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 undersökning syftar till att undersöka om och hur individuella egenskaper hos målvakter kan förklara variationen i deras genomsnittliga matchbetyg. Studien baseras på simulerad data från en virtuell fotbollsliga i spelet Football Manager 2022, där ett stort antal attribut för varje målvakt registrerats över flera säsonger. Då samma målvakt kan förekomma vid flera tillfällen över tid är datan longitudinell, vilket motiverar användningen av linjära mixade modeller för att hantera beroende mellan observationer.

Modellanpassningen genomfördes stegvis: först med en fullständig modell där samtliga tillgängliga attribut inkluderades, följt av en automatiserad reducering med hjälp av funktionen step() i R, och avslutningsvis en manuell reducering för att uppnå en tolkbar och parsimonisk slutmodell. Resultatet visar att ett flertal individuella egenskaper – däribland aggression, reflexer, snabbhet och kommunikation – har statistiskt signifikanta samband med matchbetyget. Dock är modellens totala förklaringsgrad låg (marginalt $R^2 = 0.053$), vilket innebär att en stor del av variationen i matchbetyg inte fångas av de inkluderade prediktorerna.

Den slumpmässiga effekten (intercept per målvakt) bidrog endast marginellt till modellens förklaringsgrad, vilket kan tyda på att upprepade mätningar inom samma målvakt inte är starkt korrelerade – trots att datan är longitudinell. Residualanalysen visade inga allvarliga avvikelser från modellens antaganden, men antydde viss skevhet och förekomst av outliers.

Sammanfattningsvis indikerar resultaten att även om vissa målvaktsegenskaper är betydelsefulla, så spelar sannolikt även matchspecifika och kontextuella faktorer en central roll i hur betygsättningen utformas. Detta stämmer överens med tidigare forskning som visar att subjektiva bedömningar ofta påverkar spelaromdömen.

Abstract

The aim of this study is to investigate to what extent individual goalkeeper attributes can explain variation in their average match ratings. The analysis is based on simulated longitudinal data from the game Football Manager 2022, where goalkeepers may appear repeatedly across multiple seasons. Due to the hierarchical structure of the data, a linear mixed-effects model with a random intercept for each goalkeeper was employed.

The model was developed in several steps. Initially, a full model including all available attributes as fixed effects was estimated. This was followed by stepwise model selection using the step() function from the lmerTest package in R, along with additional manual reduction. The results show that some attributes—such as aggression, communication, and reflexes—are significantly associated with match ratings. However, the model's overall explanatory power is low (marginal R^2 0.05), suggesting that most of the variation remains unexplained.

The random effect variance is small, indicating weak correlation between repeated observations within the same goalkeeper. This may suggest that match ratings are largely influenced by contextual or subjective factors, rather than by stable individual characteristics. These findings align with previous research indicating that player evaluations are often affected by match events and contextual circumstances.

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

1.1 Bakgrund

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 (Weil et al., 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 (Pierce Hugh, 2024).

En tidigare studie har visat att vissa prestationsmått har större påverkan på spelarnas betyg än andra. I studien undersöktes sambandet mellan över 70 olika prestationsmått och det tilldelade matchbetyget bland tre av de mest använda betygssättningssystemen, vilka är WhoScored, FotMob och Sofascore. Utifrån resultaten visar det att offensiva mått, såsom skott på mål, nyckelpassningar och lyckade dribb-

lingar hade starkast samband med högre betyg. På andra hand hade defensiva mått, såsom rensningar ett svagare samband men fortfarande signifikant. Dessutom visar studien även att det fanns systematiska skillnader mellan betygssättningssystemen, där WhoScored generellt ger lägre betyg än de andra (Ball et al., 2025). Trots att sådan forskning främst fokuserar på utespelare och matchstatistik som samtidigt mäts, väcker resultaten viktiga frågor om hur målvakter bedöms utifrån deras egenskaper. Eftersom målvakters prestationer ofta bygger på ett färre antal men mer avgörande händelser, är det särskilt intressant att undersöka om och hur deras egenskaper kan användas för att förklara eller prediktera deras matchbetyg.

1.2 Syfte

Syftet med denna undersökning är att analysera om och hur målvakters individuella egenskaper kan förklara variationen i deras matchbetyg. För att ta hänsyn till att samma målvakt förekommer i flera matcher används en modell med blandade effekter, vilket möjliggör att både fasta effekter, såsom målvaktens egenskaper och slumpmässiga effekter, såsom variation mellan målvakter, inkluderas i analysen.

1.2.1 Frågeställning

I vilken utsträckning kan målvakters individuella egenskaper förklara variationen i deras matchbetyg, och vilka egenskaper har ett signifikant samband när både fasta 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. Här ges också en kortfattad beskrivning av variabler och observationer i det givna datamaterialet samt av eventuella transformationer och skapande av nya variabler. Eventuellt kan tabeller och/eller diagram över data presenteras.

I denna studie används simulerad data från The Simulation Soccer League (SSL), en virtuell fotbollsliga där spelare skapas och utvecklas av användare. Data är genererad genom Football Manager 2022 och har ingen koppling till verkliga individer. Vid skapandet av en spelare i SSL används ett särskilt verktyg, ”Player Builder Tool”, där attribut poängsätts utifrån ett fast poängsystem. Två egenskaper, Natural Fitness och Stamina är alltid satta till 20, vilket är maximalt värde, medan övriga attribut tilldelas utgångsvärden som minimalt kan vara 5. Under karriären genom olika aktiviteter inom ligan kan dessa attribut utvecklas (Canadice, 2023).

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 är därmed beroende.

Det ursprungliga datamaterialet innehåller många variabler som kan delas in i två kategorier: matchspecifik statistik, exempelvis prestationer under enskilda matcher och individuella egenskaper, vilka är olika mätbara attribut som beskriver spelarens förmågor, såsom fysiska, tekniska eller mentala egenskaper. I denna undersökning fokuseras det enbart på målvakternas individuella egenskaper i enlighet med den formulerade frågeställningen.

2.1 Data beskrivning

Datamaterialet kan delas in i två typer av variabler: textvariabler och numeriska variabler. Textvariablerna innehåller information som målvaktens namn, laget och motståndarlaget och de numeriska variablerna är fler till antalet och kommer att presenteras i tabellen nedan.

Tabell 2.1: Skattningar, standardfel och p-värden från en analys av modell

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 Handling missing data

Under processen att samla in och generera data är det oundvikligt att vissa värden saknas. Det finns totalt 24 observationer med saknade värden i olika numeriska variabler. Dessa saknade värden kommer att ersättas med medelvärdet för respektive variabel som data tillhör.

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, kap. 8).

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).

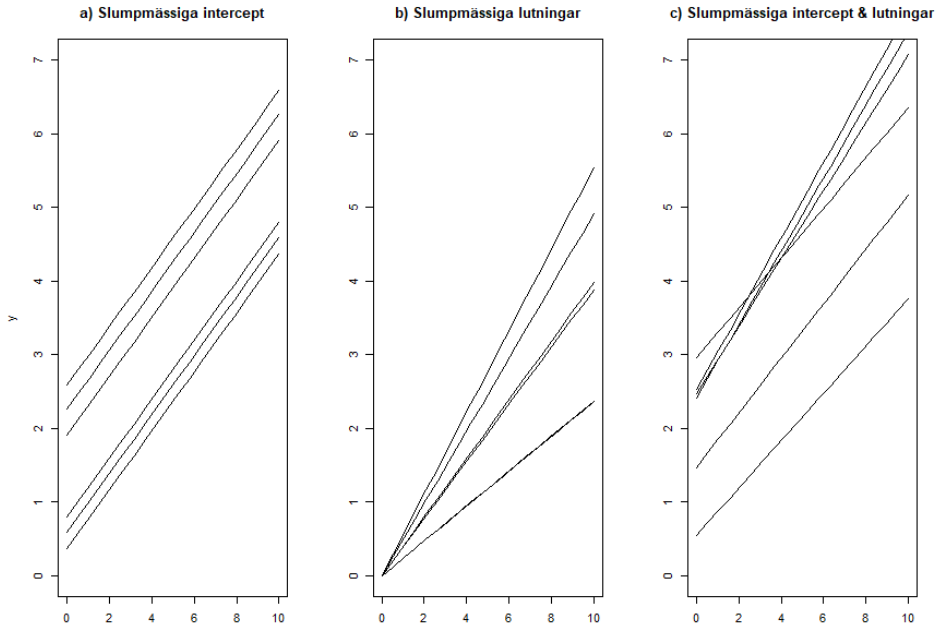
Exempel på två vanliga modeller med slumpmässiga effekter är följande (Kleinbaum, Kupper, Nizam Rosenberg, 2014):

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

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, Laird Ware, 2011, kap. 3.5).

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: $_i \sim N(0, Ri)$. 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, Laird Ware, 2011, s. 196-197). Vid modellering med linjära mixade modeller antas vanligen att mätningar från olika individer är oberoende, medan mätningar inom samma individ kan vara korrelerade. Kovariansstrukturen 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 exempelvis Compound Symmetry (CS), Unstructured (UN) och First-order Autoregressive (AR(1)) (Saarinen, 2004).

3.1.3 Residualsanalys

Efter att en modell har anpassats bör man alltid inleda med att undersöka dess lämplighet och kontrollera om förutsättningarna för modellen är uppfyllda. Det finns flera metoder för detta, men en av de vanligaste är att genomföra en residualsanalys.

Residualanalys är en central metod inom statistisk modellering som används för att bedöma om en vald modell är lämplig och om dess underliggande antaganden håller. Den baseras på analysen av residualer, det vill säga skillnaden mellan de observerade värdena och de värden som modellen förutsäger (Martin Singull, REGRESSIONSANALYS, april 2018).

Residualerna kan definieras som

$$e_i = y_j - \hat{y}_j \tag{3.4}$$

Där:

· e_i är skillnaden mellan observerade värdena och värdena som modellen förutspår (predikterat värden).
· y_i är de observerade värdena (riktigt värdena som samlas in i datamaterial).
· $\hat{y}_i$ är predikterat värdena.

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 fasta effekterna, och den konditional förklaringsgraden, som anger den andel varians som förklaras av både fasta och slumpmässiga effekter (Carcagno, S., R squared for linear mixed models, 2018).

Marginal förklaringsgrad kan identifieras som

$$R^2_{(m)} = \frac{\sigma^2_f}{\sigma^2_f + \sigma^2_r + \sigma^2_e} \tag{3.5}$$

Konditionellt förklaringsgrad kan identifieras som

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

Där:

· $R^2_{(m)}$: Marginal förklaringsgrad
· $R^2_{(c)}$: Konditionellt förklaringsgrad
· σ^2_f: Varians för de fasta effekterna
· σ^2_r: Varians för de slumpmässiga effekterna
· σ^2_e: Residualvarians

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).

$$AIC = 2k - 2\ln(L) \tag{3.7}$$

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 5 brukar indikera att det kan finnas ett problem med multikollinearitet, vilket kan påverka tolkningen av modellens koefficienter (DeRuiter, S. (n.d.). Collinearity and multicollinearity).

$$VIF = \frac{1}{1 - R_i^2} \tag{3.8}$$

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

Och

$$GVIF^{\frac{1}{2 \cdot Df}} \tag{3.9}$$

4. Resultat och diskussion

I detta kapitel redovisas resultaten från analysen som syftar till 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 har ett signifikant samband när både fasta och slumpmässiga effekter inkluderas i en linjär mixad modell.

4.1 Multikollinearitetsdiagnos för fixa effekter i den fullständiga modellen med GVIF- och justerad GVIF-värden

Tabell 4.1 visar resultaten från en multikollinearitetsdiagnos med hjälp av generalized variance inflation factor (GVIF) för fasta effekter från fullständigt modellen. De flesta värdena visar inga multikollinearitetsproblem eftersom de värdena är mindre än 5, men det finns 5 variabler ("cor","cro","fin","hea","lon") som kan inte beräknas, så de borde tas bort från modellen.
Tabellen visar Generalized Variance Inflation Factor (GVIF) samt justerade GVIF-värden för de fixa effekterna i den fullständiga modellen. Eftersom vissa variabler kan ha fler än en frihetsgrad används justerade GVIF-värden ($GVIF^{\frac{1}{2 \cdot Df}}$ som jämförbar indikator för multikollinearitet. Ett riktvärde på 5 används som tröskel, där högre värden kan indikera problematiskt hög korrelation mellan variabler.

De flesta fixa effekter har justerade GVIF-värden långt under 5, vilket tyder på att modellen i stort inte lider av allvarlig multikollinearitet. Dock rapporteras oändliga GVIF-värden (Inf) för fem variabler: cor, cro, fin, hea och lon. Dessa resultat indikerar att dessa variabler är perfekt kollineära med andra prediktorer i modellen, vilket gör att modellen inte kan särskilja deras enskilda bidrag till variationen i responsvariabeln. I praktiken innebär detta att modellen kan komma att fokusera på en av dessa variabler och felaktigt bortse från de övrigas eventuella betydelse. Denna typ av kollinearitet riskerar även att leda till uppblåsta standardfel, vilket i sin tur kan försvåra statistisk inferens och minska modellens förmåga att identifiera verkligt signifikanta effekter. Därmed bestäms det i detta fall att dessa variabler med höga justerade GVIF-värden kommer att tas bort.

Tabell 4.1: GVIF och justerade GVIF-värden för modellens prediktorer

Variabel	**GVIF**	**Df**	**GVIF**$^{1/(2\cdot Df)}$
acc	3.734509	1	1.932488
aer	2.757582	1	1.660597
agg	2.236332	1	1.495437
agi	2.932364	1	1.712415
ant	3.267977	1	1.807575
bal	2.059616	1	1.435136
bra	2.991112	1	1.729483
cmd	3.167054	1	1.779622
com	3.831786	1	1.957495
cmp	4.631327	1	2.152052
cnt	3.364728	1	1.834320
cor	2.749115e+06		$\boldsymbol{\infty}$
cro	2.749115e+06	0	∞
dec	5.649098	1	2.376783
det	2.865414	1	1.692753
dri	1.202864	1	1.096751
ecc	2.252988	1	1.500996
fin	2.749115e+06	0	$\boldsymbol{\infty}$
fir	3.186947	1	1.821728
fla	2.171448	1	1.473583
fre	1.990407	1	1.381843
han	3.334812	1	1.826147
hea	2.749115e+06	0	$\boldsymbol{\infty}$
jum	1.794900	1	1.339739
kic	2.608100	1	1.614961
ldr	2.157246	1	1.468757
lon	2.749115e+06	0	∞
otb	2.061609	1	1.435830
pac	2.305620	1	1.518427
pas	7.088629	1	2.662440
pen	2.013000	1	1.418800
pos	9.543382	1	3.088383
pun	3.116487	1	1.765358
ref	7.745470	1	2.781289
tro	4.695580	1	2.166699
sta	3.114606	1	1.765053
str	3.272970	1	1.809165
tea	2.613935	1	1.616767
tec	2.248555	1	1.499459
thr	2.385597	1	1.543950
vis	2.875886	1	1.695844
wor	3.407288	1	1.845884

Tabell 4.2: Resultat för fasta effekter i modellen 1

Variabel	**Estimate**	**Std. Error**	**df**	**t value**	**Pr(>|t|)**
(Intercept)	6.156e+00	5.034e-01	332.4	12.228	<2e-16 ***
acc	-1.414e-02	5.810e-03	344.2	-2.433	0.019086 *
aer	-2.679e-03	9.710e-03	317.2	-0.276	0.7832
agg	3.979e-02	9.853e-03	346.1	4.038	0.001007 **
agi	2.810e-02	9.441e-03	346.2	2.976	0.003153 **
ant	-7.404e-03	5.674e-03	227.9	-1.305	0.193219
bal	-6.157e-03	4.291e-03	332.1	-1.433	0.152892
bra	8.160e-03	6.122e-03	446.5	1.333	0.189437
cmd	-1.017e-02	4.746e-03	335.2	-2.143	0.032815 *
com	1.442e-03	1.963e-03	407.0	0.734	0.463566
cmp	-3.561e-03	6.930e-03	370.2	-0.514	0.607613
cnt	-7.920e-03	5.143e-03	331.7	-1.540	0.124234
cro	5.963e-03	7.413e-03	159.2	0.805	0.421420
dec	1.763e-03	6.706e-03	446.1	0.263	0.792762
det	1.429e-02	5.436e-03	270.6	2.630	0.009073 **
dri	-2.325e-03	2.506e-03	267.0	-0.928	0.980079
ecc	-1.429e-02	4.720e-03	335.6	-3.026	0.002846 **
fir	-2.727e-02	8.164e-03	102.0	-3.341	0.001057 **
fla	-1.385e-02	6.500e-03	311.0	-2.132	0.034764 *
fre	7.550e-03	3.543e-03	351.0	2.130	0.034974 *
han	8.896e-03	4.173e-03	316.2	2.130	0.034930 *
jum	-1.860e-04	2.931e-03	347.0	-0.063	0.949940
kic	-3.427e-03	6.376e-03	357.4	-0.537	0.591725
ldr	-1.145e-02	4.137e-03	330.1	-2.767	0.009044 **
otb	2.352e-03	5.379e-03	228.7	0.437	0.662294
pac	-3.121e-03	5.647e-03	332.9	-0.553	0.580644
pas	-1.635e-03	5.694e-03	311.4	-0.287	0.774126
pen	-1.066e-02	6.488e-03	312.0	-1.642	0.101947
pos	9.897e-05	8.488e-03	410.1	0.012	0.990383
pun	2.953e-03	6.158e-03	312.7	0.480	0.631379
ref	1.611e-03	5.819e-03	318.0	0.277	0.782769
tro	6.438e-02	4.643e-02	185.4	1.386	0.167809
sta	-1.740e-02	6.260e-03	330.9	-2.781	0.005731 **
str	-1.022e-02	6.374e-03	378.2	-1.604	0.111824
tea	-3.661e-03	5.507e-03	290.6	-0.665	0.506809
thr	-1.249e-02	6.016e-03	290.6	-2.078	0.039643 *
vis	-1.298e-02	6.181e-03	312.7	-2.100	0.036066 *
wor	-1.958e-02	8.551e-03	294.6	-2.290	0.024655 *

Tabell 4.3: Slumpmässiga effekter i modellen 1

Grupp	**Effekt**	**Varians**	**Std. avvikelse**
name	Intercept	0.003982	0.06311
Residual	–	0.413634	0.64314
Antal observationer: 5011, *grupper (name):* 53			

Den fullständiga modellen inkluderar samtliga tillgängliga prediktorer, dvs. alla möjliga individuella egenskaper av målvakter som finns i datamaterialet som fixa effekter i en linjär mixad modell med slumpmässigt intercept för varje målvakt. Som framgår av resultatet innehåller modellen ett stort antal målvaktsegenskaper som potentiella förklaringsvariabler till det genomsnittliga matchbetyget.

Vid modellanpassning genom funktionen lmer() i R uteslöts automatiskt fem variabler (cor, cro, fin, hea, lon) på grund av perfekt kollinearitet, vilket tidigare identifierats genom oändliga justerade GVIF-värden. Dessa variabler kunde därmed inte inkluderas i modellen då de skulle orsaka numeriska problem vid skattning.

Resultaten visar att flera egenskaper har ett statistiskt signifikant samband med matchbetyget, såsom agg, cmd, dec, ecc, pas, ref, och wor, medan övriga variabler inte uppvisar signifikans på konventionella nivåer.

Angående slumpmässiga effekter inkluderar modellen ett slumpmässigt intercept för varje målvakt (grupperade efter name), vilket möjliggör individuell variation kring det övergripande medelvärdet. Variansen för interceptet mellan målvakter är skattad till 0.00398, med en standardavvikelse på 0.0631. Det innebär att det finns en viss men begränsad spridning mellan målvakternas genomsnittliga matchbetyg, efter justering för de fixa effekterna.

Residualvariansen är betydligt större (0.41363), vilket tyder på att majoriteten av variationen i matchbetygen återstår att förklaras av faktorer som inte fångas av modellen. Totalt ingick 5011 observationer fördelade på 53 målvakter.

4.2 Utvalda modeller och utvärderingsmått

4.2.1 Stegvis modellen

Tabell 4.4: Resultat för fasta effekter i modellen 2

Variabel	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.271506	0.109921	188.72	57.055	<2e-16 ***
acc	-0.010689	0.004180	102.72	-2.557	0.012018 *
agg	0.023236	0.007952	52.78	2.922	0.005107 **
agi	0.028697	0.005461	133.46	5.255	5.71e-07 ***
bra	0.007460	0.004634	75.98	1.610	0.111545
cmd	0.023715	0.006075	152.85	3.904	0.000144 ***
cnt	0.014318	0.006217	115.79	2.274	0.024809 *
ecc	0.012356	0.004063	92.51	3.042	0.003090 **
fir	-0.018839	0.005108	109.36	-3.690	0.000496 ***
fre	-0.012689	0.009879	49.89	-2.696	0.009553 **
han	-0.014051	0.005742	104.60	-2.447	0.016073 *
otb	-0.045678	0.012647	41.87	-3.612	0.00102 **
pac	0.017259	0.004774	183.65	3.615	0.000387 ***
ref	0.011709	0.002139	232.26	5.474	1.14e-07 ***

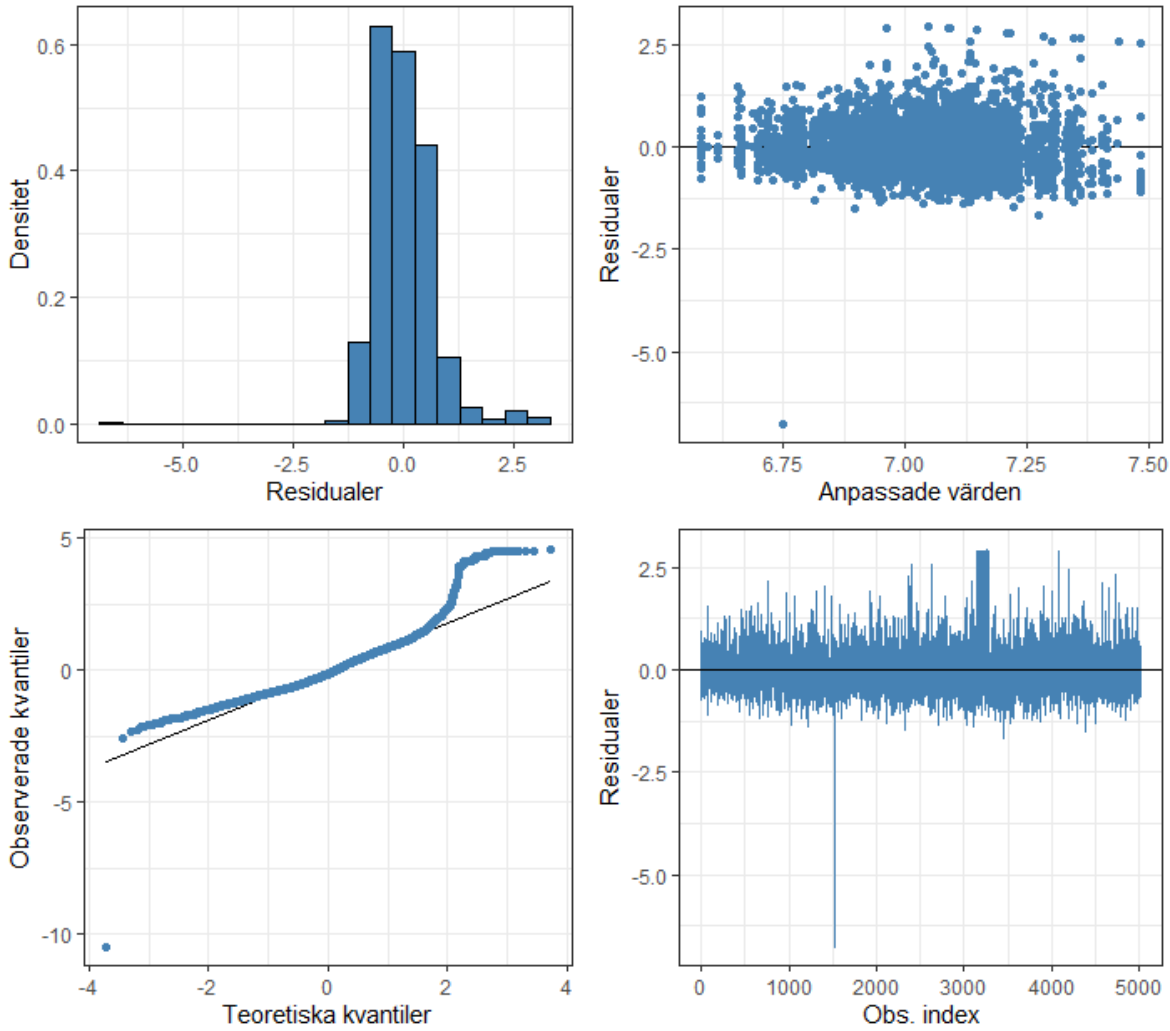
Tabell 4.5: Slumpmässiga effekter i modellen

Grupp	Effekt	Varians	Std. avvikelse
name	Intercept	0.003634	0.06028
Residual	–	0.416290	0.64521
Antal observationer: 5011, *grupper (name):* 53			

För att identifiera vilka individuella egenskaper hos målvakter som signifikant bidrar till att förklara variationen i matchbetyg, genomfördes en modellreduktion med hjälp av funktionen step() från R-paketet lmerTest. Denna funktion utför en stegvis bakåtselektering av de fixa effekterna baserat på deras statistiska signifikans. Vid varje steg utvärderas om en fix effekt kan uteslutas utan att modellen försämras signifikant. Funktionen använder sig av typ III-tester och Satterthwaites approximation av frihetsgrader. De slumpmässiga effekterna lämnas oförändrade under processen – de behålls som specificerats i den ursprungliga modellen och påverkas inte av selekteringen.

I den reducerade modellen kvarstår ett antal egenskaper som har ett signifikant samband med det genomsnittliga matchbetyget: agg, agi, cmd, cnt, ecc, fir, fre, han, otb, pac, ref och wor. Dessa prediktorer är samtliga signifikanta på 5%-nivån eller lägre, vilket indikerar att de förklarar en del av den observerade variationen mellan målvakter.

Även denna modell innehåller ett slumpmässigt intercept för varje målvakt. Variansen i interceptet är 0.00218, vilket motsvarar en standardavvikelse på 0.0467. Jämfört med den fullständiga modellen har den individuella variationen minskat något, vilket tyder på att en större del av variationen nu fångas upp av de utvalda fixa effekterna. Den kvarvarande residualvariansen är dock fortfarande relativt stor (0.4164), vilket indikerar att viss variation förblir oförklarad.



Figur 4.1: Residualdiagrammet för modellen 2

Histogrammet är inte perfekt, men det visar en liten symmetrisk fördelning och har en klockform med toppencentrerad runt väntevärdet 0. Samtidigt visar QQ-diagrammet nederst till vänster i figur 4.1 att de observerade värdena huvudsakligen följer ganska bra den inritade linjen i mitten, men tydliga avvikelser i båda svansarna, särskilt extremt låga residualer. Överst till höger visas ett diagram med syftet att kontrollera antagandet om residualernas lika varians, punkterna är spridda horisontellt kring noll, vilket är förväntat men det dock finns ett par extrema outliers, och viss klustring kring mitten.

4.2.2 Reducera modellen

Tabell 4.6: Resultat för fasta effekter i modellen 3

Variabel	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.237747	0.099486	253.82	62.699	<2e-16 ***
acc	-0.008103	0.004203	74.30	-1.928	0.057675 .
agg	0.026626	0.007932	54.58	3.357	0.001442 **
agi	0.028863	0.005156	155.70	5.233	5.32e-07 ***
cmd	0.023954	0.004068	176.85	5.896	0.000139 ***
cnt	0.013987	0.006338	131.34	2.209	0.028906 *
ecc	-0.012062	0.004419	113.30	-2.729	0.007359 **
fir	-0.017527	0.005277	127.50	-3.322	0.001167 **
fre	-0.025798	0.009616	58.69	-2.683	0.009470 **
han	-0.012158	0.005657	107.06	-2.149	0.033895 *
otb	-0.031286	0.012091	56.53	-2.586	0.000174 ***
pac	0.016776	0.004691	202.93	3.576	0.000436 ***
ref	0.011773	0.002122	222.74	5.547	8.19e-08 ***

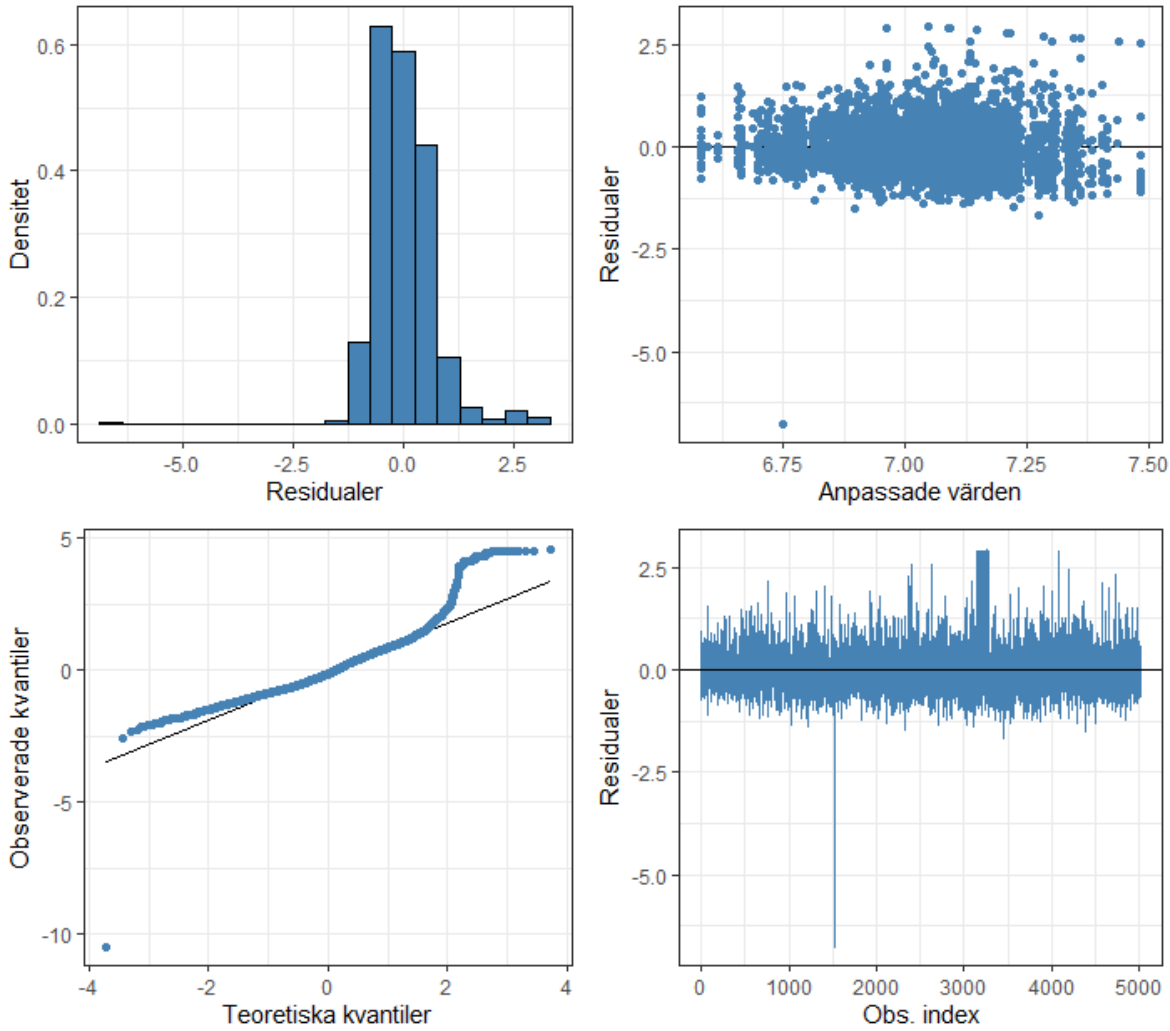
Tabell 4.7: Slumpmässiga effekter i den slutliga modellen

Grupp	Effekt	Varians	Std. avvikelse
name	Intercept	0.003634	0.06028
Residual	–	0.416290	0.64521
Antal observationer: 5011, *grupper (name):* 53			

Modell 3 bygger vidare på modell 2, som togs fram med hjälp av funktionen step() från paketet lmerTest. I modell 3 har ytterligare variabler som inte uppvisade statistisk signifikans successivt tagits bort manuellt, vilket resulterat i en slutlig, förenklad modell med bibehållen tolkningsstyrka.

Resultaten visar att flera individuella egenskaper har ett statistiskt signifikant samband med målvakters genomsnittliga matchbetyg. Bland dessa finns agg, agi, cmd, cnt, ecc, fir, fre, han, otb, pac och ref, som alla uppvisar p-värden under 0.05. Detta indikerar att dessa egenskaper har ett självständigt bidrag till förklaringen av variationen i betyg, efter att hänsyn tagits till övriga prediktorer i modellen. Variabeln acc uppvisar en svag negativ effekt ($p = 0.0577$) och är marginellt icke-signifikant, men har behållits i modellen eftersom borttagning ledde till ett försämrat AIC-värde, vilket indikerar en sämre övergripande modellpassning.

Likadant som tidigare modeller inkluderar denna modell ett slumpmässigt intercept. Variansen för dessa intercept skattas till 0.00363, medan residualvariansen är 0.41629. Detta visar att en större del av variationen finns inom samma målvakt, men att modellen fångar viss variation mellan olika målvakter. Den relativt låga variansen för det slumpmässiga interceptet kan tyda på att upprepade mätningar inom varje målvakt inte är särskilt starkt korrelerade. Trots att datan är longitudinell finns det alltså inte nödvändigtvis något tydligt beroende mellan observationerna från samma individ, vilket kan förklara varför den slumpmässiga effekten har så litet genomslag i modellen.



Figur 4.2: Residualdiagrammet för modellen 3

Histogrammet över residualer indikerar att residualerna är centrerade kring noll, men fördelningen är något vänsterskev, med några extrema negativa värden. Detta tyder på en mindre avvikelse från normalitet, men den är inte dramatisk. För majoriteten av observationerna är fördelningen relativt smal och symmetrisk. Outliers bidrar dock till avvikelser i svansen, vilket också reflekteras i Q-Q-plottens nederkant.

Spridningsdiagrammet visar ingen tydlig struktur eller trend, vilket tyder på att variansen hos residualerna är relativt konstant över hela intervallet av anpassade värden. Det finns dock tecken på viss spridningsökning vid högre anpassade värden samt ett par extrema värden. Sammantaget stödjer detta antagandet om homoskedasticitet, även om viss osäkerhet kvarstår.

Q-Q-plottens punkter följer den teoretiska normalfördelningen ganska väl i mitten, men avviker i både

nedre och övre svansarna. Detta antyder att residualerna inte är helt normalfördelade, framför allt till följd av ett fåtal outliers.

Denna graf används för att kontrollera eventuellt tidsmässigt beroende eller sekventiella mönster. Residualerna är spridda oregelbundet över observationsindex utan några tydliga mönster eller autokorrelationsstrukturer. Detta talar för att residualerna är oberoende över observationer.

4.3 Jämförelse modeller

4.3.1 AIC för modeller

Tabell 4.8: Jämförelse av AIC-värden för tre modeller

Modell	**AIC-värdet**
Model 1	9908.004
Model 2	9886.887
Model 3	9886.611

För att jämföra modellernas lämplighet användes Akaike's Information Criterion (AIC), där lägre värde indikerar en modell med bättre balans mellan modellens förklaringsförmåga och komplexitet. Som framgår av Tabell 4.8 har Modell 3 det lägsta AIC-värdet (9886.611), följt tätt av Modell 2 (9886.887) och Modell 1 (9908.004). Eftersom skillnaden mellan Modell 2 och 3 är mycket liten, men modell 3 har det lägsta värdet, valdes denna modell som slutlig. AIC användes därmed som ett objektivt kriterium för modellurval i denna studie.

4.3.2 Förklaringsgrad med modellen som har lägsta AIC-värdet

Tabell 4.9: Förklaringsgrader för den slutliga modellen

Typ av R^2	**Värde**
Marginal R^2	0,053
Konditionell R^2	0,061

Den slutgiltiga modellen har ett marginalt $\mathbf{R^2}$ på 0.053, vilket innebär att de fixa effekterna i modellen (t.ex. målvaktens egenskaper) förklarar cirka 5,3 % av variationen i matchbetygen. Det konditionella $\mathbf{R^2}$-värdet är 0.062, vilket inkluderar både de fixa effekterna och den slumpmässiga effekten (intercept per målvakt). Detta innebär att modellen som helhet förklarar cirka 6,2 % av variationen.

Skillnaden mellan marginalt och konditionell $\mathbf{R^2}$ är liten (0.009), vilket tyder på att den slumpmässiga effekten (skillnader mellan målvakter) bidrar endast marginellt till modellens totala förklaringsgrad. Det överensstämmer med tidigare resultat, där den skattade variansen för den slumpmässiga interceptkomponenten var mycket låg.

4.4 Koppling till tidigare studie

Resultaten från analysen visar att de individuella egenskaper som inkluderades i modellen endast förklarar en mindre del av variationen i målvakternas matchbetyg (marginalt $R^2 = 0{,}053$). Eftersom datan är baserad på en simulerad fotbollsliga i spelet Football Manager 2022 är det viktigt att förstå att betygssättningen i spelet kan påverkas av flera faktorer utöver de individuella attributen. 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.

5. Slutsatser

Syftet med studien var att undersöka:

1. **I vilken utsträckning målvakters individuella egenskaper kan förklara variationen i deras matchbetyg?**
2. **Vilka egenskaper som har ett signifikant samband när både fasta och slumpmässiga effekter beaktas.**

Analysen baserades på en linjär mixad modell där 12 individuella målvaktsegenskaper och *interceptet* inkluderades som fixa effekter, och ett slumpmässigt *intercept* per målvakt modellerade variation mellan individer. Modellens **marginala** R^2-värde uppgick till **0.053**, vilket innebär att de fixa effekterna förklarade cirka 5,3 % av variationen i matchbetyget. Det **konditionella** R^2-värdet var **0.061**, vilket tyder på att den slumpmässiga effekten endast bidrog marginellt till den totala förklaringsgraden. Detta indikerar att variationen i matchbetyg främst förklaras av faktorer som inte fångas av de ingående variablerna.

Trots detta identifierades flera egenskaper med statistiskt signifikanta samband med matchbetyget. Bland dessa återfinns:

- **Positivt samband:** *agg, agi, cmd, cnt, pac, ref*
- **Negativt samband:** *ecc, fir, han, otb, wor*

Variabeln *acc* uppvisade en svagt negativ effekt men behölls i modellen eftersom dess uteslutning försämrade modellens AIC-värde.

Slutsatsen är att vissa målvaktsegenskaper har ett signifikant samband med betygsättningen, 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] DeRuiter, S. (n.d.). Collinearity and multicollinearity. Retrieved May 27, 2025, from https://stacyderuiter.github.io/s245-notes-bookdown/collinearity-and-multicollinearity.html

[2] Fox, John, och Georges Monette. 1992. "Generalized Collinearity Diagnostics". Journal of the American Statistical Association 87 (417): 178–83. http://www.jstor.org/stable/2290467.

[3] Fife, D. (n.d.). flexplot: Graphically based data analysis using linear models. GitHub repository. Retrieved May 27, 2025, from https://github.com/dustinfife/flexplot

[4] Carcagno, S. (n.d.). R^2 for linear mixed models. Retrieved May 27, 2025, from https://samcarcagno.altervista.org/stat$_n$$otes/r2_lmm_jags/r_squared_lmm.html$

[5] Weil, et al. (2025). *Football (soccer).* `https://www.britannica.com/sports/football-soccer` [Hämtad 7 april 2025].

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

[7] Pierce, H. (2024, 15 november). *Player match ratings: The metrics behind soccer performance.* `https://soccerwizdom.com/2024/11/15/player-match-ratings-the-metrics-behind-soccer-performance/` [Hämtad 8 april 2025].

[8] 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), 803–812. `https://doi.org/10.1080/02640414.2025.2471208`

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

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

[11] 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`

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

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

[14] Saarinen, F. (2004). *Using mixed models in a cross-over study with repeated measurements within periods* (Examensarbete 2004:22). Stockholms universitet, Matematiska institutionen.

Bilaga

```r
### Libraries used in project
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)
require(devtools)
# install the stable version
devtools::install_github("dustinfife/flexplot")
# install the development version
devtools::install_github("dustinfife/flexplot", ref="development")
library(flexplot)



### Läsa in data set.
setwd("X:/732G56_project/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")]

### hantera med 0
only_GK[ , sapply(only_GK, is.numeric)] <- lapply(only_GK[ , sapply(only_GK, is.numeric)], function(x) {
  x[is.na(x)] <- mean(x, na.rm = TRUE)
  return(x)
})
```

Figur 1: Kod R - 01

```r
### fullständig modell
model_test <- lmer(average.rating~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+otb+pac+pas+pen+pos+pun+ref+tro+sta+
                     str+tea+tec+thr+vis+wor+(1|name),
                   data=only_GK, control = lmerControl(optCtrl = list(maxfun = 100000)),
                   REML = FALSE,
                   na.action = na.fail)

r2(model_test)
AIC(model_test)


###Model1
model1 <- lmer(average.rating~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+otb+pac+pas+pen+pos+pun+ref+tro+sta+
                 str+tea+tec+thr+vis+wor+(1|name),
               data=only_GK_over5,
               REML = FALSE,
               na.action = na.fail)

summary(model1)
AIC(model1)
r2(model1)
vif(model1) %>%
  as_tibble(rownames = NA) %>%
  rownames_to_column() %>%
  kable(
    digits = 4
  ) %>%
  kable_styling("striped")


###model2
model2 <- lmer(average.rating~acc+aer+agg+agi+ant+bal+bra+cmd+com+cmp+cnt+
                 dec+det+dri+ecc+fir+fla+fre+han+jum+kic+
                 ldr+otb+pac+pas+pen+pos+pun+ref+tro+sta+
                 str+tea+tec+thr+vis+wor+(1|name),
               data=only_GK,
               REML = FALSE,
               na.action = na.fail)
summary(model2)
vif(model2)
AIC(model2)
summary(model2) %>%
  coef() %>%
  as_tibble(rownames = NA) %>%
  rownames_to_column() %>%
  rename(
    ` ` = rowname,
    Skattning = Estimate,
    Medelfel = `Std. Error`,
    `t-värde` = `t value`,
    `p-värde` = `Pr(>|t|)`
  ) %>%
  kable(
    digits = 4
  ) %>%
  kable_styling("striped")
```

Figur 2: Kod R - 02

```r
###step() för att hitta den bäst modellen
step_model <- step(model2)
summary(step_model)

###model3
model3 <- lmer(average.rating ~ acc + agg + agi + bra + cmd + cnt +
                 ecc + fir + fre + han + otb + pac + ref + (1|name),
               data=only_GK,
               REML = FALSE,
               na.action = na.fail)
AIC(model3)
r2(model3)
summary(model3)
summary(model3) %>%
  coef() %>%
  as_tibble(rownames = NA) %>%
  rownames_to_column() %>%
  rename(
    ` ` = rowname,
    Skattning = Estimate,
    Medelfel = `Std. Error`,
    `t-värde` = `t value`,
    `p-värde` = `Pr(>|t|)`
  ) %>%
  kable(
    digits = 4
  ) %>%
  kable_styling("striped")



# 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")
  cowplot::plot_grid(p1, p2, p3, nrow = 2)
}
residualPlots(model3)
```

Figur 3: Kod R - 03

```r
###model4
model4 <- lmer(average.rating ~ acc + agg + agi + cmd + cnt +
                 ecc + fir + fre + han + otb + pac + ref + (1|name),
               data=only_GK,
               REML = FALSE,
               na.action = na.fail)
AIC(model4)
summary(model4)
r2(model4)
residualPlots(model4)
```

Figur 4: Kod R - 04