TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi Economètrica dels Valors de Mercat Futbolístic Europeu de la temporada 2017/2018

Autor: Miquel Sastre Belío

Director: Montserrat Guillen i Estany

Departament: Departament d'Econometria, Estadística i Economia Aplicada

Convocatòria: Juliol 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

Els objectius d'aquest treball són construir un model de regressió lineal múltiple per tal de determinar el valor de mercat dels futbolistes professionals a partir de mètriques de rendiment futbolístic i construir un model de predicció de gol que servirà com a base de l'*Expected Goals Method*. Les dades que s'han utilitzat fan referència a les 5 lligues europees principals (lligues anglesa, espanyola, alemanya, italiana i francesa) de la temporada 2017/2018 proporcionades per l'empresa *Wyscout* i la web *Transfermarkt*.

La part inicial del treball està formada per la introducció, la revisió de la literatura i l'apartat de metodologia i dades. En aquests apartats introductoris es posa en context al lector i es justifica el tema, es revisen els treballs més recents d'altres autors i s'explica l'estructura de les dades per a la realització del treball i la seva extracció, i la metodologia emprada.

Arribats a aquest punt, en la primera part dels resultats es construeix el model de predicció de gol, on s'avaluen un total de 40.461 xuts. S'ha trobat evidència estadística que la distància i l'angle de tir, el rol del jugador, la cama o part del cos amb la qual s'efectuï el tir, que la jugada procedeixi d'un contra-atac o d'una centrada són factors que determinen la probabilitat que un xut acabi essent gol. Finalment, pel que fa a la segona part, s'ha mesurat el valor de mercat de 2.662 jugadors, on s'ha fet una distinció segons el rol de joc (davanters, mig-campistes, defenses i porters). Per a cadascun d'aquests s'ha pogut elaborar un model de regressió per estimació robusta on les variables explicatives són mètriques de rendiment futbolístic específiques depenent del rol que es tracti.

PARAULES CLAU

Regressió lineal múltiple, regressió logística, logaritme, mínims quadrats ordinaris, regressió robusta, correcció de White, futbol, valor de mercat, rol del jugador.

CLASSIFIACIÓ AMS

62J05 Linear regression
62J12 Generalized linear models
62P20 Applications to economics

TITLE

Econometric Analysis of European Football Market Values for the 2017/2018 season

SUMMARY

The two main goals of this work are to build a multiple linear regression model in order to determine the market value of professional football players from football performance metrics and to build a goal prediction model that will be the basis for the Expected Goals Method. The data used refers to the 5 main European leagues (English, Spanish, German, Italian and French leagues) for the 2017/2018 season provided by the company *Wyscout* and the *Transfermarkt* website.

The initial part of the work is structured as follows: the introduction, the review of the literature and the methodology and data section. In these introductory sections the reader is contextualised, the topic is justified and the most recent works of other authors are reviewed. The structure of the data for the realization of the work and its extraction is presented, as well as the methodology used.

At this point, in the first part of the results section, the goal prediction model is built, where a total of 40.461 shots are evaluated. Statistical evidence shows that the distance and angle of the shot, player's role, the leg or part of the body and whether the action comes from a counterattack or from a crossing pass are factors that establish the likelihood that a shot will end up being a goal. Finally, the market value of 2.662 players is measured, where a distinction is made according to the player's role (forwards, midfielders, defenders and goalkeepers). For each of these, a robust estimation regression model has been developed where the explanatory variables are specific football performance metrics depending on the player's role.

KEY WORDS

Multiple linear regression, logistic regression, logarithm, ordinary least squares, robust regression, White's estimators, football, market value, player's role.

ÍNDEX

I. INTRODUCCIÓ 5
II. REVISIÓ DE LA LITERATURA 9
III. MATERIAL I MÈTODES 11
1. Dades 11
2. Mètriques i eines d'anàlisi 16
3. Models 18
4. Recursos informàtics 22
IV. RESULTATS 23
1. Model de gols esperats (*xGoals model*) 23
2. Anàlisi economètrica dels valors de mercat 44
V. CONCLUSIONS 91
VI. BIBLIOGRAFIA I WEBGRAFIA 96

I. INTRODUCCIÓ

Arreu del món existeixen prop de 250 esports reconeguts, dels quals es poden destacar la natació, el tennis o el basquetbol. Però, el que ressalta per damunt de tots i el que s'ha merescut el renom d'*esport rei* és el futbol. Amb aproximadament 4.000 milions de seguidors a tot el món, el futbol és l'esport on dos equips formats per onze jugadors cadascun que corren darrera d'una pilota amb l'objectiu final de ficar-la dins una porteria en més ocasions que el rival -és a dir, marcar gol- que és capaç de mobilitzar a més de la meitat de la població mundial, havent esdevingut no només una manera alternativa de fer esport, sinó en un motor social i econòmic de gran impacte sobre la societat.

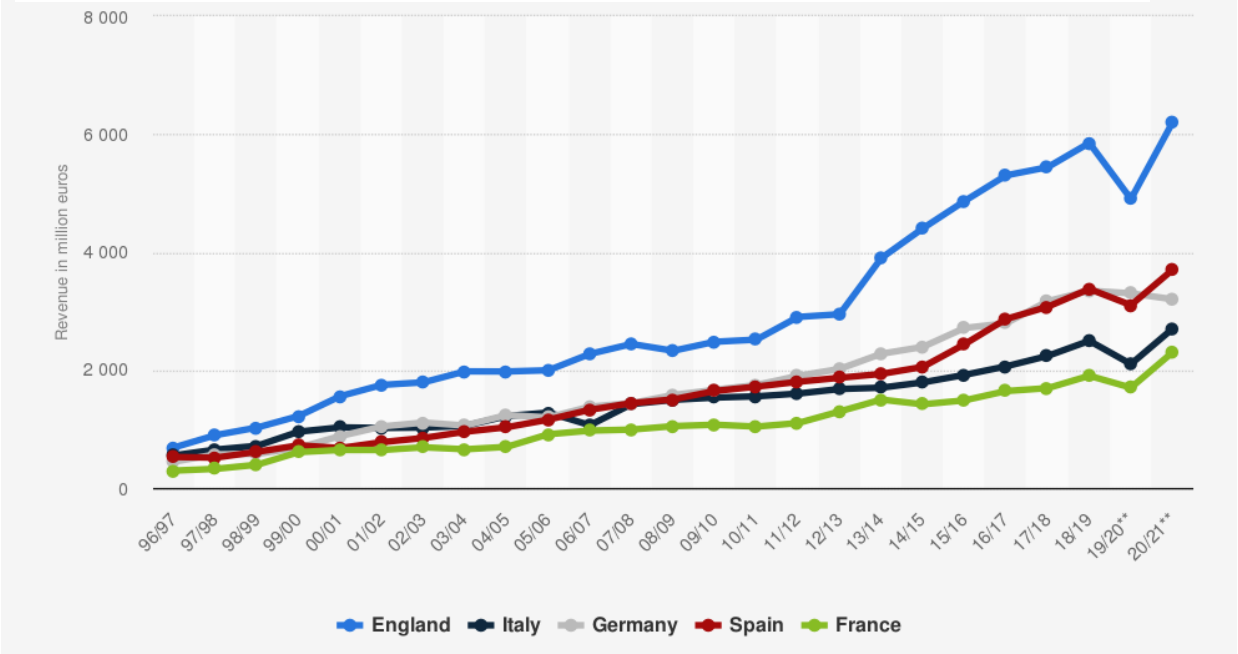
En particular, segons l'últim estudi independent de la consultora *PriceswaterhouseCooper* (PwC), on s'avalua l'impacte socioeconòmic de la indústria del futbol professional a Espanya en la temporada 2016/2017, elaborat per a *LaLiga*, organització que inclou els 20 clubs de futbol de la màxima competició i els 22 clubs de la segona màxima competició de futbol a nivell professional d'Espanya, els ingressos generats per l'esmentada indústria ascendeixen a 15.688 milions d'euros, equivalents a 1,37% del PIB; a més a més, en termes d'ocupació, la contribució del futbol professional va ser de 184.626 llocs de treball, xifra que correspon gairebé l'1% de les persones ocupades a Espanya. També, es destaca la contribució tributària de *LaLiga*, que va recaptar 4.089 milions d'euros, i l'impacte positiu sobre les relacions familiars i socials que té el futbol professional, transmetent valors tals com la responsabilitat, la integritat, la companyonia, el respecte o l'esportivitat.

Per altra banda, l'informe anual que publica, en aquest cas, la consultora *Deloitte* des de l'any 2006*, Deloitte Football Money League* (*DFML*), és una bona mostra de l'acompliment financer i de l'impacte que indueixen els 20 clubs europeus de futbol que generen majors ingressos a nivell mundial. En aquest informe es fa un peritatge de la situació econòmica-financera de cada club en particular, contextualitzant cada cas acuradament. En concret, l'informe referent a l'actual temporada 2020/2021 assenyala, principalment, l'impacte negatiu que ha provocat la pandèmia de la COVID-19, reflectida en termes generals en una reducció dels ingressos dels 20 equips de 1.100 milions d'euros, un 12% sobre la totalitat de la temporada passada, provocada majoritàriament per la caiguda dels ingressos per retransmissions. Tanmateix, a la llista que es publica dels 20 equips més poderosos segons els ingressos totals anuals hi destaca per sobre de tots el FC Barcelona (715,1 milions d'euros), seguit del Real Madrid (714,9 milions d'euros) i el FC Bayern Munich (634,1 milions d'euros).

Així doncs, és evident que la indústria futbolística ha transcendit, sobretot l'europea, i actualment és considerada una indústria rendible on els seus agents principals, clubs i jugadors, es beneficien de les grans xifres de diners que es mouen i es generen. No obstant, aquesta indústria no sempre ha acaparat la mateixa atenció i no ha tingut el mateix funcionament.

Pel que fa els clubs europeus, l'augment dels ingressos per la concessió dels drets televisius amb la creació de la *Lliga de Campions* de la *UEFA* als anys 90 va suposar un punt d'inflexió en el seu poder financer. A part, fins aquell moment l'impacte de la majoria de clubs era

Figura 1. Evolució dels ingressos de les 5 grans lligues de futbol europees (1996/1997 - 2020/2021)



Font: Deloitte. Statista 2020

d'àmbit nacional i la nova competició a nivell continental va permetre que molts clubs es donessin a conèixer. Ara bé, aquells que ja eren considerats clubs transnacionals també van adquirir encara més reconeixement, la qual cosa en certa manera va donar lloc al començament d'una certa divergència entre clubs i la creació d'un grup d'elit format per les entitats més poderoses. Conseqüentment, els salaris dels jugadors així com els preus de traspàs es veuen augmentats considerablement. Un bon exemple d'això és que el Real Madrid bat el rècord de traspàs més car per un jugador durant dos anys consecutius: l'any 2000 es paga pel jugador Luís Figo al voltant de 60 milions d'euros i l'any 2001 es traspassa Zinedine Zidane per la xifra de 72 milions d'euros. Actualment, el rècord de traspàs més car de la història el té el club parisenc *Paris Saint-Germain*, el qual va pagar al FC Barcelona 222 milions d'euros pel jugador brasiler Neymar da Silva Santos Júnior, conegut com Neymar.

Per altra banda, també als anys 90 es produeix un fet important que canviaria les circumstàncies laborals i contractuals dels jugadors. L'any 1995 el Tribunal de Justícia de la Unió Europea emet una sentència que permet als jugadors de futbol de la Unió Europea canviar de club al finalitzar el seu contracte sense l'obligació d'haver-se d'efectuar cap compensació econòmica. I és que fins aquell moment, tant si el jugador es trobava amb un contracte vigent com si no, el club que volgués adquirir els seus serveis havia de pagar un preu pel seu traspàs[1]. Aquesta sentència va rebre el nom de *Llei Bosman* atès que es deriva del cas particular que va patir el jugador belga Jean-Marc Bosman. Així, a causa de l'anterior juntament amb l'eliminació de restriccions que limitaven el nombre de jugadors estrangers per equip, o no nacionals, el mercat futbolístic va iniciar un procés de liberalització i de nova

[1] No vol dir que no hi hagués traspassos entre els equips europeus. Sí que es produïen i la qüestió és que el club comprador havia de pagar una quantitat de diners al club venedor independentment si el jugador comprat es trobava amb un contracte vigent o no.

regularització. Així és que ha sigut possible l'accés a informació, a dades referents a salaris de jugadors, preus de transferència per traspàs, etc. La qual cosa ha provocat un augment de l'interès acadèmic des de l'inici del segle XXI, especialment per tractar d'esbrinar, per exemple, quins són els fonaments que determinen els salaris dels jugadors o el seu preu de traspàs, o quines són les característiques del mercat futbolístic.

Precisament, el primer dels dos objectius principals d'aquest treball és constatar i valorar si es pot predir el valor de mercat dels jugadors de futbol per mitjà del seu rendiment esportiu de l'última temporada jugada, junt amb altres factors externs, essent aquest últim l'element més rellevant. L'anàlisi que es farà és una anàlisi economètrica amb un model de regressió estimat per mínims quadrats ordinaris sobre un total de 2.791 jugadors, on la variable endògena serà el seu valor de mercat extret de la web *Transfermarkt*. A més, una altra hipòtesi sobre la qual resideix el present escrit és que per a cada categoria de rol a la qual pot pertànyer un jugador (es fa distinció entre porter, defensa, mig-campista i davanter) s'atribueix un mercat específic definit per mètriques de rendiment diferents. És evident que a un jugador que jugui a la posició de porter no se'l valorarà per les mateixes accions que un davanter. Mentre que la principal missió del primer és defensar la porteria de l'atac rival, l'objectiu elemental del davanter és marcar gols. Per tant, el rendiment dels rols de porter i davanter no es poden mesurar de la mateixa manera, cadascun es defineix per estadístiques o mètriques diferents. Aquest és el cas més extrem que es pot donar en aquest esport, però pel que fa a la resta de posicions també presenten característiques específiques que marquen les diferències entre una i altra. És en el següent apartat on es farà una revisió de la literatura existent, però cal avançar que en cap de les anàlisis examinades s'ha fet una singularització d'aquest tipus.

El segon objectiu és el de construir i presentar una eina estadística contemporània que posa de manifest una nova manera de valoració del rendiment futbolístic, que cada vegada s'està utilitzant més pels clubs atès que ja hi ha evidències de resultats exitosos. El *Mètode dels Gols Esperats*, o *Expected Goals Method* en anglès, es podria dir que és el resultat de la combinació de gestió de bases de dades, intel·ligència artificial i estadística, amb la qual s'apunta cap a una nova direcció en termes de comprensió del joc i amb la que s'intenta deixar de banda els tòpics que sempre han estat presents al futbol i eliminar el *soroll* que sempre l'ha envoltat. Concretament, en aquest treball es farà ús dels models de resposta binària per abastar aquest tema. A més a més, es mostraran exemples de la seva aplicació amb l'ús de gràfics que complementarien l'anàlisi.

A nivell personal, m'agradaria dir que haver pogut combinar els meus estudis d'economia i estadística amb la meva passió, el futbol, ha sigut la meva motivació principal per a dur a terme aquest treball final de grau d'economia i estadística. De la mateixa manera, penso que la contemporaneïtat i l'aplicabilitat del cas fan el treball una mica més atractiu i original.

Una altra motivació que m'agradaria destacar és la possibilitat que m'ha permès aquest treball de donar una visió objectiva a un esport que, sota el meu parer, ha estat sempre envoltat d'especulació i de molts prejudicis.

Un cop presentats tant el tema com els objectius i hipòtesis del present escrit, a continuació es proposa l'estructura que tindrà per tal d'assolir un correcte desenvolupament. Després

d'actual introducció, es presentarà una revisió de la literatura relacionada, la qual es centrarà en aquells autors que han tractat de determinar els factors explicatius tant dels salaris com dels preus de traspàs dels jugadors de futbol. Seguidament, en l'apartat de *Material i mètodes* s'exposarà, per una banda, les dades que s'han utilitzat per elaborar l'estudi, així com els passos que s'han hagut de seguir per obtenir l'estructura d'aquestes de la manera més adequada; per l'altra, es farà una explicació del mètode que es segueix, tant del model de regressió lineal múltiple com el de resposta binària. Arribats a aquest punt, ja s'haurà posat en context al lector i tot seguit es procedirà a presentar les dues anàlisis elaborades juntament amb els resultats obtinguts. Finalment, el text acabarà amb l'apartat de *Conclusions* on es valorarà el treball en la seva totalitat.

II. REVISIÓ DE LA LITERATURA

Com ja s'ha comentat en l'apartat anterior, l'apertura d'accés a dades referents a salaris, preus de traspàs dels jugadors, durada dels contractes, etc. a partir dels anys 90 del segle passat va provocar l'inici per l'interès acadèmic per la indústria del futbol en l'àmbit econòmic i va ser quan es van començar a veure publicacions fins a l'actualitat. Frick (2007) recull i revisa els estudis realitzats, fins el moment de la seva publicació, sobre les diverses dimensions que conformen el mercat laboral de la indústria futbolística i sobre les característiques que determinen els salaris i preus de traspàs dels futbolistes, on es conclou que la variància tant dels salaris com dels preus de traspàs dels jugadors comparteixen un marc teòric compost per característiques dels jugadors i dels clubs.

Així mateix, Frick (2011) mitjançant models d'efectes aleatoris, regressió robusta estimada per mínims quadrats ordinaris, regressió de mediana i regressió quantílica analitza, per una banda, si la diversitat de remuneració entre jugadors es pot explicar per les diferències de rendiment individual i la capacitat de pagar que tenen els clubs i, per una altra, si existeixen diferències de comportament entre jugadors segons la durada del seu contracte vigent, utilitzant dos panels no balancejats de jugadors referents a la lliga alemanya un dels quals fa referència a 6 temporades consecutives i, l'altre, a 13 temporades. En aquest cas, s'obté per a totes les anàlisis executades que el salari d'un jugador pot ser explicat pel número de partits jugats al llarg de la seva carrera, les aparicions en partits internacionals, els gols marcats, el rol del jugador, el factor *lideratge* i la regió de naixement. A més, pel que fa a l'anàlisi que es centra en les durades dels contractes, conclou que el rendiment d'un jugador[2] incrementa significativament l'últim any de contracte i que la variància del rendiment dels jugadors és significativament més baixa, des d'un punt de vista estadístic, l'últim any de contracte. Així, aquest autor evidencia l'existència del concepte de *moral hazard* al futbol.

Altrament, Lee i Harris (2012) porten a terme un estudi semblant per investigar quins són els factors que contribueixen a la fixació dels salaris de la *Major League Soccer* (MLS), la lliga de futbol estatunidenca. Aquesta anàlisi té una particularitat i és que els equips de la lliga de futbol nord-americana tenen límits salarials inferiors i superiors. Conseqüentment, un dels mètodes que prenen com a eina és la regressió Tobit. Els resultats que s'obtenen són que el nombre de gols, assistències i minuts jugats són variables que determinen els salaris del jugadors, de manera que tenen un efecte positiu; contràriament, l'efecte del nombre de partits jugats es conclou que és dubtós ja que els coeficients associats a aquesta covariable són contraris en les diferents anàlisis que es fan. També, es destaca que els defenses reben salaris menors que els mig-campistes.

D'altra banda, Deutscher i Büschemann (2014) utilitzen dades des de la temporada 2005/2006 fins la del 2009/2010 de la Bundesliga per analitzar com afecta la consistència del rendiment dels jugadors sobre els salaris, entesa com el coeficient de variació. El seu estudi amb regressió estimada per MQO i per regressió quantílica conclou que els efectes de les variables edat, nombre de partits jugats a la Bundesliga, el rol del jugador, el nombre d'aparicions amb l'equip nacional i l'avaluació del rendiment són estadísticament significatius,

[2] Es pren com a referència la qualificació que fa la revista *Kicker*.

així com l'efecte del coeficient de variació o *consistència*. Però, contra pronòstic dels autors, el signe del coeficient associat a l'última variable s'obté negatiu, i això significa que quanta major inconsistència, major serà el salari del jugador, de manera que la inconsistència o la característica d'un jugador de ser impredictible es remunera monetàriament.

Yaldo i Shamir (2017), per altra banda, proveeixen d'un mètode quantitatiu d'estimació del salari dels jugadors de futbol basat en les seves habilitats al terreny de joc. Utilitzen algoritmes de *Machine Learning* (*Additive Regression*, *Decision Table*, *Nearest Neighbor with a weighted condition, K**, *Locally Weighted Learning with Naive Bayes and Linear Regression classifiers*, *Random Comittee*, *Random Trees*) i la base de dades recull informació de les habilitats de cadascun dels jugadors segons l'actuació de la temporada $2016/2017$, la qual s'utilitzaria pel videojoc *FIFA*, cadascuna de les quals té un domini comprès entre el 0 i el 99, essent 0 el valor més baix i 99 el valor més alt que pot tenir un jugador respecte a una habilitat determinada. Segons els resultats obtinguts, tots els algoritmes provats mostren correlació estadísticament significativa entre els valors predits i els reals. També, les habilitats més rellevants obtingudes són: capacitat de reacció, finalització, capacitat de robar una pilota, posicionament, control de pilota, entrada[3], intercepció, habilitat amb el cap, passada de llarga distància i visió de joc. A més a més, es dedueixen diferències entre lligues pel que fa al salaris, essent la *Premier League* la lliga on hi ha salaris més alts, i diferències entre rols pel que fa a les habilitats que determinen el salari.

Més recentment, García-del-Barrio i Pujol (2020) avaluen la contribució i el valor econòmic dels jugadors de futbol mitjançant mètriques *on-field* i *off-field*, és a dir, característiques de rendiment futbolístic i altres factors externs que no tenen a veure directament amb el rendiment dins del camp, com per exemple el grau de visibilitat mediàtica o l'estatus que té el jugador dins l'equip. En el seu estudi s'analitzen 5000 jugadors de més de 200 clubs. Les variables explicatives més importants del valor econòmic dels futbolistes són la visibilitat mediàtica, la duració del contracte, els anys d'experiència, el nivell de visibilitat mediàtica del jugador dins del seu equip, l'estatus de l'equip on juga, l'edat, l'edat del jugador al final del seu contracte i la lliga de l'equip en el qual juga.

[3] *Entrada* es refereix a una manera determinada d'intentar robar una pilota mitjançant el lliscament del jugador pel terra.

III. MATERIAL I MÈTODES

A continuació, es presenta el mètode que s'ha seguit i el material del qual s'ha disposat per a la realització del treball. En primer lloc, es descriurà l'extracció i la depuració de les dades. Seguidament, es farà el llistat de les mètriques de rendiment futbolístic, amb la seva definició corresponent, que s'han utilitzat per ambdues anàlisis. Així mateix, es definiran i justificaran els models que s'han usat per a l'anàlisi economètrica, juntament amb les mesures de validació apropiades. Finalment, es farà menció dels recursos informàtics que s'han utilitzat per a la realització d'aquest treball.

1. Dades

Per una banda, les dades que s'han utilitzat en aquest treball són un seguit de *datasets* lliures proporcionats per l'empresa d'anàlisi futbolística *Wyscout*, la qual els va posar a disposició pública durant la realització del *Soccer Data Challenge* l'any 2019. El conjunt de taules fa referència a la temporada jugada els anys 2017 i 2018 de les cinc competicions europees de futbol professional de referència: la lligues anglesa, espanyola, alemanya, italiana i francesa, les quals reben els noms de *Premier League, LaLiga, Bundesliga, Serie A* i *Ligue 1,* respectivament. Per a cadascuna de les anteriors, i serà d'on s'extraurà el gruix de la futura anàlisi, es disposa d'una taula la qual inclou totes les accions espai-temporals (passades, xuts a porteria, duels entre jugadors, etc.) de tots els partits, així com els jugadors que les han realitzat i amb anotacions o *tags* associades a cadascuna de les accions aportant característiques i qualitats de les pròpies accions. A més, també es disposa de les metadades tant dels equips com dels jugadors.

Per altra banda, per a dur a terme l'anàlisi econòmica s'han inclòs dues taules de dades les quals contenen, per un lloc, el valor de mercat dels jugadors i, per altra, el valor de mercat dels equips que conformaven les lligues esmentades anteriorment relatiu a la temporada dels anys 2017 i 2018. La seva extracció s'ha fet a partir de la web *Transfermarkt* la qual, val a dir, ha consistit en copiar i enganxar una per una les plantilles de cada equip ja que no hi havia cap altra alternativa millor. Finalment, pel que fa a la taula dels valors de mercat dels jugadors, s'ha hagut d'homogeneïtzar amb la taula de metadades referents als jugadors procedent de *Wyscout*, creant una taula definitiva que inclogués tant les metadades com el valor de mercat dels jugadors, anomenada *Pplayers*.

S'ha de tenir en compte que els valors de mercat que seran objecte d'anàlisi, per definició, són estimacions que *Transfermarkt* elabora. Aquestes estimacions són resultat de valorar molts factors que poden afectar el valor d'un jugador, i no tots són d'àmbit esportiu. Per exemple, el prestigi internacional, el propi sou, la jerarquia dins del vestuari o el potencial com a marca publicitària són factors que cada vegada més es tenen en compte a l'hora de mesurar el valor monetari d'un jugador. A més, com molt bé explica Tobias Blaseio, administrador a Espanya de la web *Transfermarkt*, a una entrevista del diari esportiu AS: "No s'ha de confondre valor de mercat amb preu final, ja que un jugador que quedi lliure pot canviar d'equip a cost zero, però el seu valor de mercat mai no serà zero". I afegeix: "No es valora la xifra màxima que un club pagaria per un fitxatge, sinó el que diversos clubs, en una pugna, estarien disposats a oferir". També, cal ressaltar i recordar que l'objectiu del treball és el de predir

aquests valors de mercat a partir del rendiment futbolístic de l'última temporada jugada com a part fonamental, mesurat a partir de les mètriques calculades pròpiament.

1) Base de dades

A la taula 3.1.1.1 es mostra el detall de les taules que conformen la base de dades amb la qual s'ha treballat.

Taula 3.1.1.1 Taules que formen la base de dades

Variable	Descripció
Events	
eventId	Número identificador del tipus d'acció general
subEventName	Nom del tipus d'acció en particular
tags_i, i ∈ {1,2,3,4,5,6}	Atribut o característica descriptiva de l'acció. Per a cada acció hi ha un màxim de 6 tags
playerId	Número identificador del jugador que ha executat l'acció
positions_p_i, p ∈ {x,y} , i ∈ {1,2}	Coordenades en l'espai de la posició de l'acció que es realitza.
matchId	Número identificador del partit
eventName	Nom del tipus d'acció en general
teamId	Número identificador de l'equip del jugador que realitza l'acció
matchPeriod	Període en que es realitza l'acció. Pot ocórrer a la 1a part o a la 2a part
eventSec	Temps instantani en segons en que es dona l'acció
subEventId	Número identificador del tipus d'acció en particular
id	Número identificador únic per a l'acció
Players	
weight	Pes en quilograms
firstName	Primer nom
middleName	Segon nom
lastName	Cognom
currentTeamId	Identificador de l'equip pel qual juga actualment
birthDate	Data de naixement en format *Y-m-d*
height	Altura en centímetres
wyId	Identificador del jugador
foot	Cama dominant
shortName	Àlies
currentNationalTeamId	Número identificador de l'equip nacional
passaportArea.name	Nom del país de nacionalitat
passaportArea.id	Número identificador del país de nacionalitat
passaportArea.alpha3code	Codi identificador del país de nacionalitat alpha3
passaportArea.alpha2code	Codi identificador del país de nacionalitat alpha2
role.code2	Codi identificador de la posició habitual al camp alpha2
role.code3	Codi identificador de la posició habitual al camp alpha3
role.name	Nom de la posició habitual al camp

birthArea.name	Nom del país de naixement
birthArea.id	Número identificador del país de naixement
birthArea.alpha3code	Codi identificador alpha3 del país de naixement
birthArea.aplha2code	Codi identificador alpha2 del país de naixement

Pprices	
Nom	Primer nom
Cognom1	Primer cognom
Cognom2	Segon Cognom
Cognom3	Tercer Cognom
BirthDate	Data de naixement en format *Y-m-d*
Price	Valor de mercat en la temporada 2017-2018
Lliga	Lliga a la qual competeix l'equip al qual pertany el jugador

Teams	
Variable	Descripció
city	Nom de la ciutat de l'equip
name	Nom de l'equip
wyId	Número identificador de l'equip
officialName	Nom oficial de l'equip
type	Tipus d'entitat
area.name	Nom del país al qual pertany l'equip
area.id	Número identificador del país
area.alpha3code	Codi identificador alpha3 del país
area.alpha2code	Codi identificador alpha2 del país

Competitions	
name	Nom de la competició
wyId	Número identificador de la competició
format	Tipus de competició
type	Tipus d'equips que hi competeixen
area.name	Nom del país on es juga la competició
area.id	Número identificador del país on es juga la competició
area.alpha3code	Codi identificador alpha3 del país on es juga la competició
area.alpha2code	Codi identificador alpha2 del país on es juga la competició

PlayersRank	
goalScored	Número de gols marcats pel jugador
playerankScore	Valor per a l'índex *rankScore* del jugador
matchId	Número identificador del partit
playerId	Número identificador del jugador
roleCluster	Nom de la posició on va jugar
minutesPlayed	Número de minuts jugats

Matches	
status	Estat del partit: si s'ha jugat o no
roundId	Número identificador de la jornada del partit
gameweek	Setmana en la qual es va jugar el partit
seasonId	Número identificador de la temporada
dateutc	Data en la qual es va jugar el partit
winner	Número identificador de l'equip guanyador del partit. En cas d'empat, s'escriu 0
venue	Nom de l'estadi en qual s'ha jugat el partit
wyId	Número identificador del partit
label	Marcador del resultat final del partit
dateutc	Data en la qual es va jugar el partit
referees	Dades dels àrbitres que han arbitrat el partit
duration	Temps de durada en minuts del partit
competitionId	Número identificador de la competició

Tags	
tag	Número identificador de l'atribut
Label	Nom identificador de l'atribut
Description	Descripció de l'atribut

Events Metadata	
event	Número identificador de l'acció general
subevent	Número identificador de l'acció particular
event_label	Nom identificador de l'acció general
subevent_label	Nom identificador de l'acció particular

Cprices	
Club	Nom del club
plantilla	Total de jugadors que formen l'equip
Edad	Edat mitjana de l'equip
Extranjeros	Número de jugadors estrangers
ValordeMercadoTotal	Valor de mercat del club com a suma dels preus dels jugadors que conformen la plantilla
ValordeMercado	Valor de mercat del club com a mitjana del preu per jugador de l'equip

Font: Elaboració pròpia

2) Procés d'homogeneïtzació de *Pplayers*

El procés d'unificació de la base de metadades dels jugadors, anomenada *Players*, amb la dels valors de mercat, anomenada *Pprices*, ha consistit en obtenir una taula on s'incloguessin tant les metadades dels jugadors com el valor de mercat. Inicialment, la seva homogeneïtzació presentava un obstacle evident: el número de jugadors presents a la taula *Pprices* és menor que el de la taula *Players*. Per tant, tenint en compte que els jugadors que interessa conservar són dels quals es disposa el valor de mercat per a poder fer l'anàlisi posterior (aquests es

troben a la taula *Pprices*), és d'esperar que la taula resultant sia de dimensions reduïdes. A més a més, la falta d'un nexe d'unió clar entre ambdues taules ha dificultat la seva fusió, la qual cosa ha generat diversos problemes que, finalment, han contribuït a una major reducció de les dimensions.

Retornant al procés d'unificació, aquest ha constat de tres etapes diferents. La primera ha consistit en articular les dues taules utilitzant com a vincle les dates de naixement dels jugadors. En primera instància, aquesta era l'única manera de poder vincular els jugadors d'una taula amb els de l'altra. El resultat ha estat el d'una taula conjunta on el número de files equival al número de coincidències entre jugadors segons la seva data de naixement, conservant també aquells que no han trobat la seva correspondència, i on el número de columnes equival a la suma total d'atributs menys una d'ambdues taules, les més importants de les quals seran les que fan referència als noms i cognoms dels jugadors ja que permetrà passar a la segona etapa.

A partir de la taula derivada de la primera fase, la segona etapa ha consistit en la separació entre aquelles files on els noms sí coincidissin i les que no en dues taules diferents. Aquest pas ha presentat diversos hàndicaps degut a la pròpia naturalesa de les dades. Per exemple, un cas on els noms del jugador coincideixen però que estan escrits de maneres diferents ha estat el del jugador Arturo Vidal. Havent aplicat el filtre correctament, per una banda, la taula que fa referència a aquelles observacions on el noms dels jugadors ha coincidit es manté i serà de la qual s'obtingui, finalment, la taula definitiva que s'anomenarà *Pplayers*; per altra banda, la taula que conserva aquelles observacions on no coincideixen els noms és la que s'utilitzarà a la tercera etapa.

Il·lustració 3.1.2.1 Imatge d'un dels casos on els noms no coincidien completament

	Nom <chr>	Cognom1 <chr>	Cognom2 <chr>	Cognom3 <chr>	BirthDate <date>	Price <int>	Lliga <fctr>	weight <int>	firstName <chr>	middleName <chr>	lastName <chr>	currentTeamId <chr>	height <int>	wyId <int>	foot <chr>
32	Arturo	Vidal	NA	NA	1987-05-22	35000000	Bundesliga	75	Arturo Erasmo		Vidal Pardo	2444	180	20475	right

Font: Elaboració pròpia

En la tercera i última etapa s'ha tractat de recuperar el màxim nombre d'observacions que no s'han conservat però que sí que haurien d'haver format part de la taula definitiva degut a diversos inconvenients o problemes de la pròpia naturalesa de les dades. El principal problema que s'ha detectat ha sigut que les dates de naixement d'alguns jugadors en la taula *Players* eren incorrectes. Com a conseqüència, s'han hagut de corregir 34 dates de naixement en total. Altrament, de la taula on coincidien els noms del jugadors també es van detectar alguns problemes. Concretament, es podria destacar el cas de dos jugadors bessons, Sven Bender i Lars Bender, data de naixement i cognoms dels quals coincideix i que, per tant, a l'hora d'aplicar el filtre no només coincidien amb ells mateixos sinó que també cadascun amb el seu germà.

Il·lustració 3.1.2.2 Imatge del cas especial dels bessons Sven i Lars Bender

Nom <chr>	Cogno... <chr>	Cogno... <chr>	Cogno... <chr>	BirthDate <date>	Price <int>	Lliga <fctr>	weight <int>	firstName <chr>	middleName <chr>	lastName <chr>	currentTeamId <chr>	height <int>	wyId <int>	foot <chr>	shortName <chr>
Sven	Bender	NA	NA	1989-04-27	15000000	Bundesliga	81	Lars		Bender	2446	185	14781	right	L. Bender
Sven	Bender	NA	NA	1989-04-27	15000000	Bundesliga	80	Sven		Bender	2446	186	14803	right	S. Bender
Lars	Bender	NA	NA	1989-04-27	13000000	Bundesliga	81	Lars		Bender	2446	185	14781	right	L. Bender
Lars	Bender	NA	NA	1989-04-27	13000000	Bundesliga	80	Sven		Bender	2446	186	14803	right	S. Bender

Font: Elaboració pròpia

Així doncs, el resultat definitiu d'aquest procés ha estat el d'una taula de 2791 jugadors, que ha suposat una reducció del 22,53% respecte dels 3603 jugadors potencialment analitzables. Tanmateix, a la taula 3.1.2.1 es mostra la distribució dels jugadors disponibles segons la lliga a la qual pertanyen. No es van donar coincidències de jugadors nascuts el mateix dia i el mateix any, a part dels dos bessons ja esmentats.

Tabla 3.1.2.1 Distribució per lligues dels jugadors de la taula Pplayers

Lliga	Total
Bundesliga	513
LaLiga	520
Ligue 1	584
Premier League	553
Serie A	621

Font: Elaboració pròpia

2. Mètriques i eines d'anàlisi

1) Llistat d'indicadors de rendiment futbolístic

A la taula 2.1.1 s'exposa el llistat de mètriques de rendiment futbolístic calculades a partir de les taules descrites anteriorment que s'ha pogut recollir per a cadascun dels jugadors.

Tabla 2.1.1 Definicions de les variables utilitzades com a indicadors de rendiment futbolístic

Nom	Abreviatura	Descripció
Minuts	Min	Número total de minuts jugats
Partits	Partits	Número total de partits jugats
Gols	Gols	Número total de gols "de jugada" marcats
Gols a pilota parada	Gols_pp	Número de gols marcats a pilota parada (s'inclouen els gols de falta directa i els de penal)
Gols/90 minuts	Gols_90	Ràtio de gols cada 90 minuts
Gols esperats	xGols	Suma de les probabilitats de gol associades als tirs que ha efectuat el jugador
Xuts	Xuts	Número total de xuts
Gols/Xuts	GolsXuts_pc	Ràtio de Gols per cada xut
Penals	Penals	Número total de penals
Faltes	Faltes	Número total de faltes
Assistències	Assist	Número total d'assistències

Assistències/90 minuts	Assist_90	Ràtio d'assistències cada 90 minuts
Assistències esperades	xAssist	Suma de les probabilitats d'assistència associades a les passades que ha efectuat un jugador
Passades	Pass	Número total de passades
Passades correctes	Pass_ex	Número total de passades amb èxit
Percentatge de passades exitoses	Pass_ex_pc	Percentatge de passades amb èxit
Passades Intel·ligents	Pass_intl	Número total de passades creatives i penetrants les quals intenten trencar alguna línia defensiva per aconseguir una avantatge significant en atac
Passades intel·ligents exitoses	Pass_intl_ex	Número de passades creatives i penetrants amb èxit
Percentatge de passades intel·ligents exitoses	Pass_intl_ex_pc	Percentatge de passades creatives i penetrants amb èxit (respecte de les passades intel·ligents totals)
Passades penetrants	Pass_thro	Número de passades a l'espai darrere de la línia defensiva rival
Percentatge de passades penetrants	Pass_thro_pc	Percentatge de passades a l'espai (respecte el total de passades intel·ligents)
Centres a l'àrea	Cross	Número total de centrades a l'àrea
Centres a l'àrea exitosos	Cross_ex	Número de centrades a l'àrea amb èxit
Percentatge de centres a l'àrea exitosos	Cross_ex_pc	Percentatge de centres a l'àrea amb èxit
Driblatges	Dribl	Número total de driblatges intentats (un driblatge s'entén com superar a un rival mantenint la possessió de la pilota)
Driblatges exitosos	Dribl_ex	Número total de driblatges amb èxit
Percentatge de driblatges exitoses	Dribl_ex_pc	Percentatge de driblatges amb èxit
Duels	Duels	Número total de duels defensius disputats
Duels guanyats	Duels_ex	Número total de duels defensius amb èxit
Percentatge de duels guanyats	Duels_ex_pc	Percentatge de duels defensius amb èxit
Duels guanyats per anticipació	Duels_ex_antic	Número de duels defensius amb èxit per anticipació
Percentatge de duels guanyats per anticipació	Duels_ex_antic_pc	Percentatge de duels defensius amb èxit per anticipació respecte del total de duels defensius amb èxit
Duels aeris	Duels_aeris	Número total de duels aeris disputats
Duels aeris guanyats	Duels_aeris_ex	Número de duels aeris amb èxit
Percentatge de duels aeris guanyats	Duels_aeris_ex_pc	Percentatge de duels aeris amb èxit

Pilotes perdudes perilloses	Ppp	Número total de pilotes perdudes perilloses
Intervencions	Interv	Número total d'intervencions, enteses com el número de cops que ha hagut d'intervenir el jugador en el xut d'un jugador rival
Parades	Parades	Número total de parades
Reflexes	Reflx	Número total d'intervencions que han sigut per reflexes
Percentatge de parades	Parades_pc	Percentatge de parades
Percentatge de reflexes	Reflx_pc	Percentatge d'intervencions per reflexes
Percentatge de parades per reflexes	Parades_reflx_pc	Percentatge de parades que han sigut per reflexes
Participació en termes de minuts	Participacio_Min	Quocient dels minuts jugats pel jugador i el total de minuts que podria haver jugat al llarg de la temporada
Participació en termes de xuts	Participacio_Xuts	Quocient de xuts efectuats pel jugador i el número total de xuts efectuat per l'equip al llarg de tota la temporada
Participació en termes de gols	Particiapcio_Gols	Quocient de gols efectuats pel jugador i el número total de gols efectuat per l'equip al llarg de tota la temproada
Participació en termes de xGols	Participacio_xGols	Quocient de xGols acumulats pel jugador i el número total de xGols acumulats per l'equip a llarg de tota la temporada

Font: Elaboració pròpia

2) Gràfics

A partir de les taules de dades anteriors s'ha pogut construir un seguit de gràfics que, com s'ha comentat amb anterioritat, complementen l'anàlisi i formen part del mètode dels gols esperats.

Tabla 2.2.1 Tipologia i descripció dels gràfics d'anàlisi futbolística

Gràfic	Modalitat	Descripció
xGols vs Temps	Equip	Gràfic que mostra l'evolució dels gols esperats acumulats dels dos equips rivals en el transcurs del partit
Mapa de xuts	Equip/Jugador	Mapa que mostra els xuts a porteria que hi ha hagut en un partit

Font: Elaboració pròpia

3. Models

En aquest apartat es presenta, de manera sintètica, la base de l'anàlisi economètrica que s'ha fet. La modelització principal, es recorda, és la de la variable econòmica *valor de mercat* dels jugadors, la qual s'ha portat a terme a través d'un model de regressió lineal múltiple.

Altrament, per a construir les mètriques *Gols esperats* i *Assistències esperades* s'ha utilitzat la regressió logística derivada dels models lineals generalitzats.

1) Model de regressió lineal múltiple

Siguin Y, variable objecte d'estudi o variable endògena o explicada, i X_k el conjunt de k variables exògenes o explicatives, un model de regressió lineal múltiple (MRLM) estudia la relació entre Y i X_k suposant que la seva causalitat és unidireccional i de tipus lineal. És a dir:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Els coeficients o paràmetres β seran constants per a tota la mostra i mesuren la variació en termes mitjans de la variable endògena davant un augment unitari de variable explicativa que se li correspongui. D'altra part, el terme de pertorbació o d'error, ϵ, és la part aleatòria no observable que, d'entre altres coses, recull el comportament aleatori de Y i fa que el model no sia totalment determinista.

Pel que fa a l'estimació dels coeficients a partir de la informació mostral que es disposa, el mètode que s'utilitza és el dels Mínims Quadrats Ordinaris (MQO) que consisteix en trobar aquells valors de $\hat{\beta}$ que minimitzin la suma dels quadrats dels errors. Així, sia una mostra de grandària N, el problema que es planteja és el següent:

$$Min\ SQE = Min\ (e'e) = Min\left(\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)\right) = \cdots =$$

$$= Min\left(Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right)$$

Derivant respecte els coeficients, igualant a 0 i aïllant, resulta en anotació matricial:

$$\hat{\beta}_{MQO} = (X'X)^{-1}(X'Y)$$

L'estimació dels paràmetres per MQO dona les propietats de linealitat, absència de biaix (el valor esperat coincideix amb el paràmetre poblacional), eficiència o variància mínima i consistència (quant més gran sigui la mostra, major aproximació de l'estimador al seu valor poblacional).

En quant als termes de pertorbació, per a que el model sigui lineal i es mantinguin les propietats de l'estimació per MQO, s'han de donar unes certes condicions que a priori se suposen que certament es donen. Aquests supòsits són:

- $E[\epsilon] = 0$. Els residus s'han de trobar al voltant de 0.
- $Var(\epsilon) = \sigma_u^2$. Variància constant o homoscedasticitat.
- $Cov(\epsilon_i, \epsilon_j) = Cov(\epsilon_j, \epsilon_i) = 0, \forall i \neq j$. Termes de pertorbació independents o no autocorrelació.

Resumidament, els supòsits anteriors es reflecteixen en la següent distribució:

$$\epsilon \sim N(0, I_N \sigma^2)$$

Com s'ha comentat anteriorment, en el cas que no es donessin les propietats citades del terme de pertorbació, els estimadors $\hat{\beta}_{MQO}$ perdrien la propietat d'eficiència i la seva estimació seria menys precisa. A part, la inferència que se'n derivés seria invàlida.

2) Model binomial o de resposta binària

Per a la modelització de variables de resposta binària s'ha utilitzat una extensió dels models de regressió lineal múltiple, el que es coneix com models lineals generalitzats (o per les seves sigles en anglès *GLM*). Els GLM es fonamenten en la mateixa idea que els models lineals clàssics, però en aquest cas permeten l'ús de distribucions no normals dels errors i que la variància no sigui constant. També, un altre tret destacable d'aquest tipus de modelització és que permet linealitzar la relació entre la variable endògena i les variables explicatives mitjançant la transformació de la variable resposta, que es coneix com a funció de vincle o *link function*. A més a més, aquesta permet que les prediccions del model quedin acotades (en el cas que fos necessari).

Així, l'especificació d'un GLM es realitza de la següent manera:

- Y: variable endògena o explicada. Segueix una distribució de la família exponencial. La seva esperança s'escriu com $E[Y] = \mu$.
- η: predictor lineal. Correspon a la matriu formada per les variables k independents o explicatives i els paràmetres associats a estimar.
- Funció *link*. Aquesta funció és la que posa en relació l'esperança de la variable endògena i el predictor lineal.

$$g(\mu) = \eta \Leftrightarrow \mu = g^{-1}(\eta)$$

La selecció de la funció *link* és molt important ja que es farà en funció de com es vulgui que sigui l'esperança del model, la qual també depèn de la naturalesa de les dades. Per tant, per a cada família o distribució de probabilitat existeix una funció *link* que s'anomena *canònica* i és la que s'utilitza per defecte.

Com ja s'ha comentat, en aquest treball es farà ús dels GLM per a modelitzar variables de resposta binària, les quals pertanyen a la família Binomial. Una variable de resposta binària expressa per a cada observació la presència, o no, d'una certa qualitat, condició, etc. Per tant, les observacions preses sobre aquesta podran prendre 2 valors, normalment s'utilitzen els valors 1 i 0 per indicar que presenta la qualitat, o *èxit,* i per indicar que no la presenta, o *fracàs*, respectivament. Ara bé, el que s'aconsegueix un cop construït el model és, per a cada observació, la probabilitat d'èxit o probabilitat que es doni la qualitat. Conseqüentment, si s'anota $P(Y = 1) = \pi$ com la probabilitat d'èxit de l'observació (anàlogament, la probabilitat de fracàs seria $P(Y = 0) = 1 - \pi$), l'esperança de la variable resposta és

$$E[Y] = \pi$$

Així doncs, tenint en compte que la probabilitat d'èxit està compresa entre 0 i 1, els valors del predictor lineal η tenen com a domini $\mathbb{R}$, la funció link haurà de tenir en compte aquestes condicions.

Existeixen diverses funcions link per tractar variables d'aquest tipus. Però, la funció link canònica per aquests casos és la funció *lògit* degut, entre altres arguments, a la seva facilitat d'interpretació de resultats:

$$\eta = g(\pi) = logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

En funció del predictor lineal, la probabilitat d'èxit és la que segueix:

$$\pi(\eta) = g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$$

- Amb l'anterior, a l'hora de fer les interpretacions dels paràmetres obtinguts s'ha de tenir en compte que dependrà de com es presentin o en quina escala estiguin els resultats:

- Quan el paràmetre β_k associat a la variable x_k és positiu (*negatiu*), aleshores la probabilitat condicional $\Pr(y_i = 1|x_i;\beta)$ augmentarà (*disminuirà*) quan x_j creix.
- *Odd*. L'*odd* d'un succés és la ràtio de probabilitats del succés que esdevingui i el seu complementari. Si s'apliquen logaritmes a banda i banda, s'obté el *log-odd*.
- $odd = \frac{\pi}{1-\pi} \Leftrightarrow logodd = \log\left(\frac{\pi}{1-\pi}\right)$
- Pel que fa a la interpretació del canvi de nivell d'un factor en el predictor lineal, aquesta canvia segons en quina escala s'estigui analitzant:
    - El canvi de nivell d'un factor en el predictor lineal, mentre les altres variables romanguin constants, tindrà un efecte multiplicador quan s'estigui treballant amb *odds*.
    - El canvi de nivell d'un factor en el predictor lineal, mentre les altres variables romanguin constants, tindrà un efecte de canvi de signe sobre el paràmetre, quan s'estigui treballant amb *log-odds*.
- Probabilitat. La interpretació d'un canvi del nivell d'un factor, en termes de probabilitat, no només depèn d'esmentat factor, sinó que també depèn de les demés variables. Per tant, serà una aproximació el que s'obtindrà. Si es vol calcular el canvi de nivell del factor x_2 que tindria sobre la probabilitat, s'haurà de derivar respecte x_2 i s'obté:

$$\frac{\partial \pi}{\partial x_2} = \beta_2 \pi(1-\pi)$$

3) Construcció i avaluació

En quant a la metodologia respecte la construcció i avaluació de models, s'ha seguit un mateix esquema per ambdós tipologies, sempre tenint en compte aquelles diferències significatives que hi ha entre una i altra. Doncs, els passos que s'han adoptat són els següents:

1. Anàlisi descriptiva i exploració de les dades. Aquesta fase inicial té diversos objectius, com el de trobar relacions entre la variable depenent i les variables que es disposa potencialment explicatives o com la de detectar, i posteriorment eliminar, altes

correlacions entre les candidates a ser variables explicatives que afectarien negativament al model provocant un problema de multicol·linealitat. També serà important comprovar quin és el comportament de la variable resposta. En aquesta anàlisi s'empraran, principalment, gràfics de caixa, matrius de correlacions, gràfics de barres, etc.

2. Bondat de l'ajust del model a les dades o diagnòstic del model. Un cop construït el model, primerament es comprovarà la significació estadística individual dels estimadors[4]. També, es comprovarà, en el cas de regressió lineal múltiple, les suposicions sobre els termes de pertorbació i, per tant, es farà un anàlisi dels residus. Seguidament, per mitjà de l'R^2 i, pels models binomials, el $pseudo - R^2$ es valorarà la quantitat de variabilitat que capten els models. A més a més, en aquesta fase anàlogament s'examinarà la presència de resultats individuals amb comportament especials o amb problemes que afectin als resultats del model (valors estranys, influents, atípics o *outliers*, etc.) i la possible presència de cert grau de multicol·linealitat.
3. Avaluació de models. A l'hora de comparar models, els criteris que s'utilitzaran són el test de raó de versemblança per a models encaixats, el criteri d'informació d'*Akaike*, que avalua tant la bondat de l'ajust com la complexitat del model, el criteri d'informació de Bayes i el criteri de la deviància. Tant per ambdós criteris d'informació com per la deviància, quant més petits siguin els seus valors major serà l'ajust.

Finalment, la construcció dels models es farà sempre prenent en consideració el criteri de parsimònia, el qual requereix que el model sia tan simple com sigui possible.

4. Recursos informàtics

El programa que s'ha utilitzat és el programa lliure estadístic R, amb el qual s'ha treballat durant tot el grau. També s'ha fet ús del programa $Excel$ en la fase de recollida de dades.

[4] En regressió lineal múltiple s'utilitzarà el test de *t-Student*. L'equivalent en models binomials és el test de *Wald* per a la significació individual.

IV. RESULTATS

1. Model de gols esperats (*xGoals model*)

1) Què és l'Expected goals method?

L' *Expected Goals Method* o, traduït, Mètode dels Gols Esperats és una nova metodologia que pretén avaluar el rendiment futbolístic aplicat tant a nivell col·lectiu com a nivell individual. Aquest nou mètode neix el 2012 amb la publicació d'un article per part de l'empresa britànica *Opta Sports*, que es dedica a la recopilació, tractament i anàlisi de dades esportives; i que, com bé s'explica a l'esmentat document, consisteix en la creació d'una nova mètrica anomenada *Expected Goals* (en endavant, *xG*) la qual a partir de diferents factors i covariables avalua la probabilitat d'ésser gol d'un xut a porteria.

Abans de l'aparició d'aquesta nova mètrica, les actuacions dels equips i dels jugadors es mesuraven a partir de la recollida de variables molt genèriques com el número de gols, xuts, assistències, la possessió de pilota de cada equip, etc. que, bàsicament, són resultats i aportaven, o aporten, un valor descriptiu a l'hora de dur a terme un anàlisi. Així, al finalitzar un partit, l'aficionat podia marxar amb sensacions contradictòries comparant les estadístiques finals del seu equip amb el que ell o ella havia presenciat en directe, dient frases com: "L'equip contrari ha tingut molta sort", "Si haguéssim marcat aquella ocasió el partit hauria canviat", "L'equip ha jugat de pena", etc. No obstant, els *xG* precisament el que aspiren a avaluar és l'actuació (*performance*) i no els resultats, pretén extreure el factor sort de la fórmula, l'aleatorietat del propi joc i aportar un valor predictiu de l'actuació o del rendiment. Concretament, els *xG* mesuren la perillositat de l'oportunitat de gol que s'ha donat o, d'una altra manera, diferencia entre un xut llunyà i un proper a la línia de gol, entre un xut amb la cama dominat i amb el cap, entre un xut precedit d'una jugada col·lectiva i el d'una acció individual… Per tant, el que es recull és la qualitat del tir mesurada amb la probabilitat de ser gol.

Aquesta nova mètrica és el fonament d'anàlisis que mai no s'havien pogut fer abans. Per exemple, la comparació de la suma dels gols esperats i els gols que realment ha marcat un jugador origina una mètrica que informa de si el jugador ha rendit segons el que s'esperava o no. A més, si se l'afegeix el fet que es pot recollir cada temporada, aquesta mesura temporal, juntament amb la diferència entre gols esperats i gols efectuats, permetria, per exemple, detectar aquelles actuacions de jugadors que han sigut fruït de la casualitat. També, a nivell col·lectiu, permet detectar quins equips són els que produeixen les millors oportunitats i atacs.

Amb tot, tot i que a nivell periodístic no s'ha estès gaire i existeix un cert rebuig, els equips de futbol professional de tot el món s'han anat sumant paulatinament a aquesta nova tendència, que va juntament amb l'extensió i la millora de tecnologia de recollida i anàlisi de dades de caire esportiu. Tenint present que, al cap i a la fi, el món del futbol no deixa de ser un negoci, els equips sobretot inverteixen en aquest tipus d'innovacions que els poden ajudar a millorar en eficiència i eficàcia en camps com l'*scouting* i el *recruiting*. Detectar joves promeses abans que ningú altre pot suposar un gran avantatge competitiu sobre els competidors. Per altra banda, un altre sector que es beneficia d'aquestes noves mètriques és el de les apostes esportives. Les cases d'apostes s'aprofiten de no només de l'augment de les dades que es recullen, sinó de la seva qualitat per a poder fer millors prediccions i calcular el preu de les apostes.

2) Anàlisi descriptiva

1. Base de dades: *Xuts*

A partir de les taules del tipus *events* s'ha fet una recollida de les variables potencialment explicatives del model. Aquest procés ha consistit en, un cop es disposa d'una base de dades on s'hagi recollit tots els tirs o xuts a porteria de tots els partits de les lligues disponibles, avaluar i descriure com ha sigut el xut. Les variables que s'han recollit són:

- $Goal$: variable que avalua si el tir ha sigut gol (1) o no (0).
- $partCos$: variable que recull amb Part del cos s'ha efectuat el xut. Es diferencien 3 categories:
  - *right*: El xut s'ha efectuat amb la cama dreta.
  - *left*: el xut s'ha efectuat amb la cama esquerra.
  - *headbody*: el xut s'ha efectuat amb el cap o qualsevol altra part del cos.
- GB: variable que recull si el xut s'ha fet amb la cama dominat (*good*), amb la cama no dominant (*bad*) o amb una altra part del cos que no sia les dues cames (*neutral*), que normalment serà el cap.
- $angles$: variable que recull l'angle de xut en radians que es forma des de la posició d'on s'efectua el xut respecte la porteria.
- $distancia$: variable que recullla la distància en metres que hi ha des de la posició d'on s'efectua el xut fins el centre de la porteria.
- $counter$: variable que avalua si el tir ve precedit d'un contra-atac (1) o no (0).
- $centrada$: vaaiable que avalua si el tir ve precedit d'una centrada (1) o no (0).
- Rol: variable que recull la posició o rol que té el jugador dins del camp.
- $League$: variable que recull en quina lliga s'ha efectuat el xut.

2. Variable endògena: Gol

La variable que es pretén modelitzar és la variable *Goal*, la qual informa de si el tir ha sigut gol (1) o no (0). Com bé ja indica la seva definició, aquesta variable és dicotòmica o binària, de manera que la seva modelització es farà a partir de GLM sabent que es distribueix com una Binomial amb *n = 1* (o el que és el mateix, com una Bernoulli) com ja s'ha justificat anteriorment. La grandària o mida de la mostra és de 40.461 xuts a porteria. Doncs, es pot definir

$$y_i = \begin{cases} 1, \ el\ xut\ i\ ha\ sigut\ gol. \\ 0, \ cas\ contrari \end{cases}$$

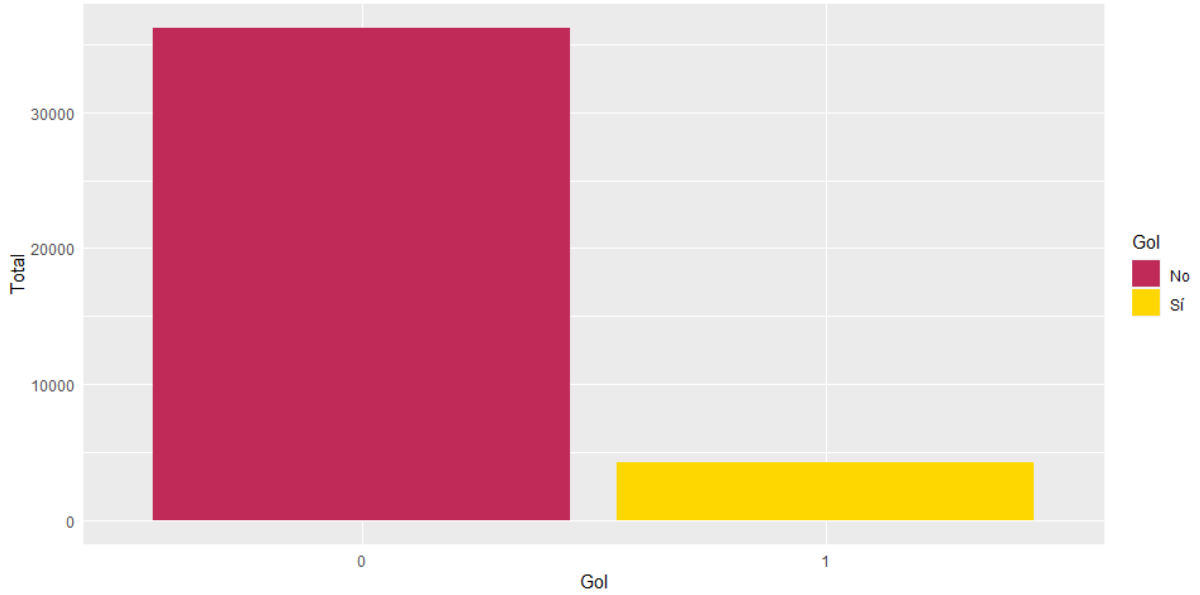
En primera instància, cal tenir en compte que per la pròpia naturalesa de l'esport, és d'esperar que hi hagi molts més tirs fallats que no pas gols. A partir de les dades que es disposa, a la taula 4.1.2.2.1 es pot veure el número total de xuts a porteria que s'han registrat al llarg de la temporada 2017-2018 en les 5 lligues europees més importants, així com la mitjana de tirs i gols per partit (no s'inclouen ni els tirs de falta ni els de penal). Tanmateix, a la figura 4.1.2.2.1 es representa en un diagrama de barres el número total de tirs que han estat *Gol* i els que no, o el que és el mateix.

Taula 4.1.2.2.1 Taula resum dels xuts i gols de les 5 lligues europees (temporada 2017/2018)

Xuts	***Gols***	***Xuts/Partit***	***Gols/Partit***
40.461	4.271	22,16	2,34

Font: Elaboració pròpia

Figura 4.1.2.2.1 Distribució de la Variable Gol



Font: Elaboració pròpia

A la taula que es mostra s'evidencia el ja s'ha assenyalat anteriorment: el número de tirs i el número de gols difereixen en gran mesura. Mentre que, en total, s'han efectuat 40.461 xuts a porteria, només 4.271 han acabat sota la xarxa. És a dir, només el $10{,}55\%$ dels tirs que s'efectuen acaben sota la xarxa. O, també, com a primera estimació de la probabilitat que té un tir de ser gol, la probabilitat que un tir sia gol és del $10{,}55\%$. No obstant, no tots els tirs s'efectuen de la mateixa manera: es poden fer amb la cama dreta, l'esquerra o amb el cap; l'altura de la pilota quan es colpeja pot ser baixa o alta, la jugada predecessora pot ser una jugada individual o col·lectiva… I no es pot oblidar que hi intervenen els jugadors de l'equip contrari, essent el més important d'ells el porter, els quals dificulten la tasca de marcar un gol. Per tant, queda clar que no tots els xuts són iguals o, dit d'una altra manera, no tots els xuts tenen la mateixa probabilitat de ser gol.

3. Variables categòriques o factors

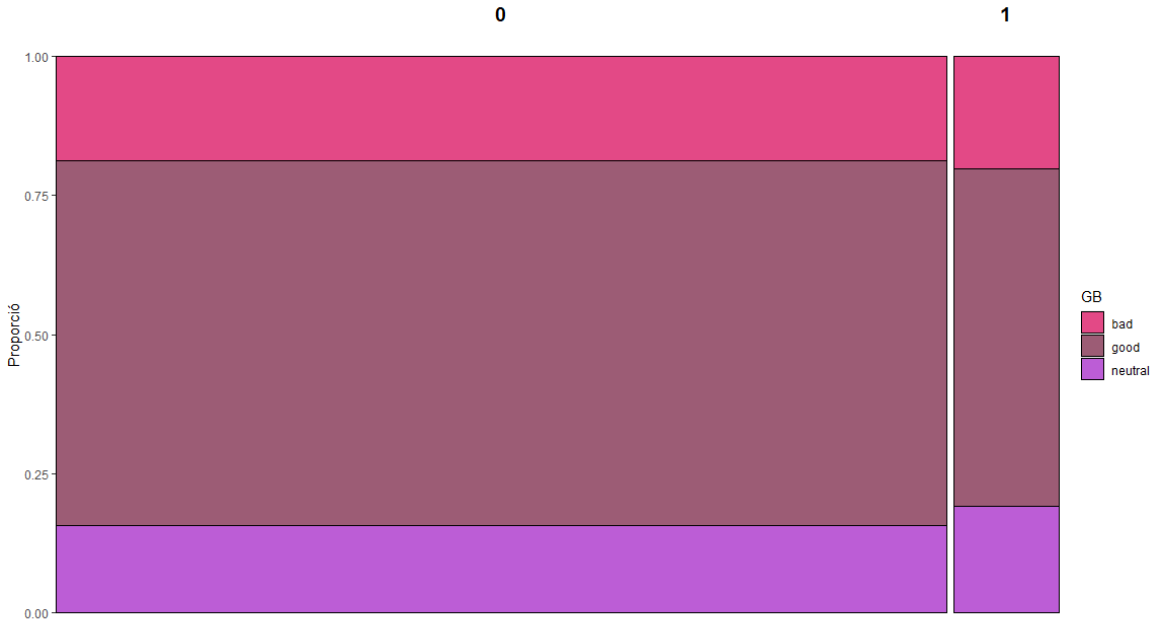
A continuació, es procedeix a fer una anàlisi descriptiva creuat de les variables factor amb la variable endògena per tal de trobar possibles patrons i relacions entre elles. Per començar, a les taules de contingència es mostren les diverses freqüències relatives en funció de les categories de la variable *GB*, *counter*, *centrada*, *Rol* i *League*. Així mateix, per tal d'observar millor el seu comportament, a les figures contigües es representen les respectives taules de contingència en forma de diagrames de Marimekko o gràfics de mosaic, adients per a la representació de la relació entre dues variables discretes:

Taula 4.1.2.3.1 Taula de contingència per a les variables Goal i GB

Goal	GB			**Total**
	bad	good	neutral	
0	6.783	23.706	5.701	36.190
1	866	2.587	818	4.271
Total	7.649	26.293	6.519	40.461

Font: Elaboració pròpia

Figura 4.1.2.3.1 Diagrama de mosaic de la variable endògena Goal en funció de la variable GB



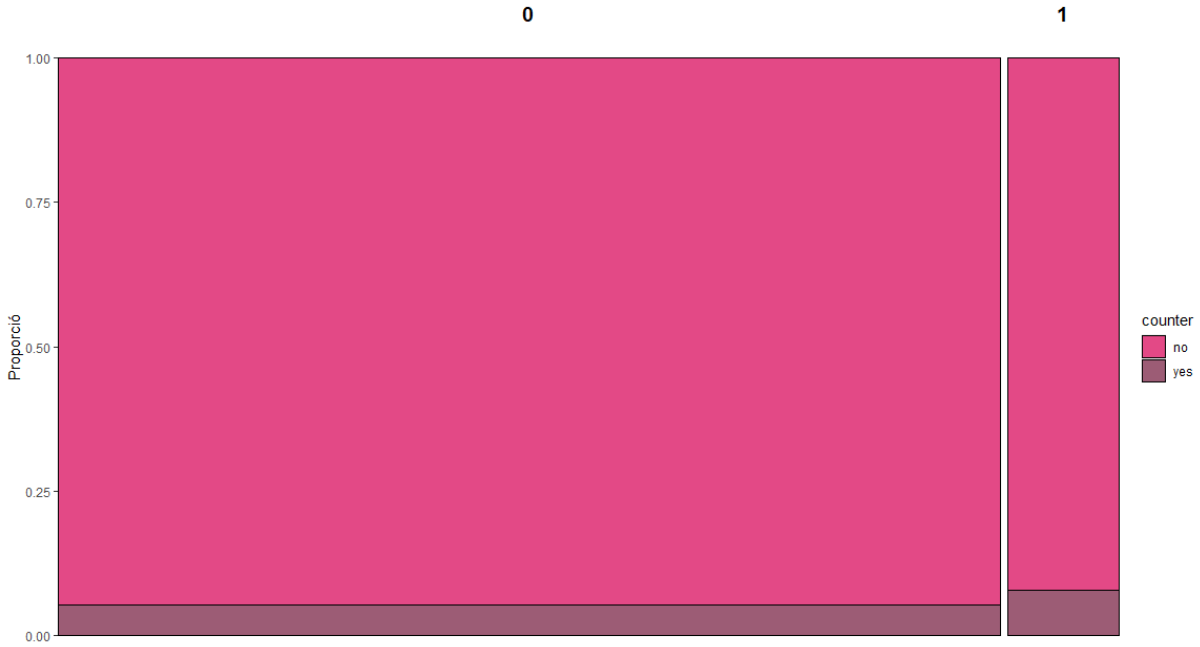
Font: Elaboració pròpia

Taula 4.1.2.3.2 Taula de contingència per a les variables Goal i Counter

Goal	*counter*		***Total***
	no	yes	
0	34.275	1.915	36.190
1	3.935	336	4.271
Total	38.210	2.251	40.461

Font: Elaboració pròpia

Figura 4.1.2.3.2 Diagrama de mosaic de la variable endògena Goal en funció de la variable counter



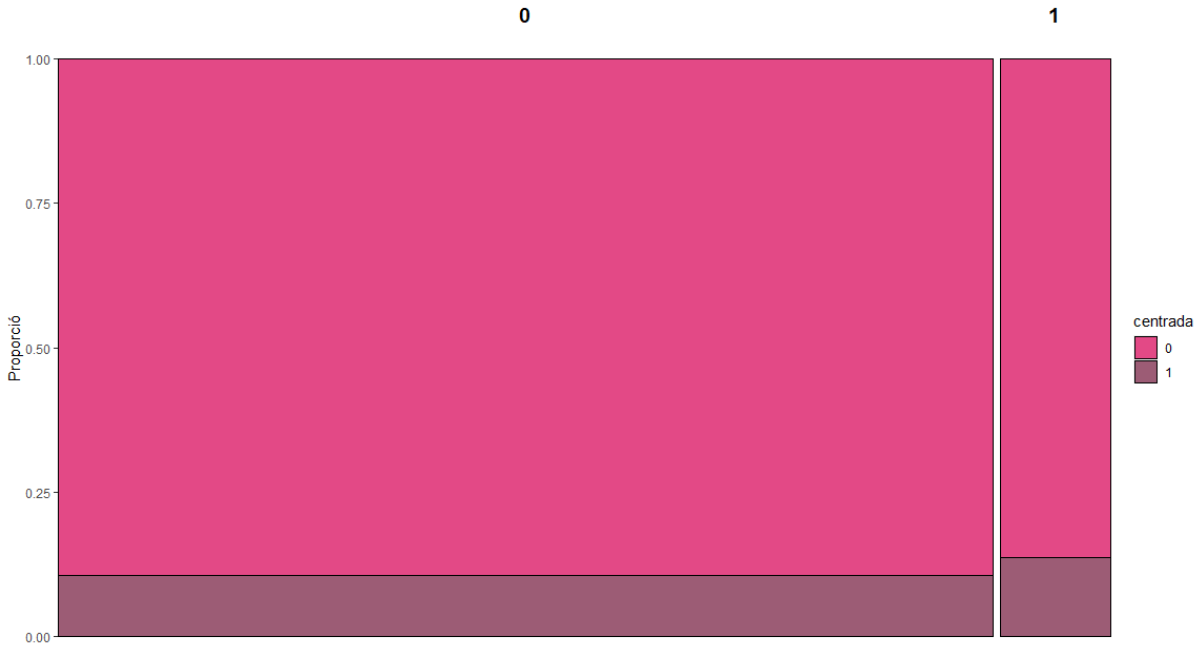
Font: Elaboració pròpia

Taula 4.1.2.3.3 Taula de contingència per a les variables Goal i centrada

Goal	centrada		**Total**
	0	1	
0	32.323	3.867	36.190
1	3.683	588	4.271
Total	36.006	4.455	40.461

Font: Elaboració pròpia

Figura 4.1.2.3.3 Diagrama de mosaic de la variable endògena Goal en funció de la variable centrada



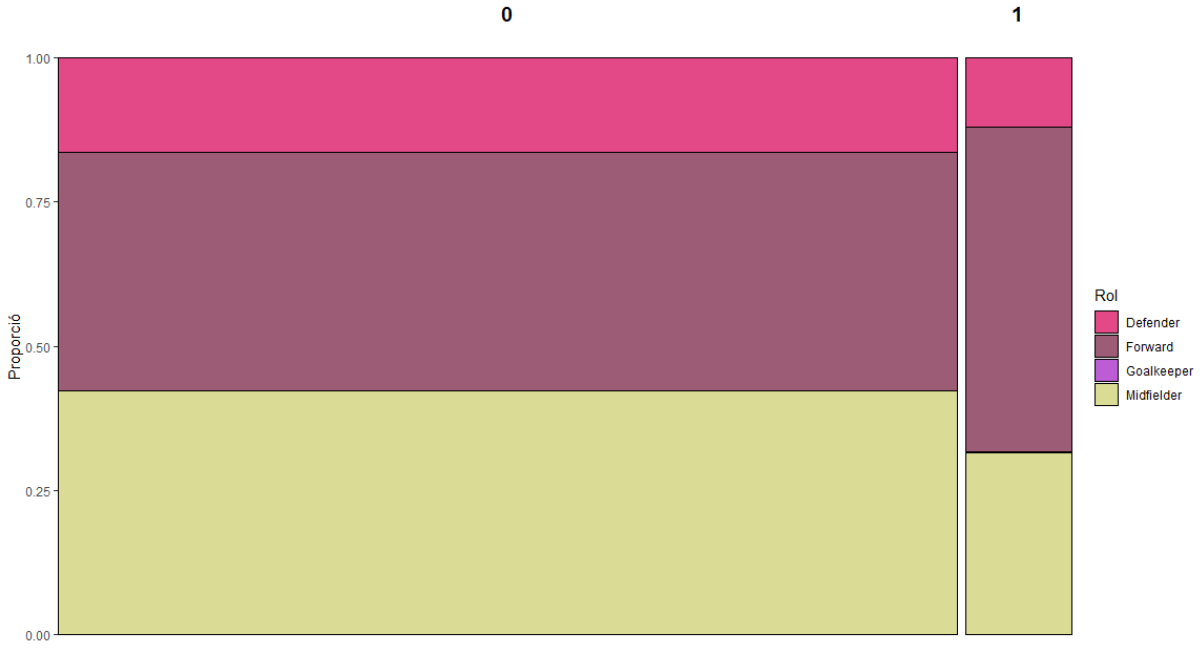
Font: Elaboració pròpia

Taula 4.1.2.3.4 Taula de contingència per a les variables Goal i Rol

Goal	Rol				**Total**
	Defender	Forward	Goalkeeper	Midfielder	
0	5.938	14.968	6	15.278	36.190
1	514	2.405	1	1.351	4.271
Total	6.452	17.373	7	16.629	40.461

Font: Elaboració pròpia

Figura 4.1.2.3.4 Diagrama de mosaic de la variable endògena Goal en funció del factor Rol



Font: Elaboració pròpia

Taula 4.1.2.3.5 Taula de contingència per a les variables Goal i League

Goal	League					**Total**
	Bundesliga	LaLiga	Ligue1	PremierLeague	SerieA	
0	6.151	7.095	7.454	7.537	7.953	3.6190
1	747	884	873	914	853	4.271
Total	6.898	7.979	8.327	8.451	8.806	40.461

Font: Elaboració pròpia

Figura 4.1.2.3.5 Diagrama de mosaic per a la variable endògena Goal en funció del factor League

Font: Elaboració pròpia

Els gràfics de mosaic mostren les freqüències relatives per a cada categoria per a cadascun dels 5 factors que es disposa respecte la variable $Goal$. En termes generals i a simple vista, no s'observen canvis desmesurats en les distribucions, però si petites diferències. Pel que fa als factors $counter$ i $centrada$, les proporcions dels xuts que són precedits d'un contra-atac o d'una centrada i que acaben en gol, $7{,}87\%$ i $13{,}77\%$ respectivament, són lleugerament majors respecte als que no acaben essent-ho, $5{,}29\%$ i $10{,}69\%$ respectivament. Per tant, ambdós factors semblen agents potencialment explicatius.

En quant a la variable *GB*, la qual indica si el xut s'ha efectuat amb la cama dominant (*good*), o no (*bad*), o amb el cap o qualsevol altra part del cos (*neutral*), a priori i com idea inicial, es podria pensar que el fet de xutar amb la cama dominant afavorís la probabilitat que el xut acabés sent gol ja que per definició és aquesta cama amb la qual es té major precisió, qualitat, etc. Però, observant el diagrama de mosaic anterior es veu que la proporció de tirs que es fan amb la cama dominant és, sorprenentment, inferior pels tirs que acaben en gol respecte els que no, un $60{,}57\%$ enfront un $65{,}50\%$ respectivament. A més a més, i contràriament al pensament inicial, les proporcions de les altres categories, *bad* i *neutral*, són majors. És a dir, en termes relatius, es marquen més gols respecte els que es fallen. Aquest comportament podria ser explicat pel fet que amb la cama dominant s'intenten molts més tirs comparat amb les altres dues categories: el $64{,}98\%$ dels tirs totals s'efectuen amb la cama "bona'.

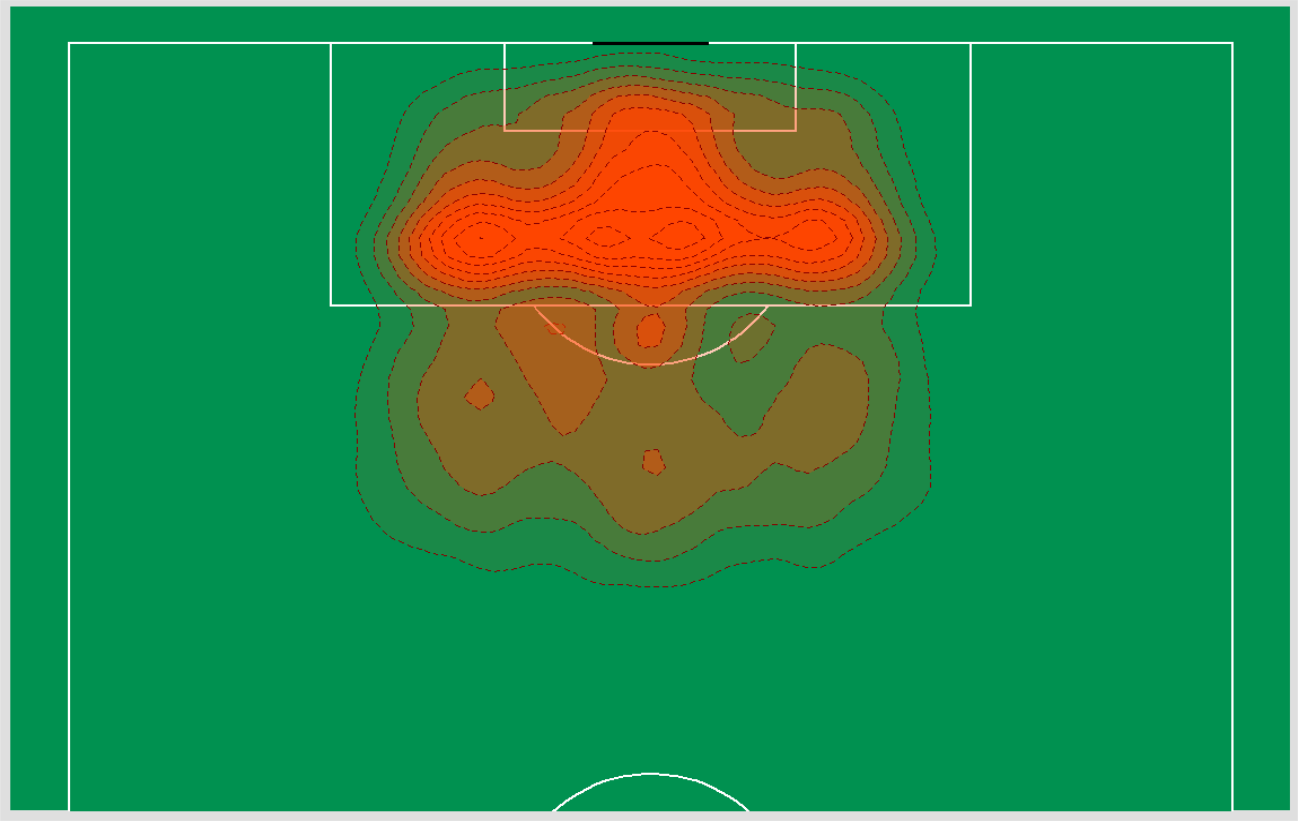
Finalment, no sembla que hi hagi diferències en la distribució dels xuts quan són gol i quan no ho són respecte el factor *Lliga*. Independentment de la lliga en la qual es jugui, la proporció de tirs que són gol i dels que no ho són és pràcticament la mateixa. Tot i això, s'ha de remarcar que la *Bundesliga*, o lliga alemanya, és la lliga on menys xuts s'efectuen, 6.898 tirs en total en front els $8.390{,}75$ de mitjana tenint en compte la resta de lligues; i que la *Serie A*, o lliga

italiana, és la lliga on més xuts s'efectuen, amb un total de 8.806 xuts a porteria. Per altra banda, pel que fa a la variable *Rol*, cal destacar que la proporció de xuts que són gol respecte dels que no varia considerablement quan es tracta de la categoria *forward*, o davanter. Fins i tot, és molt interessant ja que es pot observar clarament la diferència que hi ha entre els diferents rols: mentre que diferència de xuts efectuats entre *forwards* i *midfielders* només és de 744 xuts més assignats als davanters, els quals representen sobre el total un 1,84%, la diferència de gols entre ambdós rols és de 1.054 gols més també a favor dels davanters. Així, aquells jugadors etiquetats amb la categoria de davanter, a priori, tenen una capacitat de marcar gol major que la resta. No obstant, aquesta major precisió dels davanters a l'hora de xutar a porteria es podria veure afavorida, per exemple, per la distància de tir (s'ha de tenir en compte que els davanters estan posicionats més a prop de la porteria rival, en general i durant més temps).

4. Variables numèriques o covariables

Les principals característiques que afavoreixen que un xut sigui gol o no són la distància i l'angle de tir: quant més a prop de la porteria s'efectuï un xut, més angle de tir tindrà el jugador i, per tant, requerirà de menys precisió per marcar. Per tant, serà més fàcil.

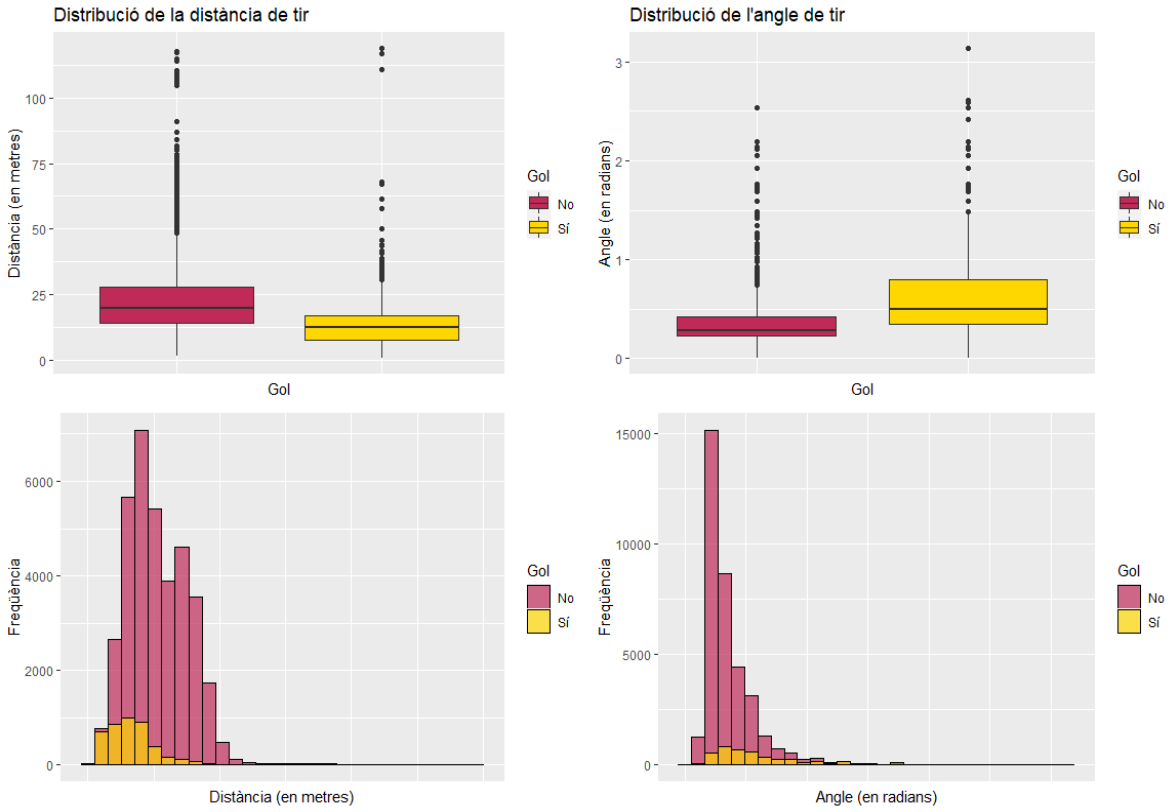
Il·lustració 4.1.2.4.1 Mapa de calor dels tirs a porteria



Font: Elaboració pròpia

La il·lustració 4.1.2.4.1 mostra la distribució dels xuts dins de les dimensions del camp en un mapa de calor. Les zones on s'expressa major tonalitat indica major densitat, és a dir, major concentració de tirs. Contràriament, aquelles zones on hi hagi poca tonalitat de color són zones del camp on és poc freqüent efectuar xuts. Com es pot observar, clarament la zona amb major concentració de tirs és una zona molt propera a la porteria, dins l'àrea. A mesura que la distància augmenta respecte l'eix vertical, la densitat disminueix significativament. També es pot destacar que els tirs que es realitzen es concentren, generalment, a la zona central del camp, delimitada per les línies de l'àrea. Per tant, la distància i l'angle de tir són motius a tenir en compte a l'hora de valorar un tir.

Figura 4.1.2.4.1 Distribucions de les variables distància i angle



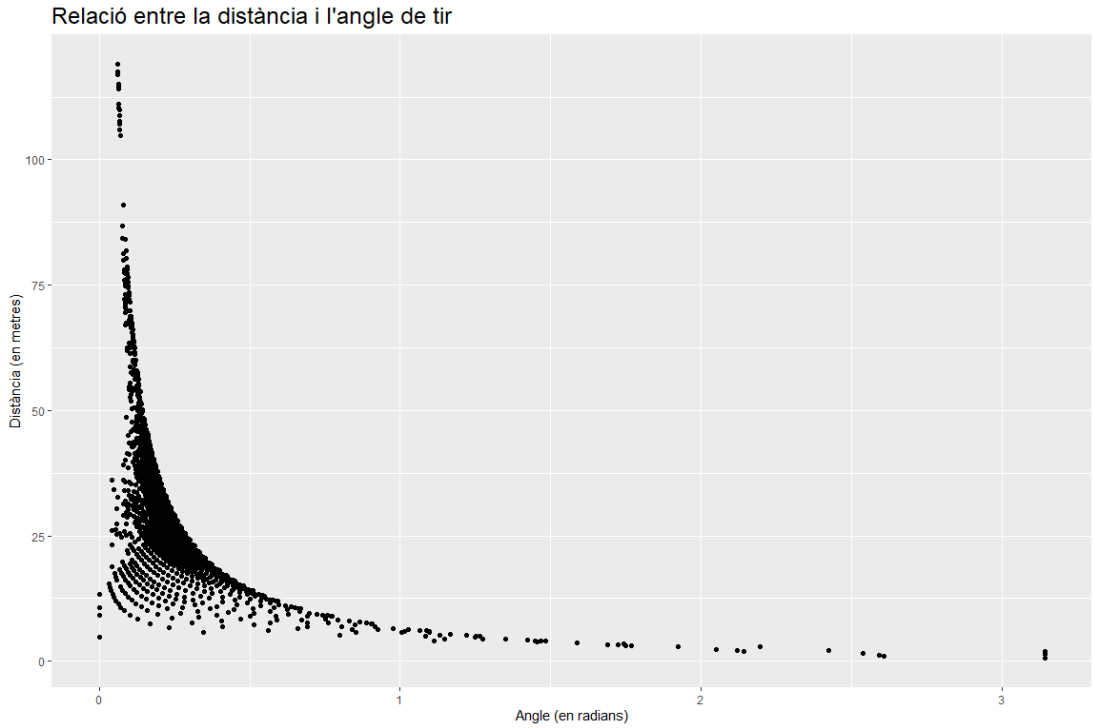
Font: Elaboració pròpia

Igualment, a través del diagrama de caixes i l'histograma a la figura 4.1.2.4.1 es pot apreciar que hi ha una diferència pel que fa a la distància de tir d'aquells entre aquells que són gol i aquells que no, essent la distància dels primers menor que la dels segons (el rang interquartílic de la distància dels tirs que són gol és menor que la dels que no ho són i, a més, la mediana també és menor). A més a més, en relació al que s'ha dit anteriorment, en quant a la variable *angles*, aquells tirs que són *Gol* es caracteritzen per haver estat realitzats amb un angle major respecte els que no ho són.

No obstant, clarament hi ha una relació existent entre ambdues covariables. No només s'ha de tenir en compte que quanta menys distància respecte la porteria major serà l'angle de tir, sinó que també degut a les propietats geomètriques i a les limitacions de les dimensions del camp, per a una mateix angle de tir es podran trobar diferents distàncies associades però limitades.

A la figura 4.1.2.4.2 precisament es pot apreciar aquestes dues propietats d'ambdues variables quan s'analitzen conjuntament. Per una banda, es mostra la relació inversa entre distància i angle de tir als extrems de la corba que es dibuixa: quanta més distància menor serà l'angle de tir. Per altra banda, la definició tan marcada de la corba que es dibuixa expressa els límits que tenen per naturalesa les dades i, a més a més, es mostra aquesta propietat geomètrica de les circumferències on per a un angle fixat hi ha diverses distàncies.

Figura 4.1.2.4.2 Diagrama de punts de les variables distància i angle



Font: Elaboració pròpia

3) Modelització

En aquest apartat es procedeix a construir el model final tenint en compte el que s'ha pogut extreure de l'anàlisi descriptiu. Com ja s'ha comentat anteriorment, la variable endògena del model és la variable $Goal$ la qual és dicotòmica i pot prendre 2 valors: 1 quan el xut és gol i 0 quan no ho és. Per la naturalesa pròpia de la variable, es tracta d'una variable que segueix una distribució de Bernoulli (o Binomial amb $n = 1$) i, per tant, se seguirà la metodologia esmentada per aquest tipus.

A partir d'aquí, a la taula 4.1.3.1 es mostren els diferents models que s'han anat construint juntament amb els seus estadístics associats. L'estratègia que s'ha seguit és l'anomenada *stepwise*, la qual consisteix en anar afegint variables i interaccions d'una en una i comprovar que, efectivament, millorin significativament el model, no obstant considerant que l'agregació d'una nova variable pot influenciar en aquelles que ja hi eren i, per tant, poden ser excloses.

Taula 4.1.3.1 Taula resum dels models construïts per determinar la probabilitat de gol d'un xut a porteria

Model	***Fórmula***	***Graus llibertat***	***Deviància***	***Info. Akaike***	***Info. Bayes***	***Pr(>Chi)***
MCV1	Goal ~ distancia	40.459	23.873,21	23.877,21	23.894,43	0,00
MCV2	Goal ~ angles	40.459	24.161,15	24.165,15	24.182,37	0,00
MCV3	Goal ~ distancia + angles	40.458	23.596,73	23.602,73	23.628,55	0,00
MCV4	Goal ~ distancia + angles + distancia:angles	40.457	23.595,62	23.603,62	23.638,05	0,29
MFCV1	Goal ~ distancia + angles + counter	40.460	23.488,48	23.496,48	23.530,91	0,00
MFCV2	Goal ~ distancia + angles + centrada	40.457	23.559,26	23.567,26	23.601,69	0,00
MFCV3	Goal ~ distancia + angles + Rol	40.455	23.436,96	23.448,96	23.500,61	0,00
MFCV4	Goal ~ distancia + angles + League	40.454	23.592,65	23.606,65	23.666,91	0,40
MFCV5	Goal ~ distancia + angles + GB	40.456	23.099,75	23.109,75	23.152,79	0,00
MFCV6	Goal ~ distancia + angles + counter + centrada	40.456	23.457,36	23.467,36	23.510,40	0,00
MFCV7	Goal ~ distancia + angles + counter + Rol	40.454	23.348,30	23.362,30	23.422,55	0,00
MFCV8	Goal ~ distancia + angles + centrada + Rol	40.454	23.407,76	23.421,76	23.482,02	0,00
MFCV9	Goal ~ distancia + angles + counter + GB	40.455	23.027,03	23.039,03	23.090,67	0,00

MFCV10	Goal ~ distancia + angles + centrada + GB	40.455	23.091,87	23.103,87	23.155,52	0,00
MFCV11	Goal ~ distancia + angles + Rol + GB	40.453	23.032,21	23.048,21	23.117,08	0,00
MFCV12	Goal ~ distancia + angles + counter + Rol + centrada	40.453	23.323,78	23.339,78	23.408,65	0,00
MFCV13	Goal ~ distancia + angles + counter + Rol + centrada + GB	40.451	22.963,60	22.983,60	23.069,68	0,00
MFCV14	Goal ~ distancia + angles + counter + Rol + centrada + GB + GB:centrada	40.449	22.956,2	22.980,20	23.083,50	0,02
MFCV15	**Goal ~ distancia + angles + counter + Rol + centrada + GB + distancia:centrada**	**40.450**	**22.953,98**	**22.975,98**	**23.070,67**	**0,00**
MFCV16	Goal ~ distancia + angles + counter + Rol + centrada + GB + GB:centrada + distancia:centrada	40.448	22.950,39	22.976,39	23.088,3	0,01
MFCV17	Goal ~ distancia + angles + counter + Rol + centrada + GB + distancia:centrada + Rol:distancia	40.447	22.951,92	22.979,92	23.100,44	0,56
MFCV18	Goal ~distancia + angles + counter + Rol + centrada + GB + distancia:centrada + Rol:angle	40.447	22.952,70	22.980,70	23.111,22	0,734

Font: Elaboració pròpia

A la taula es mostra un total de 22 models construïts seguint la metodologia *stepwise*. Com es pot apreciar, per a cadascun d'aquests s'informa del número de graus de llibertat del model, el valor de la Deviància associada, el valor pel criteri d'informació d'Akaike, el valor pel criteri d'informació de Bayes i el p-valor relatiu al test de diferència de Deviàncies entre models niats. Així, és correcte comentar que:

- Si es pren com a exemple el test de diferència de deviàncies entre els models niats $MCV3$ i $MCV4$ serviria per comprovar si la interacció $distancia:angles$ aporta informació al model. Les hipòtesis nul·la i alternativa, així com l'estadístic de prova són:

$H_0: MCV3\ s'ajusta\ millor\ vs$ $H_1: MCV4\ s'ajusta\ millor$	$\Delta D_{MCV3,MCV4}$ $= D(y, \hat{\pi}_{MCV3})$ $- D(y, \hat{\pi}_{MCV4}) \sim \chi^2_{p-q}, sota\ H_0$

  Essent $D(y, \hat{\pi}_{MCV3})$ la deviància del model $MCV3$, $D(y, \hat{\pi}_{MCV4})$ la deviància del mdeol $MCV4$, p el número de paràmetres del model ampliat (en aquest cas $MCV4$) i q el número de paràmetres del model resduït (en aquest cas $MCV3$). Ja que té un p-valor associat de $0{,}2923$, amb una confiança del 95%, es pot concloure que no hi ha evidències estadístiques per refusar la hipòtesis nul·la que el model $MCV3$ s'ajusta millor i, per tant, no s'ha d'incloure la interacció $distancia:angles$.

- Amb tot, si es para atenció als criteris d'informació d'Akaike, de Bayes i de la Deviància, es pot concloure el procés de modelització seleccionant el model $MCV15$ com aquell que té un millor ajust. El vector de variables explicatives amb el seu coeficient associat del model de predicció de la probabilitat de gol d'un tir és el que segueix:

$$x'\beta = -1{,}34 - 0{,}09distancia + 1{,}37angles + 0{,}53Counter(yes) + 0{,}40Rol(x = Forward) - 0{,}03Rol(x = Goalkeeper) + 0{,}28Rol(x = Midfielder) + 0{,}23centrada(x = 1) + 0{,}12GB(x = good) - 079GB(x = neutral) - 0{,}03distancia:centrada(x = 1)$$

- Es pot observar que no s'han calculat els coeficients per a les categories *Defender* de la variable *Rol* i *bad* de la variable *GB* ja que R les ha pres com a categories referència i es troben incloses al terme independent.
- Totes aquelles variables explicatives paràmetres de les quals siguin de signe positiu, faran augmentar la probabilitat condicional de marcar gol quan augmentin els seus valors. Aquest és el cas de les variables *angles*, *Counter (x=1), Rol (x = forward)*, *Rol (x = Midfielder)*, *centrada(x=1), GB (x = good)*. En canvi, aquelles que tinguin el signe negatiu la faran disminuir, com és el cas de les variables *distancia*, *Rol* (*x = Goalkeeper*), *centrada (x = neutral)* i *distancia:centrada (x= 1)*.
- Prenent com a referència la distància i angle mitjans d'un tir quan és gol, és interessant calcular la probabilitat que aquest tir sigui gol, sabent que el jugador juga a la posició de davanter i que ha xutat amb la seva cama dominant procedint d'una centrada i d'un contra-atac:

$$\pi = \frac{e^{(-1{,}34-0{,}09*13{,}24+1{,}37*0{,}6389586+0{,}53+0{,}40+0{,}12+0{,}23-0{,}03*0{,}6389586)}}{1 + e^{(-1{,}34-0{,}09*13{,}24+1{,}37*0{,}6389586+0{,}53+0{,}40+0{,}12+0{,}23-0{,}03*0{,}6389586)}} = 0{,}4024$$

- L'*Odds-ratio* que la jugada sigui un contra-atac respecte el fet de marcar gol és $e^{0,53} = 1,6989$. La interpretació correcta d'aquest valor seria que el fet que una jugada sigui un contra-atac augmenta el quocient de probabilitats de la mateixa que acabi sent gol en gairebé un 70%. Des del punt de vista futbolístic, té sentit que la probabilitat augmenti tant ja que una situació de contra-atac implica, en la majoria de casos, una situació de superioritat numèrica pel que fa els atacants respecte els defenses. Un contra-atac, normalment, sorgeix d'una pilota robada al contrari i quan l'equip que té la possessió de la pilota la perd, els seus jugadors es troben descol·locats pel que fa a la seva situació normal defensiva i, per tant, el rival té uns instants d'avantatge abans que l'equip es re-col·loqui per defensar bé. És, efectivament, en aquest interval de temps que l'equip que ha robat la pilota ha d'aprofitar per ser el més vertical possible i arribar a la porteria contrària el més ràpid possible.
- L'*Odds-ratio* que el xut a porteria sigui efectuat amb la cama dominant del jugador respecte el fet de marcar gol és $e^{0,12} = 1,1274$. La interpretació correcta d'aquest valor seria que el fet que el jugador xuti amb la cama dominant augmenta el quocient de probabilitats o *odds* del xut que acabi essent gol en un $12,74\%$. Des d'un punt de vista futbolístic, la cama dominant és aquella amb la qual es té major control de la pilota i major habilitat i, per tant, és indiscutible que efectuar un xut amb aquesta faci augmentar la probabilitat que acabi sota la xarxa.
- L'*Odds-ratio* que el xut, des de la distància mitjana dels xuts que han acabat sent gol, a porteria que sigui precedit d'una centrada respecte el fet de marcar gol és $e^{0,23-0,03distancia} = e^{0,23-0,03*13,24} = 0,8460$. La interpretació correcta d'aquest valor seria que el fet que el xut, des d'una distància de $13,24$ metres, procedeixi d'una centrada provoca la reducció de l'*odds* que acabi essent gol en un $15,40\%$. Per altra banda, si es vol conèixer la distància màxima per la qual el factor $centrada$ té un efecte, si més no, neutre s'hauria de resoldre l'equació següent: $e^{0,23-0,03distancia} = 1 \Leftrightarrow \ln\left(e^{0,23-0,03distancia}\right) = \ln(0) \Leftrightarrow 0,23 - 0,03distancia = 0 \Leftrightarrow distancia = 7,\hat{6}$. Així, l'efecte positiu del factor $centrada$ sobre el quocient de probabilitats que sigui gol només es donarà per aquells xuts que es facin a una distància menor de $7,6$ metres de distància respecte la porteria. Altrament, per a qualsevol xut que superi els $7,6$ metres de distància, que la jugada procedeixi d'una centrada tindrà un efecte negatiu sobre l'*odds* que aquell xut acabi essent gol.
- L'*Odds-ratio* que el xut el faci un davanter o un mig-campista respecte que el mateix acabi essent gol és $e^{0,40} = 1,4918$ i $e^{0,28} = 1,3231$, respectivament. És a dir, el fet que el jugador que faci el xut sigui un davanter o un mig-campista fa augmentar l'*odds* del xut que acabi essent gol en un $49,18\%$ i un $32,31\%$, respectivament. Aquests resultats lliguen amb el que s'havia dit a la part d'anàlisi descriptiva: els davanters tenen una major capacitat golejadora en comparació als mig-campistes, la qual com es pot veure es reflexa en l'impacte o efecte sobre el quocient de probabilitats.

1. Exemple d'anàlisi de jugadors

Com ja s'ha comentat en l'apartat inicial d'aquesta secció, la principal mesura que es pot construir amb les prediccions d'aquest tipus de model són els gols esperats d'un jugador, la qual és definida com la suma de les probabilitats associades a cada tir que ha efectuat. Així,

disposant d'aquesta característica i els gols que realment ha marcat, el que en primera instància es pot realitzar és una comparació d'aquestes dues mètriques la qual permetria veure quins són aquells que han rendit, en termes de gols, per sobre, o per sota, del que s'espera del *jugador mitjà*.

Taula 4.1.4.1 Taula del top 10 golejadors de les 5 lligues (temporada 2017/2018)

Nom	***Cognom***	***Gols***	***Gols Esperats***	***Diferència***	***Xuts***
Mohamed	Salah	31	19,03	11,97	136
Harry	Kane	27	23,75	3,25	162
Lionel	Messi	26	19,15	6,85	142
Luis	Suárez	24	21,03	2,97	103
Edinson	Cavani	24	18,43	5,57	93
Cristiano	Ronaldo	23	24,09	−1,09	151
Robert	Lewandowski	23	21,04	1,96	106
Mauro	Icardi	23	15,65	7,35	87
Ciro	Immobile	22	13,50	8,50	102
Iago	Aspas	19	13,72	5,28	85

Font: Elaboració pròpia

A la taula 4.1.4.1 es mostren els top 10 màxim golejadors d'entre les 5 lligues europees que es disposa, dels quals s'anota els gols marcats, el gols esperats, la diferència d'ambdues primeres variables i el número de tirs que ha efectuat cadascú. En general, dels 10 jugadors que es poden visualitzar, només 1 ha marcat menys gols del que s'esperava pels tirs que ha efectuat. Aquest és *Cristiano Ronaldo*, que amb una diferència d'$1{,}09$ és l'únic jugador que ha rendit, en termes de gols, per sota del que hauria fet el "jugador mitjà" amb els tirs que ha fet. Altrament, es pot destacar l'actuació del jugador egipci *Mohamed Salah*, el qual és el màxim golejador amb 31 gols i, d'entre els 10 golejadors, també es caracteritza per ser el que més s'allunya del que s'esperaria del "jugador mitjà" en executar els seus xuts, és a dir, el que millor ha rendit per sobre de les expectatives, amb una diferència positiva d'$11{,}97$.

Seguint amb l'anterior, és cert que amb la comparació dels gols efectuats i els gols esperats es pot determinar una primera anàlisi. Però, aquesta no deixa de ser superficial i insuficient. Per exemple, en el cas del jugador portuguès, que la diferència sia negativa, per una banda, efectivament, pot significar que hagi acomplit per sota de les expectatives. Però, per altra banda, el fet que la diferència sigui petita vol dir que ha estat capaç de generar oportunitats amb probabilitats altes de ser gol, la qual cosa és una característica molt positiva i que, a primera instància, no es diria . Contràriament, en el cas del davanter Salah, els seus gols esperats es troben molt per sota de la quantitat de gols que ha marcat realment, la qual cosa indicaria que la qualitat de xuts que ha efectuat no és gaire bona (tot i que ha sigut capaç de

marcar-los, la qual cosa evidentment és favorable). Per tant, es comprova que els gols esperats també permeten avaluar la qualitat de les oportunitats que té cada jugador, com ja s'havia apuntat en la seva definició.

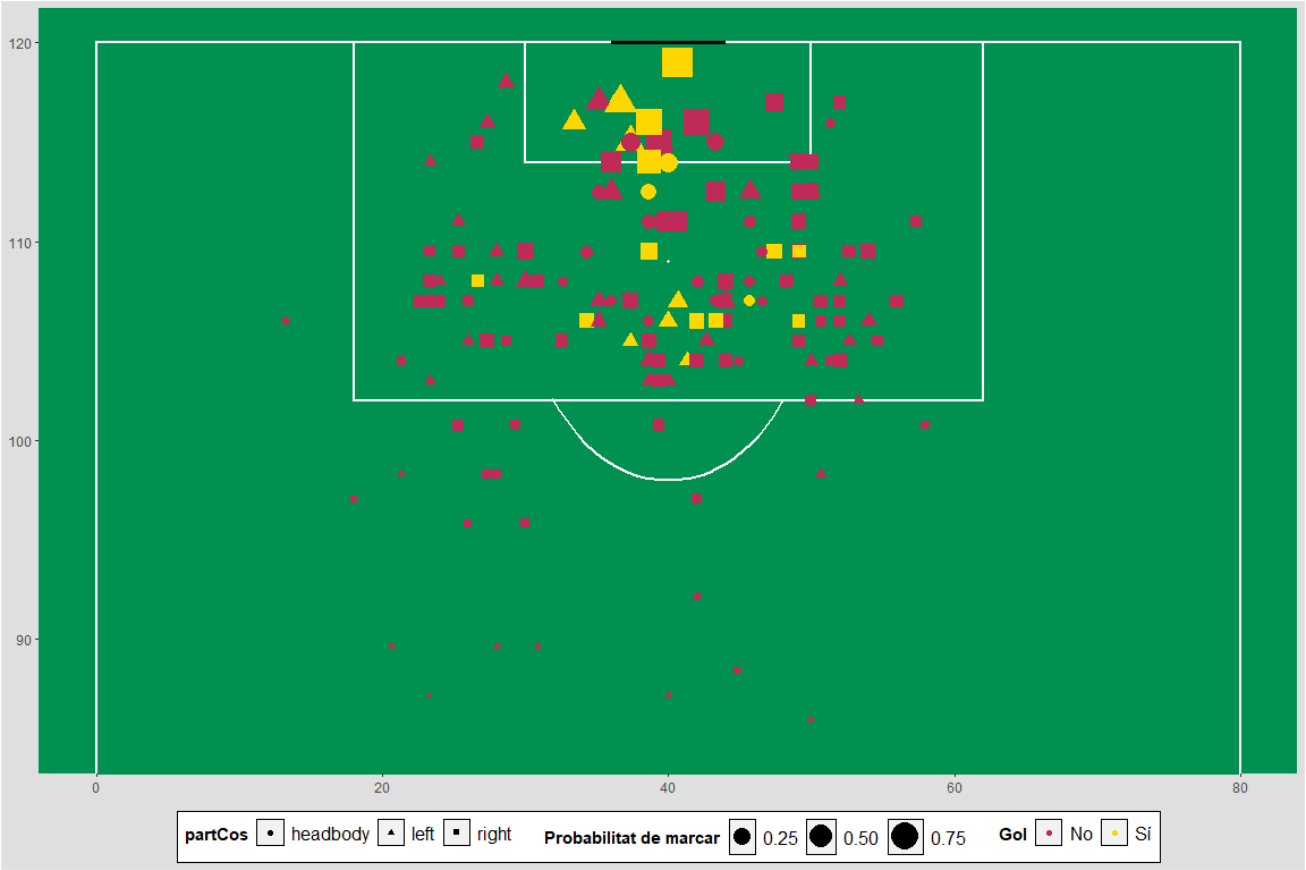
Nogensmenys, els *xG* permeten elaborar un anàlisi individual del futbolista més exhaustiu amb l'elaboració d'un mapa on es puguin visualitzar les posicions de cada xut a porteria, amb quina part del cos s'ha fet el xut, si ha sigut gol, o no, i la probabilitat que havia que acabés sota la xarxa aquella oportunitat. A la il·lustració 4.1.4.1 es mostra el mapa de xuts pel jugador del Reial Madrid *Cristiano Ronaldo*:

Taula 4.1.4.2 Taula resum dels gols i xGols per en Cristiano Ronaldo (temporada 2017/2018)

Nom	***Cognom***	***Gols***	***xGols***	***Dreta***	***Esquerra***	***Cap***	***Distància***	***Angle***	***Cama dominant***
Cristiano	Ronaldo	23	24,09	80	42	29	16,03	0.47	right

Font: Elaboració pròpia

Il·lustració 4.1.4.1 Mapa de tirs de Cristiano Ronaldo (temporada 2017/2018)



Font: Elaboració pròpia

Com es pot apreciar, en aquest mapa de xuts els punts dibuixats a sobre del camp marquen la posició exacta de tots els tirs que ha fet el jugador al llarg de la temporada. La seva forma, ja sigui triangle, circumferència o quadrat, indica la part del cos amb la qual es va fer el tir. El

color indicaria si el tir va acabar sent gol, color groc, o no, color grana. Finalment, la mida dels punts manifesta la probabilitat de marcar de cada tir determinada per les característiques que s'han establert al model: quant més gran sigui el punt dibuixat, major és la probabilitat de marcar del tir en qüestió.

Respecte l'anàlisi, com a primera anotació es pot apreciar que la majoria dels tirs del jugador es concentren dins l'àrea gran de la porteria i, per contra, no intenta xutar des de distàncies llargues. Seguint amb això, es pot observar que quant més a prop de la porteria i quant més centrada és la seva posició de tir, majors són els punts dibuixats. A més a més, s'ha de destacar que la majoria dels tirs que han acabat sent gol han sigut executats dins la zona central delimitada per l'àrea petita. També, s'ha de destacar la seva diversitat de tir quan es contempla la forma dels punts. Per tant, el que es podria extreure d'aquesta breu anàlisi és que el *Cristian Ronaldo* és un jugador capaç de posicionar-se bé per crear oportunitats amb altes probabilitats de marcar gol i que té l'habilitat de no només xutar, sinó de marcar tant amb la cama dominant (cama dreta) com amb la cama "dolenta" i el cap.

De la mateixa manera que s'ha fet una anàlisi individual, també es pot utilitzar aquesta eina gràfica per a realitzar comparacions entre futbolistes i determinar de manera més clara i visual les seves diferències i, fins i tot, detectar característiques que amb altres tipus d'anàlisi no seria pas possible.

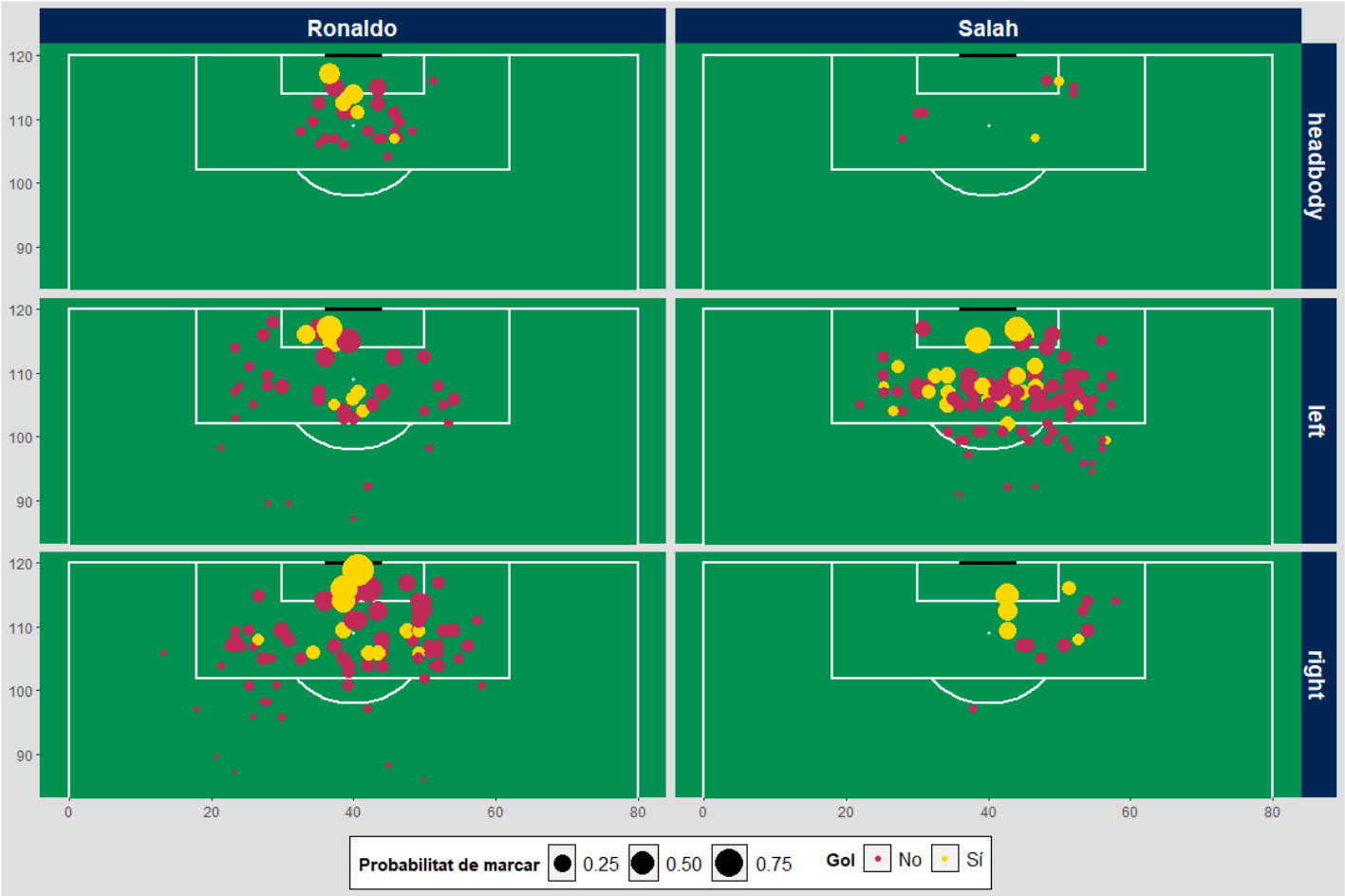
Efectivament, i seguint prenent com a exemple els futbolistes Cristiano Ronaldo i Mohamed Salah, la posada en escena dels mapes de manera conjunta en forma de matriu permet fer una anàlisi comparatiu més fàcil i visible. Se segueix mostrant els mateixos elements que en el mapa individual, però en aquest cas la disposició en columnes fa referència a cada jugador, indicant-se en el títol de cadascuna, i les files on es dibuixen els diferents mig-camps de futbol fan referència a la part del cos amb la qual han fet els tirs a porteria: la primera fila correspon a les rematades de cap, la segona amb la cama esquerra i, finalment, la tercera amb la cama dreta.

Taula 4.1.4.3 Taula resum dels gols i xGols per Cristiano Ronaldo i Mohamed Salah

Nom	***Cognom***	***Gols***	***xGols***	***Dreta***	***Esquerra***	***Cap***	***Distància***	***Angle***	***Cama dominant***
Cristiano	Ronaldo	23	24.09	80	42	29	16.03	0.47	right
Mohamed	Salah	31	19.03	14	114	8	17.04	0.38	left

Font: Elaboració pròpia

Il·lustració 4.1.4.2 Mapa de tirs de Cristiano Ronaldo i Mohamed Salah (temporada 2017/2018)



Font: Elaboració pròpia

Pel que fa a l'anàlisi comparativa dels dos jugadors, clarament es repara en la distribució de les rematades per a cadascun. Per una banda, ambdós jugadors efectuen la majoria dels seus tirs amb les seves corresponents cames dominants, Cristiano amb la dreta i Salah amb l'esquerra, i també és amb aquesta amb la qual aconsegueixen marcar més gols. Si es fixa en el mig-camp de Mohamed Salah amb la cama esquerra també es pot notar una certa concentració de xuts a la part dreta de l'àrea gran, la qual cosa determinaria la seva posició dins el terreny de joc (certament, Salah és un davanter que juga escorat a la banda dreta) i també el seu estil de joc: partint des de la banda dreta, el jugador aprofita la seva habilitat amb la seva cama dominat esquerra per ficar-se cap a dins i efectuar els seus tirs. Tanmateix, de Cristiano Ronaldo no es podria dir el mateix ja que les posicions de tir estan repartides de manera més homogènia entre la part esquerra, central i dreta, tot i que en moltes ocasions parteix des de la mateixa posició que Salah però a la inversa.

Per altra banda, la distribució de xuts segons la part del cos demostra una gran diferència entre els dos jugadors. Mentre que només el $16{,}17\%$ dels xuts de Salah són amb el cap o la cama menys dominant, en el seu cas la cama dreta, Cristiano Ronaldo xuta el $34{,}42\%$ de les vegades amb la cama menys hàbil si només es té en compte els tirs efectuats amb les cames, i el $19{,}20\%$ de les oportunitats amb el cap. En altres paraules, Cristiano Ronaldo recull una major disparitat en quant a la capacitat de rematada, és a dir, pot xutar amb ambdues cames,

independentment de si és amb la dominant o no, i també pot rematar de cap, la qual cosa el fa un total i absolut dominador de l'àrea i fa que els defenses no sàpiguen què farà i se'ls fa molt difícil predir els seus moviments.

2. Exemple d'anàlisi de partits

Una altra de les aplicacions que se'n deriven d'aquests càlculs és la d'anàlisi de partits a partir de l'exposició de l'evolució dels gols esperats acumulats d'ambdós equips al llarg del temps de durada del partit. Aquest tipus d'enfoc facilita l'explicació del que ha transcorregut realment al partit, a diferència de quan només es mostren estadístiques globals com els gols, tirs o targetes grogues.

Per a mostrar l'anterior, el primer exemple que es mostrarà serà *El Clásico*[5] que va tenir lloc a l'estadi del FC Barcelona, el *Camp Nou*, el 6 de maig del 2018 durant la jornada 36 de *LaLiga*. El Barça, 1^r classificat, rebia el Real Madrid, 3^r classificat, amb una diferència de 15 punts favorable als blau-granes. El resultat final del partit va ser un empat a 2 gols, on el Barça es va posar per davant en el marcador en dues ocasions, primer amb un gol del davanter Luís Suárez al minut 9 de partit i, més tard, amb un gol de Lionel Messi al minut 52 de la segona part, després que Cristiano Ronaldo pel Madrid empatés al minut 14 de la primera. Finalment, el minut 72 el jugador *blanc* Gareth Bale realitza el $2-2$, sent aquest el marcador final. Cal ressaltar, però, que durant tota la segona part el bloc local va jugar amb un home menys degut a l'expulsió de Sergi Roberto a les acaballes de la primera.

Així, i observant també les estadístiques del partit d'ambdós equips a la taula 4.1.5.1, a priori es podria dir que va ser un partit si més no igualat, però que el Barça tenia les de guanyar ja que es va posar per davant en el marcador fins a dues ocasions, fins i tot quan jugava amb un jugador menys. Cal comentar, però, que el club visitant va disposar de més xuts a porteria.

Ara bé, si es presta atenció al gràfic cronològic dels gols esperats acumulats (figura 4.1.5.1), o *xGols*, la interpretació del partit pren una altra forma. Com es pot observar, a l'eix horitzontal es representen els minuts i a l'eix vertical els xGols acumulats. Cada vegada que un dels equips disposa d'un xut, la seva línia dona un salt en funció de la probabilitat que tenia aquell xut de ser gol. Els punts de color daurat indiquen que el xut és gol i el punt de color vermell informa de l'expulsió per part de l'equip local. La línia vertical negra que travessa tot el plànol serveix per indicar l'inici de la segona part del partit. Així, si es presta atenció al seu contingut, la línia blanca, fent referència al Real Madrid, es troba per sobre de la del Barça, la línia de color grana, durant tot el partit des que marca el gol de l'empat. Com a conseqüència, la suma final dels xGols també resulta major pel club *blanc* (taula 4.1.5.2): mentre que els xuts del Real Madrid acumulen una probabilitat que significaria quasi bé 2 gols, $1{,}9788$, les oportunitats del FC Barcelona amb prou feines superen la unitat, $1{,}2799$. Per tant, i a diferència del que ha sigut la primera interpretació, el Real Madrid ha sigut el clar dominador del partit disposant de més i de les millors ocasions i, per tant, es podria dir que hagués sigut just que hagués acabat guanyant el partit.

[5] *El Clásico* és el nom que rep l'enfrontament entre el *Futbol Club Barcelona* i el *Real Madrid Club de Fútbol.*

Taula 4.1.5.1 Taula resum de les estadístiques de "El Clásico"(jornada 36, temporada 2017/2018)

Estadístiques	FC Barcelona	Real Madrid
Gols	**2**	**2**
Tirs	10	15
Tirs a porteria	4	5
Còrners	6	5
Faltes	19	11
Penaltis	0	0
Centrades	3	23
Passades intel·ligents	14	8
Passades penetrants	10	10
P.P.P	2	0
Targetes grogues	3	5
Targetes vermelles	1	0

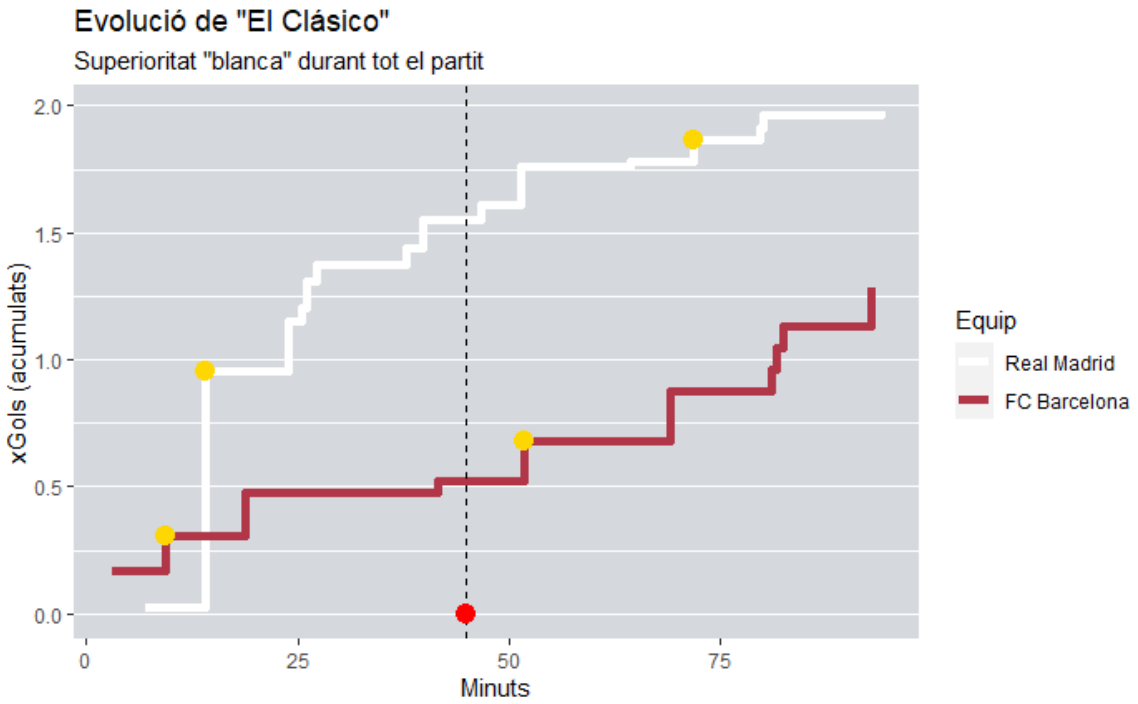
Font: Elaboració pròpia

Taula 4.1.5.2 Taula resum dels gols, gols esperats i tirs de "El Clásico" (jornada 36, temporada 2017/2018)

Estadístiques	FC Barcelona	Real Madrid
Gols	2	2
xGols	**1,279903**	**1,978899**
Tirs	10	15
Distància mitjana	15,68136	20,23618
Angulositat mitjana	0,3998583	0,4483177

Font: Elaboració pròpia

Figura 4.1.5.1 Gràfic cronològic de gols esperats de "El Clásico" (jornada 36, temporada 2017/2018)



Font: Elaboració pròpia

Continuant amb l'anàlisi anterior, una altra informació que es pot extreure del gràfic cronològic són els *moments de superioritat relativa* d'un equip sobre l'altre que es donen durant el partit. Aquestes situacions es descriuen com intervals de temps en que un equip pressiona més al rival i té major possessió de la pilota, donant com a resultat oportunitats de gol. Al gràfic s'observa aquest tipus de circumstància quan la línia de l'equip que està sent superior presenta salts molt continuats i , al mateix temps, la de l'altra equip es manté plana. Quan les línies dels dos equips són planes voldria dir que durant aquella franja de temps cap dels dos equips ha disparat a porteria. En l'exemple que s'exposa, en l'interval de temps que va des del minut 25 fins el minut 45 (final de la primera part) es pot observar que el Real Madrid va disposar de fins a 6 ocasions, mentre que el FC Barcelona només en va executar 1, moment en el qual l'equip visitant va ser superior i va disposar de bones d'oportunitats. Al final de la primera part, el Real Madrid acumulava una quantitat lleugerament superior a $1{,}5$ xGols, mentre que el FC Barcelona acumulava $0{,}5$ xGols.

2. Anàlisi economètrica dels valors de mercat

1) Anàlisi descriptiva

Per tal de designar quins són els determinants que formarien part del model economètric és convenient fer una anàlisi descriptiva de les variables que es disposa per observar les relacions que hi ha entre dites variables i, sobretot, la relació que hi ha amb la variable endògena o el valor de mercat.

1. Variable endògena: Valor de mercat

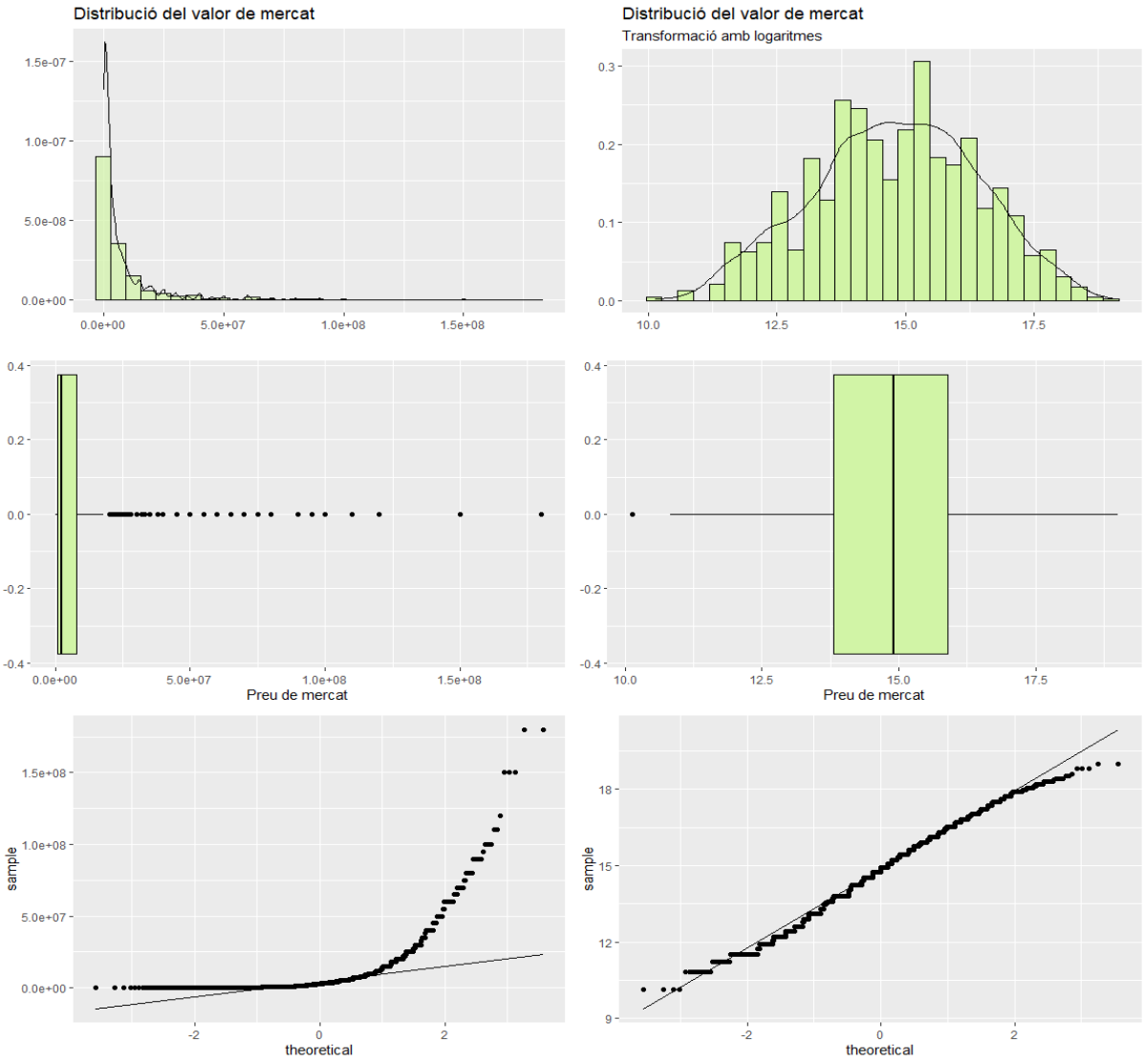
La variable que es vol modelitzar és la variable $Valor\ de\ mercat$, la qual defineix el valor de mercat del jugador a la temporada 2017/2018. Aquesta variable és una variable numèrica continua, domini de la qual són tots els números reals positius o $\mathbb{R}^+$.

Taula 4.2.1.1.1 Taula resum de la variable endògena Valor de mercat

Min.	*1st.Qu.*	*Mediana*	*Mitjana*	*3rd.Qu.*	*Max.*
0	800000	2500000	7851298.82	8000000	180000000

Font: Elaboració pròpia

Figura 4.2.1.1.1 Distribució de la variable endògena Valor de Mercat



Font: Elaboració pròpia

En els gràfics anteriors es pot observar la distribució de la variable endògena o explicada *Valor de mercat*, abans i després d'haver aplicat una transformació logarítmica. Clarament, veient l'histograma de la variable abans de transformar-la, es pot apreciar que es tracta d'una distribució que s'assembla a la d'una exponencial. El 70.226% dels jugadors es troben al primer interval, on el seu valor de mercat es troba per sota dels 6 milions d'€, freqüència que disminueix de cop al passar al següent interval i que anirà disminuint paulatinament. Un cop s'aplica la funció logarítmica a les dades, la forma tant de l'histograma com del diagrama de caixes canvia completament. Fixant-se en la corba de densitat, la distribució pren la forma d'una Normal encara que l'alçada de les barres centrals de l'histograma podrien indicar la d'una distribució bimodal. El gràfic "Normal Q-Q" apunta clarament cap a una distribució Normal de les dades. Cal comentar que quan s'aplica el logaritme s'han d'excloure aquells jugadors valor de mercat dels quals és igual a 0 ja que no es pot aplicar esmentada funció. El número total de jugadors que s'han exclòs, o en altres paraules, que tenien valor de mercat nul ha sigut de 129. En conseqüència, la mida de la mostra que s'ha utilitzat per a l'anàlisi ha quedat en 2.662 jugadors.

2. Variables categòriques o factors

Les variables categòriques o factors que es disposen són Rol i $Lliga$. Seguint la línia del que es diu a la hipòtesi inicial, a priori tindria sentit esperar que segons la posició del jugador al camp, el valor de mercat variï, sent el preu dels davanters major, en termes mitjans, que la resta de posicions.

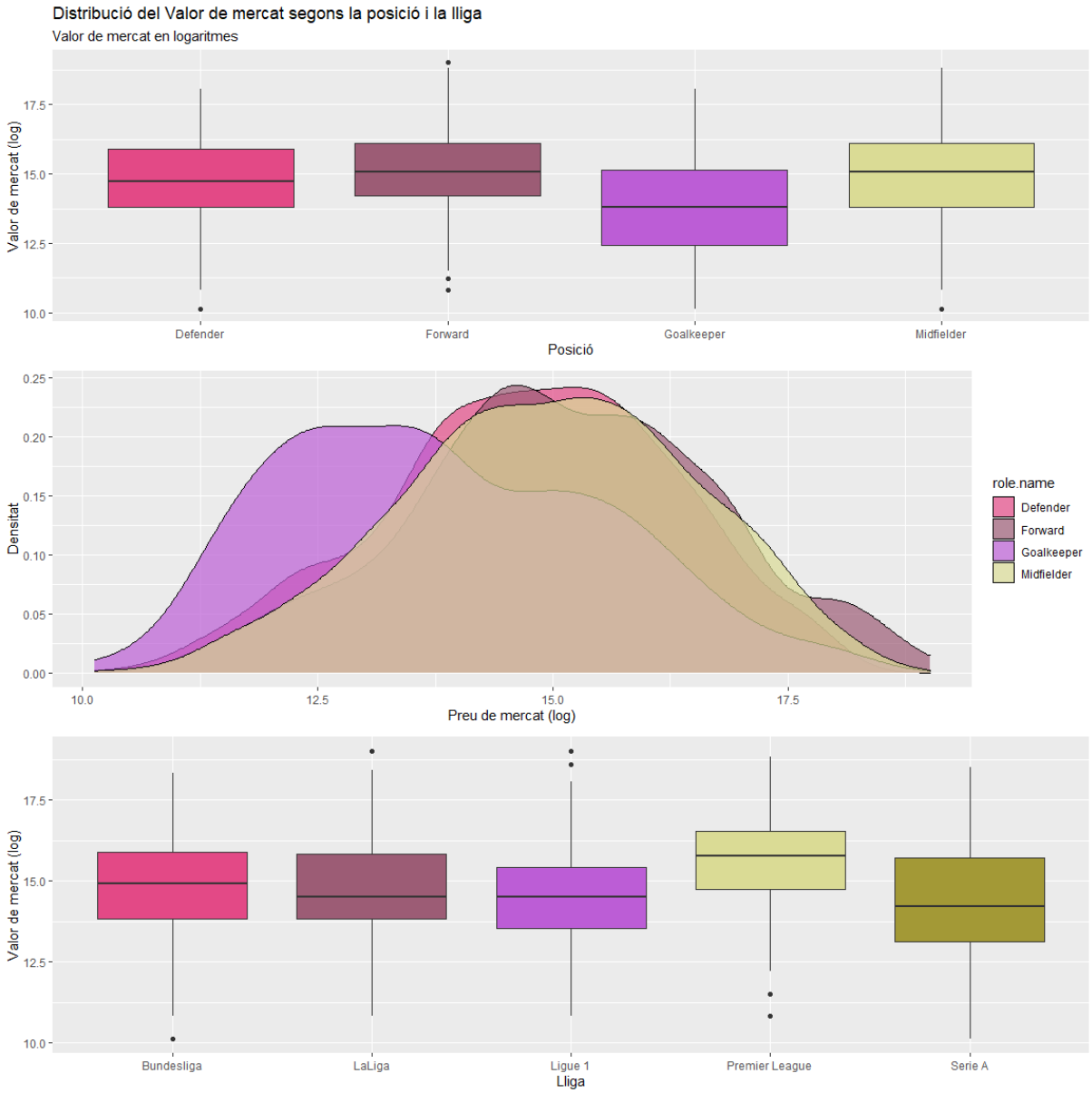
A la taula 4.2.1.2.1 s'observa que el top 12 dels jugadors més cars està format per 10 davanters (Categoria: $Forward$) i 2 Migcampistes (Categoria: $Midfielder$). Es destaca, contràriament, que fins la posició 32 no apareixen ni el porter ni el defensa més cars, *David de Gea* i *Varanne* respectivament, ambdós amb un preu de 70 milions d'€. Així mateix, en la figura 4.2.1.2.1 es representa gràficament mitjançant diagrama de caixes i corbes de densitat la distribució de la variable *Preu* segons el factor Rol i el factor $Lliga$ a la qual pertany el jugador.

Taula 4.2.1.2.1 Top 12 jugadors més valorats del mercat (temporada 2017/2018)

Nom	***Valor de mercat (en milions d'€)***	***Rol***
L. Messi	180	Forward
Neymar	180	Forward
K. De Bruyne	150	Midfielder
H. Kane	150	Forward
Mohamed Salah	150	Forward
K. Mbappé	120	Forward
P. Dybala	110	Forward
E. Hazard	110	Forward
D. Alli	100	Midfielder
Philippe Coutinho	100	Forward
A. Griezmann	100	Forward
Cristiano Ronaldo	100	Forward

Font: Elaboració pròpia

Figura 4.2.1.2.1 Diagrama de caixes de la variable endògena "Valor de mercat" en funció dels factors "Rol" i "Lliga"



Font: Elaboració pròpia

Pel que fa al rol del jugador, es pot observar que la posició de porter és la concentra valors de mercat més baixos, seguida pels defenses, dels mig-campistes i, finalment, dels davanters. La mitjana del valor de mercat per a jugadors que juguen en la posició de porter és de menys de la meitat dels davanters. En referència a aquests últims, no sembla que destaqui per sobre dels migcampistes, tot i que la diferència de mitjanes entre ambdues categories és de quasi bé 3 milions d'euros, com es pot observar a la taula 4.2.1.2.2. En general, el diagrama de caixes per a cadascuna de les posicions es caracteritzen per un rang interquantílic semblant (amplitud de la caixa) i una distribució simètrica (indicada per la posició central de la mediana). Es detecta major presència d'outliers en la posició de davanter.

Taula 4.2.1.2.2 Taula resum de la variable endògena "Valor de mercat" en funció del factor "Rol"

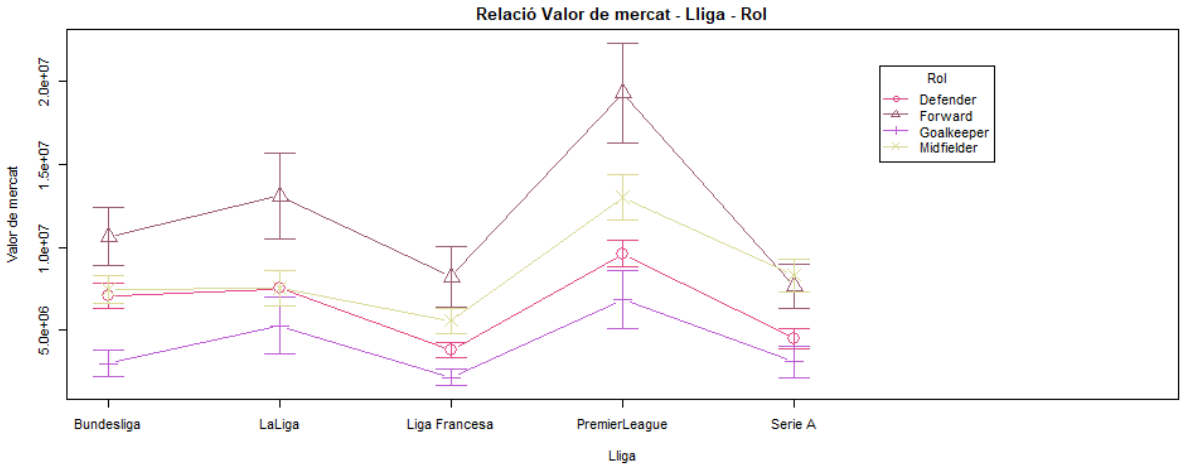
Rol	***Freq. Absoluta***	***Min.***	***Mitjana***	***Max.***	***Desv. Tip.***	***CV***
Defender	926	0	6.426.268,90	70.000.000	10.276.127,76	1,60
Forward	578	0	11.328.806,23	180.000.000	22.656.996,62	2,00
Goalkeeper	317	0	4.003.312,30	700.00.000	9.688.029,53	2,42
Midfielder	970	0	8.399.123,71	150.000.000	14.268.135,15	1,70

Font: Elaboració pròpia

Per altra banda, quan s'agrupen els valors de mercat segons la lliga dels jugadors, clarament es veu que els valors de mercat dels jugadors que juguen a la Lliga anglesa són els més cars, seguit de les lligues espanyola i alemanya. A més a més, es detecta asimetria en la majoria de lligues i presència de valors atípics en totes menys a la lliga italiana. Seguint amb l'anàlisi descriptiva, per detectar un possible efecte creuat i evitar confusions és interessant relacionar els dos factors anteriors i analitzar-los de manera conjunta i amb diferents covariables, a part de la variable endògena.

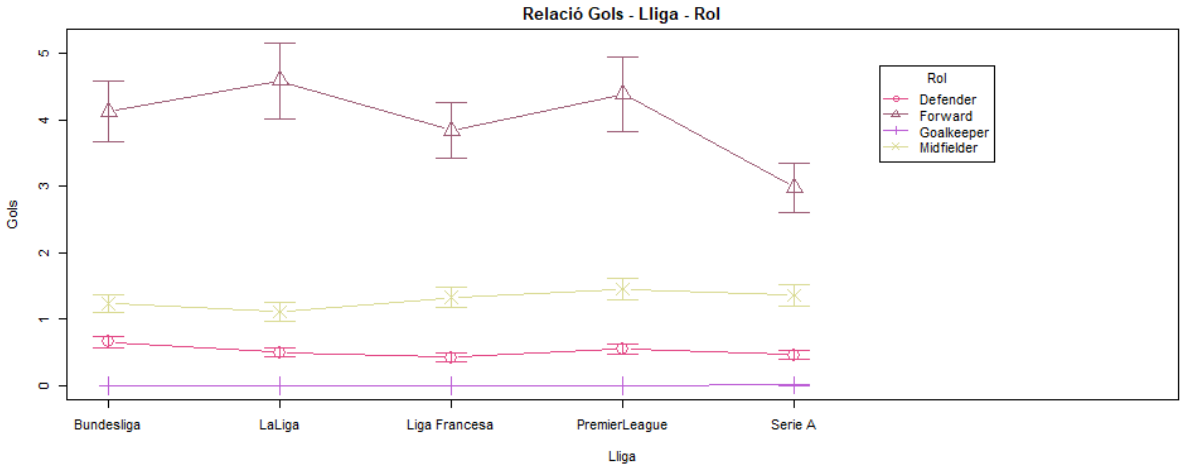
A la figura 4.2.1.2.3 es mostren les mitjanes de la variable endògena $Valor\ de\ mercat$ en funció dels factors $Lliga$ i Rol. A l'eix de les abscisses es representen les categories del factor $Lliga$, a l'eix d'ordenades es representen les mitjanes del valor de mercat i el color i símbol representen les diferents categories del factor Rol. Com ja s'ha vist anteriorment, els valors de mercat de la lliga anglesa (categoria $PremierLeague$) són els més elevats independentment del rol del jugador i, altrament, els davanters (categoria $Forward$) són els jugadors més valorats en termes monetaris independentment de la lliga que es tracti, seguit dels mig-campistes, defenses i, finalment, porters. En general, s'observen els mateixos patrons que ja s'han vist. Es podria destacar, però, que a les lligues alemanya (categoria $Bundesliga$) i espanyola (categoria $LaLiga$) els preus dels defenses (categoria $Defender$) i mig-campistes (categoria $Midfielder$) es superposen, igualment que els davanters i mig-campistes a la lliga italiana. És a dir, sembla ser que el valor dels mig-campistes sí que varia en funció de la lliga en la que es jugui: en termes relatius, a les lligues alemanya i espanyola es podria dir que els migcampistes estan infravalorats o pitjor valorats ja que els seus preus són semblants als dels defenses i, en general, els preus d'aquests últims són menors; per altra banda, a la lliga italiana els migcampistes tenen preus semblants als davanters, així que es podria dir que estan sobrevalorats o millor valorats.

Figura 4.2.1.2.5 Gràfic de la mitjana de la variable "Min" en funció dels factors "Rol" i "Lliga"



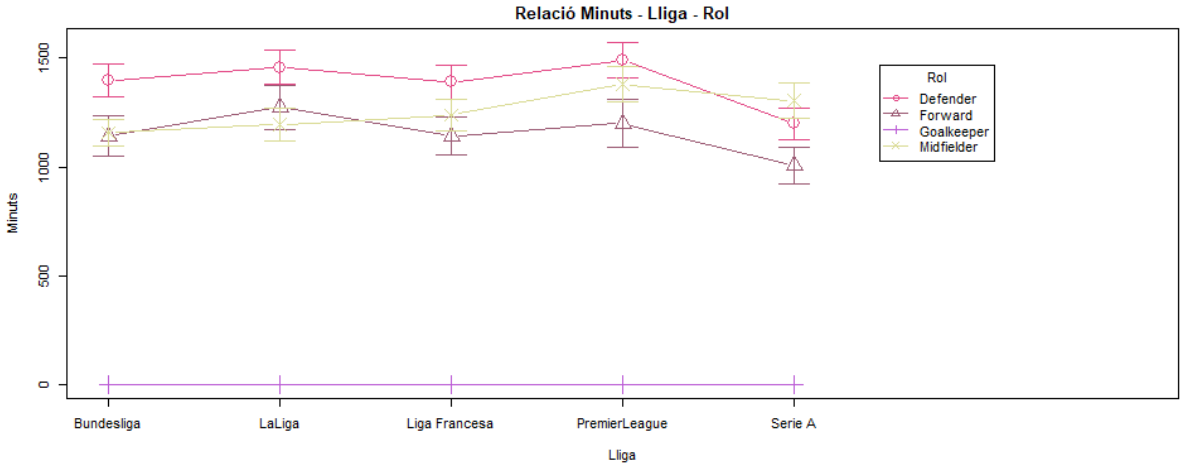
Font: Elaboració pròpia

Figura 4.2.1.2.4 Gràfic de la mitjana de la variable "Gols" en funció dels factors "Rol" i "Lliga"



Font: Elaboració pròpia

Figura 4.2.1.2.3 Gràfic de la mitjana de la variable endògena "Valor de mercat" en funció dels factors "Rol" i "Lliga"



Font: Elaboració pròpia

Per altra banda, s'ha volgut observar la relació d'aquests dos factors amb la variable $Gols$, la qual informa del total de gols que ha marcat el jugador a la lliga regular durant tota la temporada, i amb la variable Min que indica el minuts jugats durant tota la temporada per a cada jugador i comparar-ne les estructures i patrons per tal de detectar possibles efectes confosos. Pel que fa a la primera, a la figura 4.2.1.2.3 s'observa una disposició de la variable $Gols$ molt similar a la de la variable $Valor\ de\ mercat$. A part que els davanters són els que més gols marquen (els més cars del mercat), també s'observa lleugerament la mateixa disposició pel que a les lligues: mentre que a les lligues alemana i espanyola les diferències de gols entre mig-campistes i defenses són les més petites, a la lliga italiana esdevé entre davanters i mig-campistes (patró si més no semblant al dels valors de mercat). Pel que fa a la variable Min, la disposició de la distribució de la variable diferenciada per Rol i per $League$ no s'assembla pas a la dels preus. En tot cas, es mostra una relació contrària a la dels preus i gols marcats: mentre que els defenses (i, òbviament els porters) són els que més minuts juguen per temporada a totes les lligues (exceptuant a la lliga italiana), són els més barats (tret dels porters).

Concloentment, a l'hora de construir el model s'haurà de tenir en compte que l'efecte del factor $League$ sobre els valors de mercat és clar individualment. Pel que fa a la seva interacció amb el factor Rol, com s'ha vist, només es tindria en compte per la categoria $Midfielder$ ja que en termes generals no s'observen indicacions de confusió. Nogensmenys, s'ha pogut observar que l'efecte del factor Rol es podria veure confós amb la variable potencialment explicativa $Gols$, sobre la qual cosa també s'hi haurà de prestar atenció.

3. Variables numèriques o covariables

En primera instància, cal comentar hi ha una clara singularitat respecte a com avaluar el rendiment futbolístic específicament d'una de les categories del factor Rol. Els porters, categoria *Goalkeeper*, per la naturalesa de la seva posició i funció, s'hauran de valorar a partir de les mètriques específiques que s'han recollit per aquest grup i, per tant, des del començament se separaran de la resta i formaran un *mercat* propi. Les variables o mètriques de rendiment específiques pel grup de porters són: $Interv$, $Parades$, $Parades\_pc$, $Reflx$, $Reflx\_pc$ i $Parades\_reflx\_pc$.

Pel que fa a l'anàlisi descriptiva de les variables numèriques, el que es fa és calcular el coeficient de correlació de Pearson (ρ) per a detectar la presència, o no, de relació entre dites variables i la variable endògena $Valor\ de\ mercat$ i també la seva direcció, positiva o negativa. Cal recordar, però, que l'existència de correlació no implica causalitat. Es pot afirmar que correlació simplement significa associació de fet. Així mateix, se seleccionaran aquelles variables on $\rho \geq 0{,}3$ o $\rho \leq -0{,}3$ com a covariables potencialment explicatives del model.

Taula 4.2.1.3.1 Taula de coeficients de correlació de Pearson de la variable endògena "Valor de mercat" respecte totes les variables potencialment explicatives numèriques pels jugadors amb rol "Davanter", "Mig-campista" i "Defensa"

Variable	$\rho_{Valor\ de\ mercat,Variable}$
Valor de mercat	1,00
Gols	0,59
xAssist	0,59
Pass_thro	0,57
Pass_intl	0,57
Pass_intl_ex	0,57
Xuts	0,57
Assist	0,56
VMTC	0,55
xGols	0,53
Dribl_ex	0,48
Dribl	0,47
Partits	0,46
Min	0,46
Participacio_Xuts	0,41
Participacio_Minuts	0,40
Faltes	0,37
Participacio_Gols	0,36
Participacio_xGoals	0,35
Gols_90	0,35
Gols_pp	0,34
Pass_ex	0,28
Penals	0,27
Pass	0,27
Cross	0,26
Ppp	0,26
Cross_ex	0,25
Assist_90	0,24
Duels	0,23
GolsXuts	0,22

Pass_thro_pc	0,22
Duels_ex	0,21
Pass_intl_ex_pc	0,20
Duels_aeris	0,17
Dribl_ex_pc	0,17
Duels_aeris_ex	0,15
Cross_ex_pc	0,13
Pass_ex_pc	0,11
Duels_ex_antic	0,10
Duels_ex_pc	0,10
Duels_aeris_ex_pc	0,08
Weight	0,06
Height	0,04
Parades_pc	0,04
Duels_ex_antic_pc	0,03
Parades	0,03
Interv	0,02
Reflx	−0,00
49Reflx_pc	−0,01
Parades_refl_pc	−0,01
Edat	−0,01
MVMC	−0,02

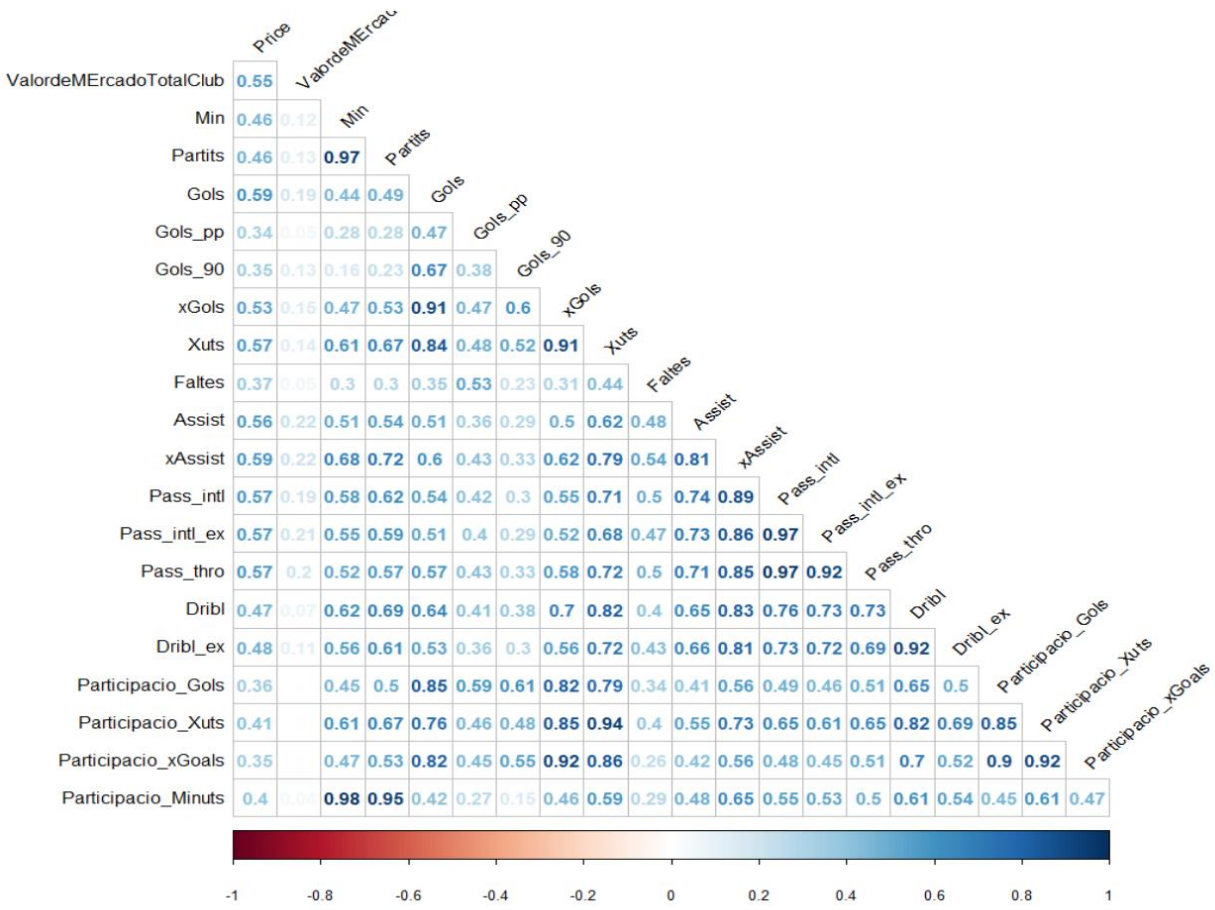
Font: Elaboració pròpia

Taula 4.2.1.3.2 Taula de coeficients de correlació de Pearson de la variable endògena "Valor de Mercat" respecte totes les variables potencialment explicatives numèriques pels jugadors amb rol "Porter"

Variable	$\rho_{Valor\ de\ Mercat,Variable}$
Valor de Mercat	$1{,}00$
xAssist	$0{,}49$
Parades	$0{,}46$
Reflx_pc	$0{,}43$
Interv	$0{,}43$
VMTC	$0{,}40$
Reflx	$0{,}40$
Ppp	$0{,}37$
Pass_ex	$0{,}34$
Pass	$0{,}34$
Duels_aeris	$0{,}34$
Parades_reflx_pc	$0{,}33$
Duels_aeris_ex	$0{,}33$
Parades_pc	$0{,}33$
Duels_aeris_ex_pc	$0{,}33$
Duels	$0{,}26$
Duels_ex_antic	$0{,}22$
Duels_ex_antic_pc	$0{,}22$
Pass_ex_pc	$0{,}20$
Duelx_ex	$0{,}19$
Duels_ex_pc	$0{,}18$
Weight	$0{,}10$
Height	$0{,}07$
Pass_intl	$-0{,}01$
MVMC	$-0{,}02$
Edat	$-0{,}05$

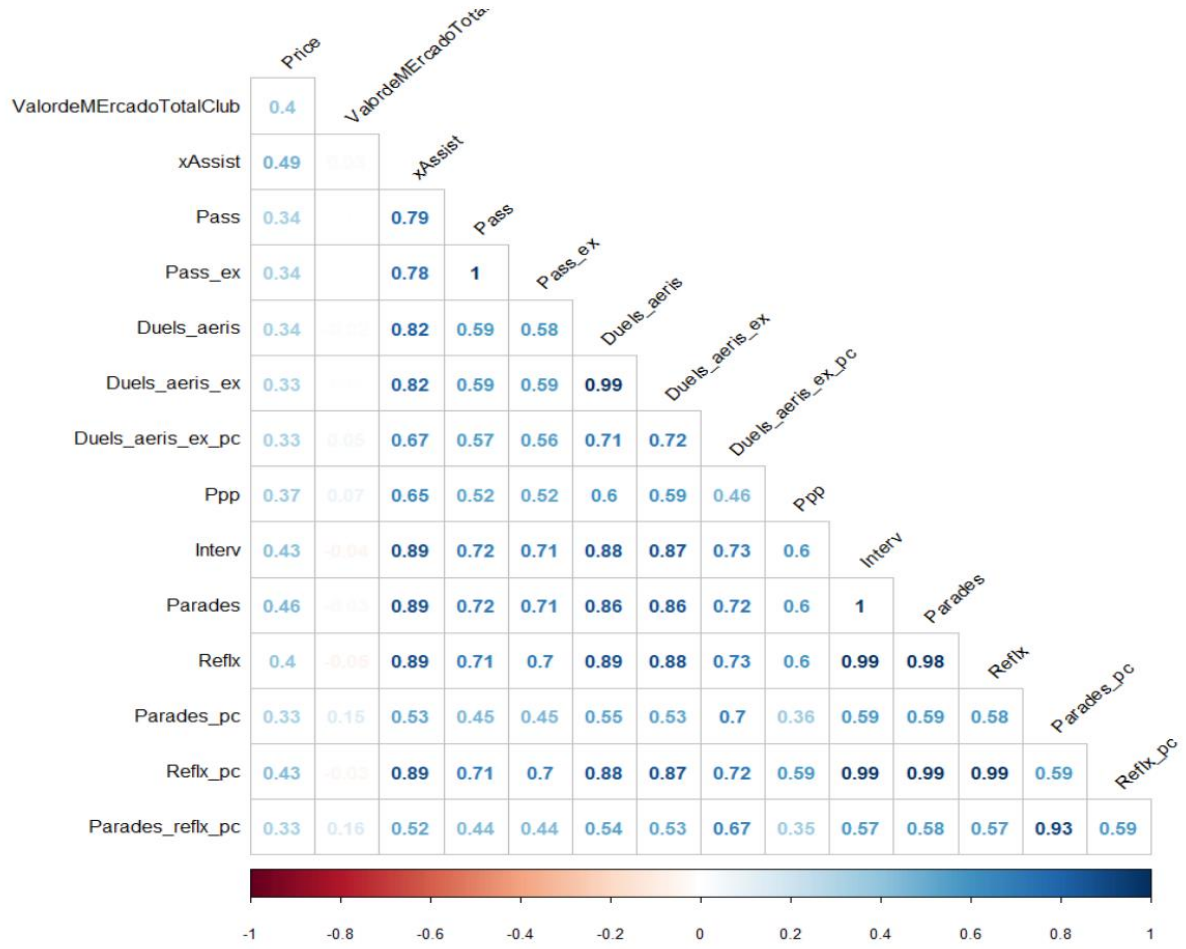
Font: Elaboració pròpia

Il·lustració 4.2.1.3.1 Matriu de correlacions de Pearson de les variables potencialment explicatives pels jugadors amb rol "Davanter", "Mig-Campista" i "Defensa"



Font: Elaboració pròpia

Il·lustració 4.2.1.3.2 Matriu de correlacions de Pearson de les variables potencialment explicatives numèriques pels jugadors amb categoria "Porter"



Font: Elaboració pròpia

Tal i com es pot observar, de les taules 4.2.1.3.1 i 4.2.1.3.2 on es mostren els coeficients de correlació de Pearson calculats per a la variable endògena i totes les covariables disponibles, havent diferenciat entre jugadors de camp i porters, s'han escollit 20 variables pel conjunt de jugadors davanters, mig-campistes i defenses, i 14 variables pel conjunt de porters, aquelles que complien el criteri de selecció; i s'ha construït una matriu de correlacions (il·lustracions 4.2.1.3.1 i 4.2.1.3.2) associada a cadascuna de les taules, de les quals es mostren els elements per sota de la diagonal. A través d'aquestes, es poden advertir futurs problemes de multicol·linealitat que, com es pot veure, existeixen (les variables independents estan altament correlacionades entre elles) i s'hauran de tenir en compte i tractar.

Respecte els resultats, tant pels jugadors de camp com pels porters, totes les variables independents que s'han seleccionat mantenen una relació positiva amb la variable endògena. Respecte del conjunt format per davanters, mig-campistes i defenses, és interessant comentar que la majoria de variables seleccionades són mètriques que mesuren la part atacant del joc, que tenen a veure amb gols marcats, xuts, assistències i driblatges (en total 12 variables). Per tant, de nou es passa pel cap la possibilitat de confondre l'efecte d'aquestes variables amb el factor Rol donat que els jugadors caracteritzats amb valors alts d'aquest tipus de variables són, en general, davanters i mig-campistes. Conseqüentment, això porta a la idea de la possible necessitat d'haver de diferenciar i fer anàlisis separades en funció de la posició dels

jugadors, a part de la diferenciació que ja s'ha fet amb el grup de porters, i veure si per a cada categoria de Rol existeix un mercat específic. Precisament, per tal d'esbrinar si és necessari fer aquesta distinció, a continuació se separa la base de dades que fa referència als jugadors de camp i es calcula igualment el coeficient de correlació de Pearson per la variable $Valor\ de\ mercat$ respecte els diversos indicadors de rendiment. A la taula 4.2.1.3.3 es mostra l'anterior on cada columna fa referència a una categoria del factor Rol diferent:

Taula 4.2.1.3.3 Taula de correlacions de Pearson de la variable endògena "Valor de mercat" respecte les variables potencialment explicatives numèriques diferenciant per les 3 categories del factor rol "Davanters", "Mig-campistes" i "Defenses"

Variable	***Davanters***	***Mig-Campistes***	***Defenses***
Valor de mercat	1,00	1,00	1,00
Weight	0,01	0,08	0,12
Height	−0,03	0,08	0,09
VMTC	0,62	0,57	0,59
MVMC	−0,02	−0,02	−0,01
Edat	0,06	−0,03	−0,05
Min	0,53	0,51	0,45
Partits	0,48	0,51	0,45
Gols	0,71	0,53	0,36
Gols_pp	0,44	0,19	0,08
Gols_90	0,51	0,17	0,21
xGols	0,63	0,53	0,40
Xuts	0,63	0,55	0,40
GolsXuts	0,28	0,17	0,21
Penals	0,36	0,10	0,06
Faltes	0,51	0,29	0,07
Assist	0,66	0,59	0,31
Assist_90	0,25	0,24	0,14
xAssist	0,65	0,62	0,38
Pass	0,35	0,34	0,36
Pass_ex	0,36	0,35	0,37
Pass_ex_pc	0,07	0,12	0,20
Pass_intl	0,64	0,60	0,33
Pass_intl_ex	0,64	0,59	0,35
Pass_intl_ex_pc	0,21	0,19	0,28

Pass_thro	0,64	0,60	0,30
Pass_thro_pc	0,20	0,23	0,17
Cross	0,36	0,34	0,17
Cross_ex	0,38	0,33	0,16
Cross_ex_pc	0,16	0,17	0,11
Dribl	0,51	0,48	0,24
Dirbl_ex	0,59	0,44	0,22
Dribl_ex_pc	0,26	0,16	0,18
Duels	0,24	0,35	0,30
Duels_ex	0,21	0,35	0,34
Duels_ex_pc	0,09	0,16	0,21
Duels_ex_antic	0,12	0,26	0,28
Duels_ex_antic_pc	0,09	0,08	0,12
Duels_aeris	0,06	0,23	0,34
Duels_aeris_ex	0,04	0,20	0,34
Duels_aeris_ex_pc	0,11	0,13	0,20
Ppp	0,29	0,34	0,35
Participacio_Gols	0,40	0,30	0,19
Participacio_Xuts	0,44	0,39	0,24
Participacio_xGoals	0,38	0,34	0,22
Particiapcio_Minuts	0,47	0,44	0,38

Font: Elaboració pròpia

Com es pot observar, en general per a la categoria *Davanters* els valors de ρ calculats són més elevats comparant-los amb les dues categories restants, menys per a aquells indicadors de caire defensiu: *Duels*, *Duels_ex*, *DuelsÇ_ex_pc*, *Duels_ex_antic*, *Duels_ex_antic_pc*, *Duels_aeris*, *Duels_aeris_pc* i *Duels_aeris_ex_pc*. Com era d'esperar, el valor dels coeficients de correlació que posen en correspondència el preu i el rendiment defensiu dels jugadors pren valors majors (i positius) per les categories *Mig-campistes* i *Defenses*. S'ha de comentar, però, que per aquests dos rols els indicadors de caire atacant continuen tenint valors elevats i positius, de manera que segueixen tenint importància. També cal remarcar, pel cas del grup de davanters, que variables que abans no havien superat el llindar establert ara sí que ho fan: són els casos de les varaibles *Cross* i *Cross_ex*.

Amb tot, es conclou que se separarà l'anàlisi de manera que per cada categoria de la variable Rol es construirà un model economètric.

2) Anàlisi economètrica

D'entrada, per a cadascuna de les categories de la variable Rol es construirà un model de regressió lineal múltiple additiu on la variable endògena serà el logaritme del $Valor\ de\ mercat$, que inclourà, per una banda, el factor $Lliga$ i, per una altra, com a covariables explicatives aquelles els coeficients de correlació de les quals tinguin valors majors que $0{,}3$.

A la taula 4.2.2.1 es mostren els primers models construïts amb la sortida de la funció $summary$ d'R, on es pot veure per a cadascun dels models construïts un resum que inclou els valors dels coeficients i la seva significació individual, entre d'altres. Respecte el darrer, s'observa que moltes de les variables no són estadísticament significatives. S'ha de tenir en compte que s'han inclòs dites variables segons un únic criteri, el referent al coeficient de correlació de Pearson, i per tant cal fer un escrutini detallat per a cadascun dels models per tal de construir-los el millor possible. A part, es pararà esment en el concepte de multicol·linealitat a l'hora de seleccionar les variables adients. Existeixen diverses variables per explicar un mateix concepte, qualitat o fet, de manera que són combinació lineal. A efectes de predicció, no pren molta rellevància ja que, al cap i a la fi, el que informa és que hi ha un excés d'informació redundant. La capacitat predictora del model es mantindria, els estimadors són consistents i l'estimació puntual seria vàlida. No obstant, quanta major relació hi hagi entre les mateixes variables explicatives del model, major serà la variància dels estimadors calculats, la qual cosa afectarà a l'amplitud dels intervals i, com sembla ser que succeeix, es podria indicar artificialment que les variables no són significatives individualment[6]. A més a més, un altre símptoma evident de col·linealitat múltiple és que els paràmetres no siguin individualment significatius, però sí de manera conjunta. Aquest concepte ja se n'ha fet esmena amb anterioritat i ja s'ha vist, en taules anteriors, que moltes de les variables potencialment explicatives que s'han obtingut estan altament correlacionades. Més endavant, és clar que s'haurà d'aturar a la comprovació que els supòsits dels models, efectivament, es compleixin.

Per tal d'intentar corregir l'existent multicol·linealitat dels models s'analitzarà els factors d'inflació de la variància, o *VIF* en les seves sigles en anglès, que és un altre mètode per detectar la multicol·linealitat del model. Els *VIF*'s quantifiquen la intensitat de la multicol·linealitat i es defineixen com:

$$VIF_j = \frac{1}{1 - R_j^2}$$

On R_j^2 és el coeficient de determinació o coeficient de bondat de l'ajust de la regressió múltiple del j −èssim regressor sobre la resta, és a dir, per a cada variable exògena del model es construeix un altre model de regressió múltiple essent aquesta la variable endògena i es calcula el seu R^2 associat. Per la forma de càlcul del quocient, el valor mínim del VIF és 1. Un $VIF > 10$ pot indicar existència de multicol·linealitat severa.

[6] L'existència de multicol·linealitat implica, a efectes matemàtics, que $|X^T X| \approx 0$, que está inclosa a l'estimació de la variància dels estimadors per MQO: $\widehat{var}\left(\hat{\beta}_{MQO}\right) = \hat{\sigma}^2 (X^T X) = \hat{\sigma}^2 \frac{Adj(X^T X)}{|X^t X|}$

Però, prèviament es considera l'extracció d'algunes variables de cada model amb les justificacions corresponents següents:

- Variable *Partits*. Aquesta variable explica la mateixa noció que la variable *Min*, la qual explica el temps total que ha jugat el jugador en tota la temporada. Es deixa de tenir en compte *Partits* ja que és més general.
- Variable *Pass*. Aquesta variable quantifica el número de passades que ha fet el jugador, correctes i incorrectes, la qual cosa es considera una variable més descriptiva que no pas explicativa.
- Variable *Xuts*. Aquesta variable recull el nombre de xuts a porta que ha fet el jugador. Es considera que el seu efecte es veu explicat per altres variables com *Gols* o *xGoals*.
- Variable *Gols_pp*. Aquesta variable. específica pels davanters, quantifica els gols per partit del jugador. L'efecte de la qual es veu contemplat per la variable *Gols_90*.
- Variable *Interv*. Aquesta variable, específica pels porters, quantifica el número de vegades que el jugador ha hagut d'intervenir en la jugada, entesa com un xut, independentment de si ha parat el xut o no. Es considera que és una variable més descriptiva que no pas explicativa.

Un cop eliminades les variables, s'elaboren els models per a cada categoria de nou i s'avaluen els VIF. S'ha de tenir en compte que els valors associats a cada variable majors de 10 es considera, doncs, que hi ha multicol·linealitat alta.

Taula 4.2.2.1 Taula resum de la primera estimació per mínims quadrats ordinaris separant per categories del factor "Rol"

	Davanters			Mig-campistes			Defenses			Porters		
Predictors	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***
(Intercept)	13,18 ***	12,95 – 13,42	**< 0,001**	13,13 ***	12,97 – 13,30	**< 0,001**	13,01 ***	12,81 – 13,21	**< 0,001**	12,06 ***	11,76 – 12,37	**< 0,001**
Lliga [LaLiga]	−0,18	−0,43 – 0,07	0,157	−0,43 ***	−0,60 – −0,25	**< 0,001**	−0,31 **	−0,52 – −0,10	**0,004**	0,24	−0,17 – 0,65	0,258
Lliga [Ligue 1]	−0,20	−0,44 – 0,05	0,119	−0,40 ***	−0,58 – −0,22	**< 0,001**	−0,32 **	−0,53 – −0,11	**0,003**	0,12	−0,26 – 0,49	0,539
Lliga [Premier League]	0,60 **	0,29 – 0,91	**< 0,001**	0,42 **	0,19 – 0,64	**< 0,001**	0,39 *	0,11 – 0,67	**0,007**	0,78 **	0,35 – 1,20	**< 0,001**
Lliga [Serie A]	−0,28 *	−0,52 – −0,04	**0,020**	−0,34 ***	−0,52 – −0,16	**< 0,001**	−0,54 ***	−0,74 – −0,34	**< 0,001**	−0,24	−0,60 – 0,11	0,176
VMTC	0,00 **	0,00 – 0,00	**< 0,001**	0,00 **	0,00 – 0,00	**< 0,001**	0,00 **	0,00 – 0,00	**< 0,001**	0,00 **	0,00 – 0,00	**< 0,001**
Min	−0,00	−0,00 – 0,00	0,289	−0,00 **	−0,00 – −0,00	**0,003**	−0,00 ***	−0,00 – −0,00	**0,001**			
Partits	0,04 *	0,02 – 0,06	**0,001**	0,06 **	0,04 – 0,08	**< 0,001**	0,12 **	0,07 – 0,16	**< 0,001**			
Gols	0,02	−0,04 – 0,08	0,520	0,03	−0,02 – 0,08	0,309	0,05	−0,04 – 0,14	0,266			
Gols_pp	−0,01	−0,20 – 0,18	0,931									
Gols_90	1,53 **	1,02 – 2,03	**< 0,001**									
xGols	0,03	−0,10 – 0,16	0,622	0,07	−0,13 – 0,27	0,469	0,03	−0,18 – 0,24	0,771			
Xuts	0,00	−0,01 – 0,02	0,588	0,01	−0,01 – 0,03	0,182	0,02	−0,00 – 0,04	0,065			

Penals	−0,05	−0,21 – 0,12	0,570									
Faltes	−0,00	−0,04 – 0,03	0,816									
Assist	0,04	−0,03 – 0,10	0,244	0,08 **	0,03 – 0,12	**0,001**	0,07 *	0,00 – 0,15	**0,047**			
xAssist	−0,02	−0,24 – 0,20	0,860	−0,15 *	−0,29 – −0,01	**0,037**	−0,07	−0,30 – 0,16	0,568	15,12	−3,55 – 33,80	0,112
Pass	−0,00	−0,01 – 0,00	0,292	−0,00 **	−0,01 – −0,00	**0,006**	−0,00	−0,00 – 0,00	0,166	−0,02	−0,05 – 0,01	0,165
Pass_ex	0,00	−0,00 – 0,01	0,200	0,00 **	0,00 – 0,01	**0,002**	0,00	−0,00 – 0,01	0,077	0,02	−0,01 – 0,05	0,141
Pass_intl	0,02	−0,01 – 0,05	0,214	−0,00	−0,02 – 0,02	0,835	0,00	−0,04 – 0,05	0,826			
Pass_intl_ex	−0,01	−0,06 – 0,03	0,554	−0,01	−0,04 – 0,01	0,333	0,00	−0,05 – 0,06	0,991			
Pass_thro	−0,04 *	−0,07 – −0,00	**0,028**	0,02	−0,00 – 0,05	0,097	−0,02	−0,07 – 0,04	0,583			
Cross	−0,01	−0,02 – 0,01	0,330	−0,00	−0,01 – 0,01	0,950						
Cross_ex	0,01	−0,02 – 0,05	0,410	−0,00	−0,02 – 0,02	0,766						
Dribl	−0,00	−0,00 – 0,00	0,700	0,00	−0,00 – 0,00	0,454						
Dribl_ex	0,01	−0,00 – 0,01	0,069	0,00	−0,00 – 0,01	0,263						
Participacio_Gols	−0,36	−2,92 – 2,20	0,783									
Participacio_Xuts	0,41	−6,83 – 7,65	0,912	−1,97	−10,09 – 6,15	0,633						
Participacio_xGoals	−2,18	−8,02 – 3,65	0,462	−1,75	−9,31 – 5,81	0,650						
Participacio_Minuts	1,56	−0,08 – 3,20	0,062	0,98	−0,16 – 2,11	0,092	0,38	−0,80 – 1,57	0,525			
Duels				0,00	−0,00 – 0,01	0,095	0,00	−0,00 – 0,00	0,599			

Duels_ex		−0,00	−0,01 – 0,01	0,855	0,01	−0,00 – 0,02	0,227		
Ppp		0,00	−0,04 – 0,04	0,847	0,01	−0,03 – 0,06	0,622	0,24 *	0,04 – 0,44	**0,019**
Duels_aeris					0,00	−0,01 – 0,01	0,697	−0,07	−0,21 – 0,08	0,345
Duels_aeris_ex					−0,00	−0,01 – 0,01	0,797	0,09	−0,07 – 0,25	0,263
Duels_aeris_ex_pc								0,00	−0,00 – 0,01	0,181
Interv								0,00	−0,08 – 0,08	0,918
Parades								0,02	−0,07 – 0,10	0,657
Reflx								−0,03	−0,12 – 0,05	0,437
Parades_pc								0,01 **	0,00 – 0,02	**0,004**
Reflx_pc								0,02	−0,07 – 0,12	0,642
Parades_reflx_pc								−0,00	−0,01 – 0,01	0,636
Observations	548		930			889			295	
R^2 / R^2 adjusted	0,733 / 0,718		0,727 / 0,718			0,631 / 0,622			0,698 / 0,679	
AIC	1419,027		2304,690			2425,560			839,440	
F-Statistic	49,09202 ***		88,7455 ***			64,40058 ***			35,50873 ***	

** p<0,05 ** p<0,01 *** p<0,001*

Font: Elaboració pròpia

Taula 4.2.2.2 Taula dels FIV per a la categoria "Davanters"

Variable	***GVIF***	***Df***	***GVIF,,1,,2,Df,,***
Lliga	2,16	4	1,10
VMTC	1,97	1	1,40
Min	15,13	1	3,89
Gols	15,80	1	3,97
Gols_90	2,67	1	1,63
xGols	25,93	1	5,09
Penals	2,10	1	1,45
Faltes	2,19	1	1,48
Assist	3,63	1	1,90
xAssist	22,86	1	4,78
Pass_ex	4,66	1	2,16
Pass_intl	77,40	1	8,80
Pass_intl_ex	22,01	1	4,69
Pass_thro	32,80	1	5,73
Cross	18,25	1	4,27
Cross_ex	14,45	1	3,80
Dribl	16,04	1	4,00
Dribl_ex	15,95	1	3,99
Participacio_Gols	15,23	1	3,90
Participaci_Xuts	22,19	1	4,71
Particiapcio_xGoals	38,42	1	6,20

Font: Elaboració pròpia

Taula 4.2.2.3 Taula dels FIV per a la categoria "Mig-campistes"

Variable	***GVIF***	***Df***	***GVIF,,1,,2,Df,,***
Lliga	2,41	4	1,12
VMTC	1,78	1	1,33
Min	15,42	1	3,93
Gols	4,02	1	2,00
xGols	12,92	1	3,59
Assist	3,51	1	1,87
xAssist	14,51	1	3,81
Pass_ex	3,92	1	1,98
PAss_intl	55,72	1	7,46
Pass_intl_ex	16,84	1	4,10
Pass_thro	24,17	1	4,92
Cross	17,07	1	4,13
Cross_ex	14,14	1	3,76
Dribl	20,05	1	4,48
Dribl_ex	15,62	1	3,95
Duels	22,11	1	4,70
Duels_ex	14,62	1	3,82
Ppp	2,16	1	1,47
Participacip_Xuts	9,31	1	3,05
Participacio_xGoals	13,81	1	3,72

Font: Elaboració pròpia

Taula 4.2.2.4 Taula dels FIV per a la categoria "Defenses"

Variable	***GVIF***	***Df***	***GVIF,,1,,2,Df,,***
Lliga	2,62	4	1,13
VMTC	1,44	1	1,20
Min	13,65	1	3,70
Gols	2,01	1	1,42
5xGols	3,27	1	1,81
Assit	2,24	1	1,50
xAssist	6,73	1	2,59
Pass_ex	4,12	1	2,03
Pass_intl	21,72	1	4,66
Pass_intl_ex	8,40	1	2,90
Pass_thro	7,94	1	2,82
Duels	16,64	1	4,08
Duels_ex	12,88	1	3,59
Duels_aeris	51,04	1	7,14
Duels_aeris_ex	39,54	1	6,29
Ppp	1,66	1	1,29

Font: Elaboració pròpia

Taula 4.2.2.5 Taula dels FIV per a la categoria "Porters"

Variable	***GVIF***	***Df***	***GVIF,,1,,2,Df,,***
Lliga	1,94	4	1,09
VMTC	1,32	1	1,15
xAssist	7,87	1	2,81
Pass_ex	3,64	1	1,91
Duels_Aeris	73,91	1	8,60
Duels_aerus_ex	70,25	1	8,38
Duels_aeris_ex_pc	3,37	1	1,83
Ppp	1,84	1	1,36
Parades	50,00	1	7,07
Reflx	86,83	1	9,32
Parades_pc	9,31	1	3,05
Reflx_pc	102,54	1	10,13
Parades_reflx_pc	8,96	1	2,99

Font: Elaboració pròpia

Donats els valors els quals es poden contemplar a les taules 4.2.2.2, 4.2.2.3, 4.2.2.4 i 4.2.2.5 es contempla l'eliminació d'aquelles variables amb valors molt elevats (com a mínim per sobre de 10) i que, per tant, causen multicol·linealitat severa als models. Un cop suprimides, es construeixen de nou els models, els quals es poden analitzar a la taula 4.2.2.4. És correcte comentar:

- Es pot observar que no s'ha estimat el coeficient associat a la categoria $Bundesliga$ del factor $Lliga$ ja que R ha pres aquesta categoria com a categoria de referència o de base.

  A la tercera columna pel que fa a cada model es mostren els $p-valors$ associats als tests de significació individual per a cadascun dels coeficients. Les hipòtesis nul·la i alternativa que es plantegen a cadascun d'aquests tests, juntament amb la definició de l'estadístic de contrast associat, són:

$H_0 : \hat{\beta} = 0 \; vs \; H_1 : \hat{\beta} \neq 0$	$t = \frac{\hat{\beta}}{S.d(\hat{\beta})} \sim t_{n-k} \, , sota \; H_0$

Havent realitzat aquest contrast d'hipòtesis per a cadascun dels coeficients estimats dels models, es pot observar que pel coeficient de la variable $Cross\_ex$ del model que fa referència als jugadors *Mig-campistes* no es pot rebutjar la hipòtesis nul·la que el coeficient sigui diferent de 0 amb un nivell de confiança del 95%, ja que s'ha obtingut un p-valor de $0{,}205$. Per altra banda, també es pot observar que per a certes categories de la variable categòrica $Lliga$ – es dona als models pel que fa als *Davanters* i *Porters* – els coeficients resulten en no ser estadísticament significatius individualment. No obstant, pels casos de factors amb diverses categories el que caldria tenir en compte és la significació global o general del factor per comprovar si realment el seu efecte és significatiu o no. Per tal de comprovar-ho, es realitza el test *anova* (o test de raó de versemblança) que es basa en comparar models encaixats, on un dels models est inclòs (model *reduït*) dins d'un altre (model *ampliat*). Les hipòtesis nul·la i alternativa, juntament amb l'estadístic de proba, són:

$H_0: \widehat{\beta_{Lliga}} = 0 \; vs \; H_1: \widehat{\beta_{Lliga}} \neq 0$	$F = \dfrac{\dfrac{SS_R(ampliat) - SS_R(reduït)}{p_1 - p_0}}{MS_{Res}} \sim F_{p_1-p_0, n-p_1}, sota\ H_0$

On $SS_R(ampliat)$ és la suma de quadrats dels errors de la regressió del model ampliat, és a dir, amb el factor $Lliga$ inclòs; $SS_R(reduït)$ és la suma de quadrats dels errors de la regressió del model reduït, p_1 és el número de coeficients que hi ha en el model ampliat, p_0 és el número de coeficients que hi ha en el model reduït i MS_{Res} és l'estimació de σ^2 en el model ampliat.

Un cop realitzats els contrastos d'hipòtesis, pel model dels *Davanters* i pel model dels *Porters*, amb els valors dels estadístics $F = 12{,}61$ i $F = 8{,}78$ p-valors associats dels quals són $8{,}18^{-10}$ i $1{,}06^{-6}$, respectivament; amb un nivell de confiança del 95% es conclou que hi ha evidències estadístiques per a rebutjar les hipòtesis nul·les i, per tant, la inclusió del factor $Lliga$ millora el model reduït.

- També es realitza el test de significació conjunta de cada model, hipòtesis nul·la i alternativa dels quals, així com l'estadístic de contrast són els següents:

$H_0: \hat{\beta}_1 = \cdots = \hat{\beta}_k \; vs \; H_1: Alguna\ \hat{\beta}_i \neq 0, on\ i = 1, \dots, m$[7]	$F = \dfrac{\dfrac{SS_R(restringit) - SS_R(estimat)}{m}}{\dfrac{SS_R(estimat)}{n-k}} \sim F_{m,n-k}, sota\ H_0$

On $SS_R(estimat)$ és la suma dels quadrats dels errors del model estimat; $SS_R(restringit)$ és la suma dels quadrats dels errors del model restringit ($y = \beta_0 + e$); m és el nombre de coeficients del model (excloent el terme constant β_0); n és el nombre d'observacions que s'han utilitzat per a l'estimació del model i k és el nombre de variables predictores d'estimació.

[7] S'ha de tenir en compte que per a cadascun dels 4 models el valor de k és diferente, essent $m_{Davamters} = 11$, $m_{Mig-campistes} = 14$, $m_{Defenses} = 11$ i $m_{Porters} = 10$.

Un cop executat, s'obtenen els valors corresponents a l'última fila de la taula i es conclou que hi ha evidències significatives estadísticament per a rebutjar la hipòtesis nul·la amb un nivell de confiança del 99% per a cadascun dels models. D'aquesta manera, es pot dir que hi ha algun paràmetre β diferent de 0 estadísticament.

- Tenint en compte els coeficients de determinació o bondat de l'ajust R^2 dels models estimats, s'han obtingut per valor de $0{,}722$, $0{,}739$, $0{,}652$ i $0{,}683$ pels models de *Davanters*, *Mig-campistes*, *Defenses* i *Porters* respectivament. Aquest estadístic informa de quin percentatge de la variabilitat de la variable resposta està explicat amb el model. Quant més a prop d'1, més a prop estarà el model d'explicar la totalitat de la variabilitat. Amb els valors obtinguts, es pot dir que els models capturen una notable quantitat de variabilitat.

Una vegada s'han obtingut i comentat els resultats dels models anteriors, seria adient comprovar que es donessin els supòsits que comporta l'estimació per *MQO* dels models de regressió lineal múltiple. No obstant, abans de verificar-ho es decideix analitzar si la inclusió extra de variables quadràtiques afavoriria al model. Durant la lectura de la literatura existent sobre el tema en qüestió s'ha vist que prenen importància i que són significatives estadísticament, per exemple el cas de la variable *edat*; i és per aquest motiu que es determina fer aquest pas.

Per a portar a terme l'anterior, es realitzarà el test de curvatura de Tukey, basat en l'estadístc $t - statistic$, implementat a la funció $residualPlots$ d'R, la qual a més a més dibuixa els residus envers cada variable explicativa del model. El test consisteix en afegir un terme quadràtic i contrastar estadísticament que sigui igual o diferent a 0.

Taula 4.2.2.7 Taula resum del test de curvatura de Tukey per a la categoria "Davanters"

Variable	***Test,stat***	***P-value***
Edat	-8,19202	0,00000
VMTC	-4,88306	0,00000
Min	-5,40751	0,00000
Gols_90	-4,90895	0,00000
Penals	-0,29587	0,76744
Pass_ex	-2,67771	0,00764
Dribl_ex	-3,43917	0,00063

Font: Elaboració pròpia

Taula 4.2.2.6 Taula resum de la segona estimació per mínims quadrats ordinaris separant per categories del factor "Rol"

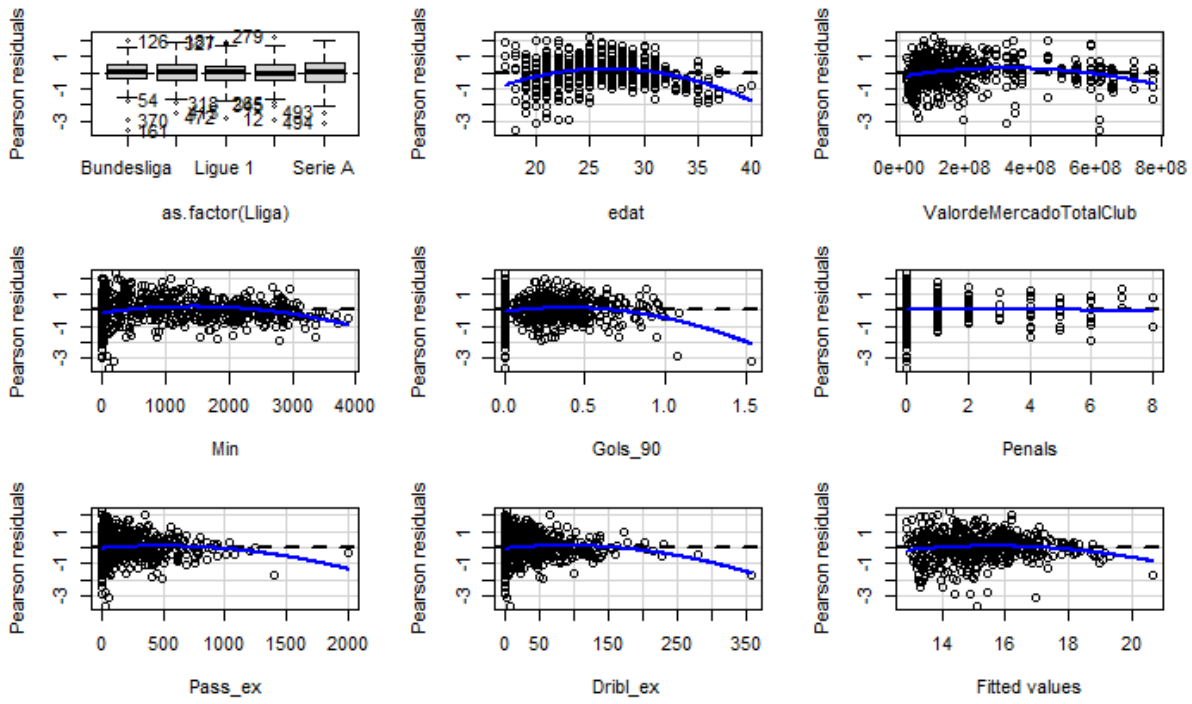
	Davanters			Mig-Campistes			Defenses			Porters		
Predictors	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***
(Intercept)	14,15954 ***	13,66502 – 14,65406	**< 0,001**	14,09073 ***	13,75311 – 14,42835	**< 0,001**	13,99574 ***	13,62470 – 14,36677	**< 0,001**	13,01705 ***	12,37701 – 13,65710	**< 0,001**
Lliga [LaLiga]	−0,23978 *	−0,47808 – − 0,00149	**0,049**	−0,38609 ***	−0,54914 – − 0,22303	**< 0,001**	−0,26923 **	−0,46694 – − 0,07152	**0,008**	0,19051	−0,20572 – 0,58673	0,345
Lliga [Ligue 1]	−0,21632	−0,44980 – 0,01716	0,069	−0,40594 ***	−0,56967 – − 0,24221	**< 0,001**	−0,27898 **	−0,47205 – − 0,08591	**0,005**	0,12358	−0,24246 – 0,48961	0,507
Lliga [Premier League]	0,60073 ***	0,31652 – 0,88494	**< 0,001**	0,54531 ***	0,34231 – 0,74830	**< 0,001**	0,76004 ***	0,51962 – 1,00046	**< 0,001**	1,02475 ***	0,59915 – 1,45034	**< 0,001**
Lliga [Serie A]	−0,34490 **	−0,56780 – − 0,12199	**0,002**	−0,33187 ***	−0,49726 – − 0,16649	**< 0,001**	−0,43448 ***	−0,61939 – − 0,24958	**< 0,001**	−0,13679	−0,48308 – 0,20950	0,438
edat	−0,03652 ***	−0,05524 – − 0,01781	**< 0,001**	−0,04725 ***	−0,05984 – − 0,03465	**< 0,001**	−0,05719 ***	−0,07150 – − 0,04287	**< 0,001**	−0,04164 ***	−0,06348 – − 0,01980	**< 0,001**
VMTC	0,00000 ***	0,00000 – 0,00000	**< 0,001**	0,00000 ***	0,00000 – 0,00000	**< 0,001**	0,00000 ***	0,00000 – 0,00000	**< 0,001**	0,00000 ***	0,00000 – 0,00000	**< 0,001**
Min	0,00063 ***	0,00049 – 0,00077	**< 0,001**	0,00039 ***	0,00028 – 0,00050	**< 0,001**						
Gols_90	1,89746 ***	1,53713 – 2,25779	**< 0,001**									
Penals	−0,08270 **	−0,14535 – − 0,02005	**0,010**									
Pass_ex	0,00079 **	0,00027 – 0,00132	**0,003**	0,00074 ***	0,00054 – 0,00094	**< 0,001**	0,00093 ***	0,00071 – 0,00116	**< 0,001**	0,00190 *	0,00043 – 0,00337	**0,012**
Dribl_ex	0,00367 **	0,00101 – 0,00633	**0,007**	0,00622 ***	0,00381 – 0,00862	**< 0,001**						
Gols				0,07296 ***	0,03964 – 0,10627	**< 0,001**	0,12533 ***	0,05584 – 0,19482	**< 0,001**			
Assist				0,04913 **	0,01346 – 0,08481	**0,007**	0,12583 ***	0,07616 – 0,17550	**< 0,001**			

Cross_ex		−0,00537	−0,01368 – 0,00293	0,205					
Pass_thro_pc		0,00863 ***	0,00652 – 0,01074	**< 0,001**					
Duels_aeris_ex_pc		0,00564 ***	0,00274 – 0,00853	**< 0,001**	0,01557 ***	0,01250 – 0,01863	**< 0,001**	0,00519 *	0,00105 – 0,00933	**0,014**
Duels_ex					0,01767 ***	0,01340 – 0,02193	**< 0,001**			
xAssist								29,44813 ***	16,85362 – 42,04263	**< 0,001**
Parades_pc								0,01508 ***	0,01038 – 0,01978	**< 0,001**
Observations	548	930			889			295		
R^2 / R^2 adjusted	0,722 / 0,716	0,739 / 0,735			0,652 / 0,648			0,683 / 0,672		
AIC	1403,231	2233,218			2348,716			835,899		
F-Statistic	126,3 **	185,2 **			149,5 **			61,28 **		

$*\ p < 0{,}05$ $**\ p < 0{,}01$ $***\ p < 0{,}001$

Font: Elaboració pròpia

Figura 4.2.2.1 Diagrames de punts dels residus de Pearson i els valors de la mostra per a les diferents variables del model referent a la categoria "Davanters"



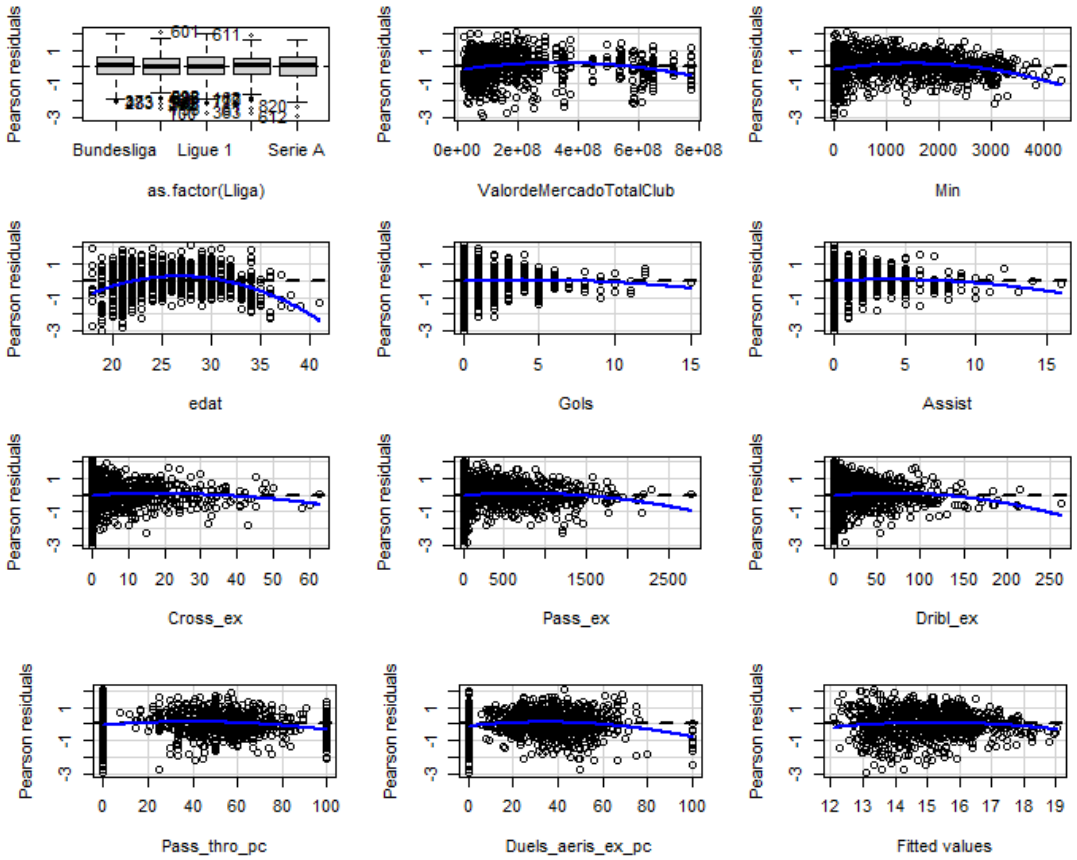
Font: Elaboració pròpia

Taula 4.2.2.8 Taula resum del test de curvatura de Tukey per a la categoria "Mig-Campistes"

Variable	Test.stat	P-value
VMTC	−5,59149	0,00000
Min	−7,27879	0,00000
Edat	−12,70576	0,00000
Gols	−1,32882	0,18424
Assist	−2,05042	0,04061
Pass_ex	−3,34318	0,00086
Dribl_ex	−3,81429	0,00015
Pass_thro_pc	−5,98915	0,00000
Duels_aeris_ex_pc	−6,11958	0,00000

Font: Elaboració pròpia

Figura 4.2.2.2 Diagrames de punts dels residus de Pearson i els valors de la msotra per a les diferents variables del model referent a la categoria "Mig-Campistes"



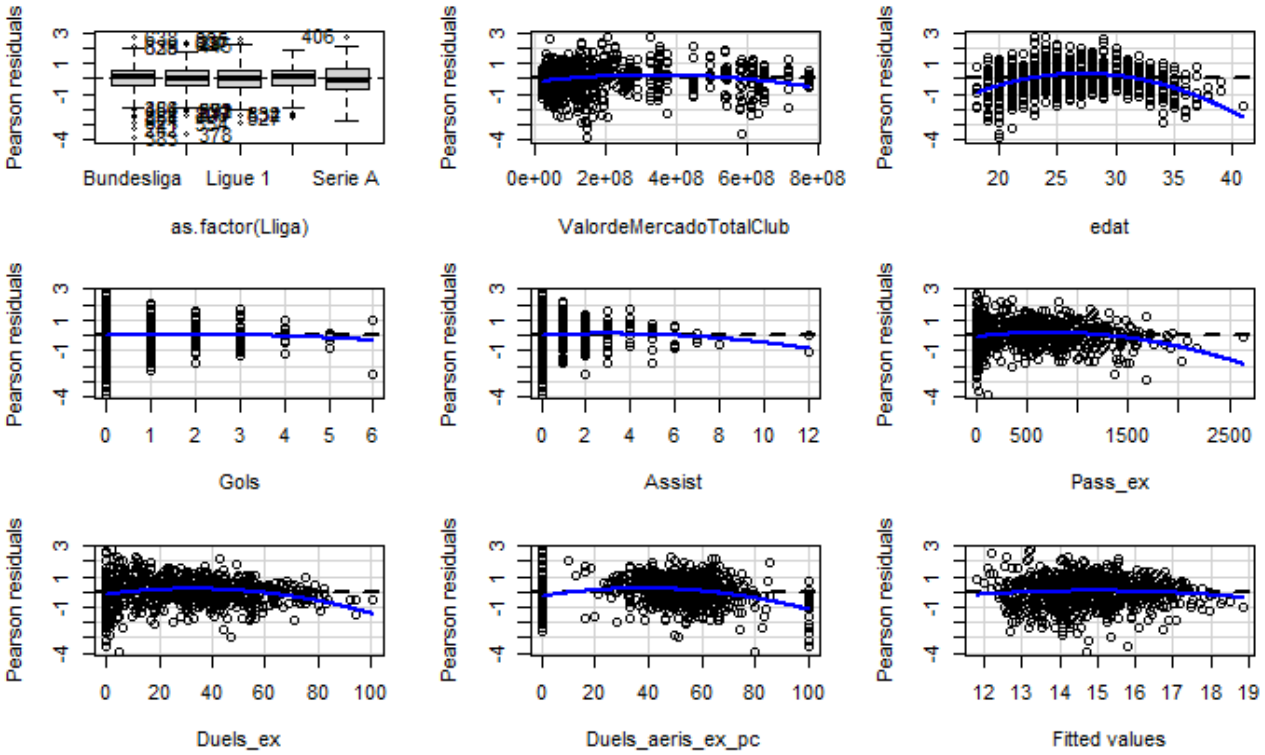
Font: Elaboració pròpia

Taula 4.2.2.9 Taula resum del test de curvatura de Tukey per a la categoria "Defenses "

Variable	***Test.stat***	***P-value***
VMTC	−4,77893	0,00000
Edat	−13,80987	0,00000
Gols	−1,09553	0,27359
Assist	−1,84642	0,06517
Pass_ex	−6,16290	0,00000
Duels_ex	−6,92006	0,00000
Duels_aeris_ex_pc	−9,57437	0,00000

Font: Elaboració pròpia

Figura 4.2.2.3 Diagrames de punts dels residus de Pearson i els valors de la mostra per a les diferents variables del model referent a la categoria "Defenses "



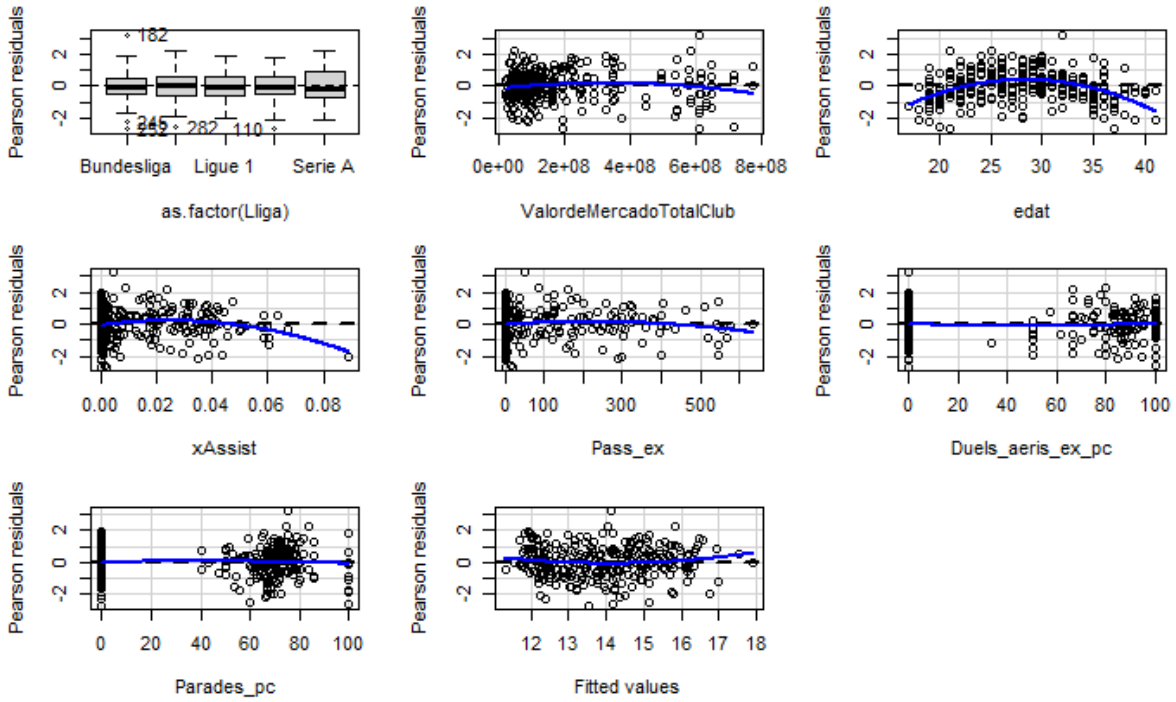
Font: Elaboració pròpia

Taula 4.2.2.10 Taula resum del test de curvatura de Tukey per a la categoria "Porters"

Variable	***Test.stat***	***P-value***
VMTC	-2,02092	0,04423
Edat	-9,20555	0,00000
xAssist	-3,60972	0,00036
Pass_ex	-1,84545	0,06602
Duels_aeris_ex_pc	0,87215	0,38387
Parades_pc	-0,54101	0,58892

Font: Elaboració pròpia

Figura 4.2.2.4 Diagrames de punts dels residus de Pearson i els valors de la mostra per a les diferents variables del model referent a la categoria "Porters "



Font: Elaboració pròpia

Observant els resultats de les taules 4.2.2.7, 4.2.2.8, 4.2.2.9 i 4.2.2.10, així com els gràfics associats (figures 4.2.2.1, 4.2.2.2, 4.2.2.3 i 4.2.2.4) , amb una confiança del 95%, hi ha evidències estadístiques significatives per a valorar la inclusió dels termes al quadrat de les variables $VMTC$, $edat$, Min, $Gols\_90$, $Dribl\_ex$, $Assist$[8],, $Pass\_ex$[9], $Pass\_thro\_pc$, $Duels\_aeris\_ex\_pc$[10], $Duels\_ex$ i $xAssist$. Cal comentar que la variable $Cross\_ex$ no es tindrà en compte ja que el seu coeficient no era estadísticament signficatiu al model referent als *Mig-campistes* que es mostra a la taula 4.2.2.6.

A continuació, es construeixen els models de regressió múltiple pels diferents rols tot havent inclòs els termes quadràtics anteriorment mencionats[11]. Tanmateix, abans de finalitzar i comentar els resultats del que serien els models finals, és convenient revisar i comprovar que es mantinguin els supòsits que fan que els models siguin lineals i que romanguin les propietats dels estimadors per MQO. D'aquesta manera, val la pena recordar que els supòsits citats són:

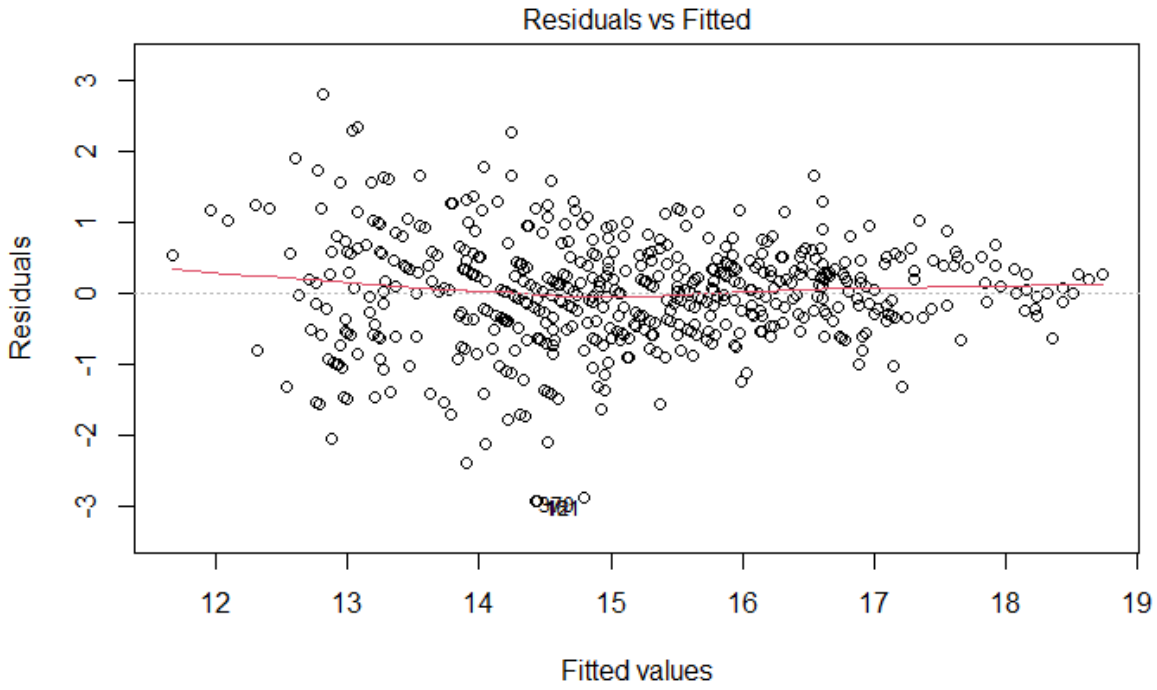
- Mitjana nul·la o residus al voltant de 0, és a dir, $E[e] = 0$.
- La variància del residu (i per tant de y) és independent dels valors de les variables explicatives i, a més a més, és constant, és a dir, $Var(e) = \sigma^2$.

[8] Pel que fa al model de *Mig-campistes*
[9] Pel que fa als models de *Mig-campistes* i *Defenses.*
[10] Pel que fa als models de *Mig-campistes* i *Defenses.*
[11] La taula que msotra els resultats dels models (coeficients, p-valors, estadístics, etc.) es troba a l'annex (**ANNEX**).

- Els termes de residu de dues observacions diferents estan incorrelacionats, és a dir, $Cov(e_i, e_j) = Cov(e_j, e_i) = 0, \forall i \neq j$.

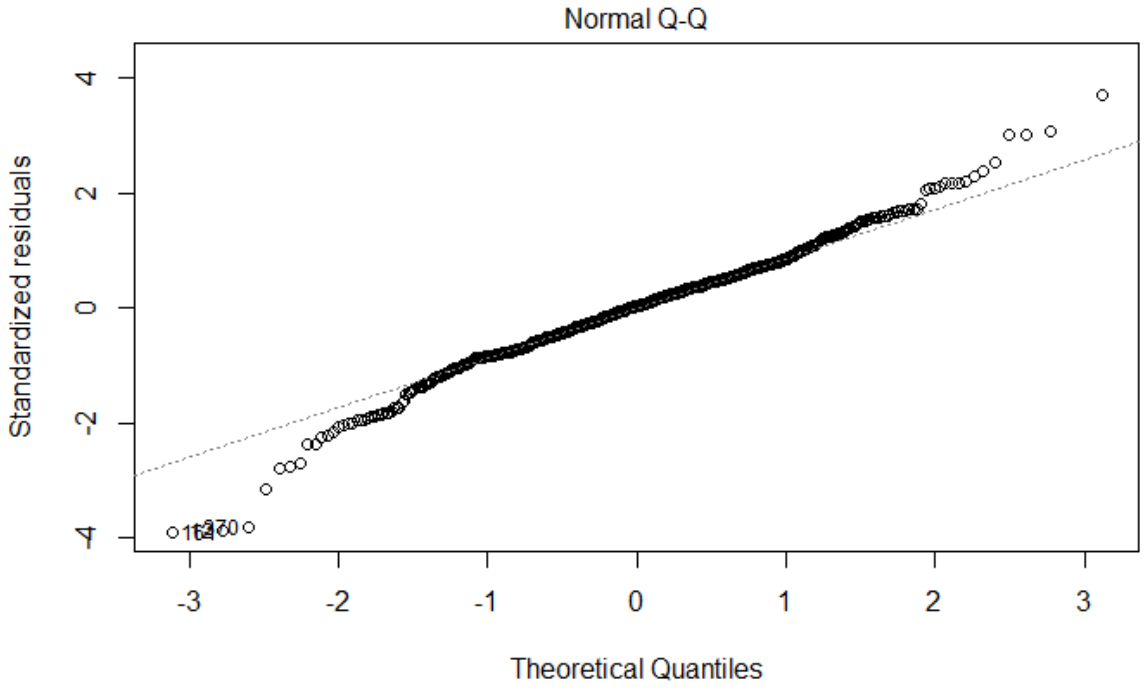
Seguidament, es mostra els gràfics pertinents a l'hora d'analitzar els residus obtinguts i verificar els supòsits anteriors. Per a cadascuna de les 4 regressions múltiples construïdes es presenten, per una banda, el gràfic dels residus envers les estimacions (*Residuals vs Fitted*), el qual a l'eix d'abscisses es representa les estimacions de la variable endògena i a l'eix d'ordenades l'error associat (és a dir, $\hat{y} - y$); i, per l'altra, el gràfic de quantils teòrics (*Normal Q-Q*), el qual consisteix en comparar els quantils de la distribució dels residus estandarditzats calculats (eix d'ordenades) amb els quantils teòrics d'una distribució normal amb la mateixa mitjana i desviació estàndard que les dades (eix d'abscisses).

Figura 4.2.2.5 Diagrama "Residuals vs Fitted" per a la categoria "Davanters"



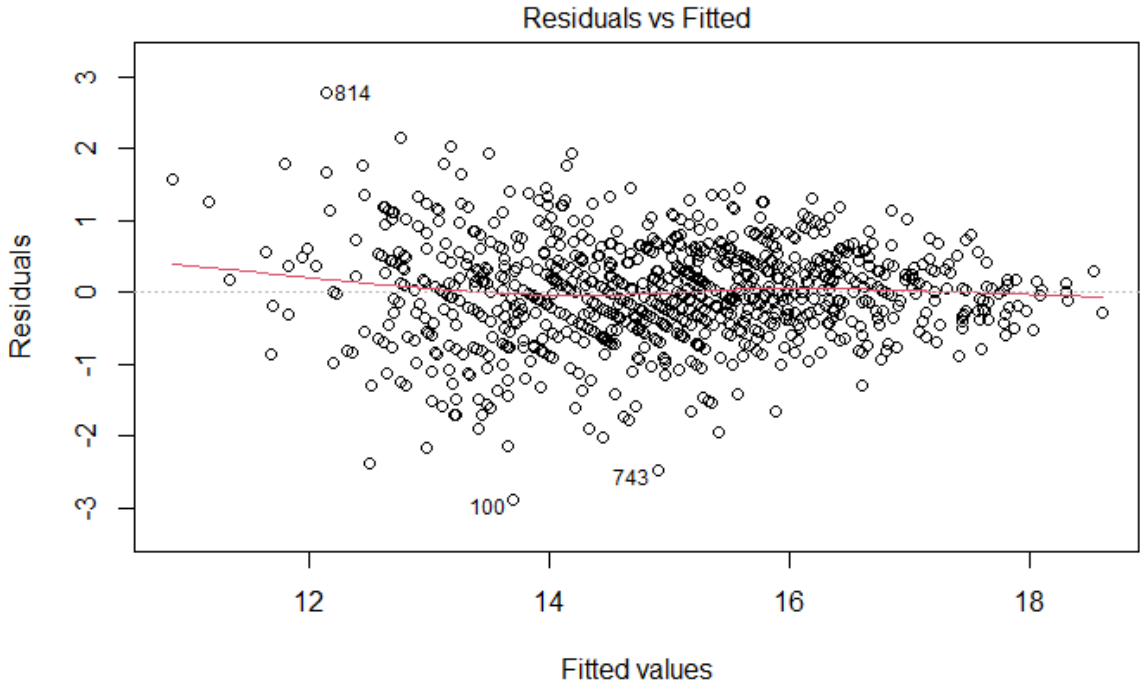
Font: Elaboració pròpia

Figura 4.2.2.6 Gràfic "Normal Q-Q" per a la categoria "Davanters"



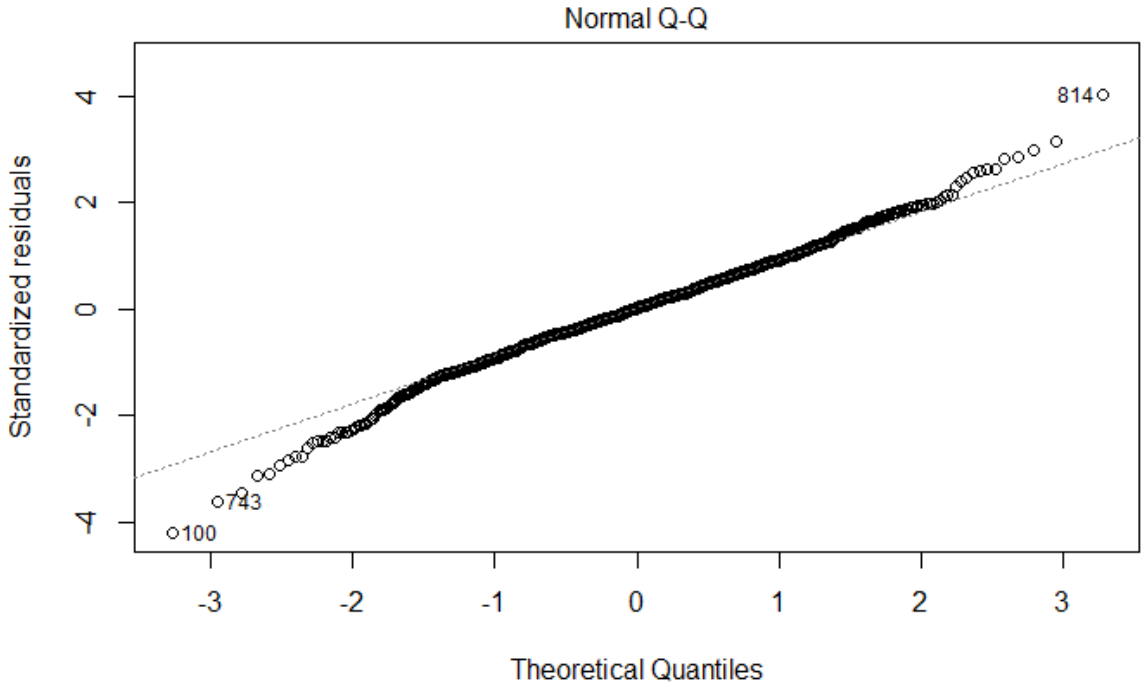
Font: Elaboració pròpia

Figura 4.2.2.7 Gràfic "Residuals vs Fitted" per a la categoria "Mig-campistes"



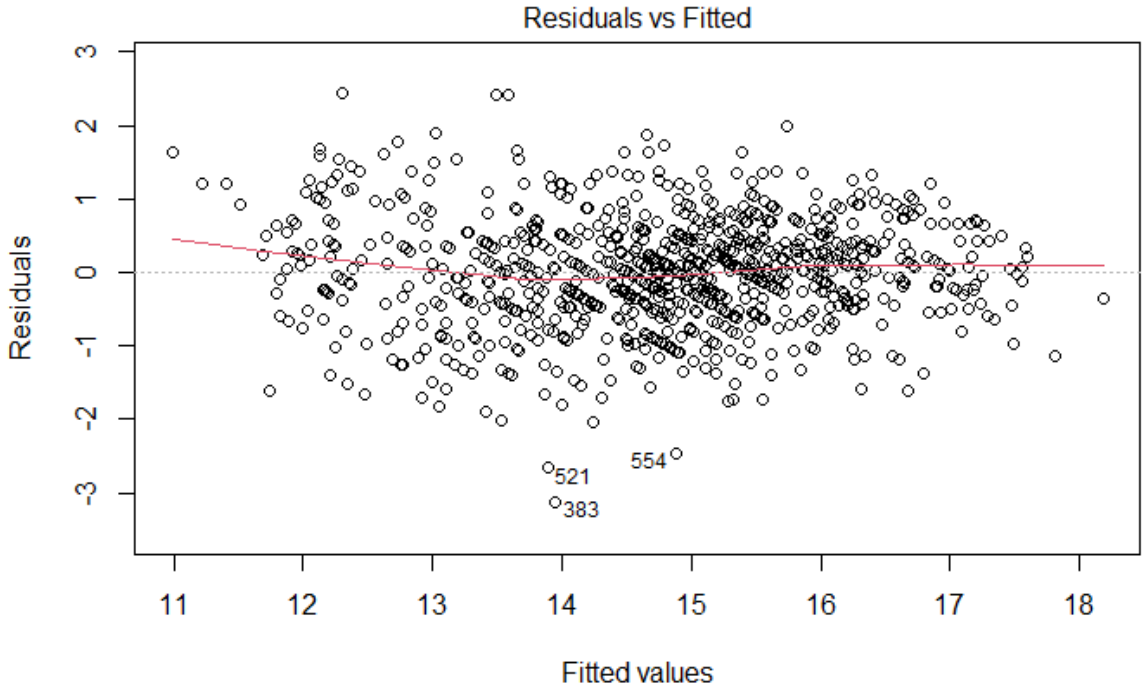
Font: Elaboració pròpia

Figura 4.2.2.8 Gràfic "Normal Q-Q" per a la categoria "Mig-campistes"



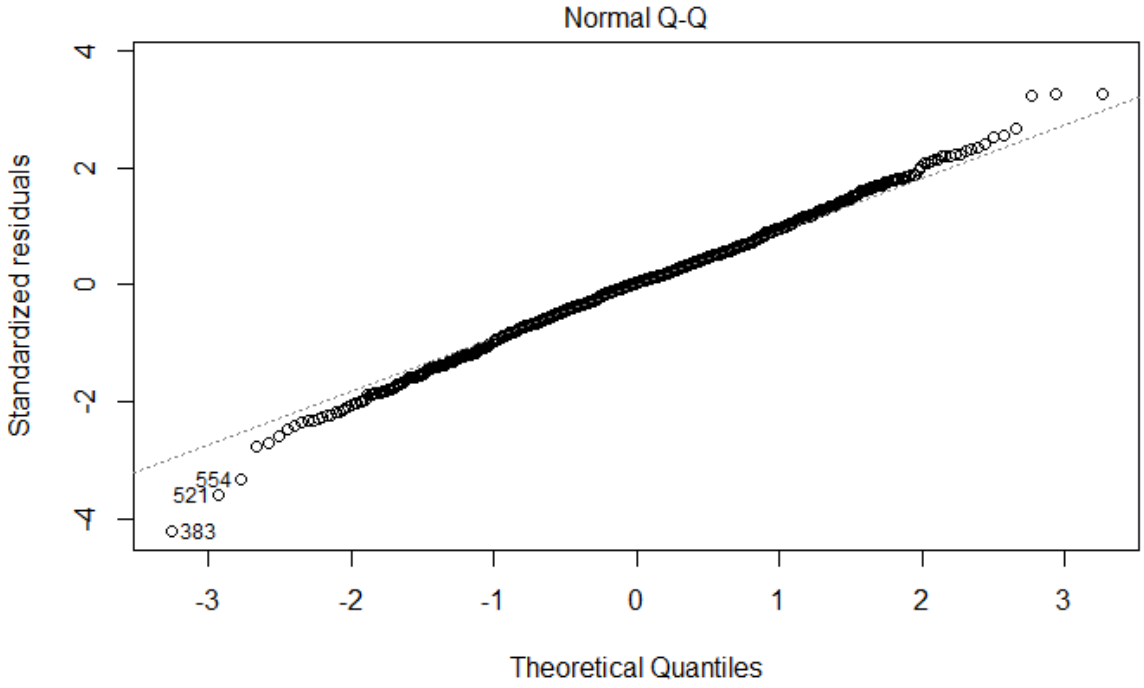
Font: Elaboració pròpia

Figura 4.2.2.9 Gràfic "Residuals vs Fitted" per a la categoria "Defenses"



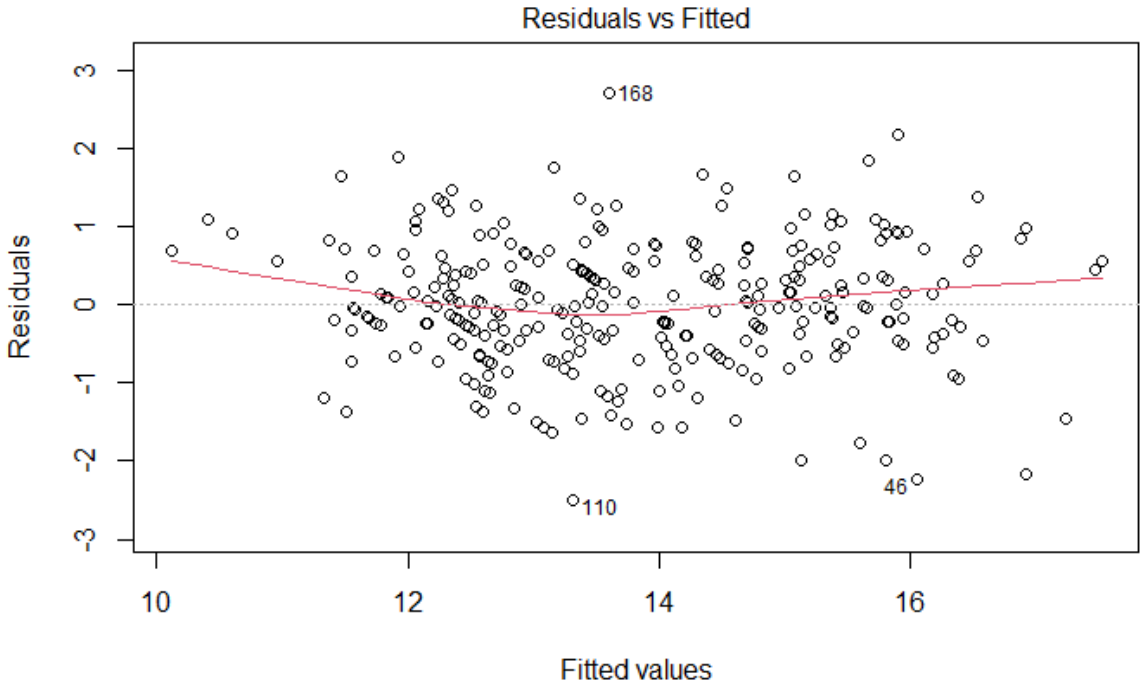
Font: Elaboració pròpia

Figura 4.2.2.10 Gràfic "Normal Q-Q" per a la categoria "Defenses"



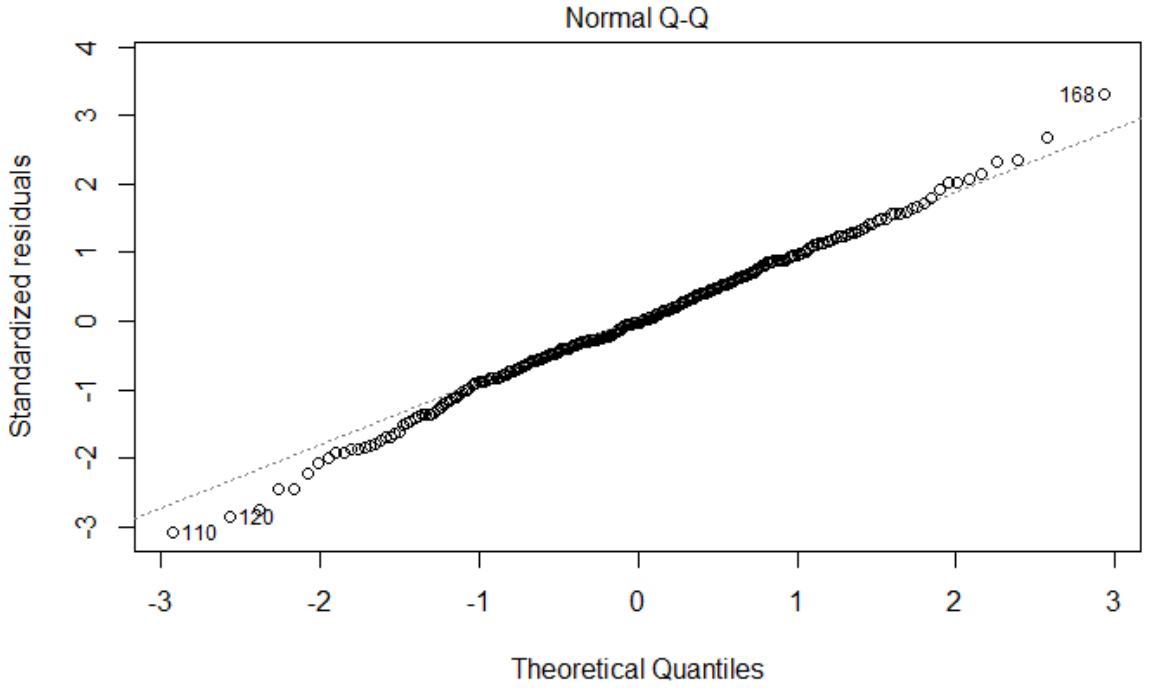
Font: Elaboració pròpia

Figura 4.2.2.11 Gràfic "Residuals vs Fitted" per a la categoria "Porters"



Font: Elaboració pròpia

Figura 4.2.2.12 Gràfic "Normal Q-Q" per a la categoria "Porters"



Font: Elaboració pròpia

En primer lloc, si s'observen els gràfics *Normal Q-Q* (figures 4.2.2.6, 4.2.2.8, 4.2.2.10 i 4.2.2.12) es pot advertir que per a la totalitat dels 4 models, però sobretot pels que fan referència als *Davanters* i *Mig-campistes*, hi ha problemes pel que fa a l'estimació dels valors més petits i més grans de la variable endògena *Valor de Mercat*, ja que aquests es troben molt allunyats del que marca la línia que haurien de seguir si, efectivament, seguissin la distribució normal. Per tal corroborar-ho, es decideix realitzar el test de *Jarque-Bera* el qual també contrasta la normalitat dels errors, hipòtesis nul·la i alternativa així com l'estadístic de prova del qual són:

$H_0: X \sim Normal \; vs \; H_1: X \; no \; Normal$	$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \sim \chi^2_2$

On S és la desviació estàndard de les dades de les quals es vol comprovar la normalitat, K és el coeficient de curtosi corresponent i n el número d'observacions. Un cop definit el contrast d'hipòtesis, els resultats obtinguts que es mostren a la taula 4.2.2.11 són:

Taula 4.2.2.11 Taula resum dels resultats del test de Jarque-Bera

Model	***JB Statistic***	***P-value***
Davanters	51,72	$< 2{,}2^{-16}$
Mig-campistes	42,88	$< 2{,}2^{-16}$
Defenses	12,01	0,0075
Porters	1,68	0,3875

Font: Elaboració pròpia

Amb un nivell de confiança del 95%, amb els p-valors obtinguts hi ha evidències estadísticament significatives per rebutjar la hipòtesis nul·la de normalitat pels casos dels *Davanters*, *Mig-Campistes* i *Defenses*. En aquests casos on s'incompleix la hipòtesis de normalitat, cal comentar que els estimadors per MQO deixen de ser eficients (és a dir, ja no són estimadors de variància mínima), la qual cosa suposaria la invalidesa de la inferència posterior que es pogués fer. A part, els intervals de confiança així com els contrastos de significació deixen de ser vàlids.

Per tal d'intentar corregir l'anterior, el que es decideix és avaluar la discrepància i la influència d'aquelles observacions que puguin afectar als resultats de les estimacions dels models mitjançant el càlcul dels *hat values,* els residus *esutdentitzats*, i la distància de Cook:

- *Hat values*: $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$. El càlcul dels *hat values* fa referència a l'avaluació del *leverage* o "apalancament" d'una observació en concret, la qual mesura com de lluny es troba el predictor de la resta de predictors. El màxim i mínim valor que pot prendre és 1 i -1, respectivament

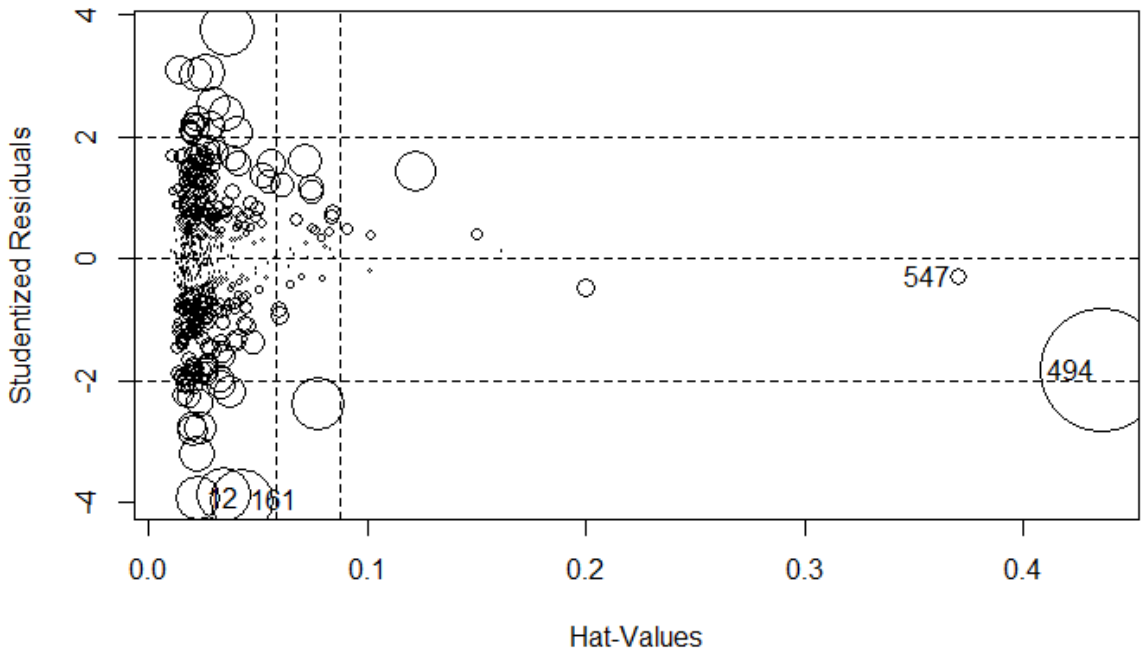
- Residus *estudentitzats*: $e_i^* = \frac{e_i}{SE_{(-i)}\sqrt{1-h_i}}$. Els residus es calculen ajustant un model sense el cas per al qual es calcula el residu, i després escalant el residu (E_i) mitjançant una estimació de la desviació típica ($SE_{(-i)}$) i el *hat value* (h_i). Mitjançant aquest indicador es detecten valors atípics o *outliers*.

- Distància de Cook: $D_i = \frac{e_i^2}{p^2 EQM^2}\left[\frac{h_i}{(1-h_i)^2}\right]$, on e_i és el residu de l'observació i, p és el numero de paràmetres del model, EQM fa referència a l'error quadràtic mitjà i h_{ii} és el *hat value*.

Aquests tres indicadors serviran per analitzar i avaluar tant la discrepància com la influència de cada observació. Cal ressaltar que discrepància i influència són dos conceptes diferents i que cal tenir-los en compte alhora: una observació pot discrepar en gran mesura però sense influència (valor d'y molt allunyat, però els seus predictors són molt propers a $\bar{X}$). De la mateixa manera, una observació pot presentar apalancament degut als valors dels seus predictors (X), però en canvi no tenir cap influència ja que el seu valor en la variable endògena (y) és molt proper a la seva predicció.

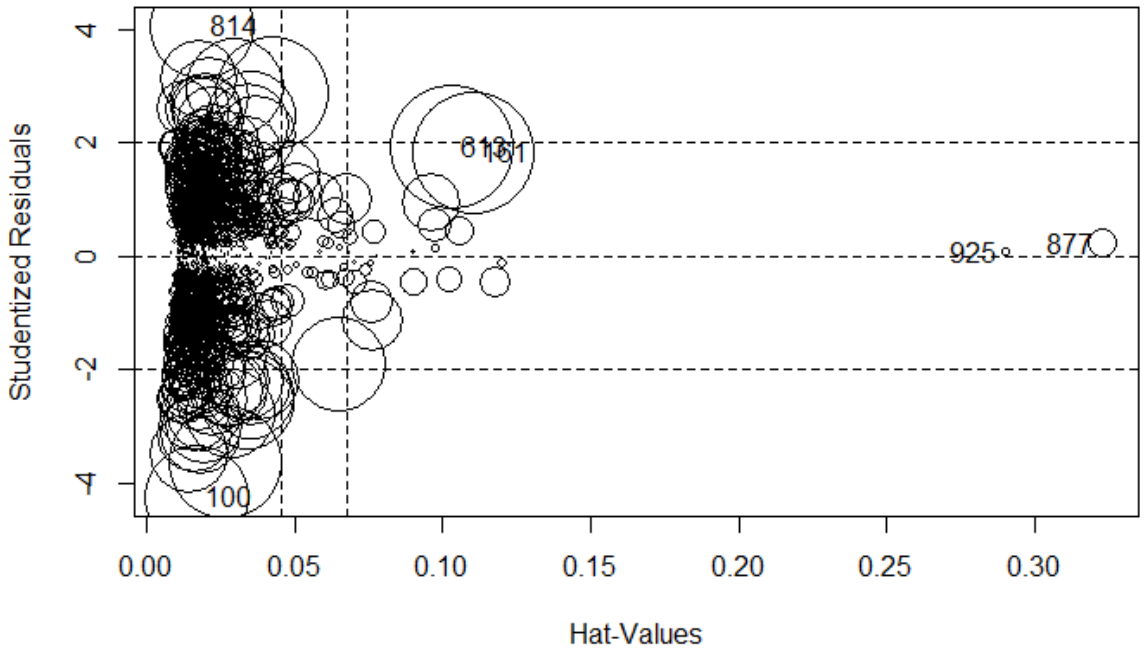
Una vegada s'ha aclarit l'anterior, a les figures 4.2.2.13, 4.2.2.14 i 4.2.2.15 es mostra una part de la sortida de la funció d'R $influencePlot$ la qual calcula i avalua la influència de les observacions a partir dels indicadors esmentats.

Figura 4.2.2.13 Gràfic "Hat-Values vs Residus estudentitzats" per a la categoria "Davanters"



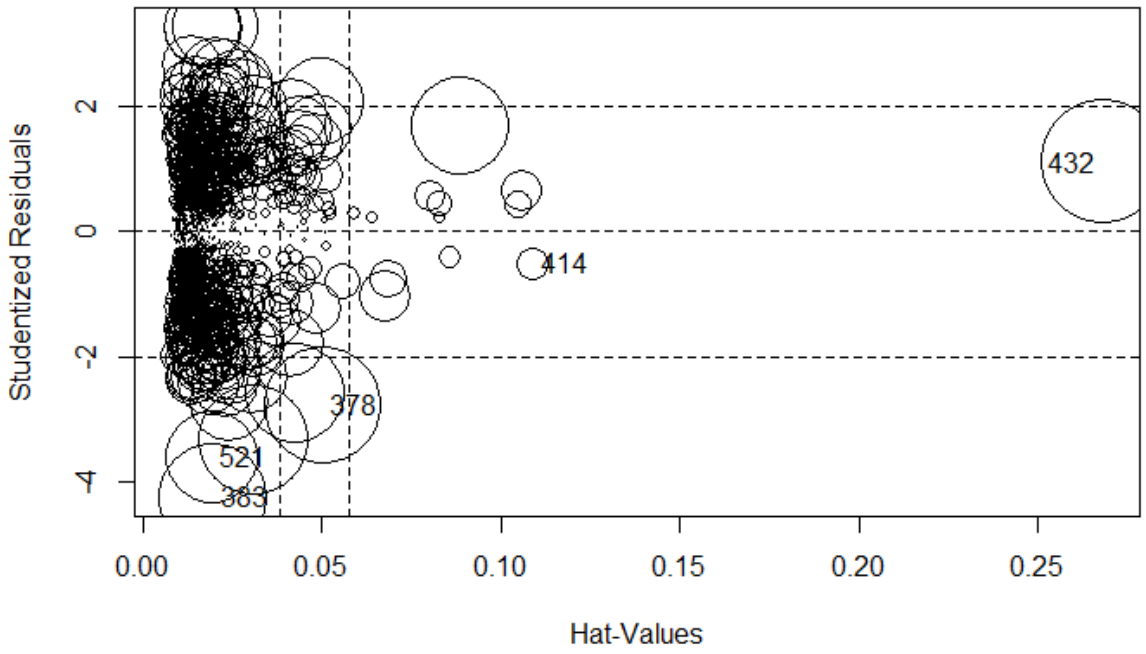
Font: Elaboració pròpia

Figura 4.2.2.14 Gràfic "Hat-Values vs Residus estudentitzats" per a la categoria "Mig-Campistes"



Font: Elaboració pròpia

Figura 4.2.2.15 Gràfic "Hat-Values vs Residus estudentitzats" per a la categoria "Defenses"



Font: Elaboració pròpia

Com ja s'ha comentat, els gràfics que s'observen just a dalt avaluen els residus *estudentitzats* (eix vertical), els *hat values* (eix horitzontal) i la distància de Cook (mida de les circumferències). Amb aquests tres índex de mesura de la influència es poden determinar quins són els valors influents per a cadascun dels models, és a dir, que afecten a l'estimació dels coeficients. Aquests són els que s'indiquen a les figures amb el número d'observació.

A partir d'aquí, es recorda que l'anterior s'ha fet per intentar corregir la no normalitat dels residus dels models. Com a conseqüència, es contrasta de nou la normalitat dels residus a través del test de *Jarque-Bera* com s'ha fet prèviament, resultats dels quals es troben a la següent taula 4.2.2.12.

Taula 4.2.2.12 Taula resum del test de Jarque-Bera

Model	***JB Statistic***	***P-value***
Davanters	31,18	$< 2{,}2^{-16}$
Mig-campistes	18,39	0,001
Defenses	1,74	0,4085

Font: Elaboració pròpia

Clarament, es pot veure que els valors pel que fa l'estadístic i, conseqüentment, els p-valors associats han canviat dràsticament. Amb una confiança del 95%, encara hi ha evidències estadísticament significatives per rebutjar la hipòtesis de normalitat pels casos dels *Davanters* i els *Mig-campistes*, però no en el cas dels *Defenses* ja que el p-valor associat supera en gran mesura el llindar que marca el $0{,}05$.

Si es torna a parar atenció als gràfics de diagnòstic dels residus 4.2.2.5, 4.2.2.7, 4.2.2.9 i 4.2.2.11, concretament als gràfics on es mostren els residus envers els valors estimats, es pot apreciar, de nou, pels casos dels models del *Davanters*, *Mig-campistes* i *Defenses* que a mesura que augmenten els valors estimats per la variable endògena (eix d'abscisses), els residus disminueixen (eix d'ordenades). Sens dubte, aquest és un indicador que els models presenten heteroscedasticitat, la qual cosa suposaria la violació d'un altre dels supòsits (presència d'homoscedasticitat o variància constant) que s'han de complir per a la correcta estimació dels paràmetres per MQO. Per altra banda, en el cas del grup dels *Porters* no sembla que s'observi cap patró i es podria dir que els residus sí presenten variància constant. D'altra manera, per corroborar la presència d'heteroscedasticitat es pren com a referència el test de *Breusch-Pagan* el qual analitza si la variància dels residus d'una regressió lineal depenen dels valors de les variables independents. Les hipòtesis nul·la i alternativa, així com l'estadístic de prova són com segueix:

$H_0: variància\ constant$ vs $H_1: variància\ no\ constant$	$nR^2 \sim \chi^2_k, sota\ H_0$ R^2 fa referència al model de regressió: $\hat{e}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + u$

Per a la realització d'aquest contrast d'hipòtesis el que es fa és ajustar una regressió lineal variable endògena i variables exògenes de la qual són els residus al quadrat i les variables del model original, respectivament. El que es contrasta realment és si $\delta_1 = \delta_2 = \cdots = \delta_k = 0$ on k és el número de variables del model original; i es pren com a referència el coeficient de determinació de la regressió nova.

Taula 4.2.2.13 Taula resum del test de Breusch-Pagan

Model	***BP Statistic***	***P-value***
Davanters	111,58	$< 2{,}2^{-16}$
Mig-campistes	152,46	$< 2{,}2^{-16}$
Defenses	123,44	$< 2{,}2^{-16}$
Porters	15,58	0,21

Font: Elaboració pròpia

Amb els resultats obtinguts que es mostren a la taula 4.2.2.13, amb una confiança del 95%, és correcte rebutjar la hipòtesis nul·la de variància constant dels residus pels models referent als *Davanters*, *Mig-campistes* i *Defenses* ja que s'han obtingut p-valors molt per sota del valor establert $0{,}05$; es confirma el que s'havia dit a partir dels gràfics. Per altra banda, pel que fa al model dels *Porters* no hi ha suficients evidències estadísticament significatives com per rebutjar la hipòtesis nul·la, és a dir, com per dir que no comparteixen una variància constant. Així doncs, de nou, esdevé un problema greu sobre l'estimació per mínims quadrats ordinaris. El fet de rebutjar la hipòtesis d'homoscedasticitat comporta errors en el càlcul la matriu de variàncies i covariàncies dels estimadors β, perdent aquesta la propietat de *no esbiaixada*.

Per tal de corregir l'heteroscedasticitat, White (1980) va proposar el següent estimador consistent per a dita matriu de variàncies i covariàncies dels paràmetres β estimats:

$$\widehat{Var(\hat{\beta})} = (X^T X)^{-1} S_0 (X^T X)^{-1}, on\ S_0 = \sum^{n} e^2 x x^T$$

Aquesta matriu és un estimador apropiat de la matriu de variàncies de $\widehat{\beta}_{MQO}$ i també es pot denominar matriu de variàncies estimada robusta. Amb la funció d'R $vcovHC$ que proporciona el càlcul d'aquesta matriu d'estimacions robusta, i amb la funció $coeftest$, que permet realitzar els tests de significació individual aplicant esmentada matriu, s'ajusta per cadascun dels rols de *Davanter*, *Mig-campista* i *Defensa* un model de regressió lineal múltiple, essent d'aquesta manera els models finals elaborats. Cal insistir que el fet d'incloure aquesta matriu permet la validesa dels contrastos d'hipòtesis, que ara també seran denominats robustos, i la futura inferència que se'n derivi.

Una vegada ajustades les dades als models, a la taula 4.2.2.14 es mostra els resultats de manera compacta. Amb tot, és adient apuntar:

- De nou, com en casos anteriors, no es calcula el coeficient $\hat{\beta}$ associat la categoria $Bundesliga$ de la variable Lliga ja que aquesta es troba inclosa al terme independent de la regressió.

  Pel que fa a la significació individual dels coeficients $\hat{\beta}$ estimats, de forma general, amb una confiança del 95%, hi ha evidències estadísticament significatives com per a rebutjar la hipòtesis nul·la on es defineix $H_0: \hat{\beta} = 0$ o, el que és el mateix, es pot concloure que els coeficients dels predictors no són iguals a 0.

- Pel que fa als coeficients de determinació dels models, si el que es vol es comparar models que tinguin diferents números de predictors, com és en aquest cas, el que és correcte és prendre com a referència l'$R^2 - ajustat$, el qual també mesura el grau de variabilitat explicada pel model però incorporant el número de predictors d'aquest, evitant l'augment de per sí quan s'agrega un nou predictor al model, fins i tot quan no hi ha millora real. Així doncs, si es comparen els $R^2 - ajustat$s $0{,}781$, $0{,}803$, $0{,}759$ i $0{,}759$, corresponents als *Davanters*, *Mig-campistes*, *Defenses* i *Porters* respectivament, amb els dels models que es troben a la taula 4.2.2.6, es pot observar que en tots s'ha augmentat. Concretament, un $6{,}3\%$, $8{,}7\%$, $11{,}1\%$ i $8{,}7\%$. Es recorda que quant més a prop de l'1 (0) millor (pitjor). Per tant, atès als resultats obtinguts, es pot concloure que els models expliquen en gran mesura la variabilitat de les dades.

  Un altre criteri per avaluar la bondat de l'ajust dels models és el criteri d'Informació d'*Akaike* (AIC) proposat per Akaike (1974). Aquest indicador s'expressa de la següent manera:

$$AIC = -\frac{2l}{n} + \frac{2k}{n}$$

Taula 4.2.2.14 Taula resum de l'estimació robusta separant per categories del factor "Rol"

	Davanters			Mig-campistes			Defenses			Porters		
Predictors	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***	***Estimates***	***CI***	***p***
(Intercept)	5,638 ***	2,941 – 8,335	**<0,001**	2,844 **	0,933 – 4,755	**0,004**	2,189 *	0,115 – 4,263	**0,039**	1,028	-1,717 – 3,773	0,462
Lliga [LaLiga]	-0,192	-0,404 – 0,021	0,077	-0,250 ***	-0,393 – -0,107	**0,001**	-0,302 ***	-0,467 – -0,137	**<0,001**	0,334	-0,006 – 0,673	0,054
Lliga [Ligue 1]	-0,112	-0,318 – 0,094	0,287	-0,140	-0,293 – 0,013	0,072	-0,153	-0,324 – 0,019	0,080	0,305 *	0,012 – 0,597	**0,041**
Lliga [Premier League]	0,501 ***	0,212 – 0,790	**0,001**	0,425 ***	0,220 – 0,630	**<0,001**	0,685 ***	0,428 – 0,942	**<0,001**	1,032 ***	0,635 – 1,430	**<0,001**
Lliga [Serie A]	-0,243 *	-0,452 – -0,034	**0,023**	-0,222 **	-0,367 – -0,077	**0,003**	-0,436 ***	-0,596 – -0,277	**<0,001**	0,196	-0,109 – 0,501	0,207
edat	0,586 ***	0,383 – 0,788	**<0,001**	0,785 ***	0,641 – 0,928	**<0,001**	0,802 ***	0,644 – 0,959	**<0,001**	0,817 ***	0,621 – 1,013	**<0,001**
edat^2	-0,012 ***	-0,015 – -0,008	**<0,001**	-0,016 ***	-0,018 – -0,013	**<0,001**	-0,016 ***	-0,019 – -0,013	**<0,001**	-0,015 ***	-0,018 – -0,012	**<0,001**
VMTC	6,7265e-09 ***	0,000 – 0,000	**<0,001**	6,510e-09 ***	0,000 – 0,000	**<0,001**	6,881e-09 ***	0,000 – 0,000	**<0,001**	4,402e-09***	0,000 – 0,000	**<0,001**
VMTC^2	-5,6062e-18 ***	-0,000 – -0,000	**<0,001**	-4,900e-18 ***	-0,000 – -0,000	**<0,001**	-5,461e-18 ***	-0,000 – -0,000	**<0,001**	-3,781e-18*	-0,000 – -0,000	**0,034**
Min	0,001 ***	0,001 – 0,001	**<0,001**	5,691e-04 ***	0,000 – 0,001	**<0,001**						
Min^2	-1,7190e-07 ***	-0,000 – -0,000	**<0,001**	-9,122e-08 ***	-0,000 – -0,000	**<0,001**						
Gols_90	1,648 ***	1,270 – 2,027	**<0,001**									
Pass_ex	0,001 ***	0,000 – 0,001	**<0,001**	0,001 ***	0,000 – 0,001	**<0,001**	0,002 ***	0,001 – 0,002	**<0,001**	0,002 *	0,000 – 0,003	**0,034**
Dribl_ex	0,004 ***	0,003 – 0,006	**<0,001**	0,003 ***	0,002 – 0,005	**<0,001**						
Gols				0,065 ***	0,043 – 0,087	**<0,001**	0,113 ***	0,061 – 0,166	**<0,001**			

Assist		0,051 ***	0,029 – 0,073	**<0,001**	0,069 ***	0,035 – 0,104	**<0,001**		
Pass_thro_pc		0,017 ***	0,009 – 0,024	**<0,001**					
Pass_thro_pc^2		-1,221e-04 **	-0,000 – -0,000	**0,001**					
Duels_aeris_ex_pc		0,017 ***	0,008 – 0,025	**<0,001**	0,042 ***	0,032 – 0,051	**<0,001**	0,006 **	0,002 – 0,009	**0,002**
Duels_aeris_ex_pc^2		-1,734e-04 ***	-0,000 – -0,000	**0,001**	-3,548e-04 ***	-0,000 – -0,000	**<0,001**			
Pass_ex^2					-5,226e-07 ***	-0,000 – -0,000	**<0,001**			
Duels_ex					0,008 ***	0,004 – 0,011	**<0,001**			
xAssist								28,122 ***	12,418 – 43,826	**<0,001**
Parades_pc								0,011 ***	0,006 – 0,015	**<0,001**
Observations	543	924			884			295		
R^2 / R^2 adjusted	0,786 / 0,781	0,807 / 0,803			0,763 / 0,759			0,769 / 0,759		
AIC	1238,150	1935,782			1984,651			741,734		

** p<0,05 ** p<0,01 *** p<0,001*

Font: Elaboració pròpia

On l és el logaritme de la funció de versemblança dels valors estimats dels coeficients. El criteri que regeix aquest indicador és que quan més baix sigui el seu valor, o més a prop del $-\infty$ ja que no té cota inferior ni superior, millor ajust del model. Finalment, un cop definit, segons el criteri d'informació d'Akaike amb els resultats obtinguts, també s'arriba a la conclusió que els models s'ajusten millor.

- Pel que fa a la interpretació dels coeficients de les regressions ajustades, cal recordar, en primer lloc, que la variable endògena $Valor\ de\ mercat$ es troba transformada en logaritmes, la qual cosa implica que les interpretacions dels efectes marginals o parcials es faran en termes d'*elasticitat*, essent aquesta la sensibilitat de variació d'una variable -la variable endògena- respecte els canvis experimentats per una altra -variable explicativa o independent. Així, en termes general, la interpretació dels regressors és:

$$EMg = \frac{dy}{dX} = \%\triangle y = \left(100\,\hat{\beta}\right)\triangle X$$

Per exemple, si es vol saber l'efecte marginal que té l'augment d'una unitat en la variable $Gols\_90$ sobre el $Valor\ de\ mercat$ dels $Davanters$ s'obté:

$$EMg^{Davanters}_{Valor\ de\ mercat/Gols\_90}\% = 164{,}8$$

És a dir, la diferència en 1 gol més cada 90 minuts d'un davanter respecte un altre, essent aquests dos jugadors idèntics en les demés variables, provoca que el primer sigui un $164{,}8\%$ més car que el segon. Com que la variable $Valor\ de\ mercat$ varia en major mesura que la variable $Gols\_90$, en termes d'elasticitat, es diu que la relació entre aquestes dues variables és elàstica.

Seguint amb l'anterior, pel que fa a la interpretació correcta del factor $Lliga$ pels diversos models (l'única variable categòrica), aquesta seria: si es pren com a exemple la categoria $Premier\ League$, el fitxatge d'un jugador per un equip de la lliga anglesa, mantenint la resta de variables explicatives iguals, provocaria un augment del seu $Valor\ de\ mercat$ en un $50{,}1\%$ en el cas que fos un $Davanter$, un $42{,}4\%$ en el cas que fos un $Mig-campista$, un $68{,}5\%$ en el cas que fos $Defensa$ i un $103{,}2\%$ en el cas que el jugador tingués el rol de $Porter$. Aquests resultats ja eren d'esperar ja que en l'apartat d'anàlisi descriptiva s'ha pogut observar que, en general, els jugadors estan major valorats a la lliga anglesa. Per altra banda, si es fes el mateix, l'efecte sobre el valor de mercat d'un $Davanter$, $Mig-campista$ o $Defensa$ que és fitxat per algun equip de les lligues $LaLiga$ (lliga espanyola), $Serie\ A$ (lliga italiana) i $Ligue\ 1$ (lliga francesa) és perjudicial degut al signe negatiu dels coeficients associats i, per tant, el seu valor de mercat es veuria reduït pel simple fet de passar a formar part a alguna d'aquestes lligues. En canvi, pels $Porters$ sempre serà beneficiós canviar de lliga ja que tots els coeficients associats a les categories de la variable $Lliga$ són positius.

També, cal prestar atenció a aquells coeficients els quals presenten termes quadràtics. La inclusió d'esmentats termes serveix per captar la disminució o l'augment dels efectes marginals. Així, en presència d'un terme quadràtic, l'efecte marginal, prenent derivades, es defineix com:

$$EMg = \frac{dy}{dX} = \hat{\beta} + 2\widehat{\beta}'X$$

En el cas que $\hat{\beta}$ i $\widehat{\beta}'$ fossin de signes contraris i si igualem a 0 l'equació anterior s'obté el punt de canvi:

$$x^* = \frac{\hat{\beta}}{2\widehat{\beta}'}$$

Si $\hat{\beta} > 0$ i $\widehat{\beta}' < 0$, l'efecte marginal d'x sobre y serà positiu fins que x superi a x^*. Si $\hat{\beta} < 0$ i $\widehat{\beta}' > 0$, l'efecte marginal d'x sobre y serà negatiu fins que x superi a x^*. Així doncs, es pot exemplificar amb la variable $edat$. L'efecte marginal d'aquesta és:

$$EMg^{Davanters}_{Valor\ de\ mercat/edat}\% = 58{,}6 - 2 \times 1{,}2 \times edat$$

Així, en el cas d'un davanter amb 20 anys d'edat, el guanyar 1 any d'experiència li suposaria incrementar el seu valor de mercat en un 10,6%. Si s'iguala a 0 i s'aïlla la variable $edat$, es troba que a partir dels 24 anys aproximadament l'edat deixa de tenir un efecte positiu sobre el valor de mercat dels davanters.

- Una de les idees centrals sobre la qual es basa el present escrit és que, dins del mercat futbolístic, existeix una diferenciació de mercats segons quin rol es tracti atès que el rendiment futbolístic de cadascun dels rols es mesura per mitjà de mètriques específiques. Així doncs, els resultats obtinguts recolzen la idea anterior. Els quatre models ajustats per a cada rol estan definits per un conjunt de variables comunes i un altre d'específiques.

  El conjunt de variables comunes que comparteixen els quatre models són $Lliga$, $VMTC$, $VMTC^2$, $edat$, $edat^2$ i $Pass\_ex$, les quals expressen la lliga a la qual pertany el jugador, el valor de mercat de l'equip al qual pertany el jugador (i el seu terme quadràtic), l'edat del jugador i el nombre de passades exitoses que ha fet el jugador, respectivament. Efectivament, aquest grup de variables es caracteritzen per ser variables simples i de caràcter general. No obstant, per una banda, si es compara entre els diferents rols, en el cas de la variable $edat$ s'observa que, mentre que pels $davanters$ i $mig-campistes$ l'edat màxima que anul·la l'efecte positiu de la variable sobre el valor de mercat del jugador és aproximadament de 24 anys, la dels $defenses$ i $porters$ és major, 25 i 27 aproximadament, respectivament. En altres paraules, si es pren l'edat com un indicador de l'experiència que adquireix el jugador, la qual és una característica positiva, es podria dir que els $davanters$ i $mig-campistes$ assoleixen la màxima experiència als 24 anys, abans que els $defenses$ i $porters$ que ho fan als 25 i 27 respectivament. A partir d'aquestes edats límit, cada any que passa suposa una pèrdua sobre el valor de mercat. Per altra banda, en el cas del factor $Lliga$, es pot observar que només pel cas dels $porters$ el fet de jugar en qualsevol de les 5 lligues té un efecte positiu sobre el seu valor de mercat. En canvi, per la resta de rols, només la $Permier\ League$ fa augmentar-lo.

  Després, prestant atenció a les variables específiques de cada rol, les que defineixen el valor de mercat dels $Davanters$ són les variables Min, Min^2, $Gols\_90$ i $Dribl\_ex$, que defineixen els minuts jugats al llarg de la temporada (i el seu terme quadràtic), la ràtio de gols cada 90 minuts i el número de driblatges exitosos que ha fet el jugador,

respectivament. Cal ressaltar la variable $Gols\_90$, la qual mesura més aviat l'eficiència del jugador a l'hora de marcar gols que no pas, per exemple, el total de gols que marca, mesurat per la variable $Gols$ que s'inclou als models dels mig-campistes i defenses. Així, quan es tracti de mesurar el valor de mercat d'un davanter s'haurà de tenir en compte la seva ràtio de gols cada 90 minuts i no els gols totals que ha fet. Pel que fa a la variable $Dribl\_ex$, s'ha de considerar com un indicador de la qualitat tècnica del jugador. Els davanters són els jugadors més atacants d'un equip i en moltes ocasions han de sobrepassar els defenses de l'equip contrari amb la pilota als peus, driblant-los. Per tant, és raonable que un factor explicatiu del valor de mercat dels davanters sigui aquesta capacitat de driblatge, mesurada pel nombre de driblatges exitosos. També, s'ha de prestar atenció a l'absència de variables de caire defensiu, com $Duels\_ex$ o $Duels\_aeris$, i de variables que tenen a veure amb l'elaboració i construcció del joc, com $Pass\_thro$. D'alguna manera, es pot dir que dels davanters només s'espera que marquin gols i que siguin bons tècnicament.

Pel que fa als $Mig - campistes$, les variables específiques del seu model són Min, Min^2, $Dribl\_ex$, $Gols$, $Assist$, $Pass\_thro\_pc$, $Pass\_trho\_pc^2$, $Duels\_aeris\_ex\_pc$ i $Duels\_aeris\_ex\_pc^2$ que defineixen el nombre de minuts jugats al llarg de la temporada (i el seu terme quadràtic), el nombre de driblatges exitosos, el número de gols marcats, el número de assistències fetes, el percentatge de passades a l'espai respecte el total de passades intel·ligents fetes pel jugador (i el seu terme al quadrat) i el percentatge de duels aeris guanyats respecte el total de duels aeris jugats pel jugador (i el seu terme quadràtic), respectivament. Prestant atenció a la naturalesa futbolística de les variables, els mig-campistes són els jugadors més complets de tots o, si més no, són aquells que se'ls requereix que siguin bons tant ofensiva com defensivament, a més de ser bons creadors i directors de joc. La capacitat ofensiva es veu definida amb les variables $Gols$, que seria un indicador de la capacitat golejadora del jugador, i $Dribl\_ex$. La capacitat defensiva es veuria definida amb la variable $Duels\_aeris\_ex\_pc$. Atès que la posició natural dels mig-campistes és la zona central del camp i és on es donen molts casos de jugades on s'ha de disputar una pilota enlairada, se'ls exigeix que no només les competeixin, sinó que també les guanyin. També, aquesta variable podria ser un indicador de les facultats físiques del jugador ja que, a part de l'alçada, ser un bon rematador de cap exigeix força, capacitat de salt i presència física. Finalment, la capacitat de direcció i creació de joc es veuria explicada per les variables $Assist$ i $Pass\_thro\_pc$. Els mig-campistes són els jugadors que es troben posicionament entre els davanters i els defenses, així que s'encarreguen de connectar ambdues parts fent arribar la pilota als atacants i assistint-los amb passades que faciliten molt la resolució de la jugada en gol.

Si s'observa el model referent als $Defenses$, els factors explicatius específics per aquest grup de jugadors són $Pass\_ex$, $Pass\_ex^2$, $Gols$, $Assist$, $Duels\_aeris\_ex\_pc$, $Duels\_aeris\_ex\_pc^2$ i $Duels\_ex$, que es defineixen com número de passades exitoses (i el seu terme quadràtic), els gols marcats, el número de passades fetes, el percentatge de duels aeris guanyats (i el seu terme quadràtic) i el número de duels exitosos, respectivament. Després dels mig-campistes, els defenses conformen el segon grup de jugadors més complet. Destaquen les variables de caire defensiu com a factors explicatius d'aquest rol com són $Duels\_ex$ i $Duels\_aeris\_ex\_pc$. Els defenses són els que, en última instància, i deixant de banda el porter, s'encarreguen de defensar la

porteria i, per tant, és adient que els factors explicatiu vagin en aquesta direcció. Tanmateix, el seu valor de mercat també es veu incrementat pels gols que marquen i les assistències que donen als seus companys: per cada gol i assistència que fan de més el seu valor de mercat augmenta en un $11{,}3\%$ i $6{,}9\%$, respectivament. No és sorprenent que l'increment sobre el valor de mercat del marcar un gol més i fer una assistència més sigui major que la resta de coeficients. L'objectiu final del futbol com a joc és marcar gols i és per això que el marcar gols, o fer l'última passada abans que sigui gol, per a un jugador que "teòricament" no està destinat a fer-ho tingui tant de valor.

Finalment, les variables específiques que determinen una part del valor de mercat dels $Porters$ són $Duels\_aeris\_ex\_pc$, $xAssist$ i $Parades\_pc$, les quals defineixen el percentatge de duels aeris exitosos, la suma de les probabilitats que una passada sigui assistència i el percentatge de parades sobre el total d'intervencions al llarg de la temporada. S'ha de tenir en compte que un "duel aeri" per un porter no és exactament el mateix per un jugador que jugui en qualsevol altra posició. La gran majoria de duels aeris en els quals intervé un porter són dins l'àrea que envolta la porteria, zona de major incidència dels jugadors d'aquest grup, on són els únics que poden utilitzar les mans per atrapar la pilota. Així, un duel aeri guanyat per un porter és perquè ha atrapat la pilota amb les mans. Quan un porter destaca en aquest tipus d'accions es diu que té capacitat o presència aèria. Altrament, la variable $xAssist$ recull la suma de les probabilitats associades de totes les passades d'un jugador que dites passades acabin essent una assistència de gol, la qual cosa és un indicador de l'aportació del jugador al joc de l'equip. Amb tot, és raonable que, per un costat, el valor de mercat d'un porter vingui determinat per la seva capacitat aèria i la seva qualitat per parar els xuts del rival ja que és la seva funció específica; i per un altre, que també es valori la seva capacitat d'incidir en el joc col·lectiu de l'equip de manera positiva i que esdevingui un factor explicatiu i diferenciador sobre el seu valor de mercat.

V. CONCLUSIONS

En aquest treball final de grau s'ha treballat amb un conjunt de dades de tipus futbolístic, les quals proporcionaven el conjunt d'accions o esdeveniments espai-temporals de tots els partits de la temporada 2017/2018 de les 5 lligues europees de futbol principals - *Premier League*, *LaLiga*, *Bundesliga*, *Serie A* i *Ligue 1*-, a part de les metadades referents als partits, equips i jugadors que les composen. Juntament amb l'agregació dels valors de mercat dels jugadors extrets de la web *Transfermarkt.com*, referent a la temporada analitzada, amb el conjunt total de dades s'ha pogut elaborar, per una banda, una anàlisi economètrica amb models de regressió lineal múltiple estimats per mínims quadrats ordinaris on la variable endògena és el valor de mercat, de tal manera que s'ha determinat per a cada categoria de rol de jugador (diferenciant entre *davanters*, *mig-campistes*, *defenses* i *porters*) un model específic, o *mercat específic*, definits cadascun per variables comunes entre els models i variables singulars i pròpies. Per altra banda, s'ha construït un model de predicció de gol el qual avalua la probabilitat de gol que té un xut en funció de les seves característiques com per exemple la distància i l'angle de tir respecte el centre de la porteria, el qual permet enfocar l'anàlisi futbolístic des d'un nou i singular punt de vista.

La revisió de la literatura ha consistit en repassar els últims treballs que s'han dut a terme enfocats a l'estudi dels factors explicatius i determinants dels preus de traspassos de jugadors, salaris de jugadors i valors de mercat dels jugadors. S'ha pogut observar, per un costat, que la metodologia que generalment empren és la de models lineals estimats per mínims quadrats ordinaris, amb l'agregació d'alternatives com models de regressió quantílica o model Tobit. Per altre costat, s'ha vist que tot i treballar amb bases de dades diferents i metodologies que poden diferir, els autors troben que els factors i variables explicatives que determinen tant els salaris com preus de traspàs o valors de mercat dels jugadors són molt similars, de les quals es poden destacar el número de gols que marca el jugador, el número de partits que juga, el número de vegades que ha estat seleccionat per a representar el seu país en competicions internacionals, l'edat, el club pel qual juga, el rol del jugador, la duració del contracte del jugador... Així doncs, s'ha pogut advertir que totes les aproximacions existents tenen una clara infrautilització de mètriques de rendiment futbolístic (únicament s'han utilitzat els gols, les assistències i les passades com a mètriques de rendiment del jugador) i, a part, que aquest enfocament podria ser més específic pel que fa al rol del jugador atès que en funció de quina posició es tracti el jugador, aquest té unes funcions determinades definides pel propi esport. Eventualment, aquest fet es transmet en la caracterització del jugador per diferents mètriques en funció del rol que prengui.

La primera part del treball fa referència a la construcció del model predictiu de resposta binària que avalua la probabilitat de gol que sia un xut en funció de les seves característiques. Com diu la pròpia definició, la variable resposta que es pretén ajustar és binària, en aquest cas avalua si un xut ha sigut gol (1) o no (0). La mostra que s'ha utilitzat ha consistit en un total de 40.461 tirs recollits a partir de les taules que es disposava referents a les 5 lligues europees principals la temporada 2017/2018.

La primera de les fases ha consistit en l'avaluació i la recol·lecció de les característiques dels tirs, les quals conformarien les variables explicatives del model. Aquestes característiques o variables que s'han recollit són: distància i angle de tir ($distancia, angles$), la cama (o part del cos) amb la qual s'ha executat el tir ($partCos$) i si era la cama dominant del jugador (GB), si el xut venia precedit d'un contra-atac ($counter$), si el xut venia precedit d'una centrada

($centrada$), el rol del jugador que ha fet el tir (Rol) i la lliga ($League$). Seguidament, s'ha procedit a fer una anàlisi descriptiva conjunta dels factors anteriors amb la variable resposta i s'ha obtingut que, efectivament, a priori tots semblarien factors explicatius del model excepte el factor $League$. També, s'ha pogut observar que els xuts es concentren a la part central de l'àrea que envolta la porteria i que quanta menys distància i quant més angle de tir es tingui, millor.

A la fase de modelització s'ha arribat a ajustar un total de 22 models mitjançant la metodologia *stepwise* i s'ha arribat a la conclusió que el model de predicció de gol que millor s'ajusta a les dades té com a factors explicatius les variables $distancia$, $angles$, $counter$, Rol, $centrada$, GB i la interacció $distancia{:}\,centrada$. Els resultats obtinguts indiquen, per una banda, que quant més llunyà sigui el xut menys probabilitat de gol i que si el jugador que l'executa és porter també la condiciona, de manera que la fa reduir. Per altra banda, s'extreu que el fet que la jugada precedent del xut sigui un contra-atac fa augmentar el quocient de probabilitat (*odds*) que sigui gol en un 70%; que el fet que el xut sigui realitzat amb la cama dominant del jugador fa augmentar l'*odds* de gol en un $12{,}74\%$; que el fet que el jugador que xuti a porteria jugui a la posició de davanter i mig-campista fa augmentar l'*odds* que el xut sigui gol en un $49{,}18\%$ i un $32{,}91\%$.

Finalment, aquesta primera part del treball ha finalitzat amb l'exposició de dos exemples d'aplicacions d'anàlisi que es podria fer a partir de la construcció d'un model d'aquest tipus. S'ha presentat la principal mètrica que es pot construir que resulta en la suma de les probabilitats de gol de cada tir, anomenada *xGols* o gols esperats, la qual permet avaluar el rendiment d'un futbolista (especialment aquells que juguen en la posició de davanters) quan es compara amb els gols realitzats: si la suma dels gols marcats supera els *xGols*¸ es pot dir que ha rendit per sobre del que s'esperava; contràriament, quan els *xGols* superen els gols realitzats es diria que el seu rendiment ha sigut inferior a les expectatives. Això no obstant, disposant de la probabilitat de gol de cada tir ha permès la construcció del *Mapa de tir* d'un jugador, el qual conformaria una de les aplicacions. Aquest mapa de tir mostra la posició dels xuts sobre les delimitacions del camp de futbol d'un jugador, d'un equip, etc. amb la seva probabilitat de gol associada i distingint amb quina part del cos s'ha efectuat el xut. Aquest tipus d'il·lustració permet analitzar i avaluar més detallada i exhaustivament l'estil de joc i, el més interessant, permet la comparació de diversos jugadors. La segona de les aplicacions que s'han presentat és l'anàlisi d'un partit. Amb el *gràfic cronològic de gols esperats* d'un partit es pot observar els *xGols* acumulats de cada equip al llarg dels minuts que dura un partit (més o menys 90 minuts): quan la corba d'un equip es troba per sobre de la del seu rival vol dir que el primer ha estat superior al segon (i viceversa). Amb l'exposició d'altres estadístiques complementàries del partit, l'anàlisi que es pot fer d'aquest és molt més exhaustiu i detallat respecte l'anàlisi que es faria amb només la presentació d'estadístiques com els gols o la possessió de pilota.

Abans de prosseguir amb la segona part del treball, cal comentar que per la pròpia naturalesa de les dades, el model de predicció de la probabilitat de gol d'un tir és limitat. Per experiència pròpia es té consciència que hi ha variables i factors que afectarien i modificarien els resultats que no s'han tingut en compte. La posició dels jugadors que defensen el xut, la posició del porter en el moment del tir, la posició relativa de la pilota respecte el terra (és a dir, l'altura del xut) o l'element del peu amb el qual es xuta (interior, exterior o l'empenya[12]) són exemples

[12] Part superior del peu corresponent al metatars.

d'elements del joc que intervenen directament en que un xut acabi sota la xarxa o no, és a dir, sobre la probabilitat de gol.

La segona part d'aquest treball final de grau s'ha compost de l'anàlisi economètrica del valor de mercat dels jugadors de les 5 lligues europees més importants, ajustant models de regressió lineal múltiple estimats per mínims quadrats ordinaris. La mida de la mostra inicialment era de 2971 jugadors, però per manca de dades i com a conseqüència de l'aplicació del logaritme sobre el seu valor de mercat s'han hagut de suprimir 129, de manera que la mida mostral final ha quedat en 2662 futbolistes.

La primera etapa ha consistit en l'anàlisi descriptiva de les dades per tal de detectar patrons o relacions entre les diverses variables que caracteritzen cadascun dels jugadors. S'ha disposat de 48 variables potencialment explicatives de la variable resposta $Valor\ de\ mercat$. Aquest conjunt de variables, tret dels factors $Lliga$ i Rol, conformen un llistat de mètriques de rendiment futbolístic que recull, entre d'altres, el número de gols marcats i el percentatge de duels aeris guanyats. Aquestes mètriques s'han aplegat a partir de les accions individuals dels jugadors que s'ha anat donant al llarg de la temporada 2017/2018 que es disposen a la base de dades descrita anteriorment. Amb tot, s'ha pogut determinar que els jugadors més valorats del mercat són els davanters, seguit dels mig-campistes, defenses i porters; i que, en general, a la lliga anglesa les valoracions dels jugadors són majors que la resta de lligues. Cal ressaltar, però, que a les lligues alemanya i espanyola els defenses i mig-campistes tenen valoracions semblants, igual que a la lliga italiana pel que fa els davanters i mig-campistes.

Pel que fa a les variables numèriques, el criteri que s'ha seguit ha consistit en calcular el coeficient de correlació de Pearson d'aquestes respecte la variable endògena i seleccionar aquelles on el coeficient, en valor absolut, fos major que $0{,}3$. En primer lloc, s'ha fet una distinció entre els porters i la resta de jugadors de camp, però ràpidament s'ha cregut necessària una diferenciació més detallada per categories del factor Rol sobre el grup dels jugadors de camp, entre davanters, mig-campistes i defenses ja que moltes de les variables que prenien importància només tenien en compte característiques ofensives del joc (la qual cosa fa pensar en la influència del grup dels davanters), la qual cosa feia perdre l'impacte de les demés variables. Així, eventualment, s'han calculat els coeficients de correlació de Pearson pels 3 grups de jugadors i s'han seleccionat les variables d'acord amb el criteri establert amb anterioritat.

Arribats a aquest punt, s'ha fet la primera estimació per mínims quadrats ordinaris separant per categories del factor Rol, on la variable endògena, cal recordar, és el valor de mercat havent aplicat logaritmes. D'aquesta s'ha observat un problema existent de multicol·linealitat severa, de manera que s'ha decidit eliminar aquelles variables que la generaven. Val la pena comentar que, tot i que es justifica l'eliminació de certes variables, sempre hi haurà el debat sobre si l'elecció s'ha fet correctament i si la reducció de la multicol·linealitat i la dispersió es fa a canvi de la introducció de cert biaix.

A continuació, s'ha valorat la possibilitat de la inclusió de certes variables amb termes quadràtics mitjançant el test de curvatura de Tukey. Seguidament, un cop construïts els models incloent les variables necessàries amb termes quadràtics, s'ha volgut comprovar els supòsits elementals de l'estimació per mínim quadrats ordinaris. Per una banda, s'ha pogut detectar la no normalitat dels residus en els models referents als grups de davanters, mig-campistes i defenses; i s'ha intentat corregir mitjançant l'exclusió de valors influents i

d'"apalancament". Contrastant la normalitat dels residus de nou (havent fet l'exclusió), s'ha pogut aconseguir la correcció pel que fa al grup dels defenses, però no a la resta. Per altra banda, també pels jugadors amb el rol de davanters, mig-campistes i defenses, s'ha observat heteroscedasticitat. Per tal de corregir-la, s'ha optat per l'estimació robusta que proposa White (1980) substituint la matriu de variàncies i covariàncies dels estimadors β per una estimació de la mateixa.

Amb tot, a partir de l'ultima estimació robusta s'obté com a variables explicatives:

Taula 5.1 Variables explicatives per a cadascun dels models estimats

Rol	***Variables***
Davanter	$LLiga, edat, edat^2, VMTC, VMTC^2, Min, Min^2, Gols\_90, Pass\_ex, Dribl\_ex.$
Mig-Campistes	$Lliga, edat, edat^2, VMTC, VMTC^2, Min, Min^2, Pass\_ex, Dribl\_ex, Gols, Assist, Pass\_thro\_pc, Pass\_trho\_pc^2, Duels\_aeris\_ex\_pc, Duels\_aeris\_ex\_pc^2.$
Defenses	$Lliga, edat, edat^2, VMTC, VMTC^2, Pass\_ex, Pass\_ex^2, Gols, Assist, Duels\_aeris\_ex\_pc, Duels\_aeris\_ex\_pc^2, Duels\_ex.$
Porters	$Lliga, edat, edat^2, VMTC, VMTC^2, Pass\_ex, Duels\_aeris\_ex\_pc, xAssist, Parades\_pc.$

Font: Elaboració pròpia

En conclusió, les variables que conformen el valor de mercat per a les diferents categories de rols dels jugadors són diferents i específiques definides per la funció específica que implica cada posició determinada. Mentre que pels davanters les variables explicatives són els gols que fan cada 90 minuts o el número de driblatges exitosos que fan, dels mig-campistes se'ls valoren altres mètriques d'àmbits diferents com pot ser el percentatge de passades penetrants que fan o el percentatge de duels aeris que guanyen. Tanmateix, els R^2 obtinguts per cadascun dels grups de jugadors es valoren de manera positiva, de manera que s'ha aconseguit explicar un $78{,}6\%$, $80{,}7\%$, $76{,}3\%$ i $76{,}9\%$ de la variabilitat de les dades pel que fa als davanters, mig-campistes, defenses i porters.

Caldria comentar que hagués estat interessant poder incloure altres variables relacionades amb l'estat físic del jugador (com per exemple els quilòmetres recorreguts) que, per experiència pròpia, són factors que intervenen en el rendiment del futbolista. També, com s'ha vist l'apartat de revisió de la literatura, el número d'anys restants del contracte del jugador és un factor a tenir en compte en el seu rendiment, de manera que el concepte de *moral hazard* entraria en joc, en el sentit que un jugador podria millorar el seu rendiment coneixent el seus anys restants de contracte i, per tant, influir en la seva valoració en el mercat. Independentment de l'anterior, també cal ressaltar l'obtenció de no normalitat dels residus quan s'estima per mínims quadrats ordinaris en el cas dels davanters i mig-campistes. Una de les solucions que es proposa seria l'estimació d'una regressió quantílica per millorar la qualitat de les estimacions.

Per finalitzar, a mode de suggeriment, la construcció de models de predicció de la probabilitat de gol i del valor de mercat dels jugadors, conformaria un conjunt d'eines de suport

complementàries per als clubs de futbol professional que els ajudaria a la part d'avaluació i presa de decisions a l'hora de fitxar un jugador. També, seria interessant aplicar la metodologia d'anàlisi de components principals (*PCA*) sobre el conjunt de dades, de manera que s'aconseguís agrupar els jugadors segons les seves característiques. A més, permetria la comparació de jugadors i observar quins s'assemblen més entre ells, la qual cosa ajudaria, en un cas hipotètic, a un club quan ha de trobar el millor jugador del mercat que reemplaci a una baixa de la seva plantilla.

VI. BIBLIOGRAFIA I WEBGRAFIA

1. *Ajadi, T., Bridge, T., Hansons, C., Hammond T. and Udwadia, Z. (2021). Testing times. Football Money League. https://www2.deloitte.com/content/dam/Deloitte/es/Documents/tecnologia-media-telecomunicaciones/Deloitte-ES-tmt-Football-Money-League-2021.pdf*
2. *Bernabé, D. (2021, març). El tenis, la F1 y el baloncesto, los deportes más vseguidos en España después del fútbol. Kantar, https://www.kantar.com/es/inspiracion/consumidor/el-tenis-la-f1-y-el-baloncesto-los-deportes-mas-seguidos-en-espana*
3. *Coluccia, D. Fontana, S. and Solimene, S. (2018) 'An application of the option-pricing model to the valuation of a football Player in the 'Serie A League'', Int. J. Sport Management and Marketing. 18(1-2). 155-168.* https://doi.org/10.1504/IJSMM.2018.091345
4. *Deutscher, C., and Büschemann, A. (2016). Does Performance Consistency Pay Off Financially for Players? Evidence From the Bundesliga. Journal of Sports Economics, 17(1), 27–43. https://doi.org/10.1177/1527002514521428*
5. *Felipe, J., Fernandez-Luna, Á., Burillo, P., de la Riva, L., Sánchez-Sánchez, J. and García-Unanue, J. (2020). Money Talks: Team Variables and Player Positions that Most Influence the Market Value of Professional Male Footballers in Europe. Sustainability 2020. 12(9), 3709. https://doi.org/10.3390/su12093709*
6. *Frick, B. (2007). The Football Players' Labor Market: Empirical Evidence from the Major European Leagues. Scottish Journal of Political Economy. 54(3), 422-446.* https://doi.org/10.1111/j.1467-9485.2007.00423.x
7. *Frick, B. (2011). Performance, Salaries and Contract Length: Empirical Evidence from German Soccer. International Journal of Sport Finance. 6. 87-118*
8. *Garcia-del-Barrio, P. and Pujol, F. (2021), "Recruiting talent in a global sports market: appraisals of soccer players' transfer fees", Managerial Finance, 47(6), 789-811. https://doi.org/10.1108/MF-04-2020-0213*
9. *Gómez, E. (27/08/2019). Así calcula Transfermarkt el valor de los jugadores de fútbol. Diario AS. https://as.com/futbol/2019/08/27/mas_futbol/1566863263_499442.html*
10. *Lee, S. and Harris, J. (2012) Managing excellence in USA Major League Soccer: an analysis of the relationship between player performance and salary, Managing Leisure. 17(2-3), 106-123. https://doi.org/10.1080/13606719.2012.674389*
11. *Llei Bosman. (12 abril 2021). https://ca.wikipedia.org/w/index.php?title=Llei_Bosman&oldid=26955634*
12. *Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. and Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. Nature Scientific data, 6(1), 1-15. https://www.nature.com/articles/s41597-019-0247-7*
13. *Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., and Giannotti, F. (2019). PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transactions on Intelligent Systems and Technology (TIST), 10(5), 1-27.* https://doi.org/10.1145/3343172

14. *PricewaterhouseCoopers Asesores de Negocios, S.L. (2018). Impacto económico, fiscal y social del fútbol profesional en España. https://newsletter.laliga.es/upload/media/multimedia/0001/45/9478d0ec2c82057e0b41c762f06581ef2e434d04.pdf*
15. *Tippett, J. (2019). The Expected Goals Philosophy: A Game-Changing Way of Analysing Football. Publicació independent.*
16. *Transfermarkt.com [últim accés 28/2/2021]*
17. *Tunaru, R., Clarke, E. and Viney, H. (2005) 'An option pricing framework for valuation of football players', Review of Financial Economics, 14(3-4), 281-295.*
18. *Unió d'Associacions Europees de Futbol (25 abril 2021). https://ca.wikipedia.org/w/index.php?title=Uni%C3%B3_d%27Associacions_Europees_de_Futbol&oldid=27040196*
19. *Yaldo,L. and Shamir,L.(2017).Computational Estimation of Football Player Wages. International Journal of Computer Science in Sport, 16(1) 18-38. https://doi.org/10.1515/ijcss-2017-0002*