Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Ocorrência de Gols em Jogos de Futebol via Cadeias de Markov em Tempo Contínuo

André Felipe Brusco

Orientador: Professor Antônio Eduardo Gomes

Brasília

2018

André Felipe Brusco

Ocorrência de Gols em Jogos de Futebol via Cadeias de Markov em Tempo Contínuo

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Professor Antônio Eduardo Gomes

Brasília

2018

André Felipe Brusco
Ocorrência de Gols em Jogos de Futebol via Cadeias de Markov em Tempo Contínuo/ André Felipe Brusco. – Brasília, 2018-
li p. : il. (algumas color.) ; 30 cm.

Orientador: Professor Antônio Eduardo Gomes

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação, 2018.

1. Futebol. 2. Cadeia de Markov. 3. Kaplan-Meier. 4. Exponencial. 5. Campeonato Brasileiro.

André Felipe Brusco

Ocorrência de Gols em Jogos de Futebol via Cadeias de Markov em Tempo Contínuo

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. Brasília, 25 de Junho de 2018:

Professor Antônio Eduardo Gomes
Orientador

Jhames Matos Sampaio
Membro da Banca

Eduardo Yoshio Nakano
Membro da Banca

Brasília
2018

Agradecimentos

Gostaria de dedicar este trabalho, primeiramente, a minha família, minha namorada e a meus amigos queridos. Não teria chegado onde estou sem eles. Agradeço também a meus docentes do Departamento de Matemática e Estatística, por tantos ensinamentos. Sou grato por estar ao lado de meus colegas de "sala", que me acompanharam nessa jornada.

Agradeço à ESTAT por tantos ensinamentos ao longo de meus dois anos e meio de participação. Também agradeço ao professor Antônio Eduardo, por me dar a oportunidade de aplicar um Trabalho de Conclusão de Curso em uma paixão minha, o futebol. Sou muito grato a todo o pessoal que compõe este departamento excepcional de Estatística da Universidade de Brasília, onde fui tão feliz em estar.

Por fim e mais importante, agradeço a Deus, não por me dar um caminho fácil, mas por sempre me dar forças. Desejo a todos longos dias e belas noites.

“Ka is a wheel. If we fail,
it put us again on
the same path.“
(Roland Deschain)

Resumo

O objetivo deste trabalho é estimar a ocorrência de gols em campeonatos de futebol por meio de Cadeias de Markov. Posteriormente, analisar a taxa de falha do tempo de saída de cada gol utilizando técnicas de Análise de Sobrevivência.

Palavras-chave: Futebol, Markov, Sobrevivência, Gol, Estatística.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

TRV	Teste de Razão de Verossimilhança
KM	Kaplan - Meier
GGama	Gama Generalizada
TTT	Tempo Total em Teste
EXP	Exponencial
CM	Cadeia de Markov
K-S	Kolmogorov - Smirnov
WEIB	Weibull

Sumário

Introdução

Com o passar do tempo, a estatística vem se desenvolvendo graças a capacidade de armazenamento e processamento de dados. Assim, sua aplicabilidade se estendeu para diversas áreas, entre elas, esportes em geral.

Como sabemos, o esporte mais difundido atualmente é o futebol. Sua utilização na estatística vem desde análises de dados históricos até modelos de previsão de resultados das partidas, por exemplo.

A principal competição nacional do país é o Campeonato Brasileiro de Futebol, mais conhecido como Brasileirão, que teve seu início na era dos pontos corridos no ano de 2003. Hoje em dia, o campeonato conta com quatro divisões: A, B, C e D. Em seu primeiro ano de novo regulamento (quando o formato dos pontos corridos foi implementado), o Brasileirão contava com 26 times participantes para a série A ou série de elite. O campeonato, atualmente com 20 times na série A, é divido em dois turnos. Ao longo destes dois turnos, todos os times se enfrentam uma vez dentro de casa(time mandante) e uma vez na casa do adversário (nesse caso o time será visitante).

Ao final de cada partida, o time vencedor receberá 3 pontos e o time perdedor 0 pontos. Em caso de empate, ambos os times receberão 1 ponto. É declarada campeã a equipe que acumular o maior número de pontos ao longo do campeonato.

Diferentemente de outro artigos propostos, como (MARTINS, 2014) e (JUNIOR, 2014), que visam prever os resultados das partidas, o objetivo desse trabalho de conclusão de curso é descrever justamente a intensidade da saída de placares em jogos de futebol, especificamente no campeonato Brasileiro (série A), na era dos pontos corridos. Veja a tabela a seguir:

Tabela 1 – Exemplificação Inicial

Ano	Jogo	Time	Fora	Minuto	Metade	Placar	Intervalo de Saída
2004	364	Santos	0	8	1	(1x0)	10
2004	364	Santos	0	18	1	(2x0)	24
2004	364	Fluminense	1	42	1	(2x1)	55+

Como iremos mostrar na seção metodológica deste relatório, a Tabela 1 será utilizada como base para a construção de nossos trabalhos. O estudo a ser apresentado utiliza Cadeias de Markov para modelar a ocorrência de gols a tempo contínuo (como vemos, nosso minuto de jogo) ao longo das partidas e, posteriormente, estimar não parametricamente o tempo de espera até o próximo gol (Intervalo de Saída para o próximo placar). Para casos em que o placar permanecer inalterado até o final do jogo, será utilizado o conceito de dados censurados à direita (tempo de 55+, visto na última linha).

Por fim, será feita uma modelagem paramétrica dos intervalos de tempos de saída dos placares e, assim, compararemos com a estimativa não paramétrica. Vejamos a seguir nosso referencial teórico.

1 Referencial Teórico

1.1 Processos Estocásticos

Definição 1: Um processo estocástico é uma família $Z = \{Z(s), s \in T\}$, onde, para cada $s \in T$, $Z(s)$ é uma variável aleatória. Em geral, quando $T \subset \mathbb{R}$, dizemos que $Z(s)$ é um processo estocástico a tempo contínuo.

1.1.1 Espaço de Estados

O conjunto de valores, individualmente assumidos por $Z(s)$, são chamados de *estados*. Assim, temos:

Definição 2: O conjunto de valores assumidos por nosso processo $Z(s)$ é denominado espaço de estados. Nosso espaço será denotado por ξ.

1.1.2 Cadeias de Markov

Definição 3: (Propriedade de Markov) Sejam nossos estados $\{i_1, i_2, \ldots, i_n, j\} \in \xi$. Para os tempos $0 < t_1 < t_2 < \ldots < t_n < t$, temos a seguinte propriedade de Markov, definida em (DURRETT, 2011):

$$P(Z_{s+t} = j \mid Z_s = i_n, Z_{s-t_{n-1}} = i_{n-1}, \ldots, Z_{s-t_1} = i_1) = P(Z_{s+t} = j \mid Z_s = i_n)$$

para $s > 0$.

Definição 4: Uma cadeia de Markov(CM) a tempo contínuo é um processo estocástico tal que, para $Z = \{Z(s), s \in T\}$,

1. $T \subset \mathbb{R}$,
2. ξ é *espaço de estados discreto*,
3. $Z(s)$ satifaz a *Propriedade de Markov*.

1.1.3 Probablidade de Transição de Estados:

Definimos a probabilidade de transição:

$$p_t(i,j) = P(Z_{t+s} = j \mid X_s = i)$$

com $t > 0, i, j \in \xi$.

1.2 Estimação Não Paramétrica

A partir da presença de dados censurados à direita, será utilizado para o trabalho o estimador não paramétrico de Kaplan-Meier(KAPLAN; MEIER, 1958) para obteção da função empírica de sobrevivência de determinado estado $j \in \xi$.

- **Estimador de Kaplan-Meier:**

  Para a obteção do estimador de Kaplan-Meier, vamos definir $t_{(k)}$ o k-ésimo valor não censurado, distinto e ordenado em nosso conjunto de tempos observados para o estado j. Definimos então nosso estimador:

$$\hat{S}(t)_{KMj} = \prod_{k:t_{(k)} \leq t} \left[1 - \frac{d_k}{n_k}\right]$$

  - d_k = número de acontecimentos de interesse no tempo t (inclusive) para o estado j;
  - n_k = número de tempos suscetíveis ao evento de interesse a partir de k para o estado j.

1.3 Estimação Paramétrica

Definimos $D_j = \{(t_x, \delta_x); x = 1, 2, \ldots, n_j\}$ o conjunto de dados referentes ao tempo de permanência no estado j. Isto é, se $\delta_x = 1$, t_x é observado. Alternativamente, se $\delta_x = 0$, t_x é um tempo censurado à direita.

- **Função de Verossimilhança:**

  Assim, através de nosso conjunto D_j definimos:

$$L(\theta) \propto \prod_{x=1}^{n_j} [f(t_x \mid \theta)]^{\delta_x} [S(t_x \mid \theta)]^{1-\delta_x} \tag{1.1}$$

  nossa função de verossimilhança.

- **Distribuição Exponencial:**

  Modelaremos parametricamente o tempo a partir da seguinte parametrização da distribuição exponencial:

$$f(t) = \frac{1}{\alpha}\mathrm{e}^{-\frac{t}{\alpha}}, \alpha > 0,\ t > 0. \tag{1.2}$$

  Se $U \sim Exponencial(\alpha)$, temos:

  1. $E[U] = \alpha$ e $Var[U] = \alpha^2$;
  2. $S(t) = e^{-\frac{t}{\alpha}}$, para $\alpha > 0,\ t > 0$;
  3. $\hat{\alpha}_{MVexp} = \dfrac{\sum_{x=1}^{n} t_x}{\sum_{x=1}^{n} \delta_x}$, o estimador de máxima verossimilhança do parâmetro α para o modelo Exponencial.

- **Distribuição Weibull:**

  Usaremos como alternativa para o não adequamento do modelo exponencial o modelo Weibull de parâmetros α e β, onde, α é dito parâmetro de escala e β é dito parâmetro de forma. Sua função de densidade é dada por:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, \alpha,\ \beta > 0,\ t > 0. \tag{1.3}$$

  Se $U \sim Weibull(\beta, \alpha)$, temos:

  1. $E[U] = \alpha\Gamma\left(1 + \frac{1}{\beta}\right)$ e $Var[U] = \alpha^2\left(\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right)$;
  2. $S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, t > 0$ para $\alpha, \beta > 0, t > 0$.
  3. $\hat{\alpha}_{MVweib} = \left(\dfrac{\sum_{x=1}^{n} t_x^{\beta}}{\sum_{x=1}^{n} \delta_x}\right)^{\frac{1}{\beta}}$, o estimador de máxima verossimilhança de α, em função de β, para o modelo Weibull.

  O estimador acima é calculado através da minimização não linear iterativa dos parâmetros da função de verossimilhança, (1.1). Note que, pelo item 3, construímos nossa função de tal forma que $L(\alpha, \beta) = L(\alpha(\beta), \beta)$.

1.4 Teste de Adequabilidade de Kolmogorov

Definição 5: O teste Kolmogorov-Smirnov(K-S) é utilizado para verificar se determinada amostra provém de uma população com distribução de probabilidade especificada. O teste K-S é baseado na diferença entre a função de distribuição acumulada teórica, nosso caso $\hat{S}_{KM}(t)$, e a função de distribuição acumulada empírica dos dados, nosso caso o estimado pelo modelo paramétrico.

Hipóteses: Foram formuladas as seguintes hipóteses:

H_0 : Função de sobrevivência de Kaplan-Meier é igual ao modelo paramétrico
H_A : Função de sobrevivência Kaplan-Meier não é igual ao modelo paramétrico

Estatística de Teste: A estatística de teste W é definida como o supremo das distâncias entre $\hat{S}_{KM}(t)$ e $\hat{S}_M(t)$. Ou seja:

$$W = sup_t|\hat{S}_{KM}(t) - \hat{S}_M(t)|$$

Decisão: Como, neste relatório, trabalharemos com o teste bilateral, rejeita-se H_0 a um nível γ de significância se W ultrapassa o valor do percentil de $(1-\gamma)$, indicado na Tabela A13 em (CONOVER, 1999, p. 547) .

1.5 Métodos Gráficos

Para nos auxiliar na escolha e verificação do ajuste dos modelos propostos, utiliraremos três métodos gráficos:

1.5.1 Grafico TTT

Inicialmente, iremos definir a função de risco $h(t)$. A função irá nos aulixar na construção do gráfico TTT.

1.5.1.1 Função de risco

Definição 6: Dizemos que a função de risco $h(t)$ é a probabilidade condicional de experimentar o evento de interesse no instante t, dado que o mesmo não tenha ocorrido previamente:

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},\ t \geq 0.$$

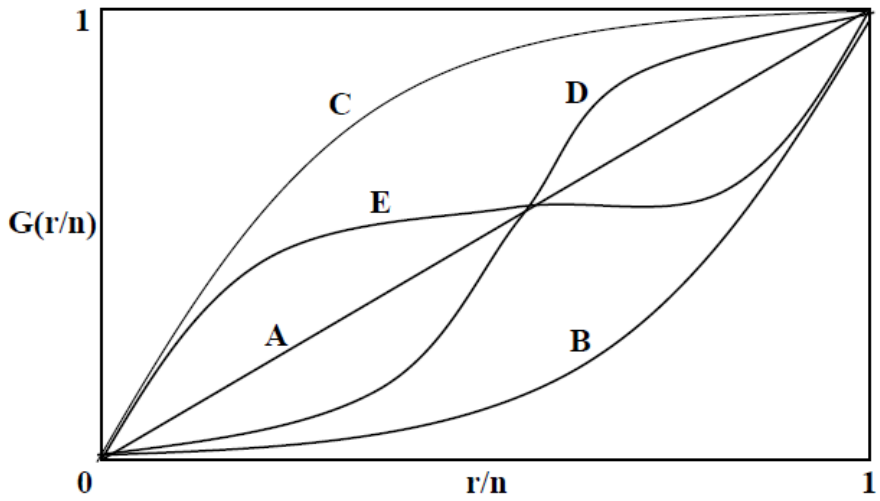
1.5.2 Determinação Empírica da forma da função de risco

Utilizaremos o Gráfico do Tempo Total em Teste (Curva TTT) para a determinação empírica da forma da função de risco. O gráfico é construído a partir das quantidades:

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{t} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{t} T_{i:n}} \quad \text{versus} \quad A = \frac{r}{n},$$

em que $r = 1, 2, \ldots$ e $T_{i:n}, i = 1, 2, \ldots$ são as estatísticas de ordem da amostra. Segue a figura abaixo, extraída da apostila (NAKANO, 2016)

Figura 1 – Formas para Gráfico TTT



em que,

- **(A)** indica uma função de risco constante, logo o modelo exponecial é adequado;
- **(B)** ou **(C)** indicam uma curva côncava ou convexa,respectivamente. Logo a função de risco é monotonicamente decrescente ou crescente e, assim, o modelo Weibull é o indicado;
- **(E)** indica uma curva côncava e posteriormente convexa, assim a função de risco é unimodal. Modelos como Log - Normal, Log - Logístico, Weibull são indicados;
- **(D)** indica função de risco em forma de U, caso inverso de **(E)**. Estudos apontam para ajustes de funções de risco multimodal, que não serão trabalhas neste relatório.

1.5.3 Sobrevivência Modelo Ajustado x Sobrevivência KM

Compara-se a função de sobrevivência ajustada pelos modelos propostos, $\hat{S}_W(t)$ ou $\hat{S}_{Exp}(t)$, com a estimativa não-paramétrica de Kaplan - Meier, $\hat{S}_{km}(t)$, através do gráfico $\hat{S}_{km}(t)$ X $\hat{S}_{Modelo}(t)$. Para um ajuste ideial, espera-se que o gráfico apresente forma linear.

1.5.4 Linearização da Função de Sobrevivência

O segundo método consiste na linearização da função de sobrevivência estimada por KM. Caso o pressuposto de linearidade seja violado, há indícios de não adequação do modelo estipulado.

- **Linearização Modelo Exponencial:** Para o modelo exponencial, será utilizado o gráfico de $-log[\hat{S}_{km}(t)]$ *versus* t, em que, a linearização da função de sobrevivência do modelo Exponencial é obtida pela equação (1.2).

- **Linearização Modelo Weibull:** Para o modelo Weibull, será utilizado o gráfico de $log\Big[-log\hat{S}_{km}(t)\Big]$ *versus* t. A linearização é obtida pela equação (1.3) .

1.6 Comparação de Modelos

Como as técnicas gráficas apresentadas anteriormente, embora úteis, têm caráter subjetivo, e K - S apresenta determinada limitação, também será realizado o Teste de Razão de Verossimilhança em modelos encaixados. O teste é realizado a partir de um modelo generalizado, tal que os modelos de interesse sejam casos particulares. Para tal, apresentaremos então o modelo Gama Generalizado:

- **Gama Generalizada:**

  Dizemos que se $U \sim GGama(\beta, \alpha, \sigma)$ para:

$$f(t) = \frac{\beta t^{\sigma-1}}{(\alpha/\sigma)^{\beta\sigma}\Gamma(\sigma)} t^{\beta(\sigma-1)} \exp -(\frac{\sigma\ t}{\alpha})^{\beta}, \tag{1.4}$$

  no qual, $\alpha, \beta, \sigma, t > 0$. Note que,

$$\sigma = 1 \Rightarrow\ U \sim\ Weibull(\beta, \alpha)$$
$$\sigma = 1,\ \beta = 1 \Rightarrow\ U \sim Exp(\alpha)$$

1.6.1 Teste da Razão de Verossimilhança

Definição 7: O teste da Razão de Verossimilhança, que pode ser visto com mais detalhes em (COLOSIMO, 2006), envolve a comparação dos valores do logarítmo da função de verossimilhança generalizada com o logarítmo do modelo proposto.

Utilizando as parametrizações (1.2), (1.3) e (1.4), para os modelos Exponencial, Weibull e GGama, respectivamente, definimos as seguintes hipóteses:

$\big(\mathbf{I}\big)$	$\big(\mathbf{II}\big)$	$\big(\mathbf{III}\big)$
$H_0 : \sigma = 1, \beta = 1$	$H_0 : \sigma = 1$	$H_0 : \sigma = 1, \beta = 1$
$H_A : \sigma = 1, \beta \neq 1$	$H_A : \sigma \neq 1$	$H_A : \sigma \neq 1, \beta \neq 1$

- $\big(\mathbf{I}\big)$: Weibull *versus* Exponencial;
- $\big(\mathbf{II}\big)$: Gama Generalizada *versus* Weibull;
- $\big(\mathbf{III}\big)$: Gama Generalizadas *versus* Exponencial.

Estatística de Teste: Seja θ_0 o vetor de parâmetros sob H_0, ou seja, o vetor de parâmetros do modelo generalizado e θ o vetor de parâmetros do modelo a ser ajustado. A estatística de teste do TRV, $\mathbf{R}$, é definida como:

$$R = 2\log\Big[L(\hat{\theta}) - L(\hat{\theta_0})\Big]$$

em que R possui distribuição χ^2_l, com l a diferença do número de parâmetros dos modelos comparados.

Decisão: H_0 é rejeitada a um nível γ de significância se $R > \chi^2_{l,1-\gamma}$.

2 Metodologia

2.1 Banco de Dados

Os dados foram extraídos através da utilização de expressões regulares e WebScraping em tabelas do Campeonato Brasileiro de Futebol, exclusivamente da série A, fornecidas pelo site *http://www.rsssfbrasil.com/*, na era dos pontos corridos.

Constam em nosso banco de dados as seguintes variáveis:

1. Ano do Campeonato;
2. Número de jogo do Campeonato;
3. Time que marcou gol;
4. Variável Dummy indicando se o time está jogando fora de casa (0 - não , 1 - sim);
5. Minuto da ocorrência do gol;
6. Metade do Jogo (primeira ou segunda metade);
7. Placar de jogo no instante que ocorreu o gol;
8. Intervalo de tempo de jogo da saída do placar atual para o próximo placar;

Vale ressaltar que o campeonato de futebol na forma de pontos corridos, no Brasil, teve início no ano de 2003. Porém, no site utilizado para obtenção dos dados, os anos de 2009, 2014 e 2017 estavam indisponíveis. Assim, foram coletados os resultados de 18.293 placares para os anos de $2003 - 2008, 2010 - 2013, 2015$ e 2016. Segue abaixo o cabeçalho do banco de dados construído:

Tabela 2 – Cabeçalho Banco de Dados

Ano	Jogo	Time	Fora	Minuto	Metade	Placar	Intervalo de Saída
2008	13	Figueirense	1	74	2	(0x1)	17
2008	13	Figueirense	1	91	2	(0x2)	6+
2004	24	Corinthians	0	10	1	(1x0)	48
2004	24	Corinthians	0	58	2	(1x0)	39+
2004	364	Santos	0	8	1	(1x0)	10
2004	364	Santos	0	18	1	(2x0)	24
2004	364	Fluminense	1	42	1	(2x1)	55+

Vamos supor que o jogo 13 no ano de 2008, observado na Tabela 2, seja Figueirense *versus* São Paulo. Note que, aos 74 minutos na segunda metade da partida, o time Figueirense, no caso visitante, abriu o placar para (0x1). Posteriormente, aos 91 minutos, o

time Figueirense aumentou o placar para (0x2). Note que, o instante do acontecimento do primeiro gol até a saída para o placar (0x2), 17 minutos, é nosso intervalo de tempo para a saída do placar. Note ainda que, caso todos os jogos do campeonato fossem acrescidos de 7 minutos na segunda metade da partida e em 0 minutos na primeira metade, o nosso conjunto de tempo de jogo seria de 97 minutos. Assim, ao encerrar a partida no minuto 97, o intervalo de tempo para a saída do placar (0x2) seria de 6 minutos, porém, censurados. Ou ainda, 6+.

O parágrafo acima exemplifica a metodologia aplicada para contabilizar os intervalos de tempo de jogo. Foi utilizado como tempo de jogo, nosso conjunto T, os 90 minutos mais o maior tempo acrescido ao longo dos campeonatos coletados. Ou seja, se nos anos observados o maior acréscimo dado foi de 7 minutos, temos $T = [0, 97]$.

É verdade que, se o time Figueirense marcasse seu segundo gol aos $45 + 2$, ou seja, no minuto extra 2 do primeiro tempo e, porventura, o time São Paulo descontasse o placar aos 45 minutos da segunda metade, o intervalo de tempo para a saída do placar (0x2) para o placar (1x2) seria negativo em duas unidades, e portanto, inconsistente. Contudo, ao longo dos 12 anos de campeonato Brasileiro observados, apenas um caso foi encontrado e removido da base. Assim, não houve a necessidade, para fins de cálculo, da diferenciação de gols ocorridos nos acréscimos da primeira metade e de gols no início da segunda etapa.

2.2 Espaço de Estados

Vamos considerar um possível placar (i, j) do nosso espaço de estados, em que i irá representar o número de gols do time jogando em casa e j o número de gols do time visitante. Nosso espaço de estados de possíveis placares é dado por $\xi = \mathbb{N}^2 = \mathbb{N}\text{x}\mathbb{N} = \{(i, j) : i, j \in \mathbb{N}\}$ com $\mathbb{N} = \{0, 1, 2, \ldots\}$.

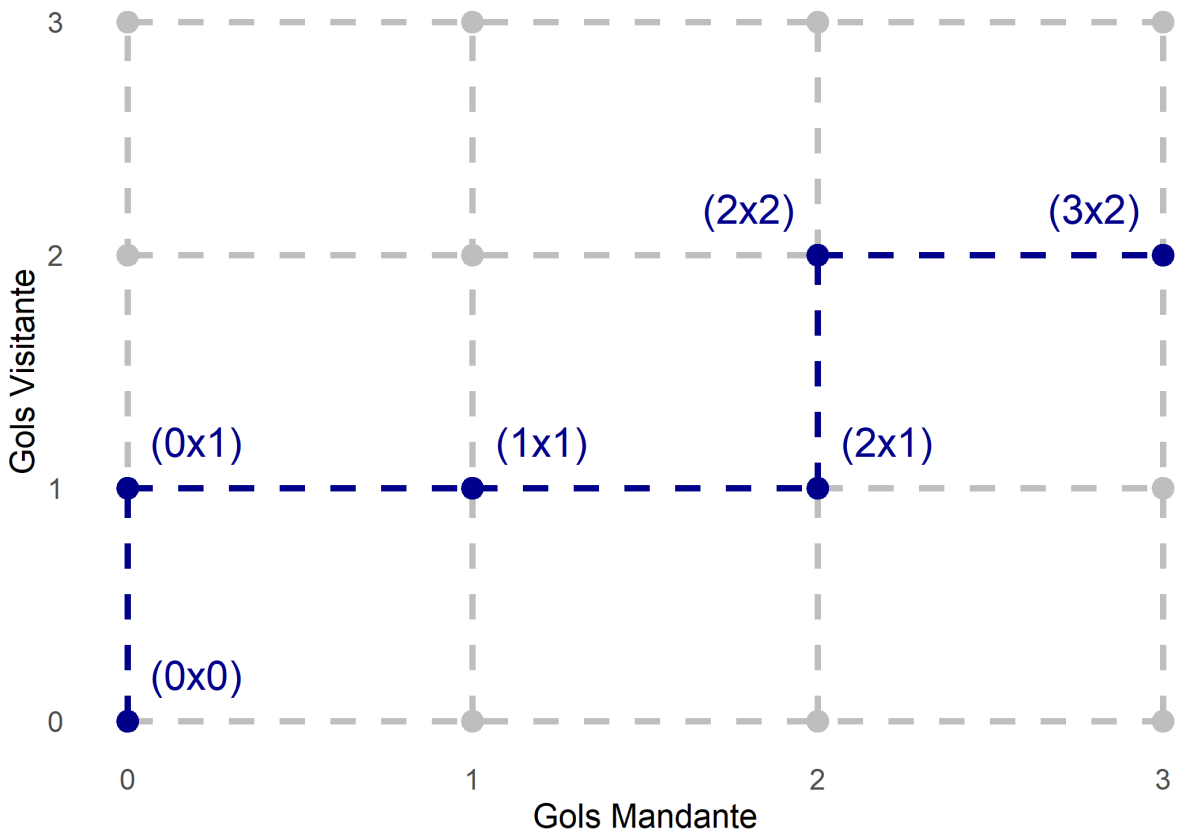
2.3 Transição entre Placares

Dado como conhecimento prévio que placares de futebol contabilizam apenas um gol para o time visitante ou um gol para o time mandante por vez, temos que nossas possíveis transições entre estado são:

$$(i, j) \to (i, j + 1)$$
$$\text{ou}$$
$$(i, j) \to (i + 1, j)$$

Veja a figura abaixo:

Figura 2 – Possível trajetória do placar $(0, 0)$ para o placar $(3, 2)$



Para ilustrar o que foi dito anteriormente, observe na Figura 2 uma possível trajetória do estado (0x0) para o estado (3x2) (destacada em azul), dentre todas as possíveis trajetórias (destacadas em cinza).

2.4 Probabilidade de Transição de Placares

Seja T o conjunto do tempo do jogo. Como dito previamente, como o maior tempo de acréscimo observado em todas as competições foi de 7 minutos, temos $T = [0, 97]$. Definimos então, a probabilidade de transição entre estados:

$$p_t(a, b) = P(X_{s+t} = b \mid X_s = a), t > 0, a = (i, j), b = (k, l) \ \in \ \xi$$

Neste caso, t é o intervalo de tempo de jogo da saída do estado a para o estado b (variável 8 do banco de dados) e s é o tempo de jogo transcorrido (variável 5 do banco de dados). Para o cálculo estimado das probabilidades de transição do estado (i, j) para

o estado $(i+1, j)$, foi utilizado o número de placares que saíram do estado (i, j) para o placar $(i+1, j)$ sobre o total de placares que deixaram (i, j), ou seja:

$$\hat{P}\Big((i,j) \to (i+1,j)\Big) = \frac{n\Big[(i,j),(i+1,j)\Big]}{n\Big[(i,j),(i+1,j)\Big] + n\Big[(i,j),(i,j+1)\Big]} \tag{2.1}$$

Analogamente, para o cálculo da probabilidade de transição de (i, j) para $(i, j+1)$, temos:

$$\hat{P}\Big((i,j) \to (i,j+1)\Big) = \frac{n\Big[(i,j),(i,j+1)\Big]}{n\Big[(i,j),(i+1,j)\Big] + n\Big[(i,j),(i,j+1)\Big]} \tag{2.2}$$

no qual, $n[a, b]$ denota a frequência de placares que saíram de a e chegaram em b.

2.5 Estimação com tempos Censurados à direita

Como exemplificado na Seção 3.1, quando o placar (i, j) permaner inalterado até o final da partida, consideraremos o tempo t^+ dado pela diferença entre o tempo final do jogo (em nosso caso 97 minutos) menos o tempo de chegada no placar (i, j) como tempo de censura à direita. Para o caso particular $(i, j) = (0, 0)$, os tempos censurados à direita foram dados pelos tempos finais de jogo.

Assim, como dito previamente no referencial teórico, definimos o nosso conjunto $D_a = \{(t_x, \delta_x); x = 1, 2, \ldots, n_a\}$ como o conjunto de dados referentes ao tempo de permanência no placar a. Então, através de nosso conjunto D_a, estimaremos através de Kaplan-Meier a função de sobrevivência para o placar $a = (i, j)$ ao longo do tempo de jogo.

$$\hat{S}_{KMa} = \prod_{k:t_k \leq t} \left[1 - \frac{d_{k_a}}{n_{k_a}}\right]$$

- d_{j_a} = número de saídas do placar a no tempo j (inclusive);
- n_{j_a} = número de tempos suscetíveis a saída de a a partir de j.

2.6 Escolha de Placares para Modelagem Paramétrica

Ao longo dos 12 anos de campeonato coletados, foram observados 50 placares, ou seja, $\#\xi = 50$. Como sabemos, jogos com grande número de gols não são recorrentes, e assim, suas estimativas paramétricas de tempo de permanência no placar não seriam satisfatórias. Por outro lado, muitos tempos de censura podem, porventura, limitar a

estimação do modelo. Portanto, foram escolhidos escores com mais de 60 observações, nos quais, pelo menos 40% dos tempos coletados não sejam tempos de censura.

2.7 Modelagem Paramétrica

Por fim, modelaremos parametricamente o tempo de permanência de jogo em cada placar a escolhido ao longo dos campeonatos via:

- **Modelo Exponencial:** [Equação (1.2)]

$$f_a(t) = \frac{1}{\alpha_a} \mathrm{e}^{-\frac{t}{\alpha_a}}, \quad t,\ \alpha_a > 0$$

  Dizemos que α_a é o tempo médio de permanência no placar a. Ainda, para fins de interpretação, dizemos que α_a^{-1} é a taxa de saída de a.

- **Modelo Weibull:** [Equação (1.3)]

$$f_a(t) = \left(\frac{\beta_a}{\alpha_a}\right)\left(\frac{t}{\alpha_a}\right)^{\beta_a - 1} \exp\left\{ -\left(\frac{t}{\alpha_a}\right)^{\beta_a}\right\}, t > 0$$

  Dizemos que α_a é o tempo médio de permanência em a. Quando $\beta_a > 1$, temos a taxa da saída do placar estritamente crescente ao longo do tempo. Em contrapartida, quando $\beta_a < 1$, temos a taxa da ocorrência do placar estritamente decrescente ao longo do tempo. Por fim, quando $\beta_a = 1$, dizemos que a taxa de saída do placar é constante, e assim, temos o modelo exponencial.

2.7.1 Taxa de Saída de Gol

Inicialmente, para cada placar $a = (i, j) \in \xi$ que atende os critérios estipulados na Seção 3.6, foram estimados os parâmetros α_{exp} para o modelo exponencial, utilizando $\hat{\alpha}_{MVexp}$, e os parâmetros $\beta_{weib}, \alpha_{weib}$ do modelo Weibull através da otmização não-linear iterativa de $\hat{\beta}_{MVweib}$ e $\hat{\alpha}_{MVweib}$.

2.7.2 Escolha dos Modelos

Utilizaremos, por fim, para analisar os resultados encontrados:

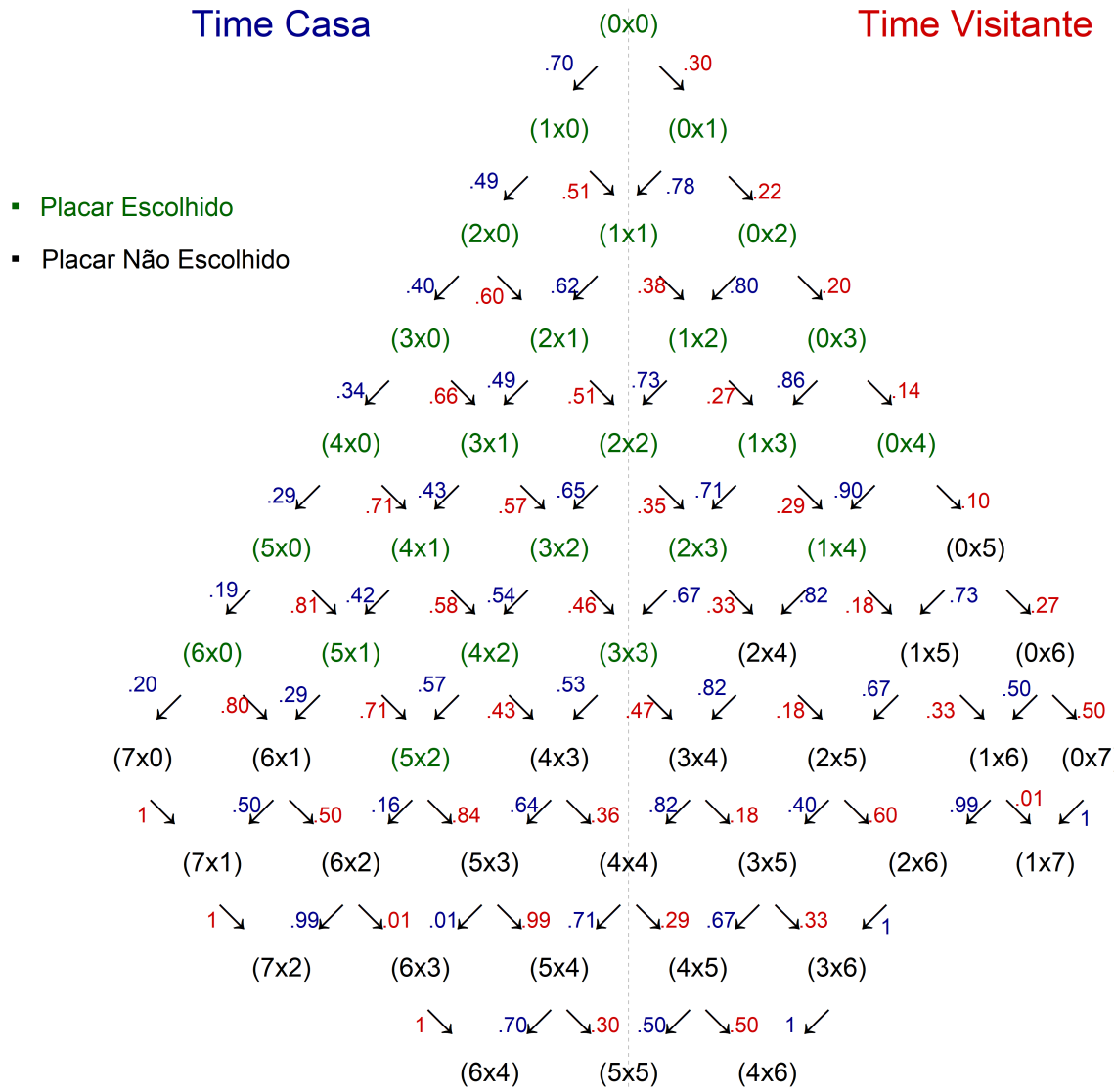
- O **Gráfico TTT** para uma ideia inicial de qual modelo pode vir a ser escolhido;
- **TRV** para assisitir na escolha do modelo mais adequado (Exponencial, Weibull, Modelo Generalizado) para cada placar escolhido;
- O teste **K - S** para verificar o ajuste do modelo paramétrico em comparação com a função de sobrevivência de KM;
- Demais técnicas gráficas apresentadas para nos auxiliar visualmente no ajuste do modelo.

3 Resultados

3.1 Probabilidade de Transição entre Placares

A partir das equações (3.1) e (3.2), foram calculadas as probabilidades da saída de cada placar. Para nosso trabalho, iremos denominar a Figura 3 a seguir como nossa **Árvore de Transição de Placares**. (DURRETT, 2011) denota a Figura 3 como a **Matriz de Rotas** dos estados.

Figura 3 – Probabilidade de Transição entre Placares



Pode-se notar que a probabilidade do time mandante abrir o marcador é maior que a do time visitante. Por outro lado, em situações de ampla vantagem para qualquer lado, a probabilidade do time em desvantagem descontar o marcador é maior. Tal fato, intuitivamente, pode ser explicado pela provável garantia dos *3 pontos* no campeonato, levando o time que está à frente no marcador a um certo comodismo.

Outro ponto interessante é que placares empatados "favorecem"times mandantes. Com excessão de (3x3), que aparenta certo equilíbro, para todos os placares $(i, j) = \{(i, j) \in \xi \ : i = j\}$ escolhidos, $\hat{P}\Big((i, j) \to (i + 1, j)\Big) > \hat{P}\Big((i, j) \to (i, j + 1)\Big)$. Um motivo plausível para tal evento é o fator *casa*. Times com mando de jogo normalmente desfrutam de apoio extra vindo das arquibancadas.

Veja que, para placares ao final e aos extremos de nossa Árvore de Transição, há apenas a probabilidade de transição para um determinado estado. Resultados como (7x0), (7x1),(0x7) apresentam probabilidade de 100% de transição para apenas um resultado. É compreensível imaginar que resultados com grandes diferenças de gols de um time para o outro ocorrem já ao final da partida, quando ocorrem. Assim, não há mais tempo de ampliar o marcador, ou, quando há, ambos os times estão conformados com o cenário final do jogo.

Ressaltamos também que placares com ausência de transição para outros resultados não foram estimados por apenas apresentarem tempos de censura, isto é, não houve, nos anos coletados do Campeonato Brasileiro de Futebol, tempo hábil para os times ampliarem ou descontarem placares como (6x4) ou (1x7)), ou decidirem o jogo após chegarem ao estado (5x5).

3.2 Modelagem dos Placares

Vamos apresentar agora os resultados relativos a modelagem paramétrica do tempo de permanência de cada placar. Como dito anteriormente, a presença de censuras e ausência de dados podem ser um problema na estimação dos parâmetros e avaliação do modelo. Portanto, a partir dos preceitos expostos na Seção 3.6, escolhemos os placares em verde na Figura 3 para serem modelados. Veja a Tabela 3 abaixo:

Tabela 3 – Resultados Modelagem Paramétrica

Placar	Exponencial	Weibull		**III**	**II**	**I**	P-valor K-S Teste	
	$\hat{\alpha}_{exp}$	$\hat{\beta}_{weib}$	$\hat{\mu}_{weib}(t)$	P-valor	P-valor	P-valor	Exponencial	Weibull
(0x0)	40,34	1,16	39,58	1,00	0,98	0,84	0,63	1,00
(1x0)	35,34	1,15	34,12	0,99	0,95	0,06	1,00	0,55
(2x0)	29,56	1,12	28,42	0,99	0,92	0,05	0,93	0,18
(3x0)	28,19	1,20	26,01	0,99	0,91	0,09	0,99	0,52
(4x0)	22,60	1,14	21,41	0,99	0,87	0,10	0,32	0,99
(5x0)	21,34	1,06	20,73	0,99	0,87	0,97	1,00	0,98
(6x0)	30,86	1,65	22,37	0,96	0,95	0,21	0,34	0,61
(1x1)	40,10	1,03	39,43	0,99	0,91	0,17	0,06	0,002
(1x2)	37,71	1,17	34,2	0,99	0,93	0,07	0,77	0,61
(2x2)	36,36	1,11	33,2	0,99	0,90	0,05	0,16	0,03
(2x1)	40,73	1,02	40,31	0,99	0,89	0,10	0,00	0,00
(3x1)	35,25	0,95	36,7	0,98	0,85	0,02	0,00	0,00
(4x1)	38,13	0,89	42,37	0,98	0,87	1,00	0,10	0,16
(5x1)	28,70	1,06	27,67	0,99	0,92	1,00	0,97	0,97
(3x2)	42,27	1,03	41,41	0,99	0,88	0,006	0,41	0,40
(4x2)	45,37	0,96	47,35	0,99	0,86	1,00	0,21	0,21
(5x2)	41,00	0,97	46,01	1,00	0,95	0,01	0,99	0,99
(3x3)	31,90	1,05	30,38	1,00	0,89	0,98	0,96	0,96
(0x1)	36,17	1,09	35,26	0,99	0,94	0,04	1,00	0,67
(0x2)	33,28	1,10	31,8	0,99	0,91	0,05	0,88	0,47
(0x3)	34,21	1,91	28,37	0,99	0,91	0,92	0,77	0,58
(0x4)	37,31	1,34	29,31	0,96	0,82	0,87	0,45	0,77
(1x3)	36,50	1,17	32,25	0,99	0,93	0,93	0,94	0,07
(1x4)	41,11	1,11	36,46	0,99	0,99	0,95	1,00	0,99
(2x3)	45,96	1,05	43,33	0,98	0,87	0,97	0,08	0,24

Tempo de Permanência no Placar:

Como sabemos, a estimativa do tempo médio de permanência do placar para o modelo Exponencial, $\mu_{exp}(t)$, é dada pela nossa estimativa do nosso parâmetro de escala, $\hat{\alpha}_{exp}$. Já para nosso modelo Weibull, nossa estimativa para o tempo médio de saída do placar, $\mu_{weib}(t)$, é dada pela expressão apresentada no item 1, no subtópico do modelo Weibull da Seção 1.3.

Analisando primeiramente as estimativas de nosso tempo médio de permanência nos placares, observe que placares iniciais como (0x0),(1x0) e (0x1) apresentam tempo médio de permanência próximos ou maiores que 35 minutos . Tal fato pode ser explicado por jogos amarrados, nos quais ambos os times têm dificuldade para abrir o placar, ou quando conseguem abri-lo, decidem segurar o resultado até o final da partida.

Novamente, ao analisarmos as estimativas da Tabela 3, note que os placares em que o time mandante está na frente e que o visitante não marcou têm menor tempo médio de permanência que os demais. Este evento pode explicar o evento chamado *Goleada*, que mostra determinado embalo do time mandante para marcar consecutivas vezes em pouco tempo. Porém, como analisado na seção anterior, ao final da partida, para placares de larga vantagem, a probabilidade do time visitante descontar o marcador é maior. Este evento pode ser visto no placar (5x0), cujo tempo médio de permanência é o menor estimado.

Sob outra perspectiva, observe que, para placares que o time visitante amplia seu marcador consecutivamente, o tempo de permanência no placar aumenta a medida que o resultado se alarga, o que mostra a dificuldade do time em aumentar sua vantagem fora de casa e, possivelmente, da desistência do time mandante em buscar o resultado sob o olhar de sua torcida.

Outro ponto interessante é que o placar (1x1), mesmo sendo de baixa escala, apresenta tempo de permanência alto, para ambos os modelos, de aproximadamente 40 minutos. Porém, uma explicação plausível seria pelo fato que (1x1) é o segundo placar com mais tempos de censura, isto é, o segundo placar, entre todos os coletados, que as partidas mais terminaram, perdendo apenas para (1x0).

Além do mais, as estimativas dos nossos parâmetros de forma para os placares (1x0), (2x0),(3x0),(4x0) são todas maiores que uma unidade. Tal fato mostra que, pelo modelo Weibull, nos placares apresentam taxa de saída crescente ao longo da partida, reforçando o conceito de determinado embalo do time. Alternativamente, as estimativas dos parâmetros de forma para os placares (0x3), (0x4), somadas ao conhecimento da probabilidade alta de transição destes placares para os estados (1x3) e (1x4), respectivamente, podem reforçar o conceito do time em casa vir a descontar o placar, dado o apoio de seus torcedores.

Por fim, note que os placares (4x2) e (5x2) (para o modelo Weibull) apresentam tempo de permanência maior que 45 minutos, ou seja, se o placar aconte na primeira metade da partida ou no início da segunda etapa, em média, o jogo permanecerá nele.

Estimativa Exponencial x Weibull:

Em geral, pode-se dizer que as estimativas apresentadas para os modelos Weibull e Exponencial não aparentam grandes divergências. Comparativamente, apenas os placares (6x0), (5x2),(0x3),(0x4), (1x3) e (1x4) apresentaram grandes mudanças nas estimativas do parâmetro de escala e parâmetro de forma diferentes de um, e consequentemente, no tempo de permanência no placar.

Para nosso modelo Weibull, apenas os placares (0x3), (0x4) e (6x0) apresentaram estimativas que levam a considerar a hipótese de que a taxa de saída destes placares é crescentes ao longo do jogo. Novamente, associamos com a idéia anterior que, ao chegar ao final da partida, o aumento da saída do placar é crescente dado o desconto do time perdedor ou o embalo do time vencedor.

TRV:

Observe que, para todos os placares, o **TRV**, sob as (III) e (II), rejeitou a hipótese da utilização de um modelo generalizado, a um nível $\gamma = 5\%$ de confiança. Por outro lado, para os placares (3x1), (3x2), (5x2),(0x1) e possivelmente o placar (2x0) (p-valor exato de $0,05$), o teste, sob (I), também a um nível $\gamma = 5\%$ de confiança, rejeitou a hipótese de tempo médio constante de permanência ao longo do jogo.

K-S:

Também a um nível de confiança de $\gamma = 5\%$, nosso teste rejeita os dois modelos para os placares (2x1) e (3x1), e rejeita o modelo Weibul para os empates (1x1) e (2x2). Em contrapartida, para os placares (0x0), (4x0), (6x0) e (2x3), houve uma melhora significativa do p-valor ao utilizarmos o modelo Weibull. Levaremos tal informação em consideração para o modelo a ser escolhido, dado que, como dito previamente no Referêncial Teórico, o teste K-S apresenta problemas de sensibilidade nas caudas. Para os demais placares, nosso teste aceitou o ajuste para ambos os modelos.

Após as análises apresentadas, veremos a seguir os resultados via técnicas gráficas dos modelos. Caso apresentássemos as quatro técnicas para cada um dos 25 placares escolhidos, teríamos neste relatório um total de cem gráficos. A fim de otimizar a apresentação deste documento, foram escolhidos para o modelo Exponencial um placar inicial, (1x0), um placar em que há ampla vantagem do time mandante, (4x0), um placar de pequena vantagem do time mandante, (2x1), um placar de empate, (3x3), e um placar que o time visitante está na frente, (0x1).

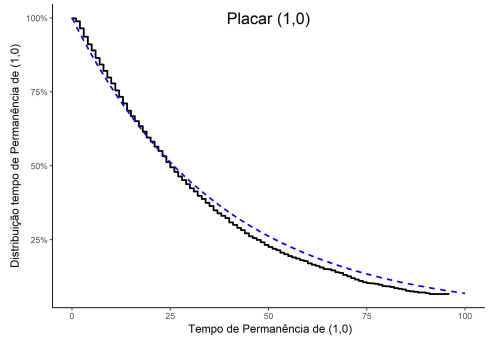
Já para o modelo Weibull, foram escolhidos os placares que apresentaram aspectos de melhora no modelo, em comparação ao Exponencial, ou taxa de saída crescente ao longo da partida.

Por fim, concluiremos os resultados e comentaremos algumas das características dos placares cujo modelo Generalizado aparenta ser mais adequado, dado que ambos os modelos anteriores não apresentaram bom ajuste.

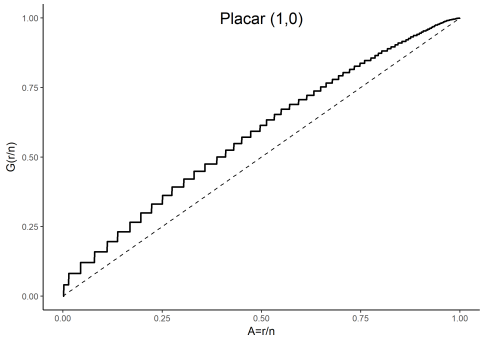
3.2.1 Modelos Exponencial

Figura 4 – Placar (1x0)

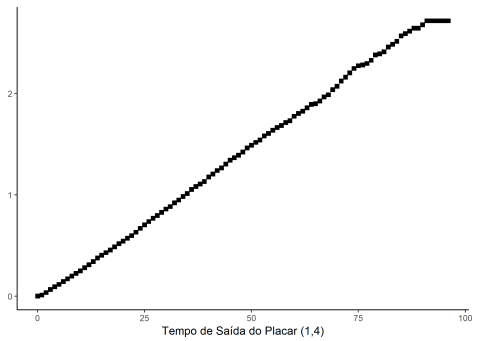
(**a**) KM e Ajuste



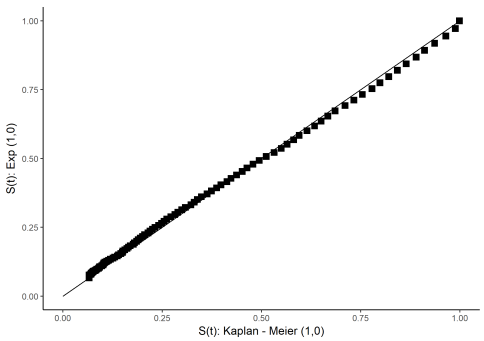
(**b**) Gráfico TTT



(**c**) Linearização Modelo Exponencial



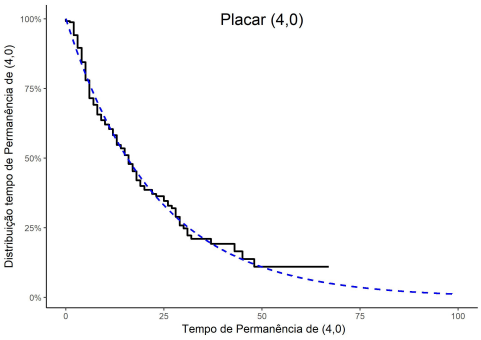
(**d**) $\hat{S}_{exp}(t)$ X $\hat{S}_{km}(t)$



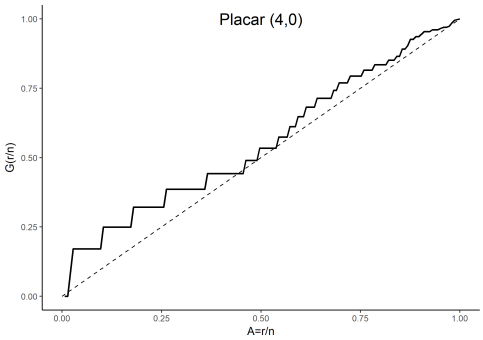
Veja que, a partir de nosso Gráfico TTT, o modelo Weibull seria indicado. No entanto, o ajuste visual sobre a função de sobrevivência parece adequado. Os gráficos **(c)** e **(d)** apresentam aspecto linear. Dizemos, associados aos resultados anteriores, um bom ajuste utilizando o modelo proposto.

Figura 5 – Placar (4x0)

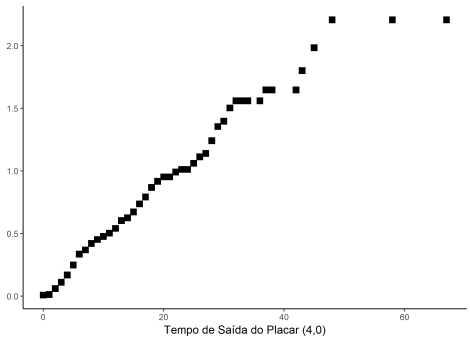
(**a**) KM e Ajuste



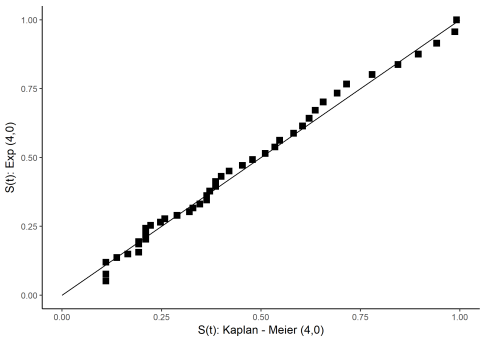
(**b**) Gráfico TTT



(**c**) Linearização Modelo Exponencial



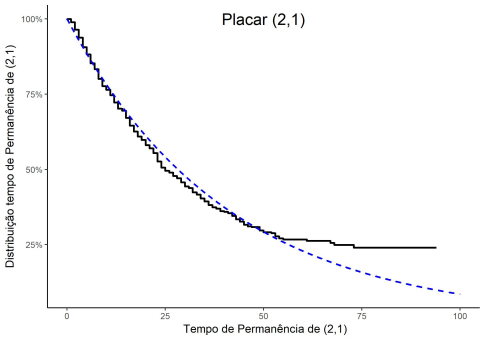
(**d**) $\hat{S}_{exp}(t)$ X $\hat{S}_{km}(t)$



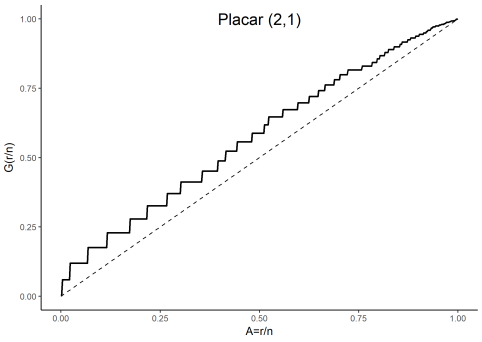
Olhando para o Gráfico TTT, tanto o modelo EXP quanto o modelo WEIB parecem adequados para modelar a taxa do placar (4x0). O gráfico **(d)** aparenta forma linear, porém, o modelo aparenta leve desajuste em sua cauda, evidenciado em **(c)**. Em geral, ao olharmos para seu ajuste em **(a)**, o modelo Exponencial aparenta ser adequado, com um leve desajuste em sua cauda. Ao olharmos na Tabela 2, vemos que o **TRV** sob a hipótese (I) aceitou a hipótese do modelo EXP. Ainda, ao realizamos o teste K-S, o p-valor do modelo WEIB é menor que o do modelo atual. Assim, dizemos que o ajuste através do modelo exponencial parece adequado.

Figura 6 – Placar (2x1)

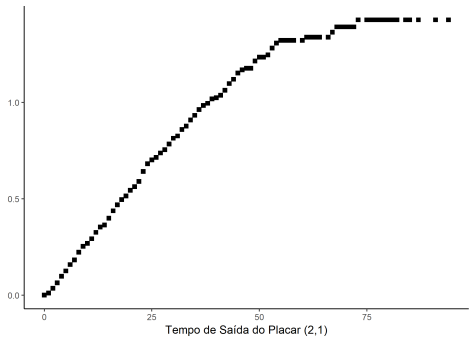
(**a**) KM e Ajuste



(**b**) Gráfico TTT



(**c**) Linearização Modelo Exponencial



(**d**) $\hat{S}_{exp}(t)$ X $\hat{S}_{km}(t)$

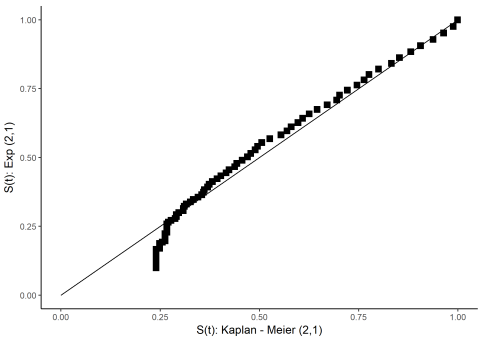
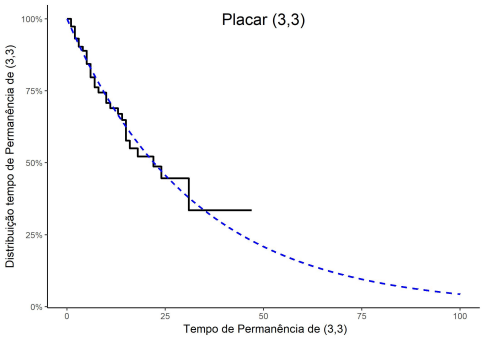
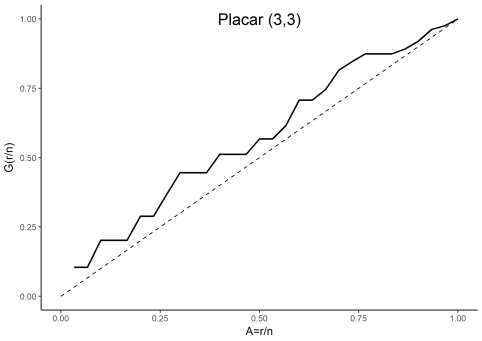


Figura 7 – Placar (3x3)

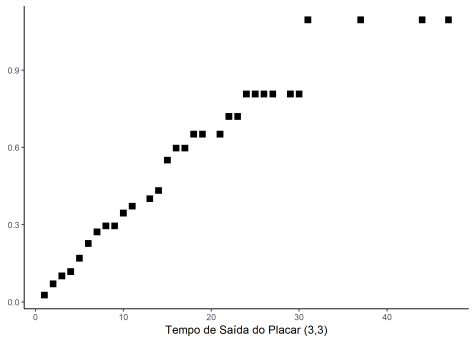
(**a**) KM e Ajuste



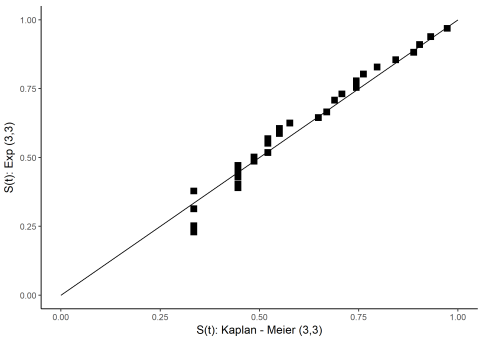
(**b**) Gráfico TTT



(**c**) Linearização Modelo Exponencial



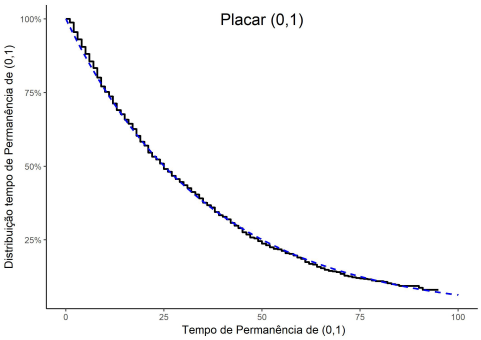
(**d**) $\hat{S}_{exp}(t)$ X $\hat{S}_{km}(t)$



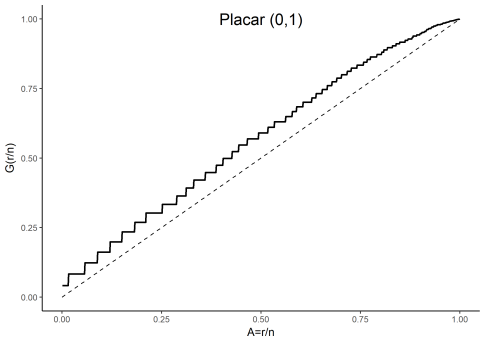
Ambos os placares, através de **(b)**, aparentam estar adequadamente ajustados por EXP ou WEIB, porém, principalmente para o placar (2x1), o modelo Exponencial não está bem ajustado para a cauda da função de sobrevivência de KM. Tal desajuste não está tão evidente em (3x3), talvez, pela pouca presença de censuras no placar. Também, para o placar (2x1), o teste K-S rejeitou a hipótese de adequabilidade de ambos os modelos. Assim, concluímos que para ambos os placares o modelo não parece adequado.

Figura 8 – Placar (0x1)

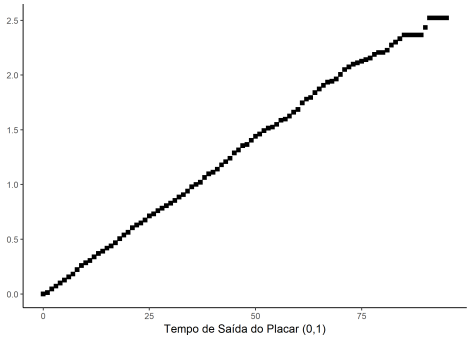
(**a**) KM e Ajuste



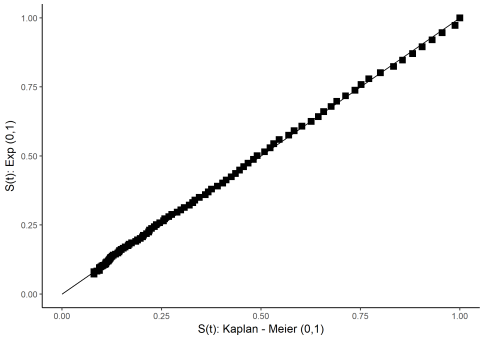
(**b**) Gráfico TTT



(**c**) Linearização Modelo Exponencial



(**d**) $\hat{S}_{exp}(t)$ X $\hat{S}_{km}(t)$



Para o placar (0x1), a forma da curva TTT mostra que o modelo WEIB parece adequado. Analisando **(c)** e **(d)**, observamos uma aparente forma linear. Através da modelagem paramétrica vista em **(a)**, associados aos resultados encontrados anteriormente na Tabela 2, dizemos que o modelo EXP é dito satisfatório para este marcador.

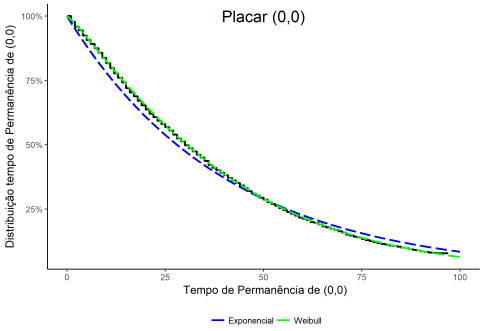
Assim, concluímos as análises para os cinco placares selecionados. Em geral, os maiores problemas encontrados foram de superestimação da função de sobrevivência, que pode ser vista de forma sutil na Figura 4.**(a)**, e principalmente desajustes em relação a cauda da função empírica de KM, que pode ser vistas nas Figuras 6.**(a)** e 7.**(a)** e levemente na Figura 4.**(a)**.

Em seguida, analisaremos os placares cujas taxas de saídas foram modeladas por Weibull.

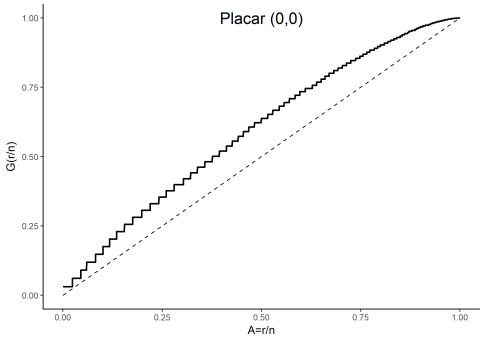
3.2.2 Modelos Weibull

Figura 9 – Placar (0x0)

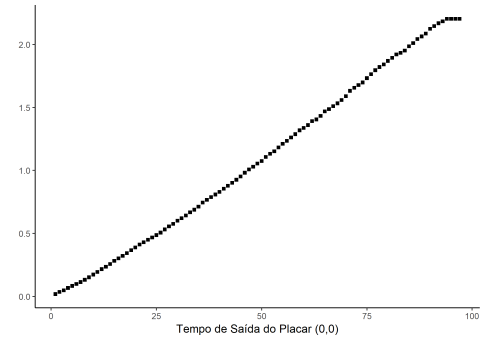
(**a**) KM e Ajustes



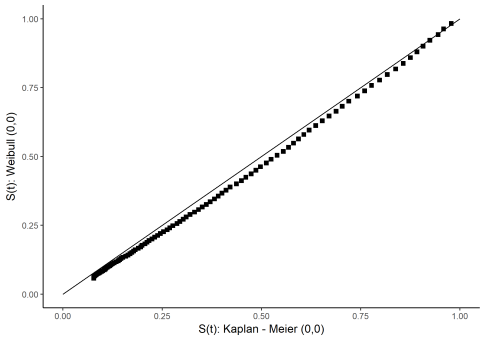
(**b**) Gráfico TTT



(**c**) Linearização Modelo Weibull



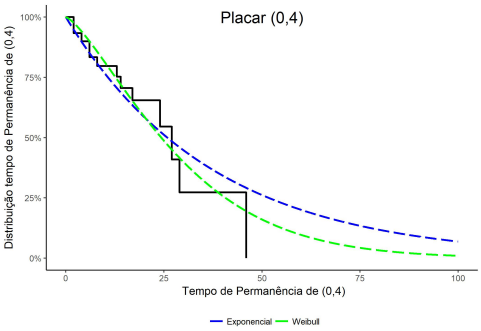
(**d**) $\hat{S}_w(t)$ X $\hat{S}_{km}(t)$



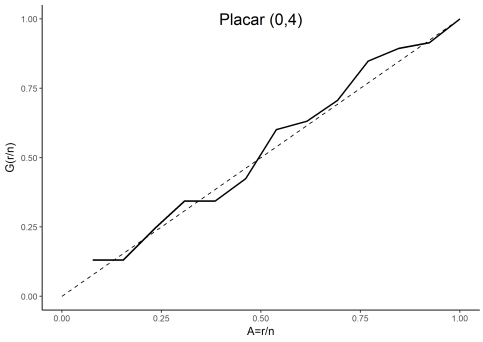
Nosso gráfico **(a)** nos mostra um ajuste visual mais satisfatório do modelo WEIB em relação ao modelo EXP, que apresenta leve subestimação para $t < 50$ e leve superestimação para $t > 50$. Nossa curva de tempo em teste sugere o modelo WEIB como adequado e nossas Figuras 9.**(c)** e **(b)** mostram certa forma linear. Associado a uma melhora no p-valor do teste K-S, concluímos que o segundo modelo proposto parece adequado. Veja que, agora, assumimos o tempo de saída do placar crescente.

Figura 10 – Placar (0x4)

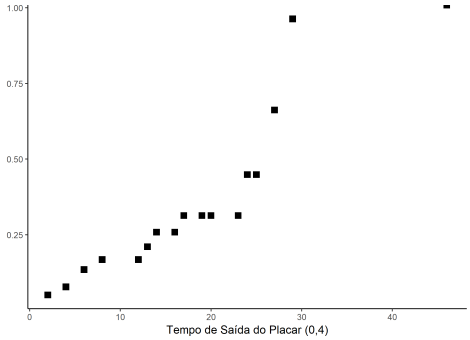
(**a**) KM e Ajustes



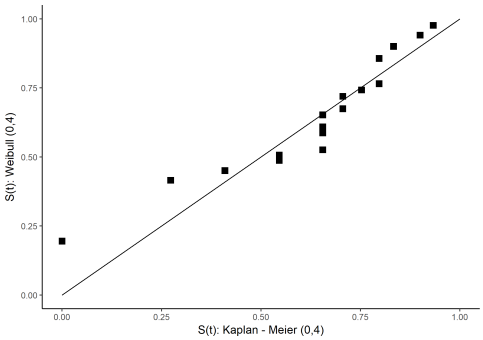
(**b**) Gráfico TTT



(**c**) Linearização Modelo Weibull



(**d**) $\hat{S}_w(t)$ X $\hat{S}_{km}(t)$

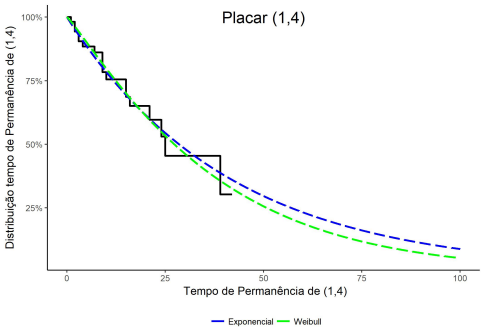


Para o placar (0x4), dizemos que, de certa forma, a curva TTT aparenta certa linearidade e o modelo EXP seja satisfatório. Os gráficos **(b)** e **(c)** são inconclusivos, no qual, este último, não aparenta certa linearidade, com desajuste para tempos de função de sobrevivência baixos. Tal fato pode ser evidenciado em **(a)**, cujo modelo WEIB aparenta melhor adequabilidade, porém com desajuste da curva de sobrevivência para a cauda. Porém, ao associarmos ao resultados anteriores já expostos na Seção 4.2.2, dizemos que o modelo proposto parece adequado.

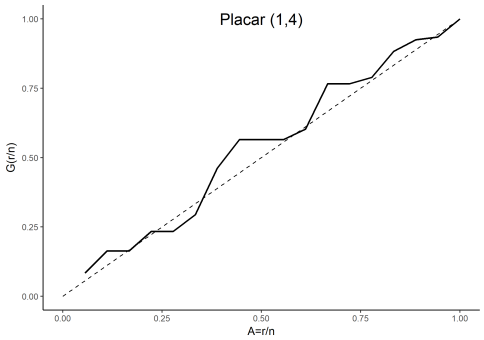
Dessa forma, dizemos agora que o tempo médio de ocorrência do placar (0x4) é dada por 29,31 minuto, com intensidade crescente ao longo do jogo.

Figura 11 – Placar (1x4)

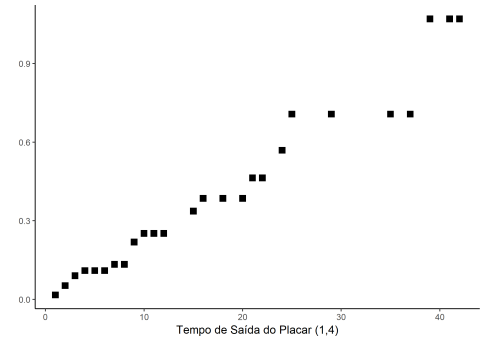
(**a**) KM e Ajustes



(**b**) Gráfico TTT



(**c**) Linearização Modelo Weibull



(**d**) $\hat{S}_w(t)$ X $\hat{S}_{km}(t)$

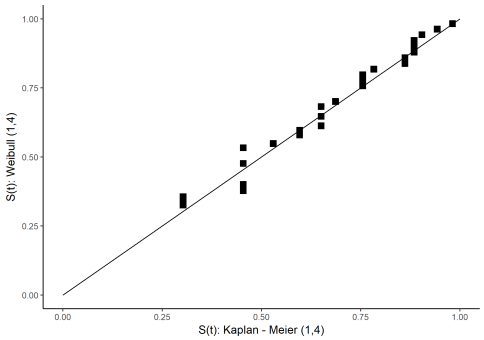
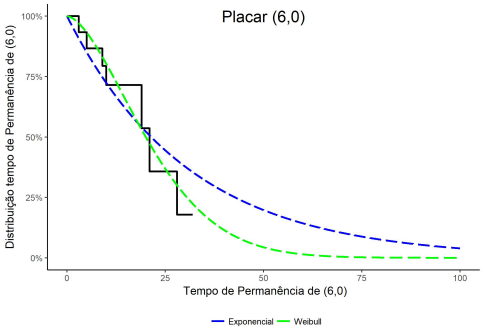
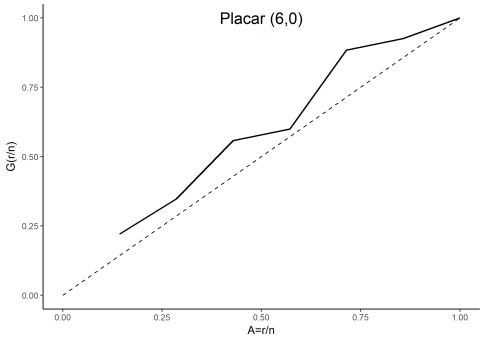


Figura 12 – Placar (6x0)

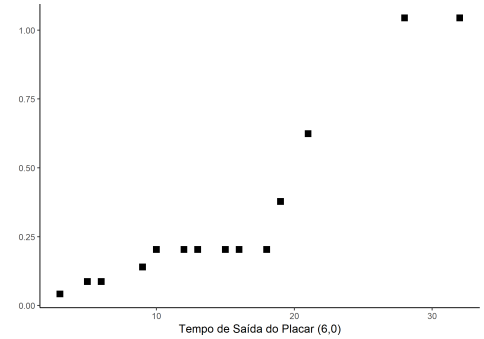
(**a**) KM e Ajustes



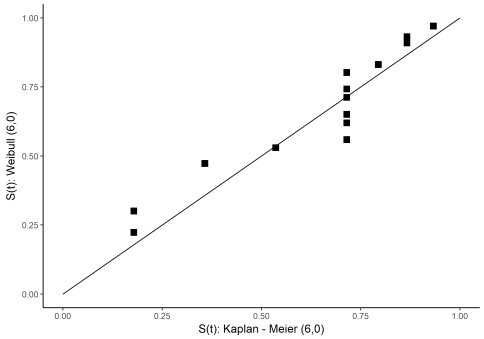
(**b**) Gráfico TTT



(**c**) Linearização Modelo Weibull



(**d**) $\hat{S}_w(t)$ X $\hat{S}_{km}(t)$



Por fim, ao estudarmos os placares (6x0) e (1x4) observamos similaridade nas análises . Ambas as Figuras 11.**(b)** e 12.**(b)** aparentam certa linearidade, o que implica que o modelo EXP seja adequado. Contudo, as Figuras 11.**(a)** e 12.**(a)** mostram um ajuste mais adequado quando feito pelo modelo com mais parâmetros. Os gráficos 11.**(d)** e 12.**(d)** também aparentam linearidade, contudo, ambas as figuras **(c)** mostram leve desajuste. Assim, associados novamente aos resultados da Seção 4.2.2, dizemos que os modelos parecem relativamente bem ajustados.

Observe então que assumimos os tempos médios de permanência dos placares (6x0) e (1x4) como sendo 22,37 e 36,46 minutos; respectivamente, sendo este primeiro um dos menores estimados.

3.2.3 Modelos Escolhidos

A seguir, mostraremos as escolhas dos modelos e suas estimativas finais:

Tabela 4 – Modelo escolhido para cada placar

Placar	Quantidade	Censuras	Modelo Escolhido	$\hat{\mu}(t)$	Intensidade	$\hat{\beta}$
(0x0)	4.597	7,4%	Weibull	39,58	Crescente	1,16
(1x0)	3.200	20,06%	Exponencial	35,34	Constante	1,00
(2x0)	1.601	27,23%	Exponencial	29,56	Constante	1,00
(3x0)	682	35,04%	Exponencial	28,19	Constante	1,00
(4x0)	240	39,58%	Exponencial	22,60	Constante	1,00
(5x0)	74	36,49%	Exponencial	21,34	Constante	1,00
(6x0)	15	53,33%	Weibull	22,37	Crescente	1,65
(1x1)	1.616	36,26%	Exponencial	40,10	Constante	1,00
(1x2)	636	46,23%	Exponencial	37,71	Constante	1,00
(2x2)	484	54,96%	Generalizado	36,36	-	-
(2x1)	1.034	46,23%	Generalizado	40,73	-	-
(3x1)	470	48,72%	Generalizado	35,25	-	-
(4x1)	184	54,35%	Generalizado	38,13	-	-
(5x1)	65	43,08%	Exponencial	28,70	Constante	1,00
(3x2)	246	60,38%	Generalizado	42,27	-	-
(4x2)	89	60,04%	Generalizado	45,37	-	-
(5x2)	39	61,67%	Generalizado	41,00	-	-
(3x3)	75	60,00%	Exponencial	31,90	Constante	1,00
(0x1)	1399	21,44%	Exponencial	36,17	Constante	1,00
(0x2)	443	35,44%	Exponencial	33,28	Constante	1,00
(0x3)	129	51,94%	Generalizado	34,21	-	-
(0x4)	30	56,67%	Weibull	29,31	Crescente	1,34
(1x3)	179	55,31%	Generalizado	36,50	-	-
(1x4)	54	59,67%	Weibull	36,46	Crescente	1,11
(2x3)	133	60,17%	Generalizado	45,96	-	-

Na Tabela 4, encontram-se os placares, o número de passagens pelo placar (Quantidade), a porcentagem dos tempos de censura, o modelo escolhido dentre os propostos, o tempo médio de permanência a partir do modelo proposto e por fim, o indicativo da intensidade de saída do placar, ao longo da partida, e sua quantificação.

Pode-se ver através da Tabela 4 que, dentre os placares propostos, onze deles apresentaram um modelo paramétrico para a curva de sobrevivência satisfatório através do modelo Exponencial, quatro deles para o modelo Weibull e em dez casos nenhum dos dois se apresentou adequado. Nosso maior tempo de permanência foi de 45,96 minutos, no placar (2x3), e menor estimado foi de 21,34 minutos, curiosamente, do placar (5x0), como já interpretado no início da seção anterior. Em relação a presença de censuras, o placar (0x0) apresentou o menor percentual de términos da partida sem alteração do placar.

Repare que utilizamos as taxas estimadas pelos respectivos modelos escolhidos, para cada placar. Apresentamos também as conclusões das intensidades das taxas de saída ao longo da partida e sua quantificação através da estimativa do parâmetro de forma do modelo Weibull.

Especificamente, para o **TRV**, utilizados o modelo Gama Generalizado para comparação da adequabilidade do modelo atual com modelo proposto por meio da função de verossimilhança. Em geral, como dito anteriormente, alguns placares apresentaram inadequabilidade, seja por desajuste nas caudas, que apresentaram forma constante após determinado tempo t, seja pela forma da função de sobrevivência. Assim, quando houve desajuste por parte dos dois modelos propostos, decidimos optar por um modelo generalizado para ser realizado por trabalhos futuros. Dessa forma, não consideramos na Tabela 3 as estimativas dos outros dois modelos, embora elas venham a ser úteis para efeitos comparativos.

4 Conclusão

Como comentado anteriormente, o principal intuito deste trabalho foi estudar a ocorrência de gols, em jogos de pontos corridos, no campeonato brasileiro série A.

Utilizando os conceitos de Cadeias de Markov a tempo contínuo, foi possível estimar as probabilidades de saída de determinado placar, mostrado por meio de nossa Árvore de Transição de Placares. Os resultados obtidos possibilitaram a interpretação de determinados eventos típicos do futebol, como: *Goleadas*, *jogos truncados*, vantagens do time mandante devido ao fator *casa*.

Ao modelarmos as taxas de saída de determinado placar, obtivemos uma estimativa média do tempo necessário para a ocorrência de um gol em determinado estado e compreensão da intensidade de saída do mesmo ao longo do jogo, sendo ela crescente ou constante.

Em relação as dificuldades encontradas, o principal problema enfrentado na estimação nas extremidades das caudas dos modelos pode ser explicado pela pouca ocorrência de determinado placar ou alta presença de censuras. Uma possível solução seria utilizar o conceito de *Fração de Cura*, no qual a partir de determinado tempo t, a nossa S(t) é constante. Porém, como estamos utilizando o conceito em jogos de futbol esta opção não é factível. Outra alternativa seria a utilização de modelos mais sofisticados, como o próprio modelo GGama.

Outro ponto importante seria a utilização de adaptações do teste de Kolmogorov - Smirnov que diminuíssem a sensibilidade nas extremidades das caudas para melhor avaliação do ajuste paramétrico.

Concluímos este relatório então, abrindo o precedente para, em trabalhos posteriores, a estimação da taxa de ocorrência de gols em jogos de mata-mata, para efeitos comparativos de resultados.

Referências

COLOSIMO, S. R. G. E. A. *Análise de Sobrevivência Aplicada*. [S.l.: s.n.], 2006. Citado na página xxvii.

CONOVER, W. C. *Practical Nonparametric Statistics*. [S.l.: s.n.], 1999. Citado na página xxiv.

DURRETT, R. *Essentials of Stochastics Processes*. [S.l.: s.n.], 2011. Almost Final Version of 2nd Edition. Citado 2 vezes nas páginas xxi e xxxv.

JUNIOR, D. G. Oswaldo Gomes de S. Previsão de partidas de futebol utilizando modelos dinâmicos. *XXXVI-SBPO*, 2014. Citado na página xix.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página xxii.

MARTINS, A. R. D. Helgem de S. R. Modelagem dinâmica de partidas de futebol. *Revista da Estatística UFOP*, v. 2, 2014. Citado na página xix.

NAKANO, E. Y. *Um Curso de Análise de Sobrevivência*. [S.l.], 2016. Citado na página xxv.