Escolha de Modelos Para Predição de AIDS

Basílio B. Pereira e Helio S.Migon

UFRJ

Summary

In this paper tests of separated families of hypothesis and information critirias are used to choose among growth models for the number of diagnosed cases of AIDS in Brazil.

The forecast profiles obtained from the models also illustrates the well know problem of unreliable data for more recent periods.

Sumário

Neste artigo diferentes testes de famílias separadas de hipóteses e critérios de informação são usados para escolher entre modelos de crescimento do número de casos diagnosticados de AIDS no Brasil. As trajetórias das previsões obtidas com os modelos selecionados ilustram o problema da imprecisão dos dados de diagnóstico nos periodos mais recentes, devido aos atrasos inerentes ao processo de notificação.

1 - Introdução

Os primeiros casos de AIDS no mundo foram diagnosticados no início dos anos 80 e em torno de 1983 o virus tipo I (HIV1) foi identificado pela primeira vez. No Brasil os primeiros casos foram notificados em 82 e hoje já se tem um total de 12405 (junho de 1990) casos diagnosticados. De setembro de 1985 para cá o crescimento do número de casos notificados foi da ordem de 30 vezes.

O sistema compulsório de notificações do Ministério da Saude-DST/AIDS registra os casos segundo a data do diagnóstico e também da notificação. Alguns autores preferem usar as contagens de casos diagnosticados ao invés de casos notificados, pois o padrão internacional da doença baseia-se nos casos diagnosticados. A desvantagem de usar este tipo de dado é que exige manipulações extras para cuidar do atraso na notificação. Diferentes estudos mostram que estes atrasos afetam o número de casos até aproximadamente 18 meses atrás. Apesar das críticas, neste artigo os modelos são calibrados com base no número de casos diagnosticados e serão testadas diferentes especificações de curvas de crescimento.

São várias as razões para se estudar a evolução da epidemia de AIDS. A projeção do número de casos de AIDS é essencial para se desenvolver políticas de saúde pública no sentido de controlar a evolução do vírus e até, quem sabe, para erradicá-lo. Muitos esforços tem sido desenvolvidos no sentido de compreender a doença e descrever sua evolução. Os números especiais do *Journal of the Royal Statistical Socity - série B - 1988 e do Journal of Medical Statistics - 1989* evidenciam a importância que o assunto vem merecendo.

Na seção 2 são apresentados diferentes curvas de crescimento. Observe-se que cada especificação inclue as curvas logística, Gompertz e exponencial modificado como casos particulares para valores especiais de seus parâmetros (Ord e Young, 1988). Um resumo de testes de hipóteses separadas é apresentado na Seção 3, ficando os resultados, obtidos usando-se o pacote DatFit (Pesaran e Pesaran, 1987), para serem discutidos na Seção 4. Os resultados descritos neste artigo são preliminares e sofrem de várias deficiências. Os dados de diagnóstico introduzem novas fontes de variabilidade em razão do atraso no processo de notificação sendo portanto preferível trabalhar-se com os casos notificados. Por outro lado, as transformações lineares dos modelos de crescimento impedem a estimação do parâmetro que caracteriza o valor assintótico do processo de crescimento. As conclusões obtidas neste artigo são similares àquelas de Migon e Gamerman (1990).

2 - Modelos de Crescimento

Dado uma variável estoque Y_t, o fluxo ou aumento líquido será denotado por $y_t = \frac{dY}{dt} \sim \Delta Y_t = Y_t - Y_{t-1}$. Os modelos a serem considerados são:

$$\log \frac{d}{dt} Y_t = \log y_t = \beta_0 + \beta_1 t + \beta_2 \log Y_t \tag{1}$$

$$\frac{d}{dt} \log \left(\frac{d}{dt} Y_t \right) = \Delta \log y_t = \beta_0 + \beta_2 \Delta \log(Y_t), \quad \text{ou mais geralmente}$$

$$\log(y_t) = \beta_1 + \beta_2 \Delta \log(Y_t) + \beta_3 \log(y_{t-1}) \tag{2}$$

$$\frac{d}{dt} \log(Y_t) \simeq \frac{y_t}{Y_t} = \beta_0 + \beta_1 Y_t + \beta_2 Y_t^{-1} + \beta_3 log(Y_t), \quad \text{ou seja}$$

$$\log(y_t) = \log(\beta_0 Y_t + \beta_1 Y_t^2 + \beta_2 + \beta_3 Y_t log(Y_t)) \tag{3}$$

Uma aproximação linear para (3) é dada por:

$$\log(y_t) = \gamma_0 + \gamma_1 Y_t + \gamma_2 Y_t^2 + \gamma_3 Y_t \log(Y_t) \tag{4}$$

Segundo o excelente 'review' sobre modelos para curvas de crescimento de Ord e Yóung (1988), (1) e (2) são devidos a Harvey (1984) e (3) a Levenbach e Reuter (1976), extendido por Kendal, Stuart e Ord (Ord e Young, 1988).

3 - Escolha de Modelos

A escolha entre modelos será feita utilizando testes de hipóteses separadas e critérios de informação, bem como as trajetórias das previsões.

3.1 - Testes de Hipóteses separadas

Considere os seguintes modelos de regressão:

$$M_1 : \underline{y} = X\underline{\beta}_1 + \underline{u}_1, \quad \underline{u}_1 \sim N[0, \sigma^2 I_n]$$

$$M_2 : \underline{y} = Z\underline{\beta}_2 + \underline{u}_2, \quad \underline{u}_2 \sim N[0, \omega^2 I_n]$$

onde $\underline{y}$ é um vetor de observações da variável dependente $n \times 1$, X e Z são matrizes $n \times K_1$ e $n \times K_2$ de observações dos regressores dos modelos M_1 e M_2. Os coeficientes das regressões são denotados por β_1 e β_2, vetores de dimensão k_1 e $k_2 \times 1$, respectivamente e $\underline{u}_1$ e $\underline{u}_2$ são vetores $n \times 1$ de erros aleatórios.

Grosseiramente os modelos M_1 e M_2 são ditos não encaixados se os regressores de M_1 (resp. M_2) não puderem ser expressos como combinação dos regressores em M_2 (resp. M_1). Uma revisão da literatura em testes de hipóteses separadas é encontrada em McAleer e Pesaran (1986) e McAleer (1986) e uma bibliografia sobre o assunto em Pereira (1977 e 1981).

O teste utilizado neste artigo foi o teste de Cox modificado (ver McAleer e Pesaran, 1986), o qual é denominado N-test ou NT-test.

3.2 - Critério de Informação

Denote por L_1^* e L_2^* o máximo da função de verossimilhança dos modelos M_1 e M_2, respectivamente. O critério de informação de Akaike (AIC) para escolha entre os modelos M_1 e M_2 é calculado como:

$$\text{AIC}(M_1 : M_2) = L_1^* - L_2^* - [k_1 - k_2]$$

Se AIC($M_1 : M_2$) > 0 o modelo M_1 é preferido ao modelo M_2, caso contrário M_2 é preferível.

4 - Análise dos Dados

Foram realizadas duas análises: a primeira com dados de 11/1985 a 11/1988 (37 observações) e a outra com dados de 11/1985 a 12/1987 (26 observações) e previsões para 1988.

4.1 - Testes de Hipóteses Separadas

Os resultados do teste NT são apresentados na tabelas I e II e devem ser comparados com o valor de uma $N(0,1)$

	M_1	M_2	M_3
M_1	–	−1.27	1.47
M_2	−8.85	–	−9.02
M_3	−2.74	−2.05	–

Tabela I - Teste NT, 26 observações

	M_1	M_2	M_3
M_1	–	−5.94	−2.85
M_2	−1.94	–	−0.65
M_3	−7.35	−9.22	–

Tabela II - teste NT, 37 observações

Da tabela I verificamos que os teste de M_1 (nula) contra M_2 (-1.27) e contra M_3 (1.47) não permitem rejeitar M_1. Os testes de M_2 (nula) contra M_1 (-8.85) e contra M_3 (-9.02) indicam a rejeição de M_2. Os testes de M_3 (nula) contra M_1 (-2,74) e contra M_3 (-2,05) indicam a rejeição de M_3. Portanto com 26 observações M_1 parece ser o modelo mais apropriado.

Da tabela II verificamos que os testes de M_1 (nula) contra M_2 (-5,94) e contra M_3 (-2,85) apontam a rejeição de M_1. Os testes de M_2 (nula) contra M_1 (-1.94) e contra M_3

(-0.65) não permitem rejeitar M_2 (embora -1.94 seja quase significante). Os testes de M_3 (nula) contra M_1 (-7,35) e contra M_2 (-9,22) indicam a rejeição de M_3. Portanto com 37 observações M_2 parece ser o modelo mais apropriado.

4.2 - Critérios de Informação

As tabelas III e IV apresentam os máximos das verossimilhanças e os valores dos critérios AIC e confirmam as escolhas dos testes de hipóteses separadas.

$L_1^* = 27,15$	AIC$(M_1 \times M_2) = 11,63$ favor M_1
$L_2^* = 15,53$	AIC$(M_1 \times M_3) = 4,02$ favor M_1
$L_3^* = 24,77$	AIC$(M_2 \times M_3) = -8,24$ favor M_2

Tabela III - Versossimilhanças e AIC, 26 observações

$L_1^* = 7,32$	AIC$(M_2 \times M_3) = 18,95$ favor M_2
$L_2^* = 15,44$	AIC$(M_2 \times M_1) = 8,12$ favor M_2
$L_3^* = 1,70$	AIC$(M_1 \times M_3) = 10,83$ favor M_1

Tabela IV - Versossimilhanças e AIC, 37 observações

5 - Ajustamentos, Previsões e Conclusões

As figuras I, II e III apresentam os dados e o comportamento dos ajustamentos (37 observações) e dos ajustamentos (26 observações) mais as previsões (11 valores).

A figura Ib apresenta as melhores previsões entre os três modelos enquanto a Figura IIa apresenta o melhor ajustamento. Entretanto devido aos erros de notificação (os dados de 1988 ainda não haviam sido corrigidos) observa-se mais uma vez a dificuldade em se trabalhar com casos notificados (a serem posteriormente realocados em outros instantes de tempo).

Por este motivo parece que o modelo mais indicado para predição de AIDS é o modelo M_1 que corresponde a uma tendência determinística linear para o $\log y_t$ em lugar de uma tendência estocástica representada por M_2.

O modelo M_1 é obtido de (Ord e Young, 1988)

$$Y_t^\theta = \frac{L^\theta}{1 + \alpha\theta e^{-\beta t}}$$

que inclui os casos: $\theta = 1$ logístico, $\theta \to 0$ Gompertz e $\theta = -1$ exponencial modificada. Após alguma manipulação obtém-se M_1

$$\log y_t = \beta_0 + \beta_1 t + \beta_2 \log Y_t + \varepsilon_t$$

$$\beta_0 = \log(L^{-\theta}\alpha\beta), \quad \beta_1 = -\beta \quad e \quad \beta_2 = (1+\theta)$$

As estimativas de mínimos quadrados são:

$$\log y_t = -20.09 - 0.21t + 3.87 \log Y_t$$
$$(6.16) \quad (0.07) \quad (1.01)$$

e um intervalo de 95% de confiança para β_2 é (5,85;1.89) e para θ é (4,85;0.89) portanto incluindo a logística e excluindo a Gompertz e a exponencial modificada. Resultado semelhante ao obtido por Migon e Gamerman (1990).

Bibliografia

Harvey, A C (1984) - Time series forecasting based on the logistic curve. J. Op. Res. Soc., vol 35, pp. 641-646.

Levenbach, H & Reuter, B E (1976) - Forecasting trending time series with relative growth rate models - Technometrics, vol. 18, pp 261-268.

McAleer, M (1986) - Specification tests for separate models: a survey - In - ML King & DE Agiles (Eds) - Specification Analysis in the Linear Model. Routledge & Kegan.

McAleer, M & Pesaram, M H (1986) - Statistical inference in non-nested econometric models. Appl. Math. and Comp., vol. 20, pp 271-311.

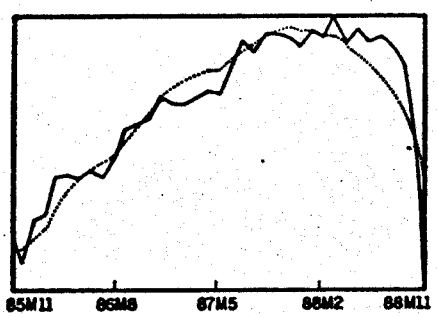
Migon, H S & Gamerman, D (1990) - Exponential growth modelling- a Bayesian approach. Rel. Tec. - LES - UFRJ, n. 41.

Ord, K & Young, P (1988) - Time series models for technological forecasting, Manuscript - Preliminar version.

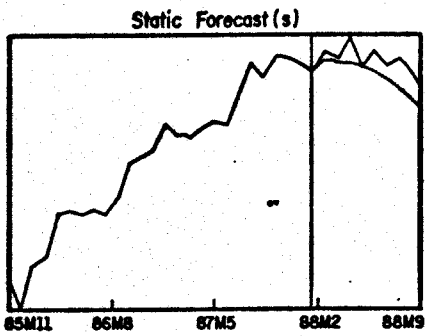
Pereira, B. B. (1977) - Discriminating among separate models: a bibliography. Intern. Statist. Review, vol. 45, pp 163-172.

Pereira, B. B. (1981) - Discriminating among separate models: an additional bibliography, Intern. Statist. Information, vol. 6, pp 3.

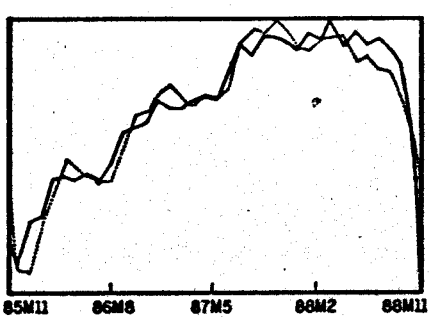
Pesaran, M. H. and Pesaran, B. (1987) - DataFit: An Iterative Econometric Software Package. Oxford, Oxford University Press.



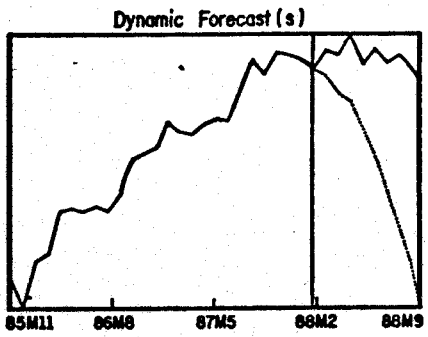
Ia A



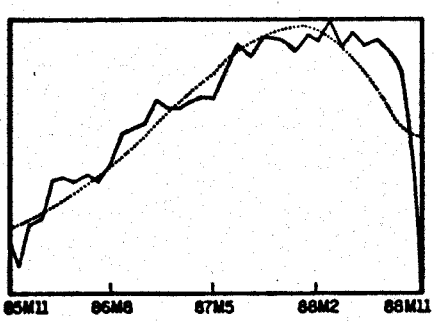
Ib A



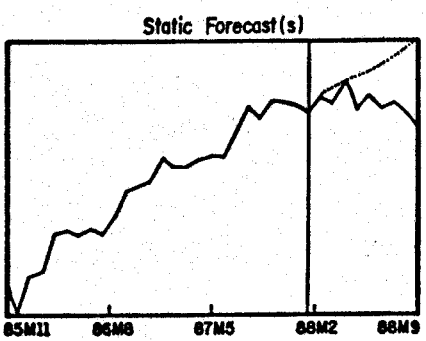
IIa B



IIb B



IIIa C



IIIb C

Observed
Fitted

Observed
Forecast