Rúben Filipe Caldeira Lavrador

Previsões em Futebol

Dissertação de Mestrado em Métodos Quantitativos em Finanças, orientada pela Professora Doutora Maria da Graça Santos Temido Neves Mendes e apresentada ao Departamento de Matemática da Faculdade de Ciências e Tecnologia e à Faculdade de Economia da Universidade de Coimbra.

2017

· U C ·

Universidade de Coimbra

Previsões em Futebol

Rúben Lavrador

Universidade de Coimbra

Mestrado em Métodos Quantitativos em Finanças

Master in Quantitative Methods in Finance

Dissertação de Mestrado | MSc Dissertation

March 2017

Agradecimentos

Agradeço, em primeiro lugar, a Deus, que me deu vida e saúde.

Agradeço aos meus pais, Florentino Lavrador Francisco e Benvinda Maria da Silva Caldeira Lavrador, pelo apoio e incentivo que me deram ao longo da vida.

Apresento os meus profundos agradecimentos, em particular, à minha orientadora, Doutora Maria da Graça Santos Temido Neves Mendes, que aceitou orientar-me de forma constante e assídua não poupando esforços para que conseguíssemos chegar ao resultado final.

Resumo

Esta dissertação é dedicada ao estudo de modelos de Poisson univariados e bivariados que permitem estudar probabilidades e odds relativas a jogos de futebol. São estimadas estas quantidades para alguns jogos do campeonato português e do campeonato inglês

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

"In God we trust, all others must bring data"
William E. Deming

A previsão de resultados em futebol é um assunto que tem recebido a atenção de muitos estatísticos devido à sua importância, quer no contexto dos sistemas de apostas quer no impacto que este desporto tem na sociedade em geral. De facto, existe já uma vasta literatura cujo conteúdo é dedicado a tais previsões, partindo da introdução de novos modelos ou simplesmente aplicando os já existentes à multiplicidade de dados que surgem constantemente.

Entre muitos autores citamos [1–4].

Existe em tal literatura a tendência generalizada em considerar a lei de Poisson para modelar o número de golos marcados por uma equipa num jogo de futebol. Por um lado, porque o modelo é simples, por outro, porque os dados assim o determinam. Mais do que o número de golos por jogo, considerado unilateralmente, interessam-nos probabilidades de vitórias, empates, derrotas, isto é, há que estudar conjuntamente as duas variáveis associadas ao número de golos das duas equipas em jogo.

Numa primeira análise somos levados a acreditar que os dados disponíveis em qualquer *site* sobre futebol nos permitem chegar a resultados fidedignos. No entanto, quando o objetivo é o de prever probabilidades associadas a determinado confronto, por exemplo Benfica – Belenenses, para podermos prever o resultado de forma correta seria preciso uma amostra de resultados anteriores entre as duas equipas. Esses dados não refletiriam o momento atual dessas equipas uma vez que os resultados disponíveis estão associados muitas vezes a jogadores, treinadores, favoritismos e staff muito diferentes. Globalmente, são condições bastante distintas das de hoje. Para o mesmo par não dispomos de uma amostra bivariada com regularidade estatística suficiente para proceder à inferência dos resultados pretendidos. Sendo assim, a solução passa por considerar os jogos anteriores das duas equipas quando estas afrontam outras equipas.

Depois da escolha do modelo e do cálculo das probabilidades de vitória ou empate das equipas quando afrontam uma outra equipa, estas probabilidades são depois convertidas em *odds* as quais são o inverso da probabilidade estimada. Em geral, a *odd* de um evento é o inverso da probabilidade desse evento ocorrer:

$$Odd_i = \frac{1}{P(i)}$$

As casas de apostas recebem as probabilidades dos eventos de especialistas internacionais. O valor final das odds a apresentar na oferta de eventos desportivos tem depois a correção proveniente das margens que cada casa de apostas define para a sua exploração.

Para perceber melhor esta situação apresentamos um exemplo simples.

Consideremos o lançamento de uma moeda equilibrada, como a probabilidade de cada acontecimento é 0.5, então as odds serão iguais a 2 para cada acontecimento. Sendo assim, se um apostador apostar 1 euro no evento "Face" e outro apostador 1 euro no evento "Cara", o vencedor receberia 2 euros, que são o euro que ele apostou e o euro do outro apostador, deixando assim a casa de apostas sem ganhos. É por esta razão que as casas de apostas definem odds mais pequenas. Assim, no nosso exemplo, a casa de apostas definiria uma *odd* de 1.9. Neste caso, o vencedor recebe 1.9 euros e a casa de apostas 0.1 euros. Aqui, as probabilidades que correspondem às *odds* na equação anterior são chamadas de probabilidades implícitas.

Estamos, então, interessados em saber como estimar estas probabilidades. Para isso, estabelecemos primeiro os modelos que vamos usar e, de seguida, vamos transformar as nossas probabilidades em odds e compará-las com os valores projetados pelas casas de apostas.

O campeonato inglês é composto por 20 equipas que jogam entre elas duas vezes por época desportiva, dando um total de 380 jogos.

O campeonato português é composto por 18 equipas que jogam entre elas duas vezes por época desportiva, num total de 360 jogos.

O campeonato inglês é mais homogéneo do que o campeonato português. No campeonato português existem 2 ou 3 equipas que são as mais fortes e normalmente o campeonato acaba por se decidir entre estas. Quanto à liga inglesa os resultados são extremamente variáveis . A título de exemplo, sabemos que o Leicester em 2014/2015 acabou o campeonato em 14º lugar e acabou por ser campeão no ano a seguir.

Os dados que vamos usar serão os resultados dos jogos da época 2015/2016 e os resultados da época atual 2016/2017 até à data de 31/1/2017. Encontram-se em http://www.football-data.co.uk/data.php.

Consideramos as quatro variáveis aleatórias PortCasa, PortFora, InglCasa e InglFora que descrevem o número de golos da equipa da casa e da equipa de fora, em ambos os campeonatos.

A escolha do modelo estatístico é de extrema importância. Este trabalho, que reflete parcialmente o estudo que realizamos durante os últimos seis meses, começou naturamente pelo estudo de artigos científicos. Como já dissemos, o seu conteúdo tem subjacente um modelo de Poisson univariado ou bivariado. No segundo caso a dependência recebe na literatura um tratamente muito deversificado, baseado na autocovariância ou em funções de dependência.

Apesar de tal diversidade de trabalhos e de versões do modelo de Poisson, em muito poucos são apresentados com clareza os métodos de estimação dos parâmetros envolvidos. Mesmo assim, em alguns destes a aplicação dos métodos propostos aos nossos dados, conduziu a resultados não conformes.

Passamos a uma breve descrição de alguns destes trabalhos.

Em Maher [4] é estudado o modelo de Poisson univariado e bivariado adaptado ao futebol, uma vez que para cada jogo, com as equipas i e j, se considera a média das duas variáveis de Poisson como sendo o produto $\alpha_i\beta_j$ e $\gamma_i\delta_j$, com

$\alpha_i =$ medida da capacidade de ataque da equipa i em casa,

$\beta_j =$ medida da capacidade de defesa da equipa j fora,

$\gamma_i =$ medida da capacidade de defesa da equipa i em casa

e

$\delta_j =$ medida da capacidade de ataque da equipa j fora.

Maher estabelece os estimadores de máxima verosimilhança e faz uma aplicação ao campeonato inglês de 1982. Interessámo-nos por este artigo e construímos as nossas previsões a partir dos seus resultados, como veremos nos capítulos 2 e 3.

Dixon and Coles [1] apresenta um modelo de Poisson bivariado, com médias descritas como em Maher, que considera dependência apenas nos resultados (0,0), (0,1), (1,0) e (1,1). As equações de verosimilhança têm forma demasiado complexa, não sendo completamente trabalhadas no artigo. Os resultados, muito semelhantes ao de Maher, dizem respeito ao campeonato inglês.

Em Mwembe et al. [5] é também aplicado um modelo de Poisson aos dados do campeonato do Zimbabwe. Aqui considera-se que os parâmetros α_i, β_j, γ_i e δ_j podem ser determinados por um modelo autorregressivo de ordem k, isto é, envolvendo os k valores anteriores dos mesmos parâmetros. Por exemplo, no momento t

$$\alpha_i \equiv \alpha_{i,t} = \phi_{1,t}\alpha_{i,t-1} + \phi_{2,t}\alpha_{i,t-2} + \ldots + \phi_{k,t}\alpha_{i,t-k}$$

e o mesmo para os outros três parâmetros. A forma como escolhem a ordem da série temporal é pouco rigorosa. Os dados(neste caso o número de golos) não são dependentes ao longo do tempo, ou seja o número de golos que uma equipa marca numa semana não é dependente do número de golos que a mesma equipa marcou na semana anterior. Obtemos resultados muito diferentes consoante a ordem do modelo autorregressivo.

A questão da inflação de zeros nos dados encontra-se modelada em muitos trabalhos. Em [2] é apresentado um modelo de Poisson inflacinado nos empates (0,0) e nos resultados $(i,0)$ e $(0,j)$, com $i,j \geq 1$, a que chama *Bivariate Zero Inflacted Model*. Os métodos de estimação apresentados, quando aplicados aos nossos dados, não produziram resultados que pudessemos considerar.

Ainda assim, a referida inflação de zeros levou-nos a estudar o modelo de Poisson univariado inflacionado em zero. Tal estudo é apresentado na Secção 2.2. e permitiu melhorar significativamente o ajustamento tradicional com o modelo de Poisson. Tal facto e uma observação mais cuidada dos dados globais de cada campeonato, levou-nos a introduzir e estudar dois modelos de Poisson "corrigidos" de outras formas. A saber, trata-se do modelo de Poisson deflacionado em dois e do modelo de Poisson inflacionado em zero e deflacionado em dois. São estudados nas secções 2.3 e 2.4. Tanto quanto sabemos não se encontram na literatura de probabilidades com aplicações em desporto. A aplicação aos nossos dados deu evidência a um ajustamento muito melhor. Uma vez que são aplicados aos dados globais e não por equipas, julgamos que seria de todo o interesse a introdução

nestes modelos dos fatores de ataque e de defesa, isto é, das médias escritas na forma $\alpha_i\beta_j$ e $\gamma_i\delta_j$. Só assim se poderão fazer previsões para os pares de equipas em cada jogo.

O Capítulo 2 termina com duas secções dedicadas ao modelo de Poisson bivariado (com dependência) e com o modelo de Poisson bivariado inflacionado nos empates. A aplicação deste modelo ao campeonato inglês, considerando os empates em $(0,0)$ e em $(2,2)$ mostrou a relevância do modelo.

Há ainda a questão do momento da equipa. Ao longo deste estudo entendemos que o método de estimação das probabilidades implícitas, utilizando os modelos previamente estudados, deve ter uma modificação na análise dos dados, de modo a que possamos tomar em consideração o momento da equipa. Assim, damos mais valor aos dados mais recentes e menos valor aos dados mais antigos refletindo deste modo tal momento da equipa. Esta alteração, que reflete uma sobrevalorização dos dados mais recentes, pode ser realizada da seguinte forma: consideramos um valor k (número de semanas, por exemplo) e calcula-se novos dados $x^*_{i,j}$ tais que

$$\sum_{1\leq i,j\leq n} x^*_{ij} = \sum_{1\leq i,j\leq n-k} (1-p)x_{ij} + \sum_{n-k+1\leq i,j\leq n} (1+p)x_{ij},$$

estimando p e k convenientemente. Entendemos que esta ideia precisa de mais reflexão.

Do que foi dito, concluímos que este trabalho se divide em três capítulos, sendo o último dedicado à previsão de probabilidades e de odds com base no artigo de Maher.

Acrescentamos que os modelos descritos ao longo do texto são modelos estatísticos que não tomam em conta outros parâmetros como fatores psicológicos da equipa e dos jogadores individualmente, suspensões, lesões, ou mesmo condições atmosféricas, uma vez que algumas equipas tem um tipo de jogo muito desfavorável ao mau tempo. Mais, algumas formas de jogar das equipas encaixam muito bem noutras, há influência dos resultados dos jogos de outra competição na qual a equipa esta incluída, a fadiga de jogos acumulados (devido a mais competições algumas equipas tem mais jogos), entrada ou saída de jogadores, treinadores, staff, etc.

Terminamos informando que os gráficos incluídos, a maximização de funções e a resolução de equações, quando não tratadas analiticamente, se devem ao *software Mathematica*.

A figura da capa foi retirada de http://whattodoinmadrid.com/blog/vicente-calderon-stadium/

Capítulo 2

Modelos

"Matemática, onde estás?
Estou aqui!
Onde?
No remate que antecede o golo!!"

(Autor desconhecido)

2.1 Poisson univariado

Uma variável aleatória X discreta tem distribuição ou lei de Poisson de parâmetro $\lambda > 0$ se tem suporte N_0 e a sua função de probabilidade se escreve na forma

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, x \in \mathrm{N}_0$$

Usamos a notação $X \sim \mathscr{P}(\lambda)$. Prova-se que $E(X) = \lambda$ e $Var(X) = \lambda$.

Num jogo onde a equipa i defronta a equipa j representemos por

$$X = \text{Número de golos da equipa i}$$

e por

$$Y = \text{Número de golos da equipa j.}$$

Se uma equipa i defronta a equipa j então os resultados serão da forma (x_{ij},y_{ij}).

Em Maher [4] considera-se que X tem como média $\alpha_i\beta_j$ e Y tem como média $\gamma_i\delta_j$. Os parâmetros α_i,β_j,γ_i e δ_j, tal como referido na introdução, são definidos da seguinte forma:

$\alpha_i =$ medida da capacidade de ataque da equipa i em casa,

$\beta_j =$ medida da capacidade de defesa da equipa j fora,

$\gamma_i =$ medida da capacidade de defesa da equipa i em casa

e

$\delta_j =$ medida da capacidade de ataque da equipa j fora.

De seguida, apresentamos como se estimam estes quatro parâmetros.

Considerando inicialmente que as variáveis X e Y são independentes, então a estimação de α e de β só depende da amostra de X e a estimação de δ e γ só vai depender da amostra de Y.

Uma vez que se tem

$$P(X = x_{ij}) = \frac{(\alpha_i\beta_j)^{x_{ij}}e^{-\alpha_i\beta_j}}{x_{ij}!},$$

então a função de verosimilhança é dada por

$$\begin{aligned} L(\underline{\alpha},\underline{\beta}) &= \prod_{i\geq 1}\prod_{j\neq i} P(X = x_{ij}) \\ &= \prod_{i\geq 1}\prod_{j\neq i} \frac{(\alpha_i\beta_j)^{x_{ij}}e^{-\alpha_i\beta_j}}{x_{ij}!} \end{aligned}$$

onde $\underline{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n)$ e $\underline{\beta} = (\beta_1, \beta_2, ..., \beta_n)$.

Aplicando a função logarítmo temos

$$\begin{aligned} \log L(\alpha,\beta) &= \log(\prod_{i\geq 1}\prod_{j\neq i} \frac{(\alpha_i\beta_j)^{x_{ij}}e^{-\alpha_i\beta_j}}{x_{ij}!}) \\ &= \sum_{i\geq 1}\sum_{j\neq i}(x_{ij}\log(\alpha_i\beta_j) - \alpha_i\beta_j - \log(x_{ij}!)) \\ &= \sum_{i\geq 1}\sum_{j\neq i}(-\alpha_i\beta_j + x_{ij}\log(\alpha_i\beta_j) - \log(x_{ij}!)) \end{aligned}$$

Assim, para qualquer $i \in \{1, 2, ..., n\}$, fixado previamente, obtemos o seguinte sistema

$$\begin{cases} \frac{\partial \log L}{\partial \alpha_i} = \sum_{j\neq i}(-\beta_j + \frac{x_{ij}}{\alpha_i}) \\ \frac{\partial \log L}{\partial \beta_i} = \sum_{j\neq i}(-\alpha_i + \frac{x_{ij}}{\beta_i}) \end{cases}$$

Igualando as duas equações a zero, o sistema

$$\begin{cases} \frac{\partial \log L}{\partial \alpha_i} = 0 \\ \frac{\partial \log L}{\partial \beta_i} = 0 \end{cases}$$

é equivalente a

$$\begin{cases} \sum_{j\neq i}(-\beta_j + \frac{x_{ij}}{\alpha_i}) = 0 \\ \sum_{j\neq i}(-\alpha_i + \frac{x_{ij}}{\beta_i}) = 0 \end{cases}$$

ou seja

$$\begin{cases} \sum_{j\neq i}\beta_j = \frac{1}{\alpha_i}\sum_{j\neq i}x_{ij} \\ \sum_{j\neq i}\alpha_i = \frac{1}{\beta_i}\sum_{j\neq i}x_{ij} \end{cases}$$

Os estimadores da máxima verosimilhança verificam então

$$\widehat{\alpha}_i = \frac{\sum_{j\neq i}x_{ij}}{\sum_{j\neq i}\widehat{\beta}_j} \quad \text{e} \quad \widehat{\beta}_i = \frac{\sum_{j\neq i}x_{ij}}{\sum_{j\neq i}\widehat{\alpha}_j}$$

Para obter os valores $(\widehat{\alpha}_i, \widehat{\beta}_j)$, usamos um método iterativo, como por exemplo o método de Newton-Raphson. Da mesma forma se obtém $(\widehat{\gamma}_i, \widehat{\delta}_j)$ substituindo x_{ij} por y_{ij}.

No Capítulo 3 usamos estes valores para estimar probabilidades e Odds associados a alguns pares de equipas.

No que se segue usamos os dados globais das quatro variáveis PortCasa, PortFora, InglCasa e InglFora. Para cada uma destas variáveis realizamos o teste do qui-quadrado de ajustamento com o objetivo de averiguar se o modelo de Poisson se adequa à sua distribuição. Concretamente, testamos a hipótese

$$H_0 : X \sim \mathscr{P}(\lambda)$$

com λ estimado previamente, considerando as classes $\{0\},\{1\},\{2\},\{3\},\{4\},\{5\},\{6,7,...\}$. Representando por n_i a frequência absoluta da classe $\{i\}$ e por e_i a respetiva frequência esperada sob a validade de H_0, a estatística de teste é

$$\chi^2 = \sum_{i=1}^{7}\frac{(n_i - e_i)^2}{e_i},$$

a qual segue assintoticamente a lei do qui-quadrado com 6 graus de liberdade. O valor observado da estatística de teste será representado por χ^2_{obs}. Tomamos a decisão com base no p-valor do teste, nomeadamente, mantemos H_0 sempre que p-valor> 0.05 e rejeitamos H_0 sempre que p-valor≤ 0.05.

Comecemos com a variável aleatória InglCasa. Tem-se $\bar{x} = 1.545$, pelo que se testa a distribuição $\mathscr{P}(1.545)$. As frequências observadas e as frequências esperadas, sob a validade de H_0, encontram-se na Tabela 2.1. O teste do qui-quadrado fornece $\chi^2_{obs} = 5.6307$ e p-valor igual a 0.47. Mantemos então a hipótese H_0.

Relativamente a variável InglFora tem-se $\bar{y} = 1.215$. Procedendo como no caso da variável InglCasa foram obtidos os valores que se encontram na Tabela 2.2. O valor observado da estatística de teste é $\chi_{obs} = 6.1546$ tendo-se $p-valor = 0.39$. É também mantida a hipótese H_0.

Relativamente as variáveis PortCasa e PortFora o procedimento foi rigorosamente o mesmo. As tabelas 2.3 e 2.4 contêm as frequências esperadas, sob a validade da hipótese H_0, onde se usou λ estimado por $\bar{x} = 1.456$ e $\bar{y} = 1.139$, respetivamente. No primeiro caso obtemos $\chi^2_{obs} = 11.109$ e $p-valor = 0.09$ e no segundo caso $\chi^2_{obs} = 6.2108$ e $p-valor = 0.37$. Em ambos os casos se mantém H_0.

	Freq obs	**Freq esp**
0	133	127.987
1	201	197.740
2	136	152.754
3	84	78.668
4	33	30.386
5	8	9.389
6	5	2.418

Tabela 2.1 Frequências observadas e frequências esperadas relativas à variável InglCasa

	Freq obs	**Freq esp**
0	193	178.026
1	203	216.302
2	120	131.403
3	56	53.218
4	23	16.165
5	4	3.928
6	1	0.795

Tabela 2.2 Frequências observadas e frequências esperadas relativas à variável Inglfora

	Freq obs	**Freq esp**
0	124	110.555
1	153	160.934
2	102	117.136
3	64	56.838
4	18	20.685
5	9	6.022
6	4	1.461

Tabela 2.3 Frequências observadas e frequências esperadas relativas à variável Portcasa

	Freq obs	**Freq esp**
0	161	151.710
1	166	172.833
2	93	98.449
3	33	37.386
4	17	10.648
5	3	2.426
6	1	0.461

Tabela 2.4 Frequências observadas e frequências esperadas relativas à variável Portfora

2.2 Poisson inflacionado em zero

As tabelas 2.1-2.4 apresentadas nos exemplos da secção anterior sugerem que a frequência de zeros na amostra é significativamente superior à respetiva frequência esperada gerada pelo modelo de Poisson. *Grosso modo*, podemos afirmar que as equipas "marcam mais 0 golos" do que aquilo que o modelo prevê. Neste contexto, será de todo o interesse verificar se o modelo de Poisson inflacionado em zero (modelo ZIP - do inglês Zero-Inflated Model) será adequado para corrigir significativamente tal discrepância. Antes de apresentarmos uma aplicação adequada passamos a caracterizar a distribuição subjacente a este modelo, bem como os estimadores obtidos pelo Método dos Momentos e pelo Método da Máxima Verosimilhança. Este modelo é sobejamente conhecido em Teoria das Probabilidades (veja-se por exemplo [6]).

Uma variável aleatória X tem distribuição de Poisson inflacionada em zero se tem suporte $\mathbb{N}_0$ e

$$P(X=x) = \alpha \mathbb{1}_{\{0\}}(x) + (1-\alpha)P(Z=x),\ x \in \mathbb{N}_0,$$

onde $\alpha \in]0,1[$ e $Z \sim \mathscr{P}(\lambda)$, com $\lambda > 0$. Uma vez que se tem

$$\begin{cases} E(X) = (1-\alpha)\lambda \\ E(X^2) = (1-\alpha)(\lambda+\lambda^2) \end{cases}$$

pelo método dos momentos obtemos

$$\begin{cases} \widehat{\lambda} = \dfrac{\sum_{i=1}^{n} X_i^2}{\overline{X}} - 1 \\ \widehat{\alpha} = 1 - \dfrac{\overline{X}}{\widehat{\lambda}} = \dfrac{\sum_{i=1}^{n} X_i^2 - \overline{X}^2 - \overline{X}}{\sum_{i=1}^{n} X_i^2 - \overline{X}} \end{cases}$$

ou seja

$$\begin{cases} \widehat{\lambda} = \frac{M_2}{M_1} - 1 \\ \widehat{\alpha} = \frac{M_2 - M_1^2 - M_1}{M_2 - M_1} \end{cases}$$

Nestre trabalho representamos por M_k o momento empírico de ordem k de X, isto é $M_k = \sum_{i=1}^{n} X_i^k$.

Por outro lado, para a mostra observada $(x_1, x_2, ..., x_n)$, a função de log-verosimilhança é

$$\log L(\alpha,\lambda) = \sum_{x_i=0} \log(\alpha + (1-\alpha)e^{-\lambda}) + \sum_{x_i \neq 0} \log((1-\alpha)e^{-\lambda}\frac{\lambda_i^{x}}{x_i!})$$

Derivando a função anterior em função de α obtemos:

$$\frac{\partial \log L}{\partial \alpha}(\alpha,\lambda) = n_0 \frac{1-e^{-\lambda}}{\alpha + (1-\alpha)e^{-\lambda}} - \frac{n-n_o}{1-\alpha}$$

e derivando em função de λ vem

$$\begin{aligned}\frac{\partial \log L}{\partial \lambda}(\alpha,\lambda) &= n_0 \frac{-(1-\alpha)e^{-\lambda}}{\alpha+(1-\alpha)e^{-\lambda}} + \sum_{x_i \neq 0}(-1+\frac{x_i}{\lambda}) \\ &= n_0 \frac{-(1-\alpha)e^{-\lambda}}{\alpha+(1-\alpha)e^{-\lambda}} - (n-n_0) + \frac{1}{\lambda} n\bar{x}\end{aligned}$$

onde n_0 representa a frequência absoluta de $\{0\}$. Doravante denotamos por n_i a frequência absoluta das observações $\{i\}$ e por f_i a correspondente frequência relativa.

De modo a obter o ponto crítico da função $\log L$, igualamos as expressões das suas derivadas parciais a zero.

$$\begin{aligned}&\frac{\partial \log L}{\partial \alpha}(\alpha,\lambda) = 0 \\ &\qquad \Leftrightarrow n_0(1-e^{-\lambda})(1-\alpha) - (n-n_0)e^{-\lambda} - (n-n_0)\alpha(1-e^{-\lambda}) = 0 \\ &\qquad \Leftrightarrow (1-e^{-\lambda})(n_0 - n\alpha) - (n-n_0)e^{-\lambda} = 0 \\ &\qquad \Leftrightarrow n_0 - n\alpha = (1-\alpha)ne^{-\lambda}\end{aligned}$$

$$\begin{aligned}&\frac{\partial \log L}{\partial \lambda}(\alpha,\lambda) = 0 \\ &\qquad \Leftrightarrow n_0 \frac{-(1-\alpha)e^{-\lambda}}{\alpha+(1-\alpha)e^{-\lambda}} - (n-n_0) + \frac{1}{\lambda} n\bar{x} = 0 \\ &\qquad \Leftrightarrow -n_0(1-\alpha)e^{-\lambda} - (n-n_0)\alpha - n(1-\alpha)e^{-\lambda} + n_0(1-\alpha)e^{-\lambda} + \frac{1}{\lambda} n\bar{x}\alpha \\ &\qquad\qquad\qquad + \frac{1}{\lambda} n\bar{x}(1-\alpha)e^{-\lambda} = 0 \\ &\qquad \Leftrightarrow n_0 - n\alpha = (1-\alpha)ne^{-\lambda}\end{aligned}$$

Obtemos então o seguinte sistema

$$\begin{cases} n_0 - n\alpha = (1-\alpha)ne^{-\lambda} \\ -(n-n_0)\alpha - n(1-\alpha)e^{-\lambda} + \frac{1}{\lambda} n\bar{x}\alpha + \frac{1}{\lambda} n\bar{x}(1-\alpha)e^{-\lambda} = 0 \end{cases}$$

equivalente a

$$\begin{cases} f_0 - \alpha = (1-\alpha)e^{-\lambda} \\ \dfrac{\bar{x}}{\lambda} = 1-\alpha \end{cases}$$

que por sua vez é equivalente a

$$\begin{cases} \frac{\overline{x}}{\lambda}(1-e^{-\lambda}) = 1 - \frac{n_0}{n} \\ \alpha = 1 - \frac{\overline{x}}{\lambda} \end{cases}$$

Há que verificar que a solução do sistema anterior maximiza a função de log-verosimilhança. Para tal usamos o teste da segunda derivada. A saber, no ponto crítico $(\widehat{\alpha},\widehat{\lambda})$ (que anula as primeiras derivadas) se o determinante da matriz Hessiana for positivo e se $\frac{\partial^2 \log L}{\partial \alpha^2}(\widehat{\alpha},\widehat{\lambda}) < 0$, então $\log L$ terá um máximo nesse ponto.

Calculamos de seguida as derivadas parciais de segunda ordem da função $\log L(\alpha,\lambda)$. Tem-se,

$$\frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\lambda) = \frac{-(1-e^{-\lambda})^2}{(\alpha+(1-\alpha)e^{-\lambda})^2} - \frac{n-n_0}{(1-\alpha)^2}$$

$$\frac{\partial^2 \log L}{\partial \alpha \partial \lambda}(\alpha,\lambda) = n_0 \frac{e^{-\lambda}}{(\alpha+(1-\alpha)e^{-\lambda})^2}$$

$$\frac{\partial^2 \log L}{\partial \lambda \partial \alpha}(\alpha,\lambda) = n_0 \frac{e^{-\lambda}}{(\alpha+(1-\alpha)e^{-\lambda})^2}$$

$$\frac{\partial^2 \log L}{\partial \lambda^2}(\alpha,\lambda) = n_0 \frac{(1-\alpha)\alpha^{-\lambda}}{(\alpha+(1-\alpha)e^{-\lambda})^2} - n\frac{\overline{x}}{\lambda_2}$$

Note-se que $\frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\lambda) < 0$ para quaisquer λ e α. No ponto crítico $(\widehat{\alpha},\widehat{\lambda})$ o determinante da matriz Hessiana é então

$$H(\hat{\alpha},\hat{\lambda}) = \begin{vmatrix} -\frac{n_0}{f_0^2}(\frac{1-f_0}{\hat{\lambda}}\bar{x})^2 - \frac{n-n_0}{\bar{x}^2}\hat{\lambda}^2 & \frac{n_0 e^{-\hat{\lambda}}}{f_0^2} \\ \frac{n_0 e^{-\hat{\lambda}}}{f_0^2} & -\frac{n_0\hat{\alpha}(\hat{\alpha}-f_0)}{f_0^2} - n\frac{\bar{x}}{\hat{\lambda}^2} \end{vmatrix}$$

$$= \frac{n_0^2}{f_0^4}(1-f_0)\left[\frac{\overline{x}_2}{\widehat{\lambda}^2}f_0(1-\hat{\alpha})(f_0-\hat{\alpha}) + \hat{\alpha}(f_0-\widehat{\alpha})\widehat{\lambda}^2 f_0 + \frac{\overline{x}^3}{\hat{\lambda}^4}f_0(1-f_0) + \bar{x}\hat{\lambda}f_0^2\right]$$

Tem-se então $H(\widehat{\alpha},\widehat{\lambda}) > 0$, pois a igualdade $f_0 = \hat{\alpha} + (1-\hat{\alpha})e^{-\hat{\lambda}}$ implica $f_0 > \widehat{\alpha}$.

Relativamente à variável InglFora, o contraste observado na Tabela 2.2 entre n_0 e $nP(Y=0)$ sugere-nos a utilização deste modelo.

Utilizando os 600 valores desta variável que se encontra em anexo o sistema

$$\begin{cases} \bar{x}(1-e^{-\hat{\lambda}}) = \widehat{\lambda}(1-\frac{n_0}{n}) \\ \widehat{\alpha} = 1 - \frac{\bar{x}}{\widehat{\lambda}} \end{cases}$$

tem solução

$$\begin{cases} \widehat{\lambda} = 1.3059 \\ \widehat{\alpha} = 0.0696 \end{cases}$$

Assim o modelo proposto é tal que

$$\begin{aligned} P(X=x) &= \widehat{\alpha}\,\mathbb{1}_{\{0\}}(x) + (1-\widehat{\alpha})e^{-\widehat{\lambda}}\frac{\widehat{\lambda}^x}{x!} \\ &= 0.0696\,\mathbb{1}_{\{0\}}(x) + (1-0.0696)e^{-1.3059}\frac{1.3059^x}{x!},\ x \in \mathbb{N}_0 \end{aligned}$$

Com este modelo obtemos as frequências esperadas que se encontram na tabela seguinte

	Freq obs	**Freq esp**
0	193	193.001
1	203	197.513
2	120	128.963
3	56	56.136
4	23	18.327
5	4	4.786
6	1	1.042

Tabela 2.5 Frequências observadas e frequências esperadas sobre a validade de uma distribuição ZIP para a variável InglFora

Realizado o teste de ajustamento do qui-quadrado obtemos o valor observado da estatística de teste $\chi^2_{obs} = 2.0974$ ao qual corresponde o p-valor=0.925.

Concluímos assim que este modelo é mais adequado do que o modelo Poisson tradicional para proceder a previsão de estatísticas relacionadas com a variável InglFora.

2.3 Poisson deflacionado em dois

Com o mesmo objetivo, o de corrigir ou melhorar o ajustamento de um modelo tradicional de Poisson aos dados de que dispomos, nesta secção introduzimos e estudamos o modelo de Poisson deflacionado em dois. Este modelo é motivado pelo facto de se verificar uma deflação no valor dois, acentuada ou moderada, como se pode verificar nas tabelas 2.1-2.4, ou mesmo na Tabela 2.5 depois de aplicado o modelo ZIP.

Assim, introduzimos a variável aleatória com lei de Poisson deflacionado em dois, como sendo a que tem função de probabilidade caracterizada por

$$P(X=x) = -\beta\,\mathbb{1}_{\{2\}}(x) + (1+\beta)P(Z=x),\ x \in \mathbb{N}_0,$$

onde $Z \sim \mathscr{P}(\lambda)$ e os parâmetros $\beta \in]0,1[$ e $\lambda > 0$ são tais que $-\beta \mathbb{1}_{\{2\}}(x) + (1+\beta)P(Z=x)$ pertence a $]0,1[$, para qualquer x em $\mathbb{N}_0$. Concretamente $(1+\beta)P(Z=x) < 1$ e $-\beta + (1+\beta)e^{-\lambda}\frac{\lambda^2}{2} > 0$.

Na Figura 2.1 apresentamos o domínio dos parâmetros $\beta \in]0,0.4[$ e $\lambda \in]0.8,5[$ para os quais este requisito se verifica.

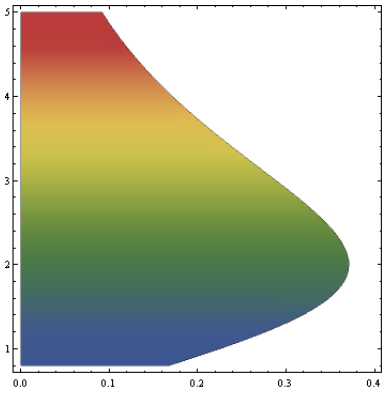

Fig. 2.1 Valores de $\beta \in]0,0.4[$ e $\lambda \in]0.8,5[$ para os quais $-\beta \mathbb{1}_{\{2\}}(x) + (1+\beta)P(Z=x) \in]0,1[$

Atendendo a que

$$\begin{cases} E(X) = -2\beta + (1+\beta)\lambda \\ \\ E(X^2) = -4\beta + (1+\beta)(\lambda^2+\lambda) \end{cases}$$

para valores de β e de λ onde estes momentos são efetivamente positivos, este sistema é equivalente a

$$\begin{cases} \dfrac{E(X)^2+4\beta}{E(X)+2\beta} = \dfrac{E(X)+3\beta+1}{1+\beta} \\ \\ \lambda = \dfrac{E(X)+2\beta}{1+\beta} \end{cases}$$

os estimadores gerados pelo método dos momentos são solução do sistema

$$\begin{cases} \dfrac{M_2+4\hat{\beta}}{M_1+2\hat{\beta}} = \dfrac{M_1+3\hat{\beta}+1}{1+\hat{\beta}} \\ \\ \hat{\lambda} = \dfrac{M_1+2\hat{\beta}}{1+\hat{\beta}} \end{cases}$$

Vamos de seguida obter os estimadores de β e de λ pelo método de máxima verosimilhança. A função de log-verosimilhança é

$$\begin{aligned}\log L(\beta,\lambda) = &\sum_{x_i=2} \log(-\beta + (1+\beta)e^{-\lambda}\frac{\lambda^2}{2}) \\ &+ \sum_{x_i\neq 2} \log((1+\beta)e^{-\lambda}\frac{\lambda^{x_i}}{x_i!})\end{aligned}$$

Derivando a função em função de β obtemos

$$\begin{aligned}\frac{\partial \log L}{\partial \beta}(\beta,\lambda) &= \frac{-1+e^{-\lambda}\frac{\lambda^2}{2}}{-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2}}n_2 + \frac{1}{1+\beta}(n-n_2) \\ &= \frac{(e^{-\lambda}\frac{\lambda^2}{2}-1)}{e^{-\lambda}\frac{\lambda^2}{2}+\beta(-1+e^{-\lambda}\frac{\lambda^2}{2})}n_2 + \frac{1}{1+\beta}(n-n_2) = 0 \\ &\iff (1+\beta)(e^{-\lambda}\frac{\lambda^2}{2}-1)n_2 + (n-n_2)(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2}) = 0 \\ &\iff (1+\beta)e^{-\lambda}\frac{\lambda^2}{2} = f_2+\beta\end{aligned}$$

Derivando a função $\log L$ em função de λ obtemos

$$\begin{aligned}&\frac{\partial \log L}{\partial \lambda}(\beta,\lambda) = 0 \\ &\iff \frac{(e^{-\lambda}\lambda - \frac{\lambda^2}{2}e^{-\lambda})(1+\beta)}{-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2}}n_2 + \sum_{x_i\neq 2}(-1+\frac{x_i}{\lambda}) = 0 \\ &\iff (e^{-\lambda}\lambda - \frac{\lambda^2}{2}e^{-\lambda})(1+\beta)n_2 + (-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})\sum_{x_i\neq 2}(\frac{x_i}{\lambda}-1) = 0\end{aligned}$$

Considerando que a solução da equação $\frac{\partial \log L}{\partial \beta}(\beta,\lambda) = 0$ é $(1+\beta)e^{-\lambda}\frac{\lambda^2}{2} = f_2+\beta$, tem-se

$$\begin{aligned}&\frac{\partial \log L}{\partial \lambda}(\beta,\lambda) = 0 \\ &\iff ne^{-\lambda}\lambda(1+\beta) - n\beta + \frac{1}{\lambda}\sum_{x_i\neq 2} x_i - n = 0 \\ &\iff \lambda(1+\beta) - 2\frac{n_2}{n} - 2\beta = \frac{1}{n}(n\bar{x}-2n_2) \\ &\iff \lambda(1+\beta) - 2\beta = \bar{x}\end{aligned}$$

Obtemos então o seguinte sistema

$$\begin{cases} \widehat{\lambda}(1+\widehat{\beta}) - 2\hat{\beta} = \bar{x} \\ (1+\widehat{\beta})e^{-\widehat{\lambda}}\dfrac{\hat{\lambda}^2}{2} = f_2 + \beta \end{cases}$$

equivalente a

$$\begin{cases} 1+\widehat{\beta} = \frac{\bar{x}-2}{\widehat{\lambda}-2} \\ \frac{\bar{x}-2}{\widehat{\lambda}-2}e^{-\widehat{\lambda}}\frac{\widehat{\lambda}^2}{2} = f_2 + \frac{\bar{x}-2}{\widehat{\lambda}-2} \end{cases}$$

ou ainda a

$$\begin{cases} \widehat{\beta} = \frac{\bar{x}-\widehat{\lambda}}{\widehat{\lambda}-2} \\ (\bar{x}-2)e^{-\widehat{\lambda}}\frac{\widehat{\lambda}^2}{2} = \widehat{\lambda}(f_2-1) + \bar{x} - 2f_2 \end{cases}$$

Provemos que a função de log-verosimilhança assume um máximo em $(\widehat{\beta},\widehat{\lambda})$.

As derivadas de segunda ordem são

$$\frac{\partial^2 \log L}{\partial \beta^2}(\beta,\lambda) = \frac{-(e^{-\lambda}\frac{\lambda^2}{2}-1)^2}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}n_2 - \frac{1}{(1+\beta)^2}(n-n_2)$$

$$\frac{\partial^2 \log L}{\partial \beta \partial \lambda}(\beta,\lambda) = \frac{e^{-\lambda}\lambda - e^{-\lambda}\frac{\lambda^2}{2}}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}n_2$$

$$\frac{\partial^2 \log L}{\partial \lambda \partial \beta}(\beta,\lambda) = \frac{e^{-\lambda}\lambda - e^{-\lambda}\frac{\lambda^2}{2}}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}n_2$$

$$\begin{aligned}\frac{\partial^2 \log L}{\partial \lambda^2}(\beta,\lambda) = {} & n_2(1+\beta)\frac{(-e^{-\lambda}\lambda + e^{-\lambda} + e^{-\lambda}\frac{\lambda^2}{2} - e^{-\lambda})(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2} \\ & - n_2(1+\beta)(1+\beta)e^{-2\lambda}(\lambda - \frac{\lambda^2}{2})^2 - \frac{1}{\lambda^2}(n\bar{x} - 2n_2)\end{aligned}$$

Assim o determinante da matriz Hessiana no ponto genérico (β,λ) é dado por

$$H(\beta,\lambda) = \begin{vmatrix} \frac{-(e^{-\lambda}\frac{\lambda^2}{2}-1)^2}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}n_2 - \frac{1}{(1+\beta)^2}(n-n_2) & \frac{e^{-\lambda}\lambda - e^{-\lambda}\frac{\lambda^2}{2}}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}n_2 \\ \frac{e^{-\lambda}\lambda - e^{-\lambda}\frac{\lambda^2}{2}}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}n_2 & \frac{n_2(1+\beta)T(\lambda,\beta)}{(-\beta+(1+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2} - \frac{1}{\lambda^2}(n\bar{x}-2n_2) \end{vmatrix}$$

onde $T(\beta,\lambda) = e^{-\lambda}(\beta(2\lambda - 1 - \frac{\lambda^2}{2} - e^{-\lambda}\frac{\lambda^2}{2}) - e^{-\lambda}\frac{\lambda^2}{2})$.

Observamos que a determinação do sinal de $H(\beta,\lambda)$ quando se substitui por $(\widehat{\beta},\widehat{\lambda})$ se mostrou bastante complexo dado que é impossível estabelecer uma relação para os valores de $n_0, n_2, \bar{x}$, etc, para qualquer amostra. Será então mais fácil determinar um domínio onde $H(\beta,\lambda) > 0$ e avaliar a inclusão de $(\widehat{\beta},\widehat{\lambda})$ no final.

Ora este determinante é igual à soma de quatro parcelas, que envolvem quatro determinantes, ou seja

$$\frac{n_2^2}{(-\beta+(1+\beta)e^{-\lambda\frac{\lambda^2}{2}})^4}\begin{vmatrix} -(e^{-\lambda}\frac{\lambda^2}{2}-1)^2 & e^{-\lambda}\lambda-e^{-\lambda}\frac{\lambda^2}{2} \\ e^{-\lambda}\lambda-e^{-\lambda}\frac{\lambda^2}{2} & (1+\beta)T(\lambda,\beta) \end{vmatrix}+$$

$$+\frac{n_2^2}{(-\beta+(1+\beta)e^{-\lambda\frac{\lambda^2}{2}})^4}\begin{vmatrix} -\frac{n-n_2}{(1+\beta)^2} & e^{-\lambda}\lambda-e^{-\lambda}\frac{\lambda^2}{2} \\ 0 & (1+\beta)T(\lambda,\beta) \end{vmatrix}$$

$$+\begin{vmatrix} -\frac{n-n_2}{(1+\beta)^2} & 0 \\ 0 & -\frac{1}{\lambda^2}(n\bar{x}-2n_2) \end{vmatrix}$$

$$+\frac{n_2^2}{(-\beta+(1+\beta)e^{-\lambda\frac{\lambda^2}{2}})^4}\begin{vmatrix} -(e^{-\lambda}\frac{\lambda^2}{2}-1)^2 & 0 \\ e^{-\lambda}\lambda-e^{-\lambda}\frac{\lambda^2}{2} & -\frac{1}{\lambda^2}(n\bar{x}-2n_2) \end{vmatrix}$$

Tem-se $n\bar{x}-2n_2>0$ equivalente a $\bar{x}>2f_2$, o que se verifica para qualquer amostra, o que permite concluir que o terceiro e o quarto determinantes são positivos.

Notemos que o segundo determinante é positivo se $T(\beta,\lambda)<0$. Por outro lado, o primeiro determinante é igual a

$$(e^{-\lambda}\frac{\lambda^2}{2}-1)^2(1+\beta)(-T(\lambda,\beta))-\left(e^{-\lambda}\lambda-e^{-\lambda}\frac{\lambda^2}{2}\right)^2$$

o qual é positivo se e só se $(1+\beta)(-T(\lambda,\beta))>D(\lambda)$, com $D(\lambda)=\left(e^{-\lambda}\lambda-e^{-\lambda}\frac{\lambda^2}{2}\right)^2/(e^{-\lambda}\frac{\lambda^2}{2}-1)^2$, onde também se exige $T(\beta,\lambda)<0$.

Estudemos a função $T(\beta,\lambda)$.

Vejamos que

$$T(\lambda,\beta)=e^{-\lambda}(\beta(2\lambda-1-\frac{\lambda^2}{2}-e^{-\lambda}\frac{\lambda^2}{2})-e^{-\lambda}\frac{\lambda^2}{2})<0$$
$$\Longleftrightarrow \beta(2\lambda-1-\frac{\lambda^2}{2}-e^{-\lambda}\frac{\lambda^2}{2})<e^{-\lambda}\frac{\lambda^2}{2}$$

Como as funções $2\lambda-1-\frac{\lambda^2}{2}-e^{-\lambda}\frac{\lambda^2}{2}$ e $e^{-\lambda}\frac{\lambda^2}{2}$ são crescentes em $]0.8,2]$, a desigualdade anterior verifica-se pelo menos se

$$\beta<\min_{0.8<\lambda<2}\frac{e^{-\lambda}\frac{\lambda^2}{2}}{2\lambda-1-\frac{\lambda^2}{2}-e^{-\lambda}\frac{\lambda^2}{2}}$$

isto é, se $\beta<0.3742$. Na Figura 2.2 é ilustrado o domínio de (β,λ) em $]0,0.3]\times[0.8,2]$, para o qual $T(\beta,\lambda)<0$ e $(1+\beta)(-T(\beta,\lambda))>D(\lambda)$. Denotemos tal domínio por $\mathscr{R}$.

Assim, em $\mathscr{R}$ estes quatro determinantes são todos positivos o que significa que $H(\widehat{\beta},\widehat{\lambda})>0$, desde que $(\widehat{\beta},\widehat{\lambda})\in\mathscr{R}$. Como $\frac{\partial^2\log L}{\partial\beta^2}(\widehat{\beta},\widehat{\lambda})<0$, de acordo com o teste da segunda derivada, fica provado que os estimadores obtidos maximizam a função de log-verosimilhança.

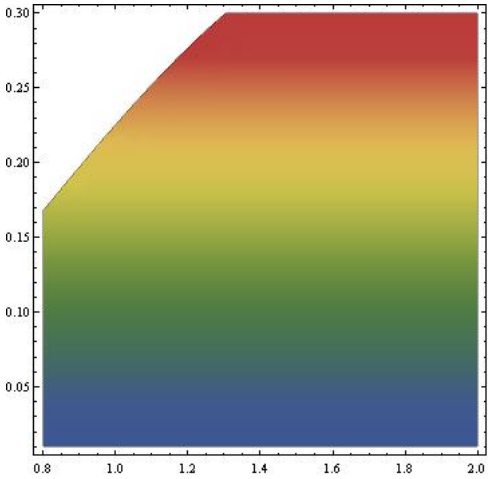

Fig. 2.2 Valores de $\beta \in]0,0.3]$ e de $\lambda \in [0.8,2]$ para os quais $T(\beta,\lambda) < 0$ e $(1+\beta)(-T(\beta,\lambda)) > D(\lambda)$.

Com as 600 observações disponíveis da variável InglCasa obtemos as estimativas da máxima verosimilhança

$$\begin{cases} \hat{\lambda} = 1.562 \\ \hat{\beta} = 0.0392 \end{cases}$$

os quais pertencem a $\mathscr{R}$. Assim o modelo proposto é descrito por

$$\begin{aligned} P(X=x) &= -\hat{\beta}\,\mathbb{1}_{\{2\}}(x) + (1+\hat{\beta})e^{-\hat{\lambda}}\frac{\hat{\lambda}^x}{x!} \\ &= -0.0392\,\mathbb{1}_{\{2\}}(x) + (1+0.0392)e^{-1.562}\frac{1.562^x}{x!}, x \in \mathbb{N}_0 \end{aligned}$$

Na Tabela 2.6 encontram-se os valores das frequências observadas e das frequências esperadas sob a validade da hipótese de a variável InglCasa seguir uma lei de Poisson deflacionada em dois com $\lambda = 1.562$ e $\beta = 0.0392$.

Realizado o teste de ajustamento do qui-quadrado obtemos os valor observado da estatística de teste $\chi^2_{obs} = 2.6686$ ao qual corresponde o p-valor=0.8536.

Como este p-valor é superior ao que se obtém considerando o modelo de Poisson tradicional (Tabela 2.1 e p-valor igual a 0.47), concluímos que o ajustamento aqui apresentado é significativamente melhor.

2.4 Poisson inflacionado em zero e deflacionado em dois

Uma análise ainda mais cuidada das tabelas de frequências já apresentas e a constatação de que ao inflacionarmos em zero podemos deflacionar acentuadamente outros valores, bem como se defla-

	Freq obs	**Freq esp**
0	133	130.742
1	201	204.241
2	136	135.999
3	84	83.071
4	33	32.443
5	8	10.136
6	5	2.639

Tabela 2.6 Frequências observadas e frequências esperadas sob a validade do modelo de Poisson deflacionado em dois para a variável Inglcasa

cionarmos em dois podemos inflacionar demasiado a ocorrência de outros resultados, nesta secção, introduzimos e estudamos um outro modelo. Trata-se do modelo de Poisson inflacionado em zero e deflacionado em dois.

A saber, uma variável aleatória X possui lei de Poisson inflacionada em zero e deflacionada em dois se o seu suporte é $\mathbb{N}_0$ e

$$P(X=x) = \alpha \mathbb{1}_{\{0\}}(x) - \beta \mathbb{1}_{\{2\}}(x) + (1-\alpha+\beta)P(Z=x), x \in \mathbb{N}_0$$

onde Z segue a lei de Poisson de parâmetro λ, sendo $\beta \in]0,1[$, $\alpha \in]0,1[$ e $\lambda > 0$ tais que $\alpha \mathbb{1}_{\{0\}}(x) - \beta \mathbb{1}_{\{2\}}(x) + (1-\alpha+\beta)P(Z=x) \in]0,1[$, para qualquer $x \in \mathbb{N}_0$. Concretamente,

$$\begin{cases} 1-\alpha+\beta > 0 \\ (1-\alpha+\beta)P(Z=x) < 1, x \notin \{0,2\} \\ \alpha + (1-\alpha+\beta)e^{-\lambda} < 1 \\ -\beta + (1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2} > 0. \end{cases}$$

Na Figura 2.3 apresentamos o domínio dos parâmetros $\alpha \in]0,0.2[$, $\beta \in]0,0.2[$ e $\lambda \in]0.8,2[$ para os quais $\alpha \mathbb{1}_{\{0\}}(x) - \beta \mathbb{1}_{\{2\}}(x) + (1-\alpha+\beta)P(Z=x)$ é efetivamente uma probabilidade.

Comecemos por observar que uma vez que se tem

$$\begin{cases} E(X) = -2\beta + (1-\alpha+\beta)\lambda \\ E(X^2) = -4\beta + (1-\alpha+\beta)\lambda \\ E(X^3) = -8\beta + (1-\alpha+\beta)\lambda \end{cases}$$

para valores de α, β e λ tais que estas expressões definem momentos positivos, o sistema anterior é equivalente a

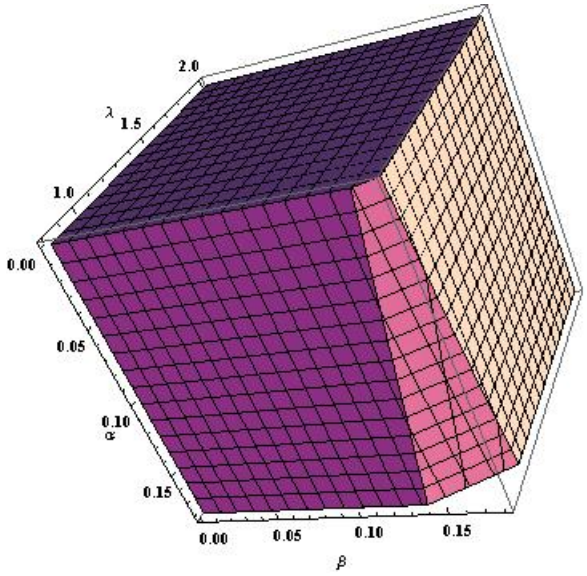

Fig. 2.3 Valores de $\alpha \in]0,0.2[$, $\beta \in]0,0.2[$ e $\lambda \in]0.8,2[$ para os quais $\alpha \mathbb{1}_{\{0\}}(x) - \beta \mathbb{1}_{\{2\}}(x) + (1-\alpha+\beta)P(Z=x) \in]0,1[$

$$\begin{cases} \dfrac{E(X^3)-2E(X^2)}{E(X^2)-2E(X)} = 1 + \dfrac{\lambda^2}{\lambda-1} \\ \\ E(X^2)-2E(X) = (1-\alpha+\beta)(\lambda^2-\lambda) \\ \\ E(X^3)-4E(X) = (1-\alpha+\beta)(\lambda^3+3\lambda^2-3\lambda) \end{cases}$$

Assim o Método dos Momentos gera os estimadores que são solução de

$$\begin{cases} \dfrac{M_3-2M_2}{M_2-2M_1} = 1 + \dfrac{\widehat{\lambda}^2}{\hat{\lambda}-1} \\ \\ M_2-2M_1 = (1-\widehat{\alpha}+\hat{\beta})(\widehat{\lambda}^2-\hat{\lambda}) \\ \\ M_3-4M_1 = (1-\widehat{\alpha}+\hat{\beta})(\widehat{\lambda}^3+3\hat{\lambda}^2-3\hat{\lambda}) \end{cases}$$

Passamos a estudar os estimadores da máxima verosimilhança. A função de log-verosimilhança é

$$\log L(\alpha,\beta,\lambda) = n_2 \log(-\beta + (1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2}) + n_0 \log(\alpha + (1-\alpha+\beta)e^{-\lambda}) + \sum_{x_i \neq 2, x_i \neq 0} \log(e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}) + \log((1-\alpha+\beta))$$

O sistema

$$\begin{cases} \frac{\partial \log L}{\partial \alpha}(\lambda,\alpha,\beta)=0 \\ \\ \frac{\partial \log L}{\partial \beta}(\lambda,\alpha,\beta)=0 \\ \\ \frac{\partial \log L}{\partial \lambda}(\lambda,\alpha,\beta)=0 \end{cases}$$

dá lugar a

$$\begin{cases} n_0\dfrac{1-e^{-\lambda}}{\alpha+(1-\alpha+\beta)e^{-\lambda}}(1-\alpha+\beta)+n_2\dfrac{e^{-\lambda}\frac{\lambda^2}{2}(1-\alpha+\beta)}{-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2}}=n-n_0-n_2 \\ \\ n_0\dfrac{e^{-\lambda}}{\alpha+(1-\alpha+\beta)e^{-\lambda}}(1-\alpha+\beta)+n_2\dfrac{(e^{-\lambda}\frac{\lambda^2}{2}-1)(1-\alpha+\beta)}{-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2}}=0 \\ \\ -n_0\dfrac{e^{-\lambda}}{\alpha+(1-\alpha+\beta)e^{-\lambda}}(1-\alpha+\beta)+n_2\dfrac{(-e^{-\lambda}\frac{\lambda^2}{2}+e^{\lambda}\lambda)(1-\alpha+\beta)}{-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2}}=n-n_0-n_2-\dfrac{1}{\lambda}(n\bar{x}-2n_2) \end{cases}$$

equivalente a

$$\begin{cases} n_2\alpha+n_0\beta=(1-\alpha+\beta)(e^{-\lambda}\dfrac{\lambda^2}{2}-e^{-\lambda}) \\ \\ n_2\dfrac{(e^{-\lambda}\lambda-1)}{f_2}(1-\alpha+\beta)+\dfrac{1}{\lambda}(n\bar{x}-2n_2)=0 \\ \\ n_2\dfrac{(1-e^{-\lambda}-e^{-\lambda}\frac{\lambda^2}{2})}{-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2}}=\dfrac{n-n_0-n_2}{1-\alpha+\beta} \end{cases}$$

o qual conduz assim aos estimadores $\widehat{\lambda}$, $\widehat{\alpha}$ e $\widehat{\beta}$ que verificam

$$\begin{cases} \dfrac{1}{\widehat{\lambda}}+\widehat{\lambda}-e^{\widehat{\lambda}}=\dfrac{f_0+2f_2-\bar{x}}{f_1} \\ e^{\widehat{\lambda}}-1-\dfrac{\widehat{\lambda}^2}{2}=\dfrac{1-f_0-f_2}{f_0-\widehat{\alpha}} \\ \widehat{\alpha}-2\widehat{\beta}+(1-\widehat{\alpha}+\widehat{\beta})\widehat{\lambda}=\bar{x} \end{cases}$$

Há que verificar que a solução do sistema anterior maximiza a função de log-verosimilhança. Calculamos de seguida as derivadas parciais de segunda ordem da função $\log L(\alpha,\beta,\lambda)$. Tem-se

$$\frac{\partial^2 \log L}{\partial \alpha \partial \beta}(\alpha,\beta,\lambda) = -n\left[f_0\frac{(e^{-\lambda}-1)}{(\alpha+(1-\alpha+\beta)e^{-\lambda})^2}e^{-\lambda} - f_2\frac{(e^{-\lambda}\frac{\lambda^2}{2}-1)e^{-\lambda}\frac{\lambda^2}{2}}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2} + (1-f_0-f_2)\frac{1}{(1-\alpha+\beta)^2}\right]$$

$$\frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\beta,\lambda) = -n\left[f_0\frac{(1-e^{-\lambda})^2}{(\alpha+(1-\alpha+\beta)e^{-\lambda})^2} + f_2\frac{(e^{-\lambda}\frac{\lambda^2}{2})^2}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2} - (1-f_0-f_2)\frac{1}{(1-\alpha+\beta)^2}\right]$$

$$\frac{\partial^2 \log L}{\partial \beta^2}(\alpha,\beta,\lambda) = -n\left[f_0\frac{(e^{-\lambda})^2}{(\alpha+(1-\alpha+\beta)e^{-\lambda})^2} + f_2\frac{(e^{-\lambda}\frac{\lambda^2}{2}-1)^2}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2} + (1-f_0-f_2)\frac{1}{(1-\alpha+\beta)^2}\right]$$

$$\frac{\partial^2 \log L}{\partial \alpha \partial \lambda}(\alpha,\beta,\lambda) = n\left[\frac{f_0(1+\beta)e^{-\lambda}}{(\alpha+(1-\alpha+\beta)e^{-\lambda})^2} + f_2\frac{\beta(-e^{-\lambda}\frac{\lambda^2}{2}+e^{-\lambda}\lambda)}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}\right]$$

$$\frac{\partial^2 \log L}{\partial \beta \partial \lambda}(\alpha,\beta,\lambda) = n\left[\frac{-f_0\alpha e^{-\lambda}}{(\alpha+(1-\alpha+\beta)e^{-\lambda})^2} - \frac{f_2(1+\beta)(e^{-\lambda}\frac{\lambda^2}{2}+e^{-\lambda}\lambda)}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}\right]$$

$$\frac{\partial^2 \log L}{\partial \lambda^2}(\alpha,\beta,\lambda) = n\left[\frac{f_0(1-\alpha+\beta)e^{-\lambda}\alpha}{(\alpha+(1-\alpha+\beta)e^{-\lambda})^2} + \frac{f_2(1-\alpha+\beta)(e^{-\lambda}\frac{\lambda^2}{2}f_0+\beta e^{-\lambda}(2\lambda-1))}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2}\right]$$

$$-nf_2\frac{(1-\alpha+\beta)^2(e^{-\lambda}\lambda)^2}{(-\beta+(1-\alpha+\beta)e^{-\lambda}\frac{\lambda^2}{2})^2} - \frac{n}{\lambda^2}(\bar{x}-f_2)$$

Recordemos que, de acordo com o teste da segunda derivada o ponto crítico $(\widehat{\alpha},\widehat{\beta},\widehat{\lambda})$ é maximizante da função de log-verosimilhança se os hessianos

$$\begin{vmatrix} \frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\beta,\lambda) & \frac{\partial^2 \log L}{\partial \alpha \beta}(\alpha,\beta,\lambda) \\ \frac{\partial^2 \log L}{\partial \beta \alpha}(\alpha,\beta,\lambda) & \frac{\partial^2 \log L}{\partial \beta^2}(\alpha,\beta,\lambda) \end{vmatrix}$$

e

$$\begin{vmatrix} \frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\beta,\lambda) & \frac{\partial^2 \log L}{\partial \alpha \partial \beta}(\alpha,\beta,\lambda) & \frac{\partial^2 \log L}{\partial \alpha \partial \lambda}(\alpha,\beta,\lambda) \\ \frac{\partial^2 \log L}{\partial \beta \partial \alpha}(\alpha,\beta,\lambda) & \frac{\partial^2 \log L}{\partial \beta^2}(\alpha,\beta,\lambda) & \frac{\partial^2 \log L}{\partial \beta \partial \lambda}(\alpha,\beta,\lambda) \\ \frac{\partial^2 \log L}{\partial \lambda \partial \alpha}(\alpha,\beta,\alpha) & \frac{\partial^2 \log L}{\partial \lambda \partial \beta}(\alpha,\beta,\alpha) & \frac{\partial^2 \log L}{\partial \lambda^2}(\alpha,\beta,\alpha) \end{vmatrix}$$

são positivos e $\frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\beta,\lambda) < 0$, nesse mesmo ponto crítico. Não tendo provado estas condições analiticamente ilustramos na Figura 2.4 o domínio onde tal se verifica

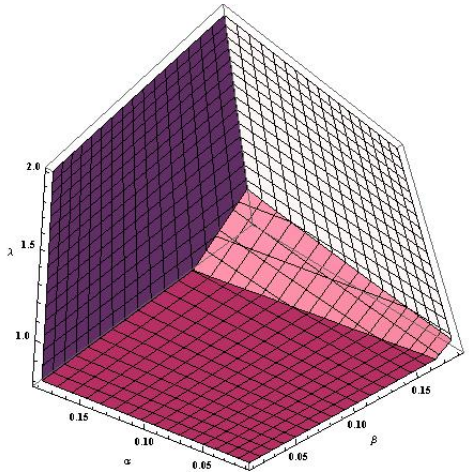

Fig. 2.4 Valores de $\alpha \in]0,0.2[$, $\beta \in]0,0.2[$ e $\lambda \in]0.8,2[$ para os quais os Hessianos são positivos e $\frac{\partial^2 \log L}{\partial \alpha^2}(\alpha,\beta,\lambda) < 0$

Com os 474 valores amostrais para a variável PortCasa obtivemos

$$\begin{cases} \hat{\lambda} = 1.513 \\ \hat{\beta} = 0.072 \\ \hat{\alpha} = 0.043 \end{cases}$$

o qual pertence à região ilustrada nas Figuras 2.3 e 2.4.

Obtemos assim o modelo descrito por

$$\begin{aligned} P(X=x) &= \hat{\alpha}\mathbb{1}_{\{0\}}(x) - \hat{\beta}\mathbb{1}_{\{2\}}(x) + (1-\hat{\alpha}+\hat{\beta})e^{-\hat{\lambda}}\frac{\hat{\lambda}^x}{x!} \\ &= 0.043\mathbb{1}_{\{0\}}(x) - 0.072\mathbb{1}_{\{2\}}(x) + (1-0.043+0.072)e^{-1.513}\frac{1.513^x}{x!}, x \in \mathbb{N}_0. \end{aligned}$$

Com este modelo obtemos as frequências esperadas que se encontram na Tabela 2.7.

	Freq	**Poisson**
0	124	127.880
1	153	162.502
2	102	88.824
3	64	61.990
4	18	23.446
5	9	7.094
6	4	1.789

Tabela 2.7 Frequências observadas e frequências esperadas sob a validade do modelo de Poisson inflacionado em zero e deflacionado em dois para a variável PortCasa

As frequências esperadas sob a validade deste modelo são agora mais próximas das observadas. Com efeito, ao realizar o teste de ajustamento do qui-quadrado obtemos os valor observado da estatística de teste $\chi^2_{obs} = 7.1957$ ao qual corresponde o p-valor=0.2994. Recordamos que no caso em que se considera o modelo de Poisson se obteve p-valor=0.09.

Com os 600 valores amostrais para a variável InglCasa obtivemos

$$\begin{cases} \widehat{\lambda} = 1.569 \\ \widehat{\beta} = 0.047 \\ \widehat{\alpha} = 0.007 \end{cases}$$

Obtemos assim o modelo descrito por:

$$\begin{aligned} P(X=x) &= \hat{\alpha}\mathbb{1}_{\{0\}}(x) - \hat{\beta}\mathbb{1}_{\{2\}}(x) + (1-\hat{\alpha}+\hat{\beta})e^{-\hat{\lambda}}\frac{\hat{\lambda}^x}{x!} \\ &= 0.007\mathbb{1}_{\{0\}}(x) - 0.047\mathbb{1}_{\{2\}}(x) + (1-0.007+0.047)e^{-1.569}\frac{1.569^x}{x!}, x \in \mathbb{N}_0. \end{aligned}$$

Com este modelo obtemos as frequências esperadas que se encontram na Tabela 2.8.

As frequências esperadas são agora mais próximas das observadas. Com efeito, ao realizar o teste de ajustamento do qui-quadrado obtemos os valor observado da estatística de teste $\chi^2 = 2.6736$ ao qual corresponde o p-valor=0.8396. Notamos que este ajustamento não é melhor do que o que foi realizado com o modelo deflacionado em dois. Para além de uma ligeira descida no p-valor, podemos observar que $\widehat{\alpha} = 0.007$ e $1 - \widehat{\alpha} + \widehat{\beta}$ é aproximadamente $1 + \widehat{\beta}$.

	Freq	**Poisson**
0	133	134.197
1	201	203.837
2	136	131.723
3	84	83.670
4	33	32.827
5	8	10.303
6	5	2.695

Tabela 2.8 Frequências observadas e frequências esperadas sob a validade do modelo de Poisson inflacionado em 0 e deflacionado em dois para a variável Inglcasa

2.5 Poisson Bivariado

O modelo de Poisson bivariado com margens dependentes (existência de uma correlação) é mais natural para estimar as probabilidades pretendidas, uma vez que um jogo de futebol consiste em duas equipas que se enfrentam.

O par (X,Y) tem lei de Poisson bivariada, $(X,Y) \sim \mathscr{BP}(\lambda_1,\lambda_2,\lambda_3)$, se a sua função de probabilidade é definida por

$$P(X=x,Y=y) = e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\sum_{i=0}^{min(x,y)}\binom{x}{i}\binom{y}{i}i!(\frac{\lambda_3}{\lambda_1\lambda_2})^i,\ (x,y)\in\mathbb{N}_0^2$$

Numa primeira abordagem referimos Maher [4], onde a estimação dos parâmetros λ_1 e λ_2 é feita recorrendo aos valores $\alpha_i\beta_j$ e $\gamma_i\delta_j$ do modelo de Poisson univariado. Concretamente, para cada jogo (i,j), Maher considera as variáveis aleatórias X e Y escritas na forma $X_{ij}=U_{ij}+W_{ij}$ e $Y_{ij}=V_{ij}+W_{ij}$ com U_{ij} e V_{ij} independentes e com médias $\alpha_i\beta_j$ e $\gamma_i\delta_j$, respetivamente. A variável comum W_{ij} introduz a dependência.

As probabilidades incluídas no Capítulo 3, foram obtidas a partir deste modelo.

Numa segunda abordagem e com um objetivo que termina na secção seguinte, vamos aplicar este modelo aos dados emparelhados globais do campeonato inglês.

Comecemos por observar que

$$\begin{cases} E(X)=\lambda_1+\lambda_3 \\ E(Y)=\lambda_2+\lambda_3 \\ Cov(X,Y)=\lambda_3 \end{cases}$$

Usando o método dos momentos, os estimadores $\widehat{\lambda_1}$, $\widehat{\lambda_2}$ e $\widehat{\lambda_3}$ são solução de

$$\begin{cases} \overline{X}=\widehat{\lambda}_1+\widehat{\lambda}_3 \\ \overline{Y}=\widehat{\lambda}_2+\widehat{\lambda}_3 \\ \widehat{Cov(X,Y)}=\widehat{\lambda}_3 \end{cases}$$

A partir da amostra disponível nas tabelas A.1 e A.2 temos $\bar{x} = 1.545$, $\bar{y} = 1.215$ e $\widehat{Cov(X,Y)} = -0.089$. Assim, tem-se

$$\begin{cases} \widehat{\lambda}_3 = -0.089 \\ \widehat{\lambda}_1 = 1.545 + 0.089 = 1.634 \\ \widehat{\lambda}_2 = 1.215 + 0.089 = 1.304 \end{cases}$$

Na Tabela 2.9 apresentam-se as frequências observadas e as frequências esperadas admitindo que o par (InglCasa,InglFora) segue a lei $\mathscr{BP}(\lambda_1, \lambda_2, \lambda_3)$. O teste do qui-quadrado de ajustamento, quando aplicado à respetiva amostra bivariada gerou o valor $\chi^2_{obs} = 27.47$ e p-valor=0.234.

	Fora						
Casa	0	1	2	3	4	5	6
0	44 37.974	46 47.917	16 14.721	20 12.745	6 3.514	0 0.914	1 0.198
1	60 57.512	68 64.542	43 43.769	19 19.658	9 6.670	2 1.804	0 0.188
2	42 38.817	42 44.845	38 32.931	9 18.089	4 6.336	1 1.771	0 0.405
3	31 21.142	28 31.024	15 22.667	7 10.998	3 3.988	0 1.152	0 0.277
4	14 10.637	11 13.143	6 9.49	1 4.997	0 1.574	1 0.559	0 0.138
5	2 2.822	5 4.449	0 3.485	1 1.810	0 0.701	0 0.216	0 0.055
6	0 0.769	3 1.254	2 1.015	0 0.545	0 0.218	0 0.069	0 0.018

Tabela 2.9 Frequências observadas e esperadas sob um modelo de Poisson bivariado para o par (InglCasa, InglFora)

2.6 Poisson Bivariado inflacionado na diagonal

Karlis e Ntzoufras [2] propuseram o modelo de Poisson Bivariado inflacionado na diagonal, o qual, neste contexto, por ser designado por inflacionado nos empates. Sendo assim o modelo é uma forma mais geral do modelo de Poisson inflacionado em $(0,0)$. O modelo é então descrito pela função de probabilidade

$$P(X = x, Y = y) = \begin{cases} (1-p)P(Z = x, W = y) & se\ x \neq y \\ (1-p)P(Z = x, W = y) + pP(D = x) & se\ x = y \end{cases}$$

onde $(x,y) \in \mathbb{N}_0^2$, o par (Z,W) segue a lei de Poisson Bivariada e D segue uma lei discreta de suporte contido em {0,1,2,....}. Para p=0 temos o modelo de Poisson Bivariado estudado previamente. Como escolha para D temos, por exemplo, uma lei de Poisson, uma geométrica ou uma simples distribuição

discreta. A distribuição geométrica pode ser de grande interesse uma vez que tem moda em 0 e decresce rapidamente quando x aumenta. A distribuição Discreta apresenta-se na forma

$$P(D=x)=\begin{cases}\theta_x & \text{se } x=0,1,...,J\\ 0 & \text{se } x\neq 0,1,...,J\end{cases}$$

onde $\sum_{x=0}^{J}\theta_x=1$. Se J=0 temos o modelo inflacionado em $(0,0)$.

Estes modelos tem duas propriedades de grande importância. Primeiro, as distribuições marginais não são Poisson, mas sim misturas de distribuições com uma componente Poisson. A lei marginal de X e de Y é definida por

$$P(X=x)=(1-p)P(Z=x)+pP(D=x)$$

e

$$P(Y=x)=(1-p)P(W=x)+pP(D=x),x\in\mathbb{N}_0$$

onde Z e W seguem uma distribuição de Poisson. Assim, a distribuição de X é então uma mistura de uma distribuição de Poisson e de uma distribuição D. Por exemplo, se consideraremos a distribuição geométrica então resulta que a distribuição marginal é uma distribuição mista com uma componente Poisson e uma componente geométrica.

As média marginais são dadas por

$$E(X)=(1-p)(\lambda_1+\lambda_3)+pE(D)$$
$$E(Y)=(1-p)(\lambda_2+\lambda_3)+pE(D)$$

onde $E(D)$ é a media da distribuição D. A variância por sua vez é dada por

$$Var(X)=(1-p)[(\lambda_1+\lambda_3)^2+(\lambda_1+\lambda_3)]+pE(D^2)-[(1-p)(\lambda_1+\lambda_3)+pE(D)]^2$$

$$Var(Y)=(1-p)[(\lambda_2+\lambda_3)^2+(\lambda_2+\lambda_3)]+pE(D^2)-[(1-p)(\lambda_2+\lambda_3)+pE(D)]^2$$

Outra característica importante é o facto de, mesmo que $\lambda_3=0$, a distribuição resultante introduz um grau de dependência entre as duas variáveis. De forma geral, o modelo Poisson Bivariado apresenta $E(XY)=\lambda_3+(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)$.

Para um modelo inflacionado na diagonal, vem

$$\begin{aligned}COV(X,Y)=&(1-p)[\lambda_3+(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)]+pE(D^2)\\&-(1-p)^2(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)\\&-(1-p)pE(D)(\lambda_1+\lambda_2+2\lambda_3)-p^2E(D)^2\end{aligned}$$

Esta fórmula é uma generalização do caso mais simples onde a inflação é só em (0,0). Se $\lambda_3 = 0$ a covariância é dada por

$$COV(X,Y) = p(1-p)\lambda_1\lambda_2 + pE(D^2) - p(1-p)E(D)(\lambda_1+\lambda_2) - p^2E(D)^2$$

o que resulta numa correlação não nula entre X e Y.

Quando a inflação é só em (0,0), obtemos $E(D) = E(D^2) = 0$ e $COV(X,Y) = p(1-p)\lambda_1\lambda_2$ que é sempre positiva.

Relativamente ao campeonato Inglês, apresentamos a comparação entre as duas situações sem e com inflação nos empates, no caso mais simples em que se admite que $\lambda_3 = 0$. Isto é, no primeiro caso consideramos que as variáveis aleatórias $X \equiv InglCasa$ e $Y \equiv InglFora$ são independentes e seguem leis de Poisson e, no segundo caso, admitimos que estas variáveis possuiem a lei $\mathscr{BP}$ inflacionada na diagonal.

Consideramos apenas os empates $(0,0)$ e $(2,2)$, sendo estes igualmente distribuídos, isto é, $P(D=0) = P(D=2) = \frac{1}{2}$. Obtemos assim $E(D) = 1$ e $E(D^2) = 2$.

Então, com $\lambda_3 = 0$, vem

$$\begin{cases} E(X) = (1-p)\lambda_1 + p \\ E(Y) = (1-p)\lambda_2 + p \\ Cov(X,Y) = p(1-p)\lambda_1\lambda_2 + 2p - p(1-p)(\lambda_1+\lambda_2) - p^2 \end{cases}$$

donde resulta

$$\begin{cases} \lambda_1 = \dfrac{E(X)-p}{1-p} \\ \lambda_2 = \dfrac{E(Y)-p}{1-p} \\ Cov(X,Y) = \dfrac{p}{1-p}(E(X)-p)(E(Y)-p) - p(E(X)+E(Y)) + p^2 + 2p \end{cases}$$

Usando o método dos momentos, os estimadores $\widehat{\lambda_1}$, $\widehat{\lambda_2}$ e $\widehat{p}$ são solução de

$$\begin{cases} \overline{\lambda_1} = \dfrac{\overline{x}-\widehat{p}}{1-\widehat{p}} \\ \overline{\lambda_2} = \dfrac{\overline{y}-\widehat{p}}{1-\widehat{p}} \\ \widehat{Cov(X,Y)} = \dfrac{\widehat{p}}{1-\widehat{p}}(\overline{x}-\widehat{p})(\overline{y}-\widehat{p}) - \widehat{p}(\overline{x}+\overline{y}) + \widehat{p}^2 + 2\widehat{p} \end{cases}$$

Com a amostra de (X,Y) obtemos neste caso as estatísticas

$$\begin{cases} \widehat{\lambda_1} = 1.552 \\ \widehat{\lambda_2} = 1.218 \\ \widehat{p} = 0.013 \end{cases}$$

Para estabelecer a comparação já referida, notemos que no primeiro caso estamos a testar a hipótese de (X,Y) possuir margens independentes com leis de Poisson de parâmetro λ_1 e λ_2 estimados por $\widehat{\lambda_1} = 1.545$, $\widehat{\lambda_2} = 1.215$. No segundo caso, a hipótese nula do teste especifica uma lei de Poisson bivariada inflacionada em $(0,0)$ e $(2,2)$, com $\lambda_3 = 0$ e os restantes parâmetros assumidos como $\widehat{\lambda_1} = 1.552$, $\widehat{\lambda_2} = 1.218$ e $\widehat{p} = 0.013$.

Na Tabela 2.10 apresentam-se as frequências observadas e as frequências esperadas em ambos os casos, respetivamente. O teste do qui-quadrado de ajustamento, quando aplicado a esta amostra bivariada produz o valor observado $\chi^2_{obs} = 32.11$ e p-valor=0.10 no primeiro caso e $\chi^2_{obs} = 26.40$ e p-valor=0.28 no segundo caso.

Concluímos, com base também nos resultados da secção anterior, que da passagem do modelo bivariado com margens independentes para o que apresenta λ_3 estimado por -0.089 e depois para este último modelo inflacinado em empates (com $\lambda_3 = 0$), o p-valor passa respetivamente de 0.10 para 0.234 e depois para 0.28. Esta alteração no p-valor dá evidência à relevância deste último modelo.

Julgamos ser nosso dever afirmar que, considerando o caso em que λ_3 é não nulo neste último modelo com empates, não obtivemos resultados satisfatórios, usando o método dos momentos.

	InglFora					
InglCasa	0	1	2	3	4	$\geq$5
0	**44** **37.975** **44.012**	46 46.140 44.291	16 28.030 26.969	20 11.352 12.948	6 3.448 3.333	1 1.043 1.062
1	60 58.671 56.450	68 71.286 68.744	43 43.306 41.859	19 17.539 16.992	9 5.327 5.173	2 1.610 1.539
2	42 45.324 43.808	42 55.068 53.349	**38** **33.454** **40.131**	9 13.549 13.186	4 4.115 4.015	1 1.244 1.217
3	31 23.342 22.665	28 28.360 27.601	15 17.229 16.806	7 6.978 6.822	3 2.119 2.077	0 0.640 0.631
4	14 9.016 8.794	11 10.954 10.710	6 6.655 6.521	1 2.695 2.647	0 1	1 0.248 0,241
≥ 5	2 3.693 3.584	8 4.494 4.269	2 2.729 2.531	1 1.106 1.173	0 0.336 0.326 0	0 0.152 0.173

Tabela 2.10 Frequências observadas e esperadas sob um modelo de Poisson bivariado – sem e com inflação nos empates –para o par (InglCasa,InglFora)

Capítulo 3

Resultados

3.1 Resultados Odds

Começamos por apresentar as estimativas dos fatores de ataque e de defesa de todas as equipas dos campeonatos português e inglês, considerando os dados retirados até ao dia 20/02/2017 Usamos, como já referimos, a teoria exposta na secção 2.1. Os resultados encontram-se nas tabelas 3.1 e 3.2.

Observamos que as equipas Tondela, Chaves, Burnley, Hull e Middlesbrough, tendo mudado de divisão, têm valores inflacionados uma vez que foram as "melhores" do campeonato da segunda divisão do ano anterior.

Equipas	α Ataque Casa	β Defesa Fora	γ Defesa Casa	δ Ataque Fora
Arouca	1.405	1.418	1.184	1.176
Belenenses	1.104	1.731	1.626	1.133
Benfica	3.214	0.526	0.761	2.253
Boavista	0.702	1.418	1.055	1.133
Braga	2.439	0.964	1.098	0.839
Chaves	1.636	1.265	0.928	1.567
Estoril	1.179	1.418	1.184	0.964
FC Porto	2.397	0.743	0.802	2.019
Feirense	1.405	1.265	1.581	1.391
Maritimo	1.405	1.771	1.271	0.964
Moreirense	1.104	1.652	1.854	1.133
Nacional	1.329	1.851	1.626	0.757
Paços	1.254	1.731	1.227	1.305
Rio Ave	1.254	1.457	1.358	1.391
Sporting	2.356	0.890	0.802	2.397
Tondela	0.846	2.054	1.536	1.006
V.Guimaraes	1.367	1.613	1.402	1.523
V.Setúbal	0.993	1.891	1.314	1.391

Tabela 3.1 Fatores de ataque e defesa das equipas do campeonato português

Equipas	α Ataque Casa	β Defesa Fora	γ Defesa Casa	δ Ataque Fora
Arsenal	1.124	0.640	0.642	1.433
Bournemouth	0.898	1.232	1.063	0.689
Burnley	2.814	2.071	1.339	2.068
Chelsea	1.448	0.465	0.339	0.948
Crystal Palace	0.547	1.139	1.009	0.948
Everton	1.079	0.729	0.489	0.740
Hull	2.979	2.388	1.744	1.433
Leicester	0.678	1.185	0.797	0.387
Liverpool	1.355	0.819	0.539	1.160
Man City	0.988	0.685	0.642	1.433
Man United	0.810	0.509	0.439	0.948
Middlesbrough	2.030	1.617	1.118	1.949
Southampton	0.591	0.729	0.693	0.689
Stoke	0.678	0.909	0.745	0.689
Sunderland	0.591	0.954	1.228	0.487
Swansea	0.765	1.139	1.453	0.638
Tottenham	1.124	0.553	0.241	1.000
Watford	0.721	1.046	0.902	0.587
West Brom	0.943	0.729	0.693	0.587
West Ham	0.591	0.909	1.118	1.000

Tabela 3.2 Fatores de ataque e defesa das equipas do campeonato Inglês

Vamos analisar as odds obtidas pelos modelos de Poisson univariado e bivariado propostos por Maher. Para isso selecionamos quatro jogos da liga inglesa e quatro jogos da liga portuguesa.

Consideramos X="Numero de golos da equipa que joga em Casa", Y="Numero de golos da equipa que joga fora" Apresentamos de forma mais detalhada o calculo das odds de um jogo de futebol.

Neste caso tem-se

$$\begin{cases} \lambda_1 = \alpha_{Estoril}\beta_{SportingCP} = 1.178715 \cdot 0.8895757 = 1.048556 \\ \lambda_2 = \gamma_{Estoril}\delta_{SportingCP} = 1.18395 \cdot 2.396519 = 2.837358 \end{cases}$$

Assim, consideramos os modelos de Poisson Bivariado com margens independentes e com covariancia estimada igual a -0.118 no caso português e igual a -0.078 no caso inglês.

Obtemos no primeiro caso

$$P(X > Y) = 0.1117406 \qquad P(X = Y) = 0.1452042 \qquad P(X < Y) = 0.7423259$$

e no segundo caso

$$P(X > Y) = 0.1004775 \qquad P(X = Y) = 0.151161 \qquad P(X < Y) = 0.7474624$$

Obtendo assim as seguintes Odds no primeiro caso

$$\begin{cases} Odd_{Estoril} = \frac{1}{0.1117406} 8.948 \\ Odd_{Empate} = \frac{1}{0.1452042} = 6.887 \\ Odd_{SportingCP} = \frac{1}{0.7423259} = 1.347 \end{cases}$$

e no segundo caso

$$\begin{cases} Odd_{Estoril} = \frac{1}{0.1198502} = 8.344 \\ Odd_{Empate} = \frac{1}{0.1433293} = 6.977 \\ Odd_{SportingCP} = \frac{1}{0.7355054} = 1.360 \end{cases}$$

Procedemos da mesma forma para os outros jogos. Apresentamos os resultados na tabela seguinte. Incluímos também as correspondentes odds do mercado retiradas de https://www.jogossantacasa.pt/web/Placard a 25/02/2017

Jogo	λ_1	λ_2	P(X>Y)	P(X=Y)	P(X<Y)	Odds Esperadas	Odds do mercado
Estoril - Sporting CP	1.0485560	2.8373581	0.1117406	0.1452042	0.7423259	8.949 - 6.887 - 1.347	7.75 - 4.25 - 1.38
	1.1663019	2.9551040	0.1198502	0.1433293	0.7355054	8.344 - 6.977 - 1.360	
Rio Ave - Paços Ferreira	2.1705217	1.7719502	0.4743805	0.2050502	0.3204617	2.108 - 4.877 - 3.120	1.72 - 3.20 - 4.62
	2.2882676	1.8896961	0.4753459	0.1988313	0.3255694	2.104 - 5.029 - 3.072	
V.Setúbal - Braga	0.9569629	1.1030001	0.3112831	0.3019473	0.3867692	3.213 - 3.312 - 2.586	2.94 - 2.83 - 2.40
	1.0747088	1.2207460	0.3224065	0.2830507	0.3945407	3.102 - 3.533 - 2.535	
Boavista - Porto	0.52117956	2.13079176	0.07743273	0.17343480	0.74905452	12.914 - 5.766 - 1.335	8.02 - 4.00 - 1.36
	0.63892546	2.24853766	0.08998807	0.16953917	0.74029248	11.113 - 5.898 - 1.351	
Tottenham - Stoke	1.02155086	0.16593771	0.58136665	0.35891828	0.05971493	1.720 - 2.786 - 16.746	1.37 - 4.16 - 6.11
	1.09935798	0.24374483	0.58246362	0.33580971	0.08172618	1.717 - 2.978 - 12.236	
Everton - Sunderland	1.02927184	0.23805213	0.56012796	0.35492105	0.08495085	1.785- 2.818 - 11.772	1.42 - 4.93 - 8.05
	1.1070790	0.3158593	0.5618473	0.3329396	0.1052126	1.780 - 3.004 - 9.505	
Chelsea - Swansea	1.64912812	0.21639871	0.74057542	0.21519220	0.04422313	1.350 - 4.647 - 22.613	1.26 - 6.02 - 13.13
	1.72693524	0.29420583	0.73388702	0.20887569	0.05721747	1.363 - 4.788 - 17.477	
West Brom - Bournemouth	1.1621996	0.4773253	0.5305226	0.3176174	0.1518596	1.885 - 3.148 - 6.585	2.17 - 3.28 - 3.75
	1.2400067	0.5551324	0.5320515	0.3016940	0.1662532	1.880 - 3.315 - 6.015	

Bibliografia

[1] Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

[2] Karlis, D., Ntzoufras, I., et al. (2005). Bivariate poisson and diagonal inflated bivariate poisson regression models in r. *Journal of Statistical Software*, 14(10):1–36.

[3] Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P. A., and Peterson, J. P. (1999). Multivariate zero-inflated poisson models and their applications. *Technometrics*, 41(1):29–38.

[4] M.J.Maher (1982). Modelling association football scores. *Statistica Neerlandia*, 1(36):109–118.

[5] Mwembe, D., Sibanda, L., and Mupondo, N. C. (2015). Application of a bivariate poisson model in devising a profitable betting strategy of the Zimbabwe premier soccer league match results. *American Journal of Theoretical and Applied Statistics*, 4(3):99–111.

[6] N. Johnson, S. K. (1969). *Discrete Distributions*. John Wiley and Sons, New York.

Anexo A

A.1 Dados

	PortCasa	PortFora	InglCasa	InglFora
0	124	161	133	193
1	153	166	201	203
2	102	93	136	120
3	64	33	84	56
4	18	17	33	23
5	9	3	8	4
6	4	1	5	1
média	1.456	1.139	1.545	1.215

Tabela A.1 Dados univariados

	Fora						
Casa	0	1	2	3	4	5	6
0	44	46	16	20	6	0	1
1	60	68	43	19	9	2	0
2	42	42	38	9	4	1	0
3	31	28	15	7	3	0	0
4	14	11	6	1	0	1	0
5	2	5	0	0	1	0	0
6	0	3	2	0	0	0	0

Tabela A.2 Dados bivariados do campeonato Inglês

	Fora						
Casa	0	1	2	3	4	5	6
0	34	50	21	9	7	2	1
1	53	50	31	14	5	0	0
2	39	33	25	3	1	1	0
3	22	22	11	5	4	0	0
4	9	5	3	1	0	0	0
5	1	6	1	1	0	0	0
6	3	0	1	0	0	0	0

Tabela A.3 Dados bivariados do campeonato Português

A.2 Código R

A.2.1 Poisson

```
setwd("C:/Users/Ruben␣Lavrador/Dropbox
/Tese␣Mestrado␣Ruben␣Lavrador/Codigo␣final/dados")

results<- read.csv(file="E0.2015.2016.csv",head=TRUE,sep=";")
E0<- read.csv(file="E0.csv",head=TRUE,sep=",")

n<-nrow(results)
m<-nrow(E0)
golos_casa<-vector()
golos_fora<-vector()

for(i in 1:n){
golos_casa[i]<-results$FTHG[i]
golos_fora[i]<-results$FTAG[i]
golos_casa<- golos_casa[!is.na(golos_casa)]
golos_fora<- golos_fora[!is.na(golos_fora)]
}

for(i in 1:m){
golos_casa[i+n]<-E0$FTHG[i]
golos_fora[i+n]<-E0$FTAG[i]
golos_casa <- golos_casa[!is.na(golos_casa)]
golos_fora<- golos_fora[!is.na(golos_fora)]
}

lambda1<-mean(golos_casa)
lambda2<-mean(golos_fora)
```

```
casa<-table(golos_casa)
fora<-table(golos_fora)
alt<-c(0,1,2,3,4,5,6)
ctrl<-c(0,1,2,3,4,5,6)

p_casa<-dpois(alt,mean(golos_casa))
p_fora<-dpois(ctrl,mean(golos_fora))

qui_casa<-chisq.test(casa, p = p_casa,rescale.p=TRUE,simulate.p.value=TRUE)
qui_fora<-chisq.test(fora, p = p_fora,rescale.p=TRUE,simulate.p.value=TRUE)
```

A.2.2 ZIP

```
setwd("C:/Users/Ruben␣Lavrador/Dropbox
/Tese␣Mestrado␣Ruben␣Lavrador/Codigo␣final/dados")

results<- read.csv(file="E0.2015.2016.csv",head=TRUE,sep=";")
E0<- read.csv(file="E0.csv",head=TRUE,sep=",")

n<-nrow(results)
m<-nrow(E0)
golos_fora<-vector()

for(i in 1:n){
golos_fora[i]<-results$FTAG[i]
golos_fora<- golos_fora[!is.na(golos_fora)]
}

for(i in 1:m){
golos_fora[i+n]<-E0$FTAG[i]
golos_fora<- golos_fora[!is.na(golos_fora)]
}

lambda2<-mean(golos_fora)

c_fora<-table(golos_fora)
n_0_fora<-getElement(c_fora, "0")/length(golos_fora)
n_1_fora<-getElement(c_fora, "1")/length(golos_fora)
n_2_fora<-getElement(c_fora, "2")/length(golos_fora)
n_3_fora<-getElement(c_fora, "3")/length(golos_fora)
n_4_fora<-getElement(c_fora, "4")/length(golos_fora)
n_5_fora<-getElement(c_fora, "5")/length(golos_fora)
```

```
n_6_fora<-getElement(c_fora, "6")/length(golos_fora)

freq_fora<-c(n_0_fora,n_1_fora,n_2_fora,n_3_fora,n_4_fora,n_5_fora,n_6_fora)

f1<-function(lambda){mean(golos_fora)-exp(-lambda)*mean(golos_fora)
-lambda+lambda*n_0_fora}
lambda_new_fora<-uniroot(f1, c(0.5, 2), tol = 0.0001)$root
alpha<-1-mean(golos_fora)/lambda_new_fora
vec<-c(0,1,2,3,4,5,6)
pois_initial_fora<-dpois(vec,lambda=lambda2)
pois_new_fora<-(1-alpha)*dpois(vec,lambda=lambda_new_fora)
add_fora<-c(alpha,0,0,0,0,0,0)
final_fora<-pois_new_fora+add_fora
res_fora<-freq_fora-pois_initial_fora

qui_fora_initial<-chisq.test(freq_fora*length(golos_fora),
 p =pois_initial_fora,rescale.p=TRUE,simulate.p.value=TRUE)
qui_fora_final<-chisq.test(freq_fora*length(golos_fora),
 p =final_fora,rescale.p=TRUE,simulate.p.value=TRUE)
```

A.2.3 Deflacionado em dois

```
setwd("C:/Users/Ruben␣Lavrador/Dropbox
/Tese␣Mestrado␣Ruben␣Lavrador/Codigo␣final/dados")
library(rootSolve)
results<- read.csv(file="E0.2015.2016.csv",head=TRUE,sep=";")
E0<- read.csv(file="E0.csv",head=TRUE,sep=",")

n<-nrow(results)
m<-nrow(E0)
golos_casa<-vector()

for(i in 1:n){
golos_casa[i]<-results$FTHG[i]
golos_casa<- golos_casa[!is.na(golos_casa)]
}

for(i in 1:m){
golos_casa[i+n]<-E0$FTHG[i]
golos_casa <- golos_casa[!is.na(golos_casa)]
}

lambda1<-mean(golos_casa)
```

```
c_casa<-table(golos_casa)
n_0_casa<-getElement(c_casa, "0")/length(golos_casa)
n_1_casa<-getElement(c_casa, "1")/length(golos_casa)
n_2_casa<-getElement(c_casa, "2")/length(golos_casa)
n_3_casa<-getElement(c_casa, "3")/length(golos_casa)
n_4_casa<-getElement(c_casa, "4")/length(golos_casa)
n_5_casa<-getElement(c_casa, "5")/length(golos_casa)
n_6_casa<-getElement(c_casa, "6")/length(golos_casa)

freq_casa<-c(n_0_casa,n_1_casa,n_2_casa,n_3_casa,n_4_casa,n_5_casa,n_6_casa)

f1<-function(lambda){exp(-lambda)*lambda^2/2
-lambda*(n_2_casa-1)/(mean(golos_casa)-2)-
(mean(golos_casa)-2*n_2_casa)/(mean(golos_casa)-2)}

lambda_new_casa<-uniroot(f1, c(0.5, 2), tol = 0.0001)$root
beta_new_casa<-(mean(golos_casa)-lambda_new_casa)/(lambda_new_casa-2)
vec<-c(0,1,2,3,4,5,6)
pois_initial_casa<-dpois(vec,lambda=lambda1)
pois_new_casa<-(1+beta_new_casa)*dpois(vec,lambda=lambda_new_casa)
add_casa<-c(0,0,-beta_new_casa,0,0,0,0)
final_casa<-pois_new_casa+add_casa

qui_casa_initial<-chisq.test(freq_casa*length(golos_casa),
 p =pois_initial_casa,rescale.p=TRUE,simulate.p.value=TRUE)
qui_casa_final<-chisq.test(freq_casa*length(golos_casa),
 p =final_casa,rescale.p=TRUE,simulate.p.value=TRUE)
```

A.2.4 Inflacionado em zero e deflacionado em dois

```
setwd("C:/Users/Ruben␣Lavrador/Dropbox
/Tese␣Mestrado␣Ruben␣Lavrador/Codigo␣final/dados")

library(rootSolve)
results<- read.csv(file="E0.2015.2016.csv",head=TRUE,sep=";")
E0<- read.csv(file="E0.csv",head=TRUE,sep=",")
n<-nrow(results)
m<-nrow(E0)

golos_casa<-vector()

for(i in 1:n){
```

```
golos_casa[i]<-results$FTHG[i]
golos_casa<- golos_casa[!is.na(golos_casa)]
}

for(i in 1:m){
golos_casa[i+n]<-E0$FTHG[i]
golos_casa <- golos_casa[!is.na(golos_casa)]
}

lambda1<-mean(golos_casa)

c_casa<-table(golos_casa)
n_0_casa<-getElement(c_casa, "0")/length(golos_casa)
n_1_casa<-getElement(c_casa, "1")/length(golos_casa)
n_2_casa<-getElement(c_casa, "2")/length(golos_casa)
n_3_casa<-getElement(c_casa, "3")/length(golos_casa)
n_4_casa<-getElement(c_casa, "4")/length(golos_casa)
n_5_casa<-getElement(c_casa, "5")/length(golos_casa)
n_6_casa<-getElement(c_casa, "6")/length(golos_casa)

freq_casa<-c(n_0_casa,n_1_casa,n_2_casa,n_3_casa,n_4_casa,n_5_casa,n_6_casa)

alt<-(n_0_casa+2*n_2_casa-mean(golos_casa))/n_1_casa
f1<-function(lambda){1/lambda+lambda-exp(lambda)-alt}
lambda_new_casa<-uniroot(f1, c(0.5, 2), tol = 0.0001)$root
alpha<-(1-n_0_casa-n_2_casa-n_0_casa*exp(lambda_new_casa)
+n_0_casa+lambda_new_casa^2/2*n_0_casa)/(1-exp(lambda_new_casa)
+lambda_new_casa^2/2)
beta<-(mean(golos_casa)-alpha-lambda_new_casa+lambda_new_casa*alpha)
/(lambda_new_casa-2)
vec<-c(0,1,2,3,4,5,6)
pois_initial_casa<-dpois(vec,lambda=lambda1)
pois_new_casa<-(1-alpha+beta)*dpois(vec,lambda=lambda_new_casa)
add_casa<-c(alpha,0,-beta,0,0,0,0)
final_casa<-pois_new_casa+add_casa

qui_casa_initial<-chisq.test(freq_casa*length(golos_casa),
 p =pois_initial_casa,rescale.p=TRUE,simulate.p.value=TRUE)
qui_casa_final<-chisq.test(freq_casa*length(golos_casa),
 p =final_casa,rescale.p=TRUE,simulate.p.value=TRUE)
```

A.2.5 Odds

```
setwd("C:/Users/Ruben␣Lavrador/Dropbox/
Tese␣Mestrado␣Ruben␣Lavrador/Codigo␣final/Odds")

## EQUIPAS##

HomeTeam<-"Tottenham"
AwayTeam<-"Stoke"

equipas<-c("Burnley","Bournemouth","Chelsea","Swansea","Everton","Watford",
"Leicester","Sunderland","Man␣United","Tottenham","Hull","Crystal␣Palace",
"Arsenal","West␣Ham","Middlesbrough","Southampton","Stoke","Liverpool",
"West␣Brom","Man␣City")

results<- read.csv(file="E0.2015.2016.csv",head=TRUE,sep=",")
E0<- read.csv(file="E0.csv",head=TRUE,sep=",")
E1<- read.csv(file="E1.2015.2016.csv",head=TRUE,sep=",")

n<-nrow(results)
m<-nrow(E0)
e<-length(equipas)
p<-nrow(E1)

homeattack_initial<-vector()
awayattack_initial<-vector()
homedefense_initial<-vector()
awaydefense_initial<-vector()

alpha<-vector()
beta<-vector()
sigma<-vector()
delta<-vector()

homegoals<-vector()
awaygoals<-vector()
homeconceded<-vector()
awayconceded<-vector()

f_zero<-vector()
for(j in 1:e){
Team<-equipas[j]
GMcasa<-vector()
GMfora<-vector()
```

```
GScasa<-vector()
GSfora<-vector()

for(i in 1:nrow(results)){
if(results$HomeTeam[i]==Team){
GMcasa[i]<-results$FTHG[i]
}
GMcasa <- GMcasa[!is.na(GMcasa)]
}

for(i in 1:nrow(E1)){
if(E1$HomeTeam[i]==Team){
GMcasa[i+n]<-E1$FTHG[i]
}
GMcasa <- GMcasa[!is.na(GMcasa)]
}

for(i in 1:nrow(E0)){
if(E0$HomeTeam[i]==Team){
GMcasa[i+p+n]<-E0$FTHG[i]
}
GMcasa <- GMcasa[!is.na(GMcasa)]
}

homegoals[j]<-sum(GMcasa)

for(i in 1:nrow(results)){
if(results$AwayTeam[i]==Team){
GMfora[i]<-results$FTAG[i]
}
GMfora <- GMfora[!is.na(GMfora)]
}

for(i in 1:nrow(E1)){
if(E1$AwayTeam[i]==Team){
GMfora[i+n]<-E1$FTAG[i]
}
GMfora <- GMfora[!is.na(GMfora)]
}

for(i in 1:nrow(E0)){
if(E0$AwayTeam[i]==Team){
```

```
GMfora[i+p+n]<-E0$FTAG[i]
}
GMfora <- GMfora[!is.na(GMfora)]
}

awaygoals[j]<-sum(GMfora)

for(i in 1:nrow(results)){
if(results$HomeTeam[i]==Team){
GScasa[i]<-results$FTAG[i]
}
GScasa <- GScasa[!is.na(GScasa)]
}

for(i in 1:nrow(E1)){
if(E1$HomeTeam[i]==Team){
GScasa[i+n]<-E1$FTAG[i]
}
GScasa <- GScasa[!is.na(GScasa)]
}

for(i in 1:nrow(E0)){
if(E0$HomeTeam[i]==Team){
GScasa[i+p+n]<-E0$FTAG[i]
}
GScasa <- GScasa[!is.na(GScasa)]
}

homeconceded[j]<-sum(GScasa)

for(i in 1:nrow(results)){
if(results$AwayTeam[i]==Team){
GSfora[i]<-results$FTHG[i]
}
GSfora <- GSfora[!is.na(GSfora)]
}

for(i in 1:nrow(E1)){
if(E1$AwayTeam[i]==Team){
GSfora[i+n]<-E1$FTHG[i]
}
GSfora <- GSfora[!is.na(GSfora)]
```

```
}

for(i in 1:nrow(E0)){
if(E0$AwayTeam[i]==Team){
GSfora[i+p+n]<-E0$FTHG[i]
}

GSfora <- GSfora[!is.na(GSfora)]
}
awayconceded[j]<-sum(GSfora)
}
################################

for( j in 1:e){
homeattack_initial[j]<-homegoals[j]/sqrt(sum(homegoals))
awayattack_initial[j]<-awaygoals[j]/sqrt(sum(awaygoals))
homedefense_initial[j]<-homeconceded[j]/sqrt(sum(awaygoals))
awaydefense_initial[j]<-awayconceded[j]/sqrt(sum(homegoals))
}

for( i in 1:20){
for( j in 1:length(equipas)){
alpha[j]<-homegoals[j]/(sum(homeattack_initial)-homeattack_initial[j])
beta[j]<-awayconceded[j]/(sum(awaydefense_initial)-awaydefense_initial[j])
sigma[j]<-homeconceded[j]/(sum(homedefense_initial)-homedefense_initial[j])
delta[j]<-awaygoals[j]/(sum(awayattack_initial)-awayattack_initial[j])
}
homeattack_initial<-alpha
awaydefense_initial<-beta
homedefense_initial<-sigma
awayattack_initial<-delta
}

## FIM DA LEITURA DOS DADOS

## Calculo das odds

for (w in 1:length(equipas)){
if(equipas[w]==HomeTeam){
i<-w}
}
```

```
for (w in 1:length(equipas)){
if(equipas[w]==AwayTeam){
j<-w}
}

lambda_hometeam<- alpha[i]*beta[j]
lambda_awayteam<- sigma[i]*delta[j]

x_draw<-c(0,1,2,3,4,5,6,7,8,9,10)
y_draw<-c(0,1,2,3,4,5,6,7,8,9,10)

x_win<-c(1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,
7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9,9,9,9,9,9)
y_loose<-c(0,0,1,0,1,2,0,1,2,3,0,1,2,3,4,0,1,2,3,4,
5,0,1,2,3,4,5,6,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,8)

x_loose<-c(0,0,1,0,1,2,0,1,2,3,0,1,2,3,4,0,1,2,3,4,5,
0,1,2,3,4,5,6,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,8)
y_win<-c(1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,
7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9,9,9,9,9,9)

## Poisson ##

draw_pois<-dpois(x_draw,lambda=lambda_hometeam)
*dpois(y_draw,lambda=lambda_awayteam)

home_win_pois<-dpois(x_win,lambda=lambda_hometeam)
*dpois(y_loose,lambda=lambda_awayteam)

away_win_pois<-dpois(x_loose,lambda=lambda_hometeam)
*dpois(y_win,lambda=lambda_awayteam)

odd_empate_pois<-1/sum(draw_pois)
odd_home_pois<-1/sum(home_win_pois)
odd_away_pois<-1/sum(away_win_pois)

odds_pois<-c(odd_home_pois,odd_empate_pois,odd_away_pois)

###### Poisson Bivariado ######

"pbivpois" <-
function(x, y=NULL, lambda = c(1, 1, 1), log=FALSE) {
```

```
if ( is.matrix(x) ) {
                var1<-x[,1]
                var2<-x[,2]
        }
        else if (is.vector(x)&is.vector(y)){
                if (length(x)==length(y)){
                        var1<-x
                        var2<-y
                }
                else{
                        stop('lengths␣of␣x␣and␣y␣are␣not␣equal')
                }
        }
        else{
                stop('x␣is␣not␣a␣matrix␣or␣x␣and␣y␣are␣not␣vectors')
        }
        n <- length(var1)
        logbp<-vector(length=n)
#
        for (k in 1:n){
                x0<-var1[k]
                y0<-var2[k]
                xymin<-min( x0,y0 )
                lambdaratio<-lambda[3]/(lambda[1]*lambda[2])
#
                i<-0:xymin
                sums<- -lgamma(var1[k]-i+1)-lgamma(i+1)
          -lgamma(var2[k]-i+1)
             +i*log(lambdaratio)
                maxsums <- max(sums)
                sums<- sums - maxsums
                logsummation<- log( sum(exp(sums)) ) + maxsums
                logbp[k]<- -sum(lambda) + var1[k] * log( lambda[1] )
              + var2[k] * log( lambda[2] ) + logsummation
        }
        if (log) { result<-    logbp }
        else     { result<-exp(logbp)  }
        result
#       end of function bivpois
}
ro<-0.2*sqrt(alpha[i]*beta[j]*sigma[i]*delta[j])
```

```
draw_bivpois<-pbivpois(x_draw,y_draw,lambda
= c(lambda_hometeam, lambda_awayteam, ro), log=FALSE)
home_win_bivpois<-pbivpois(x_win,y_loose,lambda
= c(lambda_hometeam, lambda_awayteam, ro), log=FALSE)
away_win_bivpois<-pbivpois(x_loose,y_win,lambda
= c(lambda_hometeam, lambda_awayteam, ro), log=FALSE)

odd_empate_bivpois<-1/sum(draw_bivpois)
odd_home_bivpois<-1/sum(home_win_bivpois)
odd_away_bivpois<-1/sum(away_win_bivpois)

odds_bivpois<-c(odd_home_bivpois,odd_empate_bivpois,odd_away_bivpois)
```