Postworks 1-4 (Bloque 2)

Escamilla Zepeda, Mario
Gómez Barraza, Karla Daniela
Guevara Moreno, Fernando
Muñoz Esparza, José Cruz
Rodríguez Rivera, Gil Estéfano
Rosado Martínez, Ana Elizabeth

Lo que hace el programa del **Postwork 4** es tomar tres bases de datos con información sobre la Liga Española de Primera División de las **temporadas 17/18**, **18/19** y **19/20**. Estas bases de datos cuentan con la información de todos los partidos jugados en las respectivas temporadas. Por ejemplo, tiene información sobre la fecha del partido, el equipo que jugó como local así como el que jugó de visitante, el marcador y algunos otros datos que no son de tanto interés para este estudio.

Los datos que son de interés en este trabajo son los **marcadores**, es decir, los **goles** que anotó el **equipo local** y los que metió el **equipo visitante** en cada partido, sin importar la fecha ni los equipos específicos. Las columnas en con esos datos se llaman ***FTHG*** y ***FTAG***, respectivamente. Supondremos que la probabilidad de cada marcador es independiente de la temporada que se juega (así como se ha despreciado el equipo y la fecha), por lo que, en aras de tener más datos con los cuales trabajar, se juntan las bases de datos de las tres temporadas enunciadas luego de haber escogido sólo las dos columnas citadas anteriormente (FTHG y FTAG) en cada una de las bases individuales.

Con la nueva base de datos, compuesta por dos columnas, se construye una tabla de la frecuencia de cada marcador en toda la base de datos con el comando ***table***. Esta tabla de frecuencias sirve para construir una tabla pero con probabilidades conjuntas, misma que se obtiene al dividir cada entrada entre el total de datos de la base. Por otra parte, es posible obtener las probabilidades marginales de cada variable (FTHG y FTAG) de diversas maneras; en el programa se obtuvo al agrupar la variable adecuada en la base de datos y luego dividir la frecuencia con la que aparece la variable estudiada -independientemente del valor de la otra- entre el total de datos). En el programa se pueden visualizar todas las distribuciones de probabilidad explicadas (la conjunta y las dos marginales).

Dos **variables aleatorias X y Y son independientes** dada una **distribución de probabilidad P** si su la distribución conjunta es igual al producto de las distribuciones marginales de cada variable. Esto es:

$$P(x = X, y = Y) = P_x(x = X)P_y(y = Y)$$

Donde $P_x(x = X) = \sum_y P(x = X, y = Y)$ y ocurre algo semejante para $P_y(y = Y)$, pero con una suma sobre los valores posibles de X.

De la definición anterior, se puede ver que, al dividir las probabilidades marginales en ambos lados, se obtiene un uno del lado derecho de la igualdad. Este valor puede utilizarse para estudiar en qué condiciones (y porqué) las variables FTHG y FTAG son independientes.

La segunda parte del programa tiene como objetivo determinar las **condiciones de independencia** (es decir, los marcadores en los que se puede decir que los goles del local **no dependen** estadísticamente de los goles del visitante y viceversa). Esto se hace dividiendo la tabla de las probabilidades conjuntas entre el producto de sus respectivas probabilidades marginales. Esa tabla da una idea de *cuán cercano* está cada entrada de la tabla de la independencia de variables y, si bien, puede ser una buena estimación, no es del todo certera.

Para dar certeza a la inferencia, se necesita de una desviación estándar, sin embargo, la tabla de cocientes no cuenta con una. Esta situación se puede sortear por medio del **método bootstrap**, es decir, al formar submuestras al azar a partir de la muestra original. Como nota para el método, es importante notar que, aunque se recomienda tener muestras grandes para poder tener estadísticos confiables, es importante tener en cuenta también el tamaño de la muestra a partir de la cual se aplica bootstrap: esto es para evitar crear muestras con datos repetidos.

Con cada submuestra se repite el proceso de calcular las distribuciones de probabilidad conjunta y marginales para luego conseguir una tabla de coeficientes como los descritos anteriormente para evaluar la independencia de las variables. Una vez que se tiene una tabla de coeficientes para cada submuestra, se puede calcular el promedio y la desviación estándar de cada elemento de la tabla. Esta tabla de desviaciones estándar S es útil para realizar una prueba de hipótesis.

Para la prueba de hipótesis se usará la tabla de coeficientes de la muestra original M en lugar del promedio de las submuestras porque, de acuerdo a la construcción que se hizo en el método bootstrap, la muestra original funge a modo de una -*pseudo*-población y las submuestras fungen como muestras de la población. Expresado de otro modo, es preferible usar la media poblacional en lugar de la media muestral.

La **hipótesis nula** H_0 es que las variables son independientes, es decir, que los coeficientes tienen valor 1. La prueba de hipótesis, entonces se hará, a modo de tabla (donde es claro que la expresión sólo tiene sentido hablando de entrada por entrada, y no del objeto completo, que es como opera R), de la siguiente manera:

$$Z = (1 - M)/(S/\sqrt{N})$$

Donde el 1 viene de la hipótesis nula y N es el número de submuestras utilizadas. Este estadístico de prueba se contrasta con una significancia bilateral del 5%, pues la negación de la hipótesis nula es que los coeficientes de M no sean iguales a 1, independientemente de si son mayores o menores.

Luego de la prueba, se obtienen los marcadores para los que las variables FTHG y FTAG son independientes. Estos marcadores pertenecen ambos al conjunto {0,1,2,3} además de algunas combinaciones de valores colindantes a esas celdas (si se organizara como una tabla).

Si se ignoran los valores colindantes (que pueden suponerse que, con una muestra más grande, fallarían la prueba) y la atención se centra en el *conjunto principal*, puede verse que es la región con resultados más frecuentes. Es decir, independientemente del talento de un equipo, es razonable que un equipo de menor categoría le meta un gol a uno de mayor clase, mientras que es posible que el segundo le meta uno o dos al primero. Por eso, en esas regiones los goles del visitante y los del local son independientes. Por el contrario, note la independencia no se cumple cuando se trata de marcadores más *exóticos*. Esto puede verse desde dos aristas, la primera es que son resultados suficientemente raros como para poder estudiarse adecuadamente; la segunda interpretación es que, por ejemplo, es improbable que un equipo bueno le meta 8 goles a otro mientras que el otro equipo le meta 1 o 2, aún siendo bueno: eso sólo es concebible -por lo general- cuando los dos equipos tienen un nivel desproporcionado, que ya no es un *estado independiente* de los equipos que jueguen y, por lo tanto, ya no es un marcador en el que las variables sean independientes. En casos más extremos, como un marcador de 8 a 6, sencillamente es improbable que dos equipos profesionales lleguen a esos marcadores, salvo que el portero se lesione u ocurra otro factor extraordinario.