Stockholms universitet

En publiksiffra, multipla förklaringar: en statistisk analys av publikantalet på AIK fotbolls hemmamatcher

Robert Holmlund

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:23
Matematisk statistik
Augusti 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:23**
http://www.math.su.se/matstat

En publiksiffra, multipla förklaringar: en statistisk analys av publikantalet på AIK fotbolls hemmamatcher

Robert Holmlund*

Augusti 2016

Sammanfattning

Vi har i det här examensarbetet valt att samla in information beträffande AIK Fotbolls hemmamatcher, i syfte att finna variabler som kan tänkas ha inverkan på publiksiffran. Vi valde från det insamlade datamaterialet ut 10 tänkbara förklarande variabler, som vi sedan kom att utöka till 16. Med dessa som utgångspunkt anpassade vi olika multipla linjära regressionsmodeller, där slutligen en valdes ut mest lämpad för ändamålet. Enligt den modellen var 9 av de 16 variablerna betydande för utgången av publikantalet. Det visade sig att såväl regn som samhällsekonomi kan ha inverkan på hur många åskådare som väljer att gå på en match. Vi fann även mer förväntade variabler som påverkade publiktillströmningen, såsom motståndarlag och sportslig framgång. Slutligen testade vi modellens prediktiva förmåga i praktiken. Vi lät alltså den framtagna modellen försöka prediktera publiksiffran till några matcher som ännu inte hade spelats då datamaterialet samlades in.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: robbanholmlund@gmail.com. Handledare: Jan-Olov Persson och Gudrun Brattström.

Abstract

In this text we have gathered data regarding the home games of a football team in the Swedish Allsvenskan to find variables that may affect attendance. From the collected data we chose ten possible explanatory variables, and later expanded them to 16. With these as a starting point we adjusted different multiple linear regression models, of which one finally was deemed most suitable. By that model 9 of the 16 variables affect the outcome of attendance. It appeared that rain as well as social economics may have an impact on how many fans attend a certain game. We also found variables that impacted attendance that were expected such as opposing team and previous success of the home team. Finally, we tested the predictive capacity of the chosen model in practice. That is, we let it predict attendance for games that were played after the data was gathered.

Förord

Denna uppsats utgör ett självständigt arbete om 15 hp och leder till en kandidatexamen i matematisk statistik. Jag vill rikta ett stort tack till mina handledare Gudrun Brattström och Jan-Olov Persson för goda råd och snabba återkopplingar.

Innehåll

1 Introduktion

Svensk fotboll är knappast känd för dess sportsliga framgångar och spelmässigt höga nivå, snarare tvärtom. Men det finns en sak inom den svenska fotbollen som är i världsklass och omtalad runt om i världen, nämligen supporterkulturen. Trots att Sveriges inhemska liga Allsvenskan inte är bättre rankad än 39:e plats i världen (IFFHS, 2015), så anser många att Allsvenska fotbollsfans är bland världens absolut mest hängivna. [4] Från denna fascinerande företeelse kan man bli nyfiken på de svenska fotbollssupportrarna och vad det är som driver dem. Hur kommer det sig att det en blöt och ruskig måndag i november infinner sig 40 000 fans för att titta på en fotbollsmatch, medan samma fans tre månader tidigare, en varm och solig sensommardag, var 30 000 färre på en liknande match, även fast förhållanderna kan tyckas ha varit betydligt mer gynsamma för att sitta utomhus och kolla på en match då?

I det här examensarbetet tänkte vi titta närmare på just publiksiffror, och om det finns några underliggande faktorer som påverkar hur mycket publik det kommer på ett elitfotbollslags hemmamatcher. För att utforska detta närmare har vi i den här studien valt att undersöka publiksiffran för ett av Sveriges mest folkkära fotbollslag, AIK. För det har vi samlat in information om bl.a. publiksiffra, väder, veckodag och tabellplacering för AIKs Allsvenska hemmamatcher mellan åren 2004-2015. Vi kommer med hjälp av datamaterialet försöka hitta förklarande variabler till varför publiksiffrorna varierar, och anpassa en multipel linjär regressionsmodell som på ett bra sätt beskriver data. Vi tänker också att vi slutligen mer eller mindre på något vis ska testa prediktera framtida matchers publiksiffra.

2 Datamaterial

2.1 Insamling av data

Datamaterialet är hämtat från AIK Fotboll statistikdatabas, och består av 127 observationer med 11 variabler från AIKs Allsvenska hemmamatcher mellan åren 2004-2015, exklusive år 2005. AIK spelade då inte i Allsvenskan utan i divisionen under, Superettan, och dessa observationer ansågs därför inte passa in med resterande data. Anledningen till den valda inspektionsperioden är för att vi tänker oss att ju närmare vår egen tid desto mer relevanta blir resultaten. Gällande längden på perioden så resonerar vi som så att även om många observationer är bra ur ett statistisk perspektiv, så vill vi heller inte välja en allt för lång tidsperiod, eftersom olika tidsskeden kan ha olika bidragande faktorer till varför folk väljer att gå på fotboll. Till exempel är inte graden av aktivitet på Instagram tillämpligt på data från 90-talet, eller bara den fundamentala faktorn att fotbollsintresset varierar mellan olika tidsepoker, vilket är något vi varken kan direkt mäta eller använda för vårt ändamål. [5]

Vidare valde vi även att exkludera data med alltför uppenbara och redan välkända påverkande faktorer, så som stockholmsderbyn, säsongspremieärer och guldavgörande matcher. Detta för att undvika att dessa observationer skulle ta för stor plats i analysen och överskugga de mer anonyma och okända förklarande variablerna, som vi i den här studien främst är intresserade av. Ingen skulle direkt höja på ögonbrynen om vi la fram slutsatsen att när AIK möter Djurgården verkar det komma mer publik. Således har observationer från exempelvis alla matcher där motståndarlaget varit ett stockholmslag eller hemmapremiärmatcher inte tagits med i datamaterialet.

2.2 AIK Fotbolls Statistikdatabas

Statistikdatabasen är konstruerad av AIK Fotboll tillsammans med ideella aktörer, och innehåller statistik från alla matcher och alla spelare genom AIK Fotbolls historia mellan 1896-08-08 till 2015-11-08. Databasen innehåller just nu 3 858 AIK-matcher i 239 olika sammanhang inför 26 036 931 åskådare. Autenticiteten i datan beskriver de själva såhär:

”AIK ansvarar inte för korrektheten i denna databas. AIK arbetar för att hålla materialet så korrekt som möjligt, men faktafel kan förekomma.” [5]

2.3 Beskrivning av variabler

Från den information vi funnit om varje match så har vi valt ut 11 variabler, en responsvariabel och 10 förklarande variabler att utgå ifrån i vår

analys. All information, så som väder, datum, motståndarlag m.m. är insamlad från AIK Fotbolls Statistikdatabas, förutom barometerindikatorn. Den är hämtad från www.ekonomifakta.se. [3] En översiktlig beskrivning av varje variabel följer nedan.

Publiksiffra. Den variabel som vi alltså kommer att använda som responsvariabel. Antalet uppräknade personer som gått in på arenan. Varierar mellan 7 420 och 30 999 personer, med ett medelvärde på 14 250. Eftersom vi inte har någon observation där maxgränsen uppnåtts, dvs. det har inte i vårt datamaterial förekommit en match där arenan varit fullsatt, så tar vi inte med i beräkningarna de facto att responsen kan påverkas av en övre gräns. (AIKs hemmaarena mellan 2004-2012, Råsunda Stadion, hade en kapacitet på 35 000 åskådare. Men framförallt har AIKs nuvarande arena sedan 2013, Friends Arena, en kapacitet på 50 000 åskådare. Alltså skiljer det i nuläget nästan 20 000 mellan maxgränsen och den högst observerade publiksiffran de senaste 10 åren i vårt datamaterial. Så vi känner oss ganska trygga med att förbigå den existerande övre gränsen.)

Datum. Matchdagens datum. Sträcker sig tidsmässigt från den först noterade matchen 2004-04-04 till den sista 2015-10-04. Vi vet att publiksnitt kan variera med tiden och stort mellan olika tidsepoker historiskt sett. [1] Därför tycker vi oss ha anledningen att ha med en tidsvariabel i analysen.

Tid. Tid på dygnet då matchen startar. Där den tidigaste matchen i vårt datamaterial startade kl. 12.30, och den senaste kl. 20.00. Den vanligaste tidpunkten för en matchstart är kl. 19.00, ungefär 51% av matcherna startar detta klockslag. Vi ställer oss frågan ifall vissa tider är mer attraktiva än andra i syftet att locka publik. Kanske kan vi hitta indikationer på att folk i allmänhet tycker att matchstart 18.00 en vardag anses vara för tidsknapp för att hinna dit efter jobbet? Eller tvärtom, att 20.00 på en vardag är för sent. Det är ju exempelvis många yngre barn som tillsammans med någon vuxen går på matcherna, som kanske stannar hemma ifall matcherna går försent.

Framgång_kort. Sammanlagt poängmässigt resultat från de två senast spelade matcherna innan den aktuella matchen. I en Allsvensk fotbollsmatch kan ett lag erhålla 0 poäng (förlust), 1 poäng (oavgjort) eller 3 poäng (vinst) från en match och följaktligen varierar denna variabel mellan 0 och 6. På så vis fungerar variabeln som ett mått på kortsiktig sportslig framgång. För att denna variabel ska kunna användas måste alla matcher vi tar med i datamaterialet ha två föregående matcher den aktuella säsongen, och därför var vi tvungna att utelämna hemmamatcher från omgång 2. På grund av detta gick vi miste om tre observationer när vi samlade in data från åren 2004 till 2015, då det faktiskt endast var tre hemmamatcher som spelades omgång 2. Denna variabel har vi med hjälp av erhållen information från

AIK Fotbolls Statistikdatabas skapat själva. Vi tänker att det är rimligt att anta att ett lags sportsliga framgångar påverkar intresset hos sina fans och därmed publiktillströmningen. Vi ville därför försöka skapa ett mått på kortsiktig sportlig framgång och ansåg detta som en fungerande lösning.

Framgång_lång. Den rådande tabellplacering AIK har inför matchen. Tabelläget i datamaterialet varierar mellan 1 och 14 fram till och med år 2007, och därefter mellan 1 till 16, då Allsvenskan 2008 utvidgade ligan till 16 istället för 14 lag. AIKs genomsnittliga tabellplacering under den aktuella tidsperioden är ungefär 6,1. Av samma skäl som vi angav för *Framgång_kort* motiverar vi valet av denna variabel. Gynnar till exempel en aktuell hög tabellplacering (dvs. låg siffra) publiksiffran? Vi tror att det är högst troligt. Denna kan ses som ett mått på långsiktig sportslig framgång i relation till variabeln *Framgång_kort*. Observera att en låg numerisk siffra innebär en hög tabellplacer och tvärtom. Således skulle en negativ koefficient för denna variabel innebära mer publik för en högre placering.

Omgång. Omgången för matchen. Denna variabel varierar mellan 3 och 30 i vårt datamaterial, eftersom vi valt att exkuldera hemmapremiärmatcher och avslutande matcher, samt även matcher från omgång 2 på grund av variabeln *Två senaste matcher* som vi tidigare nämnt. Vi tänker oss att intresset kan skifta under säsongen. Kanske lockar slutskedet av serien generellt mer publik än mitten?

Konjunktur. Barometerindikatorn aktuella värde den månad matchen spelas. Barometerindikatorn är skapad av konjunkturinstitutet och mäter stämningsläget i ekonomin och ger en bild av det aktuella konjunkturläget. Barometerindikatorn har ett historiskt genomsnitt på 100. Om indikatorn ligger över 100 är ekonomin för närvarande bättre än normalt och om den ligger under 100 är ekonomin svagare än normalt. Under vår tidsperiod för vår data varierar denna indikator mellan 71,5 och 123,1 med ett genomsnittligt värde på 106,2. Är fansen mer återhållsamma med sin konsumtion av biljetter vid lågkonjunktur och vice versa? [3]

Väder. Vädret den gällande matchdagen, registrerat som regn, regnskurar, molnigt, inomhus, halvklart eller soligt. Kategorinivån *inomhus* innebär att man stängt taket på Friends Arena, och åskådarna påverkas således inte av vädret under matchen. Detta alternativ fanns inte på Råsunda Station. Den mest påträffade väderleken i vårt datamaterial är soligt, som förekommer vid ungefär 46,5% av de observerade matcherna. Vi tycker det är intressant ifall vädret har någon inverkan på publiksiffran. Väljer exempelvis folk att stanna hemma om det är regnigt och kallt? Det vore i våra tankebanor en ganska rimlig företeelse.

Temperatur. Temperaturen utomhus vid matchstart, mäts i grader celsius och varierar i vårt datamaterial mellan 2 och 29 grader med ett medelvärde på ungefär 15 grader. Likt motiveringen till variabeln *Väder*, eftersom nästan alla matcher spelas i utomhusklimat, och eftersom vädret då påverkar komforten, är det inte otänkbart att extrema temperaturer kan påverka folks val att gå på match.

Veckodag. Veckodagen då matchen spelas. I vårt datamaterial förekommer alla veckans dagar, där söndagar är klart vanligast, och förekommer vid ungefär 47% av observationerna. Är vissa veckodagar bättre anpassade för att locka publik än andra? Anser supportrarna i allmänhet att de har mer tid till att gå på en match på en söndag än en måndag? Ifall det vore fallet, blir ju variabeln intressant i den mening att AIK själva skulle kunna argumentera för att lägga matcherna på de mer publik-attraktiva dagarna.

Motståndarlag. Motståndarlaget för matchen. Som vi tidigare påpekat är det redan välkänt att vissa motståndarlag, framförallt andra stockholmslag, har en betydande påverkan på publiksiffran, och som vi därför valt att inte ta med i datamaterialet. Men det vore ju intressant att undersöka ifall andra lag som är mindre dragplåster som motståndare lockar ungefär lika mycket publik sinsemellan, eller om det finns tydliga skillnader även där. Denna variabel avser de resterande lagen utöver Djurgårdens IF och Hammarby IF, och består av 22 olika lag.

3 Teori - Multipel linjär regression

3.1 Definition

Om det är troligt att en responsvariabel påverkas av två eller flera förklarande variabler kan modellen multipel linjär regression vara lämplig. Den allmänna formeln för modellen defineras som

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + \varepsilon_i, \quad i = 1, ..., N, \tag{1}$$

där Y_i är responsvariabeln och N är antalet observationer. Denna uttrycks med hjälp av de $m + 1$ okända parametrarna $\beta_0, \beta_1, ..., \beta_m$ och de m olika förklarande variablerna $x_1, ..., x_m$. Termen ε_i är en stokastisk variabel och kallas feltermen, och är med andra ord differensen $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi})$. Den antas vara normalfördelad med väntevärde 0 och sinsemellan oberoende med konstant varians, vilket kan uttryckas som

$$\varepsilon_i \sim N(0, \sigma^2). \tag{2}$$

[6]

3.2 Dummyvariabel

Ifall det förekommer kvalitativa variabler i datamaterialet kan dessa inkluderas i regressionen genom dummyvariabler, även kallat indikatorvariabler. En dummyvariabel kan endast anta värdet 0 eller 1 och används på så vis numeriskt för att koda olika kategorinivåer som observationerna är uppdelade i. Ifall den kvalitativa variabeln endast kan anta två värden, exempelvis kön, så kan en dummyvariabel skapas som antar värde 1 om det är en man och värde 0 ifall det är en kvinna. Men i många fall så kan en kvalitativ variabel anta fler än två värden, och då behöver man skapa flera dummyvariabler. Till exempel en variabel som kan anta de tre olika värdena vinst, oavgjort eller förlust. Grundregeln är då att man skapar $k - 1$ dummyvariabler, där k är antalet värden variabeln kan anta, och således används ett av värdena som referens. Vilket värde man vill använda som referens är valfritt. I vårat exempel skulle det allstå se ut på följande vis

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

där

$$x_1 = \begin{cases} 1 & vinst \\ 0 & \text{f.ö.} \end{cases}, \qquad x_2 = \begin{cases} 1 & \text{förlust} \\ 0 & \text{f.ö.} \end{cases},$$

där då basnivån eller referensnivån här representeras av värdet oavgjort. [6] [2]

3.3 Stegvis variabelselektion

Till en början vid en multipel regressionsanalys kan man besitta en stor uppsättning förklarande variabler utan att veta vilken delmängd av variablerna som är mest relevanta för ändamålet och vilka av variablerna som kan anses onödiga. För att då bespara sig tiden och beräkningsarbetet med att testa varje möjlig kombination av x-variabler så finns det olika stegvisa proceduraer man kan nyttja. Med hjälp av stegvis variabelselektion väljs de variabler ut som tillsammans bidrar till den mest lämpade modellen genom att stegvis inkludera eller exkludera en variabel i taget tills ett visst stoppkriterium är uppfyllt. Man ska dock komma ihåg att dessa procedurer genomförs av datorer och att det finns inga garantier att det är den mest optimala eller korrekta modellen som tar sig ut på andra sidan. Man bör istället se urvalet som en bra utgångspunkt för fortsatta undersökningar av variablerna. Vi går nedan igenom de tre vanligaste procedurerna, och dem som vi kommer använda i denna studie.

Backward elimination
Proceduren utgår från en modell som innefattar samtliga variabler. I första steget elimineras den variabel som enligt ett visst kriterium tjänar modellen sämst. Detta förfarande fortsätter tills modellen inte förbättras mer enligt ett visst stoppkriterium.

Forward selection
För denna procedur utgår man endast från interceptet för att sedan utvidga modellen med en variabel i taget. I varje steg adderas den variabel som vid inkludering i modellen anses förbättra modellen mest, enligt ett på förhand valt kriterium. Proceduren fortsätter tills det att modellen enligt det valda kriteriet inte kan bli bättre av att addera ytterligare variabler.

Stepwise regression
Denna procedur är en kombination av de två föregående algoritmerna, och är förmodligen den mest använda av de tre. Metoden utgår antingen från endast intercept som *forward selection* eller samtliga variabler som *backward elimination*. Det kan man välja själv. Beroende på vilken utgångpunk man valt, så antingen inkluderar eller exkluderar proceduren en variabel i första steget, utifrån något valt kriterium. I stegen som sedan följer så testar metoden både att att inkludera och exkludera alla variabler och väljer den variabel som förbättrar modellen mest efter det valda kriteriet. Detta fortsätter tills stoppkriteriet anser att modellen inte kommer förbättras mer. [6]

Vi har i den här studien valt att arbete med programspråket R. Vid användandet av de ovan beskrivna procedurerna så använder R stoppkriterium AIC (Aka-

ike Information Criterion). Procedurerna forsätter, enligt beskrivning, så länge som AIC-värdet blir lägre vid varje korrigering, annars stoppas proceduren. AIC uttrycks matematiskt som

$$AIC = -2(\text{maximum loglikelihood} - \text{antalet parameterar i modellen}), \quad (3)$$

och mäter den relativa kvalitén på en modell. Den jämför alltså modellerna i till exempel en variabelselektion med varandra, där den med lägst AIC anses mest lämpad. AIC används på så vis som ett jämförelsemått mellan modeller och är inget allmänt mått på hur bra anpassad en modell är. [2]

3.4 Transformationer

Det är inte helt ovanligt att sambandet mellan responsvariabeln och några förklarande variabler inte är linjärt, eller att residualerna saknar konstant varians. Båda problemen kan i vissa fall lösas med lämplig tranformation av en eller flera variabler. En transformation kan utföras på många olika sätt och en av de mest använda är logaritmering. Just transformation genom logaritmering är något vi kommer att testa under vår modellering för att försöka uppnå bättre resultat. Man kan i sin modell välja att transformera de eller den variabel man anser gynna modellen bäst. [6]

3.5 Modellkontroll

För att undersöka om en viss multipel regressionsmodell är lämplig för data finns många metoder och plottar man kan nyttja. Framförallt bör man kontrollera så att modellantagandena är uppfyllda, alltså de antagandena som beskrivs av (2). Är dessa inte approximativt uppfyllda är modellen inte pålitlig. Vi går följande igenom några av dessa som vi hade särskild stor nytta av i den här studien.

Ett av antagandena för modellen (1) är att residualerna (feltermerna) har konstant varians. Ett vanlig sätt att undersöka detta är att plotta residualerna mot de skattade värdena av responsvariabeln, eller mot de förklarande variablerna var för sig. Det vi vill se är att residualerna är jämnt utspridda kring 0. Man bör vara observant och spana efter tecken på systematik i residualerna. Till exempel om spridning för residualerna ökar med ett ökande värde på en förklarande x-variabel. I det fallet har inte feltermen konstant varians, vilket är en förutsättingen vi antagit för att modellen ska vara riktig. Ett annat viktigt antagande är att residualerna är normalfödelade med väntevärde 0. En metod för att undersöka det, är genom att plotta en så kallad normalfördelningsplott. I en normalfördelningsplott plottas residualerna mot normalfördelningens teoretiska kvantiler. Punkterna ska då uppträda på en rak linje om dessa är normalfördelade. I praktiken eftersträvas såklart

en approximativt rak linje, ty få om något dataset är exakt normalfördelade.

Ett annat problem som kan uppstå är multikollinearitet, vilket innebär att man har approximativt linjära samband mellan förklaringsvariablerna. Detta kan medföra komplikationer som att variabler felaktigt blir icke-signifikanta. En metod för att upptäcka multikollinearitet, är med hjälp av ett statistiskt mått kallat *VIF* (Variance Inflation Factor). VIF-faktorn beräknas som

$$VIF = \frac{1}{1 - R_j^2}, \tag{4}$$

där R_j^2 är förklaringsgraden för en multipel regressionsmodell med variabeln x_j som responsvariabel, och de andra x-variablerna som förklarande variabler. VIF-faktorn uttrycker hur mycket variansen för en skattad regressionskoefficient $\hat{\beta}_j$ ökar i en regressionsmodell med de andra x-variablerna som förklarande variabler jämfört med en modell där x_j hade varit ensam. Från formel (3) ser vi att $VIF = 1$ om förklaringsgraden $R_j^2 = 0$, dvs. att x_j är ortogonal mot de övriga förklarande variablerna. Gränsen för vilket VIF-värde man anser vara acceptabelt brukar sättas vid 5 eller 10. Vi kommer i denna studie välja VIF-värde 10 som gräns. [6]

4 Analys

Syftet med det här arbetet och den här analysen är att hitta en mutipel linjär regressionmodell som på så bra sätt som möjligt förklarar och även predikterar publiksiffran för AIK fotbolls hemmamatcher. Vi har för den här studien valt att arbeta med programspråket R.

4.1 Modifiering av variabler

Vi ska nu börja med att se över variablerna från datamaterialet, ifall det skulle vara lämpligt att modifiera några av dem. Nedan tar vi upp de variabler vi väljer att göra justeringar på.

Datum. Vi väljer att dela upp denna variabel i årstider och årtal. Detta gör vi för att utvidga undersökandet för betydelsen av tiden. Vi tänker oss att det är möjligt att det exempelvis är mer populärt att gå på fotboll på hösten än på sommaren. Det kan vi nu med den nya variabeln *Årstider* undersöka. Vidare vet vi att publiksnitt skiftar med tiden, historiskt sett. Däremot tror vi inte att detta sker månadsvis, utan vi tänker oss att sådana förändringar sker under längre tidsperspektiv. Därför komprimerar vi hela datum till årtal. Det är dessutom tekniskt enklare att räkna med årtal än hela datum. [1]

- **Årstid**. Kvalitatativ variabel innehållande klasserna *Vår*, *Sommar* och *Höst*. Eftersom denna variabel är kategorisk med tre klasser så behöver vi skapa två dummyvariabler. Dessa är:

  - *Vår*. Antar värdet 1 om matchen spelas på våren, annars 0.

  - *Sommar*. Antar värdet 1 om matchen spelas på sommaren, annars 0.

  Höst blir då basnivån, och representeras ifall båda de ovannämnda variablerna antar värdet 0.

- **Årtal**. Varierar numeriskt mellan 4 och 15. Siffran utgör antal år efter år 2000, dvs. 4 = 2004, 6 = 2006, 7 = 2007 och så vidare. Notera att siffran 5 ej finns med eftersom att AIK 2005 inte spelade i Allsvenskan.

Veckodag. Denna variabel är kvalitativ och innehar alla veckans sju dagar. Vi upptäcker dock, att det i vårt datamaterial vanligast förekommande dagarna då matcher spelas är måndagar (22%), lördagar (15%) och söndag (47,2%), som tillsammans utgör 84.2% av observationerna. Därför väljer vi att slå samman de övriga kategorinivåerna *Tisdag*, *Onsdag*, *Torsdag* och *Fredag*, som tillsammans utgör ungefär 16% av observationerna, till ett och

samma värde som vi benämner *Övriga*. Återigen har vi en kategorisk variabel och behöver därför skapa tre stycken dummyvariabler, eftersom denna har fyra värden. Vi väljer att använda det nya värdet *Övriga* som referensnivå. Dummyvariablerna blir då:

- *Måndag*. Antar värdet 1 om matchen spelas på en måndag, annars 0.

- *Lördag*. Antar värdet 1 om matchen spelas på en måndag, annars 0.

- *Söndag*. Antar värdet 1 om matchen spelas på en måndag, annars 0.

När alla dessa tre antar värdet 0 representeras övriga dagar.

Väder. Denna variabel utgörs av klasserna *Soligt*, *Halvklart*, *Inomhus*, *Molnigt*, *Regnskurar* och *Regn*. Vi väljer här att slå ihop flera klasser, delvis på grund av att vi har en väldigt ojämn fördelning av observationerna på de olika klasserna. Vissa klasser har endast en eller ett fåtal observationer medan exempelvis nästan 79% av observationerna noteras av klasserna *Soligt* och *Halvklart*. Dessutom tänker vi oss att om vädret ska påverka någon att gå på en match eller ej, så torde det framförallt vara ifall det regnar eller ej. Vi föreställer oss att en del av fansen möjligtvis drar sig för att sitta utomhus och bli blöta och därför väljer att stanna hemma när det regnar. Det är rimligtvis ensamt den väderlek som borde påverka mest ifall vädret har någon inverkan. Vi bildar därför två nya värden av de ursprungliga sju, en för när det regnat och en för när det inte regnat. De ursprungliga värden som kommer ingå i det nya värdet *Regn* är *Regnskurar* och *Regn*. Resterande fem tidigare kategorinivåer kommer att ingå i det nya värdet *Ej_regn*. Det visar sig att genomsnittliga publiksiffran när det regnar är 13 890, och när det inte regnar är 14 293. Så våra föraningar verkar inte vara helt fel ute. Denna variabel är kvalitativ med två klasser och vi skapar således en dummyvariabel:

- *Regn*. Antar värdet 1 ifall det regnat under matchen, annars 0.

Ej_regn fungerar då som referens.

Motståndarlag. Denna kvalitativa variabel består av 22 olika värden, varje värde representerar ett motståndarlag. För att förenkla kommande beräkningar och tolkningar så vill vi minska utbudet av olika värden. Av de 22 lag som förekommer i vårt dataset så är det känt (för de hyfsat fotbollskunniga) att det är tre lag som står ut något från de andra. Det handlar om lagen: Malmö FF, IFK Göteborg och IF Elfsborg, som anses som storlag i Allsvenska mått mätt, och därför rivaler till AIK. Vi väljer därför att bunta ihop de övriga 19 lagen till en ny kategorinivå som vi kallar för *Övriga*. Således har vi nu

istället fyra olika värden att jobba med istället för 22. Även här verkar den nya uppdelningen vara befogad, då det visar sig att medelvärdena för våra nuvarande kategorinivåer är 20 102 *(Malmö FF)*, 16 968 *(IFK Göteborg)*, 15 408 *(IF Elfsborg)* och 13 261 *(Övriga)*. Kollar man dessutom på de övriga lagens inviduella medelväden så figurerar de mellan ungefär 11 500 och 14 000 åskådare. För de kvarvarande fyra värdena bildar vi de tre nedanstående dummyvariablerna.

- *Malmö FF*. Antar värdet 1 om motståndarlaget för matchen är Malmö FF, annars 0.

- *IFK Göteborg*. Antar värdet 1 om motståndarlaget för matchen är IFK Göteborg, annars 0.

- *IF Elfsborg*. Antar värdet 1 om motståndarlaget för matchen är IF Elfsborg, annars 0.

Kategorinivån *Övriga* får då stå för basnivån, som blir närvarande när de tre övre dummyvariablerna alla antar vädet 0.

4.2 Samspel

Det kan existera samspel mellan de förklarande variablerna som påverkar responsen. Det kan dock vara svårt att finna vilka dessa samspel är när variabelutbudet är av vår storlek. Vi har definitivt inte möjlighet att testa alla möjliga kombinationer, vilket är väldigt många. Därför nöjer vi oss med att testa de samspel som vi intuitivt tänker oss skulle kunna ha en betydande inverkan på publikantalet. Vi kommer fram till fem interaktionstermer och de presenteras här under.

Framgång_kort × **Framgång_lång**. Vi tänker oss att kombinationen av kortstiktig framgång och långsiktig kan ha inverkan på publiksiffran. Till exempel kanske fansen triggas extra mycket att sluta upp kring matcherna ifall AIK har vunnit de två senaste matcherna och samtidigt har en hög placering.

Framgång_lång × **Omgång**. Det borde vara troligt att sambandet av vilken placering laget har, när under säsongen påverkar publiktillströmningen. Anta exempelvis att AIK har en hög tabellplacering i slutet av serien och har chans på att vinna guld. Det borde högst sannolikt öka fansens intresse och därav vore det troligt att fler ansluter till matcherna. Eller anta tvärtom, om det är ett dött lopp de sista 5-6 matcherna, då kommer det förmodligen färre fans än normalt.

Temperatur × **Väder**. Kan kombinationen regn och dessutom kallt avskräcka fler? Vi tänker oss att möjligheten finns.

Framgång_lång × **Motståndarlag**. När ett så kallat storlag är på besök lockas ofta storpublik. Men det känns rimligt att anta att kombinationen av till exempel en sportsligt framgångsrik säsong och ett rivalmöte ger en extra skjuts på åskådartillströmningen.

Framgång_lång × **Konjunktur**. Kan det vara så att när det går bra för AIK under en säsong, så ansluter folk som normalt sett inte är så hängivna och inte brukar gå på matcherna, i större utsträckning om ekonomin är god än om inte?

Vi känner att vi nu gjort det vi på förhand kan göra med de förklarande variablerna, och är redo att forstätta framåt i analysen.

4.3 Korrelation mellan förklarande variabler

De variabler vi nu har som underlag sammanställs i tabell 1 nedan.

Tabell 1: Underlaget av variabler

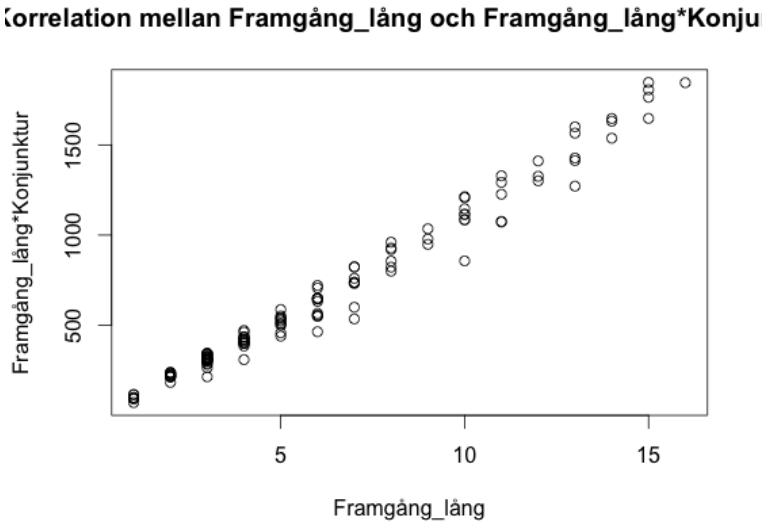
Variabel		**Beskrivning**
Publiksiffra		Responsvariabel.
Årtal		Årtalet -2000 då matchen spelas.
Tid		Tidpunk på dygnet.
Omgång		Spelomgång för matchen.
Konjunktur		Indikatorbarometern aktuell månad.
Temperatur		Temperatur vid matchstart.
Framgång_lång		Tabellplacering inför match.
Framgång_kort		Poäng två senaste matcherna.
Väder	Regn	Dummyvariabel.
	Ej regn	Basnivå.
Veckodag	Måndag	Dummyvariabel.
	Lördag	Dummyvariabel.
	Söndag	Dummyvariabel.
	Övriga	Basnivå.
Årstid	Vår	Dummyvariabel.
	Sommar	Dummyvariabel.
	Höst	Basnivå.
Motståndarlag	Malmö FF	Dummyvariabel.
	IFK Göteborg	Dummyvariabel.
	IF Elfsborg	Dummyvariabel.
	Övriga	Basnivå.
Framgång_kort	× **Framgång_lång**	Samspelsterm.
Framgång_lång	× **Omgång**	Samspelsterm.
Temperatur	× **Väder**	Samspelsterm.
Framgång_lång	× **Motståndarlag**	Samspelsterm.
Framgång_lång	× **Konjunktur**	Samspelsterm.

Vi börjar med att undersöka ifall det föreligger korrelation mellan de förklarande variablerna. Det är som vi berättade i avsnitt 3.5 i så fall inte önskvärt och kan störa andra undersökningar. Vi utgår från modellen med samtliga variabler och utreda detta genom att titta på variablernas VIF-värden. Resultaten visas i tabell 2.

Tabell 2: VIF-värden för x-variablerna. $R^2_{adj} = 0.538$ för modellen.

Variabel	VIF-värde
Måndag	2.22
Lördag	3.32
Söndag	4.40
Sommar	5.45
Vår	11.28
Årtal	1.28
Tid	2.61
Omgång	11.27
Konjunktur	5.22
Regn	10.10
Temperatur	2.93
Framgång_lång	213.33
Malmö FF	4.13
IFK Göteborg	5.19
Elfsborg	5.76
Framgång_kort	4.69
Framgång_lång × Framgång_kort	4.82
Framgång_lång × Omgång	8.67
Framgång_lång × Malmö FF	4.40
Framgång_lång × IFK Göteborg	5.08
Framgång_lång × Elfsborg	6.07
Regn × Temperatur	10.13
Framgång_lång × Konjunktur	232.51

Som vi kan se från tabell 2 korrelerar *Framgång_lång* starkt med samspelstermen *Framgång_lång* × *Konjunktur*. Detta illustreras även i figur 1 nedan, där vi plottar dessa mot varandra.



Figur 1: *Framgång_lång* plottad mot samspelstermen *Framgång_lång* × *Konjunktur*.

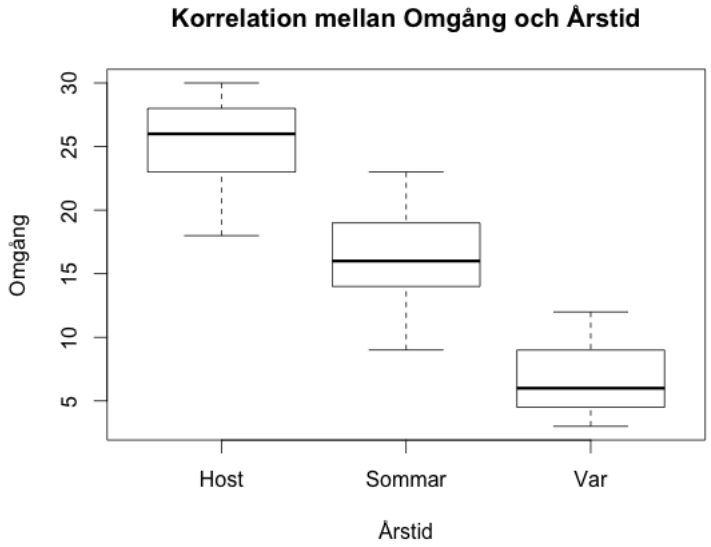
Någon av dessa blir vi definitivt tvugna att ta bort. Valet blir ganska enkelt, då *Framgång_lång* ingår i fler samspelstermer. Vi eliminerar därför *Framgång_lång* × *Konjunktur* och beräknar på nytt VIF-värdena på det variabler som är kvar, se resultatet i tabell 3 nedan.

Tabell 3: VIF-värden för modell utan variabel *Framgång_lång* × *Konjunktur*. R^2_{adj} = *0.539 för modellen.* .

Variabel	VIF-värde
Måndag	2.16
Lördag	3.32
Söndag	4.32
Sommar	5.42
Vår	11.24
Årtal	1.27
Tid	2.55
Omgång	11.20
Konjunktur	1.28
Regn	10.08
Temperatur	2.92
Framgång_lång	9.82
Malmö FF	4.13
IFK Göteborg	5.19
Elfsborg	5.55
Framgång_kort	4.13
Framgång_lång × Framgång_kort	4.41
Framgång_lång × Omgång	7.53
Framgång_lång × Malmö FF	4.38
Framgång_lång × IFK Göteborg	5.06
Framgång_lång × Elfsborg	5.87
Regn × Temperatur	10.13

Från tabellen ser vi att det blev mycket bättre, även justerad förklaringsgrad ökade marginellt. Men det är fortfarande fyra variabler som överstiger den VIF-värdesgräs, 10, som vi satte i avsnitt 3.5. Det handlar om årstidvariabeln *Vår*, *Omgång*, *Regn* och samspelsvariabeln *Regn* × *Temperatur*. Det är inte jättesvårt att gissa att det är variablerna *Regn* och *Regn* × *Temperatur* som hör ihop. Vi testar att ta bort den samspelande termen och ser att VIF-värdet för *Regn* minskar till 1.21. Dessutom noterar vi att $R^2_{adj} = 0.543$, och alltså ökar den ytterligare något för denna reducerade modell. Vi kan därför anta att samspelstermen *Regn* × *Temperatur* inte heller var vidare viktigt för inverkan på responsen. Vidare hade vi även en trolig korrelation mellan kategorivariabeln *Årstid* och *Omgång*. Det vore i så fall inte en så stor överraskning att just dessa korrelerar, eftersom den allsvenska säsongen alltid startar på våren och avslutas på hösten. Matchomgångarna följer på så vis årstiderna ganska konsekvent. Hade vi till exempel mätt månader

istället för årstider, så hade nog korrelationen varit ännu större. Vi plottar variablerna mot varandra för att bekräfta det vi misstänker, se figur 2.



Figur 2: *Omgång* plottad mot *Årstid*.

Som vi ser i figur 2 finns det ett tydligt linjärt samband mellan de två variablerna. Vi bör därför utesluta någon av dessa variabler. Valet här är inte fullt lika enkelt som det vi gjorde nyss för *Regn* och *Regn* × *Temperatur*. Eftersom *Årstid* är en kategorivariabel och *Omgång* kontinuerlig beskriver de lite olika saker, och vi kan gå miste om information. Någon av dem bör dock tas bort anser vi, och då vi har en samspelsterm innehållande variabeln *Omgång* lutar vi mot att ta bort den andre. Men för att vara på den säkrare sidan så testar vi att ta bort de båda var för sig och jämföra modellerna som uppstår. Det visar sig att vi tappar en hel del i justerad förklaringsgrad i modellen med *Årstid* jämfört med modellen innehållande *Omgång* istället. Vi anser oss nu ha två starka argument för att välja att ha kvar *Omgång* av de två. Vi tittar på nytt på VIF-värdena för modellen med variablerna *Årstid* och *Regn* × *Temperatur* exkluderade. Se nedan i tabell 4 de nya värdena.

Tabell 4: VIF-värden för modell med variablerna *Årstid* och *Regn* × *Temperatur* exkluderade. $R^2_{adj} = 0.545$.

Variabel	VIF-värde
Måndag	2.11
Lördag	3.18
Söndag	4.21
Årtal	1.21
Tid	2.47
Omgång	4.22
Konjunktur	1.25
Regn	1.19
Temperatur	1.48
Framgång_lång	9.59
Malmö FF	4.03
IFK Göteborg	5.14
Elfsborg	5.35
Framgång_kort	3.96
Framgång_lång × Framgång_kort	4.34
Framgång_lång × Omgång	7.52
Framgång_lång × Malmö FF	4.33
Framgång_lång × IFK Göteborg	5.02
Framgång_lång × Elfsborg	5.73

Resultaten i tabell 4 visar att nu är alla variabler under den tillåtna VIF-gränsen som vi valt, och vi väljer att fortsätta vidare med analysen.

4.4 Selektion av variabler

Innan vi tar hjälp av de procedurer som vi nämnde i avsnitt 3.3 för variabelselektion, så ska vi testa de kategoriska variabler som har fler än två nivåer. Det gäller de två kategorivariablerna *Veckodag* och *Motståndarlag*. Eftersom valet av referens kan påverka resultaten av vilken som väljs ut i de stegvisa metoder som vi avser att använda, så ska vi först testa hypotesen:

$$H_0 : \text{Ingen skillnad mellan kategorinivåerna,}$$

mot,

$$H_a : \text{Minst en av nivåerna skiljer sig.}$$

För att testa nollhypotesen gör vi ett F-test. Testet ger p-värde = 0.868 för kategorierna gällande variabeln *Veckodag*, dvs. vi kan inte förklasta H_0 på någon rimlig nivå. Vi utesluter därför variabeln i fortsättningen då testet antyder att den inte har någon inverkan. Resultatet av testet för *Motståndarlag* ger p-värde < 0.0001 för H_0, och vi kan därmed förkasta nollhypotesen på alla tänkabara signifikansnivåer, och variabeln behålls tillsvidare.

Nu är det dags att ta hjälp av procedurerna *Backward elimination*, *Forward selection* och *Stepwise regression* för att fortsätta selekteringen av vilka variabler vi bör ha i vår slutgiltiga modell. Det visar sig att *Backward elimination* och *Stepwise regression* väljer ut samma modell, och att *Forward selection* väljer en annan. Nedan i tabell 5 visas modellen de två förstnämnda metoderna båda valde ut, och vi kallar den för Modell 1.

Tabell 5: Modellen Stepwise regression och Backward elimination valde ut.

Modell 1		
Variabel	**Parameterskattning**	**P-värde**
Intercept	2367.26	0.4223
Årtal	-223.10	0.0061
Konjunktur	101.31	< 0.001
Regn	-1693.64	0.0369
Framgång_lång	207.31	0.2087
IF Elfsborg	4023.17	0.0318
IFK Göteborg	3623.62	0.0770
Malmö FF	10843.02	< 0.001
Framgång_kort	429.88	0.0102
Omgång	194.54	0.0018
Framgång_lång × Elfsborg	-247.27	0.3610
Framgång_lång × IFK Göteborg	64.21	0.8036
Framgång_lång × Malmö FF	-619.34	0.0053
Framgång_lång × Omgång	-36.48	< 0.001
$\hat{\sigma} = 2700$	$R^2 = 0.607$	$R^2_{adj} = 0.562$

Vi noterar i modell 1 höga p-värden för två av nivåerna för den kvalitativa samspelstermen *Framgång_lång* × *Motståndarlag*. Endast kategorinivån *Framgång_lång* × *Malmö FF* tycks vara signifikant på någon rimlig nivå. Denna företeelse tillsammans med att *Framgång_lång* × *Motståndarlag*-variablerna från tabell 4 verkar vara en bidragande orsak till en hel del av korrelationen som fanns kvar där, gör oss en aning skeptiska till variabeln. Vi testar därför att utesluta den, och analysera vad som sker. Vi kallar denna reducerade modell för Modell 2, och den syns i tabell 6 här under.

Tabell 6: Reducerad version av Modell 1, här exklusive *Framgång_lång* × *Motståndarlag*.

Modell 2		
Variabel	**Parameterskattning**	**P-värde**
Intercept	2942.8	0.3260
Årtal	-223.10	0.0050
Konjunktur	102.2	< 0.001
Regn	-1866.0	0.0245
Framgång_lång	112.9	0.4861
IF Elfsborg	2488.0	0.0074
IFK Göteborg	4158.7	< 0.001
Malmö FF	6769.0	< 0.001
Framgång_kort	433.6	0.0105
Omgång	186.2	0.0034
Framgång_lång × Omgång	-34.9	< 0.001
$\hat{\sigma} = 2770$	$R^2 = 0.575$	$R^2_{adj} = 0.539$

Signifikansen för samtliga variabler, med undantag *Framgång_lång*, förbättras och är nu signifikanta på en 5%-nivå. Gällande variabeln *Framgång_lång* så vill vi ha med den i modellen oavett, pga. samspelet med *Omgång*. Men vi tappar 3,2% i förklaringsgrad och 2,3% i justerad förklaringsgrad för denna modell jämfört med Modell 1.

Nu ska vi även titta på modellen som den stegvisa metoden *Forward selection* valde ut. Den presenteras i tabell 7 nedan, och får namnet Modell 3.

Tabell 7: Modellen *Forward selection* valde ut.

Modell 3		
Variabel	**Parameterskattning**	**P-värde**
Intercept	5918.5	0.048
Årtal	-189.9	0.028
Konjunktur	109.7	< 0.001
Regn	-2174.4	0.015
Framgång_lång	-408.8	< 0.001
IF Elfsborg	5160.7	0.011
IFK Göteborg	3148.8	0.157
Malmö FF	9940.1	< 0.001
Framgång_kort	357.7	0.044
Temperatur	-59.5	0.211
Framgång_lång × Elfsborg	-502.8	0.083
Framgång_lång × IFK Göteborg	140.3	0.616
Framgång_lång × Malmö FF	-417.8	0.087
$\hat{\sigma} = 2890$	$R^2 = 0.543$	$R^2_{adj} = 0.495$

Vi ser en tydlig försämring av förklaringsgrad och justerad förklaringsgrad jämfört med de två föregående modellerna. Vi anmärker att parameterskattningen för *Temperatur* har ett högt p-värde (0.211). Då vi inte finner fog för att behålla den, så bestämmer vi oss för att pröva att eliminera den från modellen. Resultatet för denna modell, Modell 4, visas i följande tabell 8.

Tabell 8: Reducerad version av Modell 3, variabeln *Temperatur* exkluderad.

Modell 4		
Variabel	**Parameterskattning**	**P-värde**
Intercept	5640.7	0.0590
Årtal	-178.1	0.0386
Konjunktur	102.5	< 0.001
Regn	-1953.9	0.0251
Framgång_lång	-390.2	< 0.001
IF Elfsborg	4711.0	0.0185
IFK Göteborg	3601.5	0.1020
Malmö FF	10469.8	< 0.001
Framgång_kort	328.4	0.0624
Framgång_lång × Elfsborg	-440.5	0.1230
Framgång_lång × IFK Göteborg	91.0	0.7433
Framgång_lång × Malmö FF	-505.2	0.0319
$\hat{\sigma} = 2900$	$R^2 = 0.537$	$R^2_{adj} = 0.492$

Modell 4 påvisar magrinellt sämre förklaringsgrader. Den justerade förklaringsgraden tappar exempelvis bara 0,3%. I övrigt likvärdiga p-värden för resterande variabler. Variabeln *Temperatur* som hade ett högt p-värde verkar alltså inte ha tillfört så väldigt till modellen.

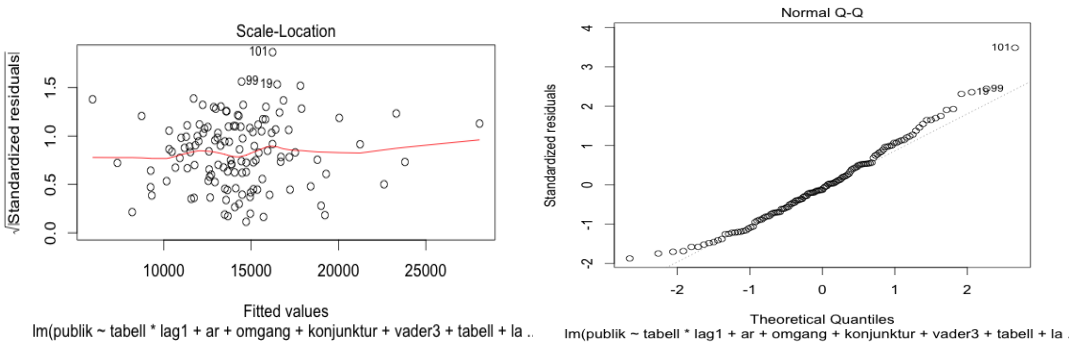
Sammanfattningsvis kan det vara värt att notera att de variabler som samtliga stegvisa procedurer plockade ut var *Årtal*, *Konjunktur*, *Regn*, *Framgång_lång*, *Motståndarlag*, *Framgång_kort* och samspelstermen *Framgång_lång* × *Motståndarlag*, och att den modell med bäst förklaringsgrad och justerad förklaringsgrad var Modell 1. Men, vi kan egentligen inte dra några slutsatser av dessa resultat förrän vi kontrollerat att modellerna är lämpliga, dvs. att modellantagandena är uppfyllda. Detta ska vi göra i nästa avsnitt.

4.5 Kontroll av modellantaganden

Något som är en förutsättning för att en modell i överhuvudtaget kan anses som användbar är att modellantagandena vi berättade om i avsnitt 3.5 approximativt är uppfyllda. För att ta reda på detta så tittar vi på residualerna plottade på olika sätt. Vi börjar med att undersöka residualerna för Modell 1, och i figur 3 plottas standardiserade residualer mot skattade responsvärden, samt en normalfördelningsplott. Det vi vill se är en jämn spridning av de standardiserade residualerna kring 0 för spridningsplotten. Det tyder i så fall på konstant varians och väntevärde 0 för residualerna, vilket är ett modellantagande. I normalfördelningsplotten önskar vi att se att de

standardiserade residualerna approximativt följer en rak ”y=x-linje”. Desto bättre denna linje följs ju bättre approximation av normalfördelningen. Se resultatet nedan.

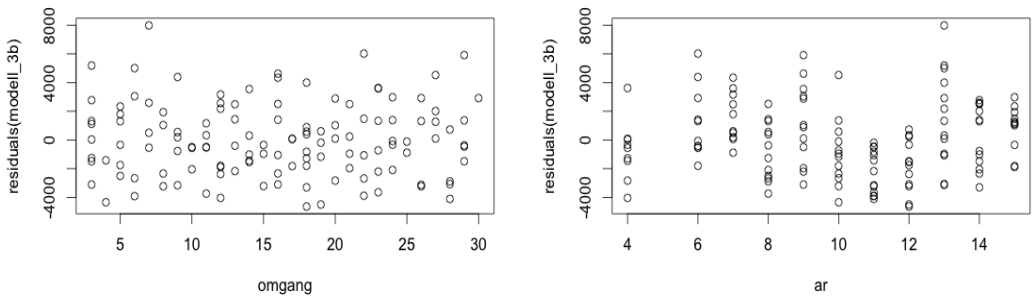
Residualer Modell 1.



Figur 3: Till vänster standardiserade residualer mot skattade värden. Till höger normalfördelningsplott.

Vi anser att spridningen av residualerna i den vänstra plotten fråg figur 3 är god nog för att godkännas. Resultatet från normalfördelningensplotten tycker vi även det är tillfredställande, och antagandet om normalfördening tycker vi vara lämpligt. För att vidare undersöka oberoendet av residualerna så plottar vi dessa mot varje förklarande variabel i Modell 1 och spanar efter systematik. Vi såg inga anmärkningsvärda tecken på detta, förutom för en variabel, nämligen variabeln *Årtal*. Nedan i figur 4 illustreras till vänster ett exempel på en av det tillfredsställande variablerna, och till höger plotten på den mer problematiska variabeln.

Residualer mot förklarande variabler, Modell 1.

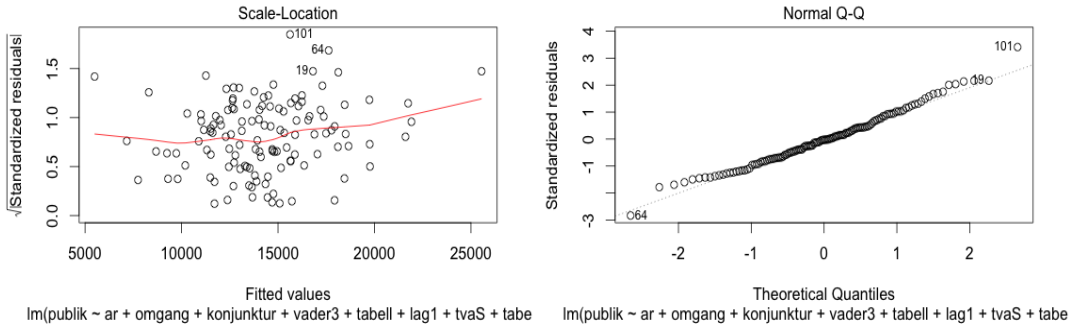


Figur 4: Till vänster residualer mot *Omgång*. Till höger residualer mot *Årtal*.

Som vi ser i den vänstra plotten från figur 4, så är spridningen bra och det syns inga tecken på systematik. Det kan vi dock uppfatta för den högra plotten, den för *Årtal*. Det verkar som att i synnerhet för år 11 (2011) och 12 (2012) men även delvis för år 4 (2004) och 10 (2010), så överskattar modellen publiksiffran. Vi fortsätter vidare för att se om detta problem uppstår för de andra modellerna också.

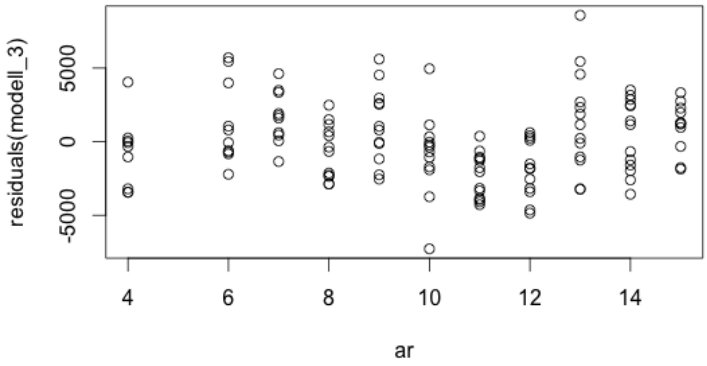
Vi kommer nu göra om samma procedurer för att undersöka modellantagandena och således residualerna för Modell 2, Modell 3 samt Modell 4. Här nedan i figur 5 och 6 följer exempel på resultaten från Modell 2. Vi utelämnar sedan plottarna tillhörande Modell 3 och Modell 4 till appendix. Läsaren får antingen ta vårt ord på att dessa var likvärdiga med plottarna för Modell 1 och Modell 2, om än något sämre, eller titta i appendix.

Residualplottar Modell 2.



Figur 5: Till vänster standardiserade residualer mot skattade värden. Till höger normalfördelningsplott.

Problematiska variabeln *Årtal*, Modell 2.



Figur 6: Residualer plottade mot *Årtal*, för Modell 2.

Alla resultat gällande undersökning av modellantaganden är likvärdiga och vi anser de vara adekvata. Något bättre approximationer för Modell 1 och Modell 2. Dock är problemet med systematik för den förklarande variabeln *Årtal* återkommande i alla modeller. Vi försöker på lite olika sätt manipulera variabeln, exempelvis genom transformation, men vi får inte bukt på problemet. Men när vi letar vidare efter sakliga förklaringar finner vi att just de åren vi anser att det är problem med, åren 4 (2004), 10 (2010), 11 (2011) och 12 (2012), är de fyra åren (eller säsongerna) som under vår tidsperiod hade lägst publiksnitt för de enskilda säsongerna. Dvs. av de 10 säsongerna vi har observerat i vårt datamaterial, så är det just dessa fyra säsonger som

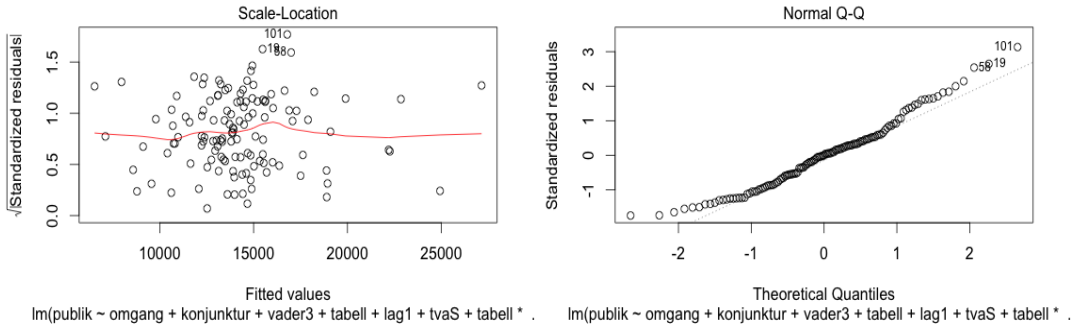
har haft lägst genomsnittlig publiksiffra. [1] Alltså fanns det just de åren något annat, utöver variablerna i modellen, som påverkade publiksiffrorna negativt. Det är troligtvis därför modellen överskattar publiksiffran just dessa årtal, och vi kommer kanske inte så mycket längre än så.

Även om spridningen för variabeln *Årtal* gällande residualerna inte är helt önskvärd, så är den heller inte katastrofal. Nu när vi tror oss ha funnit anledningen till beteendet, så överväger vi att låta variabeln vara kvar. Men för att vara på den säkra sidan gör vi en femte modell där *Årtal* exkluderas från Modell 1. Vi kallar den naturligen Modell 5, och dess värden visas i tabell 9, samt följer exempel på plottar för modellantaganden.

Tabell 9: Reducerad version av Modell 1, den problematiske variabeln *Årtal* exkluderad.

Modell 5		
Variabel	**Parameterskattning**	**P-värde**
Intercept	306.11	0.91696
Konjunktur	102.59	< 0.001
Regn	-1366.52	0.0971
Framgång_lång	180.66	0.2859
IF Elfsborg	3959.47	0.03989
IFK Göteborg	4216.38	0.0449
Malmö FF	11234.99	< 0.001
Framgång_kort	373.90	0.0280
Omgång	173.11	0.00618
Framgång_lång $\times$ Elfsborg	-224.40	0.4203
Framgång_lång $\times$ IFK Göteborg	12.95	0.9610
Framgång_lång $\times$ Malmö FF	-660.16	0.0039
Framgång_lång $\times$ Omgång	-33.37	< 0.001
$\hat{\sigma} = 2780$	$R^2 = 0.58$	$R^2_{adj} = 0.535$

Residualplottar Modell 5.



Figur 7: Till vänster standardiserade residualer mot skattade värden. Till höger normalfördelningsplott.

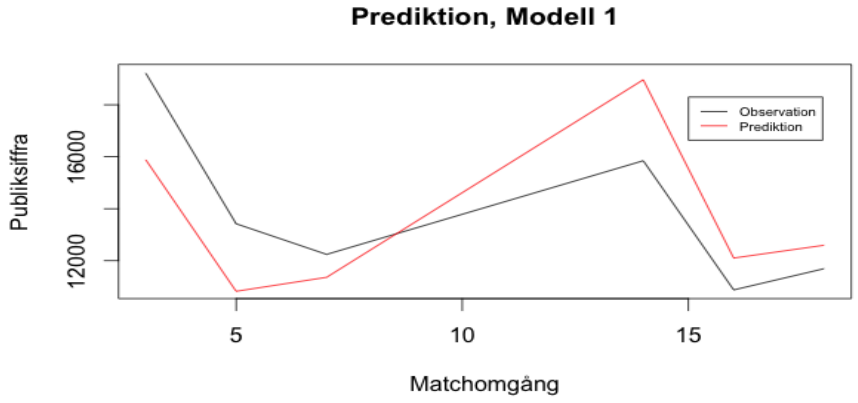
Vi ser i tabell 9 att vi får allmänt sämre p-värden. Modellen tappar ungefär 3% för båda förklaringsgraderna jämfört med Modell 1. Från figur 7 ser vi exempel på liknande approximationer för modellantaganena som för Modell 1.

5 Demonstration

Vi tänkte här testa Modell 1 prediktiva förmåga i praktiken. Eftersom vår utgångspunkt har varit att hitta en beskrivande modell i första hand så har vi inte allt för stora förväntningar, och kommer inte lägga någon stor vikt på resultatet utan anser detta som ett tillägg för den intresserade. Vi tänkte mest att det vore lite spännande att se. Matcherna som ska predikteras är från årets säsong, 2016, och datan i vårat material sträcker sig till slutet av november säsongen 2015. Se resultatet i tabell 11 och figur 8 här under.

Tabell 10

Prediktion, Modell 1			
Matchomgång	**Prediktion**	**Observation**	**Felmarginal**
3	15 889	19 222	-3 333
5	10 822	13 422	-2 600
7	11 360	12 242	-882
14	18 964	15 847	3 117
16	12 102	10 875	1 227
18	12 595	11 693	902



Figur 8: Predikterade och observerade värden för AIKs hemmamatcher säsong 2016.

Skattningarna är ungefär vad vi skulle kunnat förväntat oss utifrån Modell 1's specifikationer som standardavvikelsen som är 2 700, vilket överstämmer

ganska så bra med avvikelserna i tabellen. Från figur 8 ser vi att kurvorna har snarlika uteseenden. Att hälften av prediktionerna var inom ungefär felmarginal 1000 är ganska imponerande. Men återigen, vi lägger inte så värst stor vikt på resultatet, särskilt inte eftersom det endast handlar om 6 observationer. Hade vi varit mer inriktade på att en predikterande modell så hade korsvalidering varit ett bättre alternativ som utvärdernade metod.

6 Resultat och slutsats

Vi har under den här studien försökt anpassa en multipel linjär regressionsmodell som beskriver datamaterialet så bra som möjligt. Vi kom fram till fem alternativa modeller, och vi vill nu slutligen välja den av dem vi anser vara lämpligast för vårt syfte. Vi börjar med att sammanfatta modellerna i nedanstående tabell 10.

Tabell 11

Modell	**Variabler**	R^2	R^2_{Adj}	$\hat{\sigma}$
Modell 1	9st: *Årtal*, *Konjunktur*, *Väder*, *Framgång_lång*, *Motståndarlag*, *Framgång_kort*, *Omgång*, *Framgång_lång* × *Motståndarlag*, *Framgång_lång* × *Omgång*.	0.607	0.562	2700
Modell 2	8st: *Årtal*, *Konjunktur*, *Väder*, *Framgång_lång*, *Motståndarlag*, *Framgång_kort*, *Omgång*, *Framgång_lång* × *Omgång*.	0.575	0.539	2770
Modell 3	8st: *Årtal*, *Konjunktur*, *Väder*, *Framgång_lång*, *Motståndarlag*, *Framgång_kort*, *Temperatur*, *Framgång_lång* × *Motståndarlag*,	0.543	0.495	2890
Modell 4	7st: *Årtal*, *Konjunktur*, *Väder*, *Framgång_lång*, *Motståndarlag*, *Framgång_kort*, *Framgång_lång* × *Motståndarlag*,	0.537	0.492	2900
Modell 5	8st: *Konjunktur*, *Väder*, *Framgång_lång*, *Motståndarlag*, *Framgång_kort*, *Omgång*, *Framgång_lång* × *Motståndarlag*, *Framgång_lång* × *Omgång*.	0.58	0.535	2780

Då vårat syfte är att hitta en modell som beskriver data och variablernas inverkan så mycket som möjligt, så kommer vi att välja Modell 1 som slutgiltiga modell. Detta eftersom förklaringsgraden och framförallt den justerade förklaringsgraden är 2,3% högre än den modell med näst högst justerad förklaringsgrad. Valet var dock inte helt okomplicerat, då vi var aningen skeptiska mot två av variablerna som ingår i Modell 1 under analysen. Dels handlade det om samspelstermen *Framgång_lång* × *Motståndarlag*, där vissa nivåer inom kategorivariabeln hade väldigt höga p-värden, och dessutom bidrar variabeln till en del korrelation inom modellens förklarande variabler. Men då VIF-värdena är under den tillåtna gränsen vi satte på förhand, och att variabeln visar sig förklara ungefär 2-3% av utgången för responsen, så ser vi inte anledning nog att utesluta den. Den andra något tveksamma variabeln är *Årtal*, som uppvisade lite systematik för residualerna. Men eftersom

vi anser oss troligtvis funnit förklaringen till systematiken och alternativet vore Modell 5, där vi tappar 2,7% justerad förklaringsgrad, så väljer vi att behålla den också. Således lyder vår framtagna modell alltså:

$$\begin{aligned}\textbf{Publiksiffra} = 2367.26 &- 223.10(\textbf{Årtal}) + 101.31(\textbf{Konjunktur}) - 1693.64(\textbf{Regn}) + \\ &+ 207.31(\textbf{Framgång\_lång}) + 4023.17(\textbf{IF Elfsborg}) + 3623.62(\textbf{IFK Göteborg}) + \\ &+ 10843.02(\textbf{Malmö FF}) + 429.88(\textbf{Framgång\_kort}) + 194.54(\textbf{Omgång}) - \\ &- 247.27(\textbf{Framgång\_lång} \times \textbf{Elfsborg}) + \\ &+ 64.21(\textbf{Framgång\_lång} \times \textbf{IFK Göteborg}) - \\ &- 619.34(\textbf{Framgång\_lång} \times \textbf{Malmö FF}) - \\ &- 36.48(\textbf{Framgång\_lång} \times \textbf{Omgång}) + \varepsilon, \qquad (5)\end{aligned}$$

där, $\varepsilon \sim N(0, \sigma^2)$.

6.1 Utvalda variabler

Årtal. Som vi tidigare påpekat, så visste vi att publiksiffror historisk har varierat med tiden. Därför är det inte helt överraskande att denna variabel valdes ut. Enligt vår modell minskar antalet åskådare för varje enskild match med 223,10 personer per år, under vår tidsperiod.

Konjunktur. Enligt vår modell så verkar samhällets ekonomi kunna påverka publiksiffran. Dvs. vid högkonjunktur är fler benägna att konsumera biljetter än vid lågkonjunktur. Enligt (5) handlar det om ungefär 100 personer mer eller mindre i samma riktning som indikatorbarometern tar.

Väder. Ifall vädret kunde påverka fans att gå på en match eller ej var en av det första frågorna vi ställde oss. Vi kom fram till att väderlekarna som i så fall skulle kunna påverka var ifall det regnade eller ej. Det visade sig att enligt våra beräkningar på datamaterialet så avskräcker faktiskt regnet supportrarna för att gå på matcher. Hela 1 693,64 personer färre kommer det i snitt enligt vår modell när det regnar.

Framgång_lång. Sportsliga framgångar kändes kanske som den på förhand mest trovärdiga faktorn att ha inverkan på publiksiffror. Denna anger tabellplaceringen inför en match. Dock har den här variabeln högt p-värde, förmodligen till förmån för de två samspelstermerna i modellen, där denna är inblandad i båda. Pga, samspelet med *Motståndarlag*, som har basnivå *Övriga*, dvs. övriga lag förutom Malmö FF, IFK Göreborg och IF Elfsborg,

så tolkar vi *Framgång_lång*'s koefficient 207.31 såhär: om allting förutom dessa variabler är fixt och om Motståndarlag = Övriga, så kommer det 207.31 personer mer för varje numeriskt högre tabellplacering, dvs. längre ned i tabellen. Det kan dock tyckas vara lite förvånande, då höga tabellplaceringar, alltså det går sportsligt bättre för AIK, borde locka mer publik. Men, som sagt, p-värdet=0.2087 är hög och vi kanske inte ska lägga förstora växlar av koefficienten.

Framgång_kort. Tydligen verkar fans mer intresserade av att ta sig till matcherna ifall det har gått bra för laget de två senaste matcherna de spelat innan den aktuella matchen. Något vi på förhand tänkte skulle vara fullt rimligt. Alltså har enligt vår modell kortsiktig framgång inverkan på publiksiffran till en match. För varje poäng AIK tagit de två senaste matcherna kommer det enligt våra berälningar 429,88 personer fler.

Motståndarlag. Att motståndarlaget kan ha inverkan på publiksiffran är för de flesta fotbollsintressera känt. Vi intresserade oss dock inte för de mest uppenbara, dvs. de andra stockholmslagen. Vi undersökte istället de andra lagen, och det visade sig, inte helt förvånande heller, att andra storstäderslag och storlag som motståndare får fler att ansluta sig en sådan match. Enligt vår modell lockar IF Elfsborg 4023,17 personer, IFK Göteborg 3623,62 personer och Mamlö FF 10 843,02 personer fler än resterande lag den undersökta tidsperioden. Att just dessa lag lockade större publik än övriga var föga förvånande, men uppdelningen sinsemellan var lite överraskande. Vi hade trott att IFK Göteborg skulle locka minst lika mycket extra publik som Malmö FF om inte ännu mer. Då de flesta AIK-fansen nog anser att IFK Göteborg är en större rival. Anledningen kan vara att Malmö FF under senare delen av 2000-talet varit ett mer topplacerat lag än IFK Göteborg och kanske ansetts som det lag som är störst konkurrent om guldet senaste åren.

Omgång. Vi tänkte oss att senare matchomgångar i serien skulle kunna ha potential att locka mer publik. Eftersom det är då serien avgörs. Enligt vår analys stämmer detta, och vår modell menar att det i kommer ungefär 200 personer fler för varje ny omgång, dvs. varje enhetsökning för variabeln *Omgång*.

Framgång_lång $\times$ **Motståndarlag**. Samspelsterm mellan den aktuella tabellplaceringen och vilket motståndarlag som är på besök. Endast kategorinivån *Framgång_* $\times$ *Malmö FF* visade sig vara signifikant på någon rimlig nivå här. Enligt koefficienten för denna i modellen, –619.34, som anger att det kommer 207.31–619.34 = −412.03 färre (dvs det kommer färre eftersom resultatet av subtraktionen är < 0) för varje steg längre ned i tabellen AIK tar, om motståndarlaget är Malmö FF istället för *Övriga*.

Framgång_lång × **Omgång**. Vi tänkte oss att det är rimligt att anta att ju högre tabelläge (dvs. lägre siffra) laget har i slutet av serien, exemplevis om chanser för guld finns, så borde det öka fansens intresse och därav vore det troligt att fler ansluter till matcherna. Och tvärtom, är det ett dött lopp de sista 5-6 matcherna så kommer det förmodligen färre fans än vanligt. Enligt våra beräkningar verkar detta stämma.

7 Diskussion

I det här arbetet var utgångpunkten framförallt att finna variabler som beskriver publiksiffrorna i AIK fotbolls hemmamatcher mellan åren 2004-2015 (exklusive 2005). Vi fann en modell med 9 variabler som förklarade ungefär 60% av den totala variationen av publiksiffrorna. Variablerna som valdes ut och ingick i den slutgiltiga modellen tycker vi är rimliga. De flesta var mer eller mindre, enligt oss, förväntade att ha inverkan på publikantalet på matcher, och ingen var anmärkningsbart överraskande. Några överraskningar inom modellen fanns dock, som till exempel att IFK Göteborg enligt vår modell lockar mindre publik än IF Elfsborg och mycket mindre än Malmö FF. Vi hade på förhand förväntat oss att IFK Göteborg skulle vara den stora publikmagneten av de tre.

I början av arbetet, när insamlingen av data pågick, så letade vi efter information om AIK fotbolls reklamomsättning samt även AIKs aktivitet på sociala medier. Tyvärr fann vi inte denna information, inte i den utsträckning vi behövde i varje fall, trots några försökt att E-posta AIKs administrativa avdelningar. Det hade varit intressant att få ta del av reklamomkostnader för olika tidsperioder och utifrån det försöka undersöka om reklamen verkar vara värt kostnaden i fråga om det lockar mer folk till matcherna. På samma sätt hade det varit spännande att utreda ifall, den på 2000-talet succesivt ökande aktiviteten på sociala medier, bidrar till ökad publiksiffra. Kanske är det möjligen så, att i dagens läge så det mer lukrativt att satsa nästan enbart på att nå ut till fansen via internetsidor som *Instagram* och *Facebook*, än att lägga pengar på reklam i form av skyltning, radio, tv, osv. Vid en fortsatt studie hade vi gärna velat undersöka detta.

Mer hade man också kunnat tagit en lite annan ansats och lagt vikten på att försöka hitta en modell som predikterade kommande publiksiffror. Vore det till expempel möjligt att i praktiken göra olika val för att öka sannolikheten för en högre publiksiffra inför en match? Ifall man faktiskt skulle lyckas finna sådana variabler, som går att bemästra till fördel för en ökad åskadartillströmningen, och således öka intäkterna, torde det vara av högt intresse för alla klubblag. Och även om det skulle vara variablerna som inte gick att påverka, så vore det nog ändå av stort intresse för arrangörerna. Till exempel vore det ju nyttigt för en arrangör att ha en god gissning på hur mycket åskådare som kommer, så man kan anpassa hur mycket personal i form av vakter, kioskpersonal, entrévärdar osv. som ska plockas in.

På grund av att responsen i denna studie har omfattat publikantal, hade det varit naturligt att ta till en modell av typen GLM-Poisson. Men eftersom publiksiffrorna är så pass stora, dvs. parametern λ i en Poissonfördelning är > 1000, så är en normalfördelningsapproximation god, och därför kan även

multipel regression antas vara ett lämpligt modellval. Dessutom har vi i våra modeller inte sett något tecken på heteroskedasticitet hos residualerna, ty hade en Poissonfördelning varit lämpligare så hade variansen ökat med ökade skattade värden på responsen.

8 Referenser

[2]

Referenser

[1] AIK (2015). *Publiksnitt säsongsvis för AIKs hemmamatcher*. http://www.aikfotboll.se/TextPage.aspx?textPageID=4835

[2] Alan Agresti (2002). *Categorical Data Analysis*. Second Edition.

[3] Fredrik Carlgren (2016). *Barometerindikatorn*. http://www.ekonomifakta.se/Fakta/Ekonomi/Tillvaxt/Konjunkturen—Barometerindikatorn/

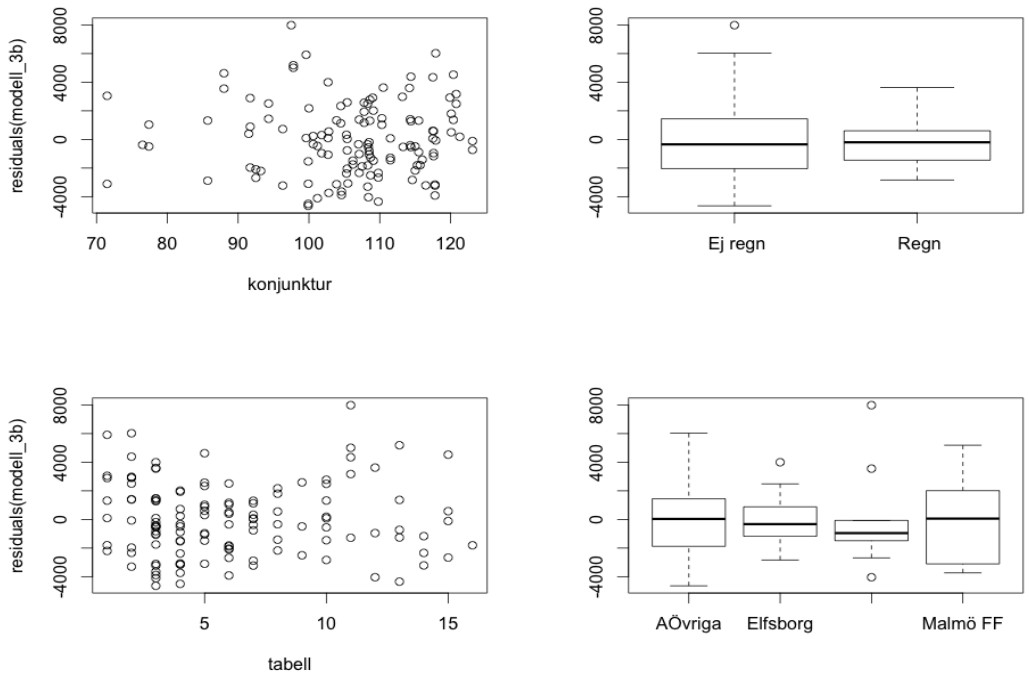
[4] IFFHS (2015). *Världsranking fotbollsligor*. http://iffhs.de/the-strongest-league-in-the-world-2015/

[5] Johan Jentell, Patrik Bodin (2016). *AIK Fotbolls Statistikdatabas*. http://www.aik.se/fotboll/statistik/

[6] Rolf Sundberg (2015). *Lineära Statistiska Modeller*.

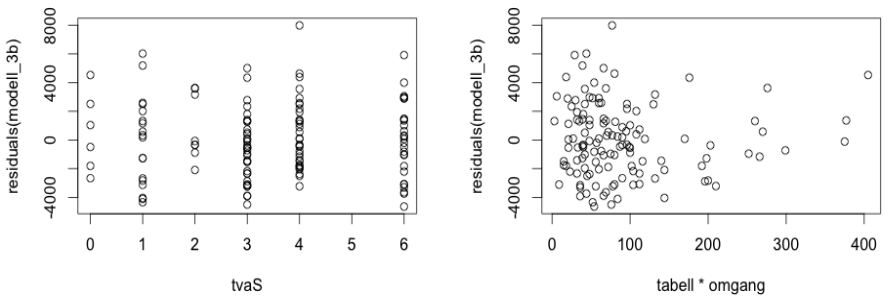
9 Appendix

Residualplottar förklarande variabler, Modell 1



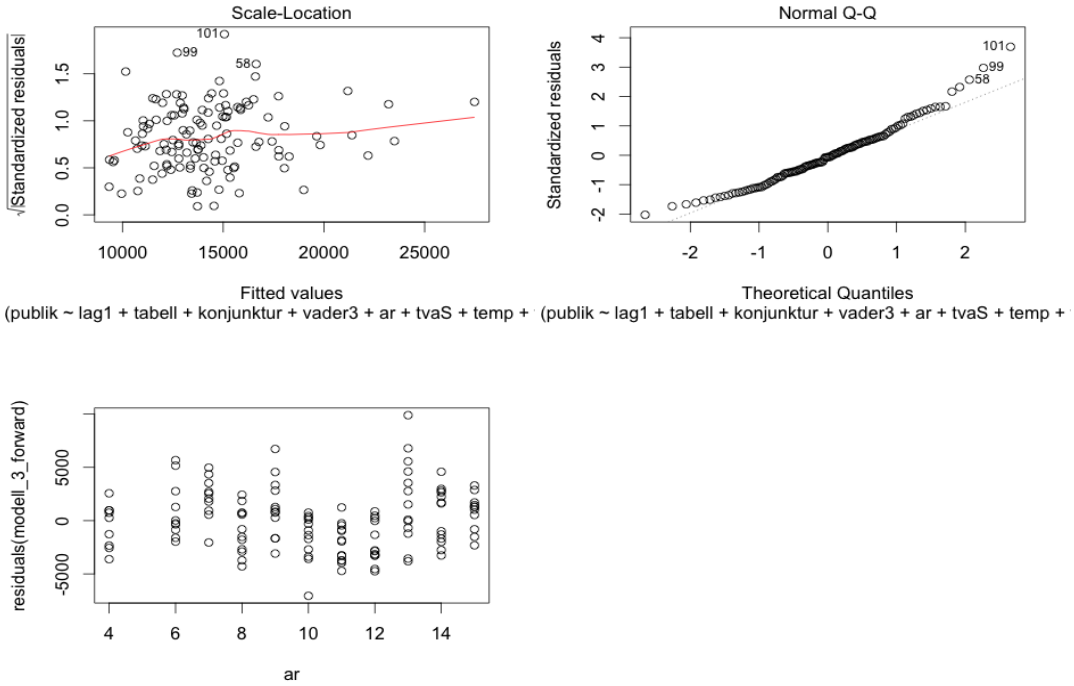
Figur 9: Residualplottar för variablerna: Konjunktur, Väder, Framgång_lång och Motståndarlag

Residualplottar förklarande variabler, Modell 1



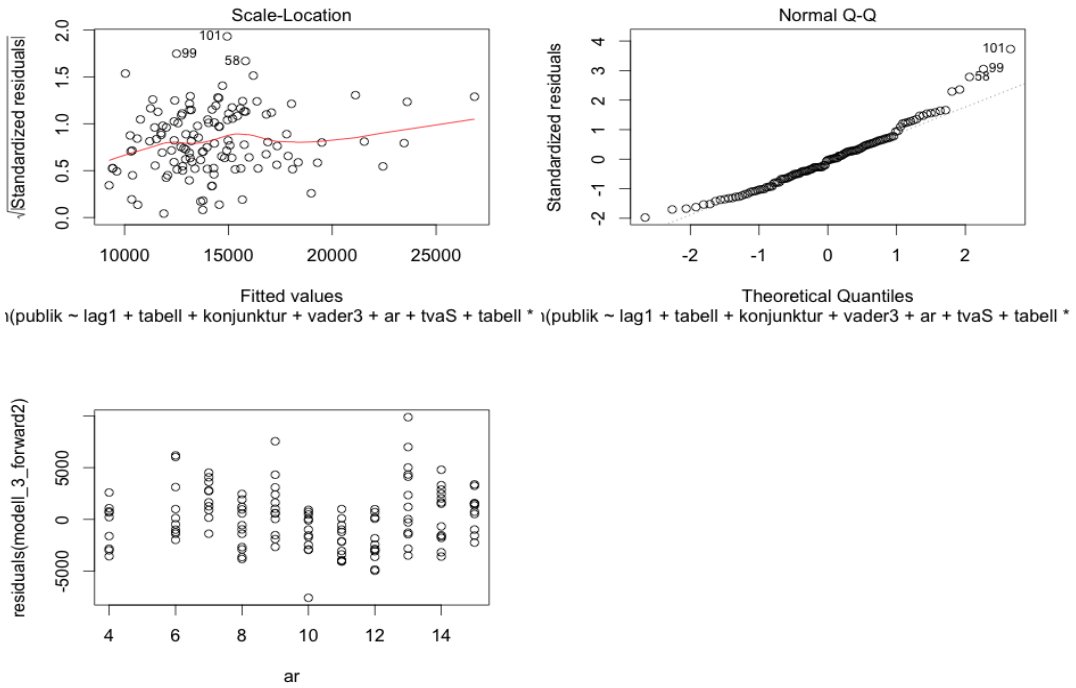
Figur 10: Residualplottar för variablerna: Framgång_kort och Framgång_lång*Omgång

Residualplottar Modell 3



Figur 11: Residualplottar.

Residualplottar Modell 4



Figur 12: Residualplottar.