Trabajo Fin de Máster

Aplicación de Modelos de Regresión de Poisson Bivariados a los resultados de los partidos de la Liga Española de Fútbol

Autora: Eva María García Quinteiro

Directora: María del Carmen Iglesias Pérez

Máster Técnicas Estadísticas
Universidad de Vigo

Fecha de presentación
15.07.2014

El presente documento que tiene como título "Aplicación de Modelos de Regresión de Poisson Bivariados a los resultados de los partidos de la Liga Española de Fútbol" recoge el trabajo realizado por Eva M. García Quinteiro como Proyecto Fin de Máster de Técnicas Estadísticas bajo la dirección de María del Carmen Iglesias Pérez.

Fdo.: María del Carmen Iglesias Pérez

Fdo.: Eva M. García Quinteiro

Índice general

Índice de figuras

Resumen

La Estadística es una ciencia aplicada de utilidad en múltiples disciplinas. Este trabajo presenta una aplicación al Deporte, con el objetivo de modelizar los resultados de los partidos de fútbol de la liga española (nº de goles obtenidos por cada equipo en un partido) mediante el uso de modelos de regresión de Poisson bivariantes. El trabajo se estructura en varios capítulos. En el primero se realiza una pequeña introducción definiendo el punto de partida y los objetivos de este trabajo. En el segundo se revisan artículos recientes sobre modelización estadística en datos de fútbol y se presentan los datos objeto de análisis pertenecientes a las temporadas 2010/2011 y 2011/2012 de la Liga Española de Fútbol. El tercer capítulo presenta el modelo de regresión de Poisson bivariante que vamos a utilizar y la estimación de sus parámetros mediante el algoritmo EM. También incluye extensiones del modelo para permitir un inflado en la diagonal que se ajuste mejor a los datos. En el cuarto se realiza una aproximación de la regresión bivariante de Poisson aplicada al fútbol y se muestran los resultados del análisis de las temporadas 2010/2011 y 2011/2012 utilizando el paquete bivpois de R. Finalmente, una conclusión final incluye una discusión de los resultados obtenidos y propone futuras líneas de trabajo teniendo en cuenta diferentes posibilidades de aplicación y relación con otras disciplinas.

1. Introducción

El punto de partida para la realización de este trabajo fin de máster fue ejecutar una búsqueda de aplicaciones estadísticas en deporte en el mes de Septiembre del 2012. Partiendo de esta idea, se hizo una búsqueda informal por Internet que dio lugar a la obtención de varios artículos relacionados con la aplicación de modelos estadísticos en el mundo del deporte y, en concreto, al fútbol.

A partir de esa búsqueda informal se acotó el objetivo de este estudio al modelo de regresión de Poisson bivariante y se procedió a una revisión más exhaustiva de la bibliografía utilizando la base de datos SCOPUS. Como resultado de todas las búsquedas efectuadas se obtuvieron un total de 40 referencias que son analizadas y discutidas en el capítulo 2 de este proyecto. Además, en dicho capítulo se presentan los datos de los partidos de fútbol que se analizan en este trabajo utilizando modelos de regresión de Poisson bivariantes. Los datos corresponden a los resultados de los partidos de la Liga Española de Fútbol de las temporadas 2010/2011 y 2011/2012 de Primera División y se obtuvieron consultando la página web de la LPF.

En el capítulo 3 se introduce el modelo de regresión de Poisson, y se estudian con más detalle las características principales del modelo de regresión bivariante de Poisson y de la regresión bivariante con inflado en la diagonal, incluyendo la estimación de sus parámetros mediante el algoritmo EM.

Posteriormente, en el capítulo 4 se presenta el análisis de los datos procedentes de las dos temporadas ya mencionadas: Temporada 2010/2011 y 2011/2012, mostrándose los resultados obtenidos. Para contextualizar dicho análisis, el capítulo comienza con un apartado previo que presenta una modelización de la regresión bivariante de Poisson aplicada al contexto futbolístico, explica la función bivpois de R utilizada en el análisis de los datos y revisa los esquemas de codificación de las variables cualitativas para explicar el enfoque utilizado en la interpretación de las variables nominales consideradas en este estudio.

Finalmente, un apartado de conclusiones resalta los hallazgos más importantes y posibles líneas de investigación sobre la aplicación de estos modelos de regresión bivariados en el análisis de los resultados deportivos.

2. Revisión de la literatura y datos

2.1. Revisión de la literatura

2.1.1. Introducción

Tal y como se comentó en el capítulo 1, el principal objetivo de este trabajo es modelizar el número de goles de los partidos de la liga española de fútbol mediante el uso de modelos de regresión de Poisson bivariantes.

El punto de partida para delimitar este tema de estudio fue realizar una búsqueda informal por Internet en el mes de Septiembre de 2012, que dio lugar a la obtención de varios artículos relacionados con la aplicación de modelos estadísticos al mundo del deporte y, en concreto, al fútbol. Como consecuencia de esa primera búsqueda informal se seleccionaron una serie de artículos que nos mostraron la diversidad de aproximaciones a la modelización en fútbol, y que incluiremos en un apartado de discusión, agrupada en torno a tres cuestiones que consideramos relevantes estadísticamente:

1. La técnica estadística utilizada para modelizar los resultados: distintos modelos de regresión, modelos mixtos, aproximaciones bayesianas,...

2. La elección de la variable dependiente a modelizar y/o la unidad muestral considerada: nº goles, resultado (ganar, perder o empatar), diferencia de goles o puntos,...de los partidos; nº total de goles, nº de puntos o clasificación de los equipos en competición, etc.

3. Variables explicativas a tener en cuenta: potencia goleadora y defensiva de cada equipo, jugar en casa, resultados en partidos anteriores, diferencia competitiva entre los equipos que se enfrentan, eventos anteriores en el propio partido,...

A partir de esa búsqueda informal acotamos nuestro objeto de estudio al modelo de regresión de Poisson bivariante y se procedió a ejecutar una búsqueda más específica en la base de datos SCOPUS.

Los criterios de inclusión fueron la fecha de publicación, desde enero de 2000 y la disponibilidad del resumen del artículo.

Se realizó una primera búsqueda con palabras clave “Sport Bivariate Poisson Models” que proporcionó dos artículos: Karlis y Tsiamyrtzis (2008) y Karlis y Ntzoufras (2003).

Seguidamente, se realizó una segunda búsqueda con las siguientes palabras clave: "Football Scores Poisson Model" y los mismos criterios de inclusión anteriormente mencionados, que proporcionó dieciséis artículos. De ellos se eliminaron ocho por no considerarse relacionados con el objetivo de este trabajo. Además, el trabajo de Karlis y Ntzoufras (2003) apareció repetido.

Los nueve artículos resultantes de la búsqueda en SCOPUS se encuentran detallados en el último apartado de esta sección.

Esos nueve artículos, los previamente seleccionados y las referencias más relevantes citadas en ellos suman un total de 40 obras de referencia que se han revisado en este trabajo. Una discusión de las mismas se incluye en el siguiente apartado.

2.1.2. Discusión

Karlis y Ntzoufras (2003) publican el artículo más relevante en relación al objetivo de este trabajo. Estos autores modelizan *el número de goles marcados por cada equipo* mediante una distribución de Poisson bivariada aplicada a los resultados de partidos de fútbol. La distribución de Poisson bivariada es más general que una Poisson doble (dos Poisson independientes) y permite un mejor ajuste a los datos observados. La estimación de máxima verosimilitud para los parámetros se realiza a través del algoritmo EM. Por otra parte, proponen modelos de inflado en la diagonal que permiten realizar de forma más precisa el ajuste de los empates. Además, estos autores tienen en cuenta como variables independientes si el equipo juega en casa o fuera y el rendimiento de defensa y ataque de cada equipo. Karlis y Ntzoufras citan los trabajos de Maher (1982) y Dixon y Coles (1997) como antecedentes de su modelo. Los aspectos más relevantes de estos dos artículos se explican a continuación.

Maher (1982) en su artículo "Modelling association football scores" propone dos variables de Poisson independientes para la modelización del nº de goles de cada equipo. Explora distintas variables explicativas relacionadas con la fuerza en defensa y ataque de los equipos cuando juegan en casa o fuera. Compara las frecuencias observadas y las esperadas de su modelo mediante pruebas de bondad de ajuste mostrando que, aunque existen algunas pequeñas diferencias sistemáticas, un modelo de Poisson independiente proporciona una descripción precisa de las puntuaciones futbolísticas. Finalmente, evita la condición de independencia proponiendo una Poisson bivariante.

Un aspecto interesante de este artículo es que argumenta a favor del uso de la distribución de Poisson, en lugar del modelo de la Binomial Negativa propuesto hasta ese momento, haciendo que la media de la Poisson varíe en función de variables explicativas. Para ello, expone las siguientes razones (pág. 109):

- La posesión es un aspecto a tener en cuenta en el fútbol puesto que, cuando un equipo tiene el balón, tiene la oportunidad de atacar y marcar.

- La probabilidad p de que un ataque termine en gol, es pequeña, pero el número de veces que un equipo está en posesión del esférico durante un partido es muy grande. Si p es constante y los ataques son independientes, el número de goles sigue una Binomial y en estas condiciones la aproximación que mejor se ajusta es la Poisson.
- La media de esta Poisson variará en consonancia con la calidad del equipo y si se considera la distribución de todos los goles marcados por todos los equipos, se debería considerar la distribución Poisson con media variable.

Dixon y Coles (1997) parten del modelo de Maher (1982) pero plantean varias modificaciones con el fin de mejorar el ajuste del modelo en partidos con bajo nº de goles y también para permitir que los parámetros de habilidad en ataque y defensa de un equipo sean dinámicos y basados en su rendimiento reciente. Para la estimación de los parámetros plantean una función de "pseudoverosimititud" que maximizan por métodos numéricos.

Recientemente, Karlis y Ntzoufras (2011) proponen una estimación robusta de un modelo similar al suyo del 2003, asumiendo dos variables de Poisson independientes y sustituyendo la función de verosimilitud por una verosimilitud ponderada, cuyos pesos producen estimaciones más robustas. Asimismo, Koopman y Lit (2012) firman un *Discussion Paper* que propone un modelo de regresión de Poisson bivariante dinámico. Este modelo generaliza el modelo de Karlis y Ntzoufras (2003) para poder introducir variables que cambian con el paso del tiempo.

En la línea de considerar el efecto temporal en la modelización del nº de goles, están también dos artículos recientes. Malcata, Hopkins y Richardson (2012) utilizan un modelo lineal mixto generalizado (Poisson) de los goles marcados durante varias temporadas por un pequeño grupo de equipos, teniendo en cuenta su rendimiento anual, su calidad, la edad de los jugadores y la ventaja de jugar en casa. Volf (2009) realiza una modelización de la secuencia de goles marcados en un partido utilizando procesos puntuales.

Volviendo a la modelización del nº de goles de los equipos en un partido mediante variables de Poisson independientes con media variable, Dyte y Clarke (2000), utilizan como variables explicativas la puntuación de la FIFA para cada equipo y el lugar del partido, en el contexto de partidos de fútbol internacionales.

Greenhough et al. (2002) analizan las distribuciones del número de goles marcados por los equipos que juegan en casa, fuera de casa y el total de goles marcados en cada partido sin considerar covariables. En este contexto, muestran que las distribuciones de valores extremos pueden mejorar el ajuste de la Poisson o la distribución binomial negativa cuando resulten inadecuadas. En relación a la elección más conveniente del tipo de distribución (Poisson, Binomial negativa, valor extremo) para la modelización del nº de goles, Bittner, Nussbaumer, Janke y Weigel (2007, 2009) sugieren modificar el proceso de Bernouilli (marcar gol en cada instante de partido) para incluir una componente que ellos denominan de *autoafirmación* (marcar un gol produce una motivación en el equipo que lo marca y una desmotivación en el

contrario), la cual permite comprender el motivo de la modelización con uno u otro tipo de distribución y consigue un buen ajuste a los datos.

Otro grupo de artículos utilizan una metodología bayesiana. Rue y Salvensen (1998) aplicaron un modelo lineal generalizado bayesiano dinámico para predecir el *número de goles* en cada partido. El modelo depende de las propiedades del equipo que juega en casa y del equipo que juega fuera. Específicamente, se tiene en cuenta la capacidad de ataque y defensa de cada uno de ellos. Además, incluyeron el efecto de que el equipo que juega en casa tiende a infraestimar la capacidad del equipo visitante si el equipo de casa es superior a éste. También tuvieron en cuenta que las variables de ataque y defensa varían a lo largo del tiempo. Skinner y Freeman (2009) discuten varios modelos en los cuales el número de goles marcados por un equipo en un partido de fútbol sigue una distribución de Poisson aunque se produce un mejor ajuste cuando se utiliza una distribución binomial negativa. Baio y Blangiardo (2010) utilizan un modelo jerárquico bayesiano que parte de una doble Poisson cuya media depende de parámetros que son a su vez variables aleatorias. Finalmente, Karlis y Ntzoufras (2009) utilizan también una aproximación bayesiana, pero se centran en modelizar *la diferencia del número de goles*, es decir, el margen de la victoria. Los autores citan como ventajas de este modelo que se elimina la correlación impuesta por el hecho de que dos equipos oponentes compiten uno contra otro y que no es necesario imponer que los goles marcados por cada equipo sean marginalmente distribuidos siguiendo una distribución Poisson.

En relación a este enfoque reciente de *modelizar la diferencia de goles entre los equipos* están los trabajos de Heuer y Rubner (2009) y de Heuer, Müller y Rubner (2010). Para estos autores la diferencia de goles es un indicador mejor que el número de puntos para establecer el mejor equipo en un campeonato. En sus trabajos analizan la evolución en el tiempo de la diferencia de goles utilizando teoría de paseos aleatorios, modelos de uso común en campos como la física (de la cual proceden estos autores).

Respecto a la modelización del *resultado del partido definido como ganar, perder o empatar,* Dobson y Goddard (2000), en su artículo sobre modelización estocástica de los resultados de partidos de fútbol, aplicaron un modelo probit ordenado con métodos de Monte Carlo. Goddard y Asimakopoulos (2004) aplicaron el mismo modelo teniendo en cuenta si los partidos se ganaban en casa, si se trataba de un empate o bien, si esa victoria era en un campo visitante. Brillinger (2006) evaluó las probabilidades de ganar, empatar o perder un partido. Estas probabilidades fueron modelizadas directamente teniendo en cuenta una variable latente que estaba representada por la diferencia de las cualidades de los dos equipos en el juego.

Por otra parte y, teniendo en cuenta la *puntuación de los equipos*, Harville (1997) empleó modelos lineales mixtos para modelizar la diferencia de puntos en un partido. Además, indicar que ya en ese trabajo se señalaba la ventaja de jugar en casa como una variable a tener en cuenta. Naim, Redner y Vazquez (2007) utilizan procesos estocásticos para estudiar diferentes resultados de los equipos, además, estudian la

distribución o la posición de cada equipo dentro de cada liga o torneo. Ben-Naim y Hengartner (2007) centrándose también en el puesto de cada equipo determinan que el rango está en función del número de equipos y del número de juegos, llegando a la conclusión que el formato de liga es un método no efectivo para determinar el mejor equipo.

Varios son los artículos que estudian la *ventaja de jugar en casa*. Jamieson (2010) publicó un meta-análisis sobre la ventaja de jugar en casa. Saavedra et al. (2012) analizan la ventaja de jugar en casa en la liga española de fútbol desde el año 1928 hasta 2011 determinando que en su estudio la ventaja de jugar en casa existe y es significativa. Lago-Peñas y Lago-Ballesteros (2011) concluyen en su estudio que los equipos locales marcan más goles y también disparan más a puerta que los equipos visitantes. Waters y Lovell (2002) realizaron un análisis de la ventaja de jugar en casa desde un punto de vista psicológico.

Panaretos (2002) estudia otras *variables explicativas relacionadas con el juego* (tiros a puerta, posesión de balón,...) que pueden influir en los goles y en los puntos que se obtienen en la Liga de Campeones.

Para concluir con esta discusión de la modelización estadística en datos de fútbol conviene citar dos referencias que hacen alusión a una *revisión de la estadística en el análisis de los resultados futbolísticos*: Emonet (2000) y Brillinger (2009).

Emonet (2000) realiza una revisión del concepto fútbol y su relación con la estadística. Cita que la probabilidad de ganar un partido sigue una distribución de Poisson o una distribución binomial negativa. También revisa los efectos de la ventaja de jugar en casa, la diferencia de jugar en campo artificial, la consecuencia de recibir tarjetas rojas, así como la influencia de las estrategias de juego.

Brillinger (2009), en un informe técnico, también refleja cuales han sido los trabajos mas importantes en el análisis de datos procedentes del fútbol. Menciona la utilización de diferentes modelos estocásticos, que incluyen distintas distribuciones específicas como la de Poisson bivariada, la exponencial, el valor extremo, la logística, binomial negativa u ordinal. Alguno de esos modelos tienen en cuenta los goles y otros los puntos de cada equipo.

Finalmente, indicar que se han revisado también algunos artículos que sin estar relacionados con el fútbol emplean modelos de regresión de Poisson bivariantes o inflados en otras disciplinas. Teniendo en cuenta una aproximación de regresión de Poisson bivariada, Jung y Winkelmann (1993) aplicaron esta aproximación en movilidad laboral. Vernic (1997) utilizó el modelo de Poisson bivariado empleando datos procedentes de seguros. Relacionado con seguros agrícolas destaca el trabajo de Wulu y Lovell (2002) que aplicaron modelos de regresión de Poisson, regresión de Poisson generalizada y binomial negativa con datos procedentes de ese campo.

Baetschamann y Rainer (2012) utilizaron los modelos de cero inflado al estudio de bajas por enfermedad o Famoye y Singh (2006) aplicaron el mismo modelo a datos obtenidos de la violencia doméstica. En el campo de los seguros se pueden señalar los

trabajos de Vernic (1997) o el de Walhin (2001) aplicando estos modelos al campo de los seguros y la biología. Otro trabajo reciente sobre modelos de regresión con cero inflado en medicina es el de Sumathi y Rao (2009) que realizaron una estimación de estos modelos. Por último, señalar el trabajo de Zhu (2009), quién aplicó modelos de cero inflado en datos procedentes de la Ecología.

2.1.3. Conclusiones

El trabajo de Karlis y Ntzoufras (2003) es uno de los artículos de referencia en la aplicación de modelos avanzados de regresión al mundo del fútbol y su enfoque inspira el presente trabajo. En concreto, estos autores basándose en los trabajos de Maher (1982) y Dixon y Coles (1997) propusieron una distribución de Poisson bivariada aplicada al número de goles marcados por cada equipo. También proponen modelos de inflado en la diagonal para modelizar mejor los resultados de empate. Además, el modelo propuesto incluye variables explicativas, como el hecho de jugar en casa o no y la capacidad defensiva y ofensiva de cada uno de los equipos.

Al igual que Karlis y Ntzoufras (2003), un gran número de artículos se centraron en la modelización de los resultados de fútbol tomando como variable dependiente el número de goles marcados por cada equipo (por ejemplo, Rue y Salvensen, 1998; Skinner y Freeman, 2009; Volf, 2009 o Baio y Blangiardo, 2010). Otros artículos, sin embargo, consideran modelos con variable respuesta la diferencia del número de goles (véase Karlis y Ntzoufras, 2009, Heuer y Rubner, 2009), los resultados de los equipos teniendo en cuenta si el resultado del mismo era ganar, perder o empatar (por ejemplo, Dobson y Goddard, 2000; Goddard y Asimkopoulos, 2004 y Brillinger, 2006) o la puntuación de los equipos (Harville, 1997, Naim et al., 2007 y Ben-Naim y Hengartner, 2007). La mayor parte de estos trabajos aplican modelos en los que el número de goles siguen una distribución de Poisson o una distribución binomial negativa. También se aplican modelos mixtos o distintos tipos de procesos estocásticos que permiten introducir efectos temporales. Asimismo, un grupo significativo de artículos utiliza un enfoque bayesiano.

Finalmente, en la revisión se incluyen algunos artículos que analizan la ventaja de jugar en casa, ventaja que se analiza en la mayoría de trabajos como una variable explicativa que estaría influyendo en la modelización de los resultados obtenidos en los partidos de fútbol, y se mencionan otros trabajos en los que se aplicaron modelos de regresión de Poisson bivariantes o inflados en el cero en otras disciplinas.

2.1.4. Artículos seleccionados de la búsqueda en SCOPUS

Palabras clave: **"Sport Bivariate Poisson Models"**

- **Karlis, D. & Tsiamyrtzis, P. (2008). Exact Bayesian modeling for bivariate Poisson data and extensions. *Statistics and Computing*, 18 (1), pp. 27-40.**

Los autores presentan una estimación bayesiana de los parámetros de un modelo de Poisson bivariante (sin covariables) previamente considerado en Karlis y Ntzoufras (2003). Demuestran que las distribuciones a posteriori son mezclas de distribuciones gamma. Además, definen una clase de distribuciones a priori que presentan buenas propiedades en este contexto.

- **Karlis, D. & Ntzoufras, L. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society Series D: The Statistician*, 52 (3), pp. 381-393.**

  Karlis y Ntzoufras (2003) modelizan el número de goles marcados por cada equipo mediante una distribución de Poisson bivariada aplicada a los resultados de partidos de fútbol. La distribución de Poisson bivariada es más general que una Poisson doble (dos Poisson independientes) y permite un mejor ajuste a los datos observados. La estimación de máxima verosimilitud para los parámetros se realiza a través del algoritmo EM. Por otra parte, proponen modelos de inflado en la diagonal que permiten realizar de forma más precisa el ajuste de los empates. Además, estos autores tienen en cuenta como variables independientes si el equipo juega en casa o fuera y el rendimiento de defensa y ataque de cada equipo. Los autores ilustran su modelo con resultados de la liga italiana de la temporada 91/92.

Palabras clave: **"Football Scores Poisson Model"**

- **Malcata, R.M., Hopkins, W.G. & Richardson, S. (2012). Modelling the progression of competitive performance of an academy's soccer teams. *Journal of Sports Science and Medicine, 11* (3), pp. 533-536.**

  La progresión del rendimiento de un equipo es la clave en el deporte competitivo, pero parece que no hay demasiadas publicaciones sobre la progresión de un equipo teniendo en cuenta períodos mas largos que una temporada. En este artículo se informa de la progresión de la puntuación de goles de tres equipos de una academia de desarrollo de jóvenes promesas durante cinco temporadas usando una nueva aproximación analítica basada en un modelo mixto generalizado. Las puntuaciones del juego se predijeron a través de un modelo mixto que asumía tener una distribución de Poisson. Todos los efectos fueron estimados como factores con una transformación log y presentados como diferencias de porcentaje en las puntuaciones. Los autores concluyen que el modelo generalizado mixto tiene mejor ajuste cuando el número de goles es mayor.

- **Karlis, D. & Ntzoufras, I. (2011). Robust fitting of football prediction models. *IMA Journal Management Mathematics, 22* (2), pp. 171-182.**

  El artículo de Karlis y Ntzoufras (2003) postula la existencia de métodos para la predicción de puntuaciones finales en el fútbol basado en modelos que

tienen en cuenta el número de goles marcados por dos equipos con estimación de parámetros por máxima verosimilitud. Aunque esta aproximación permite una predicción suficientemente exacta para el resultado final, no tiene en cuenta un gran número de resultados o tantos sorpresa que pueden deteriorar la estimación de los parámetros. Esto ocurre especialmente en el caso de competiciones con un número insuficiente de juegos para comparar a los equipos participantes (por ejemplo, Liga de Campeones). En este artículo del 2011 proponen una función de verosimilitud ponderada que permita dar poco peso a un marcador específico si se presupone que el resultado no es el habitual y falsifica de alguna forma el parámetro estimado. Esa ponderación puede ser definida subjetivamente o ser asumida por una estructura del modelo donde los parámetros pueden ser estimados por algoritmos iterativos. La estructura ponderada normalmente refleja desviaciones del modelo asumido. Este procedimiento puede proporcionar estimaciones robustas incluso si algún marcador sorprendente (bajo el modelo asumido) es observado. Los datos de la Liga de Campeones se usan para demostrar el potencial de la aproximación propuesta.

- **Baio, G. & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics, 37* (2), pp. 253-264.**

  El problema de aplicar modelos a datos futbolísticos ha ido aumentado de una forma popular en los últimos años y diferentes modelos han sido propuestos con la finalidad de estimar las características que pueden hacer que un equipo pierda o gane un juego, o predecir un marcador para un partido en particular. Los autores proponen un modelo jerárquico bayesiano para conseguir ambos objetivos y probar la fuerza predictiva basándose en el campeonato italiano Serie A 1991-1992. Para superar algunos problemas del modelo jerárquico bayesiano, se ha especificado un modelo mixto más complicado para que los resultados se ajusten mejor a los datos observados.

- **Skinner, G.K & Freeman, G.H. (2009). Soccer matches as experiments: How often does the 'best' team win? *Journal of Applied Statistics, 36* (10), pp. 1087-1095.**

  Los modelos en los cuales el número de goles marcados por un equipo en un partido de fútbol siguen una distribución de Poisson, o una próxima, se discuten en este artículo. Los autores consideran que un partido de fútbol como un experimento que permite evaluar cual de los dos equipos es superior y examina la probabilidad de que el resultado del experimento (partido) verdaderamente represente las habilidades relativas de los dos equipos. Teniendo en cuenta el resultado final es posible, utilizando una aproximación bayesiana, cuantificar la probabilidad de si era o no el caso de que el mejor equipo ha ganado. Los autores indican que habría que modificar las reglas del juego para conseguir que el mejor equipo sea el que gane.

- **Karlis, D. & Ntzoufras, I. (2009). Bayesian modelling of football**

outcomes: Using the Skellam's distribution for the goal difference. *IMA Journal Management Mathematics, 20* (2), pp. 133-145.

Karlis y Ntzoufras mencionan de nuevo como la aplicación de modelos a los partidos de fútbol se ha incrementado hoy en día tanto para los dirigentes de equipos como por el tema de las apuestas. La mayoría de la literatura existente hace referencia a aplicar modelos que tengan en cuenta los goles marcados por cada equipo. El artículo de estos autores, propone un nuevo planteamiento, en vez de tener en cuenta directamente el número de goles, se centra en la diferencia del número de goles, es decir, en el margen de la victoria. Aplicando este modelo se obtienen varias ventajas. Por una parte, se elimina la correlación impuesta por el hecho de que dos equipos oponentes compiten uno contra otro, y en segundo lugar, no es necesario asumir que los goles marcados por cada equipo sean marginalmente distribuidos con una distribución Poisson. Los autores discuten la aplicación de la metodología bayesiana para la distribución de Skellam usando covarianzas. Se ilustra este artículo usando datos reales de la Liga de Fútbol inglesa para la temporada 2006-2007. También se discuten las ventajas de la aproximación propuesta.

- **Greenhough, J., Birch, P.C., Chapman, S.C. & Rowlands, G. (2002). Football goal distributions and extremal statistics. *Physica A: Statistical Mechanics and its Applications, 316* (1-4), pp. 615-624.**

  Los autores de este artículo analizan las distribuciones del número de goles marcados por los equipos que juegan en casa, fuera de casa y el total de goles marcados en cada partido, en partidos de 169 países entre 1999 y 2001. Las funciones de densidad de probabilidad de los goles marcados presentan colas pesadas que no se ajustan correctamente por una distribución binomial negativa o Poisson, las cuales se esperan de un proceso no correlacionado. Los autores indican que las funciones de densidad de probabilidad son consistentes con aquellas que se obtienen de estadísticas extremas. Además, ilustran su modelo con partidos de la División inglesa y de la Copa de Inglaterra en las temporadas 1970/71-2000/01.

- **Dyte, D. & Clarke, S.R. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society, 51* (8), pp. 993-998.**

  En este artículo se sugiere un método para predecir la distribución de las puntuaciones en partidos de fútbol internacionales, tratando los goles de los equipos como variables de Poisson independientes según la puntuación de la FIFA de cada equipo y el lugar del partido. Los resultados de una regresión de Poisson para estimar los parámetros para este modelo se usaron para simular partidos jugados durante el torneo de la Copa del Mundo de 1998. Para que el modelo sea un predictor efectivo, se han realizado algunos ajustes en la clasificación de los datos. Las predicciones del modelo fueron colocadas en una página web para mostrar el interés de las aplicaciones de las matemáticas y

ser más popular para el público en general.

2.2. Datos de la Liga Española de Fútbol

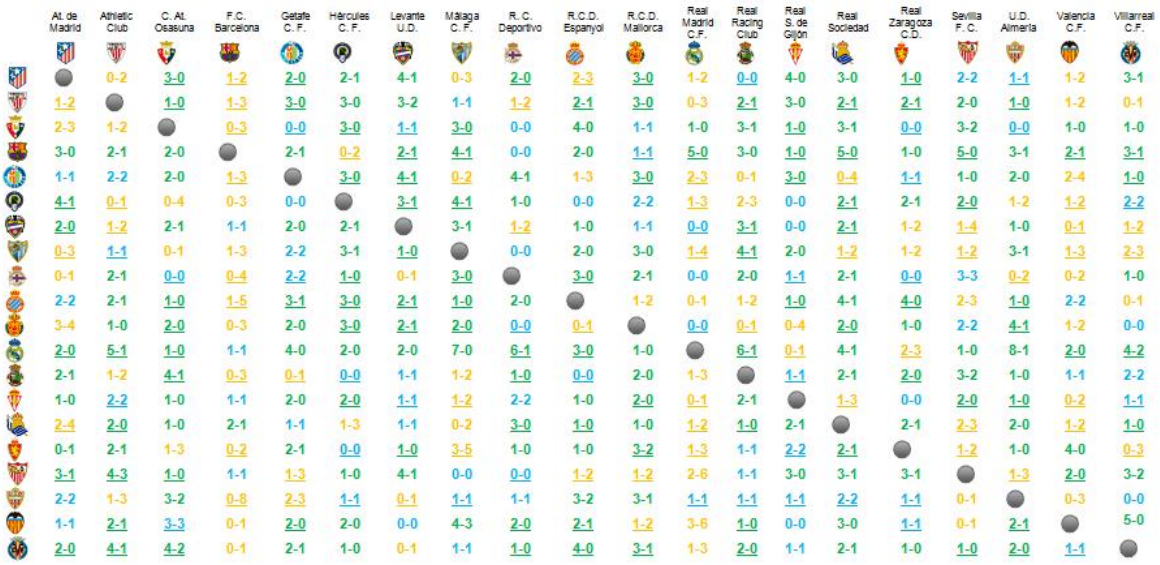
En este trabajo se utilizan los datos de la Liga Española de Fútbol. Para conseguirlos se ha consultado la página web de la Liga Nacional de Fútbol Profesional (LFP). La LFP, es una asociación deportiva de derecho privado integrada exclusiva y obligatoriamente por todas las Sociedades Anónimas Deportivas y Clubes que participan en competiciones oficiales de fútbol de ámbito estatal y carácter profesional y a la que corresponde legalmente la organización de dichas competiciones, en coordinación con la Real Federación Española de Fútbol. En dicha página se pueden encontrar las estadísticas y las fichas de los partidos celebrados en nuestro país desde el año 1929. Para la realización de este trabajo se tendrán en cuenta específicamente los resultados de la Liga Española de Fútbol de las temporadas 2010/2011 y 2011/2012 correspondientes a Primera División.

La liga de Primera División está integrada por 20 equipos. El torneo se juega entre los meses de septiembre y junio del siguiente año. Cada equipo se enfrenta en dos ocasiones a los demás en su propio campo y en el campo del adversario, según el calendario que se sortea a principios de temporada. El artículo 201 del Reglamento General de la Real Federación Española de Fútbol especifica que "la clasificación final se establecerá con arreglo a los puntos obtenidos por cada uno de los clubes contendientes, a razón de tres por partido ganado, uno por empatado y cero por perdido" (RFEF, 2012, p. 122). En caso de que al finalizar el campeonato, hubiese un empate entre dos clubes, este se resolvería por la mayor diferencia de goles a favor, sumados los en pro y en contra según el resultado de los partidos jugados entre ellos.

2.2.1. Temporada 2010/2011

2.2.1.1. Resultados de los partidos

En la temporada 2010/2011, 20 equipos jugaron en Primera División. Estos equipos fueron: At. de Madrid, Athletic Club, C. At. Osasuna, F.C. Barcelona, Getafe C. F., Hércules C.F., Levante U.D, Málaga C.F., R.C. Deportivo, R.C.D. Espanyol, R.C.D. Mallorca, Real Madrid C.F., Real Racing Club, Real S. de Gijón, Real Sociedad, Real Zaragoza C.D., Sevilla F.C., U.D. Almería, Valencia C. F y Villarreal C.F. La temporada comenzó el 28/08/2010 y finalizó el 21/05/2011. En total se celebraron 380 partidos. Los resultados de estos partidos se muestran a continuación:



	At. de Madrid	Athletic Club	C. At. Osasuna	F.C. Barcelona	Getafe C. F.	Hércules C. F.	Levante U.D.	Málaga C. F.	R. C. Deportivo	R.C.D. Espanyol	R.C.D. Mallorca	Real Madrid C.F.	Real Racing Club	Real S. de Gijón	Real Sociedad	Real Zaragoza C.D.	Sevilla F.C.	U.D. Almería	Valencia C.F.	Villarreal C.F.
At. de Madrid		0-2	3-0	1-2	2-0	2-1	4-1	0-3	2-0	2-3	3-0	1-2	0-0	4-0	3-0	1-0	2-2	1-1	1-2	3-1
Athletic Club	1-2		1-0	1-3	3-0	3-0	3-2	1-1	1-2	2-1	3-0	0-3	2-1	3-0	2-1	2-1	2-0	1-0	1-2	0-1
C. At. Osasuna	2-3	1-2		0-3	0-0	3-0	1-1	3-0	0-0	4-0	1-1	1-0	3-1	1-0	3-1	0-0	3-2	0-0	1-0	1-0
F.C. Barcelona	3-0	2-1	2-0		2-1	0-2	2-1	4-1	0-0	2-0	1-1	5-0	3-0	1-0	5-0	1-0	5-0	3-1	2-1	3-1
Getafe C. F.	1-1	2-2	2-0	1-3		3-0	4-1	0-2	4-1	1-3	3-0	2-3	0-1	3-0	0-4	1-1	1-0	2-0	2-4	1-0
Hércules C. F.	4-1	0-1	0-4	0-3	0-0		3-1	4-1	1-0	0-0	2-2	1-3	2-3	0-0	2-1	2-1	2-0	1-2	1-2	2-2
Levante U.D.	2-0	1-2	2-1	1-1	2-0	2-1		3-1	1-2	1-0	1-1	0-0	3-1	0-0	2-1	1-2	1-4	1-0	0-1	1-2
Málaga C. F.	0-3	1-1	0-1	1-3	2-2	3-1	1-0		0-0	2-0	3-0	1-4	4-1	2-0	1-2	1-2	1-2	3-1	1-3	2-3
R. C. Deportivo	0-1	2-1	0-0	0-4	2-2	1-0	0-1	3-0		3-0	2-1	0-0	2-0	1-1	2-1	0-0	3-3	0-2	0-2	1-0
R.C.D. Espanyol	2-2	2-1	1-0	1-5	3-1	3-0	2-1	1-0	2-0		1-2	0-1	1-2	1-0	4-1	4-0	2-3	1-0	2-2	0-1
R.C.D. Mallorca	3-4	1-0	2-0	0-3	2-0	3-0	2-1	2-0	0-0	0-1		0-0	0-1	0-4	2-0	1-0	2-2	4-1	1-2	0-0
Real Madrid C.F.	2-0	5-1	1-0	1-1	4-0	2-0	2-0	7-0	6-1	3-0	1-0		6-1	0-1	4-1	2-3	1-0	8-1	2-0	4-2
Real Racing Club	2-1	1-2	4-1	0-3	0-1	0-0	1-1	1-2	1-0	0-0	2-0	1-3		1-1	2-1	2-0	3-2	1-0	1-1	2-2
Real S. de Gijón	1-0	2-2	1-0	1-1	2-0	2-0	1-1	1-2	2-2	1-0	2-0	0-1	2-1		1-3	0-0	2-0	1-0	0-2	1-1
Real Sociedad	2-4	2-0	1-0	2-1	1-1	1-3	1-1	0-2	3-0	1-0	1-0	1-2	1-0	2-1		2-1	2-3	2-0	1-2	1-0
Real Zaragoza C.D.	0-1	2-1	1-3	0-2	2-1	0-0	1-0	3-5	1-0	1-0	3-2	1-3	1-1	2-2	2-1		1-2	1-0	4-0	0-3
Sevilla F.C.	3-1	4-3	1-0	1-1	1-3	1-0	4-1	0-0	0-0	1-2	1-2	2-6	1-1	3-0	3-1	3-1		1-3	2-0	3-2
U.D. Almería	2-2	1-3	3-2	0-8	2-3	1-1	0-1	1-1	1-1	3-2	3-1	1-1	1-1	1-1	2-2	1-1	0-1		0-3	0-0
Valencia C.F.	1-1	2-1	3-3	0-1	2-0	2-0	0-0	4-3	2-0	2-1	1-2	3-6	1-0	0-0	3-0	1-1	0-1	2-1		5-0
Villarreal C.F.	2-0	4-1	4-2	0-1	2-1	1-0	0-1	1-1	1-0	4-0	3-1	1-3	2-0	1-1	2-1	1-0	1-0	2-0	1-1	

Figura 2.1.: Resultados Temporada 2010/2011

De los 380 partidos, 78 terminaron con un resultado de empate a cero, uno, dos o tres goles para cada equipo (color azul en la tabla), 102 resultaron victorias en casa (color verde) y 78 victorias fuera de casa (color amarillo). En la Figura 2.1 aparecen subrayados los encuentros de la primera vuelta.

2.2.1.2. Clasificación

Al finalizar el campeonato, el F.C.Barcelona fue el equipo que se proclamó campeón de liga, consiguiendo 96 puntos (50 de ellos en partidos jugados en casa y 46 en partidos jugados como visitante). El siguiente equipo fue el Real Madrid C.F. con un total de 92 puntos (49 de ellos obtenidos cuando jugaba en casa y 43 obtenidos fuera. El tercer equipo en la clasificación fue el Valencia C.F. con un total de 71 puntos (35 se consiguieron en casa y 36 fuera). En la siguiente figura se muestran los puntos conseguidos por cada uno de los 20 equipos enfrentados en la Primera División durante esta temporada. Además de los puntos totales se incluyen el número total de partidos jugados, los partidos ganados, empatados y perdidos. También aparecen en la figura los goles a favor y los goles en contra. Esta información se proporciona para el total de partidos jugados, para los partidos jugados en casa y para los partidos jugados fuera de casa.

En la clasificación de la Figura 2.2 los equipos aparecen ordenados según el puesto alcanzado en la clasificación general según los puntos totales conseguidos por cada uno de los equipos (PT).

	TOTAL							EN CASA							FUERA						
EQUIPO	Pt	PJ	PG	PE	PP	GF	GC	Pt	PJ	PG	PE	PP	GF	GC	Pt	PJ	PG	PE	PP	GF	GC
Barcelona	96	38	30	6	2	95	21	50	19	16	2	1	46	10	46	19	14	4	1	49	11
Real Madrid	92	38	29	5	4	102	33	49	19	16	1	2	61	12	43	19	13	4	2	41	21
Valencia	71	38	21	8	9	64	44	35	19	10	5	4	34	21	36	19	11	3	5	30	23
Villarreal	62	38	18	8	12	54	44	42	19	13	3	3	33	14	20	19	5	5	9	21	30
Sevilla	58	38	17	7	14	62	61	34	19	10	4	5	35	27	24	19	7	3	9	27	34
Athletic	58	38	18	4	16	59	55	37	19	12	1	6	32	20	21	19	6	3	10	27	35
Atlético	58	38	17	7	14	62	53	33	19	10	3	6	35	20	25	19	7	4	8	27	33
Espanyol	49	38	15	4	19	46	55	35	19	11	2	6	33	22	14	19	4	2	13	13	33
Osasuna	47	38	13	8	17	45	46	36	19	10	6	3	28	14	11	19	3	2	14	17	32
Sporting	47	38	11	14	13	35	42	33	19	9	6	4	23	16	14	19	2	8	9	12	26
Málaga	46	38	13	7	18	54	68	24	19	7	3	9	29	29	22	19	6	4	9	25	39
Racing	46	38	12	10	16	41	56	30	19	8	6	5	25	21	16	19	4	4	11	16	35
Real Zaragoza	45	38	12	9	17	40	53	30	19	9	3	7	26	27	15	19	3	6	10	14	26
Levante	45	38	12	9	17	41	52	31	19	9	4	6	25	20	14	19	3	5	11	16	32
R. Sociedad	45	38	14	3	21	49	66	35	19	11	2	6	27	21	10	19	3	1	15	22	45
Getafe	44	38	12	8	18	49	60	30	19	9	3	7	33	26	14	19	3	5	11	16	34
Mallorca	44	38	12	8	18	41	56	31	19	9	4	6	25	19	13	19	3	4	12	16	37
Deportivo	43	38	10	13	15	31	47	30	19	8	6	5	22	19	13	19	2	7	10	9	28
Hércules	35	38	9	8	21	36	60	26	19	7	5	7	27	27	9	19	2	3	14	9	33
Almería	30	38	6	12	20	36	70	19	19	3	10	6	23	35	11	19	3	2	14	13	35

Figura 2.2.: Clasificación Temporada 2010/2011

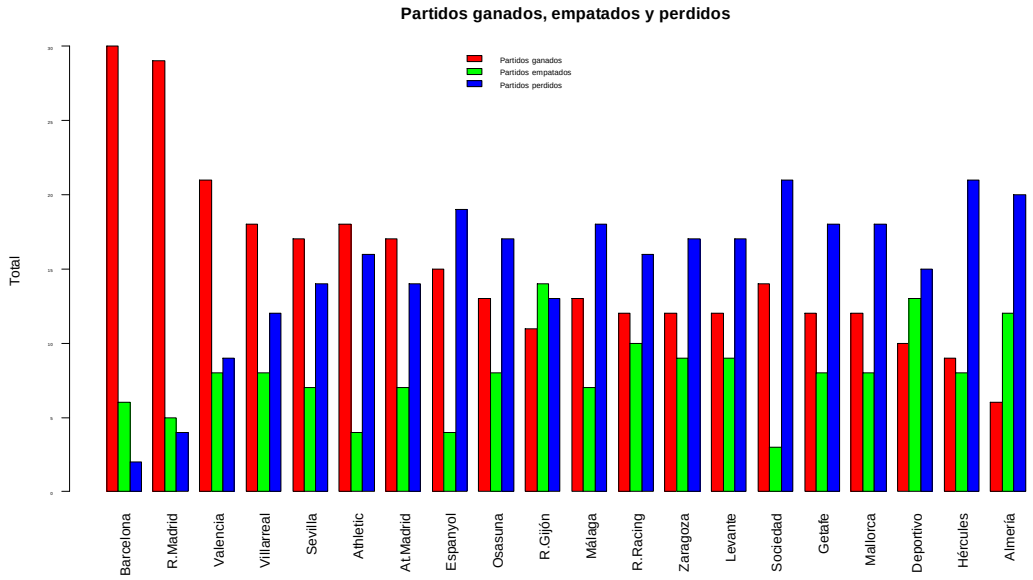


Figura 2.3.: Gráfico partidos ganados, empatados y perdidos Temporada 2010/2011

Además, se muestran el número total de partidos jugados (PJ) y de éstos, cuántos han resultado como partidos ganados (PG), cuántos empatados (PE) y cuántos partidos perdidos (PP). Finalmente, también se indica el número de goles a favor (GF) y el número de goles en contra (GC).

En la Figura 2.3 se presentan el total de partidos ganados, empatados y perdidos para cada equipo. Las barras rojas representan el número total de partidos ganados, las barras verdes los partidos empatados y por último, las barras azules el número

de partidos perdidos. Dado que los equipos están ordenados por la puntuación obtenida, se observa que los mejores clasificados ganan más partidos y ese número va descendiendo conforme los equipos se alejan en la tabla general de clasificación.

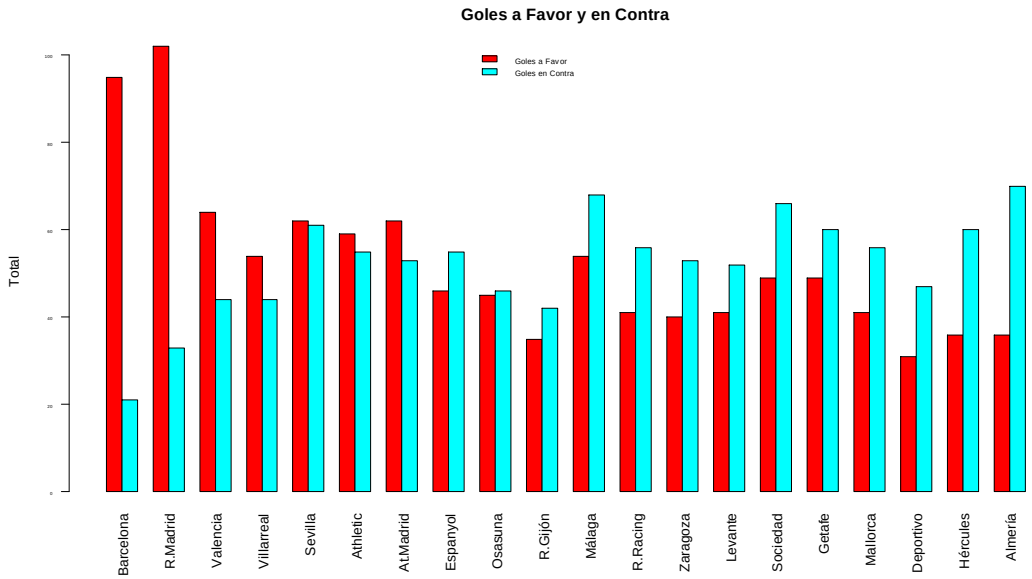


Figura 2.4.: Gráfico goles a favor y en contra Temporada 2010/2011

Una tendencia similar se aprecia en relación a los goles a favor y en contra (ver Figura 2.4). Los primeros equipos de la clasificación tienen un mayor número de goles a favor que goles en contra mientras que los equipos situados al final de la clasificación presentan más goles en contra que a favor.

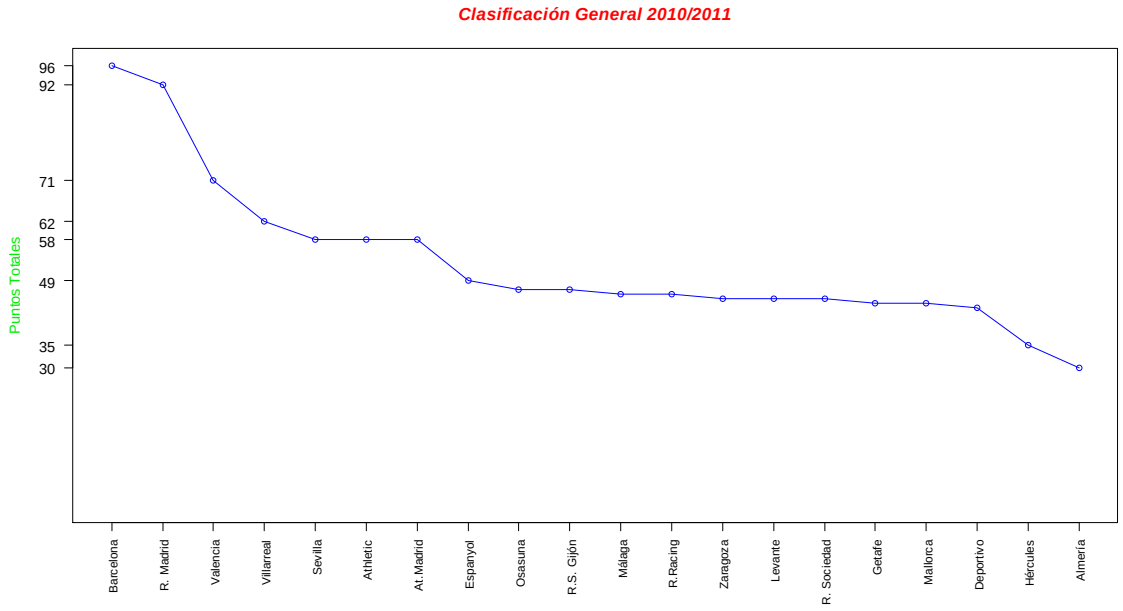


Figura 2.5.: Gráfico clasificación general según los puntos Temporada 2010/2011

En la Figura 2.5 se muestra la clasificación general de la temporada 2010/2011 teniendo en cuenta los puntos totales. En el gráfico se observa que los dos primeros equipos clasificados (F.C. Barcelona y R. Madrid) tienen una mayor puntuación que el resto de los equipos. Después sigue un segundo grupo de equipos (los 5 siguientes) que son inferiores a los dos primeros pero presentan un número mayor de puntos que los equipos que siguen a continuación. Por debajo de esos cinco equipos se situarían los restantes muy igualados a puntos.

2.2.2. Temporada 2011/2012

2.2.2.1. Resultados de los partidos

La temporada del 2011/2012 comenzó el 27/08/2011 y finalizó el 13/05/2012. Los equipos que jugaron en Primera División ese año son los mismos de la temporada a excepción del R.C.Deportivo, Hércules C.F. y U.D. Almería que descienden a Segunda División. Los equipos que ascienden esa temporada para jugar en la categoría superior son el Granada C.F., el Real Betis B.S y el Rayo Vallecano. En la siguiente tabla se muestran los resultados de las 38 jornadas:

	Real S. de Gijón	Granada C.F	Valencia C. F.	Athletic Club	At. de Madrid	Getafe C.F.	R.C.D. Mallorca	Sevilla F.C.	Real Zaragoza C.D.	F.C. Barcelona	Real Sociedad	Real Betis B.S.	Real Racing Club	Rayo Vallecano	C. At. Osasuna	Levante U.D.	R.C.D. Espanyol	Málaga C.F.	Real Madrid C. F.	Villarreal C.F.
Real S. de Gijón		2-0	0-1	1-0	1-1	2-1	2-3	1-0	1-2	0-1	1-2	2-1	0-0	2-1	1-1	3-2	1-2	2-1	0-3	2-3
Granada C.F	2-1		0-1	2-2	0-0	1-0	2-2	0-3	1-0	0-1	4-1	0-1	0-0	1-2	1-1	2-1	2-1	2-1	1-2	1-0
Valencia C. F.	4-0	1-0		1-1	1-0	3-1	2-2	1-2	1-2	2-2	0-1	4-0	4-3	4-1	4-0	1-1	2-1	2-0	2-3	1-0
Athletic Club	1-0	0-1	0-3		3-0	0-0	1-0	1-0	2-1	2-2	2-0	2-3	1-1	1-1	3-1	3-0	3-3	3-0	0-3	1-1
At. de Madrid	4-0	2-0	0-0	2-1		3-0	1-1	0-0	3-1	1-2	1-1	0-2	4-0	3-1	0-0	3-2	3-1	2-1	1-4	3-0
Getafe C.F.	2-0	1-0	3-1	0-0	3-2		1-3	0-1	0-2	1-0	1-0	1-0	1-1	0-1	2-2	1-1	1-1	1-3	0-1	0-0
R.C.D. Mallorca	1-2	0-0	1-1	1-1	2-1	1-2		0-0	1-0	0-2	2-1	1-0	2-1	1-0	1-1	1-0	1-0	0-1	1-2	4-0
Sevilla F.C.	2-1	1-2	1-0	1-2	1-1	3-0	3-1		3-0	2-0	1-0	1-2	2-2	5-2	2-0	1-1	0-0	2-1	2-6	1-2
Real Zaragoza C.D.	2-2	1-0	0-1	2-0	1-0	1-1	0-1	0-1		1-4	2-0	0-2	2-1	1-2	1-1	1-0	2-1	0-0	0-6	2-1
F.C. Barcelona	3-1	5-3	5-1	2-0	5-0	4-0	5-0	0-0	4-0		2-1	4-2	3-0	4-0	8-0	5-0	4-0	4-1	1-2	5-0
Real Sociedad	5-1	1-0	1-0	1-2	0-4	0-0	1-0	2-0	3-0	2-2		1-1	3-0	4-0	0-0	1-3	0-0	3-2	0-1	1-1
Real Betis B.S.	2-0	1-2	2-1	2-1	2-2	1-1	1-0	1-1	4-3	2-2	2-3		1-1	0-2	1-0	0-1	1-1	0-0	2-3	3-1
Real Racing Club	1-1	0-1	2-2	0-1	0-0	1-2	0-3	0-3	1-0	0-2	0-0	1-0		1-1	2-4	0-0	0-1	1-3	0-0	1-0
Rayo Vallecano	1-3	1-0	1-2	2-3	0-1	2-0	0-1	2-1	0-0	0-7	4-0	3-0	4-2		6-0	1-2	0-1	2-0	0-1	0-2
C. At. Osasuna	2-1	2-1	1-1	2-1	0-1	0-0	2-2	0-0	3-0	3-2	1-0	2-1	0-2	0-0		2-0	2-0	1-1	1-5	2-1
Levante U.D.	4-0	3-1	0-2	3-0	2-0	1-2	0-0	1-0	0-0	1-2	3-2	3-1	1-1	3-5	0-2		3-1	3-0	1-0	1-0
R.C.D. Espanyol	0-3	3-0	4-0	2-1	4-2	1-0	1-0	1-1	0-2	1-1	2-2	1-0	3-1	5-1	1-2	1-2		1-2	0-4	0-0
Málaga C.F.	1-0	4-0	1-0	1-0	0-0	3-2	3-1	2-1	5-1	1-4	1-1	0-2	3-0	4-2	1-1	1-0	2-1		0-4	2-1
Real Madrid C. F.	3-1	5-1	0-0	4-1	4-1	4-2	4-1	3-0	3-1	1-3	5-1	4-1	4-0	6-2	7-1	4-2	5-0	1-1		3-0
Villarreal C.F.	3-0	3-1	2-2	2-2	0-1	1-2	2-0	2-2	2-2	0-0	1-1	1-0	1-1	2-0	1-1	0-3	0-0	2-1	1-1	

Figura 2.6.: Resultados Temporada 2011/2012

Al igual que la temporada anterior, se jugaron 380 partidos, de los cuales 91 resultaron con un empate a cero, uno, dos, y tres goles (resultados de color azul en la tabla), 94 fueron victorias en casa (color verde en la tabla) y, finalmente, 95 fueron victorias fuera de casa (color amarillo en la tabla).

2.2.2.2. Clasificación

En la jornada 38 de la Temporada 2011/2012 la clasificación de la Primera División estaba encabezada por el Real Madrid C.F. con un total de 100 puntos (la mitad de

ellos conseguidos en casa y la otra mitad fuera). Le sigue el F.C. Barcelona con 91 puntos los cuáles obtuvo mayoritariamente en partidos jugados en casa (52 en casa y 39 fuera de ella). El tercer equipo en la clasificación fue el Valencia C.F. con un total de 61 puntos (37 en partidos celebrados en casa y 24 conseguidos como visitante). La clasificación general de la temporada 2011/2012 se muestra a continuación:

	TOTAL							EN CASA							FUERA						
EQUIPO	Pt	PJ	PG	PE	PP	GF	GC	Pt	PJ	PG	PE	PP	GF	GC	Pt	PJ	PG	PE	PP	GF	GC
Real Madrid	100	38	32	4	2	121	32	50	19	16	2	1	70	19	50	19	16	2	1	51	13
Barcelona	91	38	28	7	3	114	29	52	19	17	1	1	73	11	39	19	11	6	2	41	18
Valencia	61	38	17	10	11	59	44	37	19	11	4	4	40	20	24	19	6	6	7	19	24
Málaga	58	38	17	7	14	54	53	42	19	13	3	3	35	21	16	19	4	4	11	19	32
Atlético	56	38	15	11	12	53	46	38	19	11	5	3	36	17	18	19	4	6	9	17	29
Levante	55	38	16	7	15	54	50	36	19	11	3	5	33	19	19	19	5	4	10	21	31
Osasuna	54	38	13	15	10	44	61	36	19	10	6	3	26	19	18	19	3	9	7	18	42
Mallorca	52	38	14	10	14	42	46	32	19	9	5	5	21	15	20	19	5	5	9	21	31
Sevilla	50	38	13	11	14	48	47	31	19	9	4	6	32	25	19	19	4	7	8	16	22
Athletic	49	38	12	13	13	49	52	31	19	8	7	4	29	21	18	19	4	6	9	20	31
Getafe	47	38	12	11	15	40	51	30	19	8	6	5	24	19	17	19	4	5	10	16	32
R. Sociedad	47	38	12	11	15	46	52	33	19	9	6	4	29	17	14	19	3	5	11	17	35
Betis	47	38	13	8	17	47	56	28	19	7	7	5	28	25	19	19	6	1	12	19	31
Espanyol	46	38	12	10	16	46	56	31	19	9	4	6	31	24	15	19	3	6	10	15	32
Rayo	43	38	13	4	21	53	73	25	19	8	1	10	29	26	18	19	5	3	11	24	47
Real Zaragoza	43	38	12	7	19	36	61	28	19	8	4	7	19	24	15	19	4	3	12	17	37
Granada	42	38	12	6	20	35	56	29	19	8	5	6	22	20	13	19	4	1	14	13	36
Villarreal	41	38	9	14	15	39	53	28	19	6	10	3	26	20	13	19	3	4	12	13	33
Sporting	37	38	10	7	21	42	69	25	19	7	4	8	24	26	12	19	3	3	13	18	43
Racing	27	38	4	15	19	28	63	16	19	3	7	9	11	24	11	19	1	8	10	17	39

Figura 2.7.: Clasificación Temporada 2011/2012

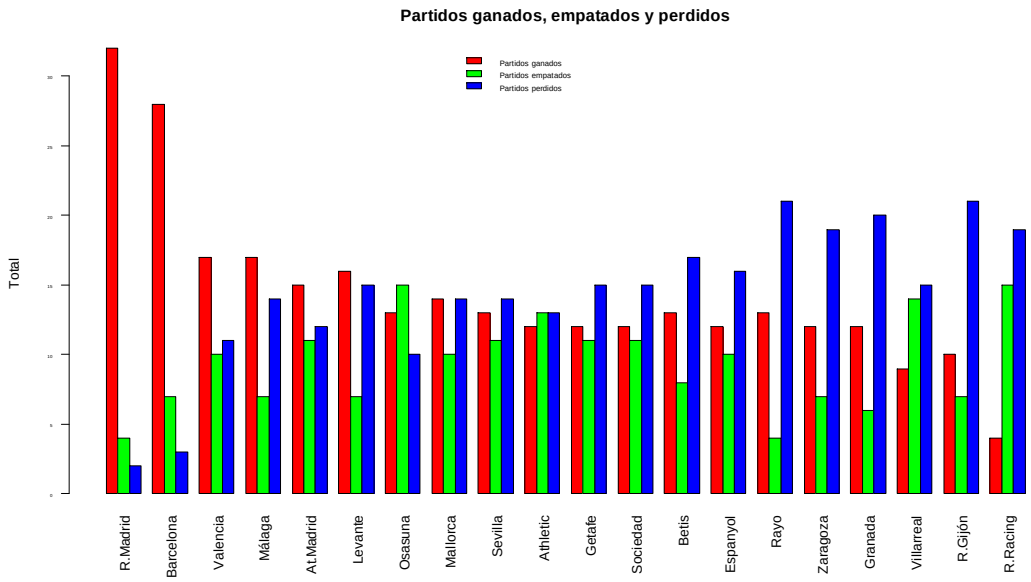


Figura 2.8.: Gráfico partidos ganados, empatados y perdidos Temporada 2011/2012

En el anterior gráfico (2.8) se muestran los partidos ganados, empatados y perdidos para los 20 equipos de esta temporada. Además, se observa como los dos primeros

clasificados (R. Madrid y Barcelona F.C.) tienen mayor número de partidos ganados que perdidos. Ocurre en sentido contrario respecto a los equipos clasificados en las últimas posiciones, en este caso, es mayor el número de partidos perdidos que los ganados.

El gráfico de barras que presenta los goles a favor y en contra para los equipos es el siguiente:

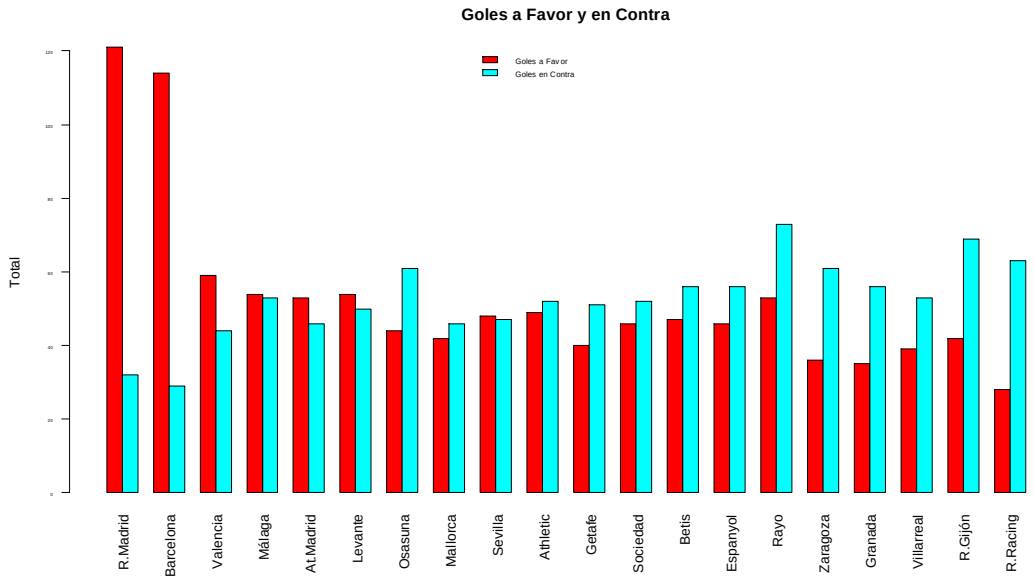


Figura 2.9.: Gráfico goles a favor y en contra Temporada 2011/2012

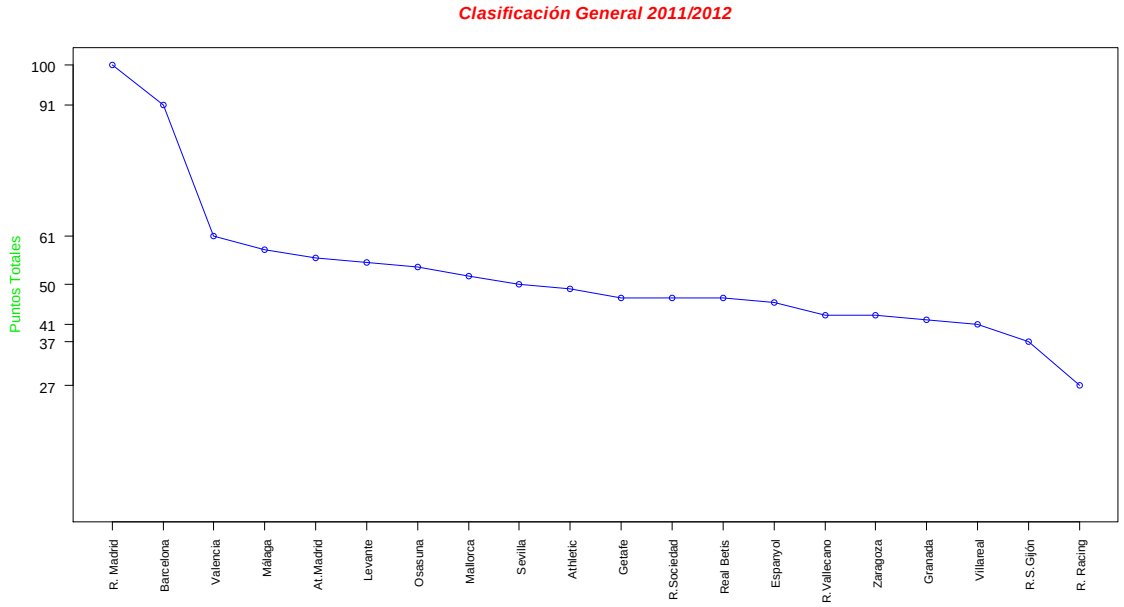


Figura 2.10.: Gráfico clasificación general según los puntos Temporada 2011/2012

Igual que ocurría en la temporada anterior, los dos primeros clasificados presentan el mayor número de goles a favor que en contra. Esta tendencia se invierte a medida que se observa los goles a favor y en contra en los últimos puestos de la clasificación. En la figura 2.10 se muestra la clasificación general teniendo en cuenta el número total de puntos para cada equipo. En la temporada 2011/2012, de nuevo los dos primeros clasificados obtuvieron una mayor puntuación que el resto de los equipos. Los equipos que se clasificaron a continuación están muy igualados, siendo los dos últimos los que presentaron una peor puntuación respecto al resto.

3. Métodos

En este capítulo se realiza una descripción de los modelos estadísticos de interés en este trabajo. En un primer punto, se introduce el modelo de regresión de Poisson para, a continuación, presentar las características principales del modelo de regresión de Poisson Bivariante. Finalmente, en un tercer punto se describe el modelo de regresión de Poisson Bivariante con Inflado en la Diagonal.

3.1. Introducción al Modelo de Regresión de Poisson

El modelo de regresión de Poisson es un tipo de Modelo Lineal Generalizado. Los Modelos Lineales Generalizados permiten incluir distintas relaciones entre las medias condicionales de las variables respuesta y las explicativas. El modelo de regresión de Poisson se utiliza para datos de conteo. Este modelo es adecuado cuando la varianza muestral es igual a la media.

En el modelo de regresión de Poisson la media (λ) se explica en términos de las variables explicativas mediante el uso de un enlace.

3.1.1. Distribución de Poisson

Sea Y una variable aleatoria discreta, se dice que Y sigue una distribución de Poisson con parámetro λ si su función de probabilidad viene dada por $f(y) = P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$, para $y = 0, 1, 2, ...$, donde,

y: es el número de veces que ocurre un evento de interés,

λ: es un parámetro positivo que representa el número de veces que se espera que ocurra el evento en un período determinado.

Las principales propiedades de la distribución de Poisson son las siguientes:

1. La media y la varianza son iguales a su parámetro, $E(Y) = \lambda = Var(Y)$.

2. Si λ crece, la masa de la distribución se desplaza hacia la derecha y $P(Y = 0)$ decrece.

3. A medida que λ crece, la distribución de Poisson se aproxima a una distribución normal por el TCL.

3.1.2. Modelo de Regresión Poisson

El modelo de regresión de Poisson surge cuando la variable respuesta es una cantidad discreta que se puede modelizar con una Poisson y se quiere estudiar si ciertas variables explicativas influyen en la variable respuesta y cómo lo hacen. Este tipo de variable respuesta suele representar el recuento de sucesos o hechos (por ejemplo, el número de goles marcados en un partido).

Por tanto, se considera una variable respuesta Y que toma los valores en $\{0, 1, 2, ..\}$, y se va estudiar su relación con otras variables explicativas X mediante un análisis de regresión. Se pretende construir un modelo para $\lambda(x) = E(Y \mid X = x)$ es decir, para la media de Y condicionada a cada valor de la variable explicativa.

Como Y nunca toma valores negativos, no procede utilizar un modelo lineal directo, y por tanto se necesita una función de enlace previa a cualquier modelo lineal. Además, la función de regresión está en el intervalo $(0, +\infty)$, el logaritmo parece la función de enlace más adecuada. Se expresaría de la siguiente forma: $g(\lambda(x, \beta)) = x'\beta$ donde lo más habitual es tomar $g(r) = log(r)$, $r \in (0, +\infty)$. Mediante $x'\beta$ se representa el producto escalar del vector de variables explicativas por el vector de parámetros. Para incluir un intercepto se considera una primera variable explicativa igual a 1.

La función de regresión del modelo de Poisson se expresaría por: $\lambda(x, \beta) = e^{x'\beta}$.

Los parámetros de este modelo, que se puede denominar modelo log-lineal, se interpretan de la siguiente forma: la exponencial del intercepto es el valor esperado de la respuesta en la categoría de referencia o cuando las variables explicativas numéricas valen cero, y las exponenciales de los coeficientes de cada variable representan tasas de incremento de la respuesta esperada al aumentar una unidad la variable si es numérica, o al pasar a la categoría correspondiente si es cualitativa. Es decir, el modelo supone efectos multiplicativos: si la componente explicativa unidimensional X_j aumenta n unidades, la media para la variable de Poisson se multiplica por la potencia n-ésima de e^{β_j}, supuestas las demás variables explicativas constantes, matemáticamente

$$\frac{\lambda((x_1, ..., x_j + n, ..., x_p), \beta)}{\lambda((x_1, ..., x_j, ..., x_p), \beta)} = \frac{e^{\beta_0 + x_1\beta_1 + ... + (x_j+n)\beta_j + ... x_p\beta_p}}{e^{\beta_0 + x_1\beta_1 + ... + x_j\beta_j + ... x_p\beta_p}} = e^{n\beta_j} = (e^{\beta_j})^n.$$

Si se tiene una muestra aleatoria simple $(X_1, Y_1), ..., (X_n, Y_n)$ de (X, Y), entonces $Y_i \in Poisson(\lambda(X_i, \beta))$ siendo $\lambda(x, \beta) = e^{x'\beta}$.

Para estimar los parámetros del modelo se utiliza la función de máxima verosimilitud, que adopta la siguiente forma: $L(\beta) = \prod_{i=1}^{n} \left[e^{-\lambda(x_i,\beta)} \frac{\lambda(x_i,\beta)^{y_i}}{y_i!}\right]$, siendo su logaritmo (en términos dependientes de β)

$$l(\beta) = \sum_{i=1}^{n} \left(y_i x_i'\beta - e^{x_i'\beta}\right) \tag{3.1}$$

Derivando dicha función e igualando a cero, se obtienen las ecuaciones de verosimilitud: $\frac{\partial logL(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i'[y_i - \lambda(x_i, \beta] = 0$,

La matriz hessiana adopta la forma: $\frac{\partial^2 logL(\beta)}{\partial \beta \partial \beta^t} = -\sum_{i=1}^{n} x_i x_i' \lambda(x_i, \beta)$.

Igual que ocurre en la regresión logística, las ecuaciones de verosimilitud no son lineales en los parámetros y es necesario recurrir a procedimientos iterativos para el cálculo de sus estimaciones. Newton-Raphson y el IRLS (Iterative Re-weighted Least Squares) son los métodos iterativos que se aplican en este modelo.

Además, es conocido que los estimadores de máxima verosimilitud son asintóticamente normales y centrados y su matriz de varianzas-covarianzas es la inversa de la matriz de información (la matriz hessiana cambiada de signo), lo cual permite hacer inferencias sobre los parámetros del modelo.

En R, los modelos lineales generalizados se ajustan con la función *glm* y en la regresión de Poisson se debe especificar *family=poisson(link=log)*.

3.2. Modelo de Regresión Bivariante de Poisson

El modelo univariado de regresión de Poisson se usa para el análisis de la relación entre un conteo observado con una distribución de Poisson y un conjunto de variables explicativas. En el caso del modelo de regresión bivariante de Poisson el vector respuesta es bidimensional, sigue una distribución bivariada de Poisson y las medias marginales son funciones de las variables explicativas (Kocherlakota & Kocherlakota, 2001).

3.2.1. Distribución de Poisson Bivariada

Sean $X_\kappa, \kappa = 1, 2, 3$ tres variables aleatorias con distribuciones de Poisson independientes y parámetros $\lambda_\kappa > 0$, respectivamente. Entonces se dice que las variables aleatorias $X = X_1 + X_3$ e $Y = X_2 + X_3$ (transformadas de las anteriores) siguen conjuntamente una distribución Bivariada de Poisson, $BP(\lambda_1, \lambda_2, \lambda_3)$. Su función de probabilidad viene dada por

$$\begin{aligned} f_{BP}(x, y) &= P(X = x, Y = y) = \\ &= exp\left\{-(\lambda_1 + \lambda_2 + \lambda_3)\right\} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{k=0}^{min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k; \\ &x, y = 0, 1, ... \end{aligned} \tag{3.2}$$

Esta distribución bivariada permite la dependencia positiva entre las dos variables aleatorias. Marginalmente cada variable aleatoria sigue una distribución de Poisson con $E(X) = \lambda_1 + \lambda_3$ y $E(Y) = \lambda_2 + \lambda_3$. Además, $cov(X, Y) = \lambda_3$ y, por lo tanto λ_3

es una medida de dependencia entre las dos variables aleatorias, X e Y. Si $\lambda_3 = 0$ las dos variables son independientes y la distribución bivariada de Poisson se reduce al producto de dos distribuciones Poisson independientes (conocida también como distribución de Poisson doble). Algunas de las propiedades matemáticas anteriores se demuestran en el apéndice A.

Es posible adoptar esta distribución para modelizar dependencia en deportes de equipo. Si X *e* Y representan el marcador conseguido por cada uno de los equipos, una interpretación natural de los parámetros de un modelo bivariado de Poisson es que λ_1 y λ_2 reflejen la habilidad de marcar cada uno de los equipos y λ_3 refleje las condiciones del juego (por ejemplo, la velocidad del juego, el clima o las condiciones del estadio).

3.2.2. Modelo de Regresión Bivariada de Poisson

Se va a considerar el caso general de un modelo de regresión de Poisson Bivariado. Para la i-ésima observación, el modelo tiene la siguiente forma:

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$

$$log(\lambda_{1i}) = w_{1i}^{'}\beta_1,$$

$$log(\lambda_{2i}) = w_{2i}^{'}\beta_2,$$

$$log(\lambda_{3i}) = w_{3i}^{'}\beta_3,$$

donde $i = 1, ..., n$, denota el número de la observación, $w_{\kappa i}$ denota el vector de las variables explicativas para la observación i-ésima usada para el modelo $\lambda_{\kappa i}$ y β_κ denota el vector correspondiente de los coeficientes de regresión, $\kappa = 1, 2, 3$. De modo que ahora usaremos la notación (X, Y) para la variable respuesta bidimensional y W para las variables explicativas.

Cada parámetro de una distribución de Poisson Bivariada puede estar influido por diferentes variables. Las variables explicativas usadas para modelizar cada parámetro $\lambda_{\kappa i}$ no tienen por que ser las mismas. Normalmente, se consideran modelos con λ_3 constante (sin covariables sobre λ_3) porque esos modelos resultan más fáciles de interpretar.

Se utiliza el algoritmo EM para obtener la estimación por máxima verosimilitud. Este es un algoritmo potente para la estimación de máxima verosimilitud para datos con valores perdidos (de individuos o variables) o que se puede considerar que tengan valores perdidos. Tal y como indica Daniel Peña (2002) “este algoritmo tiene un interés general por sí mismo para resolver la estimación de valores ausentes en

cualquier problema multivariante" (p.312). La idea es aumentar los datos observados con algunos no observados para que la maximización de la verosimilitud sea más fácil. De este modo la estimación se obtiene iterando los dos pasos siguientes:

1. Paso E. Partiendo de un estimador inicial de los parámetros, se calcula la esperanza de las funciones de los valores ausentes que aparecen en la verosimilitud completa o aumentada, con respecto a la distribución de dichos valores ausentes dados los valores observados y las estimaciones iniciales. Esta operación se denomina el paso E (de tomar valores *esperados*) del algoritmo. Cuando la verosimilitud completa es una función lineal de los valores ausentes, este paso lleva a sustituir dichos valores por sus esperanzas condicionadas a los valores observados y los parámetros estimados.

2. Paso M. Consiste en *maximizar* la verosimilitud completa donde se han sustituido los valores faltantes por ciertas estimaciones de sus valores y así obtener un nuevo estimador de los parámetros.

Con el valor obtenido en el paso M se vuelve al paso E y se itera entre ellos hasta que la diferencia entre los estimadores sea suficientemente pequeña.

Para aplicar el algoritmo EM a la estimación de los parámetros de la regresión de Poisson Bivariada se hace una reducción trivariada de la distribución de Poisson Bivariada. Esto es, se supone que para la i-ésima observación X_{1i}, X_{2i}, X_{3i} representan los datos no observados, mientras que $X_i = X_{1i} + X_{3i}$ e $Y_i = X_{2i} + X_{3i}$ son los datos observados. Si los datos no observados están disponibles la estimación debería ser sencilla: únicamente sería necesario ajustar los modelos de regresión de Poisson sobre las variables X_1, X_2 y X_3. Por lo tanto, en orden a construir el EM-algoritmo es necesario estimar las funciones de los datos no observados por esperanzas condicionadas y ajustar modelos de regresión de Poisson a los pseudovalores obtenidos en el paso E. Se denota como ϕ el vector completo de parámetros, $\phi = (\beta_1', \beta_2', \beta_3')$, la ecuación de log-verosimilitud de los datos completos viene dada por:

$l(\phi) = -\sum_{i=1}^{n}\sum_{\kappa=1}^{3}\lambda_{\kappa i} + \sum_{i=1}^{n}\sum_{\kappa=1}^{3} x_{\kappa i} log(\lambda_{\kappa i}) - \sum_{i=1}^{n}\sum_{\kappa=1}^{3} log(x_{\kappa i}!)$,

donde las $\lambda_{\kappa i}$ vienen dadas por $log(\lambda_{\kappa i}) = w_{ki}'\beta_k$, para $\kappa = 1, 2, 3$. Observemos que$l(\phi)$ es una función lineal de $x_{\kappa i}$ (el último término no depende de ϕ) y por tanto bastará sustituir en la verosimilitud completa los datos no observados por sus esperanzas condicionadas.

Así, el algoritmo EM para el modelo de Poisson Bivariado viene dado por:

Paso E: Usando los valores de los parámetros actuales de la iteración k con notación dada por $\phi^{(k)}$, $\lambda_{1i}^{(k)}$, $\lambda_{2i}^{(k)}$y $\lambda_{3i}^{(k)}$, se calculan los valores esperados de X_{3i}, para $i = 1, ..., n$ mediante (ver Anexo A, Propiedad 3):

$$s_i = E(X_{3i} \mid X_i, Y_i, \phi^{(k)}) = \begin{cases} \lambda_{3i}^{(k)} \frac{f_{BP}(x_i-1, y_i-1 | \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)})}{f_{BP}(x_i, y_i | \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)})} & if \quad min(x_i, y_i) > 0 \\ 0 & if \quad min(x_i, y_i) = 0 \end{cases} \tag{3.3}$$

donde $f_{BP}(x, y \mid \lambda_1, \lambda_2, \lambda_3)$ viene dada por Ecuación 3.2.

Además, $E(X_{1i} \mid X_i, Y_i, \phi^{(k)}) = X_i - s_i$ y $E(X_{2i} \mid X_i, Y_i, \phi^{(k)}) = Y_i - s_i$.

Paso M: Se actualizan las estimaciones por

$$\beta_1^{(k+1)} = \hat{\beta}(x - s, W_1),$$
$$\beta_2^{(k+1)} = \hat{\beta}(y - s, W_2),$$
$$\beta_3^{(k+1)} = \hat{\beta}(s, W_3),$$
$$\lambda_{\kappa i}^{(k+1)} = exp(W_{\kappa i}^{'} \hat{\beta}_{\kappa}^{(k+1)})$$

para $\kappa = 1, 2, 3$, donde $s = (s_1, ..., s_n)^{'}$ es el vector calculado en el paso E, $\hat{\beta}(x, W)$ es el estimador de máxima verosimilitud de un modelo de Poisson con respuesta el vector x y matriz de datos dada por W. Cada matriz W_κ es una matriz $n \times p_\kappa$ y $W_{\kappa i}^{'}$ es su correspondiente fila $\boldsymbol{i}$ (para $i = 1, ..., n$).

Si se desea tener parámetros comunes (o iguales) entre diferentes λ_κ se debería construir una matriz de diseño común W y su correspondiente vector de parámetros β se estimaría como $\beta^{(k+1)} = \hat{\beta}(u, W)$, con $u^{'} = (x^{'} - s^{'}, y^{'} - s^{'}, s^{'})$.

3.3. Modelo de Regresión Bivariante de Poisson con Inflado en la Diagonal

En los modelos unidimensionales, los modelos inflados se pueden construir aumentando las probabilidades de ciertos valores de la variable X en consideración. En el caso de la distribución de Poisson son bastante comunes los modelos de inflado en el cero (ver referencias citadas al final de la discusión del capítulo 2). En el caso bivariante, Dixon y Coles (1997) proponen algunos modelos de inflado para la modelización de los partidos de fútbol. Posteriormente, Karlis y Ntzoufras (2003) plantean una modificación de su modelo de regresión bivariante de Poisson para inflar las probabilidades en la diagonal (valores de X e Y iguales). Estos autores indican que los modelos de inflado en la diagonal son adecuados para los resultados de campeonatos con un exceso de empates ($X = Y$), los cuales no pueden ser ajustados por modelos de doble Poisson o incluso por modelos de Poisson bivariados. A continuación se presentan estos últimos modelos.

3.3.1. Distribución Bivariante de Poisson con Inflado en la Diagonal

Partiendo de la distribución de Poisson bivariada, el modelo con inflado en la diagonal, $IBP(\lambda_1, \lambda_2, \lambda_3)$, viene especificado por

$$f_{IBP}(x, y) = \begin{cases} (1-p) f_{BP}(x, y \mid \lambda_1, \lambda_2, \lambda_3), & x \neq y \\ (1-p) f_{BP}(x, y \mid \lambda_1, \lambda_2, \lambda_3) + p f_D(x \mid \theta), & x = y \end{cases} \tag{3.4}$$

donde $f_{BP}(x, y \mid \lambda_1, \lambda_2, \lambda_3)$ está definido en la Ecuación 3.2 y $f_D(x \mid \theta)$ es una función de probabilidad de una distribución discreta $D(x, \theta)$ definida en el conjunto $\{0, 1, 2, ...\}$ con el vector de parámetros θ. Obsérvese que para $p = 0$ se tiene el modelo de Poisson bivariado definido anteriormente en (1). Los modelos de inflado en la diagonal pueden ser ajustados usando el algoritmo EM como se verá posteriormente.

Se puede elegir para $D(x, \theta)$ cualquier distribución discreta. Algunos casos habituales son la distribución de Poisson, una distribución geométrica o distribuciones discretas simples que se denotan por *Discreta* (J). Para la *Discreta* (J) se considera la distribución con la función de probabilidad mostrada a continuación:

$$f(x \mid \theta, J) = \begin{cases} \theta_x & para \quad x = 0, 1, ..., J \\ 0 & para \quad x \neq 0, 1, ..., J \end{cases} \tag{3.5}$$

donde $\sum_{x=0}^{J} \theta_x = 1$. Si $J = 0$ entonces tenemos el caso particular de un modelo inflado en el cero.

Las propiedades más importantes de estos modelos son las siguientes:

1. Las distribuciones marginales de un modelo inflado en la diagonal no son distribuciones de Poisson sino mezclas de dos distribuciones una de las cuales sí es de Poisson:

$$f_{IBP}(x) = (1 - p) f_{P_0}(x \mid \lambda_1 + \lambda_3) + p f_D(x \mid \theta)$$

donde $f_{P_0}(x \mid \lambda)$ es la función de probabilidad de la distribución de Poisson con parámetro λ.

2. Incluso si $\lambda_3 = 0$ (distribución de Poisson Doble), la distribución inflada resultante introduce un grado de dependencia entre las dos variables consideradas.

3.3.2. Modelo de Regresión Bivariante con Inflado en la Diagonal

Una muestra de este modelo de regresión toma la siguiente forma:

$$(X_i, Y_i) \sim IBP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$

$$log(\lambda_{1i}) = w_{1i}^{'} \beta_1,$$
$$log(\lambda_{2i}) = w_{2i}^{'} \beta_2,$$
$$log(\lambda_{3i}) = w_{3i}^{'} \beta_3,$$

donde $i = 1, ..., n$, denota el número de la observación, $w_{\kappa i}$ denota el vector de las variables explicativas para la observación i-ésima usada para el modelo $\lambda_{\kappa i}$ y β_κ denota el vector correspondiente de los coeficientes de regresión, $\kappa = 1, 2, 3$.

Nótese que en $IBP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$ están implícitamente considerados p y $f_D(x \mid \theta)$ (ver Ecuación 3.4).

Para el algoritmo EM de los modelos inflados, se introduce a mayores un indicador binario latente V_i que es igual a 1 cuando se produce un inflado en la diagonal e igual a 0 en otro caso. Ahora la ecuación de log-verosimilitud viene dada por:

$$\begin{aligned} l(\phi) = & \sum_{i=1}^{n} v_i \left\{ log(p) + log(f_D(x_i \mid \theta)) \right\} \\ & + \sum_{i=1}^{n} (1 - v_i) \left\{ log(1-p) - \sum_{i=1}^{n} \sum_{\kappa=1}^{3} \lambda_{\kappa i} + \sum_{i=1}^{n} \sum_{\kappa=1}^{3} x_{\kappa i} log(\lambda_{\kappa i}) - \sum_{i=1}^{n} \sum_{\kappa=1}^{3} log(x_{\kappa i}!) \right\} \end{aligned}$$

y el algoritmo EM para el modelo inflado en la diagonal tiene los siguientes pasos:

Paso E:

(a) Usando los valores para el parámetro actual de κ iteraciones estimadas por $\phi^{(\kappa)}, \lambda_{1i,}^{(\kappa)}, \lambda_{2i}^{(\kappa)}, \lambda_{3i}^{(\kappa)}, p^{(\kappa)}$ y $\theta^{(k)}$, para $i = 1, ..., n$ se calcula:

$$\begin{aligned} v_i &= E(V_i \mid X = x_i, Y = y_i, \phi^{(\kappa)}, p^{(\kappa)}, \theta^{(\kappa)}) = \\ &= \begin{cases} \frac{p^{(\kappa)} f_D(x_i \mid \theta^{(\kappa)})}{p^{(\kappa)} f_D(x_i \mid \theta^{(\kappa)}) + (1-p^{(\kappa)}) f_{BP}(x_i, y_i \mid \lambda_{1i}^{(\kappa)}, \lambda_{2i}^{(\kappa)}, \lambda_{3i}^{(\kappa)})}, & si \quad x_i = y_i \\ 0 & si \quad x_i \neq y_i \end{cases} \end{aligned}$$

donde $f_D(x \mid \theta)$ es la función de probabilidad de la distribución inflada con un vector de parámetros θ evaluado en el valor de x.

(b) Para $i = 1, ..., n$, calculamos s_i usando Ecuación 3.3.

Paso M: Se actualizan los parámetros para

$$p^{(\kappa+1)} = \frac{1}{n} \sum_{i=1}^{n} v_i$$

$$\begin{aligned} \beta_1^{(\kappa+1)} &= \hat{\beta}_{\tilde{v}}(x - s, W_1), \\ \beta_2^{(\kappa+1)} &= \hat{\beta}_{\tilde{v}}(y - s, W_2), \\ \beta_3^{(\kappa+1)} &= \hat{\beta}_{\tilde{v}}(s, W_3) \\ \theta^{(\kappa+1)} &= \hat{\theta}_{v,D} \\ \lambda_{\kappa i}^{(\kappa+1)} &= exp(W_{\kappa i}' \beta_{\kappa}^{(\kappa+1)}) \quad para \quad \kappa = 1, 2, 3; \end{aligned}$$

donde x, y, s, v y $\hat{v}$ son $n \times 1$ vectores con elementos x_i, y_i, s_i, v_i y $\hat{v} = 1 - v_i$ para $i = 1, ..., n$, $\hat{\beta}_v(y, W)$ es la estimación de verosimilitud ponderada para β de un modelo de regresión de Poisson con respuesta y, matriz de datos W y vector de peso v, y $\theta_{v,D}$ la estimación de máxima verosimilitud de θ para la distribución $D(x; \theta)$ y

los pesos dados por el vector v. El diseño de las matrices $W_\kappa, \kappa = 1, 2, 3$ se definen como anteriormente.

Respecto a la estimación de la regresión de Poisson ponderada, presenta una log-verosimilitud similar a la dada en la Ecuación 3.1 pero ponderada por los pesos v que correspondan, esto es

$$l(\beta_v) = \sum_{i=1}^{n} v_i \left(y_i x_i' \beta - e^{x_i' \beta} \right).$$

En R, estos estimadores se pueden obtener utilizando el argumento *weights* dentro de la función *glm*.

Para la elección especifica de la distribución inflada se obtienen los siguientes estimadores:

- Distribución geométrica: Para la distribución geométrica con función de probabilidad $f(x \mid \theta) = (1-\theta)^x \theta, 0 \leq \theta \leq 1,\ x = 0, 1, ...; \theta$ se actualiza por
  $$\theta^{(\kappa+1)} = \frac{\sum_{i=1}^{n} v_i}{\sum_{i=1}^{n} v_i x_i + \sum_{i=1}^{n} v_i}$$
- Distribución de Poisson: Para una distribución de Poisson con función de probabilidad $f(x \mid \theta) = e^{-\theta}\theta^x / x!, \theta \geq 0, x = 0, 1, ...; \theta$ se actualiza por $\theta^{(\kappa+1)} = (\sum_{i=1}^{n} v_i)^{-1} \sum_{i=1}^{n} v_i x_i$
- Distribución Discreta: Para cualquier distribución discreta, *Discreta (J)*, con función de probabilidad dada por la Ecuación 3.5, los parámetros del modelo vienen dados por $\theta_j = (\sum_{i=1}^{n} v_i)^{-1} \sum_{i=1}^{n} I(X_i = Y_i = j) v_i$ para $j = 1, ..., J$ y $\theta_0 = 1 - \sum_{j=1}^{J} \theta_j$; donde $I(x)$ es el indicador de la función tomando como valor igual a 1 si x es verdad y cero en cualquier otro caso.
- Modelo inflado en cero: El modelo inflado en cero es un caso especial de la *Discreta (J)* con $J = 0$ y $\theta_0 = 1$, el cual es un resultado de inflar la casilla (0,0). A partir de aquí, no es necesario estimar parámetros adicionales excepto p , el cual es la proporción del componente de inflado.

4. Análisis de datos y resultados

En este capítulo se van a aplicar los modelos expuestos en el capítulo anterior al contexto deportivo y, tal como se mencionó anteriormente, se van a analizar los resultados de los partidos de fútbol correspondientes a dos temporadas de la Primera División de la Liga Nacional de Fútbol Española. En el conjunto de datos a analizar se han incluido cuatro variables: goles marcados por el equipo que juega en casa, goles marcados por el equipo visitante, equipo local y equipo visitante. Las dos primeras variables serían las variables respuesta, mientras que las dos segundas son las variables explicativas (de tipo cualitativo). Antes de mostrar los resultados de la aplicación práctica (apartados 2 y 3), se presenta un apartado introductorio que plantea los modelos anteriores en un contexto futbolístico, presenta el paquete específico de R para analizar esos datos y revisa la codificación de las variables explicativas nominales.

4.1. Regresión Bivariante de Poisson y modelos futbolísticos

Diferentes autores (Maher (1982), Lee (1997) y Rue y Salvensen (2000)) asumen que el número de goles marcados por cada equipo sigue una distribución de Poisson con la siguiente forma general:

$$
\begin{aligned}
&X_i \sim Poisson(\lambda_{1i}),\\
&Y_i \sim Poisson(\lambda_{2i}),\\
&log(\lambda_{1i}) = \mu + home + att_{h_i} + def_{g_i} + home.att_{h_i} + home.def_{g_i} +\\
&+att.def_{h_i g_i} + home.att.def_{h_i g_i}\\
&log(\lambda_{2i}) = \mu + att_{g_i} + def_{h_i} + att.def_{g_i h_i}
\end{aligned}
$$

para $i = 1, 2..n$, donde n es el número de partidos o observaciones, i es un indicador de partido (observación), h_i y g_i indican el equipo que juega en casa y el equipo que juega fuera para el partido i, X_i e Y_i son los goles marcados por el equipo local (h_i) y el equipo visitante (g_i) en cada partido i, λ_{1i} y λ_{2i} son el número esperado de goles correspondientes, μ es un parámetro constante, $home$ es el parámetro del efecto "jugar en casa" y finalmente, att_k y def_k engloban el rendimiento ofensivo (o de ataque) y el defensivo del equipo k. Karlis y Ntzoufras (2003) adoptan una

estructura más simple para los parámetros implicados en los predictores lineales de λ_1 y λ_2. Por lo tanto, para cada juego $i(i = 1, ..., n)$,

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$
$$log(\lambda_{1i}) = \mu + home + att_{h_i} + def_{g_i},$$
$$log(\lambda_{2i}) = \mu + att_{g_i} + def_{h_i}$$

Para lograr la identificabilidad de los parámetros del modelo, se deben usar un conjunto de restricciones estándar. Se propone usar bien la suma a cero o restricciones de ángulo, dependiendo de la interpretación que se prefiera. Los autores eligen la restriccion de la suma a cero para una interpretación más fácil. Por lo tanto, el parámetro constante general especifíca λ_1 y λ_2 cuando los equipos tiene la misma fuerza de juego en un campo neutral. Los parámetros ofensivos y defensivos se expresan como las salidas de un equipo con una habilidad ofensiva o defensiva media.

Para los parámetros de covarianza λ_{3i} se considera el siguiente predictor lineal:

$$log(\lambda_{3i}) = \beta^{con} + \gamma_1 \beta^{home}_{h_i} + \gamma_1 \beta^{away}_{g_i}$$

donde β^{con} es un parámetro constante y $\beta^{home}_{h_i}$ y $\beta^{away}_{g_i}$ son parámetros que dependen de si el equipo es local o visitante respectivamente. Los parámetros γ_1 y γ_2 son indicadores binarios dummy que toman los valores 0 o 1, dependiendo del modelo que se considere. Por lo tanto, cuando $\gamma_1 = \gamma_2 = 0$ se considera que la covarianza es constante, cuando $(\gamma_1, \gamma_2) = (1, 0)$ se asume que la covarianza depende del equipo local solamente y así sucesivamente.

El parámetro λ_3 se puede interpretar como el efecto del azar el cual actúa de forma aditiva sobre la media marginal y refleja las condiciones del juego. Una estructura alternativa de la matriz de diseño puede ser fácilmente implementada si se dispone de informacion adicional o si se asume que las habilidades de ataque son diferentes en los partidos de casa o de fuera, o si el efecto jugar en casa varía de equipo a equipo.

Los modelos de inflado en la diagonal son adecuados para campeonatos con un exceso de empates los cuales no pueden ser capturados por modelos dobles de Poisson, o incluso modelos bivariados de Poisson.

4.1.1. Paquete estadístico bivpois

Los modelos comentados en el apartado anterior y en el capítulo 3 han sido implementados en R, en el paquete denominado Modelos de Poisson Bivariantes usando el algoritmo EM (paquete “bivpois”). La versión del mismo es la 0.50-3 con fecha del 2 de febrero de 2007. Los autores son Karlis Dimitris y Ioannis Ntzoufras. El paquete bivpois puede obtenerse directamente de la siguiente dirección:

http://stat-athens.aueb.gr/~jbn/papers/paper14.htm

Aunque el paquete no se encuentra en el repositorio de R se puede conseguir el código para aplicar las funciones lm.bp, lm.dibp y pbivpois que serán utilizadas en este trabajo y que se presentan en el Anexo C. Estas funciones sirven para ajustar Modelos de Regresión de Poisson Bivariantes y con Inflado en la Diagonal usando el algoritmo EM. Las funciones lm.bp y lm.dibp utilizan algunas funciones internas, como por ejemplo, newnamesbeta y splitbeta (Karlis & Ntzoufras, 2004).

Las funciones más importantes son lm.bp, lm.dipb y pbivpois. La función lm.bp hace referencia al modelo de regresion de Poisson bivariante general. La siguiente funcion, lm.dipb está relacionada con el modelo de regresión de Poisson bivariante inflado en la diagonal. Finalmente, la función pbivpois guarda relación con la función de probabilidad de la distribución de Poisson bivariada. Estas funciones se describen en el anexo C con detalle. A continuación y a modo de ejemplo, mostramos los argumentos más importantes de la función lm.bp:

lm.bp(l1, l2, l1l2=NULL, l3=~1, data, common.intercept=FALSE, zeroL3=FALSE, maxit=300, pres=1e-8, verbose=getOption("verbose"))

l1 Fórmula de la forma "$x \sim X_1 + ... + X_p$" para los parámetros del $log\lambda_1$.

12 Fórmula de la forma "$y \sim X_1 + ... + X_p$" para los parámetros del $log\lambda_2$.

l1l2 Fórmula de la forma "$\sim X_1 + ... + X_p$" para los parámetros comunes del $log\lambda_1$ y $log\lambda_2$. Si la variable explicativa se encuentra también en *l1* y/o *l2* entonces el modelo ajusta la interacción entre los parámetros. Se pueden usar aquí términos especiales de la forma "$c(X_1, X_2)$". Estos términos implican parámetros comunes de λ_1 y λ_2 para distintas variables.

l3 Fórmula de la forma "$\sim X_1 + ... + X_p$" para los parámetros del $log\lambda_3$.

data Data frame que contiene las variables en el modelo

common.intercept Función lógica que específica si se debe usar un intercepto común sobre λ_1 y λ_2. Por defecto su valor es *FALSE*.

zeroL3 Argumento lógico que controla si λ_3 debería ser igual a cero (y por lo tanto se ajustaría el modelo de Poisson doble)

4.1.2. Esquemas de codificación de las variables cualitativas

En la estimación de un modelo de regresión se pueden diferenciar dos tipos de variables explicativas: cualitativas (también llamadas categóricas o factores) y numéricas. Las variables cualitativas hacen referencia a una característica o atributo que puede tomar un número entero de niveles o estados (Tusell, 2008). Por ejemplo, la variable equipo puede tomar los niveles: "Almeria", "Barcelona", "Betis", etc.

En R si no se especifica lo contrario, el orden de los niveles se determina por el orden alfabético de sus denominaciones. Asimismo se puede revertir el orden de los niveles mediante la funcion *rev*, sin necesidad de enumerarlos. Por otra parte, a veces, se

desea poner en primer lugar uno de los niveles, el nivel de referencia y en R, esta manipulación se realiza con la función *relevel*.

Las fórmulas en R permiten especificar de forma simple modelos de regresion, nombrando a la izquierda del símbolo ∼la variable respuesta, y a la derecha las variables explicativas. De esta forma se puede estimar un modelo de regresión lineal (función lm), regresión lineal generalizada (función glm) o regresión no lineal (función nlme). A veces, es necesario plantear modelos en los que alguna variable independiente es cualitativa como ocurre en nuestro estudio. Si ésta es una variable dicotómica, no existen dificultades en asignar un número a cada una de ellas e introducir esta variable en el modelo. Un ejemplo sería designar 0 para un grupo y 1 para otro, y el coeficiente β_1 es la diferencia entre las medias de la variable Y entre ambos grupos. Si la variable tiene más de dos categorías la solución es diferente.

Por ejemplo, si se consideran tres categorias de lesiones: sin lesión, lesión leve y lesión grave, si se define una variable $X = 0$ para no lesión, $X = 1$ para lesión leve y $X = 2$ para lesión grave y se introduce en un modelo:

$$\mu_{Y|...,X,...} = \beta_0 + \beta X + ...$$

El coeficiente β sería lo que cambia μ_Y por un aumento de la unidad de la variable X. Se está asumiendo que la media de la variable dependiente Y cambia lo mismpo por pasar X de 0 a 1 (no lesión a lesión leve) que de 1 a 2 (lesión leve a lesión grave) y, a la vez que el cambio en la media de Y por pasar de X de 0 a 2 es el doble que el provocado por pasar de 0 a 1, ó de 1 a 2. El coeficiente depende de la codificacion de la variable β sería distinto si se asigna $X = 0$ para lesión leve, $X = 1$ para sin lesión y $X = 2$ para lesión grave.

Se pueden introducir las variables cualitativas en un modelo de regresión creando tantas variables dicotómicas como categorías. En nuestro ejemplo se crearían tres variables, X_1 para lesión leve, X_2 para no lesión y X_3 para lesión grave, todas con dos posibles valores 0:no y 1:sí. El modelo sería el siguiente:

$$\mu_{Y|X_1,X_2,X_3,...} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ... + ...$$

β_1, β_2,β_3, serían, respectivamente, lo que cambia μ_Y por tener lesión leve, no lesión o lesión grave. La dificultad está en interpretar β_0, que sería μ_Y cuando las tres variables son 0, que con la codificación utilizada, no sería posible. Estas tres variables cumplen:

$$\sum_{i=1}^{3} X_i = 1 \Rightarrow X_3 = 1 - \sum_{i=1}^{2} X_i$$

es decir, X_3es la combinación lineal de X_1 y X_2 y esto hace que el modelo sea irresoluble.

La solución a este problema es crear tantas variables dicotómicas como categorías menos 1 en la variable original, denominadas también variables indicadoras o "dummy". En nuestro ejemplo se crearían dos variables X_1, X_2 con los siguientes valores:

	X_1	X_2
Sin lesión	0	0
Lesión leve	1	0
Lesión grave	0	1

Las variables X_1 y X_2 ya no son combinación lineal y el modelo es resoluble quedando de la siguiente forma:

$$\mu_{Y|X_1,X_2,...} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

y β_0 es μ_Y cuando X_1 y X_2 son ambas cero, por lo tanto para los no lesionados, $\beta_0 - \beta_1$ es μ_Y cuando X_1 es 1 y X_2 es 0; es decir lesión leve, por lo tanto, β_1 es lo que cambia μ_Y entre lesión leve y no lesión e igualmente β_2 es lo que cambia μ_Y entre lesión grave y no lesión. Utilizando esta codificación los coeficientes tienen una clara interpretación cuando una de las categorías (no lesión) se quiere usar como referencia de las demás. A dicha categoría se le asigna el valor 0 para todas las variables indicadoras.

Para variables en las que no hay una categoría natural para usarla como referencia es más útil otro esquema de codificación. Supóngase la variable lugar de juego con cuatro estadios: A, B, C y D. Se crean tres variables indicadoras (siempre una menos que categorías) con el siguiente esquema:

	X_1	X_2	X_3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

El modelo sería el siguiente:

$$\mu_{Y|X_1,X_2,X_3,...} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ... + ...$$

y por lo tanto:

$\mu_Y = \beta_0 + \beta_1 = \mu_{Y|A}$ para el estadio A

$\mu_Y = \beta_0 + \beta_2 = \mu_{Y|B}$ para el estadio B

$\mu_Y = \beta_0 + \beta_3 = \mu_{Y|C}$ para el estadio C

$\mu_Y = \beta_0 - \beta_1 - \beta_2 - \beta_3 = \mu_{Y|D}$ para el estadio D.

Si se suman las 4 ecuaciones y se divide por 4:

$$\beta_0 = \frac{\mu_{Y|A} + \mu_{Y|B} + \mu_{Y|C} + \mu_{Y|D}}{4},$$

por lo tanto β_0 es la media de Y en los cuatro lugares de juego, β_1 la diferencia del estadio A con respecto a la media, β_2 la diferencia del estadio B con respecto a la

media, β_3 la diferencia del estadio C con respecto a la media y $-\beta_1 - \beta_2 - \beta_3$ la diferencia del estadio A con respecto a la media. A diferencia del esquema utilizado anteriormente, se usa como nivel de referencia la media en todas las categorías en lugar de una de ellas (Abraira y Pérez, 1996).

En R, la forma de codificación por defecto es la denominada *contr.treatment.* Es la descrita en primer lugar: al primer nivel de la variable cualitativa se la asigna el valor 0 y los otros niveles de la variable de medida cambian desde ese primer nivel. Se puede modificar esta codificación usando la función *options( )*:

options(contrast=c("contr.sum","contr.poly").

De esta forma, la suma de los parámetros es nula y la función *contr.sum ( )* proporciona coeficientes donde se compara cada nivel con la media de todos. La función *contr.poly* proporciona también contrastes pero basados en polinomios ortogonales y se utiliza para variables con nivel de medida ordinal.

Para ver la codificación de las variables "dummy" de un factor en un archivo datos se usa *contrasts(datos[,"factor"]).*

4.2. Análisis para la Temporada 2010/2011

En este apartado se van a ajustar modelos del tipo

$$\begin{aligned} &(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \; o \; IBP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \\ &log(\lambda_{1i}) = \mu + home + att_{h_i} + def_{g_i}, \\ &log(\lambda_{2i}) = \mu + att_{g_i} + def_{h_i} \\ &log(\lambda_{3i}) = \beta^{con} + \gamma_1 \beta^{home}_{h_i} + \gamma_1 \beta^{away}_{g_i} \end{aligned} \tag{4.1}$$

(ver apartado anterior) para analizar los datos de los partidos de fútbol de la temporada 2010/2011 recopilados en el fichero de datos: "2010.txt".

> datos1 <- read.table("2010.txt",header=TRUE)

Cuatro variables se incluyen en ese conjunto de datos:

> names(datos1)

[1] "g1" "g2" "team1" "team2"

g1 y *g2* se corresponde con los goles marcados por el equipo local y visitante (variables respuesta $X = g1, Y = g2$) mientras que *team 1* y *team 2* hacen referencia al equipo local y el visitante, respectivamente, y se usarán para definir las variables explicativas (capacidad de ataque y de defensa: *att* y *def*). Una muestra de los datos se expone a continuación:

> datos1

	g1	g2	team1	team2
1	1	1	AtMadrid	Almeria
2	1	0	Athletic	Almeria
3	3	1	Barcelona	Almeria
.....				
379	1	1	Valencia	Zaragoza
380	1	0	Villarreal	Zaragoza

En total los 20 equipos han disputado 380 partidos. Como se puede observar el primer partido codificado en la matriz de datos es el Atlético de Madrid-Almería, ambos equipos empataron en ese partido. La siguiente fila se corresponde con el partido jugado en casa por el Athletic de Bilbao, el cuál ganó al Almería por un tanto. Todos los partidos se codifican de la misma forma. En nuestra base de datos la última fila se corresponde con el partido Villarreal-Zaragoza, en el que el Zaragoza perdió por un tanto fuera de casa.

Los niveles de los equipos nos proporcionan los 20 equipos en 1ª división durante esa temporada:

> levels(datos1[,3])

[1] "Almeria" "Athletic" "AtMadrid" "Barcelona" "Deportivo"

[6] "Espanyol" "Getafe" "Gijon" "Hercules" "Levante"

[11] "Malaga" "Mallorca" "Osasuna" "Racing" "RMadrid"

[16] "RSociedad" "Sevilla" "Valencia" "Villarreal" "Zaragoza"

Team1 indica los equipos que atacan para la variable *g1* mientras que *team2* indica los equipos en ataque para la variable *g2*. De forma similar *team1* y *team2* indican los equipos que defienden para las variables *g2* y *g1* respectivamente. Siguiendo el enfoque de Karlis y Ntzoufras (2003) (ver Ecuación 4.1), vamos a suponer que el efecto ataque para un equipo es independiente de la localización y del equipo rival, y lo mismo supondremos con el efecto defensivo. Para modelizar los efectos comunes de ataque y defensa se debe usar un término del tipo *c(team1, team2)* para estimar el parámetro de ataque común y un término del tipo *c(team2,team1)* para estimar el parámetro común de defensa. Por lo tanto, la fórmula para modelizar lambda 1 y lamba2 en la función *lm.bp* del paquete bivpois de R es la siguiente:

form1<-c(team1,team2)+c(team2,team1)

La codificación de las variables explicativas utiliza el siguiente código:

options(contrast=c("contr.sum","contr.poly"))

De esta forma, como se explicó en el apartado anterior, la suma de los coeficientes es 0, el coeficiente del Zaragoza se obtiene como el opuesto de la suma de todos los demás coeficientes y la interepretación de cada coeficiente es la desviación de la capacidad ofensiva (defensiva) de cada equipo respecto a las de un equipo con una habilidad ofensiva (defensiva) media.

Utilizando las funciones lm.bp y lm.dibp se han ajustado 12 modelos diferentes del tipo Ecuación 4.1: un primer modelo asumiendo la Poisson doble (respuesta BP y $\lambda_3 = 0$), cuatro modelos suponiendo que la respuesta es una Poisson Bivariante pero con distintas modelizaciones de λ_3 (constante, dependiente del equipo local, del visitante y de ambos) y siete modelos con respuesta de Poisson Bivariante con inflado en la diagonal y distintas distribuciones para el inflado (Discreta(0), Geométrica, Discretas(1),(2) y (3), Poisson y Poisson con $\lambda_3 = 0$).

El código detallado para el ajuste de estos modelos se puede ver en el Anexo D. Un ejemplo del mismo es:

ex4.m1<-lm.bp(g1~1, g2~1, l1l2=form1, zeroL3=TRUE, data=datos1)

ex4.m2<-lm.bp(g1~1,g2~1, l1l2=form1, data=datos1)

ex4.m6 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, jmax=0)

ex4.m7 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, distribution="geometric")

Las funciones anteriores proporcionan objetos entre cuyas componentes destacan: coefficients, fitted.values, residuals, loglikelihood, parameters, AIC y BIC.

El criterio de informacion de Akaike (AIC) y el Criterio de Informacion de Bayes (BIC) se utilizan para escoger el modelo cuya media global se ajusta mejor y compense el exceso de parámetros. Estos criterios se definen a continuacion:

$AIC = -2log(verosimilitud) + 2p$

$BIC = -2log(verosimilitud) + plog(n)$

El mejor modelo será áquel con el AIC o BIC más pequeño.

Los resultados del ajuste de los 12 modelos considerados se presentan en la Tabla 4.1. El primer modelo (Poisson doble) es el que tiene el mejor AIC, seguido de los modelos: Poisson bivariante constante, y el modelo inflado en la diagonal con Discreta(0). Como se puede ver, los modelos más sencillos son los que obtienen un mejor ajuste y el considerar modelos más complejos con distintas distribuciones de inflado en la diagonal no aportan una mejora sustancial. Creemos que ello es debido a que no existe un número elevado de empates en esta temporada (78 empates en 380 partidos), hecho que puede estar influido por la actual forma de puntuación asignada al resultado de cada partido: 0 perder, 1 empatar y 3 ganar.

Modelo de distribución	**Detalles adicionales del modelo**	**Log-Likelihood**	**Número de parámet-ros**	**AIC**	**BIC**
1, Poisson doble		-1070.659	40	*2221.318**	*2406.651*
Covariables en λ_3					
2, Poisson bivariante	Parámetro asociado a λ_3 Constante ($\gamma_1 = \gamma_2 = 0$)	-1071.559	41	2225.118	2415.084
3, Poisson bivariante	Equipo local ($\gamma_1 = 1, \gamma_2 = 0$)	-1069.279	60	2258.557	2536.556
4, Poisson bivariante	Equipo visitante ($\gamma_1 = 0, \gamma_2 = 1$)	-1068.691	60	2257.382	2535.381
5, Poisson bivariante	Ambos (local y visitante) ($\gamma_1 = 1, \gamma_2 = 1$)	-1065.841	79	2289.682	2655.714
6, Poisson bivariante con inflado diagonal	Discreta(0) =Constante	-1074.905	42	2233.811	2428.410
Distribución diagonal					
7, Poisson bivariante con inflado diagonal	Geométrica	-1077.429	43	2240.859	2440.092
8, Poisson bivariante con inflado diagonal	Discreta (1)	-1076.801	43	2239.601	2438.834
9, Poisson bivariante con inflado diagonal	Discreta (2)	-1076.960	44	2241.921	2445.787
10, Poisson bivariante con inflado diagonal	Discreta (3)	-1076.393	45	2242.786	2451.285
11, Poisson bivariante con inflado diagonal	Poisson	-1077.404	43	2240.808	2440.041
12, Poisson bivariante con inflado diagonal	Poisson y λ_3=0 (Poisson doble)	-1076.021	42	2236.042	2430.642

Tabla 4.1. Ajuste de los 12 modelos para la Temporada 2010/2011

A continuación se muestran los parámetros estimados para los mejores modelos: 1, 2 y 6 (véase Tabla 4.2).

	Modelo 1:DblPois		**Modelo 2: Bivpois**		**M6: DIBP**	
	Ataque	*Defensa*	*Ataque*	*Defensa*	*Ataque*	*Defensa*
Barcelona	0.62	-0.84	0.65	-0.89	0.64	-0.90
Real Madrid	0.70	-0.38	0.73	-0.42	0.73	-0.41
Valencia	0.25	-0.13	0.26	-0.14	0.25	-0.15
Villarreal	0.08	-0.14	0.08	-0.15	0.07	-0.16
Sevilla	0.23	0.20	0.24	0.20	0.26	0.20
Athletic	0.18	0.09	0.19	0.10	0.16	0.07
Atlético de Madrid	0.22	0.06	0.24	0.07	0.23	0.05
Espanyol	-0.01	0.17	0.00	0.18	-0.01	0.18
Osasuna	-0.11	-0.10	-0.11	-0.11	-0.09	-0.11
S. Gijón	0.00	0.26	0.01	0.28	-0.02	0.26
Málaga	0.10	0.30	0.11	0.32	0.11	0.33
Rácing	-0.36	-0.20	-0.38	-0.21	-0.38	-0.21
Zaragoza	-0.22	0.03	-0.24	0.03	-0.24	0.03
Levante	-0.19	0.02	-0.22	0.01	-0.22	0.02
Real Sociedad	-0.19	0.09	-0.20	0.09	-0.19	0.10
Getafe	-0.07	0.11	-0.06	0.13	-0.07	0.12
Mallorca	-0.19	0.09	-0.19	0.10	-0.19	0.11
Deportivo	-0.48	-0.10	-0.51	-0.10	-0.47	-0.03
Hércules	-0.31	0.15	-0.32	0.17	-0.31	0.18
Almería	-0.25	0.31	-0.27	0.33	-0.28	0.32
Intercepto para λ_1	0.41		0.38		0.38	
Intercepto para λ_2	0.03		-0.03		-0.03	
Intercepto para λ_3			-2.65		-2.65	
λ_3	0		0.07		0.07	
Efecto de Jugar en casa	0.38		0.41		0.41	
Proporción (p)					0.01	
$\theta_1 = P(X = Y = 1)$					0	

Tabla 4.2. Parámetros estimados para los mejores modelos (2010/2011)

En la Tabla 4.2 los equipos están colocados según su puesto en la clasificación final una vez finalizado el campeonato. Los parámetros de ataque y defensa obtenidos son acordes con la clasificación de esa temporada. El F.C. Barcelona resultó ser el mejor equipo de esa temporada, proclamándose campeón con 96 puntos. Su parámetro de ataque es el segundo más alto, solamente por debajo del Real Madrid y su parámetro de defensa es muy superior al resto de equipos. El siguiente equipo clasificado fue el Real Madrid el cual tiene el mejor parámetro de ataque y una defensa superior a todos los equipos excepto el Barcelona, que lo supera en este aspecto de forma

relevante. En cuanto a los peores equipos clasificados destaca el U.D. Almería con el peor parámetro en defensa. Aunque su parámetro en ataque no es de los peores, no consigue contrarrestar su deficiencia en defensa y este equipo descendió a Segunda División después de haber conseguido 30 puntos que no le permitieron seguir en la máxima categoría.

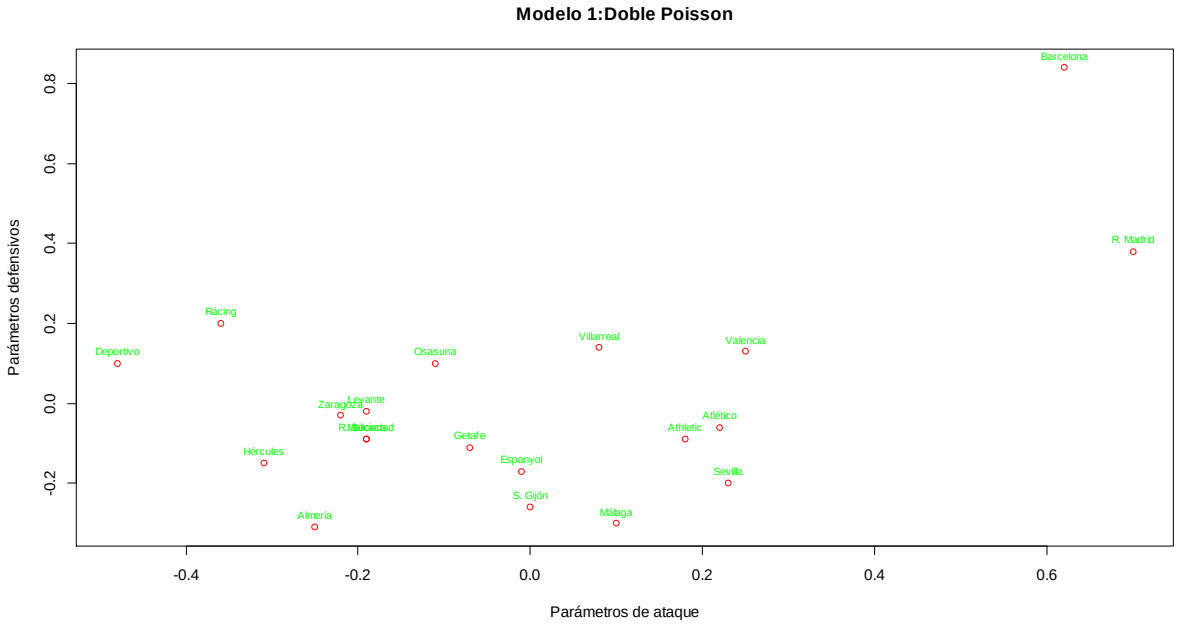


Figura 4.1.: Parámetros defensa y ataque temporada 2010/2011 (DblPois)

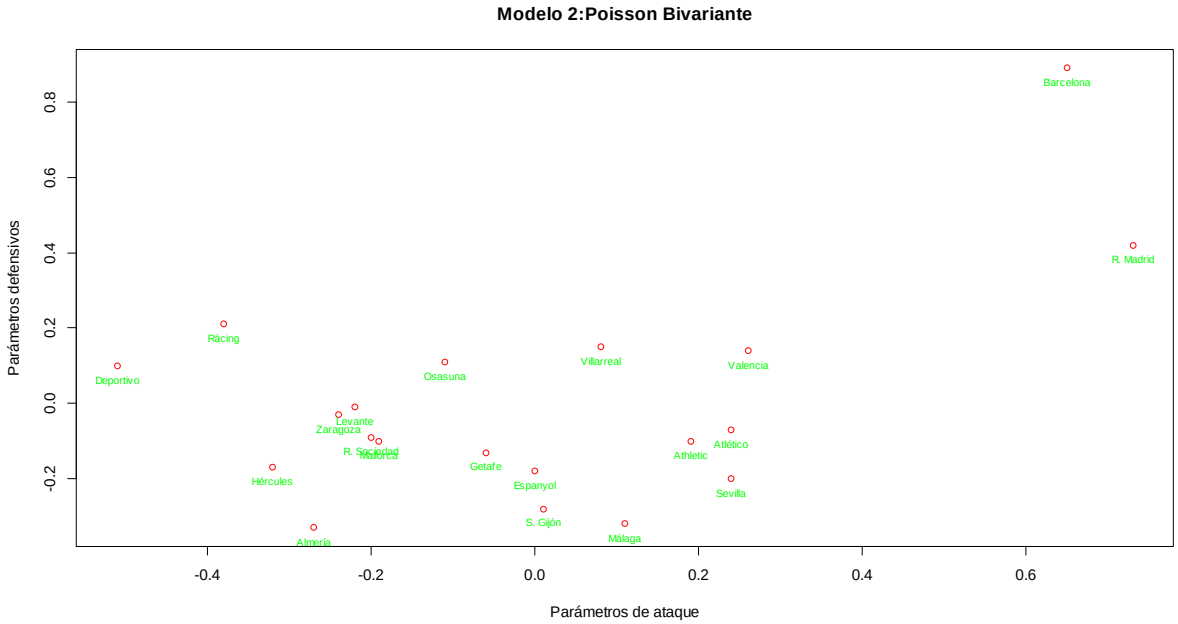


Figura 4.2.: Parámetros defensa y ataque temporada 2010/2011 (DblPois)

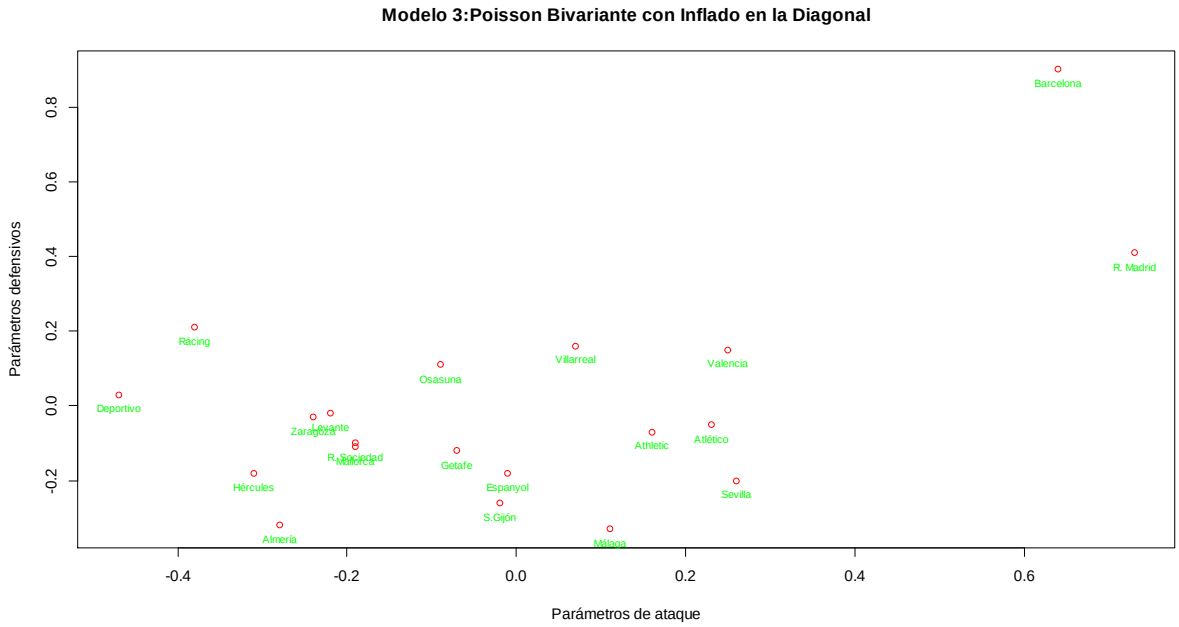


Figura 4.3.: Parámetros defensa y ataque temporada 2011/2012 (DIBP)

Otro aspecto a destacar deportivamente es el efecto de jugar en casa ($\lambda_1 - \lambda_2$), que toma valores entre 0.38 y 0.41 en esta temporada.

Los gráficos anteriores muestran una representación gráfica de los parámetros de defensa y ataque de los modelos 1, 2 y 6 (Figuras 4.1, 4.2 y 4.3). Hay que tener en cuenta que los parámetros defensivos se han multiplicado por menos uno para indicar qué equipos presentaban una mejor defensa. En la esquina superior derecha están los equipos mejor clasificados y en la esquina inferior izquierda los peores. Los tres gráficos son similares y reflejan los resultados ya comentados sobre la Tabla 4.2.

4.3. Análisis para la Temporada 2011/2012

Igual que en la temporada anterior una muestra de los datos se muestra seguidamente:

> datos2

	g1	*g2*	*team1*	*team2*
1	2	*1*	*AtMadrid*	*Athletic*
2	*2*	*0*	*Barcelona*	*Athletic*
3	*3*	*1*	*Betis*	*Athletic*
.....				
379	*1*	2	*Valencia*	*Zaragoza*
380	2	2	*Villarreal*	*Zaragoza*

En total los 20 equipos han disputado 380 partidos. Como se puede observar el primer partido codificado en la matriz de datos es el Atlético de Madrid-Athletic de Bilbao, resultando vencedor el Atlético de Madrid por un tanto. La siguiente fila se corresponde con el partido jugado en casa por el Barcelona, el cuál ganó al Athletic por dos tantos. Todos los partidos se codifican de la misma forma. En nuestra base de datos la última fila se corresponde con el partido Villarreal-Zaragoza, partido que finalizó con un empate a dos.

Los niveles de los equipos nos proporcionan los 20 equipos en 1ª división durante esa temporada:

> levels(datos2[,3])

[1] "Athletic" "Atletico" "Barcelona" "Betis" "Espanyol"

[6] "Getafe" "Granada" "Levante" "Málaga" "Mallorca"

[11] "Osasuna" "R_Gijon" "R_Madrid" "R_Sociedad" "Racing"

[16] "Rayo" "Sevilla" "Valencia" "Villarreal" "Zaragoza"

En relación a la siguiente temporada 2011/2012, los resultados del ajuste de los 12 modelos considerados para la temporada 2011/201 se presentan en la Tabla 3. El primer modelo (Poisson doble) es el que tiene el mejor *AIC* (2212.88), seguido de los modelos: Poisson bivariante constante, y el modelo inflado en la diagonal con Discreta(0). De esta forma, se vuelve a comprobar que los modelos mas sencillos obtienen un mejor ajuste con los resultados de esta temporada. En este temporada se obtuvieron 91 empates en los 380 partidos disputados.

Modelo de distribución	**Detalles adicionales del modelo**	**Log-Likelihood**	**Número de parámet-ros**	**AIC**	**BIC**
1, Poisson doble		-1066.441	40	*2212.883**	*2398.215*
Covariables en λ_3					
2, Poisson bivariante	Parámetro asociado a λ_3 Constante ($\gamma_1 = \gamma_2 = 0$)	-1065.548	41	2213.095	2403.061
3, Poisson bivariante	Equipo local ($\gamma_1 = 1, \gamma_2 = 0$)	-1061.445	60	2242.890	2520.889
4, Poisson bivariante	Equipo visitante ($\gamma_1 = 0, \gamma_2 = 1$)	-1062.503	60	2245.006	2523.005
5, Poisson bivariante	Ambos (local y visitante) ($\gamma_1 = 1, \gamma_2 = 1$)	-1057.935	79	2273.871	2639.903
6, Poisson bivariante inflado diagonal	Discreta(0) =Constante	-1067.348	42	2218.697	2413.296
Distribución diagonal					
7, Poisson bivariante inflado diagonal	Geométrica	-1069.290	43	2224.581	2423.814
8, Poisson bivariante inflado diagonal	Discreta (1)	-1068.776	43	2223.553	2422.785
9, Poisson bivariante inflado diagonal	Discreta (2)	-1067.986	44	2223.972	2427.838
10, Poisson bivariante inflado diagonal	Discreta (3)	-1067.464	45	2224.929	2433.428
11, Poisson bivariante inflado diagonal	Poisson	-1068.962	43	2223.925	2423.158
12, Poisson bivariante inflado diagonal	Poisson	-1068.795	42	2221.590	2416.189

Tabla 4.3. Ajuste de los 12 modelos para la Temporada 2011/2012

Igual que en la temporada anterior se muestran los parámetros estimados para los mejores modelos: 1, 2 y 6 (véase Tabla 4.4).

	Modelo 1:DblPois		**Modelo 2: Bivpois**		**M6: DIBP**	
	Ataque	*Defensa*	*Ataque*	*Defensa*	*Ataque*	*Defensa*
Real Madrid	0.89	-0.40	0.93	-0.45	0.93	-0.45
Barcelona	0.81	-0.44	0.86	-0.46	0.85	-0.47
Valencia	0.18	-0.14	0.20	-0.16	0.18	-0.11
Málaga	0.10	0.04	0.12	0.05	0.11	0.04
Atlético de Madrid	0.08	-0.10	0.10	-0.10	0.13	-0.07
Levante	0.10	-0.02	0.12	-0.01	0.11	-0.03
Osasuna	-0.09	0.17	-0.09	0.19	-0.05	0.21
Mallorca	-0.16	-0.11	-0.17	-0.12	-0.18	-0.14
Sevilla	0.01	-0.25	0.02	-0.27	0.03	-0.21
Athletic	-0.02	0.00	-0.01	0.00	-0.04	-0.03
Getafe	-0.33	-0.02	-0.38	-0.03	-0.37	-0.03
Real Sociedad	0.11	0.36	0.12	0.39	0.12	0.38
Betis	-0.03	0.09	-0.04	0.09	-0.07	0.06
Espanyol	-0.05	0.09	-0.05	0.10	-0.05	0.10
Rayo	-0.16	0.28	-0.16	0.30	-0.19	0.26
Zaragoza	-0.29	0.16	-0.3	0.18	-0.3	0.18
Granada	-0.33	0.08	-0.36	0.08	-0.37	0.06
Villarreal	-0.22	0.02	-0.24	0.03	-0.23	0.04
S. Gijón	-0.06	0.01	-0.05	0.02	-0.04	0.01
Rácing	-0.54	0.19	-0.63	0.19	-0.59	0.22
Intercepto para λ_1	0.43		0.36		0.40	
Intercepto para λ_2	-0.01		-0.11		-0.07	
Intercepto para λ_3			-2.45		-2.51	
λ_3	0		0.08		0.08	
Efecto de Jugar en casa	0.44		0.47		0.47	
Proporción (p)					0.03	
$\theta_1 = P(X = Y = 1)$					0	

Tabla 4.4. Parámetros estimados para los mejores modelos (2011/2012)

En la Tabla 4.4 los equipos están colocados de nuevo según su puesto en la clasificación final. Los parámetros de ataque y defensa obtenidos son acordes con la clasificación de esa temporada. El Real Madrid se clasificó como el mejor equipo de esa temporada, con el mayor parámetrode ataque. Su parámetro de defensa está muy próximo al F.C. Barcelona, equipo clasificado en segundo lugar. El F. C. Barcelona presenta el segundo parámetro de ataque. En cuanto a la cola de clasificiación señalar que el Rácing es el equipo que obtuvo el peor parámetro tanto de ataque como de defensa.

Igualmente destacamos el efecto de jugar en casa: $\lambda_1 - \lambda_2$, que toma valores entre 0.44 y 0.47, valores superiores a la temporada anterior.

A continuación se muestra una representación gráfica de los parámetros de defensa y ataque de los modelos 1, 2 y 6 (Figuras 13, 14 y 15). Los tres gráficos son similares ya que los parámetros entre modelos son muy similares tal como se mostraban en la Tabla 4.4.

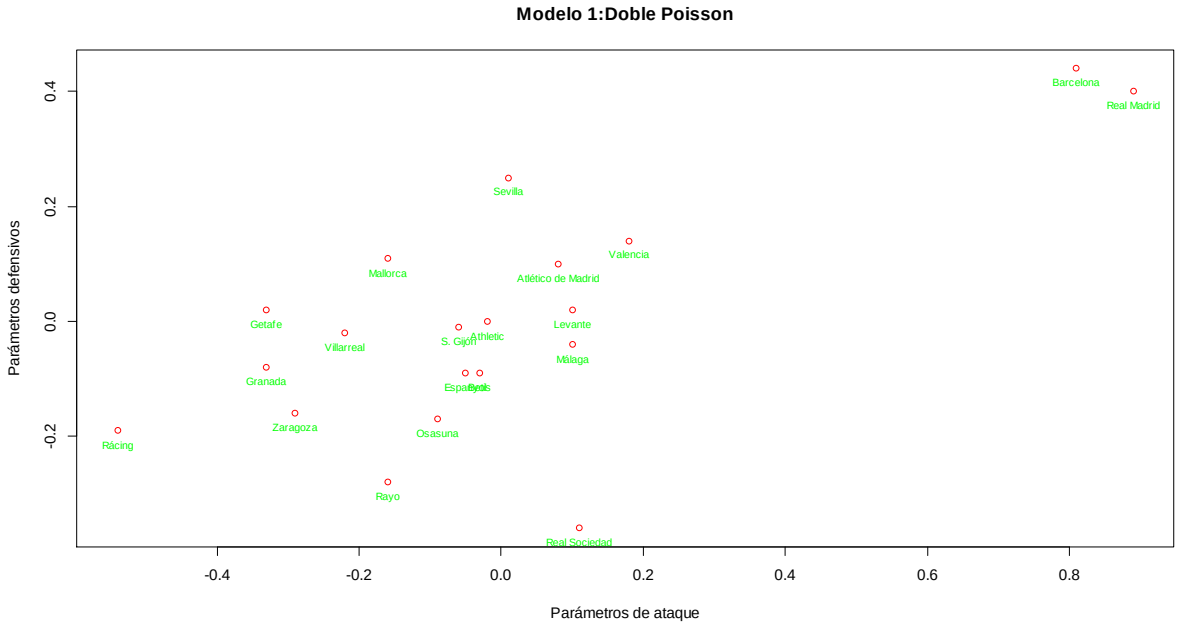


Figura 4.4.: Parámetros defensa y ataque temporada 2011/2012 (DblPois)

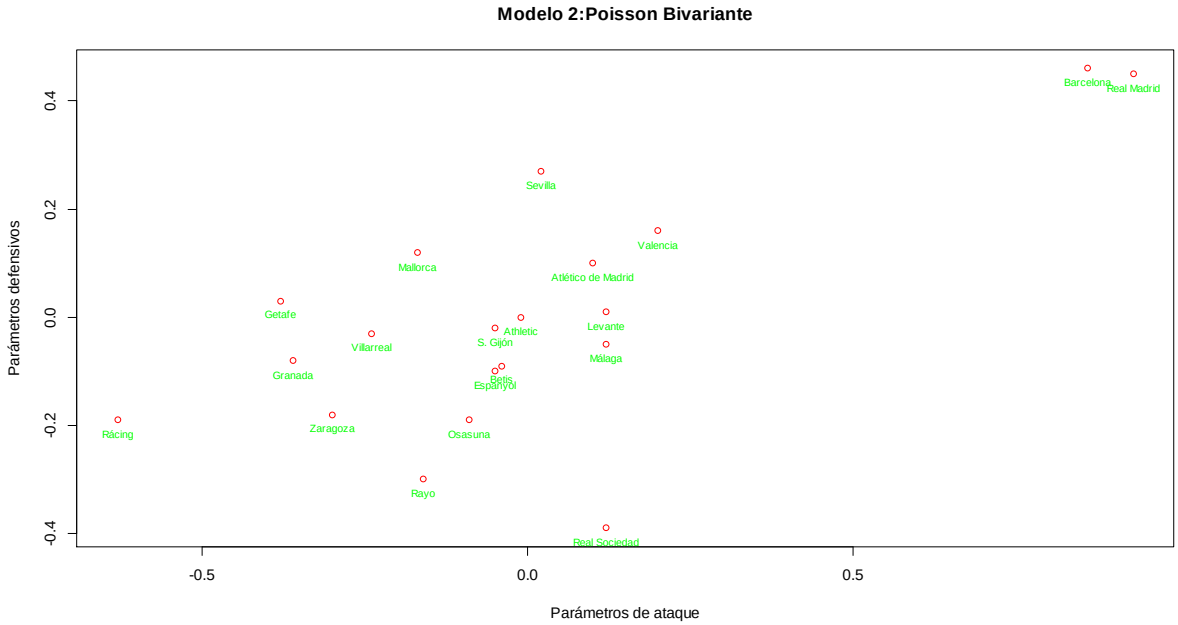


Figura 4.5.: Parámetros defensa y ataque temporada 2011/2012 (Bivpois)

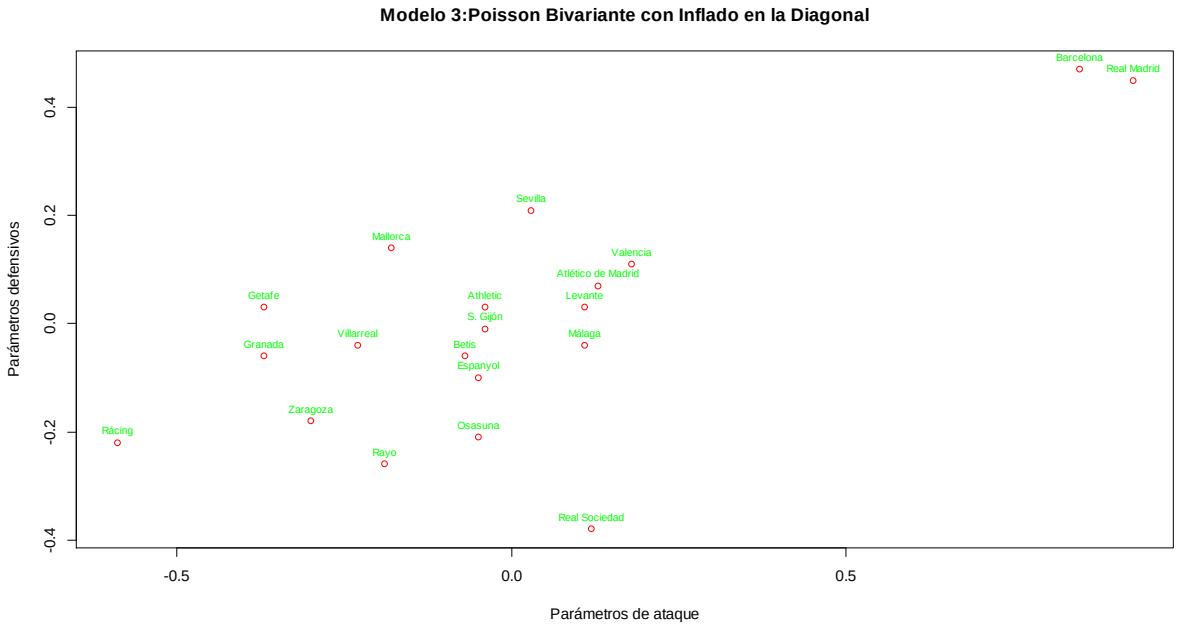


Figura 4.6.: Parámetros defensa y ataque temporada 2011/2012 (DIBP)

Comparando estos gráficos con los de la anterior temporada observamos que el Real Madrid y el Barcelona están mucho más igualados entre sí debido a que el Barcelona ya no tiene tanta ventaja en su capacidad defensiva. Además, en cuanto a la capacidad de ataque, la brecha de ambos equipos respecto al resto aumenta. También es interesante observar que en esta temporada hay una mayor correlación entre los parámetros de defensa y ataque de los equipos que en la temporada anterior. De hecho, si en la temporada 2010/2011 se eliminan Barcelona y Real Madrid se observa que no hay relación entre la capacidad de defensa y de ataque del resto de equipos.

5. Conclusión

Partiendo de una búsqueda bibliográfica de artículos relacionados con la aplicación de modelos de regresión en el fútbol se estableció que el trabajo de Karlis y Ntzoufras (2003) es uno de los trabajos de referencia en este contexto. Estos autores propusieron una regresión de Poisson bivariada aplicada al número de goles marcados por cada equipo y asimismo propusieron modelos de inflado en la diagonal para modelizar mejor los empates. Los autores tomaron como datos los partidos jugados en la Liga italiana en el año 1991. De la misma forma, y utilizando en R el paquete estadístico bivpois, en este trabajo se analizaron los resultados obtenidos en dos temporadas en la Liga Española de Fútbol (2010/2011 y 2011/2012). Los resultados de ambas temporadas se ajustaron mejor por un modelo de Poisson Doble, obteniendo este modelo los mejores valores para el *AIC* y el *BIC*. En el ejemplo analizado por Karlis y Ntzoufras los modelos con inflado en la diagonal ajustaban mejor los resultados obtenidos por los 18 equipos italianos, tal vez porque en la muestra estudiada por ambos autores, se encontraron 111 empates en 306 partidos, mientras que en nuestros datos, los 20 equipos que jugaron los 380 partidos en la temporada 2010/2011 obtuvieron un total de 78 empates y en la siguiente 91.

De la misma forma, se estimaron los parámetros de ataque y defensa para los 20 equipos participantes en cada temporada y se representaron gráficamente. Los resultados obtenidos parecen reflejar que esos parámetros de ataque y defensa ayudan a explicar los resultados de la clasificación. Aunque no se presentaron grandes diferencias en los parámetros obtenidos para los tres modelos con mejor ajuste, si se observaron diferencias en dichos parámetros entre las dos temporadas analizadas. En este sentido sería de interés ampliar los resultados de este trabajo a más temporadas para ver su ajuste con modelos de regresión como los utilizados en este trabajo y también contemplar modelos más avanzados que permitan introducir una componente temporal. De la misma forma, probablemente se podrían elegir otros deportes de equipo y ver si resulta también factible el uso de modelos de regresión de Poisson bivariados y con inflado en la diagonal.

Por otra parte, diversos estudios han contemplado la ventaja de jugar en casa (Courneya y Carron, 1992; Bray y Widmeyer, 2000; Carron, Loughhead y Bray, 2005) desde una perspectiva que tiene en cuenta variables psicológicas. La investigación, en general, parece reflejar la complejidad de la ventaja de jugar en casa (Carron et al., 2005). Esa complejidad no es universal sino que parece depender del equipo y sus jugadores. De esta forma, futuros trabajos podrían estudiar si los resultados de una temporada parecen estar influidos por la percepción que tiene cada equipo de la

ventaja de jugar en casa, es decir, dado que los resultados matemáticos reflejan que existe una ventaja de jugar en casa sería interesante observar si diferentes variables psicológicas pueden explicar esas percepciones de jugar de forma diferente debido a la circunstancia de ser equipo local o visitante.

Agradecimientos

Me gustaría dar las gracias a MariCarmen Iglesias, directora de este trabajo, por su inmensa paciencia. Valoro mucho su tiempo dedicado y todo lo que me ha enseñado. Mi agradecimiento también es, en general, para todos mis profesores y compañeros de máster, y en especial, para Marta Sestelo, Nora Martinez, Yalile Salfate y Andrea Lagoa.

A. Propiedades

A.1. Propiedad 1

Sean X_1, X_2, X_3 tres v.a. independientes de Poisson con medias λ_1, λ_2 y λ_3,respectivamente. Entonces la función de probabilidad conjunta de $X = X_1 + X_3$ e $Y = X_2 + X_3$ e

$$p(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{x!} \sum_{k=0}^{min\{x,y\}} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k$$

para $x, y = 0, 1, 2, ...$

Demostración. Se tiene que $p_{\vec{x}}(x_1, x_2, x_3) = e^{-(\lambda_1,\lambda_2,\lambda_3)} \frac{\lambda_1^{x_1}\lambda_2^{x_2}\lambda_3^{x^3}}{x_1!x_2!x_3!}$; $x_1, x_2, x_3 = 0, 1, ...$

Sean $U_1 = X_1 + X_3$, $U_2 = X_2 + X_3$ y $U_3 = X_3$. Entonces $X_1 = U_1 - U_3$, $X_2 = U_2 - U_3$ y $X_3 = U_3$.

La función de probabilidad de (U_1, U_2, U_3) viene dada por (ver p.e. Rohatgi pág. 131):

$$\begin{aligned} p_{\vec{u}}(u_1, u_2, u_3) = p_{\vec{x}}(u_1 - u_3, u_2 - u_3, u_3) &= e^{-(\lambda_1+\lambda_2+\lambda_3)} \cdot \frac{\lambda_1^{u_1-u_3}}{(u_1-u_3)!} \frac{\lambda_2^{u_2-u_3}}{(u_2-u_3)!} \frac{\lambda_3^{u_3}}{u_3!} \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)} \lambda_1^{u_1} \lambda_2^{u_2} \cdot \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{u_3} \cdot \frac{1}{(u_1-u_3)!(u_2-u_3)!u_3!} \end{aligned}$$

Además, $u_1 - u_3 \geqslant 0, u_2 - u_3 \geqslant 0 \Rightarrow u_3 \leq min\{u_1, u_2\}$. También u_1, u_2 y $u_3 = 0, 1, 2, ...$

Ahora, para $u_1, u_2 = 0, 1, ...$

$$\begin{aligned} p(u_1, u_2) = &\sum_{u_3=0}^{min\{u_1,u_2\}} p_{\vec{u}}(u_1, u_2, u_3) = \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)} \lambda_1^{u_1} \lambda_2^{u_2} \sum_{k=0}^{min\{u_1,u_2\}} \left(\frac{\lambda_3}{\lambda_1\lambda_3}\right)^k \frac{1}{(u_1-k)!(u_2-k)!k!}. \end{aligned}$$

Utilizando $\binom{u_1}{k} = \frac{u_1!}{k!(u_1-k)!}$ el sumatorio anterior se escribe

$$\sum_{k=0}^{min\{u_1,u_2\}} \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k \frac{k!}{u_1!} \binom{u_1}{k} \frac{k!}{u_2!} \binom{u_2}{k} \frac{1}{k!} = \sum_{k=0}^{min\{u_1,u_2\}} \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k \binom{u_1}{k} \binom{u_2}{k} k!$$

y se obtiene el resultado deseado.

A.2. Propiedad 2

Si (X,Y) tienen distribución bivariada de Poisson $BP(\lambda_1,\lambda_2,\lambda_3)$, marginalmente tienen distribuciones de Poisson con $E(X)=\lambda_1+\lambda_3$ y $E(Y)=\lambda_2+\lambda_3$.

Demostración.

$$P(X=x)=\sum_{y=0}^{\infty}p(x,y)=\sum_{y=0}^{\infty}e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x\lambda_2^x}{x!y!}\sum_{k=0}^{min\{x,y\}}\binom{x}{k}\binom{y}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k=$$
$$=\frac{e^{-(\lambda_1+\lambda_2+\lambda_3)}}{x!}\lambda_1^x\sum_{y=0}^{\infty}\frac{\lambda_2^y}{y!}\sum_{k=0}^{min\{x,y\}}\binom{x}{y}\binom{y}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k.$$

Supongamos $x<y$,entonces :

$$P(X=x)=e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x}{x!}\sum_{k=0}^{x}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k\binom{x}{k}k!\sum_{y=k}^{\infty}\binom{y}{k}\frac{\lambda_2^y}{y!}=$$
$$=e^{-(\lambda_1+\lambda_3)}\frac{\lambda_1^x}{x!}\sum_{k=0}^{x}\left(\frac{\lambda_3}{\lambda_1}\right)^k\binom{x}{k}k!\sum_{y=k}^{\infty}\frac{y!}{k!\,(y-k)!}\frac{\lambda_2^{y-k}e^{-\lambda_2}}{y!}=$$
$$=e^{-(\lambda_1+\lambda_3)}\frac{\lambda_1^x}{x!}\sum_{k=0}^{x}\left(\frac{\lambda_3}{\lambda_1}\right)^k\binom{x}{k}\sum_{y=k}^{\infty}e^{-\lambda_2}\frac{\lambda_2^{y-k}}{(y-k)!}$$

Teniendo en cuenta que $\sum_{y=k}^{\infty}e^{-\lambda_2}\frac{\lambda_2^{y-k}}{(y-k)!}=1$(suma de las probabilidades de una Poisson) y $\sum_{k=0}^{x}\left(\frac{\lambda_3}{\lambda_1}\right)^k\binom{x}{k}=\left(1+\frac{\lambda_3}{\lambda_1}\right)^x$ (binomio de Newton), llegamos a

$P(X=x)=e^{-(\lambda_1+\lambda_3)}\frac{\lambda_1^x}{x!}\frac{(\lambda_1+\lambda_3)^x}{\lambda_1^x}=e^{-(\lambda_1+\lambda_3)}\frac{(\lambda_1+\lambda_3)^x}{x!}$, función de probabilidad de $P(\lambda_1+\lambda_3)$.

A.3. Propiedad 3

$$E\left[X_{3i}\mid x_iy_i,\emptyset^{(k)}\right]=\begin{cases}\lambda_{3i}\frac{f_{BP}\left(x_i-1,y_i-1|\lambda_{i1}^k\lambda_{i2}^k\lambda_{i3}^k\right)}{f_{BP}\left(x_i,y_i|\lambda_{i1}^k\lambda_{i2}^k\lambda_{i3}^k\right)} & si\quad min\{x_i,y_i\}>0\\ 0 & si\quad min\{x_i,y_i\}=0\end{cases}$$

Demostración.

$$E[X_3\mid X=x,Y=y]=\sum_{x_3}x_3\cdot f(x_3\mid x,y)=$$
$$=\sum_{x_3}x_3\frac{f(x,y,x_3)}{f(x,y)}$$

Por tanto, tenemos que demostrar $\sum_{x_3} x_3 f(x,y,x_3) = \lambda_3 f\,(x-1,y-1)$.(Términos que llamaremos:$A = B$)

La distribución de (x,y,x_3) se obtiene teniendo en cuenta que es una transformación de $(x1,x2,x3)$ y que $f_x\,(x_1x_2x_3) = e^{-\lambda_1-\lambda_2-\lambda_3}\frac{\lambda_1^{x_1}\lambda_2^{x_2}\lambda_3^{x_3}}{x_1!x_2!x_3!}$. En particular,

$$\begin{array}{ll} x = x_1 + x_3 & x_1 = x - x_3 \\ y = x_2 + x_3 \text{ y} & x_2 = y - x_3 \text{ llevan a} \\ x_3 = x_3 & x_3 = x_3 \end{array}$$

$$f\,(x,y,x_3) = f_x\,(x - x3, y - x3, x3) = \\ = \frac{e^{-\lambda_1-\lambda_2-\lambda_3}\lambda_1^{x-x_3}\lambda_2^{y-x_3}\lambda_3^{x_3}}{(x-x_3)!\,(y-y_3)!x_3!}$$

para $x_3 < min\,\{x,y\}$.

Utilizando la definición de un número combinatorio, llegamos a

$f\,(x,y,x_3) = \frac{e^{-\lambda_1-\lambda_2-\lambda_3}\lambda_1^x\lambda_2^y}{x_3}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x_3}\begin{pmatrix} x \\ x_3 \end{pmatrix}\begin{pmatrix} y \\ x_3 \end{pmatrix}\frac{x_3!}{x!y!}$. Por tanto,

$$A = \sum_{x_3=0}^{min\{x,y\}} x_3 \cdot x_3!\begin{pmatrix} x \\ x_3 \end{pmatrix}\begin{pmatrix} y \\ x_3 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x_3}\left[e^{-\lambda_1-\lambda_2-\lambda_3}\frac{\lambda_1^x\lambda_2^y}{x!y!}\right]$$

$$B = \lambda_3 \cdot f\,(x-1,y-1) = \\ = \lambda_3 \cdot \left[e^{-\lambda_1-\lambda_2-\lambda_3}\cdot\frac{\lambda_1^{x-1}\lambda_2^{y-1}}{(x-1)!\,(y-1)!}\sum_{x_3=0}^{min\{x-1,y-1\}}\begin{pmatrix} x-1 \\ x_3 \end{pmatrix}\begin{pmatrix} y-1 \\ x_3 \end{pmatrix}\left(\frac{\lambda}{\lambda_1\lambda_2}\right)^{x_3} x_3!\right] \\ = \left[e^{-\lambda_1-\lambda_2-\lambda_3}\frac{\lambda_1^x\lambda_2^y}{x!y!}\right]\left(\lambda_1^{-1}\lambda_2^{-1}x\cdot y\lambda_3\right)\sum_{x_3=0}^{min\{x-1,y-1\}}\begin{pmatrix} x-1 \\ x_3 \end{pmatrix}\begin{pmatrix} y-1 \\ x_3 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x_3} x_3!$$

Para comprobar la igualdad entre A y B basta ver que $A_1 = B_1$, siendo

$$A_1 = \sum_{x_3=0}^{min\{x,y\}} x_3 \cdot x_3!\begin{pmatrix} x \\ x_3 \end{pmatrix}\begin{pmatrix} y \\ x_3 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x_3}$$

$$B_1 = \left(\lambda_1^{-1}\lambda_2^{-1}x\cdot y\lambda_3\right)\sum_{x_3=0}^{min\{x-1,y-1\}}\begin{pmatrix} x-1 \\ x_3 \end{pmatrix}\begin{pmatrix} y-1 \\ x_3 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x_3} x_3!$$

Suponiendo que $x < y$

$$A_1 = 1\cdot 1!\begin{pmatrix} x \\ 1 \end{pmatrix}\begin{pmatrix} y \\ 1 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^1 + 2\cdot 2!\begin{pmatrix} x \\ 2 \end{pmatrix}\begin{pmatrix} y \\ 2 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^2 + ...+ \\ + (x-1)\cdot(x-1)!\begin{pmatrix} x \\ x-1 \end{pmatrix}\begin{pmatrix} y \\ x-1 \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x-1} + xx!\begin{pmatrix} x \\ x \end{pmatrix}\begin{pmatrix} y \\ x \end{pmatrix}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^x$$

Teniendo en cuenta que:

$$\begin{pmatrix} x \\ k+1 \end{pmatrix} = \frac{x}{k+1} \begin{pmatrix} x-1 \\ k \end{pmatrix} \Rightarrow \begin{pmatrix} x-1 \\ k \end{pmatrix} = \frac{k+1}{x} \begin{pmatrix} x \\ k+1 \end{pmatrix}$$

B_1se puede escribir de la forma:

$$B_1 = \sum_{x_3}^{min\{x-1\}} x \cdot y \frac{(x_3+1)}{x} \begin{pmatrix} x \\ x_3+1 \end{pmatrix} \frac{(x_3+1)}{y} \begin{pmatrix} y \\ x_3+1 \end{pmatrix} x_3! \begin{pmatrix} \lambda_3 \\ \lambda_1\lambda_2 \end{pmatrix}^{x_3+1}$$

$$= \sum \begin{pmatrix} x \\ x_3` \end{pmatrix} \begin{pmatrix} y \\ x_3` \end{pmatrix} \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^{x_3`} x_3` \cdot x_3`(x_3`-1)! = A1$$

siendo $x_3` = x3+1$. Concluyendo de este modo la demostración.

B. Código R para los gráficos del Capítulo 2

B.1. Gráfico Clasificación General Temporada 2010/2011

```
puntos2010_2011<-c(96,92,71,62,58,58,58,49,47,47,46,46,45,45,45,44,44,43,35,30)
rango <- range(0,puntos2010_2011)
plot(puntos2010_2011,type="o", col="blue",ylim=rango,axes=FALSE,ann=FALSE)
title(main="Clasificación General 2010/2011", col.main="red", font.main=4)
axis(1, at=1:20, lab=c("Barcelona","R. Madrid","Valencia","Villarreal","Sevilla",
"Athletic","At.Madrid","Espanyol","Osasuna","R.S. Gijón","Málaga","R.Racing",
"Zaragoza","Levante","R. Sociedad","Getafe","Mallorca","Deportivo",
"Hércules","Almería"),las=2,cex.axis=0.80)
axis(2, las=1,at=c(96,92,71,62,58,49,35,30))
box()
title(ylab="Puntos Totales", col.lab="green2")
```

B.2. Gráfico Clasificación General Temporada 2011/2012

```
puntos2011_2012<-c(100,91,61,58,56,55,54,52,50,49,47,47,47,46,43,43,42,41,37,27)
rango <- range(0,puntos2011_2012)
plot(puntos2011_2012,type="o", col="blue",ylim=rango,axes=FALSE,ann=FALSE)
title(main="Clasificación General 2011/2012", col.main="red", font.main=4)
axis(1, at=1:20, lab=c("R. Madrid","Barcelona",
"Valencia","Málaga","At.Madrid","Levante","Osasuna","Mallorca","Sevilla","Athletic",
```

```
"Getafe","R.Sociedad","Real Betis","Espanyol","R.Vallecano","Zaragoza","Granada",
"Villareal","R.S.Gijón","R. Racing"),las=2,cex.axis=0.80)
axis(2, las=1,at=c(100,91,61,50,41,37,27))
box()
title(ylab="Puntos Totales", col.lab="green2")
```

B.3. Gráfico Partidos jugados Temporada 2010/2011

```
Barcelona<-c(30,6,2)
R.Madrid<-c(29,5,4)
Valencia<-c(21,8,9)
Villarreal<-c(18,8,12)
Sevilla<-c(17,7,14)
Athletic<-c(18,4,16)
At.Madrid<-c(17,7,14)
Espanyol<-c(15,4,19)
Osasuna<-c(13,8,17)
R.Gijón<-c(11,14,13)
Málaga<-c(13,7,18)
R.Racing<-c(12,10,16)
Zaragoza<-c(12,9,17)
Levante<-c(12,9,17)
Sociedad<-c(14,3,21)
Getafe<-c(12,8,18)
Mallorca<-c(12,8,18)
Deportivo<-c(10,13,15)
Hércules<-c(9,8,21)
Almería<-c(6,12,20)
df <- data.frame(Barcelona,R.Madrid,Valencia,Villarreal,Sevilla,Athletic,At.Madrid,Espanyol,
Osasuna,R.Gijón,Málaga,R.Racing,Zaragoza,Levante,Sociedad,Getafe,Mallorca,Deportivo,
Hércules,Almería)
df
```

```
barplot(as.matrix(df), main="Partidos ganados, empatados y perdidos",
ylab="Total", beside=TRUE, col=rainbow(3),las=2,cex.axis=0.30)
legend("top",c("Partidos ganados","Partidos empatados","Partidos perdidos"),
bty="n",cex=0.6,fill=rainbow(3))
```

B.4. Gráfico Partidos jugados Temporada 2011/2012

```
R.Madrid<-c(32,4,2)
Barcelona<-c(28,7,3)
Valencia<-c(17,10,11)
Málaga<-c(17,7,14)
At.Madrid<-c(15,11,12)
Levante<-c(16,7,15)
Osasuna<-c(13,15,10)
Mallorca<-c(14,10,14)
Sevilla<-c(13,11,14)
Athletic<-c(12,13,13)
Getafe<-c(12,11,15)
Sociedad<-c(12,11,15)
Betis<-c(13,8,17)
Espanyol<-c(12,10,16)
Rayo<-c(13,4,21)
Zaragoza<-c(12,7,19)
Granada<-c(12,6,20)
Villarreal<-c(9,14,15)
R.Gijón<-c(10,7,21)
R.Racing<-c(4,15,19)
df2 <- data.frame(R.Madrid,Barcelona,Valencia,Málaga,At.Madrid,Levante,
Osasuna,Mallorca,Sevilla,Athletic,Getafe,Sociedad,Betis,Espanyol,Rayo,
Zaragoza,Granada,Villarreal,R.Gijón,R.Racing)
df2
barplot(as.matrix(df2), main="Partidos ganados, empatados y perdidos",
```

```
ylab="Total", beside=TRUE, col=rainbow(3),las=2,cex.axis=0.30)
legend("top",c("Partidos ganados","Partidos empatados","Partidos perdidos"),
bty="n",cex=0.6,fill=rainbow(3))
```

B.5. Gráfico Goles Favor/Contra Temporada 2010/2011

```
Barcelona<-c(95,21)
R.Madrid<-c(102,33)
Valencia<-c(64,44)
Villarreal<-c(54,44)
Sevilla<-c(62,61)
Athletic<-c(59,55)
At.Madrid<-c(62,53)
Espanyol<-c(46,55)
Osasuna<-c(45,46)
R.Gijón<-c(35,42)
Málaga<-c(54,68)
R.Racing<-c(41,56)
Zaragoza<-c(40,53)
Levante<-c(41,52)
Sociedad<-c(49,66)
Getafe<-c(49,60)
Mallorca<-c(41,56)
Deportivo<-c(31,47)
Hércules<-c(36,60)
Almería<-c(36,70)
df3<- data.frame(Barcelona,R.Madrid,Valencia,Villarreal,Sevilla,Athletic,
At.Madrid,Espanyol,Osasuna,R.Gijón,Málaga,R.Racing,Zaragoza,Levante,
Sociedad,Getafe,Mallorca,Deportivo,Hércules,Almería)
df3
barplot(as.matrix(df3), main="Goles a Favor y en Contra",
ylab="Total", beside=TRUE, col=rainbow(2),las=2,cex.axis=0.30)
legend("top",c("Goles a Favor","Goles en Contra"),bty="n",cex=0.6,fill=rainbow(2))
```

B.6. Gráfico Partidos jugados Temporada 2011/2012

```
R.Madrid<-c(121,32)
Barcelona<-c(114,29)
Valencia<-c(59,44)
Málaga<-c(54,53)
At.Madrid<-c(53,46)
Levante<-c(54,50)
Osasuna<-c(44,61)
Mallorca<-c(42,46)
Sevilla<-c(48,47)
Athletic<-c(49,52)
Getafe<-c(40,51)
Sociedad<-c(46,52)
Betis<-c(47,56)
Espanyol<-c(46,56)
Rayo<-c(53,73)
Zaragoza<-c(36,61)
Granada<-c(35,56)
Villarreal<-c(39,53)
R.Gijón<-c(42,69)
R.Racing<-c(28,63)
df4 <- data.frame(R.Madrid,Barcelona,Valencia,Málaga,At.Madrid,Levante,
Osasuna,Mallorca,Sevilla,Athletic,Getafe,Sociedad,Betis,Espanyol,Rayo,
Zaragoza,Granada,Villarreal,R.Gijón,R.Racing)
df4
barplot(as.matrix(df4), main="Goles a Favor y en Contra",
ylab="Total", beside=TRUE, col=rainbow(2),las=2,cex.axis=0.30)
legend("top",c("Goles a Favor","Goles en Contra"),bty="n",cex=0.6,fill=rainbow(2))
```

C. Descripción del paquete estadistico “bivpois”

C.1. Función lm.bp

Esta función hace referencia al modelo de regresión de Poisson bivariante general. A continuacion se realiza una descripción mas detallada de la misma.

C.1.1. Descripción

Esta función produce un objeto “lista” el cuál proporciona detalles acerca del ajuste del modelo de regresión de Poisson de la forma:

$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$, para $i = 1, 2, ..., n$ con $log\lambda_1 = w_1\beta_1$, $log\lambda_2 = w_2\beta_2$ y $log\lambda_3 = w_3\beta_3$

donde

n es el tamaño de la muestra,

$\lambda_\kappa = (\lambda_{\kappa 1}, \lambda_{\kappa 2}, ..., \lambda_{\kappa n})'$ para $\kappa = 1, 2, 3$ son vectores de longitud n con una lambda estimada para cada observación,

w_1, w_2, es una matriz de datos $n \times p$ que contiene variables explicativas para λ_1 y λ_2,

w_3 es una matriz de datos $n \times p_2$ que contiene variables explicativas para λ_3,

β_1,β_2,β_3 son vectores de los parámetros usados en los predictores lineales de λ_1, λ_2y λ_3.

C.1.2. Uso

El código para la función lm.bp es el siguiente:

lm.bp(l1, l2, l1l2=NULL, l3=~1, data, common.intercept=FALSE, zeroL3=FALSE, maxit=300, pres=1e-8, verbose=getOption("verbose"))

C.1.3. Argumentos

Obligatorios.

l1 Fórmula de la forma “$x \sim X_1 + ... + X_p$” para los parámetros del $log\lambda_1$.

12 Fórmula de la forma “$y \sim X_1 + ... + X_p$” para los parámetros del $log\lambda_2$.

Opcionales.

l1l2 Fórmula de la forma “$\sim X_1 + ... + X_p$” para los parámetros comunes del $log\lambda_1$ y $log\lambda_2$. Si la variable explicativa se encuentra también en l1 y/o l2 entonces el modelo ajusta la interacción entre los parámetros. Se pueden usar aquí términos especiales de la forma “$c(X_1, X_2)$”. Estos términos implican parámetros comunes de λ_1 y λ_2 para distintas variables.

l3 Fórmula de la forma “$\sim X_1 + ... + X_p$” para los parámetros del $log\lambda_3$.

data Data frame que contiene las variables en el modelo

common.intercept Función lógica que específica si se debe usar un intercepto común sobre λ_1 y λ_2. Por defecto su valor es *FALSE*.

zeroL3 Argumento lógico que controla si λ_3 debería ser igual a cero (y por lo tanto se ajustaría el modelo de Poisson doble)

maxit Número máximo de pasos EM. Por defecto su valor es de 300 iteraciones.

pres Precisión usada para parar el algoritmo EM. El algoritmo se detiene cuando la diferencia de verosimilitud relativa es más pequeña que el valor de pres.

verbose Argumento lógico que controla si los parámetros beta tienen que ser realizados mientras se calcula EM. El valor por defecto se toma igual al valor de *options()$verbose*. Si *verbose=FALSE* entonces solo el número de iteraciones, el log de verosimilitud y su diferencia relativa de iteraciones previas se llevan a cabo. Si *verbose=TRUE* entonces los parámetros del modelo β_1, β_2y β_3 se llevan a cabo adicionalmente.

C.1.4. Valores

Al ejecutar esta función se obtiene un objeto de tipo lista con las siguientes componentes:

coefficients Estima los parámetros del modelo para β_1, β_2 y β_3 . Cuando se usa un factor, se obtienen estimadores según la codificación utilizada.

fitted.values Data frame con n líneas y 2 columnas que contiene los valores ajustados para x e y. Para un modelo de Poisson bivariado los valores ajustados son $\lambda_1+\lambda_3$ y $\lambda_2+\lambda_3$, respectivamente.

residuals Data frame con n líneas y 2 columnas que contiene los residuos del modelo para x e y. Para un modelo de Poisson bivariado los valores residuales vienen dados por $x - E(x)$ e $y - E(y)$ respectivamente; donde $E(x) = \lambda_1 + \lambda_3$ y $E(y) = \lambda_2 + \lambda_3$.

beta1, beta2, beta 3 Vectores β_1, β_2 y β_3 que contienen los coeficientes implicados en la predicción lineal de λ_1, λ_2 y λ_3 respectivamente. Cuando *zeroL3=TRUE* entonces *beta3* no se calcula.

lambda1, lambda2 Vectores de longitud n que contiene la estimación λ_1 y λ_2 para cada observación.

lambda3 Vector que contiene los valores de λ_3. Si *zeroL3=TRUE* entonces lambda3 es igual a cero y no se proporciona.

loglikelihood log-likelihood del modelo ajustado. Se da en forma de vector (un valor por iteración). Usando este vector se puede ver la evolución del log de verosimilitud en cada paso EM.

AIC, BIC AIC y BIC del modelo. Los valores también son proporcionados para el modelo de doble Poisson y el modelo saturado.

parameters Número de parámetros.

iterations Número de iteraciones.

call Argumento que proporciona los detalles exactos usados para realizar la función lm.bp.

C.2. Función lm.dibp

Este segunda función está relacionada con el modelo de regresión de Poisson bivariante inflado en la diagonal. Igual que en el apartado anterior se hace una descripción de la función.

C.2.1. Descripción

Produce un objeto "lista" el cual proporciona detalles en relación al ajuste de un modelo de regresión de Poisson bivariado inflado en la diagonal de la forma

$(X_i, Y_i) \sim DIBP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i},\ D(\theta))$ que seria el equivalente a

$(X_i, Y_i) \sim (1-p)BP(x_i, y_i \mid \lambda_{1i}, \lambda_{2i}, \lambda_{3i})$ si $x_i \neq y_i$

$(X_i, Y_i) \sim (1-p)BP(x_i, y_i \mid \lambda_{1i}, \lambda_{2i}, \lambda_{3i}) + pD(x_i \mid \theta)$ si $x_i = y_i$ para $i = 1, 2, ..., n$

con

$log\lambda_1 = w_1\beta, log\lambda_2 = w_2\beta$ y $log\lambda_3 = w_3\beta_3$

donde

n es el tamaño de la muestra,

$\lambda_\kappa = (\lambda_{\kappa 1}, \lambda_{\kappa 2}, ..., \lambda_{\kappa n})'$ para $\kappa = 1, 2, 3$ son vectores de longitud n con una lambda estimada para cada observación,

w_1, w_2, es una matriz de datos $n \times p$ que contiene variables explicativas para λ_1y λ_2,

w_3 es una matriz de datos $n \times p_3$que contiene variables explicativas para λ_3,

β es un vector de longitud p el cual es común para λ_1y $\lambda_2$en orden a permitir los efectos comunes,

β_3 es un vector de longitud p_3,

$D(\theta)$ es una distribución discreta con vector de parámetro θusada para inflar la diagonal,

p es la proporción de la inflacción utilizada en la mezcla (suma) de las distribuciones.

C.2.2. Uso

El código para la función lm.dibp es el siguiente:

lm.dibp(l1, l2, l1l2=NULL, l3=~1, data, common.intercept=FALSE, zeroL3 = FALSE, distribution = "discrete", jmax = 2, maxit = 300,

pres = 1e-08, verbose=getOption("verbose"))

C.2.3. Argumentos

Obligatorios.

l1 Fórmula de la forma “$x \sim X_1 + ... + X_p$” para los parámetros del $log\lambda_1$.

12 Fórmula de la forma “$y \sim X_1 + ... + X_p$” para los parámetros del $log\lambda_2$.

Opcionales.

l1l2 Fórmula de la forma “$\sim X_1 + ... + X_p$” para los parámetros comunes del $log\lambda_1$ y $log\lambda_2$. Si la variable explicativa se encuentra en l1 y/o l2 entonces el modelo usa la interacción entre los parámetros que se ajustan (un parámetro común para ambos predictores [efecto principal] y las diferencias entre estos para otro predictor [efecto de la interacción]. Términos especiales de la forma “$c(X_1, X_2)$ pueden ser usados aqui. Estos términos implican parámetros comunes de λ_1 y λ_2.

l3 Fórmula de la forma “$\sim X_1 + ... + X_p$” para los parámetros del $log\lambda_3$.

data Data frame que contiene las variables en el modelo

common.intercept Función lógica que específica si un intercepto común sobre λ_1 y λ_2 debería ser usada. Por defecto su valor es *FALSE*.

zeroL3 Argumento lógico que controla si λ_3 debería ser un conjunto igual a cero (y por lo tanto ajustar el modelo de Poisson doble)

distribution Especifica el tipo de distribución inflada: =”discrete”: Discrete(J=jmax),

=”poisson”: Poisson (θ) =”geometric”: Geométrica (θ).

jmax Número de parámetros usados en la distribución Discreta. Este argumento no se usa para las distribuciones de Poisson o geométrica.

maxit Número máximo de pasos EM. Por defecto su valor es de 300 iteraciones.

pres Precision usada para parar el algoritmo EM. El algoritmo para cuando la diferencia de verosimilitud relativa es más pequeña que el valor de pres.

verbose Argumento lógico que controla si los parámetros beta tienen que ser realizados mientras se calcula EM. El valor por defecto se toma igual al valor de *options()$verbose*. Si *verbose=FALSE* entonces solo el número de iteraciones, el log de verosimilitud y su diferencia relativa de iteraciones previas se llevan a cabo. Si *verbose=TRUE* entonces los parámetros del modelo β_1, β_2y β_3 se llevan a cabo adiccionalmente.

C.2.4. Valores

Al ejecutar esta función se obtiene un objeto de tipo lista con las siguientes componentes:

coefficients Estima los parámetros del modelo para β_1, β_2, β_3, p y θ.

fitted.values Data frame con n líneas y 2 columnas que contiene los valores ajustados para x e y.

residuals Data frame con n líneas y 2 columnas que contiene los residuos del modelo para x e y. Para un modelo de Poisson bivariado los valores residuales vienen dados por $x - E(x)$ e $y - E(y)$ respectivamente; donde $E(x)$ y $E(y)$ vienen dados por *fitted.values*.

beta1, beta2, beta 3 Vectores β_1, β_2 y β_3 que contienen los coeficientes implicados en la prediccion lineal de λ_1, λ_2 y λ_3 respectivamente. Cuando *zeroL3=TRUE* entonces *beta3* no se calcula.

lambda1, lambda2 Vectores de longitud n que contiene la estimación λ_1 y λ_2 para cada observacion.

lambda3 Vector que contiene los valores de λ_3. Si *zeroL3=TRUE* entonces lambda3 es igual a cero y no se proporciona.

loglikelihood log-likelihood del modelo ajustado. Se da en forma de vector (un valor por iteración). Usando este vector se puede ver la evolución del log de verosimilitud en cada paso EM.

AIC, BIC AIC y BIC del modelo. Los valores también son proporcionados para el modelo de doble Poisson y el modelo saturado.

diagonal.distribution Etiqueta usada para la distribución de la diagonal inflada.

theta Vector de parámetros de la distribucion diagonal. Para una distribución discreta la theta tiene una longitud igual a jmax con $\theta_i = theta[i]$ y $\theta_0 = 1 - \sum_{i=1}^{JMAX} \theta_i$;

para una distribucion de Poisson, la theta es la media; para una distribución Geométrica la distribucion theta es la probabilidad del suceso.

parameters Número de parámetros.

iterations Número de iteraciones.

call Argumento que proporciona los detalles exactos usados para realizar la función lm.dibp.

C.3. Función pbivpois

La función pbivpois esta relacionada con la función de probabilidad de la distribución de Poisson bivariada.

C.3.1. Descripción

Devuelve la probabilidad (o el log) de la distribución de Poisson bivariada para los valores x e y.

C.3.2. Uso

pbivpois(x, y=NULL, lambda = c(11, 12, 13), log = FALSE)

C.3.3. Argumentos

x Matriz o Vector que contiene los datos. Si x es una matriz entonces consideramos x la primera columna e y la segunda columna. Si hay columnas adicionales son ignoradas.

y Vector que contiene los datos de y. Solo se usa si x es también otro vector. Los vectores x e y deben tener igual longitud.

l*ambda* Vector (de longitud 3) que contiene los valores de los parámetros λ_1,λ_2 y $\lambda_3$de la distribución de Poisson bivariada.

C.3.4. Detalles

Esta función evalua la función de probabilidad (o su logarítmo) de una distribución de Poisson bivariada con los parámetros λ_1,λ_2 y λ_3.

C.3.5. Valor

La salida de esta función es un vector de valores de las probabilidades de $PD(\lambda_1, \lambda_2, \lambda_3)$ evaluadas en (x, y) cuando *log=FALSE* o las log-probabilidades de $PD(\lambda_1, \lambda_2, \lambda_3)$ evaluadas en (x, y) cuando *log=TRUE*.

D. Código R para el análisis de los datos

A continuación se presenta el código específico para el análisis de los datos partidos de la Liga Española de Fútbol para las temporadas 2010/2011 y 2011/2012.

D.1. Código para el ánalisis de los partidos celebrados en la Temporada 2010/2011

```
#Fichero de datos Temporada 2010/2011
datos1 = read.table("2010.txt",header=TRUE)
names(datos1)
attach(datos1)
datos1
levels(datos1[,3])
options(contrasts = c("contr.sum", "contr.poly"))
# formula for modeling of lambda1 and lambda2
form1 <- ~c(team1,team2)+c(team2,team1)
# # Model 1: Double Poisson
ex4.m1<-lm.bp( g1~1, g2~1, l1l2=form1, zeroL3=TRUE, data=datos1)
# # Models 2-5: bivariate Poisson models
ex4.m2<-lm.bp(g1~1,g2~1, l1l2=form1, data=datos1)
ex4.m3<-lm.bp(g1~1,g2~1, l1l2=form1, l3=~team1, data=datos1)
ex4.m4<-lm.bp(g1~1,g2~1, l1l2=form1, l3=~team2, data=datos1)
ex4.m5<-lm.bp(g1~1,g2~1, l1l2=form1, l3=~team1+team2, data=datos1)
# # Model 6: Zero Inflated Model
ex4.m6 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, jmax=0)
# # Models 7-11: Diagonal Inflated Bivariate Poisson Models
```

```
ex4.m7 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, distribution="geometric")
ex4.m8 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, jmax=1)
ex4.m9 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, jmax=2)
ex4.m10<-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, jmax=3)
ex4.m11<-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos1, distribution="poisson")
# # Models 12: Diagonal Inflated Double Poisson Model
ex4.m12 <- lm.dibp( g1~1,g2~1, l1l2=form1, data=datos1, distribution="poisson",
zeroL3=TRUE)
# ——————————————————————————————
# # ——————————————————————————————
# monitoring parameters for model 1: Dbl Poisson ex4.m1$coef
# all parameters ex4.m1$beta1
# model parameters for lambda1 ex4.m1$beta2 # model parameters for lambda2.
# All are the same as in beta1 except the intercept
ex4.m1$beta2[1] # Intercpept for lambda2.
ex4.m1$beta1[1]-ex4.m1$beta2[1] # estimated home effect
# estimating the effect for 20th level of attack (team1..team2) [Zaragoza]
-sum(ex4.m1$coef[2:20])
# estimating the effect for 20th level of defence(team2..team1) [Zaragoza]
-sum(ex4.m1$coef[21:39])
# # ——————————————————————————————
# monitoring parameters for model 2: BivPoisson(lamdba1,lambda2,
constant lamdba3)
ex4.m2$beta1 # model parameters for lambda1
ex4.m2$beta2 # model parameters for lambda2.
# All are the same as in beta1 except the intercept
ex4.m2$beta3 # model parameters for lambda3 (Here only the intercept)
exp(ex4.m2$beta3)
ex4.m2$beta2[1] # Intercpept for lambda2.
ex4.m2$beta1[1]-ex4.m2$beta2[1] # estimated home effect
# estimating the effect for 20th level of attack (team1..team2) [Zaragoza]
```

```
-sum(ex4.m2$coef[ 2:20])
# estimating the effect for 20th level of defence(team2..team1) [Zaragoza]
-sum(ex4.m2$coef[21:39])
# # ————————————————————————————————
# ————————————————————————————————
# monitoring parameters for model 6: Zero Inflated Model
ex4.m6$beta1 # model parameters for lambda1
ex4.m6$beta2 # model parameters for lambda2.
# All are the same as in beta1 except the intercept
ex4.m6$beta3 # model parameters for lambda3. Here beta3 has only the intercept
ex4.m6$beta2[1] # Intercpept for lambda2.
ex4.m6$beta1[1]-ex4.m6$beta2[1] # estimated home effect
# estimating the effect for 20th level of attack (team1..team2) [Zaragoza]
-sum(ex4.m6$coef[ 2:20])
# estimating the effect for 20th level of defence(team2..team1) [Zaragoza]
-sum(ex4.m6$coef[21:39])
ex4.m6$beta3 # parameters for lambda3 (here the intercept)
exp(ex4.m6$beta3) # lambda3 (here constant)
ex4.m6$diagonal.distribution # printing details for the diagonal distribution
ex4.m6$p # mixing proportion ex4.m8$theta # printing theta parameters
names(ex4.m1)
ex4.m1$coefficients
ex4.m1$fitted.values
ex4.m1$residuals
ex4.m1$beta1
ex4.m1$beta2
ex4.m1$lambda1
ex4.m1$lambda2
ex4.m1$lambda3
ex4.m1$loglikelihood
ex4.m1$iterations
```

```
ex4.m1$parameters
ex4.m1$AIC
ex4.m1$BIC
ex4.m1$call
names(ex4.m2) #Idem para ex4.m3;ex4.m4;ex4.m5;ex4.m6;ex4.m7;
ex4.m8;ex4.m9;ex4.m10;ex4.m11;ex4.m12
ex4.m2$coefficients
ex4.m2$fitted.values
ex4.m2$residuals
ex4.m2$beta1
ex4.m2$beta2
ex4.m2$beta3
ex4.m2$lambda1
ex4.m2$lambda2
ex4.m2$lambda3
ex4.m2$loglikelihood
ex4.m2$iterations
ex4.m2$parameters
ex4.m2$AIC
ex4.m2$BIC
ex4.m2$call
```

D.2. Código para el ánalisis de los partidos celebrados en la Temporada 2011/2012

```
#Fichero de datos Temporada 2011/2012
datos2 = read.table("2011.txt",header=TRUE)
names(datos2)
attach(datoa2)
datos2
levels(datos2[,3])
options(contrasts = c("contr.sum", "contr.poly"))
```

```
# formula for modeling of lambda1 and lambda2
form1 <- ~c(team1,team2)+c(team2,team1)
# # Model 1: Double Poisson
ex4.m1<-lm.bp( g1~1, g2~1, l1l2=form1, zeroL3=TRUE, data=data)
# # Models 2-5: bivariate Poisson models
ex4.m2<-lm.bp(g1~1,g2~1, l1l2=form1, data=datos2)
ex4.m3<-lm.bp(g1~1,g2~1, l1l2=form1, l3=~team1, data=datos2)
ex4.m4<-lm.bp(g1~1,g2~1, l1l2=form1, l3=~team2, data=datos2)
ex4.m5<-lm.bp(g1~1,g2~1, l1l2=form1, l3=~team1+team2, data=datos2)
# # Model 6: Zero Inflated Model
ex4.m6 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos2, jmax=0)
# # Models 7-11: Diagonal Inflated Bivariate Poisson Models
ex4.m7 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos2, distribution="geometric")
ex4.m8 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos2, jmax=1)
ex4.m9 <-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos2, jmax=2)
ex4.m10<-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos2, jmax=3)
ex4.m11<-lm.dibp(g1~1,g2~1, l1l2=form1, data=datos2, distribution="poisson")
# # Models 12: Diagonal Inflated Double Poisson Model
ex4.m12 <- lm.dibp( g1~1,g2~1, l1l2=form1, data=datos2, distribution="poisson",
zeroL3=TRUE)
# ---------------------------------------------------
# # ---------------------------------------------------
# monitoring parameters for model 1: Dbl Poisson ex4.m1$coef
# all parameters ex4.m1$beta1
# model parameters for lambda1 ex4.m1$beta2 # model parameters for lambda2.
# All are the same as in beta1 except the intercept
ex4.m1$beta2[1] # Intercept for lambda2.
ex4.m1$beta1[1]-ex4.m1$beta2[1] # estimated home effect
# estimating the effect for 20th level of attack (team1..team2) [Zaragoza]
-sum(ex4.m1$coef[2:20])
# estimating the effect for 20th level of defence(team2..team1) [Zaragoza]
```

```
-sum(ex4.m1$coef[21:39])
# # ---------------------------------------------------------
# monitoring parameters for model 2: BivPoisson(lamdba1,lambda2,
constant lamdba3)
ex4.m2$beta1 # model parameters for lambda1
ex4.m2$beta2 # model parameters for lambda2.
# All are the same as in beta1 except the intercept
ex4.m2$beta3 # model parameters for lambda3 (Here only the intercept)
exp(ex4.m2$beta3)
ex4.m2$beta2[1] # Intercpept for lambda2.
ex4.m2$beta1[1]-ex4.m2$beta2[1] # estimated home effect
# estimating the effect for 20th level of attack (team1..team2) [Zaragoza]
-sum(ex4.m2$coef[ 2:20])
# estimating the effect for 20th level of defence(team2..team1) [Zaragoza]
-sum(ex4.m2$coef[21:39])
# # ---------------------------------------------------------
# ---------------------------------------------------------
# monitoring parameters for model 6: Biv.Poisson with Dis(1) diagonal distribution
# # # monitoring parameters for model 6
ex4.m6$beta1 # model parameters for lambda1
ex4.m6$beta2 # model parameters for lambda2.
# All are the same as in beta1 except the intercept
ex4.m6$beta3 # model parameters for lambda3. Here beta3 has only the intercept
ex4.m6$beta2[1] # Intercpept for lambda2.
ex4.m6$beta1[1]-ex4.m6$beta2[1] # estimated home effect
# estimating the effect for 20th level of attack (team1..team2) [Zaragoza]
-sum(ex4.m6$coef[ 2:20])
# estimating the effect for 20th level of defence(team2..team1) [Zaragoza]
-sum(ex4.m6$coef[21:39])
ex4.m6$beta3 # parameters for lambda3 (here the intercept)
exp(ex4.m6$beta3) # lambda3 (here constant)
```

```
ex4.m6$diagonal.distribution # printing details for the diagonal distribution
ex4.m6$p # mixing proportion ex4.m6$theta # printing theta parameters
names(ex4.m1)
ex4.m1$coefficients
ex4.m1$fitted.values
ex4.m1$residuals
ex4.m1$beta1
ex4.m1$beta2
ex4.m1$lambda1
ex4.m1$lambda2
ex4.m1$lambda3
ex4.m1$loglikelihood
ex4.m1$iterations
ex4.m1$parameters
ex4.m1$AIC
ex4.m1$BIC
ex4.m1$call
names(ex4.m2) #Idem para ex4.m3;ex4.m4;ex4.m5;ex4.m6;ex4.m7;
ex4.m8;ex4.m9;ex4.m10;ex4.m11;ex4.m12
ex4.m2$coefficients
ex4.m2$fitted.values
ex4.m2$residuals
ex4.m2$beta1
ex4.m2$beta2
ex4.m2$beta3
ex4.m2$lambda1
ex4.m2$lambda2
ex4.m2$lambda3
ex4.m2$loglikelihood
ex4.m2$iterations
ex4.m2$parameters
ex4.m2$AIC
ex4.m2$BIC
ex4.m2$call
```

D.3. Código para los gráficos de parámetros de ataque/defensa

```
###Gráficos para los parámetros ataque y defensa
grafico = read.table("2010a.txt",header=TRUE) names(grafico)
plot(grafico$ataque,grafico$defensa,
xlab = "Parámetros de ataque", ylab = "Parámetros defensivos", main="Modelo 1:Doble Poisson",col=2)
text(grafico$ataque,grafico$defensa,pos=3,labels=c("Almería","Athletic",
"Atlético","Barcelona","Deportivo","Getafe","Espanyol","Hércules","Levante","Málaga",
"Mallorca","Osasuna","Rácing","R. Madrid","S. Gijón","R. Sociedad","Sevilla",
"Valencia","Villarreal","Zaragoza"),col="green",pch=0.5,cex=0.65)
################
grafico2 = read.table("2010b.txt",header=TRUE) names(grafico2)
plot(grafico2$ataque,grafico2$defensa,
xlab = "Parámetros de ataque", ylab = "Parámetros defensivos", main="Modelo 2:Poisson Bivariante",col=2)
text(grafico2$ataque,grafico2$defensa,pos=1,labels=c("Almería","Athletic",
"Atlético","Barcelona","Deportivo","Getafe","Espanyol","Hércules","Levante","Málaga",
"Mallorca","Osasuna","Rácing","R. Madrid","S. Gijón","R. Sociedad","Sevilla",
"Valencia","Villarreal","Zaragoza"),col="green",pch=0.5,cex=0.65)
##################
grafico3 = read.table("2010c.txt",header=TRUE) names(grafico3)
plot(grafico3$ataque,grafico3$defensa,
xlab = "Parámetros de ataque",
ylab = "Parámetros defensivos", main="Modelo 3:Poisson Bivariante con Inflado en la Diagonal",col=2)
text(grafico3$ataque,grafico3$defensa,pos=1,labels=c("Almería","Athletic",
"Atlético","Barcelona","Deportivo","Getafe","Espanyol","Hércules","Levante","Málaga",
"Mallorca","Osasuna","Rácing","R. Madrid","S.Gijón","R. Sociedad","Sevilla",
"Valencia","Villarreal","Zaragoza"),col="green",pch=0.5,cex=0.65)
####################
grafico4 = read.table("2011a.txt",header=TRUE) names(grafico4)
```

```
plot(grafico4$ataque,grafico4$defensa,
xlab = "Parámetros de ataque", ylab = "Parámetros defensivos", main="Modelo 1:Doble Poisson",col=2)
text(grafico4$ataque,grafico4$defensa,pos=1,labels=c("Athletic",
"Atlético de Madrid","Barcelona","Betis","Espanyol","Getafe","Granada","Levante","Málaga",
"Mallorca","Osasuna","Rayo","Real Madrid","S. Gijón","Rácing","Real Sociedad","Sevilla",
"Valencia","Villarreal","Zaragoza"),col="green",pch=0.5,cex=0.65)
######################
grafico5 = read.table("2011b.txt",header=TRUE) names(grafico5)
plot(grafico5$ataque,grafico5$defensa,
xlab = "Parámetros de ataque", ylab = "Parámetros defensivos", main="Modelo 2:Poisson Bivariante",col=2)
text(grafico5$ataque,grafico5$defensa,pos=1,labels=c("Athletic",
"Atlético de Madrid","Barcelona","Betis","Espanyol","Getafe","Granada","Levante","Málaga",
"Mallorca","Osasuna","Rayo","Real Madrid","S. Gijón","Rácing","Real Sociedad","Sevilla",
"Valencia","Villarreal","Zaragoza"),col="green",pch=0.5,cex=0.65)
######################
grafico6 = read.table("2011c.txt",header=TRUE) names(grafico6)
plot(grafico6$ataque,grafico6$defensa,
xlab = "Parámetros de ataque",
ylab = "Parámetros defensivos", main="Modelo 3:Poisson Bivariante con Inflado en la Diagonal",col=2)
text(grafico6$ataque,grafico6$defensa,pos=3,labels=c("Athletic",
"Atlético de Madrid","Barcelona","Betis","Espanyol","Getafe","Granada","Levante","Málaga",
"Mallorca","Osasuna","Rayo","Real Madrid","S. Gijón","Rácing","Real Sociedad","Sevilla",
"Valencia","Villarreal","Zaragoza"),col="green",pch=0.5,cex=0.65)
######################
```

Referencias bibliográficas

Abraira, V. & Pérez de Vargas, A. (1996) Métodos multivariantes en bioestadística. Editorial Centro de Estudios Ramón Areces. Madrid.

Aitchison, J. & Ho, C.H., The multivariate Poisson log-normal distribution (1989). *Biometrika, 76*, pp. 643-653.

Baio, G. & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics, 37*(2), 253–264.

Ben-Naim, E. & N. Hengartner, N. (2007), "Efficiency of competitions" *Phys. Rev. E*76, 026106.

Ben-Naim, E., Redner, S. & Vazquez, F. (2007). Scaling in tournaments. *EPL, 77*: 3005.

Bittner, E., Nussbaumer, A., Janke, W. & Weigel, M. (2007). Self-affirmation model for football goal distributions. *Europhys Lett 78*: 58002. doi: 10.1209/0295-5075/78/58002.

Bittner, E., Nussbaumer, A., Janke, W. & Weigel, M. (2009). Football fever: goal distributions and non-gaussian statistics. *The European Physical Journal B 67*, 459.

Bray, S.R., & Widmeyer, W.N. (2000). Athletes´ perceptions of the home advantage: an investigation of perceived casual factors. *Journal of Sport Behaviour, 23*,1-10.

Brillinger, D. R. (2007). A Potential Function Approach to the Flow of Play in Soccer, *Journal of Quantitative Analysis in Sports, 3*(1).

Brillinger, D. R. (2007). Modelling some Norwegian soccer data. En Nair, VJ (ed.), *Advances in Statistical Modelling and Inference*, pp. 3–20. World Scientific: London.

Brillinger, D. R. (2009). Soccer/world football. Technical report, 2009. URL http://www.stat.berkeley.edu/tech-reports/777.pdf.

Carron, A., Loughhead, T., & Bray, S. (2005). The home advantage in sport competitions: Courneya and Carron's (1992) conceptual framework a decade later. *Journal of Sports Sciences, 23*(4), 395-407.

Courneya, K. S., & Carron, A. V. (1992). The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology, 14*, 13–27.

Colin Cameron A. & Trivedi, P. K. (1998), Regression Analysis of Count Data. *Econometric Society Monograph, 30*, Cambridge University Press.

Dempster, A.P. Laird, N.M. & Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological) 39* (1): 1–38.

Desouza, C.M., (1992). An approximate bivariate Bayesian method for analyzing small frequencies. Biometrics, 48, pp. 1113-1130.

Dixon, M.J. & 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), pp. 265-280.

Dobson, S. & Goddard, J. (2000). Stochastic modelling of soccer match results, volume 44. Citeseer.

Dyte, D. & Clarke, S.R. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society, 51* (8), pp. 993-998.

Emonet, B. (2000). "Revisiting Statistical Applications in Soccer". *STS Report*, Ecole Polytechnique Federale de Lausanne.

Famoye, F. & Singh, K.P. (2006). Zero-inflated generalized Poisson model with an application to domestic violence data. *Journal of Data Science. 4*(1): 117-130.

Gembris, D., Taylor, J.G. & Suter, D., (2002). Sports statistics - Trends and random fluctuations in athletics, *Nature 417*.

Greenhough, J., Birch, P.C., Chapman, S.C. & Rowlands, G. (2002). Football goal distributions and extremal statistics. *Physica A: Statistical Mechanics and its Applications, 316* (1-4), pp. 615-624.

Heuer, A. & Rubnes, O. (2009). Fitness, chance, and myths: an objective view on soccer results. *The European Physical Journal B. Vol. 67*, pp. 445-458.

Heuer, A., Müller, C. & Rubner, O. (2010). Soccer: Is scoring goals a predictable Poissonian process? EPL (Europhysics Letters), 89 (3).

Heuer A, Rubner O (2012) How Does the Past of a Soccer Match Influence Its Future? *Concepts and Statistical Analysis*. PLoS ONE 7(11): e47678.

Jamieson, J. P. (2010). The Home Field Advantage in Athletics: A Meta-Analysis. *Journal of Applied Social Psychology, 40*(7), 1819–1848.

Karlis, D. & Ntzoufras, I. (1998), "Statistical Modelling for Soccer Games: The Greek League," stat-athens.aueb.gr/~jbn/tr/TR59_Greek_Soccer.ps, 23 April 2009.

Karlis D. & Ntzoufras J. (2000). On modelling soccer data. *Student 3*, 229-245.

Karlis D. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician, 52, Part 3*, 2003, 381-393.

Karlis, D. & Ntzoufras, I., (2005). Bivariate Poisson and diagonal inflated Poisson regression models in R. *Journal of Statistical Software 14*(10).

Karlis, D. & Tsiamyrtzis, P. (2008). Exact Bayesian modeling for bivariate Poisson data and extensions. *Statistics and Computing*, 18 (1), pp. 27-40.

Karlis, D. & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference. *IMA Journal Management Mathematics, 20* (2), pp. 133-145.

Koopman, Siem Jan and Lit, Rutger (2012). A Dynamic Bivariate Poisson Model for Analysing and Forecasting Match Results in the English Premier League (September 24, 2012). Tinbergen Institute Discussion Paper 12-099/III. Disponible en SSRN:http://ssrn.com/abstract=2154792.

Kocherlakota, S., Kocherlakota, K. (2001). Regression in the bivariate Poisson distribution. *Communs Statist. Theory Meth., 30*, pp. 815-827.

Maher M.J. (1982), Modelling Association Football scores. Statistica Neerlandica, 36, 109-118.

Malcata, R. M., Hopkins, W. G., & Richardson, S. (2012). Modelling the progression of competitive performance of an academy's soccer teams. *Journal of Sports Science and Medicine, 11*(3), 533-536.

Mosteller, F. (1997). Lessons from Sports Statistics. *American Statistician, 51* (4), pp. 305-310.

Li, C.S., Lu, J.C., Park, J., Kim, K., Peterson, J. (1999). Multivariate zero-inflated Poisson models and their applications. *Technometrics, 41*, pp. 29-38.

Lago-Peñas, C. & Lago-Ballesteros, J. (2011). Game location and team quality effects on performance profiles in professional soccer. *Journal of Sport Science and Medicines. 10*(3) 465-471.

Meng. X. L., Dyk, D. (1997). The EM algorithm— an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society. B* 59:511-567

Panaretos, V. (2002). "A statistical analysis of the European soccer Champions League", *Proc. Joint Statistics Meeting.*

Rohatgi, V.K. (1976). An introduction to probability theory and mathematical statistics. New York: Wiley.

Rue H. and Salvesen Ø. (1999) Predicting and retrospective analysis of soccer matches in a league. *Technical Report.* Norwegian University of Science and Technology, Trondheim.

Saavedra García, M., Gutiérrez Aguilar, O., Fernández Romero, J.J., Paulo Sá Marques, P. (2012).Ventaja de jugar en casa en el fútbol español (1928-2011). Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte.In press,pp. 1 - 14. Disponible en: http://cdeporte.rediris.es/revista/inpress/artaceptados.htm. ISSN 1577-0354

Skinner, G. & Freeman,G. (2009). Are soccer matches badly designed experiments?, *Journal of Applied Statistics, 36*, 1087.

Sumathi K and Rao A K (2009), "On estimation and tests for zero inflated regression models", INTERSTAT.

Tusell, F. (2008) Análisis de Regresión. Introducción teórica y práctica basada en R. Disponible en http://www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/estad3/nreg1.pdf

Vlastakis, N., Dotsis, G. & Markellos, R. (2008). Nonlinear modelling of European football scores using support vector machines. *Applied Economics, 40* (1), pp. 111-118.

Vernic, R. (1997) On the bivariate generalized Poisson distribution. *A S T I N Bulletbl 27,* 23-31.

Volf, P. (2009) A random point process model for score in matches. *IMA J. Management Mathematics 20*, 121-131.

Walhin, J. (2001). Bivariate ZIP models, *Biometrical Journal, 43*, 147-160.

Waters, A., & Lovell, G. (2002). An Examination of the Homefield Advantage in a Professional English Soccer Team from a Psychological Standpoint. *Football Studies, 5*(1), 46-59.

Weinberg, R. S., & Gould, D. (2007). *Foundations of Sport and Exercise Psychology* (4th ed.). Champaign, IL: Human Kinetics.

Wulu, J. T., Singh, K. P. Famoye, F. and McGwin, G. (2002). Regression analysis of count data. *Journal of the Indian Society of Agricultural Statistics 55,* 220-231.

Zeileis, Achim and Kleiber, Christian and Jackman, Simon (2007) Regression Models for Count Data in R. Research Report Series / Department of Statistics and Mathematics, 53. Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna.