PREVISÃO DE RESULTADOS EM PARTIDAS DE FUTEBOL

Universidade Federal do Rio Grande do Norte - Semana de Estatística 2013

Marcelo Leme de Arruda

Introdução

A previsão (probabilística) de resultados de partidas de futebol não é mais do que um caso particular do problema fundamental enfrentado por qualquer estatístico: a inferência sobre grandezas desconhecidas a partir de valores conhecidos e observados.

Modelos de previsões de resultados de futebol, assim como de previsões de indicadores financeiros, de taxas biológicas ou de qualquer outra grandeza que se queira estudar, se baseiam essencialmente em dois ingredientes: uma representação paramétrica dessa grandeza e um método para obtenção/estimação desses parâmetros. As duas primeiras e principais etapas deste minicurso serão voltadas à análise desses dois ingredientes.

É pertinente esclarecer que diversas fontes utilizam a palavra "modelo" para designar a representação paramétrica da grandeza em estudo ("Modelo de Poisson", por exemplo). Neste minicurso, porém, para se prevenir ambigüidades, o termo "modelo" será utilizada para designar somente o processo como um todo (a representação paramétrica mais o método de estimação/obtenção dos parâmetros)

Existe, ainda, um terceiro ingrediente, igualmente importante, mas nem sempre levado em consideração com a atenção que mereceria: a análise da qualidade do modelo. Essa análise pode ser baseada em (pelo menos) duas abordagens possíveis, as quais, acompanhadas de comentários sobre os modelos apresentados, serão objeto da terceira seção.

A quarta seção apresentará, como estudo de caso, um exemplo de aplicação concreta de um modelo existente e utilizado por um site de previsões estatísticas de resultados de futebol. Por fim, serão tecidas, na última seção, algumas considerações sobre temas atualmente em aberto na área da estatística aplicada a partidas de futebol.

1. Representação paramétrica

A representação paramétrica é a descrição matemática da grandeza que se quer estudar. No caso particular das partidas de futebol e de competições esportivas pareadas (i.e. em que os participantes se enfrentam dois a dois) em geral, há duas possíveis grandezas de interesse, cada qual relacionada a uma das duas representações paramétrica mais usualmente empregadas:

1.1. Representação para o resultado do jogo

Existem análises que utilizam como grandeza de interesse somente o resultado do jogo, i.e., a informação acerca de quem venceu ou se o jogo terminou empatado. Essas

análises essencialmente se baseiam na representação de Bradley-Terry (1952), cuja formulação simples facilitou sua popularização entre os estudiosos do assunto.

Considerando um conjunto de *N* competidores e $\pi_1, \pi_2, \ldots, \pi_N \ \ (\pi_i > 0, i = 1, 2, \ldots, N)$ parâmetros associados à força (habilidade, nível técnico, etc.). de cada competidor, então, de acordo com a **representação de Bradley-Terry**, a probabilidade de o competidor *i* derrotar o competidor *j* num confronto direto entre ambos é dada por:

$$p_{i,j} = \frac{\pi_i}{\pi_i + \pi_j}$$

Além de ser naturalmente intuitiva e de fácil compreensão, essa representação tem respaldo teórico, podendo ser derivada a partir da **Distribuição de Gumbel** (1961, também conhecida como Distribuição de Valores Extremos). Uma variável aleatória contínua *X* tem Distribuição de Gumbel, com parâmetros μ (de localização) e β (de escala), se sua função densidade de probabilidade é dada por

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} - e^{-\frac{x-\mu}{\beta}} \right) .$$

Por conseguinte, sua função distribuição acumulada é dada por:

$$F(x) = e^{-e^{-\frac{(x-\mu)}{\beta}}} .$$

Suponha-se, então, que cada um dos times em questão tenha associado a si um escore latente aleatório (e independente de quem seja o adversário) *S* com distribuição de Gumbel com parâmetros $\beta = 1$ e $\mu = \ln \pi_i$.

Então, o escore S_i do *i*-ésimo time tem distribuição acumulada.

$$F(s_i) = e^{-e^{-(s_i - \ln \pi_i)}} .$$

Definindo $\Delta_{ij} = S_i - S_j$ como o "resultado" do jogo (a "margem de vitória" a favor do time *i*), pode-se mostrar que sua distribuição acumulada é

$$F(\Delta_{ij}) = P(\Delta_{ij} \le \delta) = \frac{1}{1 + e^{(\ln \pi_i - \ln \pi_j) - \delta}} .$$

Logo, a probabilidade de o time *i* derrotar o time *j* é dada por

$$P(i \text{ vencer } j) = P(\Delta_{ij} > 0) = 1 - P(\Delta_{ij} \le 0) = \frac{1}{1 + e^{-(\ln \pi_i - \ln \pi_j)}} = \boxed{\frac{\pi_i}{\pi_i + \pi_j}} .$$

Em sua formulação padrão, a representação de Bradley-Terry se aplica somente a esportes em que não existem empates. Um dos exemplos mais difundidos da aplicação dessa representação é o xadrez, cujas principais entidades internacionais (incluindo a FIDE – Fédération Internationale des Échecs [1]) utilizam como ferramenta oficial de classificação o Ranking Elo (1978). Esse ranking equivale a representar a performance de cada enxadrista por uma Distribuição de Gumbel cujos parâmetros são definidos de uma forma específica e atualizados, após cada jogo ou série de jogos, por uma regra igualmente específica.

Existem, por outro lado, adaptações e/ou expansões da representação de Bradley-Terry, as quais contemplam, em sua formulação, fatores como:

- a possibilidade de um confronto terminar empatado;
- o efeito "vantagem do primeiro jogador" (equivalente, no futebol, ao "fator mando de campo" ou, no xadrez, à "vantagem de jogar com as brancas");
- a margem de vitória, discernindo entre placares distintos (1x0, 2x0, 2x1 etc.) ou entre "vitórias folgadas" e "vitórias apertadas";
- etc.

1.2. Representação para o placar do jogo

Análises mais abrangentes utilizam como grandeza de interesse o placar do jogo (i.e. os escores efetivamente obtidos por cada time) e não somente o resultado (a identificação do vencedor). Usualmente, nessas análises o número de gols marcados por um time é representado por uma **Distribuição de Poisson**, segundo a qual a probabilidade de o time *i* marcar *x* gols num determinado jogo é dada por:

$$P(X = x) = \frac{e^{-\lambda_i} \lambda_i^x}{x!} \ ,$$

onde $\lambda_i = E[X_i]$ é o número esperado de gols marcados por esse time *i* no jogo em questão.

É plausível considerar, contudo, que o número esperado de gols marcados por um time dependa da força do time adversário. É perfeitamente presumível, por exemplo, que contra adversários mais fracos, um time tenda a marcar mais gols do que contra adversários mais fortes. Por essa razão, uma representação mais adequada pode ser a **Distribuição de Holgate** (1964), uma classe de distribuições bivariadas de Poisson, cuja função de probabilidade conjunta é dada por:

$$P(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_{12})} \sum_{i=0}^{\min(x,y)} \frac{\lambda_1^{x-i} \lambda_2^{y-i} \lambda_{12}^i}{(x-i)!(y-i)!i!} \ .$$

Essa distribuição pode ser caracterizada da seguinte forma: sejam P_1, P_2 e P_{12} três variáveis aleatórias independentes com distribuição de Poisson com médias respectivamente iguais a λ_1, λ_2 e λ_{12}. Então, o vetor $(X, Y) = (P_1 + P_{12}, P_2 + P_{12})$ segue uma Distribuição de Holgate, com função de probabilidade igual à acima formulada.

A presença de P_{12} em ambas as somas é responsável pela existência de uma dependência entre as variáveis *X* e *Y*. Essa dependência, por sua vez, pode proporcionar uma representação mais realística das quantidades de gols marcados por um time *X* quando jogando especificamente contra o time *Y*.

Evidentemente, outras representações para o placar podem ser formuladas (usando, por exemplo, uma distribuição Binomial Negativa ou uma distribuição Gama discretizada), assim como podem ser desenvolvidas representações para o resultado diferentes da de Bradley-Terry. Porém, por serem mais freqüentemente utilizadas, a próxima seção se dedicará predominantemente às representações de Bradley-Terry e de Poisson.

2. Estimação/obtenção dos parâmetros

2.1. Estimação por máxima verossimilhança

A estimação por máxima verossimilhança é, talvez, o mais intuitivo modo de se obter os parâmetros necessários para o cálculo das probabilidades. Trata-se, numa explicação resumida, de procurar, dentre todos os valores possíveis que os parâmetros podem assumir, aqueles que maximizam a probabilidade de ocorrência dos resultados previamente observados.

Para a representação de Bradley-Terry, lembrando que a probabilidade de o competidor *i* derrotar o competidor *j* num confronto direto entre ambos é dada por:

$$p_{i,j} = \frac{\pi_i}{\pi_i + \pi_j},$$

tem-se que, para uma coleção de resultados de jogos entre diversos times, a verossimilhança conjunta é dada por:

$$L = \prod_{i=1}^{N} \prod_{\substack{j=1 \\ j \neq i}}^{N} \frac{\pi_i^{n_i}}{(\pi_i + \pi_j)^{n_{ij}}},$$

onde *N* é a quantidade total de times em estudo;

n_{ij} é o total de vitórias do time *i* em jogos contra o time *j*

e $n_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} n_{ij}$ é o total de vitórias do time *i* em jogos contra todos os demais times.

Os valores dos parâmetros $\pi_1, \pi_2, \ldots, \pi_N$ dos competidores podem, então, ser estimados por meio da maximização da verossimilhança *L*.

Não há, como regra geral, uma forma analítica fechada para os estimadores de máxima verossimilhança de Bradley-Terry. Porém, existem diversos métodos numéricos facilmente programáveis que permitem obter esses valores.

Uma observação importante que cabe ser tecida acerca desses estimadores é o fato de não serem únicos! Com efeito, se os valores $\pi_1, \pi_2, \ldots, \pi_N$ forem todos multiplicados por uma mesma constante *k*, os valores resultantes $k\pi_1, k\pi_2, \ldots, k\pi_N$ também servirão como parâmetros de Bradley-Terry, pois, chamando-se de $p'_{i,j}$ a probabilidade de vitória do time *i* sobre o time *j* calculada a partir dos novos parâmetros, é fácil perceber que:

$$p'_{i,j} = \frac{k\pi_i}{k\pi_i + k\pi_j} = \frac{\pi_i}{\pi_i + \pi_j} = p_{i,j}$$

Esse fenômeno é uma conseqüência direta da natureza multiplicativa da representação de Bradley-Terry (pois as probabilidades de vitória são diretamente proporcionais aos parâmetros π) e, para evitar ambigüidades, costuma-se usar a restrição adicional $\pi_1 + \pi_2 + \ldots + \pi_N = 1$.

Para a representação de Poisson (Holgate), por outro lado, a estimação por máxima verossimilhança não é viável, uma vez que a função de verossimilhança envolve produtos de somatórias que, por sua vez, envolvem produtos de fatoriais. Escrever essa função por si só já é complicado e procurar valores de parâmetros λ que a maximizem seria uma tarefa ainda mais complexa.

2.2. Modelos lineares (mínimos quadrados)

Outra forma possível para obtenção de parâmetros é a estimação por mínimos quadrados. Aqui, diferentemente da abordagem por máxima verossimilhança, os parâmetros são considerados como variáveis dependentes de informações observadas (variáveis explicativas) e essa dependência é explicitada sob a forma de modelos lineares. Genericamente, pode se representar essa relação de dependência como:

$$\theta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki} + \varepsilon_i \ ,$$

onde θ_i é um parâmetro (para o *i*-ésimo jogo) que se quer estimar

$x_{1i}, x_{2i}, \ldots, x_{ki}$ são variáveis explicativas de cujos valores depende o parâmetro θ;

$\alpha_1, \alpha_2, \ldots, \alpha_k$ são (hiper)parâmetros cujos valores se deseja estimar

e ε_i é um componente de erro (distância entre valores observados e previstos).

Aqui, θ_i pode ser um π_i de Bradley-Terry, um λ_i da Distribuição de Holgate, um parâmetro de outra representação ou mesmo uma função de parâmetro(s), como será exemplificado mais à frente.

O tratamento padrão para essa abordagem é buscar os valores de $\alpha_1, \alpha_2, \ldots \alpha_k$ que minimizam o erro quadrático total, ou seja, que tornam mínima a soma

$$\sum \varepsilon_i^2 = \sum [\theta_i - (\alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki})]^2 \ .$$

Uma característica útil dessa abordagem é o fato de elementos como a identificação do time adversário, o fator "jogar em casa" e quaisquer outras variáveis eventualmente pertinentes poderem ser inseridas como variáveis explicativas nesse modelo linear. Essa inserção dispensa a representação paramétrica de estruturas que explicitem a dependência desses fatores, permitindo que tal representação possa ser mais leve e manipulável.

Além disso, é importante mencionar que a estimação dos valores de $\alpha_1, \alpha_2, \ldots \alpha_k$ pode ser realizada por Mínimos Quadrados Ordinários (MQO) ou por Mínimos Quadrados Ponderados (MQP). A adoção dos MQP em lugar dos MQO permite conferir pesos diferenciados a cada jogo e, com isso, discernir entre amistosos e jogos de competição, acentuar a informação proporcionada por jogos mais recentes em comparação com jogos mais antigos ou estabelecer qualquer ênfase que se deseje a algum fator de interesse.

Por fim, embora a formulação supra exposta se baseie em modelos básicos de Regressão Linear Múltipla, existem modelos que se fundamentam em formulações mais sofisticadas, como Modelo Linear Geral (GLM) (McCullagh e Nelder, 1989), Regressão Logística (Draper e Smith, 1998) etc. O modelo de Lee (1997), baseado em GLM de Poisson, é um exemplo real da aplicação e implementação de tais formulações avançadas. Embora esses modelos se embasem em teorias mais avançadas que a regressão linear "básica", a minimização dos erros é perfeitamente factível e os principais softwares estatísticos possuem esses modelos em suas programações.

2.3 Estimação *bayesiana* e métodos iterativos

Uma outra abordagem para a obtenção de parâmetros é a atualização iterativa de valores. Isso significa que, após um determinado time disputar um jogo ou seqüência de jogos, o novo valor do parâmetro desse time será obtido diretamente a partir do "antigo" valor e do resultado desse(s) jogos disputado(s). Analiticamente, esse processo pode ser genericamente representado como

$$\theta_i^{'} = f(\theta_i, R) \ ,$$

onde $\theta_i^{'}$ é o valor atualizado do parâmetro de interesse para o *i*-ésimo time;

θ_i é o valor anterior desse parâmetro;

e R é o resultado do(s) jogo(s) disputado(s) por esse time.

Aqui, novamente θ pode ser um π_i de Bradley-Terry, um λ_i da Distribuição de Holgate, um parâmetro de outra representação etc.

Como um exemplo simples de tais métodos iterativos, pode-se citar o Ranking Elo (não-oficial) de seleções nacionais de futebol [2]. Após um jogo contra a seleção *j*, esse sistema atualiza o parâmetro da seleção *i* de acordo com a fórmula

$$\theta_i^{'} = \theta_i + K(S_o - S_e) \ ,$$

onde *K* é um peso que depende da competição por que o jogo é válido e da diferença de gols a favor do mandante;

S_o é o resultado obtido no jogo em questão pela seleção *i* (1 ponto por vitória, 0,5 por empate e 0 por derrota)

e $S_e = 1 \cdot P(\text{vitória}) + 0 \cdot P(\text{derrota}) = \dfrac{\pi_i}{\pi_i + \pi_j}$ é o resultado esperado para esse jogo.

Por fim, os parâmetros dessa representação de Bradley-Terry são definidos como:

$$\pi_i = \begin{cases} 10^{(\theta_i + 100)/400} & \text{se a seleção } i \text{ jogar em casa} \\ 10^{\theta_i / 400} & \text{caso contrário} \end{cases}$$

e, analogamente:

$$\pi_j = \begin{cases} 10^{(\theta_j + 100)/400} & \text{se a seleção } j \text{ jogar em casa} \\ 10^{\theta_j / 400} & \text{caso contrário} \end{cases} .$$

Existem diversas outras formas de atualização iterativa de parâmetros, muitas delas formuladas arbitrariamente, sem maior embasamento teórico. Existe também uma classe, mais importante e abrangente, composta pelos métodos de Estimação *Bayesiana* (O'Hagan, 1994).

Resumidamente, a estimação *bayesiana* consiste na atribuição de uma distribuição de probabilidades para o(s) parâmetro(s) da representação e na atualização dos hiperparâmetros dessa distribuição após cada jogo ou conjunto de jogos. As notações usuais para essas distribuições são:

$\pi(\theta)$ - distribuição *a priori* do parâmetro (escalar ou vetorial) θ;
$f(x \mid \theta)$ - distribuição (*verossimilhança*) de *x* condicional ao valor de θ,
$\pi(\theta \mid x)$ - distribuição *a posteriori* de θ, condicional ao valor observado de *x*.

Dadas uma distribuição *a priori* π e a verossimilhança *f* gerada pelos dados observados *x*, a distribuição *a posteriori* é definida como:

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{\int_{\Theta} \pi(\theta) f(x \mid \theta) d\theta} \propto \pi(\theta) f(x \mid \theta) .$$

Em outras palavras, a distribuição *a posteriori* é obtida a partir do produto da *priori* pela verossimilhança, sendo o denominador $\int_{\Theta} \pi(\theta) f(x \mid \theta) d\theta$ somente uma constante de normalização.

Num exemplo simplificado para facilidade de ilustração, considere-se que os gols de um time sejam representados por uma distribuição de Poisson com média λ. Então a verossimilhança gerada por um jogo em que esse time marcou x_o gols seria:

$$f(x_o \mid \lambda) = P(X = x_o \mid \lambda) = \frac{e^{-\lambda}\lambda^{x_o}}{x_o!}$$

Supondo que a distribuição a *priori* para λ seja uma Gama com parâmetros α e β:

$$\pi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \ ,$$

então a distribuição *a posteriori* para λ dado o valor observado x_o é igual a

$$f(\lambda \mid x_o) = \frac{\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \cdot \frac{e^{-\lambda}\lambda^{x_o}}{x_o!}}{\int_0^{\infty} \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \cdot \frac{e^{-\lambda}\lambda^{x_o}}{x_o!} d\lambda} \ .$$

Essa operação freqüentemente envolve integrais complicadas no denominador, o que inviabiliza a obtenção direta da *posteriori*. Contudo, para uma grande classe de distribuições (notadamente as pertencentes à Família Exponencial), esse trabalho é facilitado pela existência das classes de *prioris* conjugadas (Fink, 1997). Isso significa dizer que, para determinadas verossimilhanças, existem distribuições *a priori* que conduzem a *posterioris* da mesma família, tendo apenas os valores dos hiperparâmetros atualizados.

No exemplo acima, como a *priori* Gama é conjugada para verossimilhanças Poisson, a distribuição a *posteriori* também será uma Gama, com parâmetros $\alpha' = \alpha + x_o$ e $\beta' = \beta + 1$, de onde:

$$\pi(\lambda \mid x_o) = \frac{\beta^{\alpha+x_o}}{\Gamma(\alpha + x_o)}\lambda^{\alpha+x_o-1}e^{-(\beta+1)\lambda} \ .$$

De posse da distribuição *a posteriori*, há algumas maneiras usuais para se representar a distribuição futura da variável de interesse *x*:

a) como uma distribuição $f(x)$ com parâmetro $\theta' = E[\lambda \mid x_o]$ (a esperança *a posteriori* de $\theta \mid x_o$);

b) como uma distribuição $f(x)$ com parâmetro $\theta'' = \max[\lambda \mid x_o]$ (a moda *a posteriori* de $\theta \mid x_o$);

c) por meio da Distribuição Preditiva $DP(x) = \int_0^{\infty} \pi(\lambda \mid x_o)P(x \mid \lambda)d\lambda$.

A obtenção da Distribuição Preditiva novamente envolve integrais complicadas e difíceis de serem diretamente calculadas. Novamente, porém, o uso de classes de *prioris* conjugadas muitas vezes facilita esse trabalho. Assim, no exemplo ilustrativo, essas três distribuições seriam:

a) Distribuição de Poisson com média $\lambda' = E[\lambda \mid x_o] = \frac{\alpha + x_o}{\beta + 1}$.

b) Distribuição de Poisson com média $\lambda'' = \max[\lambda \mid x_o] = \frac{\alpha + x_o - 1}{\beta + 1}$.

c) Distribuição Preditiva: Binomial Negativa com $n = \alpha + x_o$ e $p = \frac{1}{\beta + 2}$.

De um modo geral, os modelos existentes na literatura são mais complexos que os exemplos aqui apresentados. Os modelos iterativos de Soares (1982 e 1998), por exemplo, utilizam uma representação com dois parâmetros por time (força ofensiva e força defensiva), os quais são atualizados recíproca e simultaneamente. Os modelos *bayesianos* de Suzuki *et* al (2010) e Glickman (1993) envolvem, na formulação da verossimilhança, a inclusão de parâmetros relacionados à força do(s) adversário(s), ao local de realização do jogo (fator "jogar em casa") etc.

O modelo de Glickmann, aliás, envolve até mesmo a passagem de tempo (i.e. o "envelhecimento" do time) entre os jogos, assunto que será abordado mais à frente, nos comentários finais.

2.4 Estimação direta

Uma última categoria a ser abordada é a daqueles modelos que "extraem" as probabilidades de vitória, empate e derrota diretamente a partir de estatísticas descritivas dos times em confronto, sem nem sequer formular uma representação paramétrica.

Em outras palavras, o que difere esses modelos dos demais é o fato de utilizarem não parâmetros induzidos por uma representação previamente formulada, mas indicadores externos, pré-existentes e usualmente sem vínculo com a estrutura probabilística utilizada. Exemplos de tais indicadores são o Ranking de seleções nacionais da FIFA [3] e totais cumulativos de pontos ganhos, vitórias, gols marcados etc. por cada time no campeonato em questão ou mesmo ao longo da história.

Embora essa abordagem permita contemplar, sem maiores dificuldades, a inclusão de fatores adicionais (como por exemplo o fator "jogar em casa"), com freqüência o uso de indicadores externos e desvinculados da estrutura probabilística acaba comprometendo a essência dos cálculos e conduzindo a resultados que não necessariamente refletem a realidade.

Exemplo típico dessa abordagem seria tomar R_x e R_y (os pontos de dois times, X e Y, no ranking de seleções da FIFA) e formular as probabilidades de vitória de cada time segundo a representação de Bradley-Terry (aqui, para facilidade de ilustração, a possibilidade de empate está sendo ignorada):

$$P(\text{vitória de } X) = \frac{R_x}{R_x + R_y} \text{ e } P(\text{vitória de } Y) = \frac{R_y}{R_x + R_y}$$

Existe nessa formulação uma clara incongruência conceitual, que pode levar a resultados muito pouco condizentes com a realidade técnica dos times em questão. Isso suscita uma reflexão muito importante, embora raramente efetuada, que voltará a ser abordada na próxima seção.

3. Verificação de qualidade

A qualidade preditiva de um modelo pode ser analisada sob duas visões distintas. Pode-se avaliar o modelo de forma "filosófica" com base em características de sua formulação ou pode-se estudá-lo por um viés mais numérico, com base no confronto, para uma coleção de jogos passados, entre resultados observados e probabilidades previamente calculadas.

A avaliação "filosófica" pode ser efetuada diretamente a partir da descrição teórica do modelo, mesmo antes de qualquer probabilidade ser calculada e de qualquer jogo ser iniciado. Assim, pode-se referir a essa avaliação como "análise anterior". A análise numérica por outro lado, depende da existência de um histórico de jogos cujos resultados possam ser comparados com as probabilidades anteriormente anunciadas. Evidentemente, esse histórico só poderá ser formado após o cálculo das probabilidades e a realização dos jogos. Por essa razão, esse estudo pode ser referido como "análise posterior".

3.1. Análise anterior

A análise anterior consiste basicamente numa verificação crítica da "consistência" da fundamentação do modelo. É uma reflexão qualitativa (i.e. sem a construção de indicadores numéricos) dos porquês de cada passo da construção do modelo.

É através dessa análise que se pode evidenciar falhas de conceito ou de construção que podem comprometer o aplicação do modelo e, em última instância, produzir valores irreais para as probabilidades de interesse.

Usualmente, os modelos de estimação direta, abordados na seção 2.4, são os que mais cometem equívocos detectáveis nessa análise. Como ilustração de tais equívocos, pode-se citar o exemplo apresentado naquela seção, no qual os pontos de cada seleção no ranking da FIFA foram usados como parâmetros de uma representação de Bradley-Terry. Nessa representação, as probabilidades de cada resultado são diretamente proporcionais aos valores dos parâmetros de cada time:

$$\frac{P(\text{vitória de } X)}{P(\text{vitória de } Y)} = \frac{\pi_x / (\pi_x + \pi_y)}{\pi_y / (\pi_x + \pi_y)} = \frac{\pi_x}{\pi_y} \quad .$$

Isso equivaleria, naquele exemplo, a arbitrar que as probabilidades de cada resultado fossem diretamente proporcionais aos pontos de cada seleção no ranking da FIFA. Porém, não existe nada no processo de construção desse ranking que permita afirmar que uma

seleção que tenha *k* vezes a pontuação de outra seja "*k* vezes melhor" ou tenha uma probabilidade de vitória "*k* vezes maior" que a adversária.

Outros exemplos de inconsistências detectáveis na análise anterior são os modelos que utilizam como informações dados como médias ou totais de gols marcados, gols sofridos, pontos ganhos etc., calculados isolada e independentemente para cada time e ignorando os adversários contra os quais tais números foram obtidos. Essa prática pode levar a grandes distorções, especialmente se essas médias e totais forem obtidas a partir de universos (i.e. conjuntos de adversários) substancialmente diferentes.

Como exemplo concreto, basta lembrar das Eliminatórias para a Copa do Mundo de 2002, quando a Austrália marcou 66 gols em 4 jogos contra seleções semiamadoras da Oceania e alcançou, ao final de sua participação no torneio, uma média de 9,125 gols por jogo (73 gols em 8 jogos). Como todas as outras zonas continentais eram tecnicamente mais niveladas que a Oceania, nenhuma outra seleção do planeta atingiu tamanha quantidade de gols marcados e, conseqüentemente, para um modelo que se baseasse diretamente nas médias de gols por jogo nas Eliminatórias, a Austrália seria sempre destacadamente favorita à vitória contra qualquer seleção nacional que enfrentasse.

Analogamente aos aspectos explorados nesses exemplos, outras características do modelo também podem ser estudadas e examinadas na análise anterior: a plausibilidade das variáveis explicativas escolhidas, o tratamento dedicado ao tempo decorrido entre os jogos passados (responsáveis pelas informações existentes no banco de dados) e o jogo presente (cujas probabilidades se quer calcular), a razoabilidade do modelo probabilístico formulado etc.

3.2. Análise posterior

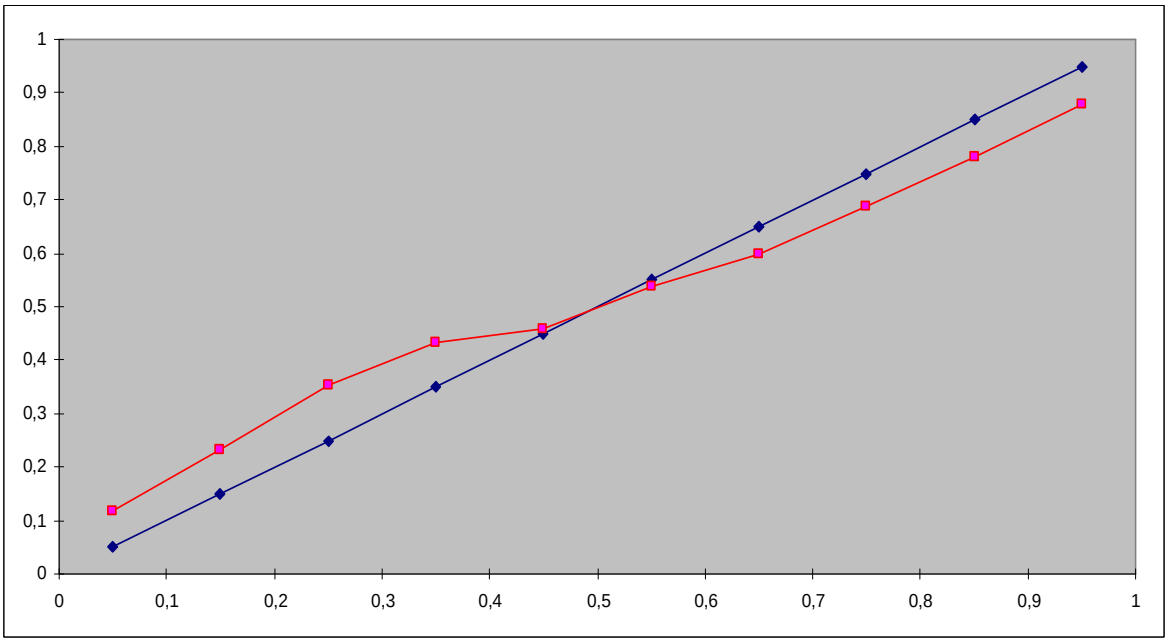
A análise posterior, que trata do confronto entre as probabilidades calculadas e os resultados efetivamente observados, pode ser feita basicamente por meio de dois atributos, cada qual com sua própria medida de qualidade.

O primeiro atributo é a confiabilidade, conceito relacionado ao comportamento dos jogos quando observados em conjunto. Nesse contexto, as probabilidades são tratadas como valores individuais (escalares) e a Medida de Confiabilidade pode ser definida em termos futebolísticos como:

$$MC = \sum_p \left(\frac{\#VO_p + \#EO_p + \#DO_p}{\#VP_p + \#EP_p + \#DP_p} - p \right)^2 \text{, onde:}$$

$\#VO_P$ = número de vitórias (com probabilidade p atribuída) ocorridas;
$\#EO_P$ = número de empates (com probabilidade p atribuída) ocorridos;
$\#DO_P$ = número de derrotas (com probabilidade p atribuída) ocorridas;
$\#VP_P$ = número de vitórias (com probabilidade p atribuída) previstas;
$\#EP_P$ = número de empates (com probabilidade p atribuída) previstos;
$\#DP_P$ = número de derrotas (com probabilidade p atribuída) previstas.

Em outras palavras, a aferição da confiabilidade consiste em comparar cada valor *p* com a freqüência observada de resultados cuja probabilidade anunciada era igual a *p*. Na prática, como *p* é um número real que pode assumir qualquer valor entre 0 e 1, as probabilidades são agrupadas em intervalos ($[0\,;0{,}1]$, $[0{,}1\,;0{,}2]$, ... , $[0{,}9\,;1]$), por exemplo) e a comparação é realizada entre o ponto médio de cada intervalo e a freqüência observada de resultados cuja probabilidade anunciada pertencia a esse intervalo, conforme se pode visualizar no gráfico abaixo:



Nesse gráfico, a diagonal azul (reta identidade) representa as probabilidades anunciadas (freqüências relativas esperadas) e a curva vermelha corresponde às freqüências efetivamente observadas em cada intervalo. A Medida de Confiabilidade equivale, então, à distância euclidiana quadrática entre essas duas curvas.

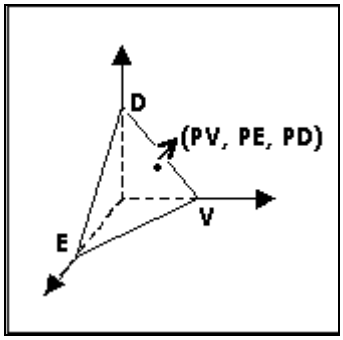
Evidentemente, quanto mais próximas as freqüências relativas observadas estiverem das probabilidades anunciadas, menor a distância entre as curvas e portanto menor será o valor de MC. Logo, a avaliação do modelo segundo esse quesito será tão melhor quanto menor for o valor a Medida de Confiabilidade, com a qualidade máxima (confiabilidade perfeita) sendo atingida quando $MC = 0$, isto é, quando as duas curvas forem coincidentes.

O segundo atributo a ser avaliado é a exatidão que, em contraste com a confiabilidade, está associado ao comportamento dos jogos quando analisados individualmente. Nessa abordagem, as probabilidades são tratadas como vetores, a partir dos quais a Distância de DeFinetti (1972) é assim definida:

$$DDF = \begin{cases} (PV-1)^2 + (PE-0)^2 + (PD-0)^2 & \text{se a equipe mandante vencer o jogo;} \\ (PV-0)^2 + (PE-1)^2 + (PD-0)^2 & \text{se a equipe mandante empatar o jogo;} \\ (PV-0)^2 + (PE-0)^2 + (PD-1)^2 & \text{se a equipe mandante perder o jogo;} \end{cases}$$

Conforme pode-se ver na ilustração a seguir, essa construção equivale a considerar um simplex contido em $\mathbb{R}^3$ como representação geométrica do conjunto das possíveis

previsões probabilísticas. Nesse simplex, os vértices correspondem às ocorrências dos resultados ($(1,0,0)$ para a vitória do mandante, $(0,1,0)$ para o empate e $(0,0,1)$ para a vitória do visitante) e os demais pontos a todas as outras possíveis previsões.



Assim, a medida de Distância de DeFinetti corresponde, geometricamente, à distância euclideana quadrática entre o ponto correspondente ao vetor de probabilidades anunciadas e o vértice correspondente ao resultado efetivamente observado. Para a análise de um conjunto de dois ou mais jogos, utiliza-se a Medida de DeFinetti, índice dado pela média aritmética das distâncias de DeFinetti calculadas para cada jogo individual.

Semelhantemente ao que ocorre com a Medida de Confiabilidade, quanto menores forem as distâncias entre as probabilidades anunciadas e os vértices (resultados anunciados), menor será o valor de DDF e, conseqüentemente, o modelo será tão melhor avaliado sob esse atributo quanto menor for a Medida de DeFinetti. Aqui, a qualidade máxima (exatidão perfeita) será atingida quando as probabilidades anunciadas sempre coincidirem com os vértices efetivamente observados (i.e. quando as probabilidades forem iguais a 1 para o resultado que de fato acontecer e 0 para os demais resultados), o que equivale a $MDF = 0$.

A avaliação da qualidade de um modelo por meio da Medida de DeFinetti pode incluir a comparação do valor dessa medida com algum padrão de referência. Para modelos de previsão de resultados de futebol, um padrão comumente utilizado é a medida obtida por um modelo que, para **qualquer** jogo, atribuísse preguiçosamente probabilidades iguais a **todos** os resultados possíveis $(PV = PE = PD = 1/3)$.

A Medida de DeFinetti para o "modelo preguiçoso" é igual a

$$(1/3-1)^2 + 2\cdot(1/3-0)^2 = 0{,}6667\,.$$

e, conseqüentemente, podem ser considerados modelos de qualidade minimamente aceitável aqueles que apresentarem Medidas de DeFinetti menores que 0,6667. Com efeito, se um modelo tiver MDF maior que 0,6667, então mais conveniente que utilizá-lo seria aderir ao "modelo preguiçoso".

Uma relação entre essas duas medidas é exibida por Murphy (1972), que estabelece uma partição da Medida de DeFinetti em duas parcelas, sendo a primeira uma versão ligeiramente modificada da Medida da Confiabilidade e a segunda relativa à "resolução" (*resolution*) do modelo (grosso modo, uma medida de dispersão das freqüências relativas observadas).

Além dessas duas medidas, um indicador importante a ser observado é a "taxa de funcionamento" do modelo, ou seja, a proporção de vezes em que o procedimento gera valores aceitáveis para os parâmetros. Tomando como exemplo um modelo baseado numa representação paramétrica de Poisson, mesmo que suas confiabilidade (MC) e exatidão (MDF) sejam excelentes, se as estimações dos parâmetros λ freqüentemente produzirem valores negativos, de pouca serventia será esse modelo.

Por fim, cabe alertar para uma "medida" freqüentemente utilizada por leigos e que na realidade constitui um grave equívoco conceitual. Trata-se da "taxa de acerto", calculada como a freqüência com que ocorre o resultado (vitória, empate ou derrota) ao qual se havia atribuído maior probabilidade. Em linguagem leiga, essa taxa equivale à verificação de quantas vezes o modelo "acertou o vencedor" dos jogos e é nessa expressão ("acertar o vencedor") que reside o equívoco: qualquer evento que tem probabilidade *p* de acontecer, também tem probabilidade $1-p$ de não acontecer; logo, as duas possibilidades (acontecer e não acontecer) estão previstas e, por conseguinte, não se pode rotular como "acerto" ou "erro" a ocorrência do evento de maior probabilidade.

O exemplo abaixo ajuda a perceber como, além de conceitualmente incorreta, essa "taxa de acerto" também pode conduzir a julgamentos inadequados acerca da qualidade dos modelos. Considere-se dois modelos hipotéticos que tivessem produzido as seguintes probabilidades para um jogo futuro entre as equipes *X* e *Y*:

	P(vitória de *X*)	*P*(empate)	*P*(vitória de *Y*)
Modelo I	0,90	0,06	0,04
Modelo II	0,35	0,33	0,32

Suponha-se também que o time *Y* tenha vencido esse jogo. Então, os dois modelos teriam "errado o vencedor" e conseqüentemente teriam "taxa de acerto" igual a zero. Todavia, é fácil perceber, numa análise mais atenta dos números, que o Modelo I, ao ter atribuído um favoritismo maior ao time *X*, "errou mais" que o Modelo II e, portanto, que índices como as Medidas de Calibração e de DeFinetti mensurariam as qualidades dos modelos com muito mais fidedignidade do que a simples "taxa de acerto".

4. Estudo de Caso

Como exemplo de aplicação de tudo o que foi apresentado nas seções anteriores, será analisada uma versão simplificada do modelo desenvolvido por Arruda (2000) e aplicado no site Chance de Gol [4]. Esse modelo se baseia num par de Distribuições de Poisson como representação paramétrica para o número *G* de gols marcados por cada equipe. Assim, num confronto entre os times *i* e *j*, as variáveis G_i e G_j têm distribuições de probabilidades:

$$P(G_i = g) = \frac{e^{-\lambda_i}\lambda_i^g}{g!} \text{ e } P(G_i = g) = \frac{e^{-\lambda_j}\lambda_j^g}{g!} .$$

A estimação dos valores dos parâmetros λ_i e λ_j se baseia essencialmente em duas observações acerca do resultado de um jogo: a diferença entre os gols marcados pelos dois times em confronto (variável indicativa de quão um time é "melhor" que o outro) e a soma

dos gols marcados pelos times em cada jogo (variável indicativa do “poder ofensivo” dessas duas equipes).

Analiticamente, isso significa uma formulação baseada em duas variáveis aleatórias, $S_{ij} = E[G_i + G_j]$ e $D_{ij} = E[G_i - G_j]$, a partir das quais podem ser facilmente obtidas as esperanças dos escores G_i e G_j de cada time:

$$\begin{cases} E[G_i] = \dfrac{E[G_i + G_j] + E[G_i - G_j]}{2} = \dfrac{S_{ij} + D_{ij}}{2} \\ E[G_j] = \dfrac{E[G_i + G_j] - E[G_i - G_j]}{2} = \dfrac{S_{ij} - D_{ij}}{2} \end{cases}$$

Dada uma coleção de resultados de jogos passados, os valores S_{ij} e D_{ij} são estimados por meio de duas equações de regressão linear múltipla:

$$\begin{cases} S_k = \alpha_1 X_{1k} + \alpha_2 X_{2k} + \ldots + \alpha_N X_{Nk} + \varepsilon_k \\ D_k = \beta_1 Y_{1k} + \beta_2 Y_{2k} + \ldots + \beta_N Y_{Nk} + \varepsilon'_k \end{cases}$$

onde: S_k é a soma dos gols marcados pelos times no *k*-ésimo jogo,

$X_{ik} = \begin{cases} 1 & \text{se o } i\text{-ésimo time participou do } k\text{-ésimo jogo} \\ 0 & \text{se o } i\text{-ésimo time não participou do } k\text{-ésimo jogo} \end{cases}$,

D_k é a diferença de gols marcados a favor do time “mandante” no *k*-ésimo jogo,

$Y_{ik} = \begin{cases} 1 & \text{se o } i\text{-ésimo time participou como "mandante" do } k\text{-ésimo jogo} \\ -1 & \text{se o } i\text{-ésimo time participou como "visitante" do } k\text{-ésimo jogo} \\ 0 & \text{se o } i\text{-ésimo time não participou do } k\text{-ésimo jogo} \end{cases}$.

$\alpha_1, \alpha_2, \ldots, \alpha_N, \beta_1, \beta_2, \ldots, \beta_N$ são (hiper)parâmetros cujos valores se deseja estimar

Aqui, os termos “mandante” e “visitante” identificam respectivamente o primeiro e o segundo nome citados no resultado do confronto (i.e., na representação usual, os nomes que aparecem à esquerda e à direita do sinal “x”), independentemente de existir ou não um time “mandante” de fato (jogando em sua própria “casa” ou detendo algum outro tipo de vantagem similar).

Como ilustração do funcionamento desse modelo, considere-se um torneio quadrangular hipotético que tenha apresentado os seguintes resultados:

Jogo 1 - Time A 2x3 Time B
Jogo 2 - Time C 5x1 Time D
Jogo 3 - Time A 4x0 Time C
Jogo 4 - Time B 1x1 Time D
Jogo 5 - Time A 0x2 Time D

O objetivo, então, é calcular as probabilidades para o Jogo 6 - Time B x Time C.

Nesse caso, ter-se-ia:

$$S = \begin{bmatrix} 5 \\ 6 \\ 4 \\ 2 \\ 2 \end{bmatrix},\ X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \text{ para a primeira regressão}$$

$$\text{e } D = \begin{bmatrix} -1 \\ 4 \\ 4 \\ 0 \\ -2 \end{bmatrix} \text{ e } Y = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \text{ para a segunda regressão.}$$

Para uma melhor visualização do conceito desse modelo, pode-se pensar na primeira regressão como a solução do sistema de equações

$$\begin{cases} \alpha_{\text{Time A}} + \alpha_{\text{Time B}} = 5 \\ \alpha_{\text{Time C}} + \alpha_{\text{Time D}} = 6 \\ \alpha_{\text{Time A}} + \alpha_{\text{Time C}} = 4 \\ \alpha_{\text{Time B}} + \alpha_{\text{Time D}} = 2 \\ \alpha_{\text{Time A}} + \alpha_{\text{Time D}} = 2 \end{cases}$$

e na segunda regressão como a solução do sistema

$$\begin{cases} \beta_{\text{Time A}} - \beta_{\text{Time B}} = -1 \\ \beta_{\text{Time C}} - \beta_{\text{Time D}} = 4 \\ \beta_{\text{Time A}} - \beta_{\text{Time C}} = 4 \\ \beta_{\text{Time B}} - \beta_{\text{Time D}} = 0 \\ \beta_{\text{Time A}} - \beta_{\text{Time D}} = -2 \end{cases} .$$

Como esses sistemas normalmente possuem mais equações (jogos) do que variáveis (times), dificilmente haverá uma solução exata. É essa a razão da utilização de modelos lineares (modelos de regressão) para a busca dos valores (α's e β's) que "mais se aproximam de solucionar os sistemas".

Assim, os valores dos parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_N$ e $\beta_1, \beta_2, \ldots, \beta_N$ são, efetivamente, estimados por meio da minimização dos erros quadráticos

$$\sum \varepsilon_k^2 = \sum [S_k - (\alpha_1 X_{1k} + \alpha_2 X_{2k} + \ldots + \alpha_N X_{Nk})]^2$$
$$\text{e } \sum \varepsilon_k'^2 = \sum [D_k - (\beta_1 Y_{1k} + \beta_2 Y_{2k} + \ldots + \beta_N Y_{Nk})]^2 \ .$$

Uma vez obtidos os estimadores de mínimos quadrados $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_N, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_N$, pode-se calcular, para qualquer jogo futuro cujas probabilidades se queira obter, as esperanças $\hat{E}[G_i + G_j]$ e $\hat{E}[G_i - G_j]$ e, por conseguinte, os parâmetros $\hat{\lambda}_i = \hat{E}[G_i]$ e $\hat{\lambda}_j = \hat{E}[G_j]$ das distribuições de Poisson associadas aos times participantes desse jogo futuro.

Voltando ao campeonato hipotético e concluindo os cálculos deste exemplo, os estimadores de mínimos quadrados para os (hiper)parâmetros são

$$\begin{cases} \hat{\alpha}_{\text{Time A}} = 1{,}25 \\ \hat{\alpha}_{\text{Time B}} = 2{,}5 \\ \hat{\alpha}_{\text{Time C}} = 4 \\ \hat{\alpha}_{\text{Time D}} = 0{,}75 \end{cases} \quad \text{e} \quad \begin{cases} \hat{\beta}_{\text{Time A}} = -0{,}125 \\ \hat{\beta}_{\text{Time B}} = 0 \\ \hat{\beta}_{\text{Time C}} = -0{,}5 \\ \hat{\beta}_{\text{Time D}} = -0{,}875 \end{cases} .$$

De posse desses estimadores e supondo que o próximo jogo (cujas probabilidades se quer calcular) seja Time B x Time C, pode-se então calcular as esperanças estimadas

$$\hat{E}[G_B + G_C] = (1{,}25 \cdot 0) + (2{,}5 \cdot 1) + (4 \cdot 1) + (0{,}75 \cdot 0) = 6{,}5 \quad \text{e}$$
$$\hat{E}[G_B - G_C] = (-0{,}125 \cdot 0) + (0 \cdot 1) + (-0{,}5 \cdot (-1)) + (0{,}875 \cdot 0) = 0{,}5 \ ,$$

e as esperanças marginais estimadas

$$\hat{\lambda}_B = \hat{E}[G_B] = \frac{\hat{E}[G_B + G_C] + \hat{E}[G_B - G_C]}{2} = \frac{6{,}5 + 0{,}5}{2} = 3{,}5$$

$$\text{e } \hat{\lambda}_C = \hat{E}[G_C] = \frac{\hat{E}[G_B + G_C] - \hat{E}[G_B - G_C]}{2} = \frac{6{,}5 - 0{,}5}{2} = 3 \ .$$

Por fim, as probabilidades de cada time marcar uma dada quantidade de gols no próximo jogo são:

$$P(G_B = b) = \frac{e^{-3{,}5}(3{,}5)^b}{b!}$$

$$\text{e } P(G_C = c) = \frac{e^{-3}3^c}{c!} \ .$$

A obtenção das probabilidades *P*(*vitória de B*), *P*(*empate*) e *P*(*vitória de C*), contudo, não é simples e direta, uma vez que não existe uma fórmula fechada para tais valores. Dois artifícios são mais usualmente empregados para a obtenção de aproximações satisfatórias desses valores:

a) Distribuição de Skellam (1946): Se G_B e G_C são variáveis independentes com distribuições de Poisson com médias λ_B e λ_C, então a diferença $G_B - G_C$ segue uma distribuição de Skellam dada por:

$$P(G_B - G_C = d) = e^{-(\lambda_B + \lambda_C)} \left(\frac{\lambda_B}{\lambda_C} \right)^{d/2} I_{|d|} \left(2\sqrt{\lambda_B \lambda_C} \right),$$

onde $I_{|d|}$ é a função de Bessel modificada de ordem $|d|$.

Não existe uma fórmula fechada para a soma dessas probabilidades para todos valores positivos ou todos os valores negativos de *d*. Porém, escolher cuidadosamente um valor de *N* (que não precisa ser muito grande) e aproximar as probabilidades de vitória de cada time pelas somas

$$P(\text{vitória de } B) = \sum_{d=1}^{N} e^{-(\hat{\lambda}_B + \hat{\lambda}_C)} \left(\frac{\hat{\lambda}_B}{\hat{\lambda}_C} \right)^{d/2} I_{|d|} \left(2\sqrt{\hat{\lambda}_B \hat{\lambda}_C} \right)$$

$$\text{e } P(\text{vitória de } C) = \sum_{d=-N}^{-1} e^{-(\hat{\lambda}_B + \hat{\lambda}_C)} \left(\frac{\hat{\lambda}_B}{\hat{\lambda}_C} \right)^{d/2} I_{|d|} \left(2\sqrt{\hat{\lambda}_B \hat{\lambda}_C} \right),$$

enquanto a probabilidade de empate pode ser calculada de forma exata:

$$P(\text{empate}) = e^{-(\hat{\lambda}_B + \hat{\lambda}_C)} I_0 \left(2\sqrt{\hat{\lambda}_B \hat{\lambda}_C} \right).$$

b) Retângulo truncado: uma alternativa mais simples (por não exigir cálculos de funções de Bessel) consiste em limitar a análise às probabilidades situadas dentro do retângulo com vértices (0,0), (0,*N*), (*N*,0) e (*N*,*N*):

$P(G_B = 0, G_C = 0)$	$P(G_B = 1, G_C = 0)$	…	$P(G_B = N, G_C = 0)$
$P(G_B = 0, G_C = 1)$	$P(G_B = 1, G_C = 1)$	…	$P(G_B = N, G_C = 1)$
⋮	⋮	⋱	⋮
$P(G_B = 0, G_C = N)$	$P(G_B = 1, G_C = N)$	…	$P(G_i = N, G_j = N)$

Por argumento análogo ao utilizado para a Distribuição de Skellam, é possível escolher um valor *N* (relativamente baixo) e aproximar as probabilidades de empate, de vitória do Time *B* e de vitória do Time *C*, respectivamente pelas somas dos valores da diagonal (destacada em cinza), do triângulo superior e do triângulo inferior.

Assim, voltando ao campeonato hipotético as probabilidades de cada resultado possível do próximo jogo (aproximadas pela Distribuição de Skellam truncada ao intervalo [-20, 20]) são iguais a:

$$\begin{cases} P(\text{vitória de } B) = 0{,}498 \\ P(\text{empate}) = 0{,}157 \\ P(\text{vitória de } C) = 0{,}345 \end{cases}.$$

Por fim, o reflexo dessas probabilidades nas medidas de qualidade do modelo será:

i) Contribuição para a Medida de Confiabilidade:

* soma de 1 ao denominador da parcela referente ao intervalo [0,4 ; 0,5];
* soma de 1 ao respectivo numerador se o Time *B* vencer o jogo e 0 em caso contrário;

* soma de 1 ao denominador da parcela referente ao intervalo [0,1 ; 0,2];
* soma de 1 ao respectivo numerador se o jogo acabar empatado e 0 em caso contrário;

* soma de 1 ao denominador da parcela referente ao intervalo [0,3 ; 0,4];
* soma de 1 ao respectivo numerador se o Time *B* perder o jogo e 0 em caso contrário.

ii) Medida de DeFinetti:

* $DDF = (0{,}498-1)^2 + (0{,}157-0)^2 + (0{,}345-0)^2 = 0{,}396$ se o Time *B* vencer o jogo;

* $DDF = (0{,}498-0)^2 + (0{,}157-1)^2 + (0{,}345-0)^2 = 1{,}078$ se o jogo acabar empatado;

* $DDF = (0{,}498-0)^2 + (0{,}157-0)^2 + (0{,}345-1)^2 = 0{,}702$ se o Time *B* perder o jogo

5. Comentários Finais

5.1. Rankings paramétricos

Modelos que sejam minimamente consistentes (de acordo, por exemplo, com os critérios da "análise anterior" descrita na seção 3.1) permitem, além das probabilidades para os jogos futuros, a formação de rankings paramétricos dos times de um dado universo ou competição. Tomando-se como exemplo o modelo de Bradley-Terry, é evidente que

$$\pi_i > \pi_j \Rightarrow \frac{\pi_i}{\pi_i + \pi_j} > \frac{\pi_j}{\pi_i + \pi_j} \Rightarrow$$
$$\Rightarrow P(i \text{ derrotar } j) > P(j \text{ derrotar } i) \Rightarrow i \text{ é "melhor" que } j$$

e, conseqüentemente, que pode-se estabelecer um ranking técnico dos times em função de seus parâmetros π.

Semelhantemente, para o modelo do site Chance de Gol, é fácil perceber que

$$\beta_i > \beta_j \Rightarrow E[G_i - G_j] > 0 \Rightarrow$$
$$\Rightarrow E[G_i] > E[G_j] \Rightarrow$$
$$\Rightarrow P(G_i > G_j) > P(G_i < G_j) \Rightarrow i \text{ é "melhor" que } j$$

e, por conseguinte, que os times podem ser tecnicamente ordenados com base em seus parâmetros β.

Para outros modelos minimamente razoáveis, é igualmente possível identificar parâmetros a partir de cujos valores se possa ranquear os times. Decorre das particularidades de cada modelo, porém, que o ranking dos times de um dado universo pode variar conforme o modelo escolhido. Em alguns casos, inclusive, tal variação pode até parecer "contraditória" ou "anti-intuitiva", como se pode ver na seção a seguir.

5.2. Resultados *versus* placares

Modelos baseados em resultados (vitória, empate ou derrota) e modelos baseados em placares (1x0, 2x0, 2x1 etc.) tendem a "enxergar" as relações técnicas entre diferentes times de formas distintas, eventualmente valorizando aspectos contrastantes das informações históricas existentes.

Por um lado, modelos baseados em placares, ao discernirem entre vitórias por 1x0 e por 8x0 permitem uma "sintonia fina" na estimação das forças de times que tenham vencido (ou perdido para) um mesmo adversário. Por outro lado, contudo, modelos baseados em resultados parecem se aproximar mais do objetivo principal de um jogo (especialmente quando dentro de um campeonato), que é a vitória: 1x0 e 8x0 valem os mesmos três pontos e, sob esse ponto de vista, a diferença entre um empate e uma "vitória magra" é muito menor que a diferença entre uma "vitória magra" e uma goleada.

Para melhor ilustração desse fenômeno, considere-se os exemplos abaixo:

Exemplo 1 (melhor de cinco jogos entre dois times):

Time M 1x0 Time N
Time M 1x0 Time N
Time M 1x0 Time N
Time M 1x0 Time N

Time N 7x0 Time M

Exemplo 2 (campeonato entre seis times jogando todos contra todos):

Time X 1x0 Time W	Time Y 8x0 Time W
Time X 1x0 Time V	Time Y 8x0 Time V
Time X 1x0 Time U	Time Y 8x0 Time U
Time X 1x0 Time T	Time Y 8x0 Time T
Time X 1x0 Time S	Time Y 8x0 Time S

Time X 1x0 Time Y

No primeiro exemplo, um modelo baseado em resultados observará que o Time M somou quatro vitórias contra apenas uma do Time N e conseqüentemente considerará o Time M como "melhor" (i.e. mais bem ranqueado e favorito à vitória num hipotético confronto futuro) que o seu adversário, enquanto um modelo baseado em placares tenderá a ressaltar o "placar total" (Time N 7x4 Time M) e a eleger o Time N como "melhor" que o seu concorrente.

Analogamente, no segundo exemplo um modelo baseado em resultados perceberá que o Time X venceu todos os seus jogos e que o Time Y teve uma derrota (justamente contra o Time X) e apontará o Time X como "melhor" que o seu oponente. Um modelo baseado em placares, por outro lado, tenderá a valorizar as intensidades das vitórias do Time Y e conseqüentemente a identificá-lo como "melhor" que o seu rival.

Esse contraste pode suscitar diversos debates filosóficos (até mesmo entre defensores do futebol "pragmático" e adeptos do jogo "ofensivo") e provavelmente seria muito bem-vinda uma abordagem "intermediária" que possa valorizar a diferença entre um empate e uma vitória e, ao mesmo tempo, distinguir as vitórias por margens distintas de gols.

5.3. Áreas de estudo

Muita coisa ainda há a ser estudada no campo das previsões aplicadas a jogos de futebol e um dos temas em aberto é justamente a busca de um "modelo intermediário" que possa conciliar a importância do resultado com a dimensão do placar.

Outra área em que se pode investir é a formulação de modelos que de alguma forma levem em consideração as individualidades (jogadores) que integram um time, permitindo que fatores como desfalques e reforços possam ser inseridos no cálculo das probabilidades.

Por fim, um tópico mais desafiador e ainda pouco explorado é a busca de modelos que permitam confeccionar "rankings históricos" de forma a comparar times que atuaram em épocas diferentes. Berry *et al* (1999) elaboraram um modelo para competições individuais, mas a sua expansão a competições entre equipes está longe de ser imediata.

6. Referências Bibliográficas

6.1. Livros e periódicos

ARRUDA, Marcelo L. (2000). *Poisson, Bayes, Futebol e DeFinetti,* São Paulo, IME-USP (Dissertação de Mestrado).

BERRY, Scott M., REESE, C. Shane, LARKEY, Patrick D. (1999). **Bridging Different Eras in Sports.** *Journal of the American Statistical Association 447 (94), 661-676.*

BRADLEY, Ralph A. e TERRY, Milton E. (1952). **The rank analysis of incomplete block designs.** *Biometrika 39, 324-345*.

DeFINETTI, Bruno (1972), *Probability, Induction and Statistics*, London: John Wiley.

DRAPER, Norman R. e SMITH, Harry (1998), *Applied Regression Analysis*, London: John Wiley.

ELO, Arpad E. (1978). ***The rating of chess players, past and present.*** Arco Publishing, New York.

FINK, Daniel (1997), *A Compendium of Conjugate Priors*. Technical Report: Montana State University.

GLICKMAN, Mark E. (1993). *Paired Comparison Models with Time-Varying Parameters.* Cambridge: Harvard University, Department of Statistics (Tese de Doutorado).

GUMBEL, Emil J. (1961). **Sommes et différences de valeurs extremes indépendantes** *Comptes Rendus Academic de Sciences 253:* 2838-2839.

HOLGATE, Philip (1964), **Estimation for the Bivariate Poisson Distribution**, *Biometrika* 51, 241-245.

LEE, Alan J. (1997), **Modeling Scores in the Premier League: Is Manchester United Really the Best?**, *Chance* 10 (1), 15-19.

McCULLAGH, Peter e NELDER, John A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

MURPHY, Allan H. (1972), **Scalar and Vector Partitions of the Probability Score: Part I: Two-State Situation**, *Journal of Applied Meteorology 11*, 273-282.

O'HAGAN, Anthony (1994), *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*, London: Edward Arnold Halsted Press.

SKELLAM, John G. (1946) **The frequency distribution of the difference between two Poisson variates belonging to different populations.** *JRSS A 109*, 296.

SOARES, José F. (1982), **Chances de vitória em uma partida de futebol**, *Atas do Sinape*, 195-198.

SOARES, José F. (1998), **Winning Odds in a Soccer Match**, *Departamento de Estatística, UFMG.*

SUZUKI, Adriano K., SALASAR, Luís E. B., LEITE, José G e LOUZADA-NETO, Francisco (2010), **Predicting 2010 Football World Cup via a bayesian approach**, *Departamento de Estatística, Universidade Federal de São Carlos.*

6.2 Sites

[1] **FIDE Chess Ratings (http://ratings.fide.com/)**
[2] **World Football Elo Ratings (http://www.eloratings.net/)**
[3] **FIFA/Coca-Cola World Ranking (http://www.fifa.com/worldranking/index.html)**
[4] **Chance de Gol (http://www.chancdegol.com.br)**