o peso igual para todas as observações em que a série seja numérica.

- O amortecimento exponencial assume que as observações extremas da série representam a aleatoriedade, e identifica o padrão através do amortecimento desses valores. A ideia geral desse método é que a média seja uma média ponderada de todas as observações, e o peso das observações decresce, isto é, quanto mais antiga a observação, menos peso ela recebe e vice-versa.

Este método é mais usado no dia a dia por ser simples de aplicar, de fácil entendimento e por gerar boas previsões.

- O modelo Box & Jenkins, se baseia na tentativa de ajustes de modelo do tipo ARIMA, que apresenta as duas ideias: o princípio da parcimônia e um ciclo iterativo para a construção do modelo.

A ideia de parcimônia se baseia na escolha do modelo que apresenta o menor número de parâmetros possíveis, para poder representá-lo matematicamente adequado.

O ciclo iterativo é a aplicação das iterações, até que obtenhamos um modelo satisfatório. A seguir, um fluxograma para explicar melhor sobre as iterações do modelo Box & Jenkins.

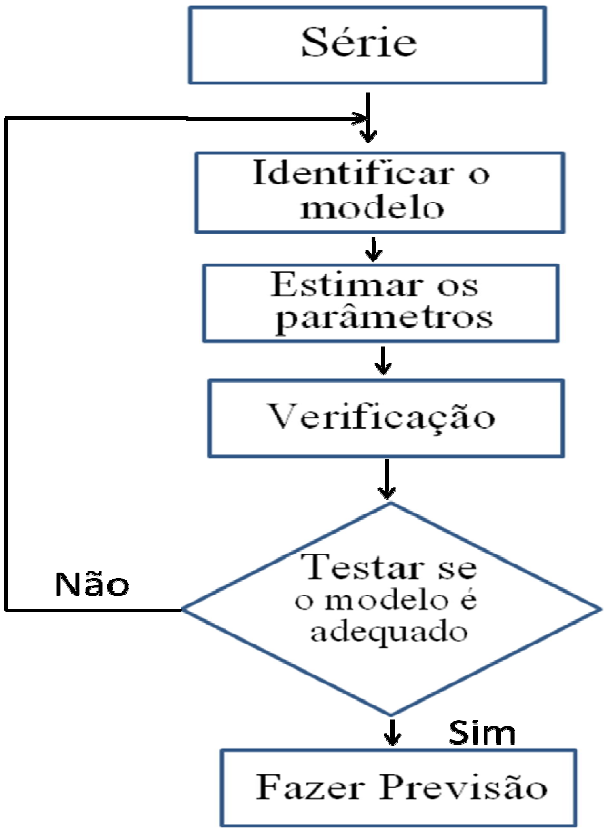


Gráfico 2.4.5 – Fluxograma das iterações do modelo Box & Jenkins.

Box & Jenkins tem como o objetivo de tornar a diferença entre os valores estimados pelo modelo e valores observados uma série de resíduos que tem um perfil de aleatoriedade próximo de zero.

2.5 - Regressão Dinâmica

Para fazer a estimação da série, foram utilizados alguns métodos, um deles é o Modelo da Regressão Dinâmica.

Para alguns autores, o modelo de regressão dinâmica é diferente dos outros modelos de regressão linear múltipla, pois, os mais usuais possuem algumas características como:

- Média zero;
- Variância constante;
- Distribuição Normal;
- Independência, que resulta a inexistência de correlação serial;

Segundo Barros (2007), a hipótese de independência do ruído não é realista quando se modela a série temporal usando o modelo de regressão, portanto, os resultados e testes usados também não são válidos.

Isso leva a algumas consequências como a subestimação da variância e do erro padrão dos coeficientes de regressão, o que induziria a uma incerteza. Portanto, o uso do modelo de regressão dinâmica na modelagem é importante.

A regressão dinâmica é uma combinação de séries com o efeito de gerar variáveis explicativas cujo objetivo é fazer uma descrição da série com os dados do seu passado usando variáveis explicativas.

O objetivo da regressão dinâmica, segundo Souza (2006), é descrever uma série em função de seu passado. Este admite também a existência de erros estruturados, isto é, erros que apresentam uma estrutura de dependência ao longo do tempo.

De acordo com Souza (2006), a estratégia mais comum na construção de um modelo de regressão dinâmica é "*bottom up*", isto é, a partir de um modelo mais simples que é usar a constante como preditor, incluem-se novas variáveis até encontrar um modelo adequado.

A elaboração de um modelo deste método geralmente é um processo difícil, pois não apenas escolhe as variáveis a serem incluídas no modelo, mas também as defasagens destas variáveis.

O *software* Forecast Pro©[1] é um programa útil para fazer as previsões dos valores através dos métodos variados, uma delas é a modelagem da regressão dinâmica, pois permite uma visualização clara de variáveis defasadas importantes no modelo.

2.6 - Simulação e estimação do modelo usando método *Markov Chain Monte Carlo*

[1] Forecast Pro for Windows version 3.00D Extended Edition.

O estudo de um sistema sempre é feito através de um modelo. A simulação e a estimação é a reprodução do comportamento de um sistema real com a ajuda do modelo.

Para a estimação, o método bayesiano é uma das formas indicadas. Aqui se trabalha com a simulação pelo *Markov Chain Monte Carlo* (MCMC) no *software* computacional WINBUGS©[2].

Markov Chain Monte Carlo

Segundo Ehlers (2003), MCMC é muito útil na resolução dos problemas dos métodos não iterativos em questões complexas, pois a sua simulação é feita através da aplicação da característica da Cadeia de Markov e, ao simulá-lo, gera uma grande quantidade de dados que serão usados para distribuição da *posteriori.*

Atualmente há dois métodos de MCMC mais utilizados: *Gibbs sampling* e *Metropolis and Hasting*.

2.6.1 - *Gibbs sampling* (amostrador de Gibbs)

[2]WinBUGS 1.4.3, desenvolvido por Spiegelhalter, David; Thomas, Andrew; Best, Nicky; Lunn, Dave.

O amostrador de Gibbs (Geman and Geman, 1984) é um dos métodos utilizados nesse aspecto cujo objetivo é simular amostra da distribuição a *posteriori*.

O esquema para esse método começa aplicando um valor inicial arbitrário, a partir deste extrai uma amostra aleatória e, repetindo este passo, nos fornecerá k parâmetros. O critério da conclusão de iterações é quando o modelo consegue convergir.

Para melhor entendimento será apresentado um esquema abaixo com os passos a seguir:

Seja θ, o conjunto de parâmetros a serem estimados.

- 1º Passo: Inicialize a iteração com j =1 e defina um vetor de valores iniciais.

$$\theta^{(0)} = \left(\theta_1^{(1)}, \theta_2^{(1)}, \dots \theta_{n-1}^{(1)}, \theta_n^{(1)}\right);$$

- 2º Passo: Fazendo simulações para obter um novo vetor $\theta^{(j)}$

$$\theta_1^j = \left(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \dots \theta_{n-1}^{(j-1)}, \theta_n^{(j-1)}\right)$$

$$\theta_2^j = \left(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots \theta_{n-1}^{(j-1)}, \theta_n^{(j-1)}\right)$$

$$\vdots$$

$$\theta_n^j = \left(\theta_n | \theta_1^{(j)}, \theta_2^{(j)}, \dots \theta_{n-1}^{(j)}\right)$$

-3º Passo: Repetir esse processo até atingir a convergência.

De maneira geral, *Gibbs* é fácil de ser praticado, no entanto, precisa-se tomar cuidado em relação ao tamanho do número *n*, pois, é preciso ter um *n* suficientemente grande para garantir a convergência do modelo.

2.6.2- *Metropolis and Hasting*

Metropolis and Hasting (Metropolis et al. 1953; Hasting 1970) é outro método a ser utilizado na simulação.

Este método pode ser dividido nos seguintes passos:

- 1º Passo: Inicialize a iteração com j =1 e defina um vetor de valores iniciais.

$$\theta^{(0)} = \left(\theta_1^{(1)}, \theta_2^{(1)}, \dots \theta_{n-1}^{(1)}, \theta_n^{(1)}\right);$$

- 2º Passo: Fazer simulações para obter um novo vetor $\theta^{(*)}$ da distribuição:

$$q\left(\theta_k^* | \theta_k^{j-1}\right), \text{com } k = 1,2,3,\dots,n$$

- 3º Passo: Sendo Z os dados e o símbolo (.) os parâmetros não explicitados. Encontre a razão

$$r\left(\theta_k^{j-1}, \theta_k^*\right) = \frac{P\left(\theta_k^* | (.)\right) q\left(\theta_k^{j-1} | \theta_k^*\right)}{P\left(\theta_k^{j-1} | (.)\right) q\left(\theta_k^* | \theta_k^{j-1}\right)}$$

tal que:

$$r\left(\theta_k^{j-1}, \theta_k^*\right) = \frac{P\left(Z|\theta_1^j, \theta_2^j, \dots, \theta_{k-1}^j, \theta_k^*, \theta_{k+1}^{j-1}, \dots, \theta_n^{j-1}\right) P(\theta_k^*) q\left(\theta_k^{j-1}|\theta_k^*\right)}{P\left(Z|\theta_1^j, \theta_2^j, \dots, \theta_{k-1}^j, \theta_k^{j-1}, \theta_{k+1}^{j-1}, \dots, \theta_n^{j-1}\right) P\left(\theta_k^{j-1}\right) q\left(\theta_k^*|\theta_k^{j-1}\right)}$$

- 4º Passo: Calcule a probabilidade de aceitação:

$$P_{AC} = min\left(1, r\left(\theta_k^{j-1}, \theta_k^*\right)\right)$$

- 5º Passo: Gerar $Z_{AC} \sim Ber(P_{AC})$

- 6º Passo:

$$\begin{cases} \theta_k^j = \theta_k^* & se \quad Z_{AC} = 1 \\ \theta_k^j = \theta_k^{j-1} & se \quad Z_{AC} = 0 \end{cases}$$

- 7º Passo: Repetir o processo até chegar a convergência, ou seja, estacionalidade da distribuição.

2.7 - Resíduos

Não é possível uma análise de regressão sem analisar os resíduos e alguns outros diagnósticos, pois, isso pode conduzir a um erro de interpretação e do ajuste do modelo.

A análise de resíduos é usada para detectar os possíveis problemas do ajuste do modelo. Tais como a detecção da presença dos pontos aberrantes, de alavanca, influentes e entre outros.

Os resíduos são definidos por:

$$r_i = Y_i - \hat{Y}_i = \varepsilon_i = Y_i - \widehat{E(Y_i)}$$

Ou seja, este é definido pela subtração de valor observado pelo valor esperado da i-ésima observação.

Para o modelo de regressão, supõe-se que:

$$\varepsilon_i \sim N(0, \sigma^2)$$

A detecção dos resíduos pode ser feita através dos gráficos de diagnóstico, como por exemplo o uso do gráfico de QQ-plot. Os pontos devem estar em torno de uma reta para validar o fato de que o pressuposto dos erros aleatórios é Normal.

Para saber se o modelo é adequado ou não, os resíduos devem satisfazer algumas suposições: Os erros devem ser normalmente distribuídos e não

correlacionados e, apresentar a média zero e variância constante independente do tempo.

A violação das hipóteses acima pode levar instabilidade ao modelo, portanto, precisa de muita atenção ao fazer a análise de resíduo.

2.8 - Regressão Beta

Existem diversos modelos de regressão na literatura estatística, cujo objetivo é a análise dos dados e da correlação entre possíveis variáveis respostas (dependentes) com variáveis independentes, a fim de estudar seus possíveis problemas, e tentar fazer a análise de previsão. No entanto, o uso da maioria dos métodos tradicionais não é adequado nos casos da análise de variáveis dependentes com dados contínuos, tais como taxas e porcentagens, pois, muitas das vezes, essas taxas ficam em torno do intervalo de (0,1) e, ajustes feitos pelos modelos clássicos geralmente excedem dos limites.

Existem diversos estudos sobre a regressão Beta na literatura estatística. Para maiores detalhes olhe Kieschnick & McCullough (2003), Brehm & Gates (1993), Johnson, Kotz & Balakrishnan (1995), Miyashiro (2008), Paolino (2001), Buckley (2003), Ferrari & Cribari-Neto (2004), Migonet al. (2011) e entre outros.

Ferrari & Cribari-Neto (2004) propuseram o modelo de regressão beta para poder analisar os dados de variáveis contínuas limitadas no intervalo [0,1].

A ideia deste modelo é a suposição dos dados seguirem as características da distribuição beta, que são os dados da variável aleatória contínua no intervalo [0,1]. A função densidade da beta é dada por:

$$p(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1$$

A média é dada pela expressão:

$$E(Y) = \frac{p}{(p+q)}$$

E a variância

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$$

Nota-se que a distribuição apresenta dois parâmetros "p" e "q", logo, a curva da densidade altera de acordo com a mudança desses dois, com isso, pode ter uma aplicação ampla em diversas questões práticas.

No modelo proposto por Ferrari & Cribari-Neto (2004) sugere-se uma reparametrização da distribuição beta, considerando a média da resposta e um parâmetro da dispersão. Estes podem ser modelados com o uso de um preditor que relaciona a média com a covariável mediante uma função de ligação.

Dado

$$E(Y_t) = \mu_t = \frac{p}{p+q} \qquad (**)$$

e

$$\phi = p + q \qquad (*)$$

Resolvendo as equações acima fazendo substituições (*) em (**), obtém-se:

$$p = \mu_t \phi$$

e

$$q = (1 - \mu_t)\phi$$

Se uma variável Y segue a distribuição beta, isto é, Y~B(μ_t, ϕ), então a função densidade é dada pela expressão abaixo:

$$f(y_t; \mu_t, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_t \phi)\Gamma\left((1-\mu_t)\phi\right)} {y_t}^{\mu_t \phi - 1}(1 - y_t)^{(1-\mu_t)\phi - 1}, \ 0 < y_t < 1,$$

onde

$$p > 0 \quad \text{e} \quad q > 0$$

Logo

$$E(Y) = \mu_t \qquad \text{e} \qquad \text{Var}(y) = \frac{V(\mu_t)}{1+\phi}$$

onde

$$V(\mu_t) = \frac{\mu_t \phi (1 - \mu_t)\phi}{\phi^2(\phi + 1)} = \frac{(1 - \mu_t)\mu_t}{\phi + 1}$$

Em que $V(\mu_t)$ é uma função da variância, μ_t é a média da variável dependente para cada coorte, no caso aqui refere-se à porcentagem de concluintes de cada coorte. E ϕ, o parâmetro da dispersão, logo, quanto maior o ϕ, menor será a variância de *Y*.

3 - Metodologia

Foram analisadas no presente trabalho as bases da PNAD pessoa do ano 1976 ~ 2009, exceto os anos da realização do Censo.

A análise foi dividida por séries históricas de escolaridades diferentes: “Conclusão do Ensino Fundamental” e “Conclusão do Ensino Médio”.

Para isso, foram utilizadas as variáveis que se relacionam com a parte da escolaridade e as características pessoais do entrevistado, tais como: idade, gênero, cor da pessoa, anos de estudo, entre outras.

3.1 - Conclusão do Ensino Fundamental

Para esta série, o método da seleção das pessoas divide-se em 2 partes.

A primeira parte da seleção é feita nas PNADs anteriores ao do ano 1992, ou seja, do ano 1976 até 1990, pois, não há dados detalhados sobre as informações de interesse do trabalho.

Portanto, a seleção foi feita utilizando a variável “Anos de Estudo” da base. Esta contém informações sobre o total de anos de estudo feito pelo entrevistado, que será categorizado como segue abaixo:

ANOS DE ESTUDO :

```
 1 - SEM INST/MENOS 1
 2 - 1 ANO
 3 - 2 ANOS
 4 - 3 ANOS
 5 - 4 ANOS
 6 - 5 ANOS
 7 - 6 ANOS
 8 - 7 ANOS
 9 - 8 ANOS
10 - 9 A 11 ANOS
11 - 12ANOS OU MAIS
12 - NÃO DETERMINADOS
13 - SEM DECLARAÇÃO
```

Figura 3.1.1 – Questão "Anos de Estudo" retirado do dicionário da pesquisa do PNAD 1976

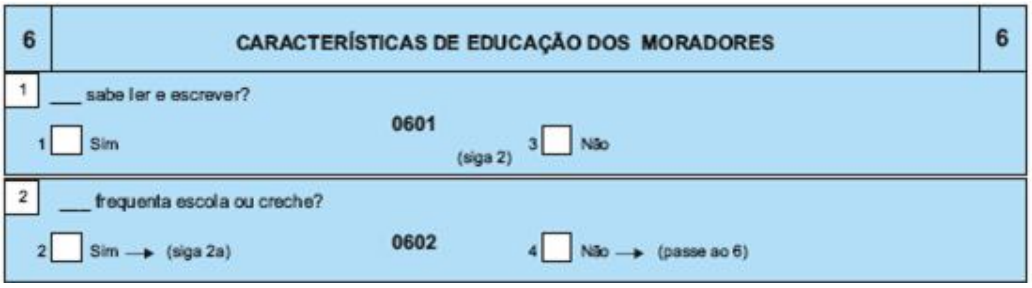
A partir disso, foram selecionadas as pessoas que completaram idade de 15 anos ou mais e tenham realizado 8 anos ou mais de estudos.

Já a segunda parte da seleção é feita nas PNADs começando a partir do ano 1992 até 2009. Nessas pesquisas mais recentes houve uma mudança estrutural no modelo e na forma de pergunta, isto é, as informações são apresentadas com mais detalhes, principalmente sobre a parte da escolaridade do entrevistado.

A seguir, as perguntas referentes à parte das características de escolaridade dos moradores. Essas são retiradas do questionário da PNAD realizada no ano 2009.

Aqui, as perguntas são divididas em duas partes: uma parte foi feita para os moradores que frequentavam a escola e a outra parte foi direcionada para os que não frequentavam mais nenhuma instituição de ensino.

Em geral, as questões são iguais nos períodos de 1992~1999, 2001~2006 e 2007~2009, o que muda entre um período e o outro são algumas das categorias das variáveis e a inclusão de novas perguntas.



6	CARACTERÍSTICAS DE EDUCAÇÃO DOS MORADORES	6

1 ___ sabe ler e escrever?

1 ☐ Sim — 0601 — (siga 2) — 3 ☐ Não

2 ___ frequenta escola ou creche?

2 ☐ Sim → (siga 2a) — 0602 — 4 ☐ Não → (passe ao 6)

Figura 3.1.2 – Parte do questionário retirado da PNAD 2009

PARA PESSOA QUE FREQUENTA ESCOLA

2a A escola ou creche que ___ frequenta é:

2 ☐ Pública → (siga 2b)

6002

4 ☐ Privada → (passe ao 3a)
(Obs: Inclusive a pessoa que frequenta estabelecimento mantido por associação de moradores, empresa, fundação, ONG etc.)

2b A escola pública que ___ frequenta é:

6020

NOVO

2 ☐ Municipal

4 ☐ Estadual

6 ☐ Federal

(siga 3a)

3a Qual é o curso que ___ frequenta?

01 ☐ Regular do ensino fundamental → (siga 3b)

02 ☐ Regular do ensino médio → (passe ao 5)

6003

03 ☐ Educação de jovens e adultos ou supletivo do ensino fundamental

04 ☐ Educação de jovens e adultos ou supletivo do ensino médio

(passe ao 4)

05 ☐ Superior de graduação → (passe ao 5)

06 ☐ Alfabetização de jovens e adultos

07 ☐ Creche

08 ☐ Classe de alfabetização - CA

09 ☐ Maternal, jardim de infância, etc.

10 ☐ Pré-vestibular

11 ☐ Mestrado ou doutorado

(Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário, passe ao 12.)

3b A duração deste curso que ___ frequenta é de:

6030

1 ☐ 8 anos

3 ☐ 9 anos

(passe ao 5)

6 CARACTERÍSTICAS DE EDUCAÇÃO DOS MORADORES 6

4 Este curso que ___ frequenta é seriado?

2 ☐ Sim → (siga 5)

0604

4 ☐ Não → (Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário passe ao 12.)

5 Qual é a série que ___ frequenta?

1 ☐ Primeira

2 ☐ Segunda

3 ☐ Terceira

4 ☐ Quarta

5 ☐ Quinta

0605

6 ☐ Sexta

7 ☐ Sétima

8 ☐ Oitava

0 ☐ Nona

(Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário, passe ao 12.)

Figura 3.1.3–Parte do questionário da PNAD para as pessoas que frequentam escola

PARA PESSOA QUE NÃO FREQUENTA ESCOLA

6 Anteriormente ___ frequentou escola ou creche?

2 ☐ Sim → (siga 7a) **0606** 4 ☐ Não → (Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário, passe ao 12.)

7a Qual foi o curso mais elevado que ___ frequentou anteriormente?

01 ☐ Elementar (primário) → (passe ao 9)

02 ☐ Médio 1º ciclo (ginasial etc.)

03 ☐ Médio 2º ciclo (científico, clássico etc.)

} (siga 8)

6007

04 ☐ Regular do ensino fundamental ou do 1º grau → (siga 7b)

05 ☐ Regular do ensino médio ou do 2º grau → (passe ao 9)

06 ☐ Educação de jovens e adultos ou supletivo do ensino fundamental ou do 1º grau

07 ☐ Educação de jovens e adultos ou supletivo do ensino médio ou do 2º grau

} (passe ao 8)

08 ☐ Superior de graduação → (passe ao 9)

09 ☐ Mestrado ou doutorado → (passe ao 11)

10 ☐ Alfabetização de jovens e adultos

11 ☐ Creche

12 ☐ Classe de alfabetização - CA

13 ☐ Maternal, jardim de infância, etc.

} (Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário, passe ao 12.)

7b A duração deste curso que ___ frequentou anteriormente era de:

6070

1 ☐ 8 anos 3 ☐ 9 anos

(passe ao 9)

6 CARACTERÍSTICAS DE EDUCAÇÃO DOS MORADORES 6

8 Este curso que ___ frequentou anteriormente era seriado:

2 ☐ Sim → (siga 9) **0608** 4 ☐ Não → (passe ao 11)

9 ___ concluiu, com aprovação, pelo menos a primeira série deste curso que frequentou anteriormente?

1 ☐ Sim → (siga 10) **0609** 3 ☐ Não → (Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário, passe ao 12.)

10 Qual foi a última série que ___ concluiu, com aprovação, neste curso que frequentou anteriormente?

1 ☐ Primeira

2 ☐ Segunda

3 ☐ Terceira

4 ☐ Quarta

5 ☐ Quinta

6 ☐ Sexta

7 ☐ Sétima

8 ☐ Oitava

0 ☐ Nona

0610

(siga 11)

11 ___ concluiu este curso que frequentou anteriormente?

1 ☐ Sim **0611** 3 ☐ Não

(Se morador de 10 anos ou mais de idade, siga quesito 11a. Caso contrário, passe ao 12.)

Figura 3.1.4 – Parte do questionário da PNAD para as pessoas que frequentam escola

A seleção dos concluintes do Ensino Fundamental, de novo, foi dividida em duas partes.

Na primeira parte se seleciona os moradores que frequentavam a escola e, estavam cursando Regular do Ensino Médio, Educação de Jovens e Adultos (EJA) ou Supletivo do Ensino Médio, Ensino Superior, Mestrado ou doutorado.

A segunda parte envolve os moradores que não frequentavam mais a escola e tinham realizado alguma das modalidades citadas na primeira parte da seleção. Além disso, se seleciona também os que frequentavam Regular do Ensino Fundamental, EJA ou Supletivo do Ensino Fundamental e tinham concluído na modalidade em que estudavam.

3.2- Conclusão do Ensino Médio

A seleção desta série seleciona nas PNADs a partir de 92, as pessoas que frequentavam "ensino superior de graduação", "pré-vestibular" e "mestrado ou doutorado" do grupo que estudava na época, e, no grupo de pessoas que não frequentavam mais a escola, além dos itens citados acima, selecionaram também os que responderam na questão "o curso mais elevado que frequentou anteriormente" as seguintes respostas: "Médio 2° ciclo", "regular do ensino médio ou 2° grau" ou "educação de jovens e adultos ou supletivo do ensino médio ou do 2° grau" e tinham concluído este curso. Aqui selecionam as pessoas que tinham 18 anos de idade ou mais.

Como os dados das PNADs anteriores ao ano 1992 não apresentaram detalhes sobre a conclusão do Ensino Médio, ele foi apresentado como “9 a 11 anos de estudo”, ou seja, só tem-se a informação de quem entrava ou estava nesta modalidade, mas não se sabe da conclusão. Portanto, para obter dados sobre a conclusão de Ensino Médio, foi feita a estimação usando os modelos de Regressão Dinâmica para obter o valor inicial de cada coorte. A partir dos valores obtidos, estimar os dados do passado e prever o futuro através do modelo que será apresentado neste trabalho.

4 - Modelagem

Dado o modelo

$$\boxed{Y_{k,t}|\mu_{k,t},\phi \sim beta(\mu_{k,t},\phi)}$$

Onde

k = coorte e t = ano

$Y_{k,t}$: percentual da coorte k no tempo t.

$\mu_{k,t}$: média de $Y_{k,t}$.

ϕ : um parâmetro de dispersão.

$$\boxed{\mu_{k,t}|Y_{k,t-1} = \alpha_{0,k} + (1 - \alpha_{0,k}) * \alpha_{1,k} * Y_{k,t-1}}$$

onde

$$\alpha_{0,k} \sim beta\ (\mu_{0,k}, H_0)$$

$$\alpha_{1,k} \sim beta\ (\mu_{1,k}, H_1)$$

$$\boxed{\mu(\alpha_{0,k}|\alpha_{0,k-1}) = \gamma_{00} + \gamma_{01} * (1 - \gamma_{00}) * \alpha_{0,k-1}}$$

$$\text{Priori} \begin{cases} \gamma_{00} \sim beta(G_{00}, H_{00}) \\ \gamma_{01} \sim beta(G_{01}, H_{01}) \end{cases}$$

$$\boxed{\mu\left(\alpha_{1,k}\middle|\alpha_{1,k-1}\right) = \gamma_{10} + \gamma_{11} * (1-\gamma_{10}) * \alpha_{1,k-1}}$$

$$\text{Priori} \begin{cases} \gamma_{10} \sim beta(G_{10}, H_{10}) \\ \gamma_{00} \sim beta(G_{11}, H_{11}) \end{cases}$$

Cálculo do limite $\mathrm{E}(\mu_t)$, note que,

$$\mathrm{E}(\mu_t|\mu_{t-1}) = \alpha_0 + (1-\alpha_0)\alpha_1\mu_{t-1}$$

Logo,

$$\mathrm{E}(\mu_2) = \alpha_0 + (1-\alpha_0)\alpha_1\mu_1$$

e,

$$\mathrm{E}(\mu_3) = \alpha_0 + (1-\alpha_0)\alpha_1\mu_2$$

$$= \alpha_0 + (1-\alpha_0)\alpha_1(\alpha_0 + (1-\alpha_0)\alpha_1\mu_1)$$

$$= \alpha_0 + \alpha_0\alpha_1(1-\alpha_0) + \alpha_1^2(1-\alpha_0)^2\mu_1,$$

$$\mathrm{E}(\mu_4) = \alpha_0 + (1-\alpha_0)\alpha_1\mu_3$$

$$= \alpha_0 + (1-\alpha_0)\alpha_1(\alpha_0 + \alpha_0\alpha_1(1-\alpha_0) + \alpha_1^2(1-\alpha_0)^2\mu_1)$$

$$= \alpha_0 + \alpha_0\alpha_1(1-\alpha_0) + \alpha_0\alpha_1^2(1-\alpha_0)^2 + \alpha_1^3(1-\alpha_0)^3\mu_1$$

Portanto,

$$\mathrm{E}(\mu_t) = \alpha_0 + \alpha_0\alpha_1(1-\alpha_0) + \alpha_0\alpha_1^2(1-\alpha_0)^2 + \alpha_0\alpha_1^3(1-\alpha_0)^3 + \cdots$$
$$+ \alpha_0\alpha_1^{t-2}(1-\alpha_0)^{t-2} + \alpha_1^{t-1}(1-\alpha_0)^{t-1}\mu_1$$

$$= \alpha_0(1 + \alpha_1(1-\alpha_0) + \alpha_1^2(1-\alpha_0)^2 + \cdots + \alpha_1^{t-2}(1-\alpha_0)^{t-2}) + \alpha_1^{t-1}(1-\alpha_0)^{t-1}\mu_1$$

$$= \alpha_0\frac{\left(1-\left(\alpha_1(1-\alpha_0)\right)^{t-1}\right)}{\left(1-\left(\alpha_1(1-\alpha_0)\right)\right)} + \alpha_1^{t-1}(1-\alpha_0)^{t-1}\mu_1$$

Onde a última igualdade é consequência da soma de progressão geométrica da razão e condição inicial.

$$\lim_{t\to\infty} E\left(\mu_{k,t}\right) = \alpha_{0,k}\frac{1}{1-\alpha_{1,k}\left(1-\alpha_{0,k}\right)}$$

$$\lim_{t\to\infty} E\left(\mu_{k,t}\right) = \frac{\alpha_{0,k}}{1+\alpha_{0,k}\alpha_{1,k}-\alpha_{1,k}} \leq 1$$

5 - Gráfico das séries históricas para a conclusão do Ensino Fundamental e Médio.

5.1 - Ensino Fundamental

A seguir, gráficos de algumas séries históricas do percentual de concluintes do Ensino Fundamental por coorte em cada ano.

Gráfico 5.1.1 – Percentual de concluintes do Ensino Fundamental nascidos em 1961.

Gráfico 5.1.2 – Percentual de concluintes do Ensino Fundamental nascidos em 1966.

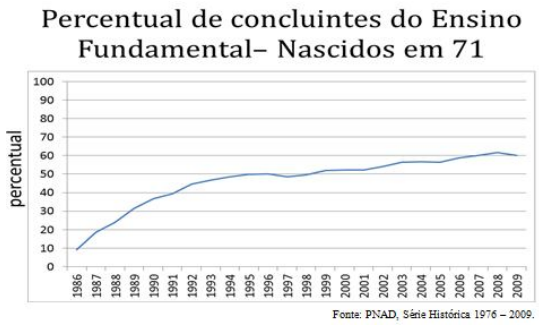


Gráfico 5.1.3 – Percentual de concluintes do Ensino Fundamental nascidos em 1971.

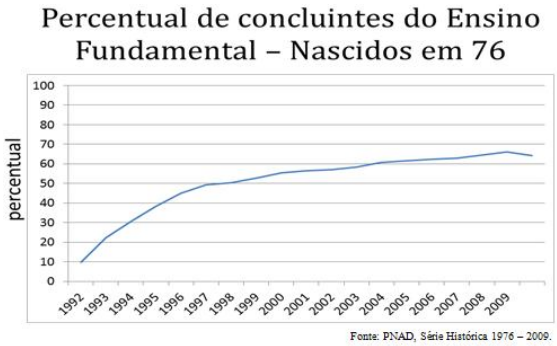


Gráfico 5.1.4 – Percentual de concluintes do Ensino Fundamental nascidos em 1976.

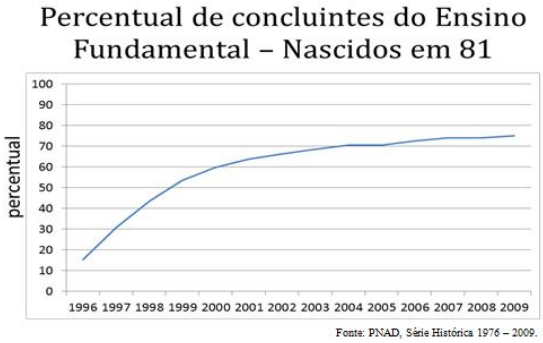


Gráfico 5.1.5 – Percentual de concluintes do Ensino Fundamental nascidos em 1981.

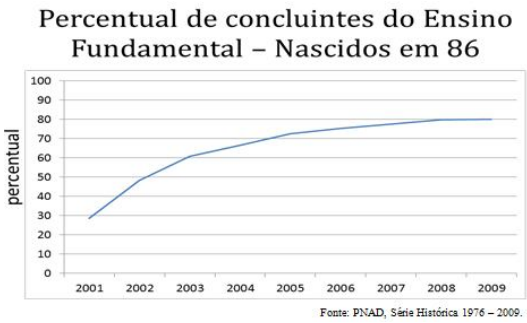


Gráfico 5.1.6 – Percentual de concluintes do Ensino Fundamental nascidos em 1986.

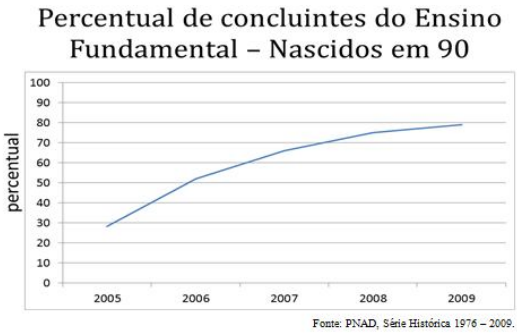


Gráfico 5.1.7 – Percentual de concluintes do Ensino Fundamental nascidos em 1990.

Observando os gráficos acima, nota-se que em todas as séries apresentam um comportamento similar, isto é, o crescimento nos primeiros anos e se estabiliza por algum período, depois volta a crescer novamente.

O gráfico abaixo é a junção de todas as coortes aqui acompanhadas:

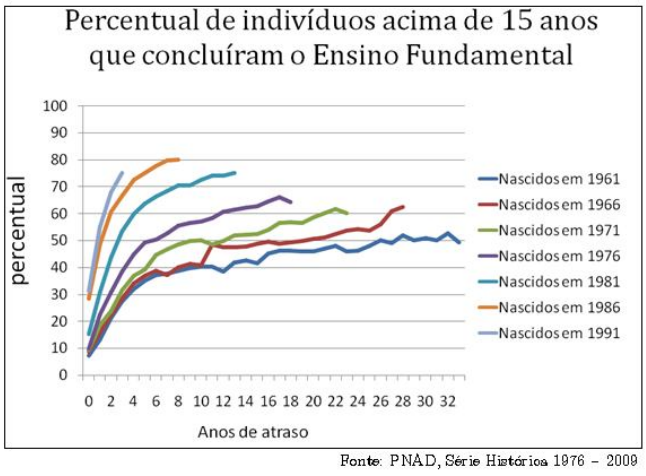


Gráfico 5.1.8 – Percentual de indivíduos acima de 15 anos que concluíram o Ensino Fundamental.

Nota-se que há um aumento no valor inicial e final de percentual dos concluintes para coortes, isto seria um bom sinal, pois, indica um aumento do grau de escolaridade brasileiro.

5.2 - Ensino Médio

A seguir, gráficos de algumas séries históricas do percentual de concluinte do Ensino Médio por coorte em cada ano.

Gráfico 5.2.1 – Percentual de concluintes do Ensino Médio nascidos em 1976.

Gráfico 5.2.2 – Percentual de concluintes do Ensino Médio nascidos em 1981.

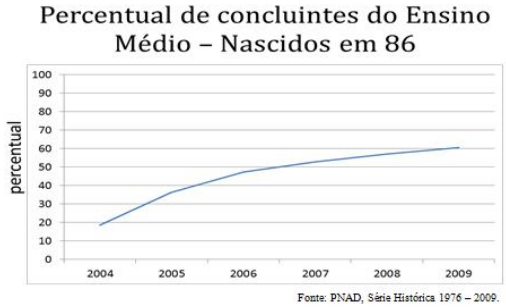


Gráfico 5.2.3 – Percentual de concluintes do Ensino Médio nascidos em 1986.

Gráfico 5.2.4 – Percentual de concluintes do Ensino Médio nascidos em 1990.

O gráfico abaixo é a junção de todas as coortes aqui analisados.

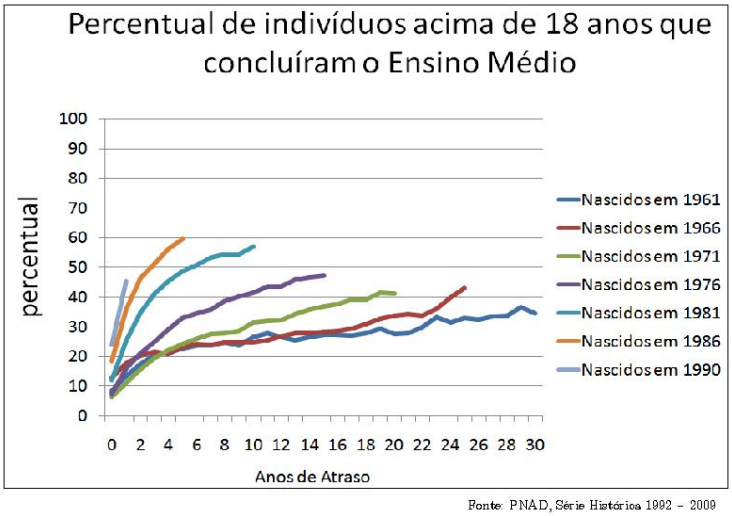


Gráfico 5.2.5 – Percentual de indivíduos acima de 18 anos que concluíram do Ensino Médio.

Nota-se que o comportamento dos valores iniciais e finais das coortes nesse gráfico do percentual dos indivíduos acima de 18 anos que concluíram o Ensino Médio, observa-se que houve um crescimento no percentual de concluintes do Ensino Médio, ou seja, nos anos recentes, aumentou a porcentagem de pessoas que concluíram esta modalidade de Ensino.

6 - Resultados da simulação

A seguir serão apresentados os resultados dos valores estimados de simulação via MCMC feitos por programas computacionais, usando o modelo mostrado anteriormente e com resultado de 45000 iterações com *Burn in* de 35000 e amostrada as 10000 a 5 iterações, obtendo um total de 2000 amostras da porteriori.

A seguir são apresentados os hiperparâmetros dos modelos do Ensino Fundamental e do Médio.

6.1 Ensino Fundamental

No gráfico de α_0 (valores iniciais, em anexo B.1.1), nota-se que este apresenta uma tendência de subida ao longo do tempo. Isto é, há um crescimento nos valores de coeficiente desse parâmetro.

Note que o gráfico dos valores de α_1(parâmetro que define a curvatura, em anexo B.1.2), para cada coorte são praticamente constantes.

Observando o gráfico da cadeia dos parâmetros (em anexo B.1.3), nota-se que todas elas não apresentam nenhuma tendência, portanto, parecem convergir.

Observando os gráficos de FAC (em anexo B.1.4), nota-se que os parâmetros não apresentam autorrelação, exceto γ_{10} e γ_{11}. Assim os percentis das posteriois são consistentemente estimados.

Ns gráficos dos parâmetros gamma(em anexo B.1.5), nota-se que as densidades da *posteriori* se assemelham com a distribuição beta.

A fim de apresentar a qualidade do ajuste do modelo é apresentado a seguir gráficos de FAC dos erros de previsão.

Analisando os gráficos da FAC dos erros (em anexo B.1.6), nota-se que as séries históricas aqui analisadas são descorrelatadas, ou seja, não apresentam auto-correlação significativa.

A partir então, foram feitas a previsão para coortes aqui utilizadas até 2025.

6.1.1- Gráficos da previsão para concluintes de Ensino Fundamental por coortes até 2025

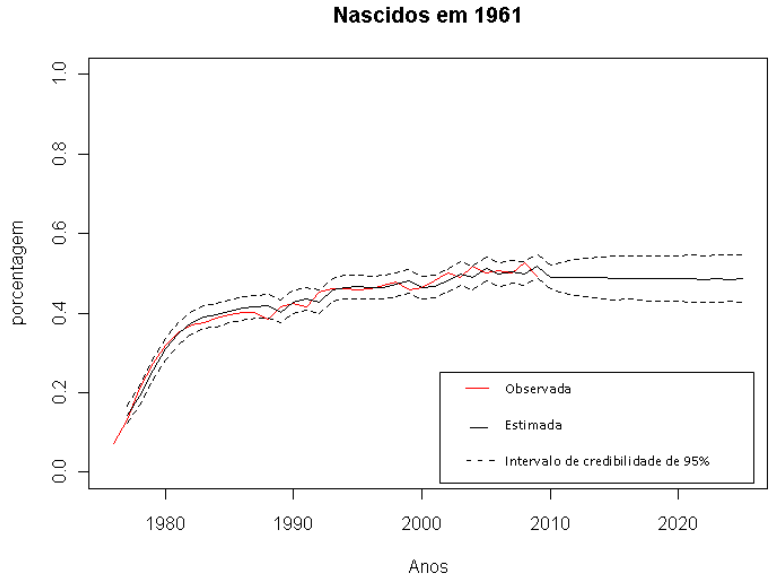


Gráfico 6.1.1.1 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1961.

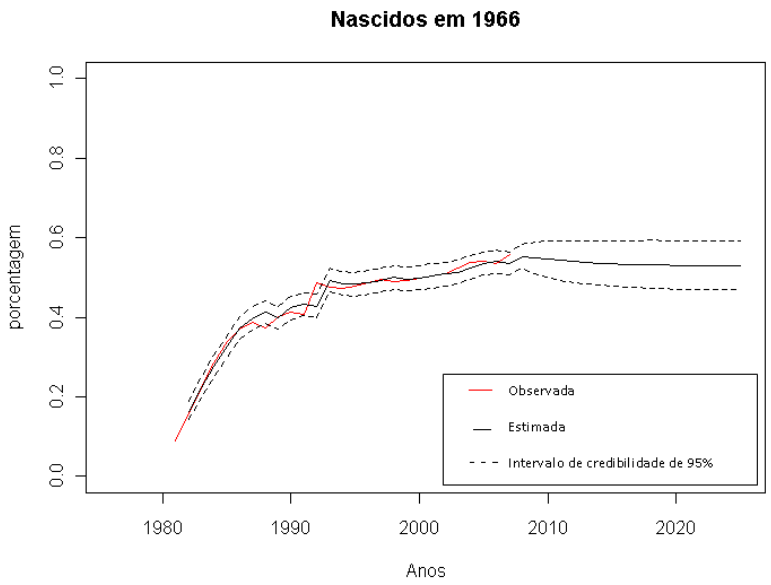


Gráfico 6.1.1.2 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1966.

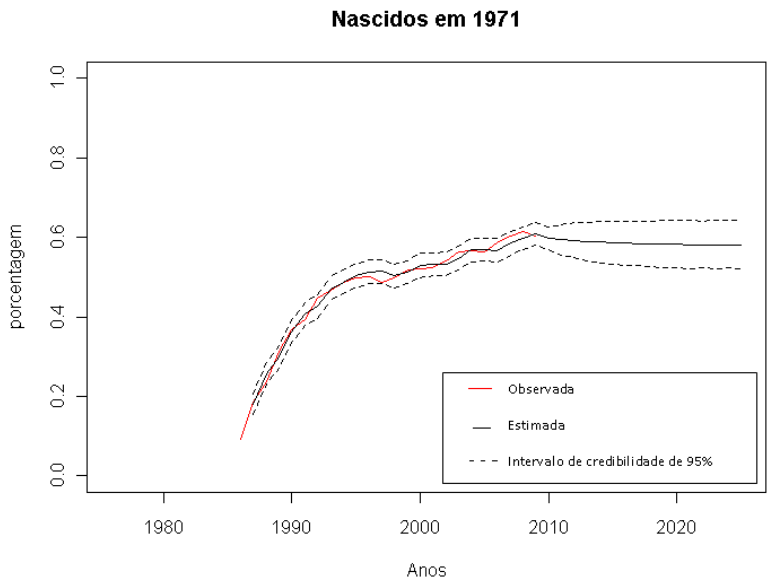


Gráfico 6.1.1.3 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1971.

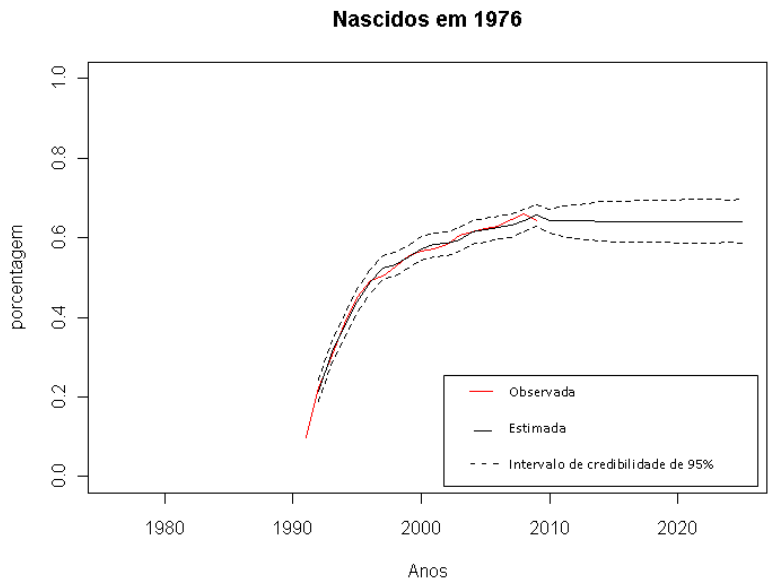


Gráfico 6.1.1.4 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1976.

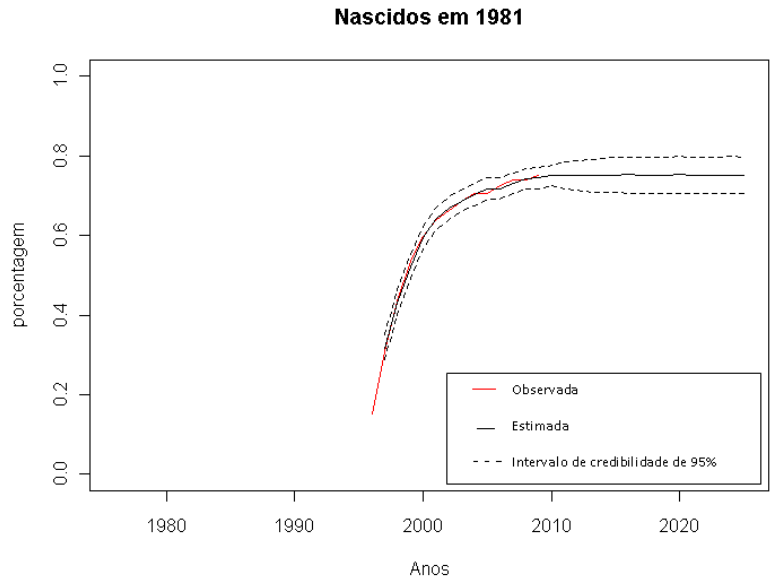


Gráfico 6.1.1.5 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1981.

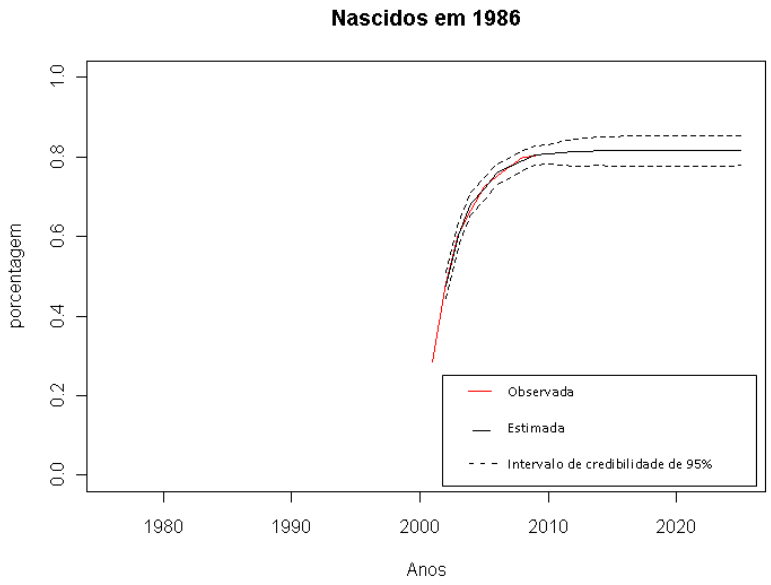


Gráfico 6.1.1.6 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1986.

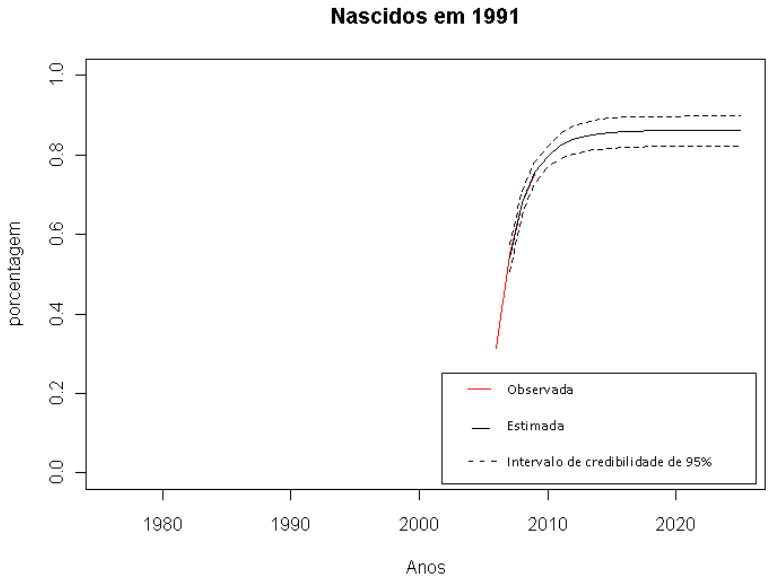


Gráfico 6.1.1.7 – A previsão da série de concluintes do Ensino Fundamental para os nascidos em 1991.

Observando os gráficos acima, nota-se que os valores estimados estão próximos dos valores observados, logo, pode-se interpretar que o modelo proposto prevê bem a série.

6.2 – Ensino Médio

No gráfico de α_0 (valores iniciais, em anexo B.2.1), nota-se que este apresenta uma tendência de subida ao longo das coortes. Isto é, há um crescimento de coeficiente desse parâmetro.

Nota-se que os coeficientes do parâmetro α_1 (parâmetro que define a curvatura da série, em anexo B.2.2) é praticamente constante ao longo das coortes.

Observando o gráfico da cadeia dos parâmetros (anexo), nota-se que todas elas não apresentam nenhuma tendência, portanto, parecem convergir.

A partir dos gráficos da FAC (anexo), observa-se que nenhum deles apresenta autocorrelação, isto é, foram encontrados os posteriori.

Observando os gráficos dos parâmetros gamma (anexo), nota-se que as densidades da posteriori se assemelham a distribuição beta.

Analisando os gráficos da FAC dos erros (anexo), nota-se que as séries históricas aqui analisadas são descorrelatadas, ou seja, não apresentam autocorrelação.

A partir então, foram feitas a estimação para coortes aqui analisadas.

6.2.1- Gráficos da previsão para concluintes de Ensino Médio por coortes até 2025

Seja a linha vermelha a série histórica observada, preta, a estimação e, as tracejadas, o intervalo de confiança de 95%.

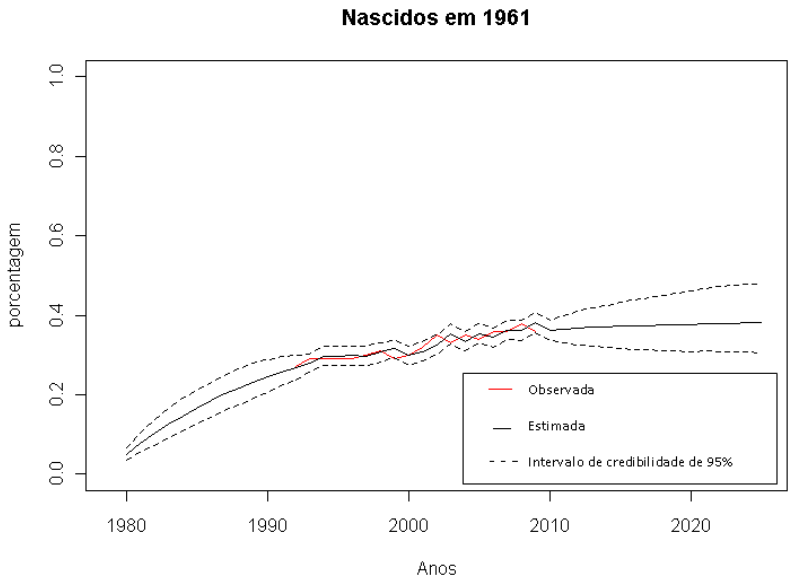


Gráfico 6.2.1.1 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1961.

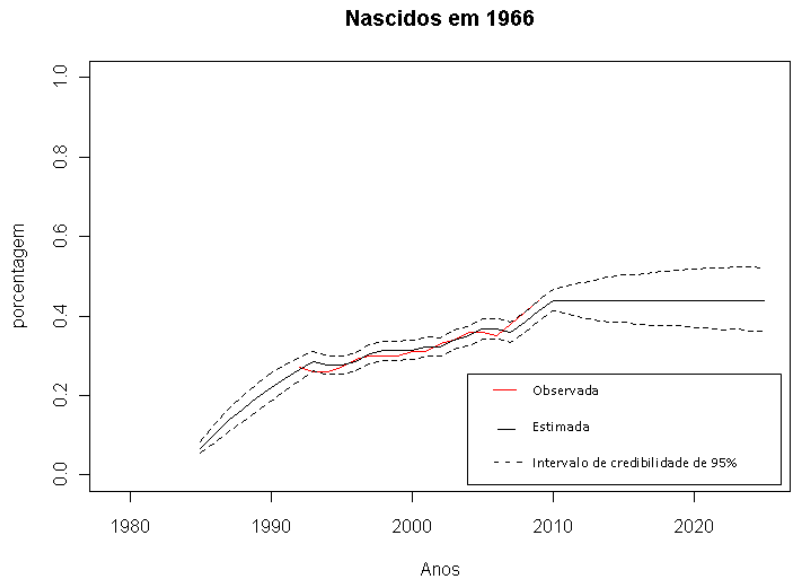


Gráfico 6.2.1.2 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1966.

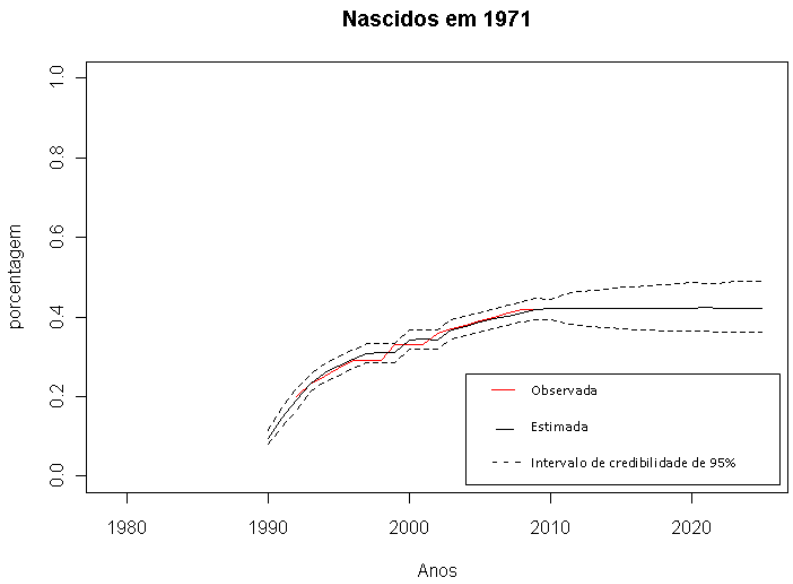


Gráfico 6.2.1.3 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1971.

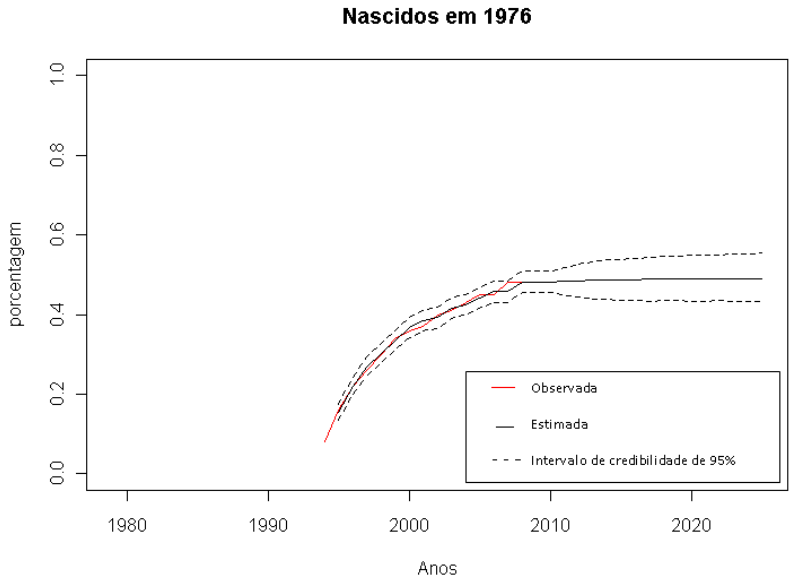


Gráfico 6.2.1.4 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1976.

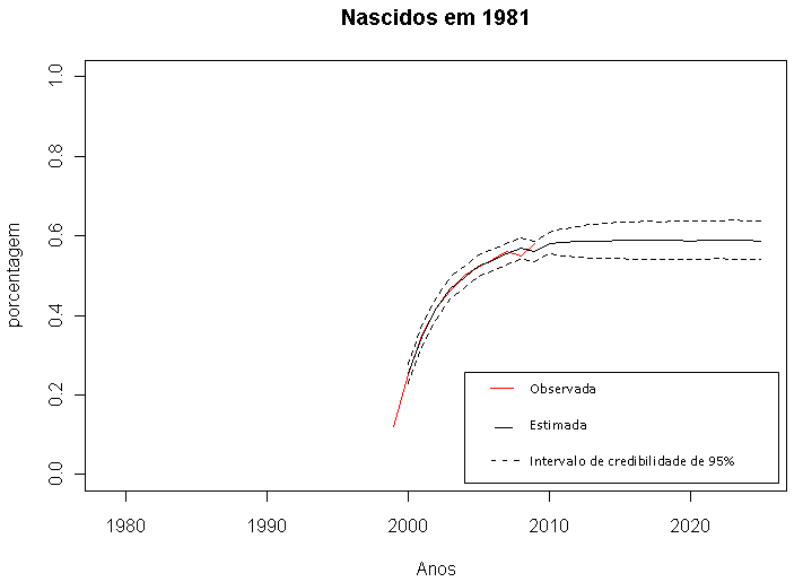


Gráfico 6.2.1.5 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1981.

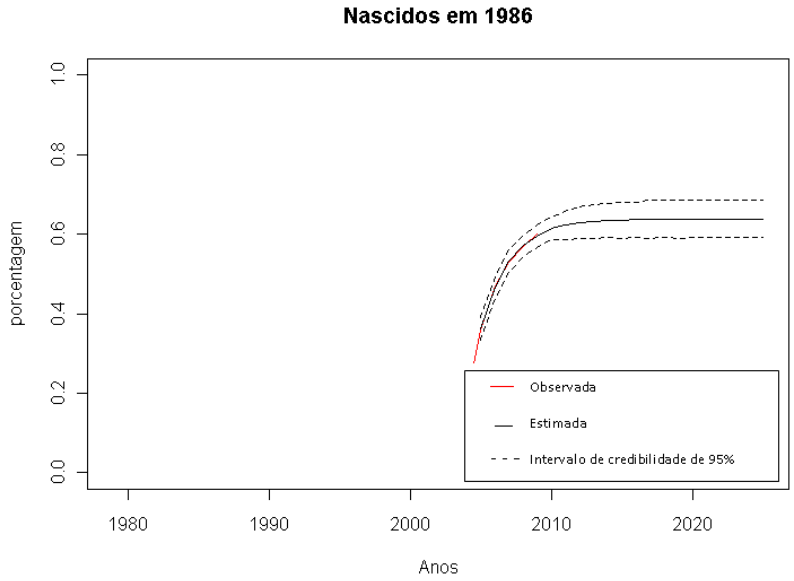


Gráfico 6.2.1.6 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1986.

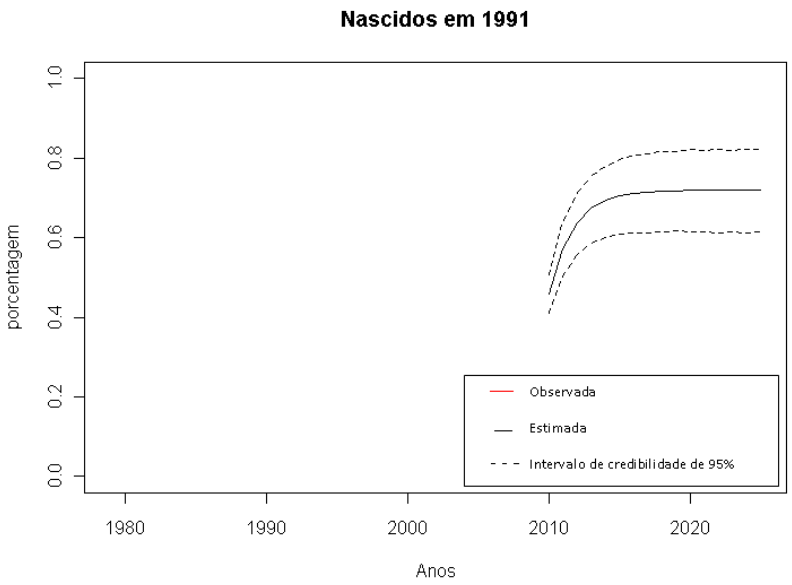


Gráfico 6.2.1.7 – A previsão da série de concluintes do Ensino Médio para os nascidos em 1991.

Observando os gráficos acima, nota-se que os valores estimados estão próximos dos valores observados, logo, pode-se interpretar que o modelo proposto prevê bem a série.

7 - Conclusões e Considerações Finais

Pelos gráficos de parâmetros, de estimação e de previsões, nota-se que o modelo proposto neste trabalho prevê bem as séries históricas da conclusão do Ensino Fundamental e Médio para cada coorte.

Com isso, pretende-se nos futuros trabalhos, aplicar o modelo nas pessoas que nasceram nos últimos anos e nas que vão nascer ainda, com a ajuda da estimação populacional feita pelo IBGE, para prever como ficaria a escolaridade dos brasileiros daqui à 10, 20 ou mais anos.

Referências Bibliográficas

Barbetta, Pedro Alberto. *Estatística Aplicada às Ciências Sociais*, Cap. 3. Ed. UFSC, 5ª Edição, 2002.

Barros, M. Módulo de Regressão e Séries Temporais. 2007.

Buckley, J. (2003), 'Estimation of models with beta distributed dependent variables: A replication and extension of Paolino (2001)', *Political Analysis* 11(1), 1–12.

Ehlers, Ricardo S.. *Introdução a Inferência Bayesiana*. Cap. 4, 2003.

Ferrari, S. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* 31, 799–815.

Geman, S. and Geman, D..*Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*. IEEE Trans. Pat. Anal.Mach. Intel. vol. 6, pp. 721-741. 1984.

Hastings, W. K. (1970), Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, vol. 57, pp. 97-109.

Johnson, N. L.; Kotz, S.; Balakrishnan, N. *Continuous Univariate Distributions*, 2nd ed.,vol. 2. New York: Wiley, 1995.

Kieschnick, R. & McCullough, B. (2003), Regression analysis of variates observed in (0,1): Percentages, proportions and fractions, *Statistical Modelling* 3, 193–213.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, vol. 21, pp. 1087-1092.

Migon, H.S.; da-Silva, C.Q.; Correia, L.T..(2011), 'Dynamic Bayesian beta models', Computational Statistics and Data Analysis 55 (2011) 2074–2089.

MIYASHIRO, E. S. Modelos de Regressão Beta e Simplex para Análise de Proporções. Dissertação (Mestrado) — Instituto de Matemática e Estatística - USP, 2008.

Morettin, Pedro A. & Toloi, Clélia M. Séries Temporais. – 2ª Ed – São Paulo: Atual, 1987.

Neto, Eufrásio de Andrade Lima & Loureiro, Juliana. Mineração de Séries Temporais e Dados Seqüenciais. 2003.

Paolino, P. (2001), Maximum likelihood estimation of models with beta-distributed dependente variables, *Political Analysis,* 9, 325–346.

PNAD. Acesso a Transferências de Renda de Programas Sociais. IBGE, 2006.

Souza, Reinaldo Castro & Camargo, Maria Emília. Análise e previsão de séries temporais: os modelos ARIMA. Ijuí:SEDEGRAF, 1996.

Souza, Reinaldo Castro; Camargo, Alexandre Zanini; Miranda, Cristina Vidigal C.. Introdução a Teoria de Previsão, 2006.

Anexos

A) Rotinas

- A rotina usada no *software* computacional WINBUGS para obter a amostra de posteriori para as séries históricas dos concluintes do Ensino Fundamental.

A rotina a seguir é utilizada também para as séries históricas dos concluintes do Ensino Médio, o que mudou na sintaxe foram apenas os valores dos dados.

```
Model{ #início do modelo
        #hiperparâmetros de alpha0[j]
            gamma00 ~ dbeta(1,1)
            gamma01 ~ dbeta(1,1)
                H0 ~ dgamma(0.01,0.01)
        #hiperparâmetros de alpha1[j]
            gamma10 ~ dbeta(5,17)
            gamma11 ~ dbeta(5,17)
                H1 ~ dgamma(0.01,0.01)
        #parâmetro geral de fi
                phi~ dgamma(0.01,0.01)
        #parâmetros alpha da coluna 1
        alpha0[1] <- 0.08
        alpha1[1] <- 0.9
#Laço em j
        for (j in 2 : K) { #início do Laço em j
```

```
#Bloco 1
mu.alpha0[j] <- gamma00+gamma01*(1-gamma00)*alpha0[j-1]
a.alpha0[j] <- mu.alpha0[j]*H0
b.alpha0[j] <- H0*(1-mu.alpha0[j])
alpha0[j] ~ dbeta(a.alpha0[j],b.alpha0[j])


# Bloco 2
mu.alpha1[j] <- gamma10+gamma11*(1-gamma10)*alpha1[j-1]
a.alpha1[j] <- mu.alpha1[j]*H1
b.alpha1[j] <- H1*(1-mu.alpha1[j])
alpha1[j] ~ dbeta(a.alpha1[j],b.alpha1[j])
                    for (i in j+1 : N) { #início do Laço em i
# Bloco 3 laço em i
mu[i,j] <- alpha0[j]+(1-alpha0[j])*alpha1[j]*Y[i-1,j]
a[i,j] <- mu[i,j]*phi
  b[i,j] <- phi*(1-mu[i,j])
  Y[i,j] ~ dbeta(a[i,j],b[i,j])
repY[i,j] ~ dbeta(a[i,j],b[i,j])


                                                } #fim do Laço em i
                                                } #fim do Laço em j
          } #fim do modelo


#dados
list(
K = 33,
N = 34,
Y=structure(
```

.Data=c(.227,.073,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.177,.131,.065,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.191,.211,.151,.075,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.2,.273,.223,.172,.095,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.2072,.3197,.287,.2588,.1752,.101,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.211,.352,.333,.295,.242,.175,.089,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.21,.371,.351,.331,.288,.239,.158,.088,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.219,.376,.375,.358,.337,.292,.224,.156,.073,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.222,.388,.377,.378,.361,.337,.285,.228,.158,.072,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.228,.397,.399,.388,.408,.377,.34,.302,.244,.18,.101,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.231,.402,.391,.408,.397,.408,.37,.352,.285,.237,.167,.092,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.235,.403,.394,.407,.399,.406,.388,.371,.34,.293,.233,.186,.088,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.224,.385,.396,.397,.405,.399,.372,.392,.353,.326,.283,.238,.166,.085,NA,NA,NA,NA,NA,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.239,.417,.432,.423,.426,.446,.401,.41,.395,.369,.347,.314,.261,.181,.092,NA,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,NA,NA,

.242,.425,.424,.438,.431,.439,.414,.427,.415,.423,.378,.368,.317,.249,.178,.087,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.2445,.4168,.412,.4265,.4277,.4275,.4071,.4198,.4153,.42,.3953,.3935,.347,.312,.253,.179,.098,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.246,.452,.455,.475,.508,.492,.486,.48,.473,.463,.458,.446,.412,.397,.347,.293,.224,.118,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.246,.461,.473,.471,.512,.502,.475,.494,.477,.471,.455,.467,.456,.429,.4,.375,.306,.232,.117,NA,NA,N
A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.2479,.4616,.4817,.476099999999999,.5094,.5084,.473699999999999,.5028,.4807,.482099999999999,.
463799999999998,.486,.478600000000003,.460199999999999,.437300000000002,.438700000000001,.
383999999999999,.331000000000001,.235500000000001,.125499999999999,NA,NA,NA,NA,NA,NA
,NA,NA,NA,NA,NA,NA,NA,
.2534,.4594,.4828,.4819,.5016,.5091,.4788,.5057,.4838,.4899,.476200000000001,.4985,.487899999999
999,.4843,.464199999999999,.481799999999999,.448999999999999,.406000000000002,.33816666666
6668,.284875000000012,.139,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.264,.46,.478,.48,.49,.502,.487,.502,.486,.488,.484,.5,.492,.495,.486,.502,.492,.457,.425,.375,.282,.153
,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.266,.469,.469,.462,.476,.485,.495,.491,.487,.47,.479,.486,.499,.486,.508,.497,.504,.484,.496,.44,.384,.
308,.163,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.274,.479,.485,.489,.498,.488,.489,.513,.482,.492,.502,.497,.516,.5,.537,.521,.527,.543,.532,.524,.494,.
435,.328,.182,NA,NA,NA,NA,NA,NA,NA,NA,NA,
.272,.459,.478,.489,.508,.501,.494,.509,.489,.487,.496,.519,.51,.526,.539,.545,.555,.552,.582,.561,.573
,.535,.465,.359,.193,NA,NA,NA,NA,NA,NA,NA,NA,
.2788,.463333333333334,.475333333333333,.487,.510833333333333,.512833333333333,.4993333333
33333,.508,.491,.489,.499,.521833333333333,.517166666666667,.5275,.545833333333333,.546833333
333333,.565666666666667,.570333333333333,.593333333333333,.5885,.62,.5985,.565333333333333,.
497166666666666,.401666666666669,.234666666666666,NA,NA,NA,NA,NA,NA,NA,
.287,.481,.482,.49,.51,.525,.505,.508,.503,.5,.511,.523,.531,.528,.555,.549,.571,.586,.594,.608,.641,.63
7,.628,.599,.547,.448,.285,NA,NA,NA,NA,NA,NA,

```
.289,.501,.503,.505,.509,.539,.511,.507,.54,.522,.532,.54,.545,.551,.564,.574,.583,.587,.612,.621,.642,.
662,.652,.667,.629,.566,.484,.322,NA,NA,NA,NA,NA,
.279,.49,.505,.514,.531,.536,.524,.534,.524,.531,.541,.565,.565,.565,.579,.577,.606,.614,.625,.659,.678
,.685,.695,.687,.695,.658,.607,.52,.349,NA,NA,NA,NA,
.302,.518,.497,.52,.538,.54,.537,.546,.548,.539,.562,.567,.559,.572,.579,.583,.614,.619,.649,.657,.679,.
705,.704,.706,.717,.706,.666,.612,.488,.291,NA,NA,NA,
.305,.5,.509,.515,.528,.541,.542,.537,.545,.546,.541,.564,.568,.566,.574,.584,.623,.625,.653,.663,.698,.
705,.727,.721,.737,.728,.726,.686,.615,.522,.284,NA,NA,
.315,.508,.526,.529,.539,.554,.536,.552,.559,.556,.56,.587,.575,.584,.6,.608,.628,.643,.66,.674,.704,.72
5,.726,.746,.752,.757,.752,.752,.707,.639,.521,.313,NA,
.325,.502,.521,.535,.546,.569,.559,.576,.586,.566,.603,.602,.609,.609,.6,.619,.645,.644,.68,.694,.713,.7
4,.737,.759,.761,.761,.776,.778,.752,.741,.659,.549,.342,
.333,.526,.533,.539,.586,.628,.61,.625,.577,.603,.594,.616,.627,.615,.631,.619,.66,.667,.687,.718,.73,.7
41,.758,.765,.781,.786,.798,.799,.799,.79,.75,.68,.556,
.331,.492,.515,.528,.576,.619,.624,.65,.659,.582,.587,.602,.606,.617,.625,.625,.642,.648,.689,.703,.739
,.751,.752,.778,.785,.79,.801,.812,.812,.808,.791,.752,.686),
.Dim=c(34,33)
)
)
#condições iniciais1
list(
alpha0=c(NA, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5,
0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5,
0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5),
alpha1=c(NA, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5,
0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5,
0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5),
gamma00=0.5,
gamma01=0.5,
```

```
gamma10=0.5,

gamma11=0.5,

H0=1,

H1=1,

phi=1

)
```

B) Gráficos da convergência

B.1) Ensino Fundamental

- Parâmetro α_0

O gráfico abaixo mostra a evolução dos valores $\alpha_0$para as diferentes coortes.

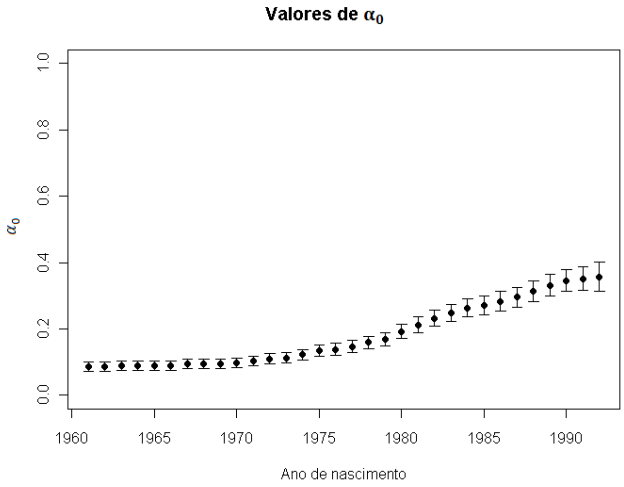


Gráfico B.1.1 – Valores do parâmetro α_0

- Parâmetro α_1

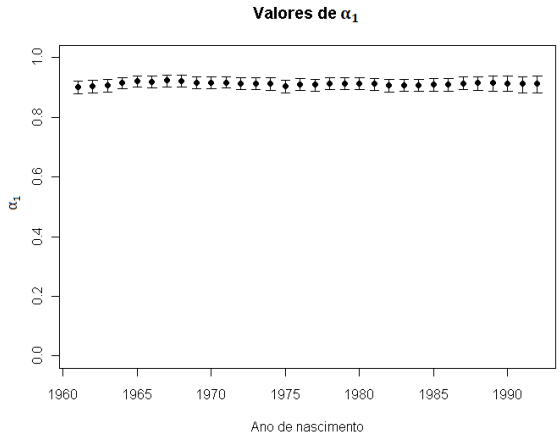
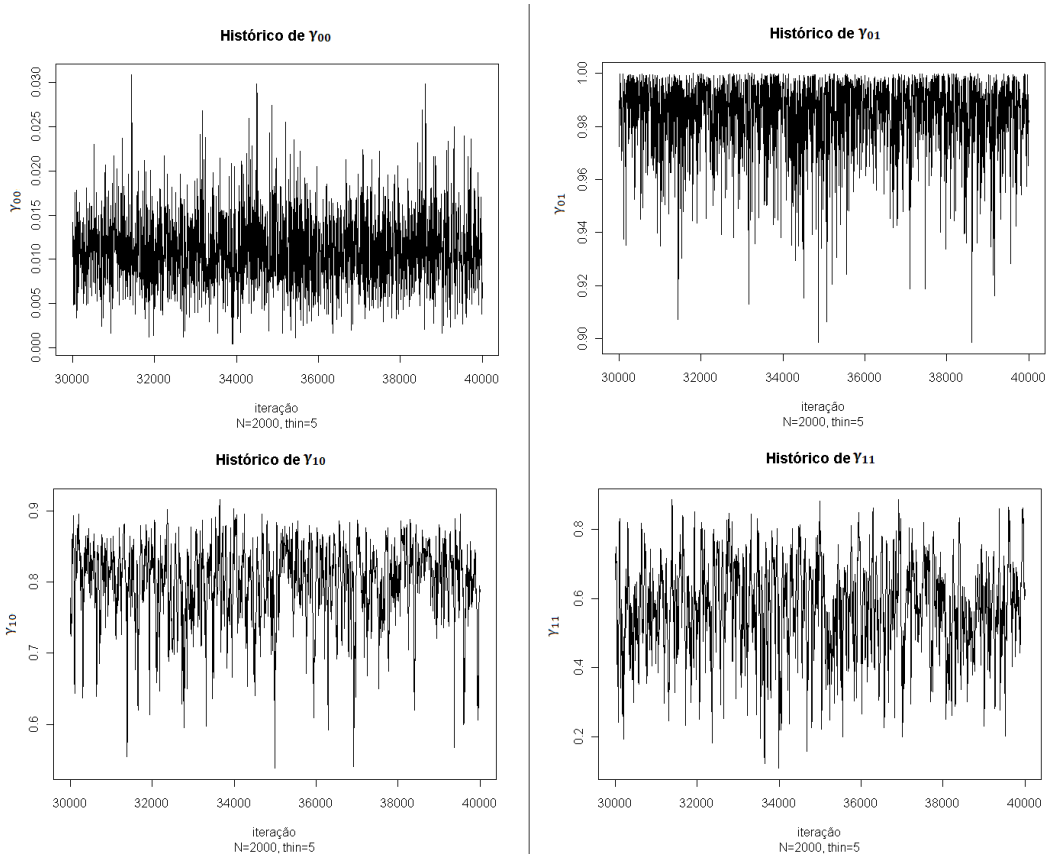
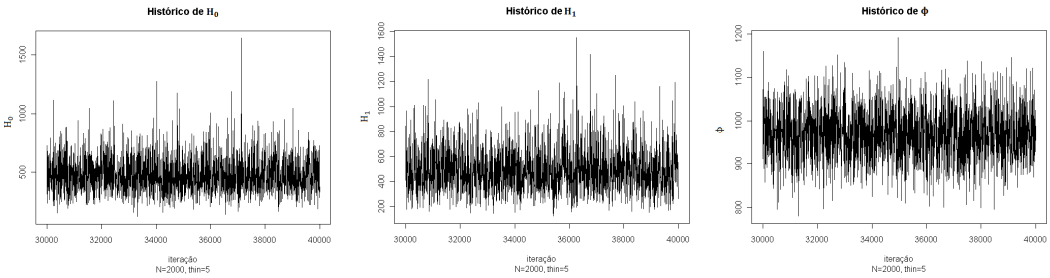


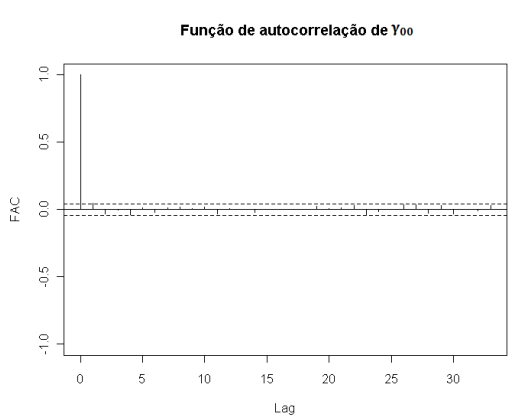
Gráfico B.1.2 – Valores do parâmetro α_1

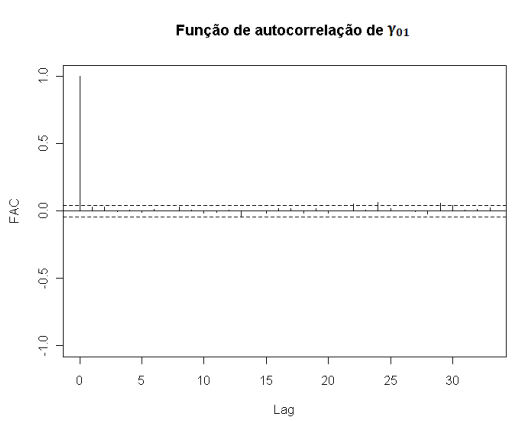
B.1.3 - Gráficos das cadeias dos parâmetros γ_{00}, γ_{01}, γ_{10}, γ_{11}, H_0, H_1 e ϕ

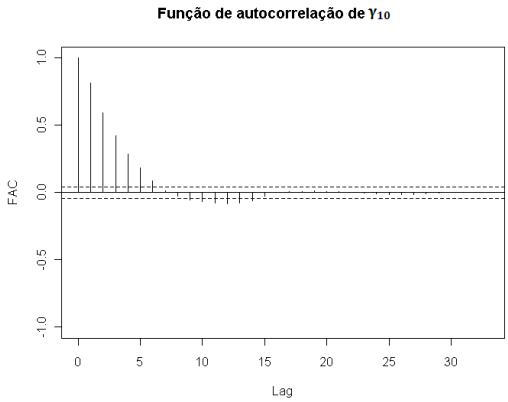


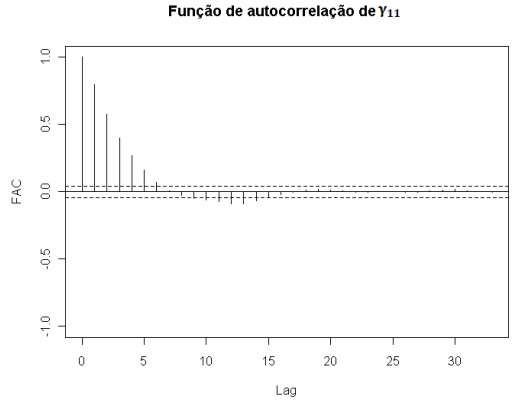


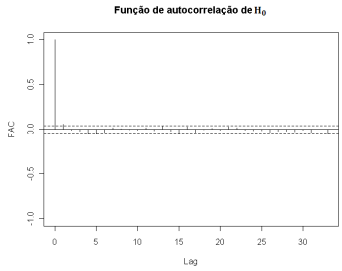
B.1.4 - Gráficos de Função de Auto Correlação (FAC) dos parâmetros γ_{00}, γ_{01}, γ_{10}, γ_{11}, H_0, H_1 e ϕ

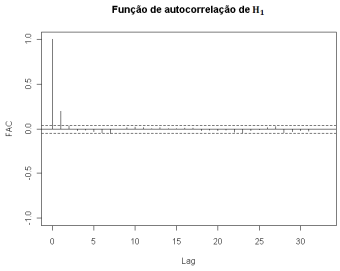


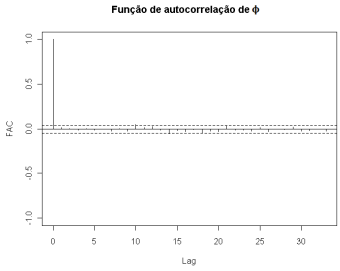




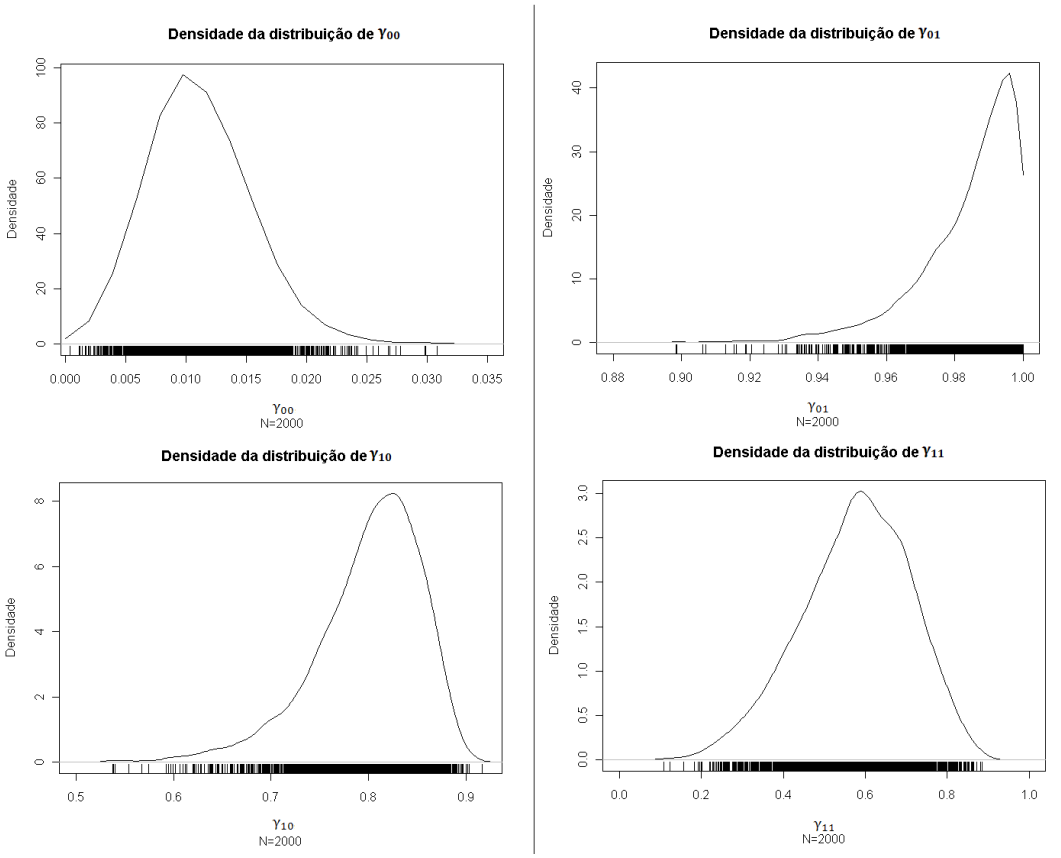


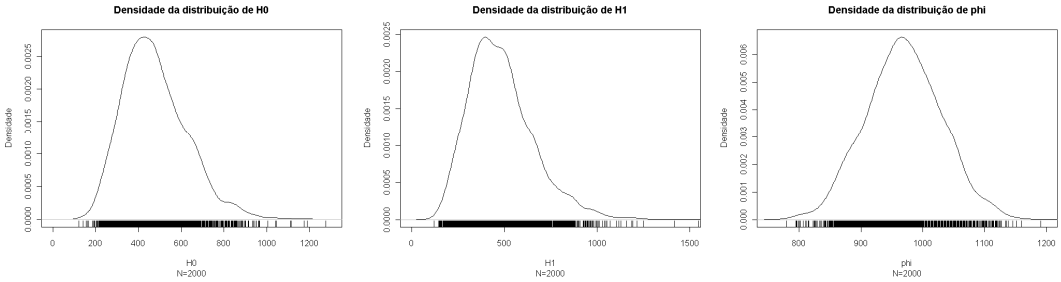






B.1.5 - Gráficos de densidade da *posteriori* γ_{00}, γ_{01}, γ_{10}, γ_{11}, H_0, H_1 e ϕ





B.1.6 - Gráficos de Função de Auto Correlação (FAC) dos erros de previsão de algumas séries históricas

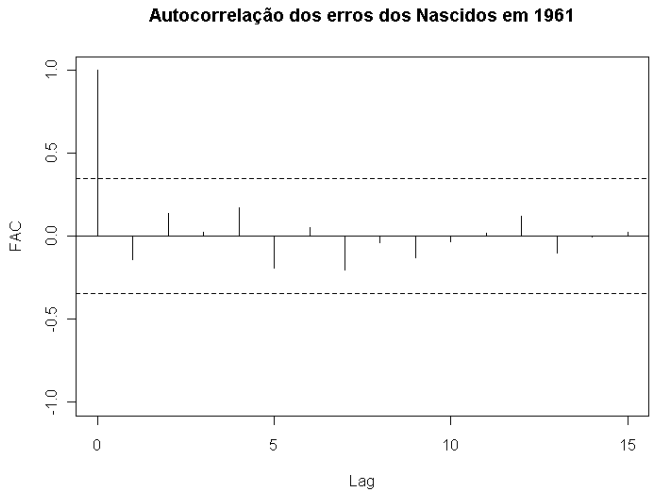


Gráfico B.1.6.1 – FAC dos erros da série “concluintes dos nascidos em 1961”.

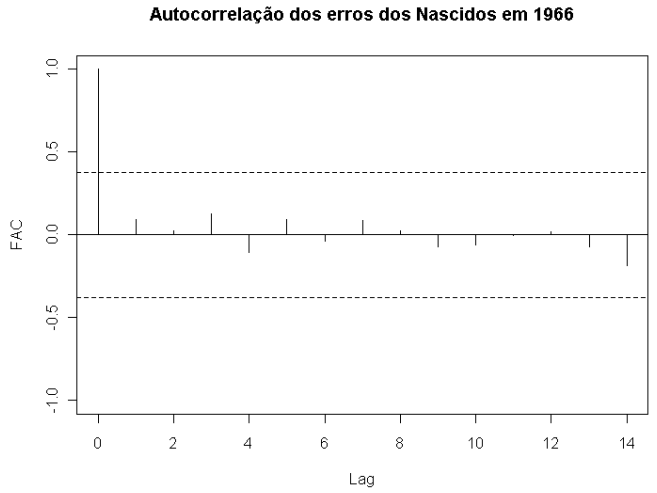


Gráfico B.1.6.2 – FAC dos erros da série “concluintes dos nascidos em 1966”.

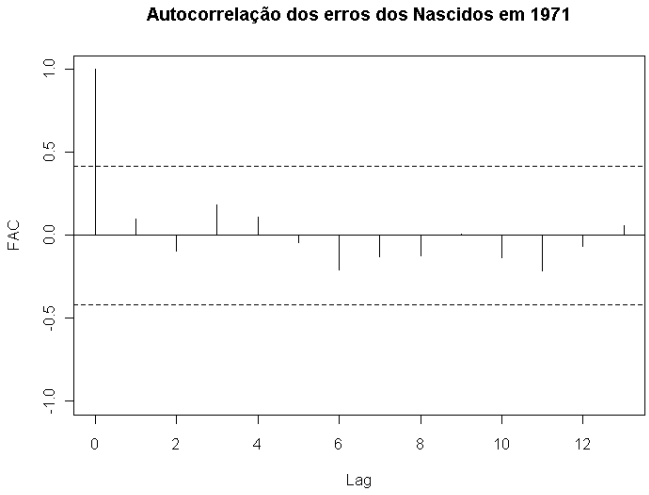


Gráfico B.1.6.3 – FAC dos erros da série “concluintes dos nascidos em 1971”.

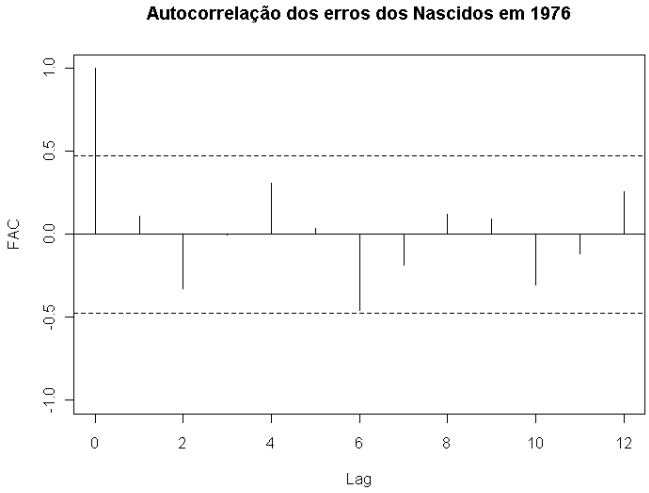


Gráfico B.1.6.4 – FAC dos erros da série “concluintes dos nascidos em 1976”.

Gráfico B.1.6.5 – FAC dos erros da série “concluintes dos nascidos em 1981”.

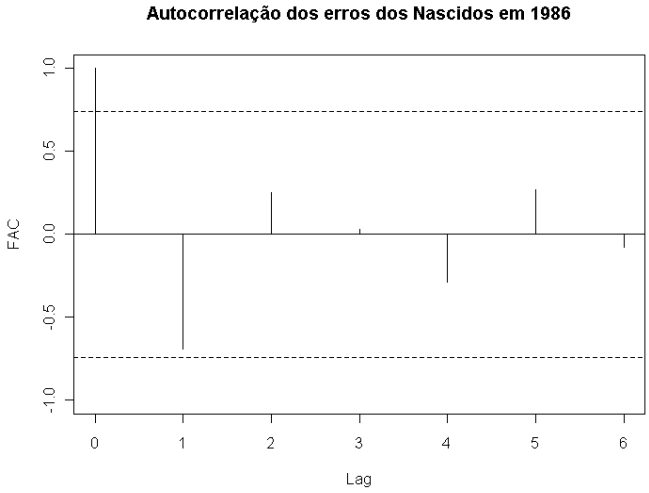


Gráfico B.1.6.6 – FAC dos erros da série “concluintes dos nascidos em 1986”.

Gráfico B.1.6.7 – FAC dos erros da série “concluintes dos nascidos em 1990”.

B.2) Ensino Médio

- Parâmetro α_0

O gráfico abaixo mostra a evolução dos valores $\alpha_0$para as diferentes coortes.

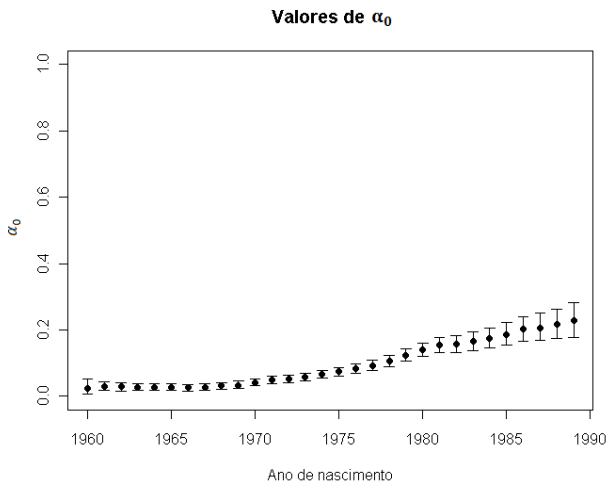


Gráfico B.2.1 – Valores do parâmetro α_0

- Parâmetro α_1

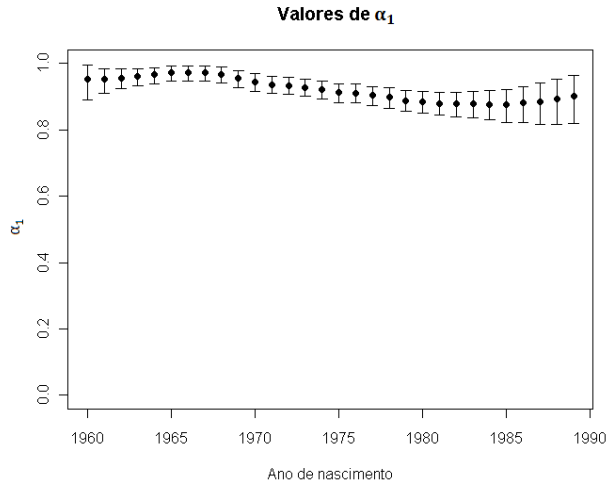
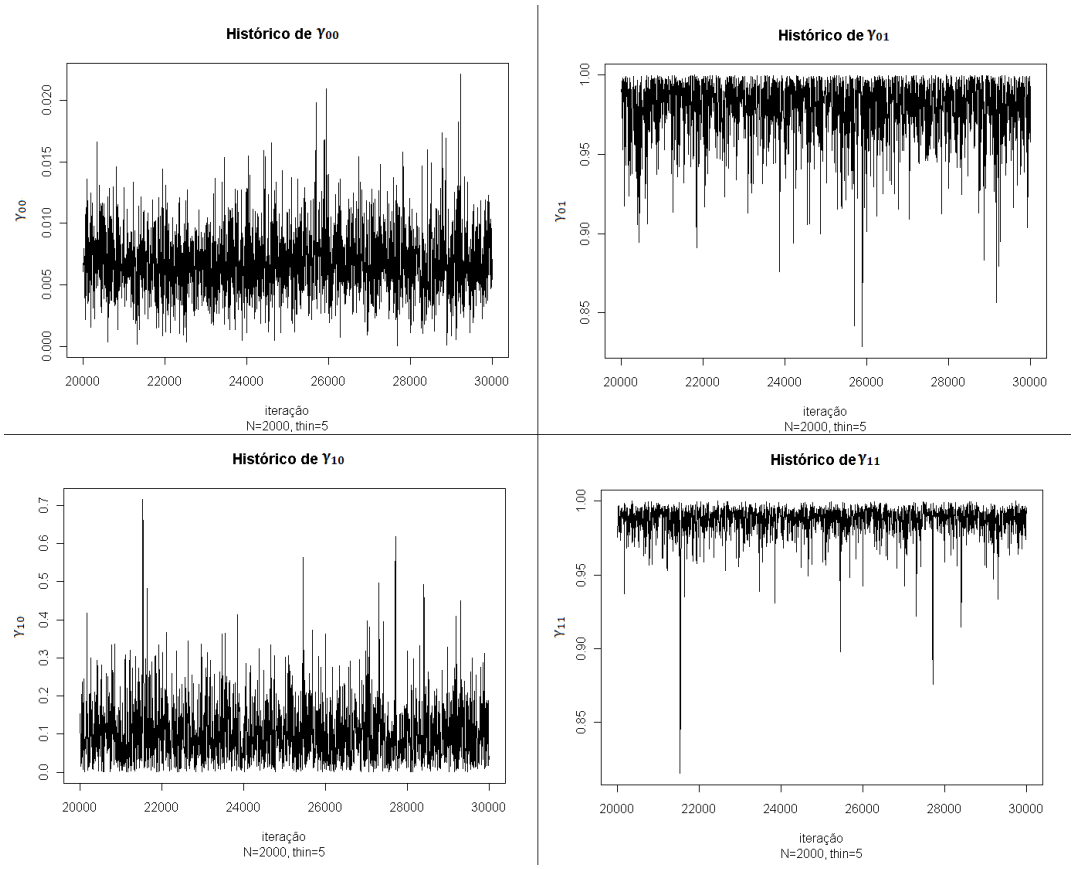
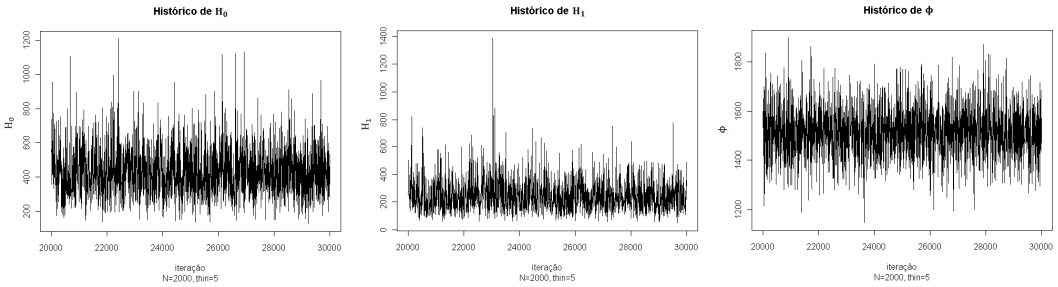


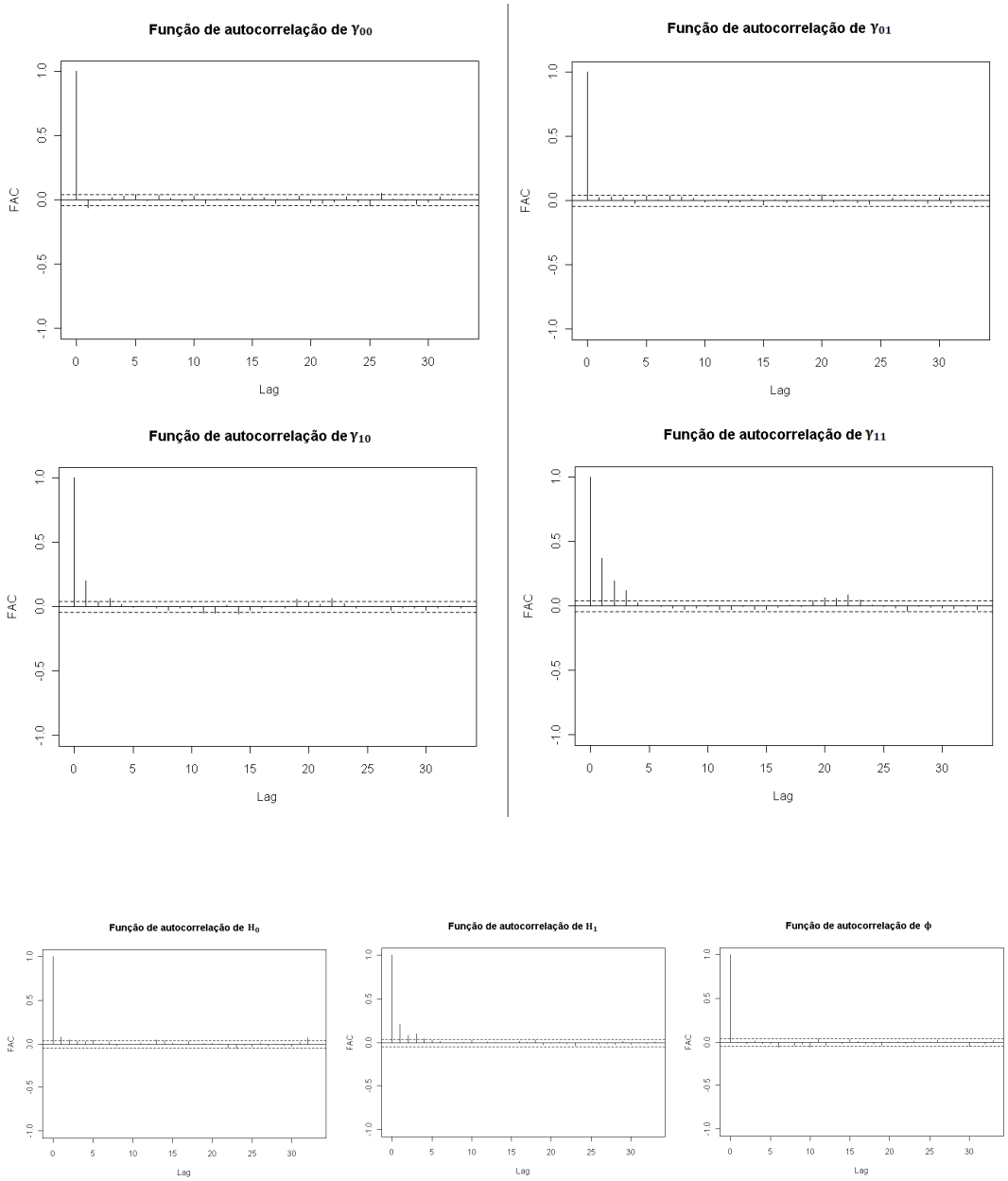
Gráfico B.2.2 – Valores do parâmetro α_1

B.2.3 - Gráfico das cadeias dos parâmetros γ_{00}, γ_{01}, γ_{10}, γ_{11}, H_0, H_1 e ϕ

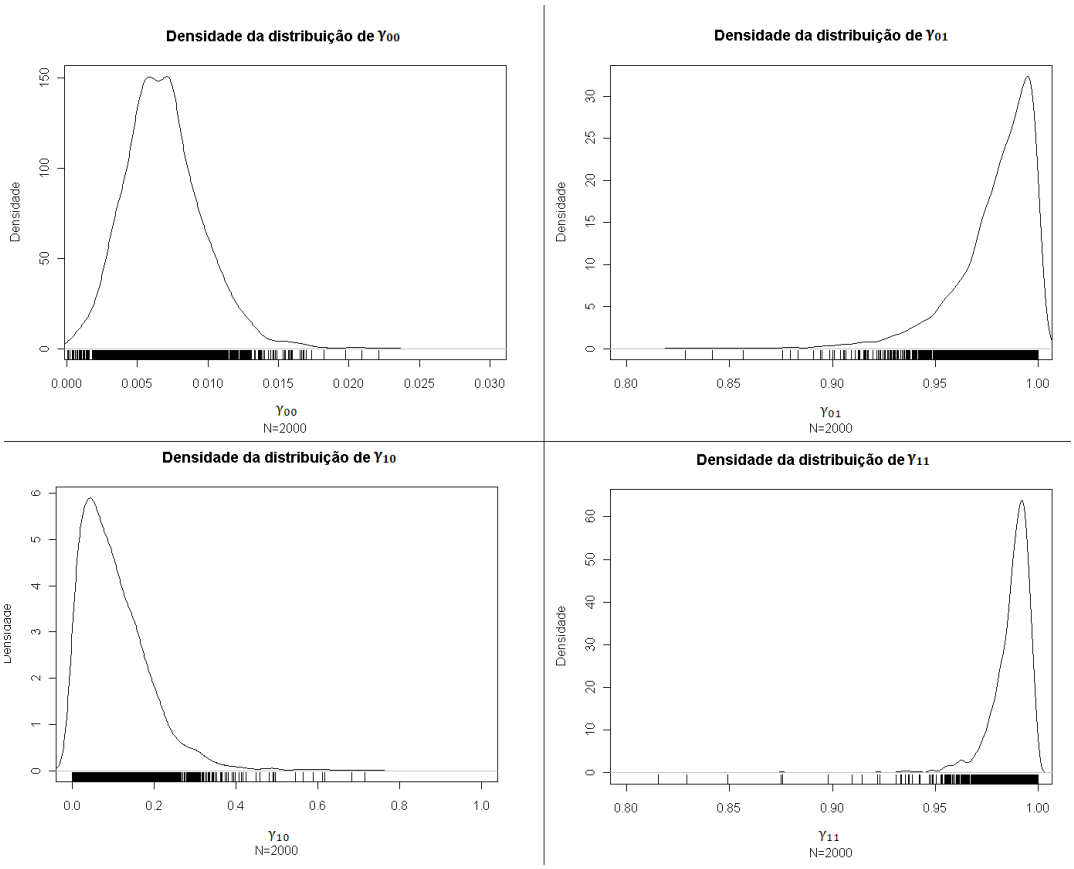


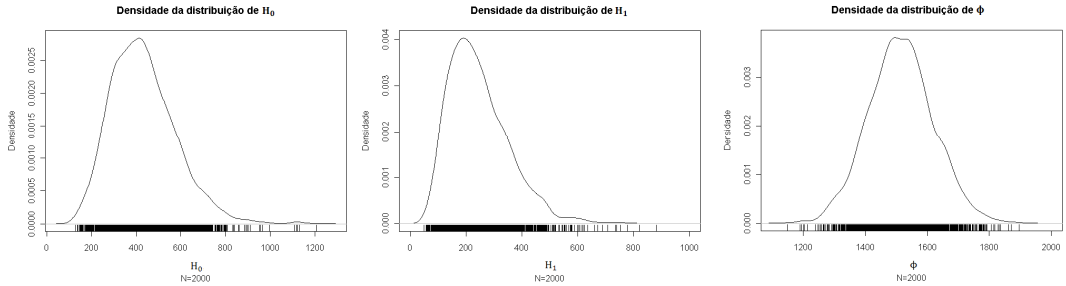


B.2.4 - Gráfico de Função de Auto Correlação (FAC) dos parâmetros γ_{00}, γ_{01}, γ_{10}, γ_{11}, H_0, H_1 e ϕ



B.2.5 - Gráfico de densidade dos parâmetros γ_{00}, γ_{01}, γ_{10}, γ_{11}, H_0, H_1 e ϕ





B.2.6 - Gráfico de Função de Auto Correlação (FAC) dos erros de previsão de algumas séries históricas

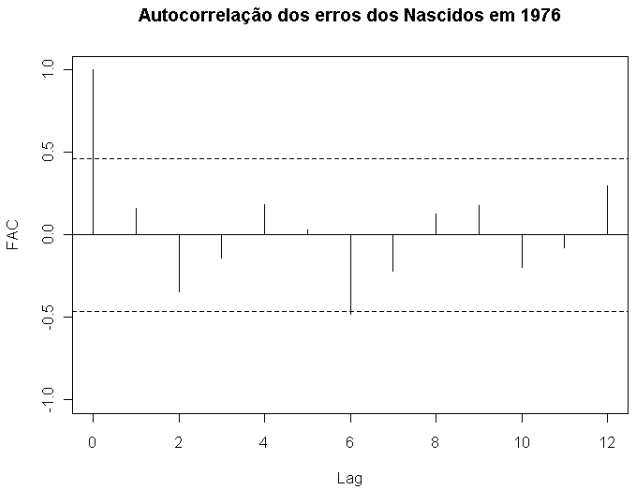


Gráfico B.2.6.1 – FAC dos erros da série “concluintes dos nascidos em 1976”.

Gráfico B.2.6.2 – FAC dos erros da série “concluintes dos nascidos em 1981”.

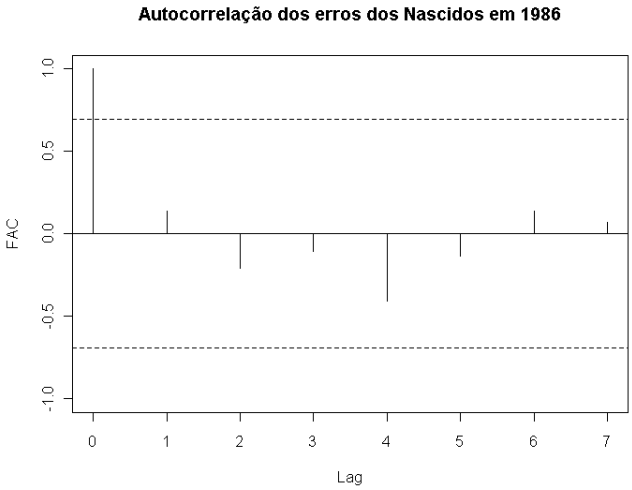


Gráfico B.2.6.3 – FAC dos erros da série “concluintes dos nascidos em 1986”.

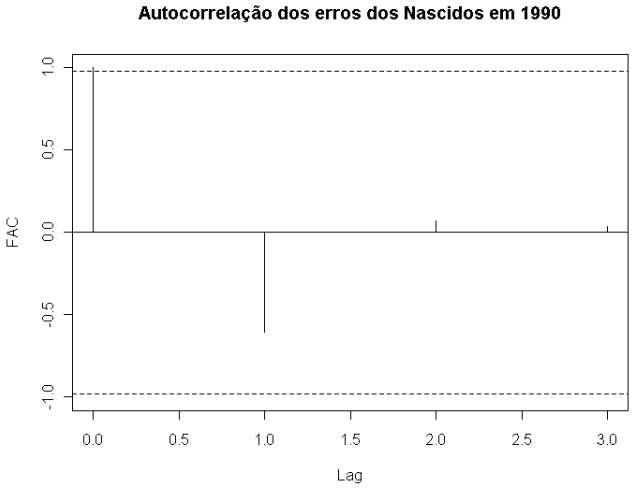


Gráfico B.2.6.4 – FAC dos erros da série “concluintes dos nascidos em 1990”.

C) Soma da progressão geométrica

$$S_n = a_0p^0 + a_0p + a_0p^2 + \cdots + a_0p^n$$

$$\Downarrow$$

$$S_n = a_0 + p(a_0 + a_0p + \cdots + a_0p^{n-1})$$

$$S_n = a_0 + pS_{n-1}$$

$$S_{n-1} = \frac{S_n - a_0}{p}$$

$$S_n = \frac{S_n - a_0}{p} + a_0p^n$$

$$pS_n - S_n = a_0p^{n+1} - a_0$$

$$S_n = \frac{(p^{n+1} - 1)}{p - 1}a_0 = \frac{(1 - p^{n+1})}{1 - p}a_0$$