När datan talar, vad avgör en säsong i Premier League

-En multipel regressionsanalys av faktorer som beskriver sportslig framgång i fotboll.

Författare: **Loran Ali**

Sammanfattning

Studien analyserar i vilken utsträckning en begränsad uppsättning prestations- och resursrelaterade variabler kan förklara variationen i säsongspoäng i Premier League samt vilken prediktionsförmåga en sådan modell uppvisar. Analysen baseras på lagdata från säsongerna 2020/2021–2024/2025 och genomförs med multipel linjär regressionsanalys.

Resultaten visar att modellen uppnår en hög justerad förklaringsgrad på cirka 0,86, där offensiva mått som skott på mål och förväntade mål har positiva och statistiskt signifikanta samband med poäng. Defensiva misstag, hög trupprotation och stor total löpdistans är negativt relaterade till poäng, medan lönekostnader uppvisar ett positivt samband. En utvärdering på data utanför skattningsperioden visar dock att modellens prediktion är begränsad, med genomsnittliga fel på cirka 8–10 poäng per lag och säsong. Studien visar därmed att enkla regressionsmodeller kan fånga övergripande mönster i säsongsdata, men att fotbollens inneboende osäkerhet begränsar möjligheten till exakta prognoser.

Nyckelord:
Fotboll, Premier League, säsongspoäng, multipel linjär regression, paneldata, prediktion.

1. Inledning

Prediktion av sportsliga resultat har alltid varit en del av fotbollens utveckling, inte minst för ekonomin kopplat till odds och spel. Tillgång till större datamängder samt mer bearbetningskapacitet har höjt kravet på förklarings- och prediktionsmodeller, vilket har drivit fram ett brett spann av metoder.

Fotboll är ett område där strukturella mönster och slumpmässiga inslag samverkar. Detta gör att modeller som försöker förklara resultat uppvisar begränsad förklaringsgrad, trots avancerade ansatser. Tidigare forskning har använt allt från traditionella regressionsmodeller till moderna maskininlärningsalgoritmer. Trots denna variation saknas fortfarande samsyn kring vilka variabler som kan förklara prestationer inom fotbollen.

Premier League utgör en lämplig miljö för att studera sambanden mellan utfall och påverkande variabler, eftersom ligan kombinerar stora mängder data med relativt stabila förutsättningar, där endast tre lag av 20 byts ut per säsong. Offensiva respektive defensiva nyckeltal tillsammans med ekonomiska variabler. Samtidigt varierar urval och kombination av variabler mellan studier.

Mot denna bakgrund är studiens mål att prediktera hur många poäng ett lag kan ta i Premier League under en hel säsong. Syftet är att analysera och identifiera vilka variabler som har betydelse för poängutfallet. Modellen har som ambition att vara enkel genom att bygga på ett begränsat antal förklarande variabler. Genom att tillämpa regressionsanalysen på paneldata över flera säsonger ska studien empiriskt pröva om exempelvis några av de intuitiva variablerna som skott på mål, bollinnehav och anfall har den relationen till poäng. Studien testar även om modellen fungerar som prediktionsmodell för säsonger som inte ingår i skattningsperioden.

Forskningsfrågan:

- ➔ Kan en begränsad uppsättning variabler förklara variationen i poängutfall per säsong i Premier League och vilken prediktionsförmåga har modellen?

2. Bakgrund och tidigare forskning

Forskningen om prestation i professionell fotboll har utvecklats i takt med att datatillgången blivit mer omfattande och att klubbar, analytiker och spelmarknader fått ett ökande behov av modeller som kan beskriva och förutsäga lagens resultat. En återkommande frågeställning är vilka variabler som förklarar variationen i lagens prestationer, samt i vilken utsträckning statistiska modeller kan fånga samband i en sport som kännetecknas av systematik och slumpmoment.

Oberstone (2009) som analyserar Premier League under säsongen 2007–08 för att undersöka vilka faktorer som påverkar utfallet av lagens poäng på en säsong. Studien utgår från 24 händelser spelarna utför på planen (pitch actions) och identifierar, genom statistisk variabelreduktion 6 förklaringsvariabler. Dessa variabler rör både offensiva och defensiva faktorer. Studien visar att det går att isolera ett antal variabler som poäng per avslut, andel mål utanför straffområdet, kvoten korta och långa passningar, inläggsprocent, insläppta mål per match och antal gula kort som fångar skillnaderna i poäng mellan lagen. Oberstone (2009) betonar dock att modellen är retroaktiv och syftar till att beskriva lagens prestationer i efterhand snarare än att fungera som en generell prediktionsmodell.

Vidare har Souza et al. (2019) analyserat åtta säsonger i LaLiga (högsta spanska ligan) och undersöker hur offensiva och defensiva nyckelvariabler förklarar antalet poäng per säsong. Med multipel regressionsanalys skattas separata modeller för offensiva och defensiva variabler, där den offensiva modellen förklarar 84 procent och den defensiva 73 procent av variationen i säsongspoäng. Resultaten indikerar att ett begränsat antal variabler står för en relativt stor del av den förklarade variationen och att både offensiv och defensiv förhållning är relevanta för att förstå poängutfall över en säsong.

För att vidga perspektivet analyserar och jämför Oberstone (2011) tre av Europas största ligor, Premier League, Serie A (högsta italienska ligan) och La Liga, för säsongen 2008–09 med multipel regressionsanalys. Studien undersöker vilka variabler som återkommer i olika nationella toppligor, vilket antyder att vissa variabler uppvisar liknande samband över landsgränser och korrelerar med höga liga poäng. Oberstone (2011) betonar dock att en unik modell som utvecklats för en ligas förutsättningar tyder på bättre resultat.

En annan utgångspunkt i litteraturen är att specifika matcher i fotboll präglas av hög oförutsägbarhet. Kundu et al. (2021) analyserar Premier League på matchnivå med data från tolv säsonger och visar att prediktionsprecisionen för matchutfall är begränsad, även när många variabler och maskininlärningsmetoder används. Författarna relaterar den måttliga träffsäkerheten till att matcher uppvisar hög grad av slump.

Sammantaget visar tidigare studier att det finns en etablerad empirisk litteratur där multipla regressionsmodeller används för att analysera prestationer i professionell fotboll. Dessa modeller har tillämpats på såväl enskilda matcher som på data aggregerad över en eller flera säsonger. Trots att fotboll kännetecknas av betydande slumpmässighet i enskilda matcher visar forskningen att modellerna kan förklara en del av variationen i lagens totala poäng över en hel säsong. Vidare indikerar resultaten i studierna att det är möjligt att identifiera ett begränsat antal variabler som återkommande uppvisar statistiskt signifikanta samband med lagens slutliga poängutfall per säsong.

3. Data

Data är hämtade från FBREF (u.å.), som innehåller statistik för professionella fotbollsligor världen över. Datamaterialet är indelat i olika kategorier och områden, vilka i denna studie delas in i offensivt, defensivt, innehav, fysiska attribut samt en ekonomisk variabel.

Studien består av data som används i lagstatistik från Premier League under säsongerna 2020/2021 till 2024/2025. Datamaterialet omfattar de 20 lag som deltog i ligan under varje säsong och inkluderar totalt 13 variabler. Variabler är valda utifrån teoretisk relevans och inspirerad av tidigare forskning om fotboll. För att säkerställa ett brett och effektivt datamaterial, som har chans att förklara variationen, har variabler från flera dimensioner av lagprestationer inkluderats.

Tabell 1. Variabler med beskrivning samt variabelnummer

Variabler	**Beskrivning**	**Variabelnummer (Parameter)**
Poäng (Beroende variabel)	Poäng samlade under säsongen	Y
SoT (Shots on Target)	Antal skott på mål under säsongen	$X_1\ (\beta_1)$
xG (Expected goals)	Ett mått på kvalitativa målchanser	$X_2\ (\beta_2)$
SCA (Shot-Creating-Actions)	Antal aktioner som leder till avslut	$X_3\ (\beta_3)$
FinalThird	Antal passningar i den offensiva tredjedelen	$X_4\ (\beta_4)$
PrgP (Progressive Passes)	Andel progressiva passningar framåt	$X_5\ (\beta_5)$
Cmp (Passningsprocent)	Andel lyckade passningar	$X_6\ (\beta_6)$
TkIW (Tackles Won)	Antal vunna tacklingar	$X_7\ (\beta_7)$
Err (Errors leading to shot/goal)	Defensiva misstag som lett till avslut eller mål	$X_8\ (\beta_8)$
TotDist (Total distans)	Total distans laget sprungit under säsongen	$X_9\ (\beta_9)$
Age	Genomsnittlig ålder i truppen	$X_{10}\ (\beta_{10})$
PI	Antal spelare som använts under säsongen	$X_{11}\ (\beta_{11})$
Wages	Totala lönekostnader för laget under säsongen (pund)	$X_{12}\ (\beta_{12})$

Dessa 13 variabler är uppdelade så att *poäng* är den beroende variabeln som förklaras av övriga 12. För förståelse vad variablerna betyder följer en detaljerad beskrivning.

Offensiva:

SoT (Shots on Target) beskriver antalet skott som antingen resulterade i mål eller att motståndarens målvakt räddar den. Denna variabel tar inte hänsyn till ribb och stolpträffar om inte målvakten rört den innan träff av målram.

xG (Expected goals) är ett mått på sannolikheten att ett avslut resulterar i mål. Variabeln är baserad på historisk data men även faktorer som spelarens position vid avslut, nick eller skott och föregående aktioner. Den tar ett värde mellan 0 och 1, ett värde närmare 1 visar på en säker målchans.

SCA (Shot-Creating-Actions) bygger på händelserna innan avslut. Den mäter de två offensiva aktionerna som leder till ett avslut. Skott, dribbling samt tacklingar vunna är några varianter av SCA.

Innehav:

FinalThird hänvisar till offensiva passningar och bollförflyttningar utanför motståndarens sista tredjedel av planen och in i den, alternativt att den når fram till motståndarens målområde. Måttet fokuserar huvudsakligen på den offensiva tredjedelen, vilket är inom motståndarens planhalva.

PrgP (Progressive Passes) är slutförda passningar mellan lagkamrater som har färdats minst 10 yards[1] mot motståndarens mål, den exkluderar passningar inom den första 40 procenten av planen från egen målvakt.

Cmp (Passningsprocent) visar passningar mellan två lagkamrater. Beräkningen av måttet är antal lyckade passningar, dividerat på antal passningsförsök.

[1.] 1 yard (UK) = 0,9144 meter

Defensiva:

TklW (Tackles Won) beskriver antalet defensiva aktioner där spelaren som genomför tacklingen eller lagkamraten, återerövrar bollinnehavet.

Err (Error leading to Goal) räknar antalet defensiva misstag som direkt leder till ett avslut eller mål för motståndarlaget.

Fysiskt:

TotDist (Total distans) mäter den totala distans som lagets spelare förflyttar sig när laget har bollinnehav, vilket mäts i yards.

Age visar lagets genomsnittliga ålder, där mätningen normalt sett sker vid säsongens start.

PI visar hur många spelare ett lag använt under en säsong. Spelare som endast spelat några minuter kommer att räknas.

Ekonomiskt:

Wages avser den totala ekonomiska ersättningen som spelarna tjänar från sina klubbar under säsongen, vilket inkluderar grundlön, bonusar och andra ersättningar, i brittisk pund.

Tabell 2. Deskriptiv data för studiens variabler

Variabler	Medelvärde	Median	Min	Max	SD	Obs
Poäng	52,5	51,5	12	93	18,34	100
SoT	162,5	159	92	264	34,99	100
xG	53,34	51,15	32,3	88,7	13,65	100
SCA	857,7	822	567	1388	170,16	100
FinalThird	1153,2	1084,5	660	2110	290,4	100
PrgP	1431	1349	859	2421	349,29	100
cmp	79,12	79,5	67	88,7	4,69	100
TklW	371,7	371	252	493	50,94	100
Err	18,17	16	5	51	8,98	100
TotDist	69 397	68 376	42 272	110 018	14 944	100
Age	26,59	26,5	23,7	29,1	1,06	100
PI	28,01	28	23	36	2,94	100
Wages	89 902 773	71 770 000	20 087 200	238 850 000	52 771 130	100

Tabell 2 presenterar deskriptiv statistik för studiens variable. För samtliga variabler redovisas medelvärde, median, minsta värde (Min), största värde (Max), standardavvikelser (SD) och antal observationer (Obs). Antal observationer är genomgående 100, vilket indikerar ett balanserat dataset utan bortfall. Variablerna uppvisar varierande spridning, där särskilt TotDist och Wages har avsevärt större värden och standardavvikelser jämfört med resterande variabler, vilket återspeglar deras större skala. I övrigt ligger medianerna nära respektive medelvärde, vilket tyder på relativt symmetriska fördelningar för flera av variablerna.

4. Metod och modell

Metoden är en multipel linjär regressionsmodell. För att säkerställa att de potentiellt förklarande variablerna är relevanta kommer vi att bearbeta data med variabelselektion och multikolinjäritetstestet (Wooldridge 2019).

Den allmänna formeln för en multipel linjär regression är följande:

1) $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Där Y är den beroende variabeln och X_{i1}, X_{i2},..., X_{iK} är de oberoende variablerna, β_0 är interceptet och β_1, β_2,..., β_k är koefficienterna för variablerna och feltermen ε.

Den empiriska analysen genomförs med en multipel linjär regressionsmodell där responsvariabeln utgörs av lagens totala antal poäng. De förklarande variablerna utgörs av olika offensiva, defensiva och speluppbyggande prestationsmått, vilka teoretiskt kan påverka lagens poängutfall. I den initiala modellen inkluderas samtliga 12 variabler enligt följande, antal skott på mål (SoT), förväntade mål (xG), skapade sekvenser som leder till avslut (SCA), bollinnehav i offensiv tredjedel (FinalThird), progressiva passningar (PrgP), passningsprecision (cmp), tacklingsprecision (TklW), individuella misstag som leder till avslut (Err), total förflyttad distans (TotDist), genomsnittlig spelarålder (Age), antal spelare som använts per säsong (PI) samt lagens totala lönekostnad (Wages).

Den empiriska modellen specificeras enligt följande:

2) $$\mathrm{Poäng}_i = \beta_0 + \beta_1 \mathrm{SoT}_i + \beta_2 \mathrm{xG}_i + \cdots + \beta_{12} \mathrm{Wages}_i + \varepsilon_i$$

Betavärden β_1, β_2,..., β_{12} skattas med minsta kvadratmetoden (OLS). Denna modell används som utgångspunkt för att identifiera vilka variabler som empiriskt och teoretiskt bidrar till variationen i poäng. Tabell 1 visar varje variabels nummer samt parameter för modellen.

Eftersom variablerna delvis mäter närliggande aspekter av lagens prestationer kan kollinearitet förväntas. För att erhålla en specifikation som uppfyller OLS-antagandena och samtidigt undviker en modell med fler parametrar än vad som är motiverat av datamaterialet, kommer den att reduceras med Variansinflationsfaktor (VIF) och Akaikes informationskriterium (AIC).

Först kontrolleras multikollinearitet mellan de förklarande variablerna med VIF:

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

* *R^2_j är förklaringsgraden när variabel X_j estimeras i en regression på alla andra variabler.*

VIF-värdena anger hur mycket variansen i respektive koefficient ökar till följd av linjära samband mellan variablerna. Höga VIF-värden indikerar osäkerhet i skattningarna och risk för överlappande informationsinnehåll.

Därefter genomförs modellreduktion med AIC.

4) $$AIC = 2k - 2\ln(L)$$ * *k = antal variabler L = modellens maximala likelihood*

AIC används för att jämföra modeller med olika antal variabler och säkerställer att den slutliga modellen balanserar förklaringsgrad och modellens komplexitet. Processen genomförs genom stegvis variabelreduktion från den fulla modellen. Således skapas en balans mellan förklaring och enkelhet i modellen som slutligen används.

Modellen tolkas som en deskriptiv regressionsmodell där varje koefficient uttrycker genomsnittligt samband mellan respektive prestationsmått och lagens poäng givet övriga variabler konstanta. Då modellen endast kan kontrollera för observerbara faktorer och eftersom prestationsmått i fotboll innehåller potentiellt samtidiga samband bör skattningarna inte ges en kausal tolkning, utan studien kan ge en fortsatt inriktning kopplat till inferens. Den slutliga modellen representerar dock den specifikation som empiriskt bäst beskriver variationen i lagens poäng inom ramen för tillgängliga data och metodologiska begränsningar.

För att utvärdera modellens prediktionsförmåga kommer root mean squared error (RMSE) och mean absolute error (MAE) beräknas som prestationsmått. Dessa mått kommer att användas på testdata utanför modelleringsperioden. MAE kommer mäta det genomsnittliga absoluta felet i predikterade poäng, medan RMSE straffar större fel hårdare.

Formlerna enligt följande:

5)
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

6)
$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

5. Resultat

Resultatet av första regressionen enligt tabell 3. Följande variabler, förutom SCA, FinalThird, PrgP och Age, uppvisar en statistiskt signifikant samband med poäng.

Tabell 3: Resultat från första regressionen.

Observationer	100
Prob> F	<0,001
R-squared	0,877
Adj R-squared	0,860

Variabler	Estimation	Std. Fel	t-värde	p-värde	Signifikans
Intercept	−20,980	35,780	−0,586	0,559	
SoT	0,273	0,078	3,481	0,001	***
xG	0,429	0,161	2,666	0,009	**
SCA	−0,028	0,017	−1,631	0,106	
FinalThird	−0,001	0,009	−0,120	0,904	
PrgP	0,009	0,009	1,061	0,291	
cmp	1,246	0,448	2,781	0,006	**
TklW	−0,043	0,016	−2,636	0,009	**
Err	−0,194	0,095	−2,043	0,044	*
TotDist	−0,0004	0,0001	−2,211	0,029	*
Age	0,025	0,737	0,035	0,972	
PI	−1,399	0,286	−4,888	<0.001	***
Wages	$5{,}934 \times 10^{-8}$	$1{,}938 \times 10^{-8}$	3,061	0,002	**

*Not. . p < 0,10; * p < 0,05; ** p < 0,01; *** p < 0,001.*

Tabell 4 visar resultatet från VIF-testet och visar variablernas multikolinjäritet. Vanliga brytpunkter är 5, 10 och 15. Där ett värde större än fem kan innebära en risk för kollinearitet, ett värde större än tio tyder på kollinearitet och slutligen ett värde större än 15 är det problematiskt att inkludera den i regressionen.

Tabell 4: VIF-värden från den första regressionen

SoT	xG	SCA	Final Third	PrgP	cmp	TklW	Err	TotDist	Age	PI	Wages
15.89	10.19	18.24	16.73	21.80	9.32	1.50	1.53	15.60	1.27	1.49	2.20

Det finns 6 variabler som har ett VIF-värde över tio. SoT, xG och TotDist uppvisar ett värde över tio, men regressionen med samtliga variabler ger ett signifikant resultat och därmed kan dessa variabler vara viktiga för modellen trots deras värden bibehålls dessa, medan SCA, FinalThird och PrgP inte visar signifikans samt ett värde över 15 kommer dessa att exkluderas för en enklare och mer effektiv modell.

Fortsättningsvis utfördes en stegvis regression som tar hänsyn till AIC-värde. AIC fokuserar på prediktionsförmåga jämfört med VIF som analyserar kollinearitet. Resultatet gav oss att några av de variabler som borde exkluderats tidigare även exkluderades i den stegvisa regressionen, medan några inte gjorde det trots höga VIF-värden. Den slutliga modellen som genereras av AIC-proceduren inkluderar SoT, xG, cmp, TklW, Err, TotDist, PI och Wages, medan variabler FinalThird, SCA, PrgP och Age exkluderas då de inte förbättrar modellen. Den reducerade modellen uppvisar en liknande justerad förklaringsgrad som den fulla modellen, samtidigt som antalet variabler är lägre och risken för kollinearitet är reducerad. Resultatet i den slutgiltiga regressionen i tabell 5 visar att samtliga variabler är statistiskt signifikanta och bidrar till att förklara lagens poängutfall över en säsong.

Tabell 5: Resultat från den stegvisa regressionsmodellen

Observationer	100
Prob> F	<0,001
R-squared	0,873
Adj R-squared	0,8618

Variabler	**Estimation**	**Std. Fel**	**t-värde**	**p-värde**	**Signifikans**
Intercept	−23,310	26,780	−0,871	0,386	
SoT	0,189	0,060	3,132	0,002	**
xG	0,431	0,137	3,134	0,002	**
cmp	1,262	0,441	2,860	0,005	**
TklW	−0,043	0,016	−2,697	0,008	**
Err	−0,202	0,091	−2,204	0,030	*
TotDist	−0,0003	0,0001	−2,151	0,034	*
PI	−1,463	0,263	−5,550	<0.001	***
Wages	$5{,}57 \times 10^{-8}$	$1{,}859 \times 10^{-8}$	2,997	0,003	**

*Not. . p < 0,10; * p < 0,05; ** p < 0,01; *** p < 0,001.*

Interceptet representerar det förväntade värdet på den beroende variabeln i en linjär regressionsmodell då samtliga förklarande variabler antar värdet noll. Värdet -23.310 kan därmed tolkas som ett tekniskt startvärde. Eftersom majoriteten av de inkluderade variablerna uppvisar positiva samband med den beroende variabeln adderas dessa effekter, vilket resulterar i predikterade poäng på realistiska nivåer, vilket framgår av tabell 7.

Tabell 6: VIF-värden för den stegvisa regressionsmodellen

SoT	**xG**	**cmp**	**TklW**	**Err**	**TotDist**	**PI**	**Wages**
9.56	7.50	9.13	1.43	1.45	10.37	1.27	2.05

Den slutgiltiga, stegvisa, regressionsmodellen visar att variablerna har en förklarande effekt på den beroende variabeln. Modellen visar ett starkt justerat R^2-värde på 0.862 och samtliga variabler är statistiskt signifikanta på fem procents nivå.

För att skapa en bättre bild av regressionens prediktionsförmåga använde vi den för att estimera lagets poäng under säsongen 2019/2020. Där är träningsdata från säsongen 2020/2021 fram till 2024/2025. Resultat enligt tabell 7.

Tabell 7: Faktiska mot predikterade poäng för säsong 2019/2020

Lag	Säsong	Poäng	Pred. Poang	Residualer
Liverpool	19/20	99	78,29	20,70
Manchester City	19/20	81	98,44	−17,44
Chelsea	19/20	66	72,82	−6,82
Manchester United	19/20	66	67,89	−1,89
Leicester City	19/20	62	64,17	−2,17
Tottenham	19/20	59	46,12	12,87
Wolves	19/20	59	57,49	1,50
Arsenal	19/20	56	49,44	6,55
Burnley	19/20	54	47,15	6,84
Sheffield United	19/20	54	36,06	17,93
Southampton	19/20	52	52,30	−0,30
Everton	19/20	49	54,91	−5,91
Newcastle United	19/20	44	35,77	8,22
Crystal Palace	19/20	43	38,67	4,32
Brighton	19/20	41	48,01	−7,01
West Ham	19/20	39	43,30	−4,30

Aston Villa	19/20	35	42,93	−7,93
Bournemouth	19/20	34	40,66	−6,66
Watford	19/20	34	42,71	−8,71
Norwich City	19/20	21	32,48	−11,48

För att utvärdera modellens prediktionsförmåga beräknades MAE och RMSE på testdata. Ett lägre värde innebär en bättre prediktionsprecision. Modellen, enligt tabell 8, visade ett MAE på cirka 8 poäng och ett RMSE på cirka 10 poäng, vilket innebär att modellen i genomsnitt missar med omkring 8-10 poäng per lag, vilket speglar begränsningar i modellens förmåga att hantera den slumpkomponenten i fotbollens poängfördelning.

Tabell 8: RMSE och MAE på testdatan.

RMSE	9,67
MAE	7,98

6. Analys och diskussion

Resultaten från den reducerade regressionsmodellen visar att en begränsad uppsättning variabler kan förklara en stor del av variationen i lagens säsongspoäng i Premier League. Den justerade förklaringsgraden uppgår till cirka 0,862, vilket är högt i relation till fotbollens osäkerhet och slumpmässighet. Detta indikerar att säsongsutfall präglas av mer stabila och systematiska mönster, vilket överensstämmer med tidigare forskning på säsongsdata i professionell fotboll (Souza et al. 2019).

Det negativa interceptet i modellen ska inte tolkas som ett empiriskt meningsfullt värde i sig, utan som ett tekniskt justeringsled. Interceptet representerar det hypotetiska utfallet då samtliga förklarande variabler antar värdet noll, ett scenario som saknar praktisk relevans då variablerna aldrig observeras i närheten av noll. Att interceptet antar ett negativt värde beror därför på att modellen förklarar poängutfallet genom de inkluderade variablerna, samtidigt som koefficienterna skattas utifrån variablernas faktiska skalor och samvariation.

De offensiva variablerna som skott på mål och kvalitativa målchanser uppvisar båda positiva och statistiskt signifikanta samband med poängutfall. Detta är i linje med teoretiska förväntningar, då dessa mått fångar både volym och kvalitet i lagens målchanser. Resultaten stödjer tidigare empiriska studier som visat att grundläggande offensiv effektivitet är en relevant variabel för ligapoäng i professionell fotboll (Oberstone 2009). Att båda variablerna kvarstår som signifikanta i den reducerade modellen tyder på att de innehåller kompletterande informationsinnehåll, trots viss samvariation.

Andelen lyckade passningar visar ett positivt och signifikant resultat. Detta kan tolkas som att lag med hög passningssäkerhet tenderar att ha bättre kontroll över matchbilden och

därigenom ackumulera fler poäng över tid. Samtidigt bör variabeln betraktas som ett övergripande mått på spelkvalitet snarare än en isolerad mekanism, då variabeln sannolikt samvarierar med lagens tekniska nivå och taktiska struktur.

På den defensiva sidan är antalet misstag som leder till avslut eller mål för motståndaren negativt relaterat till poäng, vilket är intuitivt då dåligt försvar leder till förlust. Resultatet indikerar att även ett begränsat antal misstag kan få konsekvenser för säsongsutfallet. Även vunna tacklingar uppvisar ett negativt samband med poäng, vilket vid en första anblick kan förefalla kontraintuitivt. En möjlig tolkning är dock att ett högt antal tacklingar speglar defensivt spel snarare än defensiv kvalitet, vilket innebär att lag som försvarar sig nära eget mål eller utan boll tvingas till fler defensiva aktioner.

Variabeln som mäter den totala distans laget förflyttar sig under en säsong med boll uppvisar ett negativt och signifikant samband. Detta tyder på att hög fysisk belastning inte nödvändigtvis är förknippad med bättre resultat, utan snarare kan reflektera ineffektiv bollkontroll eller defensiv matchbild. Vilket innebär ett långt springande med boll när laget väl fått kontrollen och med detta måste hämta hem förlorat djup i egen planhalva. Resultatet stödjer uppfattningen att fysiska volymmått måste tolkas i sitt taktiska sammanhang för att vara en indikator på prestation.

Den variabel som mäter antalet spelare som använts under säsongen har en negativ och statistiskt signifikant koefficienten, vilket indikerar att lag som använder fler spelare i genomsnitt tenderar att ta färre poäng. Detta kan tolkas som att trupprotation speglar skador, bristande kontinuitet eller osäkerhet kring laguppställning. Resultatet ligger i linje med intuitionen att stabilitet och kontinuitet i laget är gynnsamt, samtidigt som sambandet inte nödvändigtvis är kausalt. Det är exempelvis möjligt att svagare prestationer leder till ökad rotation snarare än tvärtom.

Lönekostnader uppvisar ett positivt och signifikant samband, vilket indikerar att ekonomiska resurser fortsatt är en faktor för framgång i fotboll. Variabeln kan tolkas som ett övergripande mått på kvalitet samt möjligheten att attrahera och behålla högpresterande spelare. Resultatet

överensstämmer med tidigare studier som visat att finansiell styrka är kopplad till prestation i europeisk toppfotboll, även om sambandet indirekt är kopplat till spelet på planen (Souza et al. 2019).

Multikollinearitetsanalysen visar att den reducerade modellen har lägre VIF-värden jämfört med den fulla modellen, vilket stärker tillförlitligheten i parameterskattningarna. Samtidigt kvarstår viss samvariation mellan offensiva variabler, vilket är svårt att undvika i prestationsdata där flera mått beskriver närliggande aspekter i fotboll.

Den genomförda utvärderingen visar att modellen i genomsnitt avviker med cirka 8–10 poäng per lag. Detta indikerar att modellen fångar övergripande strukturella mönster i säsongsdata men har begränsad precision när det gäller att förutsäga enskilda lags exakta poängutfall. Resultatet är i linje med Kundu et al. (2021), som betonar att slumpmässiga inslag och icke observerbara faktorer fortsatt spelar en betydande roll, även vid användning av avancerade statistiska metoder. Analysen visar att regressionsmodellen ur ett teoretiskt perspektiv fungerar för att beskriva samband på säsongsnivå, men att resultaten bör tolkas som deskriptiva snarare än kausala.

6.1 Begränsningar i analysen

Analysen bygger på säsongsdata, vilket innebär att variation inom säsonger, såsom skador, formsvackor och taktiska förändringar, inte fångas upp. Detta kan påverka resultaten eftersom sådana faktorer kan ha stor betydelse för hur många poäng ett lag tar, trots att de inte syns i säsongsgenomsnitt. Panelen är dessutom obalanserad på grund av upp- och nedflyttning mellan säsonger och uteblivna variabler. Nyuppflyttade lag skiljer sig från etablerade lag när det gäller resurser och spelartrupp. Vidare kan förklarande faktorer saknas i modellen, exempelvis tränarbyten, skador på nyckelspelare och deltagande i europeiska turneringar. Om sådana faktorer samvarierar med inkluderade variabler kan detta leda till snedvridna skattningar. Slutligen var målet en enkel modell för prediktion snarare än kausal tolkning. Resultaten bör därför tolkas som statistiska samband och inte som direkta orsakssamband mellan de inkluderade variablerna och säsongspoäng.

7. Slutsatser

Syftet med studien var att undersöka i vilken utsträckning en begränsad uppsättning variabler kan förklara variationen i säsongspoäng i Premier League samt att utvärdera modellens beskrivande och prediktionsegenskaper. Resultaten visar att detta är möjligt i relativt hög grad, trots sportens inneboende osäkerhet.

Studien visar att offensiva prestationsmått såsom skott på mål och kvalitativa målchanser har stabila och statistiskt signifikanta samband med poängutfall över en säsong, vilket inte är revolutionerande då mål avgör resultatet. Även passningsprecision och defensiva indikatorer bidrar till att förklara variationen i lagens prestationer. Vidare indikerar resultaten att lag med hög truppstabilitet, mätt genom ett lägre antal använda spelare under säsongen, tenderar att prestera bättre än lag med omfattande rotation. Slutligen kan ekonomiska resurser, mätt genom lönekostnader, ha fortsatt betydelse för sportslig framgång i fotboll även när prestationsmått inkluderas i modellen. Detta understryker att framgång kräver ekonomiska förutsättningar för att förklara poängutfallet på en säsong.

Samtidigt visar out-of-sample utvärderingen att modellens precision är begränsad, vilket betonar att en betydande del av variationen i poängutfall inte kan förklaras av observerbara variabler. Fotbollens slumpmässiga inslag, såsom skador, taktiska förändringar och matchspecifika händelser, utgör fortsatt begränsningar för empirisk modellering. Studien bidrar till den befintliga litteraturen genom att visa att relativt enkla regressionsmodeller fortfarande kan fånga mönster i Premier Leagues säsongsutfall, men att resultaten bör tolkas med försiktighet.

7.1 Fortsatt forskning

Vidare forskning skulle kunna utveckla analysen genom att i större utsträckning ta hänsyn till skillnader mellan lag och säsonger över tid. Genom att använda modeller som bättre fångar återkommande mönster kan man tydligare skilja mellan stabila lagrelaterade egenskaper och

förändringar i prestation från en säsong till en annan. Slutligen kan modellens giltighet prövas genom att inkludera fler ligor eller ett längre tidsspann, vilket skulle möjliggöra en bredare jämförelse.

8. Referenslista

8.1 Vetenskapliga artiklar

Kundu, T., Choudhury, A.R. & Rai, S. 2021. *Predicting English Premier League Matches Using Classification and Regression.* Bhattacharyya, K., Dey, N. & A.S. Ashour (red.). Computational Intelligence in Pattern Recognition. Singapore: Springer, s. 555-568. DOI: https://doi.org/10.1007/978-981-15-5077-5_50

Oberstone, J. 2009. Differentiating the Top English Premier League Football Clubs from the Rest of the Pack Using Pitch Actions. *Journal of Quantitative Analysis in Sports,* 5(3). DOI: https://doi.org/10.2202/1559-0410.1183

Oberstone, J. 2011. Comparing Team Performance of the English Premier League, Serie A, and La Liga for the 2008–09 Season. *Journal of Quantitative Analysis in Sports,* 7(3). DOI: https://doi.org/10.2202/1559-0410.1280

Souza, D.B., López-Del Campo, R., Blanco-Pita, H., Resta, R. & Del Coso, J. 2019. A new paradigm to understand success in professional football: Analysis of match statistics in LaLiga for 8 complete seasons. *International Journal of Performance Analysis in Sport,* 19(4), s. 543–555. DOI: https://doi.org/10.1080/24748668.2019.1632580

8.2 Böcker

Wooldridge, J.M. (2019). *Introductory Econometrics: A Modern Approach*. 7:e uppl. Boston: Cengage Learning

8.3 Datakällor

FBREF (u.å.) *Squad Standard Stats, Premier League.* https://fbref.com/en/comps/9/stats/Premier-League-Stats [2025-11-17]