Kandidatuppsats i Statistik

Individuella egenskapers betydelse för målvakters matchbetyg

En analys med linjär mixad modellering av data från Football Manager 2022

Duy Thai Pham
Viet Tien Trinh

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Vårterminen, 2025

Handledare: Isak Hietala, universitetsadjunkt
Examinator: Isak Hietala, universitetsadjunkt

Sammanfattning

Denna studie undersöker i vilken utsträckning individuella målvaktsegenskaper i fotboll kan förklara variationen i deras genomsnittliga matchbetyg. Analysen bygger på longitudinell, simulerad data från spelet Football Manager 2022, där 51 målvakter följdes över 19 säsonger, vilket resulterade i totalt 4 597 observationer. För att hantera datans hierarkiska struktur, där samma målvakt förekommer upprepade gånger, användes linjära mixade modeller (LMM) med ett slumpmässigt intercept för varje målvakt.

Analysen inleddes med en full modell som inkluderade 35 spelaregenskaper, inklusive flera som är specifika för målvaktspositionen. För modellurval användes funktionen step() från paketet lmerTest i R, vilken genomför bakåtseliminering baserat på AIC. Eftersom step() returnerar en klassisk linjär modell utan slumpmässiga effekter, och datan är tydligt hierarkiskt strukturerad, skattades även en linjär mixad modell för att bättre representera datans beroendestruktur.

Den slutgiltiga modellen innehöll 14 förklaringsvariabler, varav 13 visade ett statistiskt signifikant samband med matchbetyg ($p < 0.05$). Positiva samband observerades för aggression (agg), agility (agi), bravery (bra), command of area (cmd), concentration (cnt), free kick (fre), pace (pac) och reflexes (ref). Negativa samband identifierades för acceleration (acc), eccentricity (ecc), first touch (fir), handling (han) samt off the ball (otb). Endast determination (det) uppvisade inget signifikant samband.

Trots dessa fynd var modellens totala förklaringsgrad låg. De fixa effekterna förklarade endast 5,8% av variationen i betygen (marginal $R^2 = 0.058$), och den konditionella förklaringsgraden, som även inkluderar slumpmässiga effekter, var 6,1% ($R^2 = 0.061$). Den slumpmässiga interceptvariansen var mycket liten (0.0017), vilket tyder på att skillnader mellan målvakter har en begränsad inverkan på betygssättningen.

Residualanalysen visade inga allvarliga brott mot modellens antaganden, men viss skevhet och enstaka outliers noterades. Sammantaget indikerar resultaten att även om vissa individuella egenskaper påverkar betygen, spelar matchspecifika och kontextuella faktorer, såsom motståndarnivå, matchsituationer eller subjektiva bedömningar troligen en större roll. Detta är i linje med tidigare forskning som betonar kontextens betydelse vid spelarevaluering

Abstract

This study investigates the extent to which individual goalkeeper attributes can explain variation in their average match ratings. The analysis is based on longitudinal simulated data from Football Manager 2022, covering 51 goalkeepers over 19 seasons, resulting in a total of 4,597 observations. To account for the hierarchical structure of the data, where each goalkeeper appears repeatedly, linear mixed-effects models (LMM) with random intercepts per goalkeeper were used.

The analysis began with a full model including 35 attributes, encompassing both general player characteristics and goalkeeper-specific traits. For model selection, the step() function from the lmerTest package in R was applied, which performs backward elimination based on AIC. Since step() yields a standard linear regression model without random effects, a final linear mixed-effects model was estimated to appropriately reflect the nested structure of the data.

The final model included 14 predictor variables, of which 13 showed statistically significant associations with match ratings ($p < 0.05$). Positive associations were found for aggression, agility, bravery, command of area, concentration, free kick, pace, and reflexes. Negative associations were observed for acceleration, eccentricity, first touch, handling, and off the ball. Only determination was found to be non-significant.

Despite these significant effects, the model's explanatory power was low. The fixed effects accounted for only 5.8% of the variance in ratings (marginal $R^2 = 0.058$), and the conditional R^2, including both fixed and random effects was 6.1%. The variance of the random intercept was minimal (0.0017), indicating that differences between goalkeepers contributed little to explaining variation in ratings.

Residual diagnostics revealed no major violations of model assumptions, although slight skewness and a few outliers were present. Overall, the results suggest that while individual goalkeeper attributes play a role, contextual and match-specific factors, such as opponent strength, match situations, or subjective elements of the rating system likely have greater influence. These findings are consistent with previous research highlighting the central role of context in player evaluation.

Förord

Vi vill uttrycka vårt varma tack till vår handledare Isak Hietala för hans värdefulla vägledning, stöd och konstruktiva återkoppling under arbetets gång.

Innehåll

Figurer

Tabeller

1. Introduktion

Fotboll är en av världens mest populära sporter med en global publik och en växande industri kring prestationsbedömning och spelaranalys. I en fotbollsmatch består det av två lag med elva spelare vardera som får använda vilken del av kroppen som helst förutom händer och armar för att manövrera bollen in i motståndarlagets mål. Ett undantag är målvakten som får hantera bollen men endast inom det avgränsade straffområdet framför målet. Det lag som gör flest mål under matchen vinner (Encyclopaedia Britannica, 2025).

Traditionellt finns det tre huvudområden på planen, vilka är försvar, mittfält och anfall. Varje position är tydligt definierad, där försvarare försvarar, anfallare attackerar och mittfältare fungerar som en länk mellan de två med inslag av båda rollerna. Emellertid, i takt med matchutvecklingen kan spelarnas roller bli mer komplexa och dynamiska och gränserna mellan positionerna är numera inte lika tydliga, vilket medför att individuella roller är mer flexibla (Soccer Coaching Pro, 2023).

Målvaktens huvudsaklig uppgift handlar om att förhindra att bollen hamnar i det egna målet. Dock sträcker sig rollen i dagens fotboll långt bortom detta grundläggande ansvar. Målvakter förväntas numera inte bara vara den sista utposten i försvaret, utan även bidra aktivt i uppbyggnaden av spelet. I ett lag som spelar med hög press och högt försvar krävs det att målvakten är trygg med bollen vid fötterna och kan fungera som en spelbar passningspunkt. Dessutom för att kunna hantera snabba avslut från olika vinklar och avstånd krävs goda reflexer, snabb reaktionsförmåga, smidighet och koordination (Soccer Coaching Pro, 2023).

I professionell fotboll spelar målvakten en avgörande roll för lagets defensiva stabilitet. Trots det är deras insats ofta svår att bedöma objektivt, eftersom traditionella statistikmått som mål och assist inte fångar deras prestationer på ett rättvist sätt.

En fotbollsspelares matchbetyg är numeriska poäng som hen får efter en match. Betyget baseras på olika prestationsmått som exempelvis passningsprecision, gjorda mål, vunna tacklingar och spelarens övergripande bidrag till laget. Syftet med dessa betyg är att kvantifiera hur väl en spelare presterade i en specifik match, vilket möjliggör mer objektiva jämförelser mellan spelare och prestationer över tid. En kombination av avancerad statistik, expertanalyser och ibland även fansens åsikter används av de flesta betygssättningssystem. Denna kombination ger en helhetsbild av prestationen och fångar både mätbara insatser och mer svårdefinierade faktorer som positionsspel, speluppfattning och inflytande (Hugh, 2024).

1.1 Bakgrund

Tidigare forskning visar att vissa prestationsmått har större påverkan på spelarnas matchbetyg än andra. I en studie av Ball et al. (2025) undersöks sambandet mellan över 70 olika prestationsmått och

det tilldelade matchbetyget bland tre av de mest använda betygssättningssystemen, vilka är WhoScored, FotMob och Sofascore. Resultaten visar att offensiva mått, såsom skott på mål, nyckelpassningar och lyckade dribblingar har starkast samband med högre betyg. Defensiva mått, såsom rensningar uppvisar ett svagare men fortfarande signifikant samband. Studien visar även att det finns systematiska skillnader mellan betygssättningssystemen, där WhoScored generellt ger lägre betyg än de andra.

Trots att denna typ av analys är väletablerad för utespelare, är forskningen kring målvakter fortfarande begränsad. De flesta tidigare studier bygger på matchstatistik, exempelvis antal räddningar, passningar eller rensningar snarare än på spelarnas individuella egenskaper. För målvakter, vars prestationer ofta består av färre men mer avgörande moment, kan sådan statistik vara otillräcklig för att ge en rättvisande bild av deras insats. Det finns därför ett behov av att undersöka hur både generella egenskaper som är relevanta för alla fotbollsspelare, samt egenskaper som är specifika för målvaktsrollen, såsom reflexer, kommunikation och speluppfattning kan bidra till att förklara variationen i matchbetyg. Få studier har hittills fokuserat på sambandet mellan sådana egenskaper och matchbetyg, vilket gör detta till ett viktigt område som kan ge en mer nyanserad förståelse för hur målvakters prestationer bör bedömas.

Vidare visar tidigare forskning att matchbetyg även har kommersiell betydelse. He, Cachucho och Knobbe (2015) fann ett positivt samband mellan spelares genomsnittliga matchbetyg och deras marknadsvärde, vilket tyder på att betygen påverkar spelarens ekonomiska värde på transfermarknaden. Detta stärker relevansen av att studera hur matchbetyg uppstår, särskilt för målvakter där sådan forskning är begränsad.

1.2 Syfte

Syftet med denna undersökning är att analysera om och hur olika spelaregenskaper kan förklara variationen i målvakters matchbetyg. Fokus ligger både på generella egenskaper som är relevanta för alla fotbollsspelare och egenskaper som är specifika för målvaktsrollen. För att ta hänsyn till att samma målvakt förekommer i flera matcher används en modell med mixade effekter, vilket möjliggör att både fixa effekter, såsom målvaktens egenskaper och slumpmässiga effekter, såsom variation mellan målvakter, inkluderas i analysen.

1.2.1 Frågeställning

I vilken utsträckning kan både generella spelaregenskaper och målvaktsspecifika egenskaper förklara variationen i målvakters matchbetyg, och vilka av dessa egenskaper har ett signifikant samband när både fixa och slumpmässiga effekter beaktas?

1.3 Etiska och samhälleliga aspekter

I denna studie används simulerad data, vilket innebär att inga riktiga personer ingår i materialet. Därmed finns det ingen risk att enskilda individer kan identifieras eller att personuppgifter exponeras. På grund av att datan är fiktiv, behöver inga särskilda åtgärder vidtas för att skydda deltagarnas integritet.

Ur ett samhälleligt perspektiv kan studiens resultat bidra till en bättre förståelse för hur olika egenskaper påverkar målvakters prestationer enligt etablerade betygssättningssystem. En sådan förståelse kan vara värdefull både för idrottsanalytiker, tränare och rekryterare som vill fatta mer informerade beslut baserat på objektiva bedömningsgrunder.

2. Data

Här ges en beskrivning av var data kommer ifrån och en kortfattad beskrivning av variabler och observationer i det använda datamaterialet samt av eventuella transformationer och skapande av nya variabler. Eventuellt kan tabeller och/eller diagram över data presenteras.

2.1 Data beskrivning

I denna studie används simulerad data från The Simulation Soccer League (SSL), en virtuell och communitydriven fotbollsliga där spelare skapas, utvecklas och deltar i matcher över flera säsonger. SSL är inte en officiell del av Football Manager-serien, men ligan bygger sin matchsimulering på Football Manager 2022, ett avancerat fotbollssimuleringsspel utvecklat av Sports Interactive och publicerat av SEGA.

Football Manager är ett avancerat spel där användaren tar rollen som manager för ett lag och ansvarar för allt från taktik, träning och laguttagning till scouting, spelarövergångar och ekonomi. Varje spelare representeras genom ett stort antal attribut som beskriver deras tekniska, mentala och fysiska förmågor, vilka bedöms på en skala från 1 till 20, där högre värden motsvarar högre skicklighet (Football Manager, 2023). Spelet är internationellt välkänt för sin höga grad av realism och har även uppmärksammats för sin användning inom scouting och prestationsanalys på professionell nivå (Stuart, 2014).

I SSL genereras spelarstatistik och matchbetyg genom att spela upp matcher i Football Manager, där spelare först konstrueras via externa verktyg såsom Player Builder Tool och importeras till spelets databas. Football Manager används alltså som en simuleringsmotor, medan själva ligastrukturen, matchadministrationen och datahanteringen sker inom SSL:s egen plattform (Canadice, 2023; Simulation Soccer League, 2024).

Enligt SSL:s regelverk ska spelarnas attribut sättas med ett lägsta initialvärde på 5. Undantaget är egenskaperna Natural Fitness och Stamina, vilka alltid är satta till det maximala värdet 20 för alla nya spelare. Under karriären kan attributen därefter utvecklas genom träning och matchprestation (Simulation Soccer League, 2024).

Datamaterialet omfattar 19 säsonger. Denna undersökning fokuserar enbart på målvakter, vilka utgör de primära observationsenheterna. Eftersom flera målvakter deltar under flera säsonger och varje målvakt kan spela flera matcher, är datamaterialet longitudinellt och observationerna kan därmed vara beroende.

Datamaterialet innehåller variabler som kan delas in i fyra kategorier. Först finns identifierande variabler, såsom spelarens namn, säsong och matchnamn som används för att särskilja observationer. Den andra kategorin består av matchrelaterade information, exempelvis motståndarlag och matchresultat. Den tredje kategorin rör matchspecifik statistik, som antalet spelade minuter och det betyg spelaren

fått för sin prestation. Slutligen finns individuella egenskaper, vilka beskriver spelarens fysiska, tekniska och mentala färdigheter. Alla variabler som används i denna studie presenteras i tabellen nedan.

Tabell 2.1: Beskrivning av de variabler i datamaterialet, inklusive deras förkortningar i R och definitioner.

Variabler	Beskrivning
Average.rating	Spelares prestation betygsatt av FM
acc	Acceleration, hur snabb spelaren är på att öka tempo
aer	Aerial Reach, Beskriver hur högt en målvakt når i luften
agg	Aggression, hur aggressiv spelaren är i olika situationer
agi	Agility, hur kvick spelaren är på att byta riktning
ant	Anticipation, hur snabbt spelaren kan reagera på händelser
bal	Balance, hur stabil spelaren är vid fr.a. luftdueller
bra	Bravery, hur ofta spelaren sätter sig i situationer som kan orsaka skada
cmd	Command of Area, målvaktsegenskap
com	Communication, målvaktsegenskap
cmp	Composure, hur lugn spelaren är i stressiga situationer
cnt	Concentration, hur fokuserad spelaren är under en match
cor	Corner, hur bra spelaren är att genomföra en hörna
cro	Crossing, hur bra spelaren är att genomföra inlägg till målområdet
dec	Decisions, hur smart spelaren är att ta rätt beslut på planen
det	Determination, hur benägen spelaren är att lyckas på planen
dri	Dribbling, hur bra spelaren är på att komma förbi en motståndare med bollen
ecc	Eccentricity, målvaktsegenskap
fin	Finishing, hur bra spelaren är på att träffa målet
fir	First Touch, hur bra spelaren är på att ta emot och ta kontroll på bollen
fla	Flair, hur ofta spelaren försöker genomföra oförutsedda saker med bollen
fre	Free Kick, hur bra spelaren är på att genomföra frisparkar
han	Handling, målvaktsegenskap
hea	Heading, hur bra spelaren är att vinna dueller i luften
jum	Jumping Reach, hur högt en spelare når i luften
kic	Kicking, målvaktsegenskap
ldr	Leadership, hur inflytelserik spelaren är på planen
lon	Long Shots, hur bra spelaren är på att träffa mål med skott utanför målområdet
l.th	Long Throws, hur bra spelaren är på att kasta långa inkast
mar	Marking, hur bra spelaren är på att markera och följa en motståndare med boll
nat	Natural Fitness, spelarens naturliga kondition (låst till 20)
otb	Off the Ball, förmåga att placera sig i rätt position utan boll i offensiva situationer
pac	Pace, spelarens maxfart

pas	Passing, spelarens förmåga att träffa rätt med en passning
pen	Penalty Kicks, hur bra spelaren är på att genomföra en straff
pos	Positioning, förmåga att placera sig i rätt position utan boll i defensiva situationer
pun	Tendency to Punch, målvaktsegenskap
ref	Reflexes, målvaktsegenskap
tro	Tendency to Rush Out, målvaktsegenskap
sta	Stamina, spelarens kondition (låst till 20)
str	Strength, spelarens förmåga att använda sin kropp mot motståndaren
tck	Tackling, hur bra spelaren är på att vinna bollen från en motståndare
tea	Team Work, hur bra spelaren är på att jobba tillsammans med lagkamrater
tec	Technique, hur bra spelaren är med bollen
thr	Throwing, målvaktsegenskap
vis	Vision, hur bra spelaren är på att se öppna lagkamrater för att slå passningar
wor	Work Rate, hur benägen spelaren är att genomföra sitt arbete till 100%

2.2 Tillgängliga och saknade värden

I datamaterialet förekommer ett mindre antal observationer med saknade värden i samtliga egenskapsvariabler. Totalt rör det sig om 24 av de 5011 observationerna, vilket motsvarar mindre än 0,5% av datan. Ett exempel visas i Tabell 2.2 nedan. Dessa fall innehåller information om spelare, klubb och motståndare, men saknar samtliga individuella attribut.

Tabell 2.2: Exempeltabell över målvakter och deras värden (NA = saknas)

average.rating	**name**	**acc**	**aer**	**agg**	**agi**	**ant**	**bal**	**bra**	**...**
8.08	Luke Laraque	NA	NA	NA	NA	NA	NA	NA	...
6.77	Tony Yeboah	NA	NA	NA	NA	NA	NA	NA	...
6.79	Elmis The Heretic	NA	NA	NA	NA	NA	NA	NA	...
7.29	Koschei Oakdown	NA	NA	NA	NA	NA	NA	NA	...
6.20	Jannik Andersen	NA	NA	NA	NA	NA	NA	NA	...
7.51	Scott Sterling	NA	NA	NA	NA	NA	NA	NA	...
6.61	Radek Soboda	NA	NA	NA	NA	NA	NA	NA	...
7.40	Adam Rage	NA	NA	NA	NA	NA	NA	NA	...
6.12	Certified Problem	NA	NA	NA	NA	NA	NA	NA	...
...	...	...	...	...	...	...	...	...	...

2.3 Hantering av ogiltiga värden

Observationer där attributvärden understiger det lägsta tillåtna värdet i SSL, det vill säga värden under 5, betraktas som ogiltiga eftersom de strider mot ligans definierade regler för hur spelare får konstrueras. I tabell 2.3 kan man exempelvis se att för målvakten Jannik Andersen från Athênai F.C när han möter LON, så finns det 9 värden som är 0 och inga värden som är 2, 3 eller 4. Dessa värden kan inte ha uppstått genom regelrätt användning av det attributsystem som ligger till grund för spelardata och bedöms därför vara ett resultat av fel vid datagenerering eller inläsning. I kvantitativ metodologi är det grundläggande att säkerställa att de data som ligger till grund för analysen är giltiga och tillförlitliga. Som en del av datarensningsprocessen bör omöjliga eller otillåtna värden aldrig lämnas orörda. I de fall där en korrekt värde kan identifieras bör det ersätta det felaktiga, men om detta inte är möjligt bör observationen istället tas bort. Van den Broeck et al. (2005) formulerar detta tydligt: ”Impossible values are never left unchanged, but should be corrected if a correct value can be found; otherwise they should be deleted.” Eftersom det i detta fall inte finns något tillförlitligt sätt att återställa de felaktiga värdena, har dessa observationer exkluderats från analysen. Detta utgör en metodologiskt försvarbar åtgärd som minskar risken för snedvridna resultat och säkerställer att analysen vilar på en valid databas.

Tabell 2.3: Frekvenstabell av värdena som mindre än 5 i varje målvakt.

Name	Club	Opponent	1	2	3	4
Jannik Andersen	Athênai F.C.	LON	9	0	0	0
Tony Yeboah Inter	London	ATH	9	0	0	0
Arcueid Brunestud	FC Rio	HOL	4	2	3	0
Scott Sterling	Hollywood FC	RIO	9	0	0	0
Pingu Nootazuki	Tokyo S.C.	CAI	3	5	1	0
Scott Sterling	Hollywood FC	CAI	9	0	0	0
Arcueid Brunestud	FC Rio	ATH	4	3	2	0
Jannik Andersen	Athênai F.C.	RIO	9	0	0	0
Pingu Nootazuki	Tokyo S.C.	LON	3	5	1	0
Tony Yeboah	Inter London	TOK	9	0	0	0
...	...	...	...	...	...	...

2.4 Beskrivande statistik

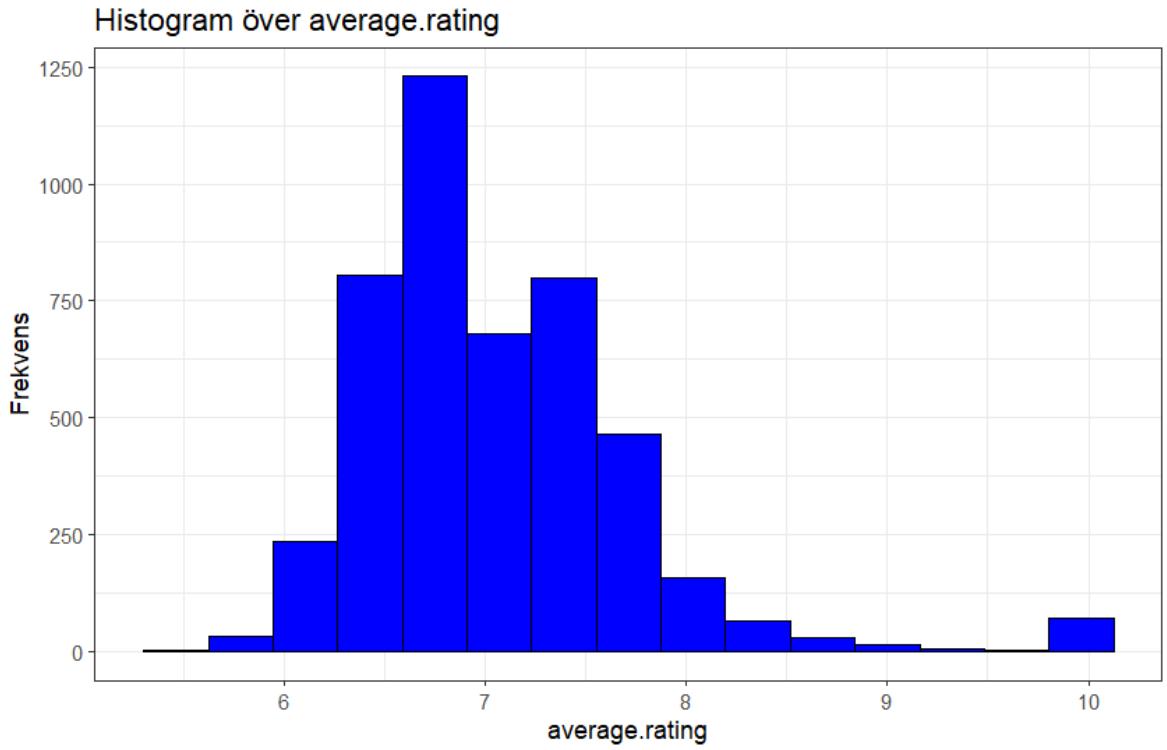
I kapitlet om beskrivande statistik presenteras en sammanfattande översikt av de variabler som ingår i undersökningen. Syftet är att ge en övergripande förståelse för datamaterialet. Tabellen nedan visar medelvärde, standardavvikelse samt minsta och största observerade värde för respektive variabel.

Tabell 2.4: Beskrivande statistik för samtliga numeriska variabler

Variabel	**Medelvärde**	**StdAvvikelse**	**Minimum**	**Maximum**
average.rating	7.051	0.665	5.5	10
acc	9.830	3.688	5.0	18
aer	15.112	2.457	5.0	20
agg	5.986	1.371	5.0	14
agi	14.408	2.854	5.0	20
ant	13.503	2.693	5.0	20
bal	9.344	3.181	5.0	18
bra	9.032	3.296	5.0	16
cmd	14.390	2.914	5.0	20
com	13.242	2.633	5.0	20
cmp	11.874	3.189	5.0	18
cnt	12.630	2.460	5.0	18
dec	13.914	2.967	5.0	20
det	7.930	3.038	5.0	16
ecc	7.356	3.269	5.0	16
fir	8.822	2.809	5.0	15
fla	5.456	1.365	5.0	12
fre	5.486	1.149	5.0	12
han	14.046	2.786	5.0	20
jum	8.059	3.465	5.0	17
kic	11.809	2.697	5.0	18
ldr	6.809	2.562	5.0	15
otb	5.418	1.186	5.0	10
pac	7.527	2.661	5.0	15
pas	11.181	3.779	5.0	20
pen	5.817	2.016	5.0	13
pos	13.284	3.361	5.0	20
pun	7.845	2.840	5.0	16
ref	14.507	5.107	5.0	20
tro	10.491	3.750	5.0	18
str	8.468	3.086	5.0	17
tea	7.071	2.091	5.0	15
tec	6.868	2.544	5.0	13
thr	10.003	2.779	5.0	16
vis	8.543	2.620	5.0	15
wor	6.756	2.457	5.0	15

2.5 Explorativ dataanalys

I figur 2.1 visar histogrammet fördelningen av responsvariabeln average.rating. Fördelningen är något högerskev, med majoriteten av observationerna mellan 6 och 8. Det finns även ett fåtal höga värden runt 10, vilket tyder på förekomsten av vissa spelare med exceptionellt höga betyg.



Figur 2.1: Histogram över medelmatchbetyg

3. Metod

I denna kapitlet ges en beskrivning av våra statistiska metoder som vi kommer att använda i denna undersökning.

3.1 Linjära mixade modeller

Linjära mixade modeller (LMM) lämpar sig särskilt väl i situationer där observationerna är korrelerade, vilket ofta är fallet vid klustrade data eller upprepade mätningar på samma objekt. Exempel på detta kan vara när flera individer tillhör samma grupp, såsom elever i samma klass eller när samma individ mäts vid flera tillfällen (Statistikakademin, u.å). När upprepade mätningar sker över tid kallas studien för en longitudinell studie (Kleinbaum, Kupper, Nizam Rosenberg, 2014). Mätningar som kommer från samma källa tenderar att vara statistiskt beroende av varandra, vilket innebär att analysmetoderna måste kunna hantera dessa inbördes samband.

En linjär mixad modell bygger på antagandet att vissa regressionsparametrar kan variera slumpmässigt mellan individer. På så vis kan man ta hänsyn till individuell variation över tid, vilket är särskilt relevant i longitudinella studier där samma individer observeras upprepade gånger. Modellen består av två huvudsakliga komponenter: fixa effekter som beskriver effekter av kovariater som är gemensamma för hela populationen, samt slumpmässiga effekter som fångar upp individunika avvikelser från dessa effekter. Den genomsnittliga responsen modelleras därmed som en kombination av populationsgemensamma egenskaper och individunika variationer. Termen ”mixad” syftar på att modellen innehåller både fixa och slumpmässiga effekter (Fitzmaurice, Laird Ware, 2011).

Den linjära mixade modellen kan uttryckas i subjektspecifik skalär form enligt:

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij1} + \cdots + \beta_s X_{ijs}) + (b_0 + b_1 Z_{ij1} + \cdots + b_q Z_{ijq}) + E_{ij} \tag{3.1}$$

där:

· i=1, 2,. . . , K representerar individ i, med totalt K individer i analysen.
· j=1, 2,. . . , n_i anger observation j för individ i, där varje individ har ni observationer.
· Y_{ij} betecknar den j:te responsen för individ i.
· X_{ijg} representerar värdet av prediktorn X_g, där g=1, 2,. . . , s för den j:te responsen av individ i som ingår som fixa effekter med tillhörande β-koefficienter som skattar den genomsnittliga effekten av varje prediktor på responsvariabeln.
· Z_{ijh} anger värdet av prediktorn Z_h, där h=1, 2,. . . , q för den j:te responsen av individ i som ingår som slumpmässiga effekter med tillhörande β-koefficienter som modellerar variation mellan individer.

· E_{ij} betecknar feltermen för den j:te responsen av individ i.

I formeln har de fixa β-effekterna inte index i eller j för att dessa parametrar representerar populationsparametrar, medan de slumpmässiga b-effekterna bara har index i eftersom de representerar subjektspecifika slumpmässiga effekter som kan variera mellan individer. På så sätt, i den linjära mixade modellen finns det en blandning av fixa effekter (β-parametrar), slumpmässiga effekter (b-koefficienter) och feltermen (E_{ij}). Här antas de n_i responserna från samma individ (den i:te) vara korrelerade, vilket innebär att det klassiska linjära modellantagandet om oberoende inte gäller (Kleinbaum, Kupper, Nizam Rosenberg, 2014).

3.1.1 Fixa och slumpmässiga effekter i linjära mixade modellen

I en linjär mixad modell bör alla möjliga effekter som kan tänkas påverka responsvariabeln beaktas. En fix effekt definieras som en term i en regressionsmodell som motsvarar en faktor vars nivåer är av intresse för hela populationen och som antas vara konstanta över populationen. En sådan effekt betraktas som en populationsparameter och betecknas vanligen med grekiska bokstäver, såsom α,β eller γ. Däremot definieras en slumpmässig effekt som en slumpmässig variabel som ingår i modellen för att fånga effekten av naturlig heterogenitet mellan individer på förutsägelsen av en responsvariabel. En sådan effekt betecknas vanligen med latinska bokstäver, såsom a,b eller g (Kleinbaum, Kupper, Nizam Rosenberg, 2014).

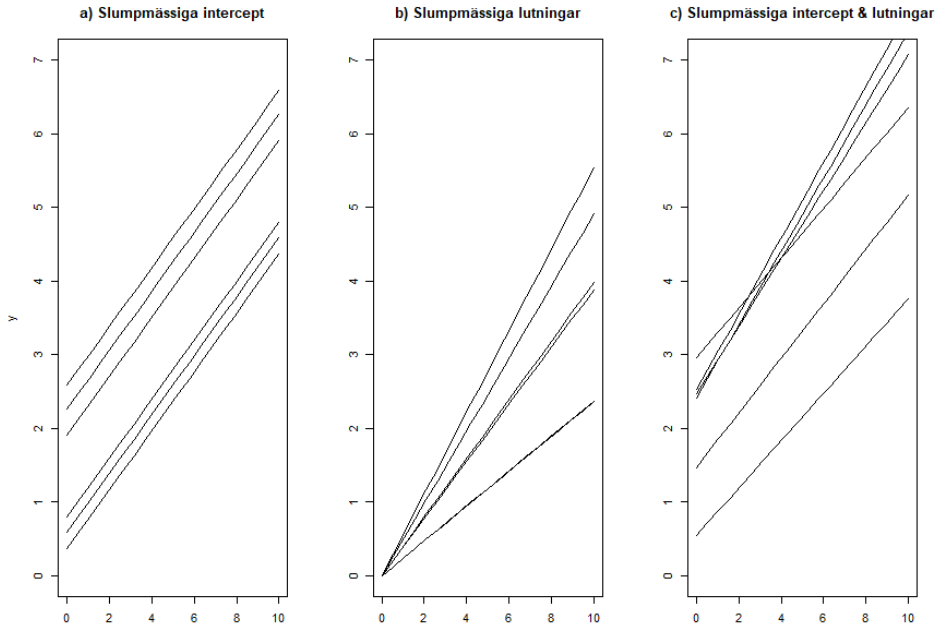
Exempel på två vanliga modeller med slumpmässiga effekter är följande (Kleinbaum, Kupper, Nizam Rosenberg, 2014):

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij1}) + b_{i0} + E_{ij} \tag{3.2}$$

Där b_{i0} representerar individunika avvikelser från det globala interceptet och

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij1}) + (b_{i0} + b_{i1} Z_{ij1}) + E_{ij} \tag{3.3}$$

Där b_{i1} anger individunika avvikelser i lutning beroende på kovariaten Z_{ij1}.



Figur 3.1: Illustration av olika typer av slumpmässiga effekter i linjära mixade modeller.

I det första fallet (a) varierar interceptet mellan individer medan lutningen är densamma, vilket resulterar i parallella linjer och detta är en modell med slumpmässiga intercept. I det andra fallet (b) är interceptet gemensamt för alla individer, men lutningen varierar, vilket innebär att linjerna har olika branthet. Detta benämns som en modell med slumpmässiga lutningar. I det tredje fallet (c) varierar både intercept och lutning mellan individer, vilket leder till att linjerna skiljer sig både i höjd och riktning och detta blir en modell med både slumpmässiga intercept och slumpmässiga lutningar.

3.1.2 Kovariansstruktur

En central utgångspunkt vid analys av upprepade mätningar är att observationer inom samma individ tenderar att vara korrelerade över tid. Denna korrelation eller mer generellt, kovariansen mellan mätningar behöver modelleras korrekt för att ge tillförlitliga slutsatser om regressionsparametrar. Genom att specificera en lämplig kovariansstruktur kan man förbättra skattningens precision och erhålla mer korrekta standardfel, särskilt när data innehåller saknade värden (Fitzmaurice et al., 2011).

För att förstå hur kovariansstruktur representeras i analysen användes en linjär mixad modell, formulerad enligt Fitzmaurice, Laird och Ware (2011) som $Y_i = X_i + Z_i b_i + _i$. Här är X_i och Z_i designmatriser för fixa respektive slumpmässiga effekter. β är en vektor av fixa regressionskoefficienter och b_i är en vektor av individunika slumpmässiga effekter. De slumpmässiga effekterna b_i antas vara oberoende av både kovariater och residualer. De modelleras vanligen som multivariata normalfördelade med väntevärde 0 och kovariansmatris G, dvs. $b_i \sim N(0, G)$. Residualerna antas även följa en multivariat

normalfördelning med väntevärde 0 och kovariansmatris R_i, vilket skrivs som: $\epsilon_i \sim N(0, Ri)$. I många tillämpningar förutsätts att residualerna är okorrelerade och har konstant varians. Under detta antagande är R_i en diagonalmatris med lika stora diagonalelement, dvs. $R_i = \sigma^2 I_{n_i}$ (Fitzmaurice et al., 2011).

För att uttrycka den totala variansen i observationerna behöver man ta hänsyn till både variansen från de slumpmässiga effekterna och från residualerna. Detta leder till den marginala kovariansstrukturen: $Cov(Y_i) = Cov(Z_i b_i) + Cov(\varepsilon_i) = Z_i G Z_i' + R_i$ Här representerar G kovariansmatrisen för de slumpmässiga effekterna, medan R_i beskriver inomindividuell variation. Matrisen G innehåller varians och samvarians mellan individunika regressionsparametrar, t.ex. slumpmässiga intercept och lutningar och är gemensam för alla individer. Denna struktur gör det möjligt att modellera både inom- och mellanindividuell variation korrekt (Fitzmaurice et al., 2011).

Observera att vid modellering med linjära mixade modeller antas vanligen att mätningar från olika individer är oberoende, medan mätningar inom samma individ kan vara korrelerade. Konvariansstrukturen inom individ specificeras genom residualernas kovariansmatris R, som beskriver hur observationer från samma individ samvarierar. Denna matris är alltid symmetrisk. Några vanliga former av kovariansstrukturer för R är exempels Compound Symmetry (CS), Unstructured (UN) och First-order Autoregressive (AR(1)) (Kleinbaum et al., 2014).

3.1.3 Maximum likelihood i linjära mixade modeller

För att använda Maximum Likelihood (ML) vid estimering i linjära mixade modeller krävs att fullständiga fördelningsantaganden görs för de uppmätta responsvektorerna. Fitzmaurice, Laird och Ware (2011) förklarar att när varje individs responsvektor, Y_i antas följa en multivariat normalfördelning bestäms hela sannolikhetsfördelningen av medelvärdesvektorn och kovariansmatrisen. Dessa antaganden möjliggör formuleringen av en sannolikhetsfunktion som kan användas för skattning av modellens parametrar.

I modellen specificeras det marginala väntevärdet som $E(Y_i) = X_i\beta$ och kovariansstrukturen som $Cov(Y_i) = \sum_i = \sum_i(\theta)$ (Fitzmaurice et al., 2011), där:

$\cdot X_i$ är designmatrisen för de fixa effekterna,
$\cdot \beta$ är en parametervektor för de fixa effekterna,
$\cdot \sum_i \theta$ är kovariansmatrisen som beror på en parametervektor θ, vilken styr variansen och samvariansen inom individ.

Under dessa antaganden härleds log-likelihoodfunktionen för hela urvalet enligt:

$$l = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta)\right\} \tag{3.4}$$

Där

· N är antalet individer,
· y_i är individens observerade responsvektor,
· $K = \sum_{i=1}^{N} n_i$ är det totala antalet observationer,
· $|\sum_i|$ är determinanten av kovariansmatrisen $\sum_i$

Skattningarna av β och θ erhålls genom numerisk maximering av denna log-likelihoodfunktion. Resultatet blir Maximum Likelihood-estimat som beskriver både medelvärdesstruktur och kovariansstruktur för modellen. ML används även i modelljämförelser, exempelvis genom likelihoodkvottester. Samtidigt påpekar Fitzmaurice et al. (2011) att ML tenderar att underskatta varianskomponenter vid små stickprov, vilket gör Restricted Maximum Likelihood (REML) till ett mer lämpligt alternativ i sådana fall.

3.1.4 Residualsanalys

Efter att en regressionsmodell har anpassats till data är det viktigt att kontrollera om modellens antaganden är uppfyllda. En central metod för detta är residualanalys. Den syftar till att bedöma om modellen är lämplig genom att analysera skillnaden mellan de observerade värdena och de värden som modellen predikterar. Dessa skillnader kallas för residualer.

Residualerna kan definieras som

$$e_i = y_j - \hat{y}_j \tag{3.5}$$

Där:

· e_i är skillnaden mellan observerade värdena och värdena som modellen förutspår (predikterade värden).
· y_i är det observerade värdet för observation i.
· $\hat{y}_i$ är predikterat värdena.

Genom att studera residualerna kan man kontroleera flera viktiga antaganden för modellen, såsom: Linearitet, dvs. om sambandet mellan variablerna är linjärt, homoskedasticitet, dvs. om residualerna har konstant varians, normalfördelning, dvs. om residualerna är ungefär normalfördelade och oberoende, dvs. om residualerna är oberoende av varandra. Om något av dessa antaganden är kraftigt avvikande kan det indikera att modellen är olämplig eller behöver justeras (Singull, 2018).

3.2 Statistiska mått

I detta avsnitt presenteras alla statistiska mått och beräkningar som används i undersökningen för att identifiera den mest lämpliga modellen för att besvara frågeställningen.

3.2.1 Förklaringsgrad

För att kolla om har modellen efter anpassat förmåga för att förklara data i verkligen, en vanligaste sätt är att skatta förklaringsgrad R2.

Förklaringsgrad ett centralt mått för att bedöma modellens anpassning till data eller kolla om hur mycket av variationen i responsvariabel (Y) som kan förklaras av de förklarande variabler (X) i en modell. Förklaringsgrad värder sträcker sig över från 0 till 1 och ju höger förklaringsgrads värdena, desto större modellensförmåga för att förklara observation och tvärtom om dessa värden är för liten, vilket indikera att modellens förklaringsgrad förmåga är svagt och modellen kan vara underanpassa eller det finns nästan ingen signifikant effekt från förklarasvariabler och responsvariabeln. I en linjär mixad modell bör man undersöka två olika mått på förklaringsgrad: den marginal förklaringsgraden, som anger den andel varians som förklaras enbart av de fixa effekterna, och den konditional förklaringsgraden, som anger den andel varians som förklaras av både fixa och slumpmässiga effekter (Carcagno, 2018).

Marginal förklaringsgrad kan identifieras som

$$R^2_{(m)} = \frac{\sigma^2_f}{\sigma^2_f + \sigma^2_r + \sigma^2_e} \tag{3.6}$$

Konditionellt förklaringsgrad kan identifieras som

$$R^2_{(c)} = \frac{\sigma^2_f + \sigma^2_r}{\sigma^2_f + \sigma^2_r + \sigma^2_e} \tag{3.7}$$

Där:

· $R^2_{(m)}$: Marginal förklaringsgrad
· $R^2_{(c)}$: Konditionellt förklaringsgrad
· σ^2_f: Detta innebär att man beräknar variansen av de värden som skulle förutsägas av de fixa effekterna ensamma, dvs. modellen utan slumpmässiga intercept och residual.
· σ^2_r: Varians mellan grupper från slumpmässiga effekterna
· σ^2_e: Residualvarians eller varians inom varje grupp

Till skillnad från en vanlig linjär regression innehåller en linjär mixed modell flera källor till variation i responsvariabeln. Den totala variationen kan delas upp i tre huvudsakliga varianskomponenter: en som förklaras av när modellen finns enbart fixa effekter i modellen, en som förklaras av de slumpmässiga effekterna (t.ex. skillnader mellan grupper), och en som kommer från residualvariansen (variation inom grupper)(Carcagno, 2018).

3.2.2 AIC

AIC (Akaike Information Criterion) är mått som balanserar modellens passform mot dess komplexitet. Den används för att jämföra olika statistiska modeller och hitta den lämpligaste modell. Med lägre AIC värder indikera bättre eller lämpligare modell(Isabella Roos, Leonard Persson Norblad,2021).

$$AIC = 2k - 2\ln(L) \tag{3.8}$$

Där:

· k är antal parametrar i modellen
· L är maximala värdet på likelihood-funktionen (hur bra modellen passar data)

3.2.3 VIF och GVIF

VIF (Variance Inflation Factor) är ett mått som används för att identifiera multikollinearitet mellan förklarande variabler. Ett VIF-värde över 10 brukar indikera att det kan finnas ett problem med multikollinearitet, vilket kan påverka tolkningen av modellens koefficienter (DeRuiter, 2019).

$$VIF = \frac{1}{1 - R_i^2} \tag{3.9}$$

Där:
· R_i^2 är determinationskoefficienten från en regression av X_j mot alla andra förklarande variabler.

Och

$$GVIF^{\frac{1}{2 \cdot Df}} \tag{3.10}$$

Där:
· Df är antalet frihetsgrader av varibel.

3.3 Bortfallsmekanismer

Vid statistisk analys är saknade värden ett vanligt problem som kan påverka både resultatens tillförlitlighet och statistisk styrka. Beroende på hur bortfallet uppstått t kan det delas in i tre huvudmekanismer (Schober Vetter, 2020):
· Missing Completely At Random (MCAR): Bortfallet är helt slumpmässigt och oberoende av alla andra variabler, inklusive den variabel som saknas. I detta fall påverkas inte resultaten systematiskt.
· Missing At Random (MAR): Bortfallet är systematiskt men kan förklaras av andra observerade variabler.
· Missing Not At Random (MNAR): Sannolikheten för att ett värde saknas beror på själva värdet som saknas, vilket gör hanteringen mer osäker.

3.3.1 Listwise deletion

Listwise deletion innebär att hela observationer med minst ett saknat värde utesluts från analysen. Denna metod är den vanligaste standardmetoden i många statistiska program. Den ger korrekta och obiaserade resultat endast om bortfallet är MCAR. Vid MAR eller MNAR kan däremot listwise deletion introducera bias (Schober Vetter, 2020).

Trots dessa begränsningar är listwise deletion ofta försvarbar om andelen bortfall är liten. Enligt Schober och Vetter (2020) anses en bortfallsnivå under 5% i regel vara trivial, vilket innebär att bortfallet inte påverkar analysresultaten i någon större utsträckning. Bennett (2001) menar att även bortfall upp till 10% kan vara acceptabelt, särskilt om bortfallet inte är systematiskt snedfördelat och datamaterialet i övrigt är omfattande.

Därför kan listwise deletion betraktas som en metodologiskt rimlig strategi i situationer där bortfallet är litet i omfattning, de bortfallna observationerna innehåller osäkra eller icke-tillförlitliga värden och analysen syftar till att bevara hög validitet snarare än maximal datavolym.

3.4 Implementation i R

Detta delkapitel beskriver hur den statistiska analysen har implementerats i R. Fokus ligger på hur data har hanterats, hur modeller har specificerats samt vilka paket och funktioner som använts i analysarbetet. Arbetet har genomförts i utvecklingsmiljön RStudio (version 2024.12.01).

3.4.1 Databearbetning och transformation

Databearbetning och transformation genomfördes med hjälp av paket dplyr, tidyverse, stringr och readr vilka användes för at filtrera, omkoda och förbereda variabler inför analys. Beskrivande statistik beräknades med grundläggande R-funktioner och presenteras i tabellform med hjälp av kableExtra.

3.4.2 Modellimplementering

För att anpassa de linjära mixade modellerna i studien användes funktionen lmer() från paketet lme4. Denna funktion möjliggör skattning av så kallade linear mixed-effects models (LMM), vilket är en generalisering av linjära modeller som inkluderar både fixa effekter (t.ex. individuella spelarprestationer) och slumpmässiga effekter (t.ex. variation mellan spelare). Syntaxen för lmer() bygger på en formelnotation där de fixa effekterna anges som vanliga regressionsled, medan de slumpmässiga effekterna specificeras med en vertikalstrecknotation, exempelvis (1|name) för att modellera ett slumpmässigt intercept för varje spelare.

En central aspekt av lmer() är att modellen inte enbart skattar regressionskoefficienterna, utan även den kovariansstruktur som motsvarar de slumpmässiga effekterna. Varians- och kovarianskomponenterna estimeras på en så kallad Cholesky-parametriserad form (theta-parametrar), vilket möjliggör numeriskt stabil och effektiv optimering (Bates et al., 2025). Denna skattning sker automatiskt utifrån modellformeln och används för att modellera beroendet mellan observationer inom samma grupp. Skattningen sker genom numerisk optimering av modellens likelihoodfunktion, där algoritmer som BOBYQA eller Nelder-Mead används, beroende på modellens komplexitet och angivna inställningar via lmerControl() (Bates et al., 2025). I denna studie användes maximum likelihood (dvs. REML =

FALSE) vid modelljämförelser, eftersom AIC inte är jämförbart mellan modeller skattade med olika metoder. För att säkerställa konvergens specificerades även kontrollparametrar med ökat antal tillåtna iterationer.

3.4.3 Modellurval och jämförelser

För att genomföra modellurval användes funktionen step() från paketet lmerTest, som implementerar en tvåstegs bakåtseliminering i linjära mixade modeller (lmerModLmerTest). I ett första steg utvärderas den slumpmässiga effektsstrukturen, där icke-signifikanta slumpmässiga termer tas bort. Därefter genomförs bakåtseliminering av de fixa effekterna, baserat på Akaike's Information Criterion (AIC) (Kuznetsova, Brockhoff, Christensen Jensen, 2022).

Vid varje steg rapporteras även p-värden för de fixa effekterna. Dessa beräknas med hjälp av t-tester där frihetsgraderna approximeras enligt Satterthwaite-metoden, vilket är standardinställning i både summary() och step() i lmerTest. Denna metod möjliggör inferens i modeller med komplexa slumpstrukturer där exakta frihetsgrader inte kan härledas (Kuznetsova et al., 2022).

Utöver den automatiska urvalsproceduren användes även manuella jämförelser mellan olika modeller genom att direkt jämföra deras AIC-värden. Lägre AIC indikerar en bättre balans mellan modellens komplexitet och förklaringskraft.

Slutligen beräknades modellernas förklaringsgrad med hjälp av funktionen r2() från paketet MuMIn, där marginal R^2 anger hur stor andel av variatRionen som förklaras av de fixa effekterna, medan konditionell R^2 inkluderar både fixa och slumpmässiga effekter. Dessa mått användes som ytterligare kriterier för att utvärdera modellens prediktiva kapacitet.

3.4.4 Multikollinearitet

För att undersöka multikollinearitet mellan de fixa effekterna användes funktionen vif() från paketet car. Resultaten presenterades i tabellform med hjälp av kable() och kable$_s tyling() frn paketet kable Extra$.

3.4.5 Residualanalys

Residualdiagnostik genomfördes med hjälp av en egen funktion i R. Funktionerna residuals() och fitted() från paketet lme4 användes för att extrahera residualer och anpassade värden. Visualiseringarna skapades med hjälp av ggplot2, och plottarna kombinerades med cowplot.

4. Resultat

I detta kapitel presenteras resultaten från analysen, med syftet att besvara studiens frågeställning. Fokus ligger på att undersöka i vilken utsträckning målvakters individuella egenskaper kan förklara variationen i deras genomsnittliga matchbetyg, samt att identifiera vilka egenskaper som uppvisar ett signifikant samband när både fixa och slumpmässiga effekter beaktas i en linjär mixad modell. Kapitlet är upplagt så att resultaten tolkas löpande.

4.1 Datamaterialet

Det slutgiltiga analysmaterialet omfattar 4597 observationer fördelade på 51 unika målvakter. Ursprungligen innehöll datamaterialet 5011 observationer, men en mindre andel exkluderades i samband med datarensningen.

Totalt 24 observationer ($\simeq$0,48%) hade saknade värden i alla egenskapsvariabler och uteslöts därför från analysen. Som beskrivs i avsnitt 3.3.1 tillämpades listwise deletion för att hantera dessa fall. Enligt Schober och Vetter (2020) är denna metod mest ämplig vid bortfall som är Missing Completely At Random (MCAR), men även under andra bortfallsmekanismer, såsom Missing At Random (MAR) eller Missing Not At Random (MNAR) med en bortfallsnivå under 5%, då nivån i regel är för låg för att påverka analysens resultat i någon nämnvärd utsträckning. Det mycket begränsade bortfallet i detta fall bedöms därmed inte utgöra något problem för analysens validitet.

Utöver detta identifierades 390 observationer där någon egenskap hade ett ogiltigt värde, definierat som mindre än 5 på den interna skalan i spelet Football Manager. Dessa observationer bedömdes sakna tillräcklig informationskvalitet och togs därför bort före analysen. Efter dessa exkluderingar återstår ett omfattande och representativt material som ger god grund för statistisk modellering.

4.2 Fullständig modell

Tabell 4.1: ett avsnitt av modellens ingående variabler

modell	**variabler**
fullständig modell	Average.rating acc aer agg agi ant bal ...

Tabell 4.2: Skattningar av de fixa effekterna i fullständig modellen

Variabel	**Estimate**	**Std. Error**	**df**	**t value**	**Pr(>|t|)**
(Intercept)	6.143e+00	1.113e-01	4597	46.804	$< 2e$-16 ***
acc	-1.426e-02	5.155e-03	4597	-2.767	0.005684 **
aer	1.649e-03	7.135e-03	4597	0.231	0.81724
agg	3.606e-02	7.230e-03	4597	4.984	6.4333e-07 ***
agi	2.399e-02	6.840e-03	4597	3.508	0.000454 ***
ant	1.136e-02	6.482e-03	4597	1.752	0.079806 .
bal	-2.356e-02	7.043e-03	4597	-3.344	0.000831 ***
bra	7.003e-03	6.496e-03	4597	1.078	0.28121
cmd	-6.223e-03	7.210e-03	4597	-0.863	0.38832
com	1.058e-02	7.610e-03	4597	1.390	0.16452
cmp	6.312e-03	7.482e-03	4597	0.843	0.39957
cnt	8.194e-03	6.602e-03	4597	1.241	0.21463
dec	7.859e-03	7.403e-03	4597	1.062	0.28851
...	...	...	...	...	...

Tabell 4.3: Skattade varianskomponenter för de slumpmässiga effekterna

Groups	**Name**	**Variance**	**Std.Dev.**
name	(Intercept)	0.0000	0.0000
Residual		0.4155	0.6446

Utifrån den beskrivande statistiken i avsnitt 2.3 har totalt 11 av de ursprungliga 46 variablerna uteslutits från den fortsatta modelleringen. Dessa variabler är: cor, cro, dri, fin, hea, lon, l.th, mar, nat, sta och tck. Sex av dessa variabler uppvisar en standardavvikelse på noll, vilket innebär att de saknar variation och därmed inte kan bidra till att förklara variationen i responsvariabeln. De övriga har mycket

låg standardavvikelse (mindre än 0,6), vilket tyder på begränsad variation och låg förklaringspotential.

Den fullständiga modellen inkluderar 35 förklaringsvariabler och ett slumpmässigt intercept för målvakter. Resultatet visar att flera av variablerna uppvisar statistiskt signifikanta samband med målvaktens genomsnittliga betyg, medan andra inte är signifikanta vid 5%-nivån. Exempelvis är variabler som agg, agi, cmp, cnt, ref och fre signifikanta, medan flera andra, såsom aer, kic, pas och tec, inte uppvisar någon tydlig effekt i modellen.

En varning om singular fit genererades vid modellskattningen. Detta återspeglas i att variansen för den slumpmässiga effekten är noll, vilket tyder på att skillnader mellan målvakter inte fångas upp av modellen efter att de fixa effekterna har inkluderats. En djupare diskussion kring dessa resultat och modellens tolkbarhet ges i nästkommande kapitel.

4.2.1 Kontroll för multikollinearitet

Tabell 4.4: Exempel på värden per variabel

Variabel	Värde
acc	3.9986
aer	3.4370
agg	3.1134
agi	3.7800
ant	3.8909
bal	2.0759
bra	4.8117
cmd	4.5709
com	4.6196
cmp	4.2739
cnt	4.0959
dec	5.0742
det	3.6151
ecc	2.8057
fir	3.7322
fla	2.5031
fre	3.9628
han	3.9621
jum	2.0667
kic	2.0732
ldr	2.3429
otb	2.3291
pac	2.8582
pas	7.0152
pen	4.2206
pos	4.7386
pun	2.2380
ref	3.9555
tro	5.0541
sta	3.6404
str	3.9549
tea	3.6434
tec	1.7610
thr	3.1610
vis	2.9237
wor	4.4514

För att undersöka förekomsten av multikollinearitet beräknades Variance Inflation Factor (VIF) för samtliga förklaringsvariabler. Samtliga variabler har VIF-värden under 10 som tumregeln, vilket tyder på att allvarlig multikollinearitet inte föreligger.

4.3 Modell med bakåtseliminering från lmerTest::step() i R

Tabell 4.5: Modellens ingående variabler

modell	**variabler**
Reducera Modell	Average.rating
	acc
	agg
	agi
	bra
	cmd
	cnt
	det
	ecc
	fir
	fre
	han
	otb
	pac
	ref
	name

Vid bakåtseliminering från lmerTest::step()-funktion reduceras den fullständiga modellen till en multipel linjär modell med 14 förklaringsvariabler utan slumpmässig effekt.

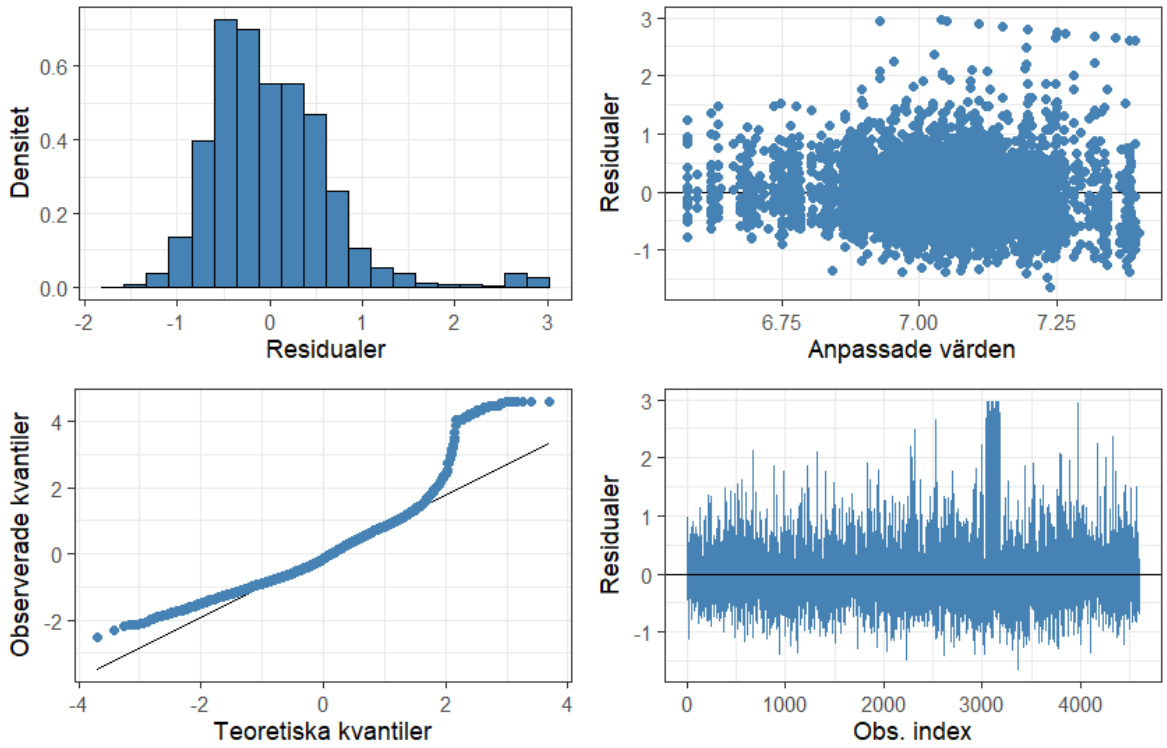
4.3.1 Linjär regression modell

Tabell 4.6: Koefficientskattningar från linjär modell

Variabel	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Signifikans**
(Intercept)	6.247418	0.105316	59.321	$< 2e$-16	***
acc	-0.013608	0.003945	-3.450	0.000567	***
agg	0.018238	0.006282	2.904	0.003698	**
agi	0.028276	0.005089	5.556	2.91e-08	***
bra	0.013444	0.004690	2.866	0.004170	**
cmd	0.026092	0.005872	4.444	9.06e-06	***
cnt	0.018798	0.005496	3.420	0.000631	***
det	-0.010100	0.004667	-2.164	0.030501	*
ecc	-0.012347	0.004019	-3.072	0.002137	**
fir	-0.014988	0.004685	-3.199	0.001388	**
fre	0.026300	0.008292	3.139	0.001704	**
han	-0.018967	0.005515	-3.439	0.000588	***
otb	-0.035658	0.010258	-3.476	0.000513	***
pac	0.018236	0.004367	4.176	3.02e-05	***
ref	0.011665	0.002100	5.555	2.94e-08	***

Denna modell har genererats genom bakåtseliminering med hjälp av funktionen step() från paketet lmerTest, som iterativt reducerar den fullständiga modellen baserat på AIC för att identifiera en förenklad modell med bibehållen förklaringskraft. Resultatet visar att flera individuella egenskaper uppvisar statistiskt signifikanta samband med det genomsnittliga matchbetyget. Variablerna acc, agg, agi, bra, cmd, cnt, det, ecc, fir, fre, han, otb, pac och ref är samtliga signifikanta på 5%-nivån eller lägre. De skattade koefficienterna anger riktningen på sambandet mellan respektive egenskap och responsvariabeln, där negativa värden innebär ett negativt samband och positiva värden ett positivt. Exempelvis för varje enhetsökning av variabeln ref (reflexes), så ökar genomsnittligt matchbetyg med cirka 0.012.

4.3.1.1 Residualanalys



Figur 4.1: Kombination för residualanalys

Histogrammet över residualer visar att dessa är centrerade kring noll, men fördelningen uppvisar en viss högerskevhet. Det förekommer flera positiva residualer med höga värden, vilket skapar en förlängd svans åt höger. Detta tyder på en mindre avvikelse från normalitet, även om huvuddelen av residualerna är symmetriskt fördelade kring medelvärdet.

Spridningsdiagrammet mellan residualer och anpassade värden uppvisar ingen tydlig systematisk struktur, vilket är önskvärt. Dock kan man ana något större variation i residualerna vid vissa nivåer av de anpassade värdena, särskilt i intervallet mellan 7.0 och 7.25. Det finns också ett fåtal extremvärden. Sammantaget tyder detta på att antagandet om homogen varians (homoskedasticitet) är rimligt uppfyllt, även om viss spridningsvariation förekommer.

Q-Q-plottens punkter följer den teoretiska normalfördelningen relativt väl i mitten, men visar tydliga avvikelser i båda svansarna, särskilt i den övre. Detta tyder på att residualerna inte är helt normalfördelade, främst på grund av ett antal outliers i datan.

Plottens fjärde ruta visar residualer mot observationsindex. Residualerna verkar vara slumpmässigt spridda utan tydliga mönster eller systematiska svängningar över tid. Detta ger stöd för antagandet om oberoende mellan observationerna.

4.3.2 Modell med bevarad slumpmässig effekt utifrån studiens syfte

Tabell 4.7: Resultat från linjär mixad modell: skattningar av slumpmässiga och fixa effekter

Random effects:

Groups	Name	Variance	Std.Dev.
name	(Intercept)	0.001672	0.04089
Residual		0.416256	0.64518

Fixed effects:

Variabel	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.268154	0.112705	178.958030	55.616	$< 2e$-16 ***
acc	-0.012955	0.004708	36.753766	-2.752	0.009152 **
agg	0.021840	0.007479	34.623772	2.920	0.006113 **
agi	0.028041	0.005460	173.510952	5.135	7.51e-07 ***
bra	0.012424	0.005363	58.531445	2.317	0.024046 *
cmd	0.025936	0.006249	214.460623	4.150	4.79e-05 ***
cnt	0.018925	0.006093	87.019850	3.106	0.002562 **
det	-0.010158	0.005276	67.191849	-1.925	0.058431 .
ecc	-0.011454	0.004559	66.599410	-2.513	0.014412 *
fir	-0.017301	0.005209	83.738281	-3.322	0.001328 **
fre	0.025177	0.009499	46.386210	2.650	0.010959 *
han	-0.018636	0.005309	158.400304	-3.143	0.002000 **
otb	-0.040102	0.011833	43.253999	-3.389	0.001500 **
pac	0.019657	0.004929	86.225496	3.988	0.000139 ***
ref	0.011008	0.002258	164.783764	4.875	2.54e-06 ***

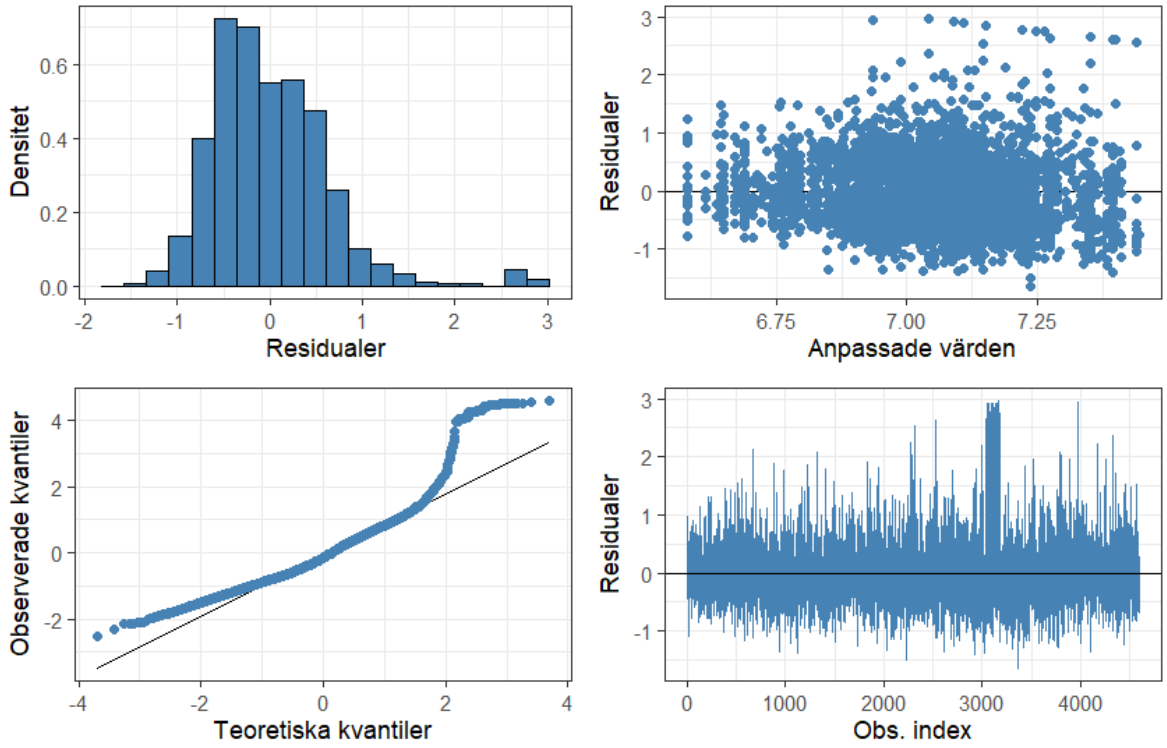
Den skattade modellen bygger på en linjär mixad modell där ett urval av individuella målvaktsegenskaper har inkluderats som fixa effekter, samt ett slumpmässigt intercept för varje målvakt (grupperade efter name) för att beakta upprepade mätningar per individ. Modellen är ett resultat av förenkling från den fullständiga modellen, där endast de mest relevanta prediktorerna har behållits.

Resultatet visar att flera spelaregenskaper har ett statistiskt signifikant samband med målvaktens genomsnittliga matchbetyg. Bland dessa återfinns exempelvis acc (acceleration), agg (aggressivitet), agi (agility), cmd (command of area), cnt (concentration), ecc (eccentricity), fir (first touch), fre (free kick), han (handling), otb (off the ball), pac (pace) och ref (reflexes). Dessa egenskaper kan därmed tolkas som viktiga faktorer för att förklara variationen i målvakters prestation enligt matchbetyg. Några variabler, såsom det (determination), ligger nära gränsen till signifikans.

Modellen inkluderar ett slumpmässigt intercept för varje målvakt, vilket tillåter individuell variation kring det övergripande medelvärdet. Variansen för denna slumpmässiga effekt är skattad till 0.00167, med en standardavvikelse på 0.0409. Detta indikerar att det finns viss men små mellanindividuell variation i genomsnittligt matchbetyg även efter att de fixerade egenskaperna har beaktats.

Residualvariansen uppgår till 0.4163, vilket innebär att en stor del av variationen i matchbetyg fortfarande återstår att förklara. Totalt ingår 4597 observationer från 51 målvakter i modellen.

4.3.2.1 Residualanalys



Figur 4.2: Kombination för residualanalys

Residualanalysen för denna modell uppvisar ett mönster som i stort överensstämmer med den tidigare modellen. Fördelningen är liknande, med svag skevhet och vissa outliers, men utan tydliga mönster som bryter mot modellens antaganden

4.4 Modelljämförelse

I detta delkapitel används olika statistiska mått för att jämföra de valda modellerna.

4.4.1 AIC för modeller

Tabell 4.8: Jämförelse av AIC-värden för två reducera modeller

Modell	**AIC-värdet**
Linjär regression modell	9065.121
Linjär mixed effekt modell	9064.984

För att jämföra modellernas lämplighet användes Akaike's Information Criterion (AIC), där ett lägre värde indikerar en bättre balans mellan modellens förklaringsförmåga och komplexitet. Som framgår av resultaten har den linjära modellen ett AIC-värde på 9065.121, medan den linjära mixade modellen har ett något lägre AIC på 9064.984. Skillnaden mellan modellerna är dock mycket liten ($\simeq$0.14), vilket innebär att båda modellerna uppvisar likvärdig prestanda enligt detta kriterium. AIC användes därmed som ett objektivt mått vid modelljämförelsen.

4.4.2 Förklaringsgrad för modeller

Tabell 4.9: R^2 och justerat R^2 för linjär regression

Mått	**Värde**
R^2	0.056
Justerat R^2	0.053

För den linjära regressionsmodellen utan slumpmässiga effekter är det totala $\mathbf{R^2}$-värdet 0,056 och det justerade $\mathbf{R^2}$-värdet 0,053. Det innebär att modellen som helhet förklarar cirka 5,6% av variationen, men justerat för antalet förklaringsvariabler är förklaringsgraden något lägre. Det justerade värdet är direkt jämförbart med det marginella $\mathbf{R^2}$ från den mixade modellen.

Jämförelse mellan den linjära mixade modellen och den linjära regressionsmodellen visar att de uppvisar mycket liknande förklaringsgrad. Den mixade modellen har ett marginellt $\mathbf{R^2}$ på 0,058, medan den linjära modellen har ett justerat $\mathbf{R^2}$ på 0,053. Det konditionella $\mathbf{R^2}$-värdet för den mixade modellen är något högre (0,061), vilket återspeglar det tillskott som den slumpmässiga interceptkomponenten bidrar med. Skillnaden mellan modellerna är dock liten, både sett till förklaringsgrad och modellens totala anpassning till data.

Tabell 4.10: R^2-mått för linjär mixad modell

Mått	**Värde**
Marginalt R^2	0.058
Konditionellt R^2	0.061

För den linjära mixade modellen uppgår det marginella R^2-värdet till 0,058, vilket anger att de fixa effekterna ensamt förklarar cirka 5,8% av variationen i målvakters matchbetyg. Det konditionella R^2 är något högre (0,061), vilket inkluderar både fixa och slumpmässiga effekter. Skillnaden mellan dessa är liten (0,003), vilket antyder att variationen mellan målvakter (den slumpmässiga interceptkomponenten) endast har en marginell inverkan på modellens totala förklaringsgrad.

5. Diskussion

Vid datarensningen uteslöts observationer med saknade eller ogiltiga värden för att säkerställa att de statistiska analyserna kunde genomföras utan tekniska problem. De saknade värdena utgjorde endast 0,48% av hela datamängden, vilket är en mycket liten andel och därmed inte förväntas påverka resultaten i någon större utsträckning. Däremot togs 390 observationer bort på grund av ogiltiga värden, vilket ledde till att två målvakter uteslöts helt ur analysen.

Trots detta kvarstod ett omfattande datamaterial bestående av 4597 observationer och 51 målvakter, vilket ger en god grund för statistisk modellering. Det slutgiltiga urvalet bedöms därför vara tillräckligt stort och representativt för att dra meningsfulla slutsatser, även om det inte kan uteslutas att bortfallet påverkat resultaten i viss utsträckning. För framtida analyser kan mer avancerade metoder för hantering av saknad data, såsom multipel imputering, övervägas – särskilt vid större bortfall eller om mönster i bortfallet misstänks.

Med en rensad och balanserad datamängd kunde den fullständiga modellen estimeras utifrån ett brett urval av målvaktsegenskaper. Trots att variabler med låg variation uteslöts återstod 35 förklaringsvariabler. Vid modellskattningen uppstod dock en varning om singular fit, vilket visade sig i att variansen för den slumpmässiga interceptkomponenten blev noll. Det tyder på att modellen inte fångar någon ytterligare variation mellan målvakter utöver de fixa effekterna, vilket kan bero på överanpassning till följd av många förklaringsvariabler i förhållande till datamängden.

För att hantera problemet med överanpassning användes funktionen step() från paketet lmerTest, som genom bakåtseliminering föreslog en enklare modell baserad på AIC. Denna modell exkluderade flera icke-signifikanta variabler och resulterade i en förbättrad modellstruktur utan varningar om singularitet. Dock innebar modellurvalet att den slumpmässiga effekten uteslöts, vilket står i viss kontrast till studiens ursprungliga syfte att beakta variation mellan målvakter.

Vidare är det viktigt att reflektera över metodvalets lämplighet utifrån studiens syfte och datans struktur. Eftersom varje målvakt förekommer upprepade gånger i materialet är det statistiskt motiverat att använda en modell med mixade effekter. Även om den slumpmässiga variansen skattades nära noll i vissa modeller, ger den mixade modellen ändå möjlighet att beakta potentiell klustring och undvika underskattning av standardfel. Detta kan ses som en styrka, särskilt i sammanhang där individberoende observationer förekommer. Därför, trots att step()-funktionen föreslog en modell utan slumpmässiga effekter valdes det att även behålla en modell med slumpmässigt intercept för målvakter i den fortsatta analysen. Det är också möjligt att den svaga och otydliga effekten av den slumpmässiga interceptkomponenten beror på att relevanta förklaringsvariabler saknas i modellen. De individuella egenskaper som ingår fångar sannolikt endast en del av den variation som påverkar matchbetyget. Om kontextuella faktorer såsom matchstatistik (t.ex. antal räddningar eller insläppta mål) hade inkluderats, är det möjligt att variationen mellan målvakter blivit tydligare, vilket i sin tur hade kunnat öka betydelsen av den slumpmässiga effekten i en mixad modell. Detta pekar på att framtida analyser bör överväga att kombinera spelaregenskaper med matchspecifik information för att bättre förklara variationen i

betygssättning och även att fånga den slumpmässiga effekten mer tydligt.

Vid jämförelse mellan den linjära mixade modellen och den vanliga linjära regressionsmodellen framkom endast små skillnader i både modellernas informationskriterium och förklaringsgrad. AIC-värdet för den mixade modellen var något lägre (9064.984) än för den linjära modellen (9065.121), vilket antyder att båda modellerna uppvisar likvärdig anpassning till data. Även skillnaden i förklaringsgrad var marginell; den mixade modellen hade ett marginellt R^2 på 0,058 och ett konditionellt R^2 på 0,061, medan den linjära modellen uppvisade ett justerat R^2 på 0,053. Den lilla skillnaden mellan marginalt och konditionellt R^2 i den mixade modellen tyder på att variationen mellan målvakter bidrog endast i begränsad utsträckning till den totala variationen i matchbetyg.

Samtidigt visar den låga förklaringsgraden i samtliga modeller att en stor del av variationen i matchbetyg sannolikt beror på faktorer som inte fångas av de individuella egenskaperna i denna studie. Det är därför troligt att matchspecifik statistik, såsom antal räddningar, insläppta mål eller matchens svårighetsgrad spelar en avgörande roll för hur betygen sätts. Den relativt låga förklaringsgraden tyder på att individuella egenskaper i sig inte räcker för att förklara variationen i genomsnittligt matchbetyg. Detta stämmer väl överens med tidigare forskning av Ball, Huynh och Varley (2025), som visar att kommersiella betygssystem ofta influeras av både subjektiva bedömningar och kontextuella faktorer.

6. Slutsatser

Syftet med denna studie var att undersöka i vilken utsträckning både generella spelaregenskaper och målvaktsspecifika egenskaper kan förklara variationen i målvakters matchbetyg, samt att identifiera vilka av dessa egenskaper som uppvisar ett statistiskt signifikant samband när både fixa och slumpmässiga effekter beaktas.

Analysen baserades på en linjär mixad modell där 14 individuella spelaregenskaper inkluderades som fixa effekter och ett slumpmässigt intercept användes för att modellera variation mellan målvakter. Den skattade variansen för den slumpmässiga effekten var låg (ca 0.0017), vilket innebar att skillnader mellan målvakter bidrog endast marginellt till den totala variationen. Modellens marginala R^2 uppgick till 0.058 och det konditionella R^2 till 0.061, vilket visar att huvuddelen av variationen förklaras av de fixa effekterna i modellen.

Flera egenskaper uppvisade ett statistiskt signifikant samband med matchbetyget på 5%-nivån. Bland de variabler som hade en positiv effekt återfanns Aggression (agg), Agility (agi), Bravery (bra), Command of Area (cmd), Concentration (cnt), Free Kick (fre), Pace (pac) och Reflexes (ref). Negativa samband observerades för Acceleration (acc), Eccentricity (ecc), First Touch (fir), Handling (han) och Off the Ball (otb). Variabeln Determination (det) låg nära signifikansgränsen (p 0.058) och inkluderades i modellen med viss försiktighet. Resultatet visar att både generella spelaregenskaper och målvaktsspecifika egenskaper kan ha betydelse för variationen i matchbetyg, även om sambanden varierar i styrka och riktning. Dessa resultat visar att både generella spelaregenskaper (t.ex. agg, agi, bra, cnt, pac) och målvaktsspecifika egenskaper (t.ex. cmd, han, ref, ecc) kan ha betydelse för hur spelare bedöms i matchbetyg.

Slutsatsen är att vissa målvaktsegenskaper har ett signifikant samband med betygsättning, men att den totala variationen i betyg i hög grad påverkas av faktorer utanför de individuella variabler som ingick i modellen. Det tyder på att bedömning av målvakter i matchsammanhang sannolikt också påverkas av kontextuella, subjektiva eller matchrelaterade faktorer som inte fångats upp i analysen.

Litteraturförteckning

[1] Ball, D., Huynh, N., & Varley, M. C. (2025). Comparing player rating systems as a metric for assessing individual performance in soccer. *Journal of Sports Sciences, 43*(7), 676–686. https://doi.org/10.1080/02640414.2025.2471208

[2] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E., Jagan, M., & Boylan, R. D. (2025). *lme4: Linear mixed-effects models using Eigen and S4* (Version 1.1-35) [R package manual]. CRAN. https://cran.r-project.org/web/packages/lme4/lme4.pdf

[3] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2025). *lme4: Linear Mixed-Effects Models using Eigen and S4* (Version 1.1-37) [R package manual]. CRAN. https://cran.r-project.org/package=lme4

[4] Encyclopaedia Britannica. (2025). *Football (soccer)*. Hämtad 7 april 2025 från https://www.britannica.com/sports/football-soccer

[5] Canadice. (2023, 21 juni). *Simulation Soccer League – Sign Up* [Forum-inlägg]. Sports Interactive Community. Hämtad från https://community.sports-interactive.com/forums/topic/576227-simulation-soccer-league-sign-up/

[6] Carcagno, S. (n.d.). *R^2 for linear mixed models*. Hämtad 27 maj 2025 från https://samcarcagno.altervista.org/stat_notes/r2_lmm_jags/r_squared_lmm.html

[7] DeRuiter, S. (n.d.). *Collinearity and multicollinearity*. Hämtad 27 maj 2025 från https://stacyderuiter.github.io/s245-notes-bookdown/collinearity-and-multicollinearity.html

[8] Fife, D. (n.d.). *flexplot: Graphically based data analysis using linear models* [GitHub repository]. Hämtad 27 maj 2025 från https://github.com/dustinfife/flexplot

[9] Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2 uppl.). John Wiley & Sons.

[10] Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178–183. http://www.jstor.org/stable/2290467

[11] Sports Interactive. (2023, 28 november). *What is Football Manager 2024?* Hämtad 20 juni 2025 från
https://www.footballmanager.com/what-is-fm

[12] He, M., Cachucho, R., & Knobbe, A. (2015). Football players' performance and market value. I J. Davis, J. Van Haaren, & A. Zimmermann (Red.), *Machine learning and data mining for sports analytics: Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics* (s. 87–95). CEUR Workshop Proceedings, Vol. 1970.
https://ceur-ws.org/Vol-1970/paper-11.pdf

[13] Hugh, P. (2024, 15 november). *Player match ratings: The metrics behind soccer performance.* Hämtad 8 april 2025 från
https://soccerwizdom.com/2024/11/15/player-match-ratings-the-metrics-behind-soccer-performance

[14] Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). *Applied regression analysis and other multivariable methods* (5th ed.). Cengage Learning.

[15] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2022). *lmerTest: Tests in linear mixed effects models* (Version 3.1-3) [R-paketdokumentation]. CRAN. Hämtad 20 juni 2025 från
https://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf

[16] Schober, P., & Vetter, T. R. (2020, november). Missing data and imputation methods. *Anesthesia & Analgesia, 131*(5), 1419–1420.
https://doi.org/10.1213/ANE.0000000000005068

[17] Singull, M. (2018). *Regressionsanalys.* Linköpings universitet. Hämtad från
https://courses.mai.liu.se/GU/TAMS24/JT-2018/Dokument/regressionsanalys.pdf

[18] Soccer Coaching Pro. (2023, 20 februari). *Soccer positions, numbers, and roles (Full breakdown).* Hämtad 7 april 2025 från
https://www.soccercoachingpro.com/soccer-positions/

[19] Simulation Soccer League. (2023, 1 april). *Create-a-player guide* [Foruminlägg]. Hämtad 20 juni 2025 från
https://forum.simulationsoccer.com/showthread.php?tid=159

[20] Statistikakademin. (u.å.). *Mixade modeller i SPSS.* Hämtad 5 maj 2025 från
https://statistikakademin.se/vara-kurser/spss/mixade-modeller/

[21] Stuart, K. (2014, 12 augusti). *Why clubs are using Football Manager as a real-life scouting tool.* Hämtad 20 juni 2025 från
https://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool

[22] Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine, 2*(10), e267.
https://doi.org/10.1371/journal.pmed.0020267

Bilaga

```r
install.packages(c("dplyr","ggplot2","tidyverse","nnet","caret",
                   "readr","stringr","kableExtra","lme4","lmerTest",
                   "car","performance","devtools","MuMIn"))
library(dplyr)
library(ggplot2)
library(tidyverse)
library(nnet)
library(caret)
library(readr)
library(stringr)
library(kableExtra)
library(lme4)
library(lmerTest)
library(car)
library(MuMIn)
library(performance)



### Läsa in data set.
setwd("C:/Users/timce/OneDrive/Dokument/Rcode_2024_2025/732G56/datamaterial")
data = read.csv("utespelare.csv", sep=";")
data$average.rating <- as.numeric(gsub(",",".",data$average.rating))
only_GK <- data[data$position == "GK",]
only_GK$matchday <- str_extract(only_GK$matchday,"S\\d+")
only_GK$matchday <- as.numeric(str_remove(only_GK$matchday,"S"))
only_GK <- only_GK[,c("matchday","average.rating","name","club","opponent","acc","aer",
"agg","agi","ant","bal","bra","cmd","com","cmp","cnt","cor","cro","dec","det","dri",
"ecc","fin","fir","fla","fre","han","hea","jum","kic","ldr","lon","l.th","mar","nat"
,"otb","pac","pas","pen","pos","pun","ref","tro","sta","str","tck","tea","tec","thr",
"vis","wor")]

# NA värder
na_rows <- only_GK[rowSums(is.na(only_GK[,6:ncol(only_GK)])) == (ncol(only_GK) - 5),]
# filtrera bort alla observationer som innehålla na värder
only_GK <- only_GK[!(rowSums(is.na(only_GK[, 6:ncol(only_GK)])) == (ncol(only_GK) - 5)), ]
```

```r

# Identifiera numeriska kolumner
numeric_cols <- sapply(only_GK, is.numeric)
# Uteslut 'matchday' från de numeriska kolumnerna
numeric_cols["matchday"] <- FALSE
# Filtrera bort rader där något av de utvalda numeriska värdena är < 5
df_filtered <- only_GK[!apply(only_GK[ , numeric_cols] < 5, 1, any), ]
# Spara alla observationer som innehålla värdena som mindre än 5
df_removed <- only_GK[apply(only_GK[ , numeric_cols] < 5, 1, any), ]


beskrivning_data <- df_filtered[,-1]
# Välj bara numeriska kolumner
numeric_df <- beskrivning_data[sapply(beskrivning_data, is.numeric)]
numeric_df <- numeric_df %>%
  select(-c("cor", "cro", "dri", "fin", "hea", "lon", "l.th", "mar", "nat", "sta", "tck"))


# Skapa en data.frame med medelvärde, min och max för varje variabel
summary_stats <- data.frame(
  Medelvärde = sapply(numeric_df, mean, na.rm = TRUE),
  StdAvvikelse = sapply(numeric_df, sd,na.rm = TRUE),
  Minimum    = sapply(numeric_df, min, na.rm = TRUE),
  Maximum    = sapply(numeric_df, max, na.rm = TRUE)
)

# Visa resultatet
summary_stats %>%
  round(3)%>%
  kable()

# hist() för målvakters medelmatchbetyg
ggplot(data = df_filtered, aes(x = average.rating)) +
  geom_histogram(bins = 15, color = "black", fill = "blue") +
  labs(title = "Histogram över average.rating",
       x = "average.rating",
       y = "Frekvens") + theme_bw()

### fullständig modell
fullständig_model <- lmer(average.rating~acc+aer+agg+agi+ant+bal+bra+cmd+com+cmp+cnt+
                      dec+det+ecc+fir+fla+fre+han+jum+kic+
                      ldr+otb+pac+pas+pen+pos+pun+ref+tro+
                      str+tea+tec+thr+vis+wor+(1|name),
                    data=df_filtered, control = lmerControl(optCtrl = list(maxfun = 100000)),
                    REML = FALSE,
                    na.action = na.fail)
```

```r

r2(fullständig_model)
AIC(fullständig_model)
summary(fullständig_model)
# Beräkna GVIF och justerad GVIF värder
vif(fullständig_model) %>%
  as_tibble(rownames = NA) %>%
  rownames_to_column() %>%
  kable(
    digits = 4
  )

# båkateliminering
lmerTest::step(fullständig_model)
get_model(step(fullständig_model))

### Funktionen för att skapa en kombination av 4 olika plot för residualanalys
### Funktionen kräver endast ett argument, modellen som anpassats
residualPlots <- function(model) {
  residualData <-
    data.frame(
      residuals = residuals(model),
      # Responsvariabeln finns som första kolumn i modellens model-objekt

      yHat = fitted(model)
    )
  p1 <- ggplot(residualData) +
    aes(x = residuals, y = after_stat(density)) +
    geom_histogram(bins = 20, fill = "steelblue", color = "black") +
    theme_bw() +
    labs(x = "Residualer", y = "Densitet")
  p2 <- ggplot(residualData) +
    aes(x = yHat, y = residuals) +
    geom_hline(aes(yintercept = 0)) +
    geom_point(color = "steelblue") +
    theme_bw() +
    labs(x = "Anpassade värden", y = "Residualer")
  p3 <- ggplot(residualData) +
    # Använder standardiserade residualer
    aes(sample = scale(residuals)) +
    geom_qq_line() +
    geom_qq(color = "steelblue") +
    theme_bw() +
    labs(x= "Teoretiska kvantiler", y = "Observerade kvantiler")
  p4 <- ggplot(residualData) +
    aes(x = 1:nrow(residualData), y = residuals) +
```

```
    geom_line(color = "steelblue") +
    theme_bw() +
    labs(x = "Obs. index", y = "Residualer") +
    geom_hline(
      aes(yintercept = 0),
      color = "black")
  cowplot::plot_grid(p1, p2, p3,p4, nrow = 2)
}



### reducerad modell1
reducerad_model1 <- lm(average.rating ~ acc + agg + agi + bra + cmd + cnt + det +
                         ecc + fir + fre + han + otb + pac + ref,
                       data=df_filtered)

summary(reducerad_model1)
r2(reducerad_model1)
AIC(reducerad_model1)

# residual plot
residualPlots(reducerad_model1)



### reducerad modell2
reducerad_model2 <- lmer(average.rating ~ acc + agg + agi + bra + cmd + cnt + det +
                           ecc + fir + fre + han + otb + pac + ref + (1|name),
                         data=df_filtered,
                         REML = FALSE,
                         na.action = na.fail)

summary(reducerad_model2)
AIC(reducerad_model2)
r2(reducerad_model2)

# Residual plot
residualPlots(reducerad_model2)
```