ANÁLISE DA RELAÇÃO ENTRE O ÍNDICE DE EMPREGO E PRODUÇÃO INDUSTRIAL BRASILEIROS VIA MODELOS DE REGRESSÃO LINEAR SEGMENTADA.

Vinícius D. Mayrink, Rosângela H. Loschi, Frederico R. B. da Cruz e Vander L. Aguiar

Departamento de Estatística – ICEx - Universidade Federal de Minas Gerais
Av. Presidente Antônio Carlos, 6627, Campus Pampulha, Belo Horizonte, Minas Gerais, CEP: 31270-901
E-mail: mayrink@ufmg.br, loschi@est.ufmg.br e fcruz@est.ufmg.br

Resumo: Neste trabalho consideraremos modelos de regressão segmentada para estimar a relação, ao longo do tempo, entre o índice de emprego e a produção industrial brasileiros, no período compreendido entre janeiro de 1985 e abril de 2001. No período considerado percebemos que a relação entre estas duas variáveis se modifica ao longo do tempo e, portanto, um modelo de regressão linear simples não é adequado para descrever tal comportamento. Utilizaremos, então, um modelo bayesiano de regressão segmentada onde o número e a posição dos segmentos são considerados como variáveis aleatórias, removendo assim a arbitrariedade dos modelos usuais. Para fins de comparação, propomos uma modificação no modelo clássico usual: assumimos a partição mais provável indicada pelo modelo bayesiano e ajustamos o modelo clássico de regressão segmentada. Concluímos que ambos modelos são bastante razoáveis para os dados considerados. O modelo bayesiano indica que a relação entre o índice de emprego e a produção industrial brasileiros se modifica três vezes entre janeiro de 1985 e abril de 2001.

Palavras-chave: variáveis dummy, regressão linear segmentada, modelo partição produto.

Abstract: In this paper, we consider segmented regression models to estimate the relation along the time between the Brazilian Employment Index and Industrial Production, recorded monthly from January, 1985 to April, 2001. We perceived that the linear relation between these two variables change through the time and, consequently, a simple regression model is not appropriate to describe such behavior. In the usual segmented regression models, we should arbitrate the number of segments and the observation in each segment. To approach this problem, we propose a Bayesian segmented regression model which consider as random variables the number of segments as well as the instants when a change occur, thus removing the ad hoc character of the segmented regression models. We also propose a modification in the classic segmented regression model. In this modification we assume the blocks of observations, which are indicated by the most probable partition provided by the Bayesian model. We concluded that the both models are reasonable for explaining the relation between the Brazilian Employment Index and Industrial Production and the Bayesian model indicates that this relation change three times in the period analyzed.

Keywords: dummy variables, segmented linear regression, product partition model.

1. Introdução.

Dificuldades na estimação dos parâmetros do modelo de regressão podem surgir quando os dados parecem ter sido gerados não por um único, mas por vários modelos lineares de regressão como acontece, por exemplo, com o índice de emprego e a produção industrial brasileiros mostrados na Figuras 1 e 2.

Normalmente, para modelarmos o comportamento dos dados neste tipo de situação, definimos arbitrariamente o número de modelos, blocos ou segmentos existentes, identificamos as observações em cada segmento e, posteriormente, utilizamos os estimadores de Bayes ou mínimos quadrados para estimar os parâmetros de cada modelo. (Ver, por exemplo, McGee e Carleton (1970), Holbert (1982),

Montgomery (1992), entre outros). Este tipo de procedimento é chamado de *ad-hoc*, pois não indica a maneira de escolhermos o número de segmentos, nem de como identificarmos tais segmentos. Alguns métodos não paramétricos usando *splines* para tratar este tipo de problema podem ser encontrados em Craven and Wahba (1979), Dias (1999) e outros.

Figura 1: Índice de emprego e produção industrial brasileiros.

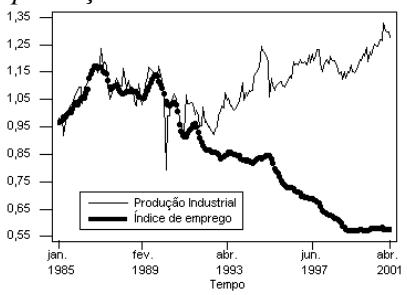
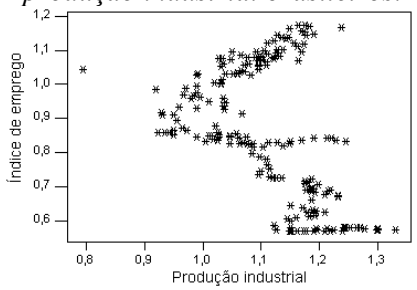


Figura 2: Relação entre o índice de emprego e produção industrial brasileiros.



O objetivo deste trabalho é analisar a relação entre o índice de emprego e a produção industrial brasileiros, registrados mensalmente entre janeiro de 1985 e abril de 2001, utilizando a abordagem bayesiana para o modelo de regressão segmentada desenvolvida em Aguiar (2003) que considera como aleatório tanto o número de segmentos B como também, os instantes ρ em que há mudanças na relação linear entre as séries. Ao usarmos este tipo de abordagem, estamos removendo a arbitrariedade existente nos modelos de regressão segmentada usuais, uma vez que estamos descrevendo nossa opinião *a priori* sobre B e ρ através de uma medida de probabilidade. Compararemos os resultados obtidos através do uso deste modelo com os resultados obtidos através do modelo clássico assumindo como segmentos o valor mais provável de ρ indicado pelo modelo bayesiano, ou seja, estamos propondo uma modificação no modelo clássico.

Este trabalho está organizado como segue: Na seção 2 definimos o modelo de regressão linear segmentada considerando o modelo partição produto definido por Barry e Hartingan (1992). Descreveremos ainda nesta seção, a proposta para o modelo clássico de regressão linear segmentada modificado. Na seção 3 aplicaremos os modelos bayesiano e clássico para estimar a relação existente entre o índice de emprego e a produção industrial brasileiros. Na seção 4, apresentamos algumas conclusões e trabalhos futuros.

2. Modelo de Regressão Linear Segmentada via modelo partição produto (MPP).

2.1 Modelo Partição Produto para Coesões de Yao.

Seja $X_1, X_2, ..., X_n$ uma seqüência de dados observados e considere o conjunto de índices $I=\{1, ..., n\}$. Considere uma partição aleatória $\rho = \{i_0, i_1, ..., i_b\}$ do conjunto de índices I tal que $0 = i_0 < i_1 < ... < i_b = n$, e uma variável aleatória B que denota o número de blocos em ρ. Considere que cada partição divide a seqüência $X_1, X_2, ..., X_n$ em B = b subseqüências contíguas, as quais serão denotadas aqui por $X_{[i,j]} = (X_{i+1}, ..., X_j)$.

Seja C_{ij} a coesão *a priori* associada com o bloco $[i, j] = \{i+1, ..., j\}$ *para* $i, j \in I \cup \{0\}$, e $j > i$, que representa o grau de similaridade entre as observações em $X_{[i,j]}$ e pode ser interpretada aqui como a probabilidade de transição na cadeia de Markov definida por pontos de mudança, Barry e Hartigan (1992).

Considere p, para $0 \le p \le 1$, a probabilidade de que uma mudança ocorra em um instante qualquer. A coesão *a priori* para o bloco [i , j] proposta por Yao (1984) é dada por:

$$C_{ij} = \begin{cases} p(1-p)^{j-i-1}; & j < n \\ (1-p)^{j-i-1}; & j = n \end{cases} \qquad \forall\, i, j \in I,\ i < j. \qquad [2.1.1]$$

Essas coesões *a priori* implicam que a seqüência de pontos de mudança estabelece um processo de renovação, com tempos de ocorrência idêntica e geometricamente distribuídos.

Seja $\theta_1,...,\theta_n$ uma seqüência de parâmetros desconhecidos, de forma que condicionalmente em $\theta_1,...,\theta_n$ a seqüência de variáveis aleatórias $X_1, X_2,..., X_n$ têm densidades condicionais marginais $f_1(X_1 \mid \theta_1),..., f_n(X_n \mid \theta_n)$, respectivamente. A distribuição *a priori* de $\theta_1,...,\theta_n$ é construída da seguinte maneira: dado uma partição $\rho = \{i_0,...,i_b\}$, para $b \in I$, temos que $\theta_i = \theta_{[i_{r-1},i_r]}$, para todo $i_{r-1} < i \le i_r$, $r = 1, \ldots, b$, e que $\theta_{[i_0,i_1]},...,\theta_{[i_{b-1},i_b]}$ são independentes entre si e de p, com $\theta_{[i,j]}$ tendo distribuição *a priori* $\pi_{[i,j]}(\theta)$, $\theta \in \Theta_{[i,j]}$, onde $\Theta_{[i,j]}$ é o espaço paramétrico correspondente ao parâmetro comum, ou seja, $\theta_{[i,j]} = \theta_{i+1} = ... = \theta_j$, que indexa a densidade condicional de $X_{[i,j]}$. Então, seguindo Barry e Hartigan (1992), definimos o modelo partição produto para coesões *a priori* de Yao como segue:

i) dado *p*, a distribuição *a priori* de ρ é a seguinte distribuição produto:

$$P(\rho = \{i_0, i_1,..., i_b\} \mid p) = p^{b-1}(1-p)^{n-b} \qquad [2.1.2]$$

para toda partição $\{i_0,..., i_b\}$, satisfazendo $0 = i_0 < i_1 < ... < i_b = n$.

ii) condicionalmente em $\rho = \{i_0,..., i_b\}$ e p, os elementos da seqüência $X_1,..., X_n$ são independentes de *p* e tem a densidade conjunta dada por:

$$f(X_0, X_1,..., X_n \mid \rho = \{i_0, i_1,...i_b\}) = \prod_{j=1}^{b} f_{i_{(j-1)}i_j}(X_{i_{(j-1)}i_j}) \qquad [2.1.3]$$

onde $f_{[i,j]}(X_{[i,j]}) = \int_{\Theta_{[i,j]}} f_{[i,j]}(X_{[i,j]} \mid \theta)\pi_{[i,j]}(\theta)d\theta$ é chamado de fator dado.

Com base na definição i) podemos expressar um resultado análogo para a variável *B*. Devemos entretanto considerar todas as possibilidades de formarmos *b* blocos, com um conjunto de *n* observações. O número de possibilidades pode ser obtido através da combinação: $\binom{n-1}{b-1}$

Desta forma teremos:

$$P(B = b \mid p) = \binom{n-1}{b-1} p^{b-1}(1-p)^{n-b} \qquad [2.1.4]$$

Barry e Hartigan (1992) também mostram que a esperança *a posteriori* (ou estimativa produto) para θ_k, *k = 1, ..., n,* é dada por:

$$E(\theta_k \mid X_1,..., X_n) = \sum_{i=0}^{k-1}\sum_{j=k}^{n} r^*_{[ij]}\, E(\theta_k \mid X_{[ij]}) \qquad [2.1.5]$$

onde $r^*_{[ij]} = P([i, j] \in \rho \mid X_0,..., X_n)$ denota a relevância *a posteriori* para o bloco [*i* , *j*].

2.2 Modelo bayesiano de regressão linear segmentada.

Nesta seção descreveremos a construção do modelo de regressão linear segmentada via modelo partição produto apresentada por Aguiar (2003). Considere para o conjunto de valores observados seqüencialmente, (X_i, Y_i) onde $i = 1, ..., n$, o seguinte modelo estatístico:

$$Y_i = g(X_i) + \varepsilon_i$$

Isto é, o comportamento de *Yi* é explicado em parte por *Xi*, através da função *g(Xi)* e, em outra parte não captada por essa função, representada por ε_i. Várias opções para *g(Xi)* podem ser utilizadas, mas a que define o modelo de regressão linear simples é:

$$g(X_i) = \alpha + \varsigma X_i$$

Portanto, o modelo de regressão que assumiremos para modelar o comportamento de uma variável resposta *Y* que está associada a uma variável explicativa *X,* é dado pela seguinte equação:

$$Y_i = \alpha_i + \varsigma_i X_i + \varepsilon_i, \quad i = 1,...,n$$

onde: $\alpha_i \in$ R denota o intercepto, $\varsigma_i \in$ R denota o coeficiente angular e ε_i denota o erro associado ao modelo. Assume-se que $\varepsilon_i \sim Normal(0, \sigma_i^2)$, onde σ_i^2 representa a variância.

O MPP será aplicado para estudar o comportamento ao longo do tempo dos parâmetros α, ς e σ^2. Condicionalmente em $(\alpha_i, \varsigma_i, \sigma_i^2)$ e X_i, temos que as variáveis resposta Y_i, para $i = 1, ..., n$, são independentes e:

$$Y_i \mid \alpha_i, \varsigma_i, \sigma_i^2, X_i \sim Normal(\alpha_i + \varsigma_i X_i \ , \ \sigma_i^2), \quad i = 1,...,n$$

Denotando por $\theta_{[i,j]} = (\alpha_{[i,j]}, \varsigma_{[i,j]}, \sigma^2_{[i,j]})$ o parâmetro comum que indexa a distribuição das variáveis pertencentes ao bloco $Y_{[i,j]}$. Temos que a distribuição de $Y_{[i,j]}$ dado $X_{[i,j]}$, $\theta_{[i,j]}$ é a distribuição Normal dada por:

$$Y_{[i,j]} \mid \theta_{[i,j]}, X_{[i,j]} \sim Normal(\alpha_{[i,j]} + \varsigma_{[i,j]} X_{[i,j]} \ , \ \sigma^2_{[i,j]})$$

As especificações *a priori* dos parâmetros $\sigma^2_{[i,j]}$, $\alpha_{[i,j]}$, $\varsigma_{[i,j]}$, a serem definidas pelo pesquisador para aplicação do MPP, são dadas por:

$$\begin{aligned} \sigma^2_{[i,j]} &\sim \ Gama\ Inversa\ \left(\vartheta_{[i,j]}/2; d_{[i,j]}/2\right) \\ \alpha_{[i,j]} \mid \sigma^2_{[i,j]} &\sim \ Normal\ \left(a_{[i,j]}; \tau_0^2 \sigma^2_{[i,j]}\right) \\ \varsigma_{[i,j]} \mid \sigma^2_{[i,j]} &\sim \ Normal\ \left(z_{[i,j]}; \gamma_0^2 \sigma^2_{[i,j]}\right) \end{aligned} \qquad [2.2.1]$$

onde: $a_{[i,j]} \in$ R, $\tau_0^2 > 0$, $z_{[i,j]} \in$ R, $\gamma_0^2 > 0$, $\vartheta_{[i,j]} > 0$ e $d_{[i,j]} > 0$.

A distribuição *a priori* de $\alpha_{[i,j]}$ é a distribuição t_1 com parâmetros $m = a_{[i,j]}$, $C = \tau_0^2$, $D = \vartheta_{[i,j]}$ e $d = d_{[i,j]}$. A distribuição *a priori* para $\varsigma_{[i,j]}$ também é uma distribuição t_1 com parâmetros $m = z_{[i,j]}$, $C = \gamma_0^2$, $D = \vartheta_{[i,j]}$ e $d = d_{[i,j]}$. Ver Loschi (1998) para identificar a função densidade.

A distribuição *a posteriori* $\alpha_{[i,j]} \mid Y_{[i,j]}, X_{[i,j]}$ é uma distribuição t_1 com parâmetros $m = a^*_{[i,j]}$, $C = V^*_{[i,j]\alpha}$, $D = \vartheta^*_{[i,j]\alpha}$ e $d = d^*_{[i,j]}$.

onde:

$$G = \gamma_0^2 \sum_{k=i+1}^{j} X_k^2 + 1 \qquad [2.2.2]$$

$$\alpha^*_{[i,j]} = \frac{G\left(a_{[i,j]} + \tau_0^2 \sum_{k=i+1}^{j} Y_k\right) - \tau_0^2 \sum_{k=i+1}^{j} X_k \left(z_{[i,j]} + \gamma_0^2 \sum_{k=i+1}^{j} X_k Y_k\right)}{(G) + (j-i)\tau_0^2 (G) - \tau_0^2 \gamma_0^2 \left(\sum_{k=i+1}^{j} X_k\right)^2} \qquad [2.2.3]$$

$$d^*_{[i,j]} = d_{[i,j]} + (j-i) \qquad [2.2.4]$$

$$V^*_{[i,j]\alpha} = \frac{\tau_0^2 \left[1 + \gamma_0^2 \sum_{k=i+1}^{j} X_k^2\right]}{\left[1 + \tau_0^2 (j-i)\right]\left[\gamma_0^2 \sum_{k=i+1}^{j} X_k^2 + 1\right] - \tau_0^2 \gamma_0^2 \left(\sum_{k=i+1}^{j} X_k\right)^2} \qquad [2.2.5]$$

$$\vartheta^*_{[i,j]\alpha} = \vartheta_{[i,j]} + \frac{\left(a_{[i,j]}^2 \gamma_0^2 + \tau_0^2 \gamma_0^2 \sum_{k=i+1}^{j} Y_k^2 + z_{[i,j]}^2 \tau_0^2\right)}{\tau_0^2 \gamma_0^2} + \\ - \frac{\left[\left(\gamma_0^2 \sum_{k=i+1}^{j} X_k Y_k\right)^2 + 2 z_{[i,j]} \sum_{k=i+1}^{j} X_k Y_k + \frac{z_{[i,j]}^2}{\gamma_0^2}\right]}{G} + \\ - \frac{\left[G\left(a_{[i,j]} + \tau_0^2 \sum_{k=i+1}^{j} Y_k\right) - \tau_0^2 \sum_{k=i+1}^{j} X_k \left(z_{[i,j]} + \gamma_0^2 \sum_{k=i+1}^{j} X_k Y_k\right)\right]^2}{\left[(G) + (j-i)\tau_0^2 (G) - \tau_0^2 \gamma_0^2 \left(\sum_{k=i+1}^{j} X_k\right)^2\right]\left[\tau_0^2 (G)\right]} \qquad [2.2.6]$$

A distribuição *a posteriori* $\varsigma_{[i,j]} \mid Y_{[i,j]}, X_{[i,j]}$ é uma distribuição t_1 com parâmetros $m = z^*_{[i,j]}$, $C = V^*_{[i,j]\varsigma}$, $D = \vartheta^*_{[i,j]\varsigma}$ e $d = d^*_{[i,j]}$.

onde:

$$z^*_{[i,j]} = \frac{\left[1 + \tau_0^2 (j-i)\right]\left[\gamma_0^2 \sum_{k=i+1}^{j} Y_k X_k + z_{[i,j]}\right] - \gamma_0^2 \sum_{k=i+1}^{j} X_k \left(\tau_0^2 \sum_{k=i+1}^{j} Y_k + a_{[i,j]}\right)}{\left[1 + \tau_0^2 (j-i)\right]\left[\gamma_0^2 \sum_{k=i+1}^{j} X_k^2 + 1\right] - \gamma_0^2 \tau_0^2 \left(\sum_{k=i+1}^{j} X_k\right)^2} \qquad [2.2.7]$$

$$V^*_{[i,j]\varsigma} = \frac{\gamma_0^2 \left[1 + \tau_0^2 (j-i)\right]}{\left[1 + \tau_0^2 (j-i)\right]\left[\gamma_0^2 \sum_{k=i+1}^{j} X_k^2 + 1\right] - \tau_0^2 \gamma_0^2 \left(\sum_{k=i+1}^{j} X_k\right)^2} \qquad [2.2.8]$$

$$\vartheta^*_{[i,j]\varsigma} = \vartheta_{[i,j]} + W\left[\frac{\left[1+\tau_0^2(j-i)\right]\left[\gamma_0^2\tau_0^2\sum_{k=i+1}^{j}Y_k^2 + a_{[i,j]}^2\gamma_0^2 + z_{[i,j]}^2\tau_0^2\right]}{\tau_0^2\left[\left(1+\tau_0^2(j-i)\right)\left(\gamma_0^2\sum_{k=i+1}^{j}X_k^2+1\right)-\gamma_0^2\tau_0^2\left(\sum_{k=i+1}^{j}X_k\right)^2\right]}\right] +$$

$$-W\left[\frac{\gamma_0^2\left(\tau_0^2\sum_{k=i+1}^{j}Y_k + a_{[i,j]}\right)^2}{\tau_0^2\left[\left(1+\tau_0^2(j-i)\right)\left(\gamma_0^2\sum_{k=i+1}^{j}X_k^2+1\right)-\gamma_0^2\tau_0^2\left(\sum_{k=i+1}^{j}X_k\right)^2\right]}\right] +$$

$$-W\left[\frac{\left[1+\tau_0^2(j-i)\right]\left[\gamma_0^2\sum_{k=i+1}^{j}Y_k X_k + z_{[i,j]}\right]-\gamma_0^2\sum_{k=i+1}^{j}X_k\left(\tau_0^2\sum_{k=i+1}^{j}Y_k + a_{[i,j]}\right)}{\left[1+\tau_0^2(j-i)\right]\left[\gamma_0^2\sum_{k=i+1}^{j}X_k^2+1\right]-\gamma_0^2\tau_0^2\left(\sum_{k=i+1}^{j}X_k\right)^2}\right]^2 \quad [2.2.9]$$

$$W = \left(V^*_{[i,j]\varsigma}\right)^{-1} \quad [2.2.10]$$

A distribuição *a posteriori* por bloco para $\sigma^2_{[i,j]}$ é uma distribuição Gama Inversa (O'Hagan, 1994) com parâmetros $\frac{\vartheta^*_{[i,j]\sigma^2}}{2}$, $\frac{d^*_{[i,j]}}{2}$. Os parâmetros indicados são definidos por [2.2.4] e [2.2.11].

$$\vartheta^*_{[i,j]\sigma^2} = \vartheta_{[i,j]} + \frac{\left(1+\tau_0^2(j-i)\right)\left(z_{[i,j]}^2\tau_0^2 + \tau_0^2\gamma_0^2\sum_{k=i+1}^{j}Y_k^2 + a_{[i,j]}^2\gamma_0^2\right)}{\tau_0^2\gamma_0^2\left(1+\tau_0^2(j-i)\right)} +$$

$$-\frac{\tau_0^4\gamma_0^2\left(\sum_{k=i+1}^{j}Y_k\right)^2 + \gamma_0^2 a_{[i,j]}^2 + 2a_{[i,j]}\tau_0^2\gamma_0^2\sum_{k=i+1}^{j}Y_k}{\tau_0^2\gamma_0^2\left(1+\tau_0^2(j-i)\right)} +$$

$$-\frac{\left[\left(1+\tau_0^2(j-i)\right)E - \tau_0^2\gamma_0^2\sum_{k=i+1}^{j}Y_k\sum_{k=i+1}^{j}X_k - \gamma_0^2 a_{[i,j]}\sum_{k=i+1}^{j}X_k\right]^2}{\gamma_0^2\left(1+\tau_0^2(j-i)\right)\left[\left(1+\tau_0^2(j-i)\right)\left(\gamma_0^2\sum_{k=i+1}^{j}X_k^2+1\right)-\tau_0^2\gamma_0^2\left(\sum_{k=i+1}^{j}X_k\right)^2\right]} \quad [2.2.11]$$

$$E = \left(z_{[i,j]} + \gamma_0^2\gamma_0^2\sum_{k=i+1}^{j}X_k Y_k\right) \quad [2.2.12]$$

Para estimarmos os instantes em que ocorrem os pontos de mudança, avaliar a dimensão das mudanças e estimar o número de mudanças ocorridas, é necessário obtermos, além das distribuições *a posteriori* por bloco para os parâmetros $\alpha_{[i,j]}$, $\varsigma_{[i,j]}$ e $\sigma^2_{[i,j]}$, a distribuição marginal de $Y_{[i,j]}$. Temos que:

$$f(Y_{[i,j]} \mid X_{[i,j]}) = \frac{\left(V^*_{[i,j]y}\right)^{\frac{1}{2}}\left(\vartheta_{[i,j]}\right)^{\frac{d_{[i,j]}}{2}}\Gamma\left(\frac{d^*_{[i,j]}}{2}\right)}{\gamma_0\tau_0(\pi)^{\frac{j-i}{2}}\Gamma\left(\frac{d_{[i,j]}}{2}\right)}\left(\vartheta^*_{[i,j]y}\right)^{\frac{-(d^*_{[i,j]})}{2}} \quad [2.2.13]$$

onde $d^*_{[i,j]}$ está especificado em [2.2.4], $V^*_{[i,j]y}$ e $\vartheta^*_{[i,j]y}$ são respectivamente dados por:

$$V^*_{[i,j]y} = \frac{\left[\tau_0^2\left(1+\gamma_0^2\sum_{k=i+1}^{j}X_k^2\right)\right]\left[\gamma_0^2\left(1+(j-i)\tau_0^2\right)\right]-\tau_0^4\gamma_0^4\left(\sum_{k=i+1}^{j}X_k\right)^2}{\left[\left(1+(j-i)\tau_0^2\right)\left(1+\gamma_0^2\sum_{k=i+1}^{j}X_k^2\right)-\tau_0^2\gamma_0^2\left(\sum_{k=i+1}^{j}X_k\right)^2\right]^2} \quad [2.2.14]$$

$$\vartheta^*_{[i,j]y} = \vartheta_{[i,j]} + \frac{a^2_{[i,j]}}{\tau_0^2} + \frac{z^2_{[i,j]}}{\gamma_0^2} + \sum_{k=i+1}^{j}Y_k^2 - \left(a^*_{[i,j]}\right)^2\frac{1+(j-i)\tau_0^2}{\tau_0^2} + $$
$$- 2a^*_{[i,j]}\,z^*_{[i,j]}\sum_{k=i+1}^{j}X_k - \left(z^*_{[i,j]}\right)^2\frac{1+\gamma_0^2\sum_{k=i+1}^{j}X_k^2}{\gamma_0^2} \quad [2.2.15]$$

Os elementos $a^*_{[i,j]}$ e $z^*_{[i,j]}$ presentes nas expressões dadas acima podem ser identificados em [2.2.3] e [2.2.7], respectivamente.

Com base nas definições apresentadas até aqui e considerando [2.1.5], temos que as estimativas produto para os parâmetros que definem o modelo de regressão linear: α_k, ς_k e σ_k^2, com $k = 1, \ldots, n$, são dadas por:

$$E\left(\alpha_{[i,j]} \mid X_1,\ldots,X_n,Y_1,\ldots,Y_n\right) = \sum_{i=0}^{k-1}\sum_{j=k}^{n} a^*_{[i,j]}\,r^*_{[i,j]} \quad [2.2.16]$$

$$E\left(\varsigma_{[i,j]} \mid X_1,\ldots,X_n,Y_1,\ldots,Y_n\right) = \sum_{i=0}^{k-1}\sum_{j=k}^{n} z^*_{[i,j]}\,r^*_{[i,j]} \quad [2.2.17]$$

$$E\left(\sigma^2_{[i,j]} \mid X_1,\ldots,X_n,Y_1,\ldots,Y_n\right) = \sum_{i=0}^{k-1}\sum_{j=k}^{n} \frac{\vartheta^*_{[i,j]\sigma^2}}{d^*_{[i,j]}-2}\,r^*_{[i,j]} \quad [2.2.18]$$

onde $a^*_{[i,j]}$, $z^*_{[i,j]}$, $\vartheta^*_{[i,j]\sigma^2}$ e $d^*_{[i,j]}$ são definidos nas expressões [2.2.3], [2.2.7], [2.2.11] e [2.2.4], respectivamente.

Assumindo p como a probabilidade de ocorrência de uma mudança em um instante qualquer, asssim como foi mencionado na seção 2.1, segue que as distribuições *a posteriori* de ρ e B, são respectivamente dadas por:

$$P\left(\rho = \{i_0,\ldots,i_b\} \mid X_1,\ldots,X_n\right) \propto \prod_{j=1}^{b} C^*_{i_{(j-1)}i_j} \quad [2.2.19]$$

$$P\left(B = b \mid X_1,\ldots,X_n\right) \propto \sum_{\xi_b}\prod_{j=1}^{b} C^*_{i_{(j-1)}i_j} \quad [2.2.20]$$

onde ξ_b é o conjunto de todas as partições em blocos contíguos contendo b blocos e $C^*_{i\,j}$ é dado por:

$$C^*_{i\,j} = \begin{cases} p(1-p)^{j-i-1} f(Y_{[i,j]} \mid X_{[i,j]}); & j < n \\ (1-p)^{j-i-1} f(Y_{[i,j]} \mid X_{[i,j]}); & j = n \end{cases} \quad [2.2.21]$$

onde $f(Y_{[i,j]} \mid X_{[i,j]})$ é a distribuição preditiva apresentada em [2.2.13]

Para o computo da esperança e da variância *a priori* para a variável *B*, considere o resultado [2.2.4] e uma variável *W* com distribuição Binomial (*n* – 1 , *p*). Perceba que $P(B = w+1 \mid p)$ é igual a $P(W = w)$. Portanto *B* é igual a *W*+1 e dessa forma obtemos a esperança e a variância como segue:

$$E(B) = E(W) + 1 = (n-1)p + 1 \qquad [2.2.22]$$
$$Var(B) = Var(W) = (n-1)p(1-p)$$

2.3 Modelo clássico de regressão linear segmentada modificado.

Para finalizar esta seção descreveremos uma proposta para o ajuste de um modelo de regressão linear segmentada, realizado via método de mínimos quadrados. Na análise clássica usual, a escolha do número e da posição dos segmentos que dividem os dados em blocos, onde suspeitamos que a relação linear muda, é realizada de maneira arbitrária: "*ad-hoc*". Consideraremos aqui a proposta de utilizar a partição mais provável indicada pelo MPP na realização desse tipo de ajuste.

Suponha que a partição mais provável indique a existência de (*n – 1*) pontos de mudança ou (*n*) blocos. Para a obtenção do modelo global de regressão linear segmentada (ver Hocking, 1996), via mínimos quadrados, é necessário assumir (*n – 1*) variáveis indicadoras denotadas por $t_1, t_2, t_3, ..., t_{n-1}$, considerando que:

$$t_i = \begin{cases} 1, & \textit{se for Bloco } i \\ 0, & \textit{caso contrário} \end{cases}, \quad \text{onde } i = 1, 2, 3, \ldots, n-1.$$

Para um modelo onde *Y* representa a variável resposta, e *X* representa a variável explicativa deve ser acrescentado na equação de regressão as variáveis indicadoras $t_1, t_2, t_3, ..., t_{n-1}$ e as interações dessas variáveis com a variável explicativa: $t_1X, t_2X, t_3X, ..., t_{n-1}X$.

Obtemos dessa forma o seguinte modelo geral:

$$Y = \alpha + \varsigma_0^{(0)}X + \varsigma_1^{(0)}t_1 + \varsigma_2^{(0)}t_2 + ... + \varsigma_{n-1}^{(0)}t_{n-1} + \varsigma_1^{(1)}t_1X + \varsigma_2^{(1)}t_2X + ... + \varsigma_{n-1}^{(1)}t_{n-1}X + \varepsilon$$

onde α, $\varsigma_i^{(0)}, \varsigma_i^{(1)}$, para *i = 1 , ... , n – 1*, representam coeficientes do modelo obtidos via mínimos quadrados.

Note que se considerarmos $t_1 = 1$ e $t_i = 0$ para $i = 2,3,...,n-1$, teremos o modelo de regressão linear ajustado, via mínimos quadrados, para os dados do bloco 1:

$$Y = \left(\alpha + \varsigma_1^{(0)}\right) + \left(\varsigma_0^{(0)} + \varsigma_1^{(1)}\right)X + \varepsilon_1$$

Resultados análogos são observados para as demais variáveis indicadoras. Tome um elemento $i \in \{1,2,...,n-1\}$ e considere $t_i = 1$. Considere também que $t_j = 0$, $\forall\ j \in \{1,2,...,n-1\}$ onde $j \neq i$. Temos então que a relação linear para as observações do bloco "*i*" ficará representada por:

$$Y = \left(\alpha + \varsigma_i^{(0)}\right) + \left(\varsigma_0^{(0)} + \varsigma_i^{(1)}\right)X + \varepsilon_i$$

Perceba que, se $t_i = 0 \quad \forall\ i \in \{1,2,...,n-1\}$, obteremos o modelo ajustado via mínimos quadrados para o bloco "*n*" que será dado por: $Y = \alpha + \varsigma_0^{(0)}X + \varepsilon_n$.

3. Resultados alcançados e discussão.

Nesta seção analisaremos o comportamento do Índice de Emprego e da Produção Industrial Brasileira no período que vai de janeiro de 1985 a abril de 2001, considerando os dois modelos apresentados na seção 2. Os dados são registrados mensalmente, compondo dessa forma 196 observações, e foram divididos por 100. Esta transformação foi feita apenas para facilitar o seu manuseio. O índice de emprego é uma série que considera as pessoas envolvidas na produção (horistas e mensalistas), que exercem atividades técnico produtivas, diretamente ligadas ao processo de produção, com vínculo empregatício ou contrato de trabalho temporário na empresa. Tanto a série Produção Industrial Brasileira, que considera a produção da industrial geral, quanto a série Índice de Emprego possuem como fonte a pesquisa industrial mensal realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Os dois conjuntos de dados podem ser obtidos através do seguinte endereço na internet: www.ipeadata.gov.br. Observe que no período inicial as séries seguem um comportamento muito parecido, no momento em que uma delas sofre um aumento a outra também segue esta tendência. Pode-se notar visualmente que parece ocorrer uma mudança no comportamento dessas duas séries a partir do ano de 1992. A Produção Industrial Brasileira começa a crescer, o Índice de Emprego mostra um decrescimento.

O Índice de Emprego é influenciado pela Produção Industrial. Caso a Produção Industrial venha a cair, é de se esperar que o desemprego aumente. Por esta razão o Índice de Emprego será considerado a variável resposta do modelo e a Produção Industrial Brasileira será a variável explicativa.

Pode-se notar da Figura 2 que um modelo linear simples não descreve bem o comportamento dos dados. De fato, se o ajustássemos, apenas 20,6% da variabilidade no Índice de Emprego seria explicada pelo modelo.

3.1 Descrição dos modelos.

Para a análise dos dados via modelo bayesiano, visto que não tínhamos muita informação *a priori* sobre a relação entre o índice de emprego e a produção industrial brasileiros, consideramos distribuições pouco informativas sobre os parâmetros, a saber:

$$\sigma^2_{[i,j]} \sim G.I.(0{,}001/2\,;0{,}001/2)$$
$$\alpha_{[i,j]} \mid \sigma^2_{[i,j]} \sim Normal\ (0\ ;1\ \sigma^2_{[i,j]})$$
$$\varsigma_{[i,j]} \mid \sigma^2_{[i,j]} \sim Normal\ (0\ ;1\ \sigma^2_{[i,j]})$$

Sabe-se que a medida que os parâmetros $\vartheta_{[i,j]}$ e $d_{[i,j]}$ [ver [2.2.1] e O'Hagan (1994)] tendem para zero, a variabilidade da distribuição Gama Inversa tende a aumentar, ou seja essa distribuição se torna cada vez menos informativa. Por esse motivo foram escolhidos, para os parâmetros $\vartheta_{[i,j]}$ e $d_{[i,j]}$, o valor 0,001. Uma vez que α e ς podem assumir qualquer valor real não temos idéia de seu valor médio. Optamos por centrar as distribuições *a priori* destes parâmetros em zero. *A priori*, assumiremos que a probabilidade p de que uma mudança ocorra em um instante qualquer é 0,01. Como conseqüência desta escolha e considerando [2.2.22], temos que *a priori* o número esperado B de segmentos é 2,95 e a variância de B é 1,93.

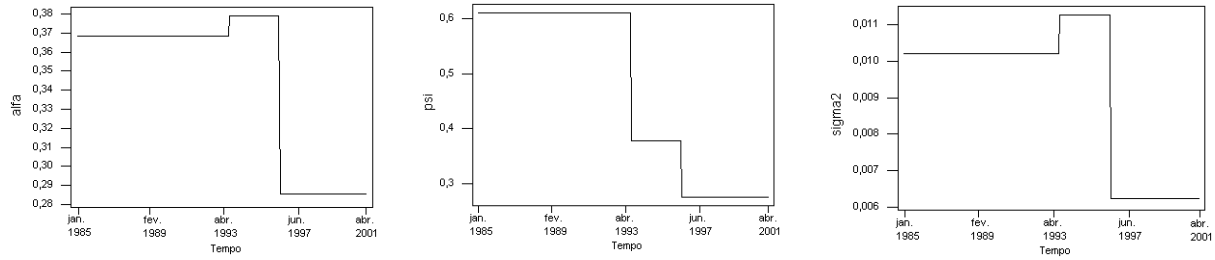
Para a análise dos dados via modelo clássico, ao invés de definirmos arbitrariamente os blocos a serem considerados, assumiremos aqueles indicados pela partição mais provável *a posteriori* fornecida pelo modelo bayesiano. Note que ao fazer isto não estamos sendo justos com este último modelo.

3.2 Análise de dados.

Considerando as especificações *a priori* apontadas na seção 3.1, observamos que as distribuições *a posteriori* de B e ρ são ambas degeneradas. A distribuição *a posteriori* de B indica que há 3 segmentos ou blocos e a distribuição *a posteriori* de ρ indica que há mudanças nos instantes 104 (agosto de 1993) e 138 (junho de 1996), ou seja $\rho = \{0, 103, 137, 196\}$.

Note da Figura 3 que os três parâmetros do modelo de regressão linear sofrem 2 mudanças cada um, e que as mudanças ocorrem em todos os parâmetros ao mesmo tempo. Perceba que a inclinação diminui e que o intercepto e a variância associada ao erro apresentam comportamento similar. A primeira mudança ocorre em agosto de 1993.

Figura 3: Estimativas a posteriori para os parâmetros do modelo de regressão – Modelo bayesiano.



Perceba que o intercepto α do modelo de regressão sofre um aumento passando do patamar de 0,3683 para 0,3791. Por outro lado, o coeficiente angular ς diminui passando de 0,6107 para 0,3778. A variância σ^2 relacionada ao erro ε do modelo de regressão tem sua magnitude aumentada de 0,0102 para 0,0113. A segunda mudança ocorre em junho de 1996, onde α sofre uma queda passando de 0,3791 para 0,2856. O parâmetro ς sofre nova queda passando 0,3778 para 0,2753. A variância relacionada ao erro ε do modelo, assim como o intercepto α, também sofre uma diminuição em seu valor: ela passa de 0,0113 para 0,0062.

Perceba que os resultados mostrados na Figura 3 indicam que a relação linear entre o Índice de Emprego e a Produção Industrial se modifica 3 vezes no período analisado. Perceba que o coeficiente ς não assume em nenhum momento um valor negativo. Concluímos com isso que as inclinações das três retas de regressão são positivas, ou seja, nos três blocos indicados pelo modelo: quanto maior a Produção Industrial maior será o Índice de Emprego. Note que a inclinação da reta, representante do conjunto de dados que forma o primeiro bloco, é maior que a inclinação das retas referentes aos dados que formam o segundo e terceiro bloco. Isso nos mostra que no período inicial, o acréscimo observado no Índice de Emprego causado pelo aumento de 1 unidade no valor da Produção Industrial é mais acentuado que nos demais períodos.

Tabela 1: Modelos ajustados por bloco via MPP e Mínimos Quadrados.

	Modelo bayesiano				*Modelo clássico*			
Bloco	α	ς	σ^2	R^2	α	ς	σ^2 =QME	R^2
1	*0,3683*	*0,6107*	*0,0102*	*22,76%*	*-0,0057**	*0,968*	*0,00437*	*56,8%*
2	*0,3791*	*0,3778*	*0,0113*	*39,28%*	*0,813*	*-0,002**	*0,00164*	*0,01%*
3	*0,2856*	*0,2753*	*0,0062*	*6,39%*	*0,949*	*-0,273*	*0,00278*	*5,5%*

* estimativa do coeficiente não significativa.

Considerando os segmentos indicados pelo modelo bayesiano, assuma as seguintes variáveis indicadoras:

$$t_1 = \begin{cases} 1 & se\ Bloco\ 1 \\ 0 & caso\ contrário \end{cases} \qquad t_2 = \begin{cases} 1 & se\ Bloco\ 2 \\ 0 & caso\ contrário \end{cases}$$

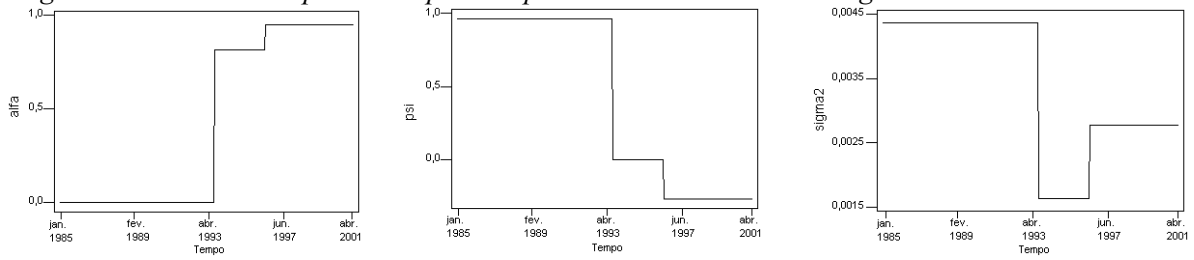
Considere as seguintes notações:

- *IE*: representa a série Índice de Emprego.
- *PI*: representa a série Produção Industrial Brasileira.

Dessa forma, o modelo clássico ajustado para os dados com os quais trabalhamos é:

$$IE = 0{,}949 - 0{,}273\ PI - 0{,}954\ t_1 - 0{,}136\ t_2 + 1{,}24\ t_1\ PI + 0{,}271\ t_2\ PI$$

Figura 4: Estimativas a posteriori para os parâmetros do modelo de regressão – Modelo clássico.

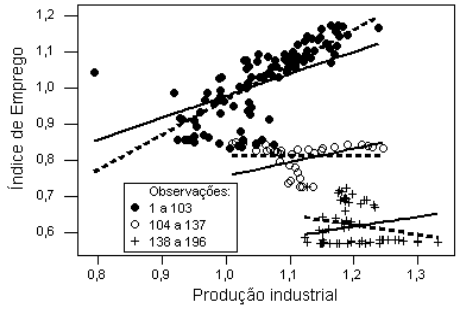


A Tabela 1 mostra as estimativas de α, ς e σ^2 para cada segmento e para ambos os modelos. Perceba da Tabela 1 e da Figura 4 que para o modelo clássico, ao compararmos o primeiro e o segundo bloco, temos que o intercepto α do modelo de regressão linear sofre um aumento passando do patamar de –0,0057 para 0,813. Por outro lado, o coeficiente angular ς sofre uma diminuição passando de 0,968 para –0,002. A variância σ^2 relacionada ao erro do modelo de regressão, também sofre uma diminuição passando da magnitude de 0,00437 para 0,00164. Comparando o segundo e o terceiro bloco, notamos que o intercepto α apresenta novo aumento passando de 0,813 para 0,949. O coeficiente angular ς sofre nova diminuição passando de –0,002 para –0,273 e a variância relacionada ao erro do modelo, após ter sofrido uma diminuição, apresenta agora um aumento passando de 0,00164 para 0,00278.

Observe pela Tabela 1 que o coeficiente angular estimado em cada bloco pelo modelo clássico é positivo apenas para o primeiro bloco. Os demais coeficientes angulares, registrados para os blocos 2 e 3, são negativos indicando que quanto maior a Produção Industrial menor será o Índice de Emprego. Temos que no período inicial há um acréscimo no Índice de Emprego causado pelo aumento de 1 unidade no valor da Produção Industrial. Nos períodos seguintes essa relação se inverte.

Na Figura 5, pode-se visualizar ambos os modelos ajustados.

Figura 5: Retas de regressão ajustadas via
Modelo bayesiano (reta contínua) e Modelo clássico (reta tracejada).



Pela Tabela 1 pode-se encontrar também os valores da estatística R^2 para os dois modelos, bayesiano e clássico, em cada bloco. Pode-se notar que para os blocos 2 e 3, a variabilidade no Índice de Emprego explicada pelo modelo para tais blocos é maior se o modelo bayesiano é considerado. No entanto, para o primeiro segmento o modelo bayesiano explica apenas 22,76% da variabilidade por bloco enquanto que o modelo clássico explica 56,8% desta variabilidade.

Utilizando o método considerado por Hocking (1996) determinamos a porcentagem da variabilidade total explicada por cada modelo. Vimos que o modelo bayesiano e clássico modificado explicam respectivamente 87,41% e 91% da variabilidade total no Índice de Emprego, o que significa dizer que ambos são bons modelos para descrever a relação entre o Índice de Emprego e Produção Industrial Brasileira no período analisado.

Note que o modelo clássico com a modificação proposta explica uma porcentagem um pouco maior da variabilidade no Índice de Emprego. No entanto, o modelo bayesiano é melhor no sentido de

que ele estabelece a probabilidade deste modelo ser o correto uma vez que fornece a distribuição *a posteriori* do número de segmentos e dos instantes onde as mudanças ocorrem.

4. Conclusões

Um modelo bayesiano de regressão linear segmentada construído utilizando o modelo partição produto foi descrito, e foi proposta uma modificação no modelo clássico de regressão linear segmentada utilizando informações provenientes do modelo bayesiano. Ambos os modelos foram utilizados para analisar a relação entre o índice de emprego e a produção industrial brasileiros. O modelo bayesiano identificou três segmentos e indicou mudanças nos três parâmetros do modelo de regressão, uma em agosto de 1993, outra em junho de 1996, as quais ocorreram ao mesmo tempo. Utilizando os segmentos indicados pelo modelo bayesiano, ajustou-se o modelo clássico. Observamos que a variabilidade explicada pelo modelo clássico é um pouco maior, mas este modelo é pior que o modelo bayesiano no sentido de que não existe uma medida de incerteza para avaliar o quão provável de ser verdadeiro é o modelo ajustado.

Acreditamos que as estimativas do modelo bayesiano ficariam melhores, se utilizássemos distribuições *a priori* mais informativas. Também percebemos que o modelo bayesisano pode ser ainda mais flexível se descrevermos a incerteza sobre p através de uma distribuição *a priori* não degenerada.

Agradecimentos: Vinícius Diniz Mayrink é bolsista do programa PIBIC – CNPq. A pesquisa de Rosângela Helena Loschi é parcialmente financiada pelo CNPq (300325/2003-7). Frederico R. B. da Cruz agradece ao CNPq (301809/96-8 e 201046/94-6) e FAPEMIG (CEX-289/98 e CEX-855/98).

Referências Bibliográficas.

- AGUIAR, V. L. *Estimativas produto para o modelo de regressão linear*. Dissertação de Mestrado, Departamento de Estatística, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil, 2003.
- BARRY, D., AND HARTIGAN, J. Product partition models for change point problems. *The Annals of Statistics* 20, 1 (1992), 260–279.
- CRAVEN, P. AND WAHBA, G. Smoothing noisy data with splines functions, Numerische Mathematik 31 (1979), 377 – 403.
- DIAS, R. Sequential adaptive non parametric regression via h-splines, *Communications in Statistics: Computational and Simulations* 28 (1999), 501 – 515.
- HOCKING, R. R. *Methods and applications of linear models – Regression and the Analysis of variance*, Wiley series in Probability and Statistics. New York: John Wiley, 1996.
- HOLBERT, D. A Bayesian analysis of a switching linear model. *Journal of Econometries* 19 (1982), 77 – 87.
- LOSCHI, R. H., *Imprevistos e suas conseqüências*. Tese de Doutorado. Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brasil, 1998.
- McGEE, V. E. AND CARLETON, W. T. Piecewise Regression. *Journal of the American Statistical Association* 65 , 331 (1970), 1109 – 1124.
- MONTGOMERY, D. C., AND PECK, E. A. *Introduction to linear regression analysis*, Wiley series in Probability and Statistics. New York: John Wiley, 1992.
- O'HAGAN, A. Kendall's Advanced Theory of Statistics 2A, Chapter Bayesian Inference. New York: John Wiley, 1994.
- YAO, Y. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annal of Statistics* 1984; 12(4): 1434–1447.