Um modelo de espaço de estados poisson para a modelagem dos confrontos de futebol entre Brasil e Argentina

Thiago R. Santos

Departamento de Estatística, Universidade Federal de Minas Gerais (UFMG).
E-mail: thiagors007@gmail.com.

Resumo: *Este artigo apresenta um modelo de espaço de estados Poisson para a modelagem dos confrontos históricos de futebol entre Brasil e Argentina. Este modelo permite o cálculo exato da função de verossimilhança marginal de fácil implementação, bem como das distribuições preditivas, suavizadas e de filtragem para a variável latente. Não há a necessidade de aproximações, que é algo muito comum na literatura se tratando de modelos de espaço de estados não-Gaussianos. A inserção de covariáveis pode ser feita, bem como o tratamento da irregularidade dos dados sem maiores dificuldades. Uma covariável denominada "Fator campo" foi inserida no modelo e foi possível verificar que a mesma influencia significativamente o resultado do jogo. Quando a seleção joga em casa, ela tem mais chance de marcar mais gols em seu adversário, algo que já era esperado em geral. Os resultados são muito satisfatórios e ilustram bem o modelo aqui proposto e desenvolvido.*

Palavras-chave: Modelo não-Gaussiano; inferências Bayesiana e clássica; função de verossimilhança exata; partidas de futebol.

Abstract: *This article presents a Poisson state space model for modeling the historical soccer matches between Brazil and Argentina. This model allows the exact calculation of marginal likelihood function in a easy fashion as well as the predictive, smoothing and filtering distributions for the latent variable. It is not necessary to use approximations, which is very common in the literature under the non-Gaussian context. The insertion of covariates and treatment of data irregularity may be done in a natural way without problems. A covariate called "Factor field", that indicates the match place (home or outside), is inserted into the model and influences the outcome of the match. When the team plays at home, it has more chance to score goals, which was already expected in general. The results are very satisfactory and illustrate well the proposed model.*

Keywords: Non-Gaussian model; Bayesian and classical inferences; exact marginal likelihood function; soccer matches.

1 Introdução

Este artigo apresenta um modelo de espaço de estados Poisson com verossimilhança exata para a modelagem dos confrontos históricos de futebol entre Brasil e Argentina. Este modelo permite o cálculo da função de verossimilhança exata, bem como das distribuições preditivas e distribuições suavizadas e de filtragem da variável latente. Não há a necessidade de aproximações, que é algo muito recorrente na literatura se tratando de modelos de espaço de estados não-Gaussianos. A inserção de covariáveis pode ser feita, bem como o tratamento da irregularidade dos dados sem maiores dificuldades.

Há alguns trabalhos na literatura para a previsão de partidas de futebol que se utilizam de métodos Bayesianos e computação intensiva, por causa complexidade do modelo que possui vários componentes latentes, dinâmicos (WEST; HARRISON, 1997; HARVEY, 1989), entre os quais pode-se citar Rue e Salvesen (2000), Souza e Gamerman (2004), e Farias (2008). Neste trabalho os modelos são de fácil implementação, porém o único componente estocástico é o nível, isto é, os demais componentes e fatores devem ser incluídos no modelo de forma determinística. Uma abordagem similar foi adotada por Harvey (1989).

1.1 Problema motivante

No site da FIFA, foram obtidos os placares, os dados estão disponíveis em http://es.fifa.com, de 93 jogos entre Brasil e Argentina no período de 1914 a 2008. Tem-se a série histórica do número de gols marcados pela seleção do Brasil por ano contra a seleção da Argentina. Logo, algumas perguntas de interesse emergem, tais como: O fator campo influencia no placar do confronto?, É possível predizer o número de gols do Brasil do próximo jogo? É possível analisar esses dados, que são irregularmente espaçados? É possível responder a essas perguntas com a metodologia que será apresentada a diante.

2 O modelo Poisson de verossimilhança exata

A formulação deste modelo e os principais resultados são herdados do trabalho de Gamerman, Santos e Franco (2013).

2.1 Definição

De uma maneira geral, define-se que a série temporal $\{y_t\}$ possui uma distribuição na família gama de modelos dinâmicos, se a sua distribuição é escrita na forma:

$$p(y_t|\mu_t, \boldsymbol{\varphi}) = a(y_t, \boldsymbol{\varphi})\mu_t^{b(y_t,\boldsymbol{\varphi})} \exp(-\mu_t c(y_t, \boldsymbol{\varphi})), \tag{1}$$

se $y_t \in H(\boldsymbol{\varphi}) \subset \Re$ e $p(y_t|\mu_t, \boldsymbol{\varphi}) = 0$, caso contrário.
As funções $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ e $H(\cdot)$ são tais que $p(y_t|\mu_t, \boldsymbol{\varphi}) \geq 0$ e $\int p(y_t|\mu_t, \boldsymbol{\varphi})dy_t = 1$.

O modelo é definido da seguinte forma:

1. a função de ligação é $\mu_t = \lambda_t g(x_t^{'}\boldsymbol{\beta})$, onde $g(\cdot)$ é uma função do preditor linear, $x_t^{'}$ é um vetor de covariáveis, β é seu coeficiente de regressão (um dos componentes de $\boldsymbol{\varphi}$) e λ_t é dado pela equação de evolução logo abaixo;

2. A equação de evolução é dada por: $\lambda_t = w^{-1}\lambda_{t-1}\varsigma_t$, onde $\varsigma_t \sim Beta\,(wa_{t-1}, (1-w)a_{t-1})$;

3. $\lambda_0|Y_0 \sim Gama(a_0, b_0)$.

É importante destacar que se o logaritmo da equação de evolução é tomado, é obtida a seguinte equação $\ln(\lambda_t) = \ln(\lambda_{t-1}) + \ln(\frac{\varsigma_t}{w})$, onde $\ln(\frac{\varsigma_t}{w}) \in \Re$. Essa equação é similar à usual equação de evolução dada por um passeio aleatório. w varia entre 0 e 1 e também compõe $\boldsymbol{\varphi}$. Em geral, deseja-se modelar a variância ou a média dessa distribuição, que será função de alguns parâmetros invariantes no tempo e do parâmetro de escala μ_t, isto é, $E(y_t|\mu_t, \boldsymbol{\varphi}) = f(\mu_t, \boldsymbol{\varphi})$.

Suponha que uma observação no tempo t é retirada de uma distribuição de Poisson com média μ_t,

$$p(y_t|\mu_t, \boldsymbol{\varphi}) = \mu_t^{y_t} \exp(-\mu_t)/y_t!, \tag{2}$$

onde $y_t = 0, 1, \ldots$, $\mu_t = \lambda_t \exp(x_t^{'}\boldsymbol{\beta})$ com uma função de ligação logarítmica. Esse modelo pertence à família Gama de modelos dinâmicos em que $a(y_t, \boldsymbol{\varphi}) = (y_t!)^{-1}$, $b(y_t, \boldsymbol{\varphi}) = y_t$ e $c(y_t, \varphi) = 1$. Logo, $\boldsymbol{\varphi} = (w, \beta)^{'}$

A distribuição *a priori* é a mesma do Teorema 1. Com as as funções $b(\cdot,\cdot)$ e $c(\cdot,\cdot)$ identificadas, utilizando o Teorema 1, a distribuição *a posteriori* de $\mu_t|\boldsymbol{Y}_t$ é dada pela distribuição Gama com parâmetros

$$\begin{aligned} a_t^* &= a_{t|t-1}^* + y_t, \\ b_t^* &= b_{t|t-1}^* + 1; \end{aligned}$$

sendo que $a_{t|t-1}^* = wa_{t-1}$ e $b_{t|t-1}^* = wb_{t-1}[exp(x_t^{'}\boldsymbol{\beta})]^{-1}$.

Logo, segue-se que $\lambda_t = \mu_t[g(x_t^{'}\boldsymbol{\beta})]^{-1}|\boldsymbol{Y}_t$ tem também distribuição Gama com parâmetros (equações de atualização)

$$\begin{aligned} a_t &= wa_{t-1} + y_t, \\ b_t &= wb_{t-1} + exp(x_t^{'}\boldsymbol{\beta}). \end{aligned}$$

Substituindo as funções $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$ e usando o Teorema 1, obtém-se a distribuição preditiva, que é Binomial negativa, dada por

$$p(y_t|\boldsymbol{Y}_{t-1},\boldsymbol{\varphi}) = \left(\begin{array}{c} a_{t|t-1}^* + y_t + 1 \\ y_t \end{array} \right) (b_{t|t-1}^*)^{a_{t|t-1}^*}(1+b_{t|t-1}^*)^{-(a_{t|t-1}^*+y_t)} \quad ,$$

em que $y_t = 0, 1, 2, \ldots$

e

$$\left(\begin{array}{c} a_{t|t-1}^* + y_t + 1 \\ y_t \end{array} \right) = \frac{\Gamma(a_{t|t-1}^* + y_t)}{\Gamma(y_t+1)\Gamma(a_{t|t-1}^*)}.$$

Das propriedades da distribuição binomial negativa, conclui-se que a média e variância da distribuição preditiva de $y_{n+1}|\mathrm{Y}_n,\boldsymbol{\varphi}$ são, respectivamente,

$$\tilde{y}_{n+1} = \mathrm{E}(y_{n+1}|\boldsymbol{Y}_n,\boldsymbol{\varphi}) = a_{n+1|n}^*/b_{n+1|n}^*$$

e

$$var(y_{n+1}|\boldsymbol{Y}_n,\boldsymbol{\varphi}) = a_{n+1|n}^*(1+b_{n+1|n}^*)/(b_{n+1|n}^*)^2.$$

A função de log-verossimilhança é obtida através das distribuições preditivas cuja forma é dada por

$$\begin{aligned} &\ln L(\boldsymbol{\varphi};\boldsymbol{Y}_n) = \sum_{t=1}^{n} \ln\Gamma(a_{t|t-1}^* + y_t) - \ln y_t! - \ln\Gamma(a_{t|t-1}^*) + \\ &a_{t|t-1}^* \ln b_{t|t-1}^* - (a_{|t-1}^* + y_t)\ln(1+b_{t|t-1}^*), \end{aligned} \tag{3}$$

onde $\boldsymbol{\varphi} = (w,\beta)^{'}$.

Os Estimadores de Máxima Verossimilhança (EMV) para os parâmetros do modelo podem ser obtidos através da maximização da função de verossimilhança acima, bem como os intervalos de confiança assintóticos usando as propriedades assintóticas dos EMV (HARVEY, 1989; GAMERMAN; SANTOS; FRANCO, 2013).

Métodos Bayesianos podem ser utilizados para se fazer inferência sobre os parâmetros estáticos do modelo. Combinando uma distribuição *a priori* para $p(\boldsymbol{\varphi})$ com a função de verossimilhança, pode se obter a distribuição *a posteriori* $p(\boldsymbol{\varphi}|\boldsymbol{Y}_n)$. A distribuição *a posteriori* nem sempre é analiticamente tratável (forma conhecida), logo deve-se lançar mão de métodos Monte Carlo (MC) e/ou Markov chain Monte Carlo (MCMC) (GAMERMAN; LOPES, 2006; GAMERMAN; SANTOS; FRANCO, 2013) a fim de obter uma amostra da distribuição *a posteriori* e, assim, proceder com as inferências sobre os parâmetros. Inferência sobre parâmetros latentes pode ser feita através das suas distribuições de filtragem e suavização (GAMERMAN; SANTOS; FRANCO, 2013).

Os resíduos de Pearson e Deviance podem ser usados para verificar se o modelo está bem ajustado. Mais detalhes em Gamerman, Santos e Franco (2013) e Harvey (1989).

3 A série temporal histórica do confronto Brasil × Argentina

Os dados consistem dos número de gols marcados pelas seleções no confronto histórico entre Brasil e Argentina no período de 1914 a 2008. Como em alguns anos não houveram jogos, têm-se dados faltantes que deverão ser tratados. Também houveram anos que ocorreram mais de um jogo em um mesmo ano.

Para contornar esse problema, terá que ser feito uma alteração nas equações da distribuição *a priori* e atualização do algoritmo recursivo.

Da Figura 1, percebe-se que a média de gols marcados pelo Brasil está em torno de 2, salvo alguns anos nos quais o número de gols foi superior a 2. Na Figura 1 alguns anos possuem valores altos, pois aconteceram mais de uma partida nos mesmos.

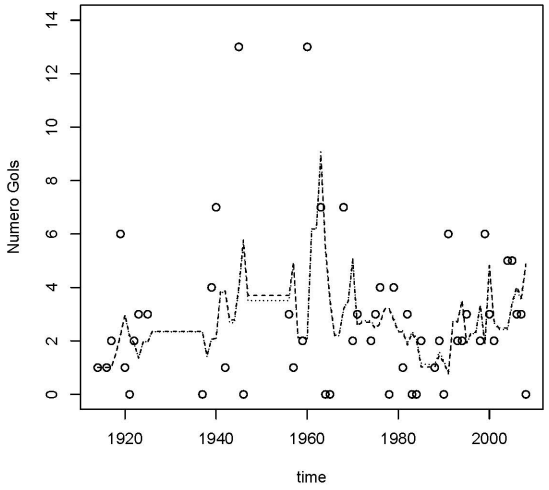

Figura 1: Série temporal do número de gols marcados pelo Brasil no confronto com a Argentina. Os pontos representam a série temporal e as linhas tracejada e pontilhada representam as médias da distribuição preditiva dos ajustes clássico e Bayesiano, respectivamente.

Como os dados referem-se a jogos no Brasil e também fora do Brasil, o modelo Poisson é ajustado com uma covariável X_t que assume o valor 1 se o jogo foi realizado no Brasil e 0 se for fora. Tem-se interesse em saber se realmente faz diferença jogar contra Argentina em casa.

Para o ajuste clássico, o algoritmo de maximização numérica BFGS foi usado para a maximizar a função de verossimlhança. Já para o ajuste Bayesiano, o algoritmo de Metropolis-Hasting (MCMC) é implementado para obter uma amostra distribuição *a posteriori* dos parâmetros. Duas cadeias de 10.000 valores foram executadas, excluindo as 8.000 primeiras como *burn-in*. Através de testes e métodos gráficos, verificou-se a convergência das cadeias (resultados omitidos aqui). As distribuições *a priori* para os parâmetros estáticos são $w \sim Unif(0,1)$ e $\beta \sim Unif(1000,-1000)$ e $\lambda_0|Y_0 \sim Gama(0,01;0,01)$ para o parâmetro latente.

Na Tabela 1, o EMV, os estimadores pontuais Bayesianos, intervalos de confiança e de credibilidade para ψ são apresentados do ajuste do modelo Poisson a série real. Nota-se que os valores do EMV e do EB-Moda ficaram bem próximos tanto para o parâmetro ω quanto para β. Os limites do intervalo de credibilidade para β são similares aos do intervalo de confiança assintótico, porém isso não acontece para ω. Observe que o limite superior do intervalo assintótico está fora do espaço paramétrico de ω. Isso pode acontecer por se tratar de um intervalo aproximado.

Observe que $\hat{\beta} = 0,485$ (maior que zero), indicando que jogar em casa confere ao Brasil uma certa vantagem, como esperado. Tomando $\exp(\hat{\beta})$, obtém-se $1,624$, isto é, o número esperado de gols marcados pelo Brasil aumenta $62,40\%$, quando o Brasil joga em casa contra Argentina.

Esse valor é significativo ao nível de $0,95$ de confiança e de credibilidade, pois o valor zero não está contido tanto no intervalo de confiança quanto no de credibilidade.

Os resíduos de Pearson padronizados foram avaliados e a variância amostral dos mesmos foi igual a 1.042. Através da análise dos gráficos dos resíduos, não observou-se nenhuma violação das suposições do modelo.

Tabela 1: Ajuste do modelo Poisson à série do número de gols marcados pelo Brasil no confronto com a Argentina.

Métodos	ω	β
EMV	1,000	0,485
Mediana	0,956	0,461
Media	0,949	0,475
Moda	0,999	0,506
Int. Cred.	[0,856; 0,997]	[0,125; 0,817]
Int. Assint.	[0,721; 1,190]	[0,153; 0,812]

A Tabela 2 apresenta as probabilidades do Brasil marcar 0, 1, 2, 3, 4 e 5 ou mais gols contra Argentina no próximo jogo em casa ou fora de casa. Nota-se que a probabilidade do Brasil fazer nenhum gol jogando fora de casa é maior do que jogando em casa.

Tabela 2: Distribuição de probabilidade preditiva do número de gols do Brasil para a próxima partida contra a Argentina.

	Fator campo	
Número de gols	Em casa	Fora de casa
0	0,125	0,277
1	0,257	0,353
2	0,267	0,227
3	0,187	0,099
4	0,099	0,033
>4	0,065	0,011

O ajuste do modelo pode ser melhorado, pois há inúmeros fatores que influenciaram essa série temporal histórica que poderiam ser discutidos e incluídos no modelo, porém, neste estudo, não será entrado nesse mérito. O intuito desta aplicação é mostrar a pontencialidade e aplicabilidade dos modelos de resposta não-Gaussiana nos mais diversos tipos de problemas. A previsão do placar da partida pode ser realizado facilmente fazendo a modelagem da série do número de gols marcados pela Argentina em confrontos contra o Brasil, usando a distribuição preditiva das observações e assumindo algumas hipóteses apropriadas.

Conclusão

Este artigo propõe um modelo para a modelagem dos confrontos entre Brasil e Argentina em partidas de futebol, podendo ser implementado de uma maneira rápida e simples. Todas as perguntas levantadas *a priori* foram respondidas.

É sabido que os resultados dos jogos são influenciados por outros fatores que não foram aqui contemplados, entretanto motivam a continuidade do trabalho, a fim de incluí-los. Por exemplo, fatores defesa, ataque e passes errados podem ser inseridos modelo. A previsão de cada uma das rodadas de um campeonato pode ser desenvolvida, baseado nos resultados deste estudo.

Agradecimentos

O autor agradece à Universidade Federal de Minas Gerais (Programa recém-doutor, Pró-reitoria de pesquisa, PrPq), ao CNPq e ao editor da revista pelo encorajamento e apoio.

Referências

FARIAS, F. F. *Análise e Previsão de Resultados de Partidas de Futebol.* Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, 2008.

GAMERMAN, D.; LOPES, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* (2nd edition). London: Chapman and Hall. 2006.

GAMERMAN, D.; SANTOS, T. R.; FRANCO, G. C. A non-Gaussian family of state-space models with exact marginal likelihood. *Journal of Times Series Analysis*, v. 34, 625-645. 2013.

HARVEY, A.C. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press. 1989.

RUE, H.; SALVESEN, O. *Prediction and retrospective analysis of soccer matches in a league.* Norwegian University of Science and Technology. Trondheim, Noruega. 2000.

SOUZA JR, O. G.; GAMERMAN, D. *Previsão de partidas de futebol usando modelos dinâmicos.* Anais do XXXVI SBPO. São João del Rey - MG. 2004.

WEST, M.; HARRISON, J. *Bayesian Forecasting and Dynamic Models.* New York: Springer. 1997.