Stockholms universitet

Statistisk analys av en bowlingspelares poängsnitt

Dennis Uygur Andersson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:7
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2018:7**
http://www.math.su.se/matstat

Statistisk analys av en bowlingspelares poängsnitt

Dennis Uygur Andersson*

Juni 2018

Sammanfattning

Syftet med detta arbete är att undersöka vad som påverkar en bowlingspelares poängsnitt, där poängsnitt inom bowling baserar sig på hur väl man har presterat under en viss tidsperiod. Till vårt förfogande har vi hjälp av ett datamaterial som är hämtat från Bowlingens IT System (BITS), vilket vi valt att analysera hur responsvariabeln poängsnitt påverkas av ett antal förklarande variabler under år 2016. Med hjälp av en kombination av multipel regression och en blandad modell, visade det sig att en bowlingspelares kön och antalet serier varje licensierad bowlare spelar, har störst inflytande på snittpoängen. Vidare verkade även åldern ha en bidragande effekt, där man i allmänhet är bäst runt 39-årsåldern. Slutligen påvisades även huruvida ett speltillfälle inträffade på våren samt på tävling gav en positiv effekt på snittpoängen. Dock kan slutresultat ifrågasättas eftersom modellförutsättningarna inte helt och hållet är uppfyllda.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: dean2175@math.su.se. Handledare: Jan-Olov Persson.

Abstract

The goal with this assignment is to examine what affects a bowling players average score, where as the average score within bowling is based on how well a licensed player has performed over a certain period of time. To our help, we have collected a data sample from the Swedish bowling's IT system (BITS), where we have chosen to analyze how the response variable (average score) is affected by a certain number of explanatory variables for the year 2016. With the help of a combination of multiple regression and mixed model effects, it turned out that a bowling players gender and the number of series the licensed bowler played, has the largest effect on the average score. Furthermore, age also had a contributing effect which generally turned out that a bowler around 39's is the optimal age. Finally, it was also shown whether a game event occurred in the spring and on the competition gave a positive effect on the average score. This result may though be questioned since the model assumptions are not fully met.

Förord

Detta självständiga arbete är en avhandling inom matematisk statistik om 15 hp. Inledningsvis vill jag rikta ett stort tack till Svenska Bowlingförbundet som hjälpt mig tillgodose med datamaterial och funnits där vid eventuella frågor. Vidare vill jag även passa på att tacka Jan-Olov Persson, som varit min handledare under arbetets gång, för hjälpsamma kommentarer och råd.

Innehåll

1 Inledning **5**
1.1 Snittpoäng 5
1.2 Frågeställning 6

2 Beskrivning av datamaterialet **6**
2.1 Allmänt 6
2.2 Begränsningar 7
2.3 Variabler 7
2.4 Exemplifiering av datamaterialet 8

3 Teori **10**
3.1 Regressionsanalys 10
3.1.1 Enkel linjär regression 10
3.1.2 Multipel linjär regression 10
3.2 Förklaringsgrad 10
3.3 Linear mixed model (Linjär blandad modell) 11
3.4 AIC 12

4 Statistisk modellering **14**
4.1 Undersökning av datamaterialet 14
4.2 Regressionsanalys 17
4.2.1 Multipel linjär regression 17
4.3 Mixed model 20
4.3.1 Val av kovariansstuktur 20
4.3.2 Modellval 21
4.3.3 Tolkning av fixa variabler 26

5 Slutsats **29**

6 Diskussion **30**

7 Appendix **32**

Referenser **33**

1 Inledning

Bowling har funnits i hundratals år, men det är inte förrän på de senare åren man försökt hitta förklaringar till vad som skiljer nybörjarna mot proffsen. Sverige har flera tusentals personer som aktivt utövar sporten, där en minoritet (<1%) är 'riktigt duktiga', med ett så kallat *poängsnitt* på över 220 poäng/serie. Man kan intuitivt ana vad som menas med ett poängsnitt, men vi kommer ändå förklara hur denna räknas ut. Vidare kommer även relevanta delar inom sportbowlingen klargöras.

För enkelhetens skull använder vi oss av ett exempel som illustrerar hur sportbowlingen fungerar. Låt oss säga att Adam gärna vill börja utöva bowling som sport. Hans första steg mot att bli en bowlingspelare är att skaffa en *licens* som i sin tur ger honom ett unikt *licensnummer*, vilket fungerar som (och faktiskt baserar sig på) ett personnummer (ex. M110387ADA03). Därefter vill Adam givetvis börja utöva sporten, där han kan välja mellan spelformerna *match* eller *tävling*. Dessa två spelsätten är ypperligt snarlika, där den enda nämnvärda skillnaden är att tävlingsformen oftast involverar fler *serier*. Oavsett vilken spelform Adam väljer, behöver han spela ett antal serier, där hans resultat för varje serie kommer variera mellan 0 och 300 poäng. Vidare kommer Adam spela ett visst antal serier, där summan av poängen för varje serie genererar en *totalpoäng*. När man undersöker totalpoängen, kan man givetvis titta på hans sammanlagda resultat under en specifik tävling, men vi kommer under arbetets gång främst vara intresserade av totalpoängen under ett helt år, vilket på motsvarande sätt beräknas genom att summera poängen man fått i varje serie över ett år.

1.1 Snittpoäng

Vi såg alltså hur totalpoängen beräknades genom att ta summan av poängen för varje serie. Denna totalpoäng kommer ligga till grund för något av större intresse, nämligen *snittpoängen*. Man kan räkna ut snittpoängen för flera saker inom bowlingen (exempelvis Adams snittpoäng för en specifik tävling), men under arbetets gång kommer vi primärt vara intresserade av två olika typer av snittpoäng:

(i) Snittpoängen på hösten respektive våren som spelats i form av både matcher och tävlingar under år 2016.

(ii) Snittpoängen under år 2016 oavsett spelform och årstid.

För (i) kommer varje bowlingspelare erhålla *fyra* olika snittpoäng, eftersom vi vill undersöka hur väl man presterat under två olika spelformer för två olika årstider. Vi beräknar då snittpoängen genom att dividera totalpoängen för en specifik spelform och årstid med korresponderande antalet serier.

När vi beräknar (ii) kommer varje individ enbart erhålla en snittpoäng, eftersom vi bortser från såväl spelformen som årstiden. Snittpoängen för en bowlingspelare beräknas då ut genom att ta totalpoängen under hela året utan hänsyn till

varken spelform och årstid, dividerat med motsvarande antal serier som hen har spelat.

1.2 Frågeställning

Vi har visat hur man beräknar ut ett personligt poängsnitt, från två olika aspekter, vilket kommer ligga i grund till arbetets frågeställning; *Vad är det som påverkar en bowlingspelares poängsnitt?* Med andra ord, vilka bakomliggande faktorer kommer påverka hur väl Adam spelar? Kan det vara *antalet serier* han spelar under en viss tid? Påverkar det faktum att han är av könet man och inte kvinna? Har årstiden någon effekt på poängsnittet? Dessa är bara ett antal av faktorer som kan inverka på poängsnittet. Gissningsvis borde antalet serier han spelar ha en större inverkan, då man i allmänhet blir bättre desto mer man utövar något. Vidare bör även åldern påverka snittet då en person över 60 år inte har samma fysiska förutsättningar som en 25 åring. Under arbetets gång kommer vi använda oss av multipel regression och blandad modell för att försöka besvara vår frågeställning.

2 Beskrivning av datamaterialet

Följande avsnitt beskriver inledningsvis allmänt om datamaterialet med en övergripande summeringsstatistik. Vidare förklaras de variabler vi kommer använda oss av under arbetets gång. Slutligen kommer exemplifiera datamaterialet efter vi gjort vissa begränsningar.

2.1 Allmänt

Allt datamaterial är hämtad ifrån Bowlingens IT System (BITS), och innehåller information om hur väl en licenserad bowlingspelare (brukar även benämnas 'bowlare') har presterat vid varje tillfälle. Varje person har ett unikt licensnummer som gör det enkelt för oss att urskilja hur mycket totalpoäng en bowlingspelare har gjort över ett antal serier för olika tävlings- och matchtillfällen. För att ett resultat ska ha blivit registrerat, måste alltså en licensierad bowlare antingen spelat en match som innefattar max 4 serier, eller en tävling sanktionerad av en förening eller ett bowlingförbund, som i allmänhet innefattar (minst) 4 serier. Vi nämnde tidigare att poängsnittet kommer vara av intresse över såväl olika spelformer (match/tävling) som för de olika årstiderna (vår/höst) för år 2016. Något som då kan vara av intresse är att se en viss summeringsstatistik över det året. Vi kommer först och främst illustrera hur många serier de två könen *i snitt* har spelat för de olika spelformerna och årstiderna. Vidare kommer vi även titta på medelvärdet och medianen för antalet serier hos de båda könen. Slutligen visar vi även det minsta respektive högsta antalet serier som har spelats för de båda könen. Följande summeringsstatistik omfattar 2814 damer och 9812 herrar för året 2016.

Tabell 1.1: Antal serier (i snitt) för de två könen.

Kön	*Tävling/match*	*Vår/Höst*	*Antal serier (snitt)*
Damer	Matcher	Vår	7.44
Damer	Tävlingar	Vår	2.92
Damer	Matcher	Höst	6.59
Damer	Tävlingar	Höst	2.32
Herrar	Matcher	Vår	8.35
Herrar	Tävlingar	Vår	2.97
Herrar	Matcher	Höst	7.48
Herrar	Tävlingar	Höst	2.35

Tabell 1.2: Medelvärdet och medianen för antalet serier samt minst respektive flest antal serier för de två könen.

Kön	*Medelvärde*	*Median*	*Minst antal serier*	*Flest antal serier*
Damer	54.63	44	1	319
Herrar	62.61	57	1	410

Notera från Tabell 1.1 att antalet serier (i snitt) varierar mellan 2.32 och 8.35, vilket kan ses som väldigt lågt med tanke på att en enda match i allmänhet innefattar 4 serier. Det innebär exempelvis att damer (i snitt) inte ens spelar två matcher per årstid. Från tabell 1.2 noterar vi dels den stora variationen mellan minst respektive flest antal spelade serier, men även variationen mellan flest antal spelade serier och medianen. Dessa delresultat indikerar på att det finns ett flertal bowlare som har spelat ytterst få serier. Faktum är att ungefär 27% av damerna (757 st) samt 21% av herrarna (2025 st) har spelat färre än 20 serier under hela året. På grund av detta är det rimligt att göra begränsningar som gör att vi 'undviker' de personerna som spelat ett fåtal serier.

2.2 Begränsningar

Vi noterade alltså andelen som spelat färre än 20 serier. Till följd av det resultatet, begränsar vi att samtliga bowlingspelare måste spelat minst 5 serier på våren respektive hösten samt att varje person spelat minst 5 serier i form av både tävling och match. Detta kommer alltså innebära att varje individ måste spelat minst 20 serier under år 2016. Något som inte har nämnts är att vissa personer har 0 i snittpoäng, vilket antagligen beror på att de enbart har spelat lagtävlingar där man inte kan beräkna ett personligt snitt. På grund av detta antar vi även att $0 < \textit{Snittpoäng} \leq 300$. Efter dessa begränsningar finns det 483 kvinnor och 1796 män för år 2016.

2.3 Variabler

Som tidigare nämnt ska vi försöka förklara snittpoängen som även blir vår responsvariabel. Vi kommer göra detta med hjälp av följande variabler:

Kön: Om personen som spelade var man eller kvinna.

Match/Tävling: Om speltillfället var en match eller en tävling.

Antal Serier: Hur många serier personen har spelat.

Period: Vilken tid på året speltillfället inföll (vår/höst).

Ålder: Hur gammal bowlaren var vid speltillfället.

Under arbetets gång kommer även *Id* dyka upp, vilket är variabelnamnet för varje licensierad bowlare.

2.4 Exemplifiering av datamaterialet

Tidigare nämnde vi intresset över två olika typer av snittpoäng för år 2016, vilket vi även visade hur man beräknade. Den ena sorten var snittpoängen hos en bowlingspelare för de båda årstiderna och spelformerna, vilket resulterade i att varje bowlingspelare erhåller fyra olika poängsnitt. Vi exemplifierar hur datamaterialet ser ut för två personer, då vi tittar just på dessa snittpoängen.

Tabell 2: Exempel på datamaterialet för två personer när vi tittar snittpoängen över de två spelformerna och årstiderna.

Id	*Kön*	*Match/tävling*	*Antal Serier*	*Period*	*Ålder*	*Snitt*
M141294DEN01	Man	Tävling	82	Vår	22	209.27
M141294DEN01	Man	Tävling	33	Höst	22	208.09
M141294DEN01	Man	Match	49	Vår	22	200.10
M141294DEN01	Man	Match	44	Höst	22	220.57
M100891MIK01	Man	Tävling	34	Vår	25	221.42
M100891MIK01	Man	Tävling	33	Höst	25	204.70
M100891MIK01	Man	Match	44	Vår	25	219.07
M100891MIK01	Man	Match	42	Höst	25	219.07

Denna sorten av snittpoäng kommer vi analysera med hjälp av en så kallad *blandad modell*, vilket vi behandlar i Avsnitt 4.3. Den andra sorten vi tittade på var snittpoängen utan hänsyn till spelform och årstid, vilket gör att varje enskild bowlingspelare enbart erhåller en snittpoäng. Vi exemplifierar även detta datamaterialet för fyra olika personer i Tabell 3.

Tabell 3: Exempel på datamaterialet för fyra personer när vi tittar på snittpoängen oavsett spelform och årstid.

Id	*Kön*	*Antal Serier*	*Ålder*	*Snitt*
K010303MAT01	Dam	102	13	112.14
K220863PIA01	Dam	52	53	183.24
M100891MIK01	Herre	153	25	216.89
M021290ERI01	Herre	133	26	198.79

Denna typen av snittpoäng kommer vi använda oss av under Avsnitt 4.1 och 4.2, det vill säga när vi undersöker datamaterialet samt använder oss av multipel regression.

3 Teori

3.1 Regressionsanalys

Linjära regressionsmodeller är en statistisk metod som används för att påvisa om en responsvariabel beror eller åtminstone beskrivs väsentligen linjärt av en eller flera andra förklarande variabler (regressorer). 3.1.1, 3.1.2 och 3.2 är taget från Rolfs kompendium (se [7]).

3.1.1 Enkel linjär regression

Den allmänna modellen för enkel linjär regression definieras enligt

$$Y_i = \alpha + \beta_1 \cdot x_i + \epsilon_i, \tag{3.1}$$

där α (intercept) och β_1 (lutningskoefficient) är okända parametrar, Y_i är responsvariabeln, x_i är förklarande variabel och ϵ_i är en slumpterm. Det gäller även att $i = (1, \ldots, N)$, där N är antalet observationer. Modellen (3.1) förutsätter att feltermerna är sinsemellan oberoende och $N(0, \sigma^2)$.

3.1.2 Multipel linjär regression

Den allmänna modellen för multipel linjär regression definieras enligt

$$Y_i = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_n \cdot x_{ni} + \epsilon_i, \tag{3.2}$$

där $\alpha, \beta_1, \beta_2 \ldots, \beta_n$ är parameterar, Y_i är responsvariabeln, $x_{1i}, x_{2i}, \ldots x_{ni}$ är förklarande variabler och ϵ_i är en slumpterm. Det gäller även att $i = (1, \ldots, N)$, där N är antalet observationer. Modellen (3.2) förutsätter att feltermerna är sinsemellan oberoende och $N(0, \sigma^2)$.

3.2 Förklaringsgrad

Ett av de vanligaste anpassningsmåttet i samband med linjära modeller och multipla regressionsmodeller är förklaringsgraden, R^2. Denna definieras som den andel av den totala variationen som modellen förklarar". Detta verktyg är användbart när man vill jämföra olika modeller och definieras enligt

$$R^2 = \frac{\text{Kvs(regression)}}{\text{Kvs(totalt)}} = 1 - \frac{\text{Kvs(residual)}}{\text{Kvs(totalt)}} \tag{3.3}$$

Förklaringsgraden, R^2, är ett tal mellan 0 och 1, där ett högre värde tyder ett starkare linjärt samband. En högre förklaringsgrad innebär även således att den skattade modellen lyckas fånga in mer av varationen.

En nackdel med förklaringsgraden är att denna ökar så fort man tillför en x-variabel till modellen, även om variabeln inte förklarar nämnvärt. Vill man tillföra en variabel kan det vara bättre att se om $\hat{\sigma}^2$ minskar. Om $\hat{\sigma}^2$ skulle

minska kan man tolka det som att man har mindre grad av slump kvar i modellen. Detta mått brukas kallas för R^2_{adj}. Den explicita formeln för adjusted R^2 ges av

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma_0}^2}, \tag{3.4}$$

där $\hat{\sigma_0}^2 = \frac{\text{Kvs(totalt)}}{N-1}$ är σ^2-skattningen då man inte har någon x-variabel i modellen.

3.3 Linear mixed model (Linjär blandad modell)

Vi säger att en linjär blandad modell är en statistisk modell som både består av fixa och slumpmässiga effekter. Denna modell används ofta när man har upprepade mätningar inom varje individ under en viss tidsperiod. Då varje individ har k upprepade mätningar för N individer skrivs den blandade modellen som

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i., \tag{3.5}$$

där $\boldsymbol{y_i} = [y_{i1}, y_{i2}, \ldots, y_{ik}]^T$ är *responsvektorn* för individ i. Det gäller även att $\boldsymbol{X}_i$ är de fixa effekternas designmatris samt att $\boldsymbol{\beta}$ är de fixa effekternas parametervektor. För modellen (3.5) är $\boldsymbol{Z}_i$ och $\boldsymbol{b}_i$ matrisen av kovariater samt korresponderande vektor av slumpmässiga faktorer. Slutligen är $\boldsymbol{\epsilon}_i$ vektorn för slumpfelet för varje individ. Vidare antas även att

$$\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{S}), \qquad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \quad \text{där} \quad \boldsymbol{b}_i \perp \boldsymbol{\epsilon}_i. \tag{3.6}$$

Med andra ord, feltermen $\boldsymbol{\epsilon}_i$ för samma grupp, är oberoende av de slumpmässiga effekterna $\boldsymbol{b}_i$. Om individ är den enda slumpmässiga effekten, ser modellen istället ut enligt

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i. \tag{3.7}$$

De slumpmässiga effekternas kovariansmatris, $\boldsymbol{R}$, är en diagonalmatris och ser ut enligt

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{R}_N \end{bmatrix}, \tag{3.8}$$

där varje R_i är $k \times k$ matriser, vilket varje k motsvarar hur många gånger mätningarna har upprepats för individ i. För matrisen (3.8) gäller även att $\boldsymbol{0}$ motsvarar blockmatriser av nollor som påvisar att kovariansen mellan mätningar hos varje individ är noll. Vidare kan kovariansstrukturen för R_i:na se olika ut. Den enklaste möjliga strukturen kallas för *compound symmetry* (CS), där man antar att samtliga korrelationer är samma för alla mätningar inom varje individ. För fyra upprepade mätningar på två individer ser kovariansstrukturen ut enligt följande

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & 1 & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \rho & \rho & \rho \\ 0 & 0 & 0 & 0 & \rho & 1 & \rho & \rho \\ 0 & 0 & 0 & 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & \rho & \rho & \rho & 1 \end{bmatrix} \tag{3.9}$$

Om man istället för varje mätning tillåter separata korrelationer och homogen varians fås något som kallas för *unstructured* (UN). För samma typ av exempel som ovan, ser denna kovariansmatrisstruktur ut enligt

$$\boldsymbol{R} = \begin{bmatrix} \sigma^2 & \sigma^2\rho_{12} & \sigma^2\rho_{13} & \sigma^2\rho_{14} & 0 & 0 & 0 & 0 \\ \sigma^2\rho_{12} & \sigma^2 & \sigma^2\rho_{23} & \sigma^2\rho_{24} & 0 & 0 & 0 & 0 \\ \sigma^2\rho_{13} & \sigma^2\rho_{23} & \sigma^2 & \sigma^2\rho_{34} & 0 & 0 & 0 & 0 \\ \sigma^2\rho_{14} & \sigma^2\rho_{24} & \sigma^2\rho_{34} & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & \sigma^2\rho_{12} & \sigma^2\rho_{13} & \sigma^2\rho_{14} \\ 0 & 0 & 0 & 0 & \sigma^2\rho_{12} & \sigma^2 & \sigma^2\rho_{23} & \sigma^2\rho_{24} \\ 0 & 0 & 0 & 0 & \sigma^2\rho_{13} & \sigma^2\rho_{23} & \sigma^2 & \sigma^2\rho_{34} \\ 0 & 0 & 0 & 0 & \sigma^2\rho_{14} & \sigma^2\rho_{24} & \sigma^2\rho_{34} & \sigma^2 \end{bmatrix} \tag{3.10}$$

En ytterligare kovariansstruktur som nämns i denna sektion kallas för *Autoregressive(1)* (AR(1)), vilket antar en struktur med homogen varians och korrelation som minskar exponentiellt mot 0, desto längre tid det är mellan mätningarna. Vi väljer att inte visa strukturen för denna, utan hänvisar istället till *Applied Mixed Models in Medicine, 2nd ed., H. Brown & R. Prescott.*

För vidare läsning om blandade modeller och ytterligare detaljer för teorin ovan, se *Linear Mixed-Effects Model Using R* av Andrezej Galecki Tomasz Burzykowski samt *Applied Mixed Models in Medicine, 2nd ed., H. Brown & R. Prescott.*

3.4 AIC

Ett användbart mått vid modellval är *Akaike Information Criterion (AIC)* som bedömer en modell genom att undersöka hur nära dess anpassade värden tenderar de sanna värdena, i form av ett visst väntevärde. Denna utgår ifrån såväl värdet på *likelihoodfunktionen* som antalet parametrar för en modell och definieras enligt

$$AIC = -2(L_M - p), \tag{3.11}$$

där L_M motsvarar värdet den maximerade loglikelihooden och p motsvarar antalet parametrar i modellen. Från (3.11) menas även att desto fler parametrar modellen har, desto lägre blir AIC-värdet. Detta innebär i sin tur att det straffar en modell med flera parametrar [1].

4 Statistisk modellering

4.1 Undersökning av datamaterialet

Vi har alltså visat hur man beräknar ut snittpoängen för en bowlingspelare, där vi dels är intresserade av snittpoängen för varje bowlingspelare över de två spelformerna (match/tävling) och årstiderna (vår/höst), men även snittpoängen oavsett spelform och årstid. Vi begränsade även att varje individ måste spelat 20 serier under året 2016, där 5 serier måste vara i form av både tävling och match samt att 5 serier måste ha spelats på såväl våren som på hösten. Under Avsnitt 4.1 och 4.2 kommer vi undersöka datamaterialet som illustrerades i Tabell 3, det vill säga snittpoängen för varje enskild bowlingspelare oavsett spelform och årstid.

En viktig del för undersökandet av datamaterialet är att dels se hur responsvariabeln *snittpoäng* förhåller sig till de förklarande kontinuerliga variablerna, det vill säga *ålder* och *antal spelade serier*, men även hur de två förklarande variablerna förhåller sig till varandra. Inledningsvis kommer vi (främst) analysera korrelationen mellan variablerna ålder och antal serier, då det kan upplysa oss ifall vi bör förvänta oss problem med multikollinearitet. Detta gör vi med hjälp av en korrelationsplot som visas i Figur 1.

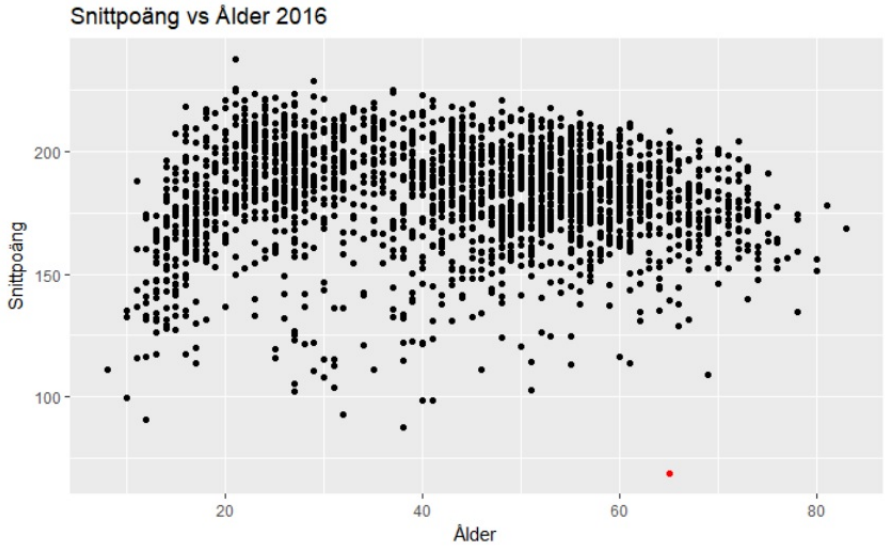
	Ålder	Antal Serier	Snitt
Ålder	1	-0.13	-0.01
Antal Serier	-0.13	1	0.46
Snitt	-0.01	0.46	1

Figur 1: Korrelationsplot mellan de kontinuerliga variablerna.

Från Figur 1 noterar först och främst att de kontinuerliga förklarande variablerna ålder och antal spelade serier har en låg negativ korrelation på -0.13. Detta

innebär att dessa två variablerna är nästintill oberoende av varandra vilket i sin tur gör att vi undviker problemet med multikollinearitet. Vidare noterar vi att korrelationen mellan variablerna snittpoäng och antal spelade serier innehar ett värde på 0.46. Detta tyder på att det finns ett någorlunda linjärt samband mellan dessa två variablerna. Vi uppmärksammar även att korrelationen mellan snittpoängen och åldern är svag (ett värde på -0.01), vilket då medför att det inte finns ett bakomliggande linjärt samband.

Vidare kan man även rita ut scatterplots med snittpoängen mot de två kontinuerliga förklarande variablerna, eftersom det såväl ger oss en helhetsbild som möjligheten att hitta avvikande observationer. Till att börja med plottar vi ut snittpoängen mot ålder.

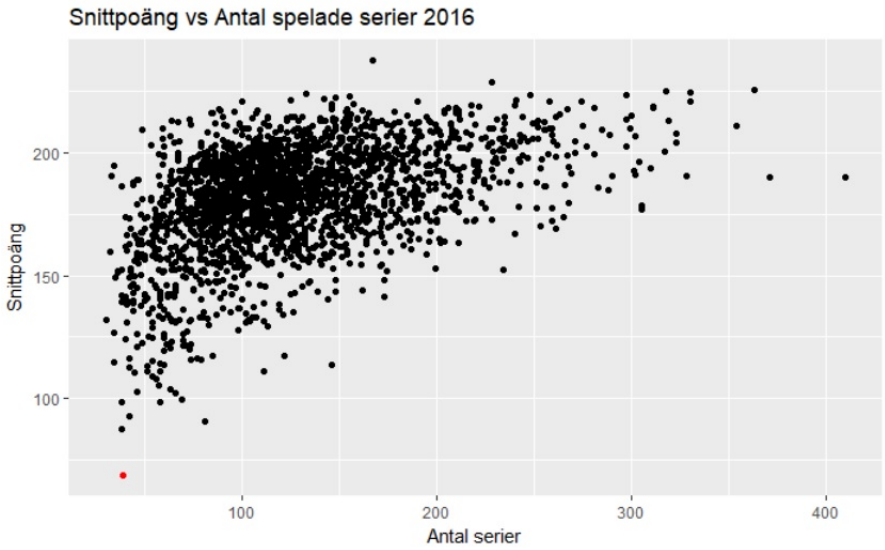


Figur 2: Scatterplot av snittpoäng mot ålder.

Tittar man på Figur 2, ser vi att det inte verkar finnas ett linjärt samband mellan snittpoängen och ålder, vilket även var förväntat då vi såg att dessa två variablerna var nästintill okorrelerade. Det verkar mer lämpligt att anpassa observationerna med ett andragradspolynom, vilket även kan verka mer logiskt. En bowlingspelare mellan 8-18 år börjar antagligen precis lära sig sporten och har därför ännu inte fått en hög snittpoäng. Vi noterar även att man verkar vara som bäst runt 25-30 års åldern. Därefter verkar snittet sjunka långsamt, vilket antagligen beror på den fysiska kapaciteten som (i allmänhet) blir sämre desto äldre man blir, vilket i sin tur förmodligen sänker snittpoängen.

Vi märker även att vissa observationer avviker mer än andra, då några bowlingspelare verkar ha ett betydligt lägre snitt än majoriteten (se rödmarkerad

observation som exempel). Vi kan dock inte anta att detta är felaktiga data, då vissa personer inte är kapabla till att prestera högre än det poängsnittet. Det kan antingen bero på att man inte utövat sporten tillräckligt länge, eller att man spelar så kallad 'parabowling' [10], där den fysiska kapaciteten är begränsad. Hursomhelst så kan vi inte anta att det är felaktiga observationer. Vidare plottar vi även ut snittpoängen mot antal spelade serier.



Figur 3: Scatterplot av snittpoängen mot antal spelade serier.

Från Figur 3, märker vi även här att snittpoängen helt och hållet inte verkar bero på antal spelade serier linjärt, vilket även såg tidigare då snittpoängen och antal spelade serier antog en korrelation på 0.46. I allmänhet verkar snittpoängen öka upp till man spelat ca 100 serier. Därefter ökar fortfarande snittet, men inte alls likt starkt, för att därefter plana ut. Detta tyder på att det istället verkar finnas något logaritmiskt samband mellan de två variablerna.

Även här kan vi inte säga något om felaktiga observationer, av samma anledning som det tidigare argumentet. Vi kan ändock notera att det låga snittet för den rödmarkerade observationen (som är densamma från Figur 2), kan bero på att hen inte spelat tillräckligt många serier.

4.2 Regressionsanalys

I detta avsnitt kommer vi använda oss av multipel regression. Vi kommer dels använda oss av den bakomliggande teorin som nämns i Avsnitt 3.1.2 och dels använda resultatet från Avsnitt 4.1, där vi såväl såg att det verkade lämpligt att anpassa ett andragradspolynom mellan variablerna snittpoängen och ålder som att det fanns ett logaritmiskt samband mellan variablerna snittpoängen och antal spelade serier.

4.2.1 Multipel linjär regression

Innan vi utför självaste regressionen, påminner vi återigen om att vi tittar på Tabell 3, vilket illustrerade hur väl en bowlingspelare presterade oavsett spelform och årstid. Vi beräknade då ut varje persons poängsnitt genom att dividera den totala poängen varje individ erhållit under år 2016 med hur många serier det motsvarar. Anledningen till att vi istället inte använder oss av poängsnittet från Tabell 2, det vill säga hur väl varje person presterat under respektive spelform och årstid, är för att vi då istället har fyra observationer per individ. För att använda multipel linjär regression, vet vi en av modellförutsättningarna är att samtliga observationer måste vara sinsemellan oberoende. Ifall vi då istället undersöker datamaterialet från Tabell 2, så uppfylls inte detta kravet på grund av att de 4 mätningarna per individ är beroende av varandra (det är trots allt samma person). Vidare kommer vi inkludera den kategoriska variabeln *kön*, som även övergår till *Dummy-variabel*, där *Man = 1* och *Kvinna = 0*. Fortsättningsvis kommer vår inledande modell se ut enligt följande

$$Y_i = \alpha + \beta_{Kön} \cdot x_{1i} + \beta_{Antal\ serier} \cdot\ x_{2i} + \beta_{Ålder} \cdot x_{3i} + \epsilon_i. \tag{4.1}$$

Vid utförande av multipel linjär regression får vi att $R^2 = 0.2905$ och $R^2_{adj} = 0.2896$ samt att p-värdet $= 2.2 \cdot 10^{-16}$. Tittar vi närmare på residualerna för modellen (se Appendix, Figur 9), är de symmetrisk fördelade runt noll med approximativt konstant varians. Dock är inte residualerna helt normalfördelade, vilket vi måste ta hänsyn till genom att använda oss resultatet ifrån Avsnitt 4.1 där vi visade att snittpoängen och våra förklarande variabler ålder samt antal spelade serier inte förhöll sig enligt ett linjärt samband. Vi konstaterade att det verkade mer lämpligt att anpassa ett andragradspolynom mellan variablerna poängsnitt och ålder. Vi utnyttjar detta i regressionen, så att vår nya modell istället ser ut som följer

$$Y_i = \alpha + \beta_{Kön} \cdot x_{1i} + \beta_{Antal\ Serier} \cdot x_{2i} + \beta_{Ålder} \cdot x_{3i} + \beta_{Ålder^2} \cdot x^2_{3i} + \epsilon_i. \tag{4.2}$$

Från detta kan vi utföra en ny regression som ger oss förklaringsgraderna $R^2 = 0.3521$ och $R^2_{adj} = 0.3509$. Detta tyder på att (4.2) är en bättre förklarande modell i jämförelse med (4.1). Återgår vi till residualerna (se Appendix, Figur 10), är de fortfarande symmetriskt fördelade runt noll med approximativt konstant varians, samtidigt som de inte är normalfördelade. Vi måste givetvis

ta det i beaktande och försöka hitta en ännu bättre modell. Då det verkade finnas ett bakomliggande (naturligt) logaritmiskt samband mellan variablerna snittpoäng och antal serier, använder vi detta med mål om att erhålla en ännu bättre modell. Den nya modellen ser då istället ut enligt

$$Y_i = \alpha + \beta_{Kön} \cdot x_{1i} + \beta_{log_e(Antal\ Serier)} \cdot log_e(x_{2i}) + \beta_{Ålder} \cdot x_{3i} + \beta_{Ålder^2} \cdot x_{3i}^2 + \epsilon_i, \quad (4.3)$$

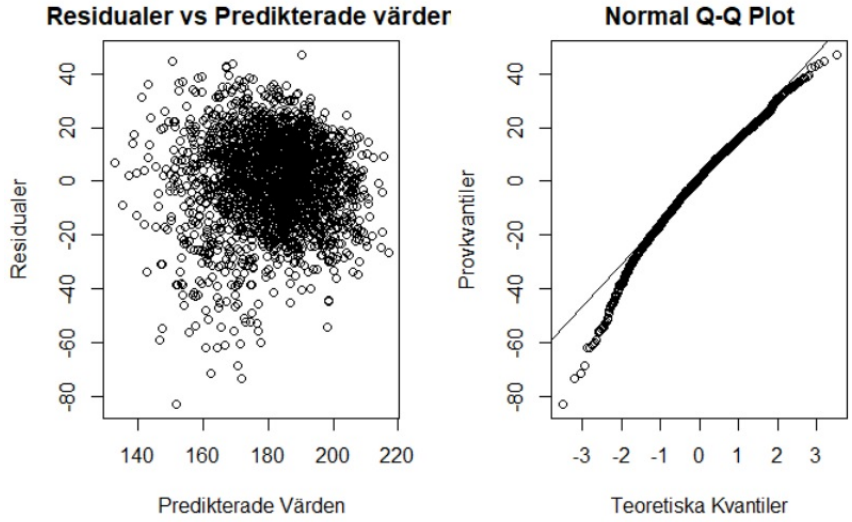
vilket återigen kommer ge ett nytt resultat för regressionen. Nu väljer vi även att illustrera variablernas parameterskattningar, med tillhörande 95%-iga konfidensintervall, tillsammans med förklaringsgraderna och residualernas standardavvikelse.

Tabell 4: Resultat av multipel linjär regression med de förklarande variablerna log_e(Antal serier), Ålder, Ålder^2 och Kön (Man = 1, Kvinna = 0).

Variabel	Parameterskattning	Undre gräns	Övre gräns	Pr($>\|t\|$)
Intercept	16.66	7.42	25.90	0.0004
Ålder	1.66	1.45	1.87	$< 2 \cdot 10^{-16}$
Ålder^2	-0.019	-0.022	-0.017	$< 2 \cdot 10^{-16}$
log_e*(Antal Serier)*	26.14	24.48	27.81	$< 2 \cdot 10^{-16}$
Kön	13.75	12.07	15.44	$< 2 \cdot 10^{-16}$

R^2	R^2_{adj}	p-värde	Residualernas standardavvikelse
0.3928	0.3918	$< 2.2 \cdot 10^{-16}$	16.65

Vi har lyckats erhålla en ännu bättre modell, ty R^2 och R^2_{adj} är högre för (4.3) i jämförelse med (4.2). Vi ser även att skattningarna för samtliga parametrar är signifikanta (på 5%-nivån). Tittar man återigen på residualerna för modellen ser de ut enligt följande



Figur 4: Scatterplot med residualerna mot predikterade värdena samt tillhörande QQ-plot för modell (4.3)

Från högra plotten i Figur 4 ser man att residualerna verkar närma sig en normalfördelning. Vidare ser vi även från den vänstra plotten, att residualerna är symmetrisk fördelade runt noll med approximativt konstant varians. En av de viktigaste frågorna just nu är ifall vi kan erhålla en ännu bättre modell.

Om vi prövar utföra ytterligare transformationer (både för förklarande variabler och responsvariabeln), sjunker både R^2 och R^2_{adj}, vilket tyder på att (4.3) fortfarande är den hittills bästa modellen. Försöker vi istället ta med samspel mellan de förklarande variablerna, som exempelvis *kön* och log_e*(Antal serier)*, så blir ingen av dessa signifikanta samtidigt som förklaringsgraderna minskar. Slutligen nämner vi även att ifall vi försöker använda oss av så kallade *splines* [6], det vill säga en kontinuerlig funktion med styckvisa polynom och kontinuerlig förstaderivata, så erhålls ingen bättre slutmodell. Med andra ord verkar (4.3) fortfarande vara den bäst lämpade modellen vi kommit fram till genom multipel regression.

Slutligen kvarstår att tolka parameterskattningarna från Tabell 4. Inledningsvis skattades *Ålder* och *Ålder*^{2} till 1.66 och -0.019 respektive, vilket kan ses som att en person bli bättre fram tills att han eller hon är 43 år gammal. Därefter kommer poängsnittet i allmänhet sjunka vilket antagligen beror på att den fysiska kapaciteten börjar bli begränsad. Vidare är log_e*(Antal Serier)* skattad till 26.14, vilket kan tolkas som att snittpoängen ökar med 26.14 då man spelar $e \approx 2.72$ gånger fler serier. Slutligen erhöll *Kön* skattningen 13.75 vilket kan ses som att män i allmänhet har 13.75 högre snittpoäng än vad kvinnor har, vilket kan bero på att damerna har en mer begränsad fysisk kapacitet.

Genom multipel regression kom vi alltså fram till resultatet vi precis beskrev, men vi vill även försöka inkludera de förklarande variablerna *period* och *match/tävling* för att se om såväl årstiderna som typen av spelform ger någon bidragande effekt på poängsnittet. För att ta med dessa förklarande variablerna måste vi använda oss av en *blandad modell*, vilket görs i nästa avsnitt.

4.3 Mixed model

Tidigare beskrevs vilka två olika typer av poängsnitt vi är intresserade över under arbetets gång. Den ena var poängsnittet för varje bowlingspelare oavsett spelform (match/tävling) och årstid (vår/höst), för år 2016. Denna snittpoängen använde oss av i föregående avsnitt när vi anpassade en modell med hjälp av multipel regression. Den andra sorten av snittpoäng visade hur väl varje bowlingspelare presterat under år 2016 med hänsyn till de två spelformerna och årstiderna. För denna data erhöll varje individ fyra olika snittpoäng för varje kombination av årstid och spelform, vilket vi beräknade genom att dividera den totala poängen för en specifik spelform och årstid med antalet serier det motsvarade (se Tabell 2 för exemplifiering av datamaterialet). Fördelen med denna sorten av snittpoäng är att vi kan inkludera de kategoriska variablerna *period* och *match/tävling*, men till kostnaden av att var och en av observationerna för varje enskild individ är beroende av varandra (det är trots allt samma person). Som tur är, finns det statistiska metoder som löser problemet med beroendet. En av dessa kallas för en *blandad modell* (mixed model), vilket är vad följande avsnitt kommer handla om.

En blandad modell är en statistisk modell som både består av fixa och slumpmässiga variabler, vilket kan jämföras med multipel regression som enbart behandlar fixa faktorer. Denna modell används ofta när man har upprepade mätningar inom varje individ under en viss tidsperiod. När vi undersöker snittpoängen med fyra observationer per person, kan man se dessa som fyra upprepade mätningar hos varje individ under år 2016. Med våra variabler, kommer vi se *Id* som slumpmässiga faktor och *kön, period, match/tävling, ålder* samt *antal spelade serier* som fixa variabler. Eftersom individ är den enda faktorn som är slumpmässig kommer modellen se ut enligt

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{4.4}$$

där $\boldsymbol{y}_i = [y_{i1}, y_{i2}, y_{i3}, y_{i4}]^T, \quad (i = 1, \ldots, 2279)$ eftersom vi har 4 upprepade mätningar för 2279 individer. Vidare antas även att $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i)$, där man måste specificera kovariansstrukturen för $\boldsymbol{R}$.

4.3.1 Val av kovariansstuktur

Kovariansstrukturen för $\boldsymbol{R}$ kan se olika ut beroende på ett antal faktorer. En viktig del av analysen är att bestämma vilken struktur som bör användas. Det finns en flertals kovariansstrukturer som inte nämns i Avsnitt 3.3, där många

av dessa (likt AR(1)-strukturen) beror på mätningarna mellan tid. I vårt fall kan vi bortse från tidsaspekten eftersom det inte är specificerat när självaste mätningarna är gjorda, utan vi vet enbart ifall ett speltillfälle inträffade på våren eller hösten. Istället kommer analysera de mer 'vanliga' strukturerna, CS (Compound Symmetry) och UN (Unstructured). För att avgöra vilka av dessa två vi ska använda oss av, kommer AIC-värdena (se Avsnitt 3.4) vara användbara. Vi ställer upp olika modeller och jämför vilken av kovariansstrukturerna som ger lägst AIC-värde. Detta kommer även göras med flera modeller bara för att säkerställa att vi väljer rätt kovariansstruktur. Vid utförande fås följande

Tabell 5: Skillnad i AIC-värden mellan kovariansstrukturerna Compound Symmetry (CS) och Unstructured (UN).

Modell	Kovariansstruktur	AIC
Ålder, Ålder^2, log_e *(Antal serier), Kön, Period, Match/tävling*	CS	71350
	UN	**71125**
Ålder, Ålder^2, log_e *(Antal serier), Period, Match/tävling*	CS	71570
	UN	**71362**
Ålder, Ålder^2, log_e*(Antal serier), Kön, Match/tävling*	CS	71615
	UN	**71450**
Ålder, Ålder^2, log_e*(Antal serier), Kön*	CS	71880
	UN	**71717**
Ålder, Ålder^2, log_e*(Antal serier)*	CS	72098
	UN	**71935**

Vi noterar omedelbart från Tabell 5 att den ostrukturerade kovariansstrukturen ger lägst AIC-värden, oavsett vilken av modellerna vi kollar på. Alltså kommer vi använda oss av denna i fortsättningen när vi ska bestämma parameterskattningarna för de fixa variablerna.

4.3.2 Modellval

Vidare behöver vi välja modellens fixa parametrar. Vi har redan konstaterat att *Id* är vår slumpmässiga faktor samt att variablerna *ålder, antal spelade serier, kön, match/tävling* och *period* är våra fixa effekter. Den observante har även noterat att vi har behållit våra tidigare transformeringar, nämligen anpassningen av ett andragradspolynom mellan variablerna *snittpoäng* och *ålder*, men även att vi använde oss av faktumet att det finns ett logaritmiskt samband mellan responsvariabeln och faktorn *antal serier*. Den naturliga frågan vi måste ställa oss nu är; ska vi ta bort någon/några förklarande variabler? Den modell vi har just nu är

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{4.5}$$

med följande variablers parameterskattningar (samt tillhörande 95%-iga konfidensintervall) och standardavvikelser

Tabell 6: Resultat vid utförande av regression för en blandad modell, med Id som slumpmässig faktor samt kön (man = 1, kvinna = 0), period (vår = 1, höst = 0), match/tävling (tävling = 1, match = 0), ålder och log_e(antal serier) som fixa variabler.

	Id	Residualer
Standardavvikelse	18.61	8.25

Variabel	Parameterskattning	Undre gräns	Övre gräns	p-värde
Intercept	133.84	128.83	138.84	< 0.0001
Ålder	1.43	1.18	1.67	< 0.0001
Ålder^2	-0.018	-0.020	-0.015	0.0081
log_e*(Antal serier)*	2.76	2.41	3.11	< 0.0001
Kön	14.81	12.89	16.73	< 0.0001
Period	3.17	2.83	3.51	< 0.0001
Match/tävling	2.79	2.39	3.18	< 0.0001

Vi väntar tills vidare med att kommentera skattningarna för de fixa faktorerna i Tabell 6. Istället observerar vi att samtliga variabler är signifikanta (på 5%-nivån). Vanligtvis är det naturligt att eliminera en parameter som inte är av signifikans och jämföra AIC-värdena mellan de två olika modellerna. Men eftersom samtliga skattningar är av signifikans, så är det omotiverat att göra något sådant. Vi kan ändock nämna att, om vi skulle ta bort en parameter i taget ifrån den fulla modellen (4.5), så ökar AIC-värdet markant, vilket är ytterligare ett resultat som stödjer att borttagning av parametrar är omotiverat. Fortsättningsvis är det å andra sidan intressant att undersöka samspelet mellan vissa variabler. I regressionsavsnittet (Avsnitt 4.2.1), när variablerna *Period* och *Match/tävling* inte var inkluderade, undersökte vi samspelet mellan variablerna men såg att ingen av samspelsfaktorerna var av signifikans. Om vi istället undersöker samspelet mellan variablerna *Period* och *Match/tävling* så visar det sig att denna är signifikant på 5% nivån. Dock bör man notera att AIC-värdet gick ner till 71098 i jämförelse med 71125 för modell (4.5), vilket inte är en stor minskning. Vi bör dels tänka på att AIC-värdet alltid ökar när man tillför fler parametrar (se formel (3.11)), men även att vi får en mer lättolkad modell utan samspelsfaktorn. Till följd av detta, kommer vi inte ta med samspelet mellan de två variablerna i vår modell. Med andra ord så verkar (4.5) vara en lämplig kandidat som en slutgiltig modell.

Vidare noterar vi standardavvikelserna i Tabell 6. Vi ser först och främst att standardavvikelserna *inom* varje observation (8.25) är lägre än *mellan* (18.61) individerna. Det betyder med andra ord att snittpoängen varierar mer mellan varje bowlingspelare än vad den gör för varje individuellt *Id*. Något som även är

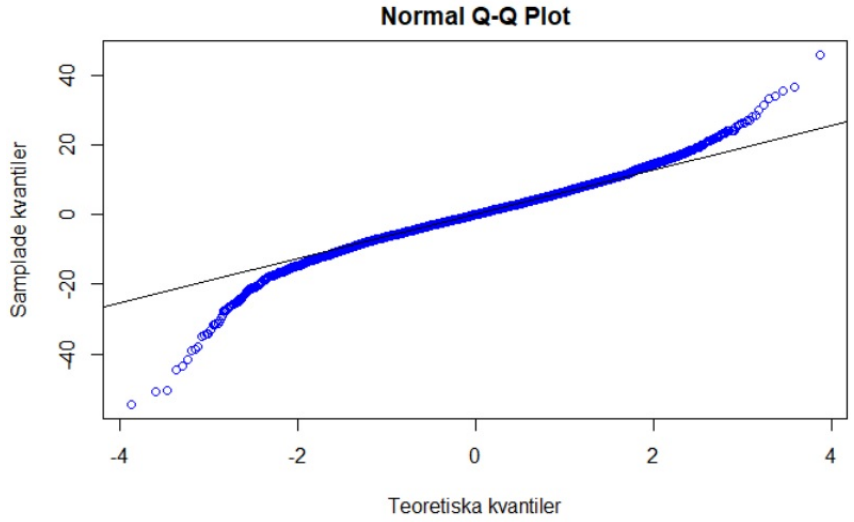
av intresse är de *obetingade* korrelationerna inom varje individ, vilket avslöjar ifall det var befogat att anta att observationerna inom individ är beroende av varandra. I vårt fall kommer vi erhålla 6 stycken korrelationer, där var och en representerar beroendet mellan samtliga kombinationer av spelformerna och årstiderna. För modell (4.5) får vi följande korrelationer

$$\begin{cases} \rho_{Höstmatch,\ Vårmatch} = 0.88 \\ \rho_{Höstmatch,\ Hösttävling} = 0.82 \\ \rho_{Höstmatch,\ Vårtävling} = 0.85 \\ \rho_{Vårmatch,\ Vårtävling} = 0.80 \\ \rho_{Vårmatch,\ Vårtävling} = 0.86 \\ \rho_{Hösttävling,\ Vårtävling} = 0.81 \end{cases} \tag{4.6}$$

Eftersom samtliga korrelationer är höga, så visar det sig att observationerna inom individ är beroende av varandra, vilket alltså innebär att vårt antagande var befogat. Fortsättningsvis är det även intressant att undersöka korrelationerna *betingat* på *Id*. När vi beräknar dessa, och det exempelvis visar sig att korrelationen mellan tävlingar och matcher på hösten antar en hög korrelation, så innebär det att en hög snittpoäng för matcher under hösten även bör indikera på en hög snittpoäng för tävlingar under hösten. Med 'hög' menar vi då högre än den totala individuella snittpoängen. Vi beräknar dessa och erhåller följande korrelationer

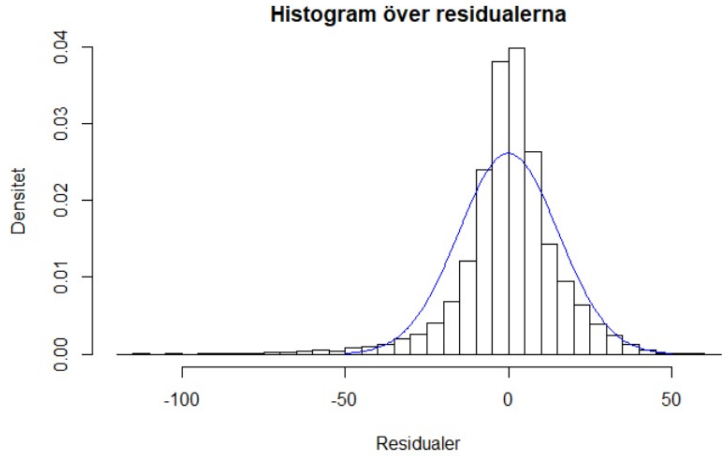
$$\begin{cases} \rho_{Höstmatch,\ Vårmatch|Id} = 0.264 \\ \rho_{Höstmatch,\ Hösttävling|Id} = -0.103 \\ \rho_{Höstmatch,\ Vårtävling|Id} = 0.099 \\ \rho_{Vårmatch,\ Vårtävling|Id} = -0.203 \\ \rho_{Vårmatch,\ Vårtävling|Id} = 0.114 \\ \rho_{Hösttävling,\ Vårtävling|Id} = -0.175 \end{cases} \tag{4.7}$$

Vi noterar från (4.7) att korrelationerna dels är låga men antar även såväl positiva som negativa värden. Detta gör det svårt för oss att se det bakomliggandet sakskälet. Vidare behöver man, likt det som gjordes i Avsnitt 4.2.1, undersöka modellens residualer som ska uppfylla vissa antaganden. Under detta avsnitt har vi antagit att de slumpmässiga felen följer en multivariat normalfördelning med ett medelvärde som ska ligga vid 0 och en specifik kovariansstruktur. Residualerna bör då följa en normalfördelning, vilket kan undersökas genom en QQ-plot enligt



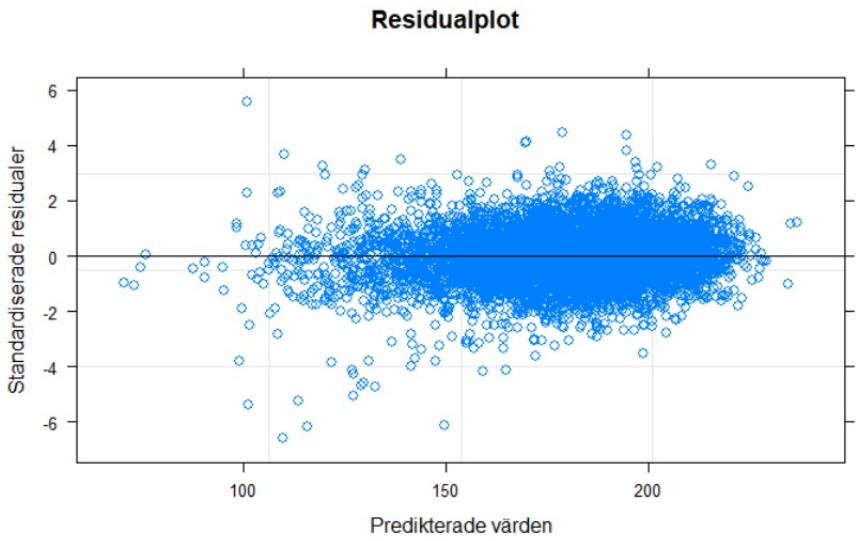
Figur 5: Normalkvantilplot för modell (4.5).

Från Figur 5, som illustrerar en QQ-plot för residualerna, ser man att det inte följer en rak linje, vilket alltså innebär att normalfördelningsantagandet inte är uppfyllt. Istället verkar linjen följa någon slags sigmoid form, vilket innebär att residualerna antagligen följer en symmetrisk fördelning som inte är normalfördelad *(Applied Mixed Models in Medicine, 2nd ed., H.Brown R. Prescott)*. Till följd av detta brukar det vara motiverande att utföra någon slags transformation av responsvariabeln eller någon av de förklarande variablerna. Vi försökt utföra detta i tidigare avsnitt utan något lyckatresultat, vilket dessvärre (och som förväntat) blir detsamma i detta fall. Vi noterar även från Figur 5 att det inte finns någon indikation av avvikande värden. Vidare plottar vi istället ut ett histogram med residualerna för att få en tydligare bild av fördelningen, vilket illustreras i Figur 6.



Figur 6: Histogram över residualer.

Från Figur 6 ser man tydligare att residualerna inte normalfördelade och att det till och med finns en mindre skevhet. En viktig fråga är *varför* vi inte uppnår en normalfördelning. En av anledningarna kan vara att det finns en 'maxgräns' i vad man kan få för poäng inom bowlingen (nämligen 300), som gör att snittpoängen inte kan variera hur mycket som helst. En annan faktor kan vara att snittpoängen hos de olika licensierade bowlarna varierar olika beroende på vilken snittpoängsnivå man ligger på. En 'lågpoängssnittare' som (oftast) är äldre, brukar alltid ha en snittpoäng runt 150 oavsett speltillfälle medan en spelare med ett högre snitt (>210) kan ha en dålig dag och därför snitta 150 vid ett specifikt speltillfälle [8]. Detta skapar en större varation för spelare med högre snitt, som i sin tur kan ligga i grund till skevheten i fördelningen. Ett bra sätt att undersöka detta, är genom att plotta residualerna mot de predikterade värdena. Om vår hypotes att personer med högre snitt har större variation mellan speltillfällena bör det visa sig i denna figur. Denna plot visar även hur väl residualernas varians ligger kring noll, vilket är ytterligare en förutsättning som behöver vara uppfyllt.



Figur 7: Residualer plottade mot predikterade värdena.

Från Figur 7 är det svårt med blotta ögat att undersöka spridningen mellan de som har en låg respektive hög snittpoäng, då de flesta punkterna är samlade som en klump. Vidare ser vi ändock att residualerna är symmetrisk fördelade runt noll med en approximativ konstant varians.

Det finns givetvis andra påverkande faktorer som kan förklara varför vi inte uppfyller normalfördelningsantagandet. Det finns en risk att vår kovariansstruktur inte är korrekt eller att det finns vissa transformationer som vi missat att göra. Sedan finns det naturligtvis andra metoder som kan behandla detta, vilket man kan läsa mer om i diskussionsavsnittet (Avsnitt 6).

4.3.3 Tolkning av fixa variabler

Denna del kommer behandla tolkningar av de fixa variablerna. Vi såg att eftersom samtliga skattningar var av stor signifikans, så var borttagning av några fixa faktorer inte givande. Vi såg även att varken såväl transformeringar av variabler som samspelseffekt mellan faktorer gav ett bättre resultat. Alltså är det parameterskattningarna som illustrerades i Figur 6 som gäller, vilket vi återigen visar nedan.

Tabell 6: Resultat vid utförande av regression för en blandad modell, med Id som slumpmässig faktor samt kön (man = 1, kvinna = 0), period (vår = 1, höst = 0), match/tävling (tävling = 1, match = 0), ålder och log_e(antal serier) som fixa variabler.

Variabel	Parameterskattning	Undre gräns	Övre gräns	p-värde
Intercept	133.84	128.83	138.84	< 0.0001
Ålder	1.43	1.18	1.67	< 0.0001
Ålder^2	-0.018	-0.020	-0.015	0.0081
log_e*(Antal serier)*	2.76	2.41	3.11	< 0.0001
Kön	14.81	12.89	16.73	< 0.0001
Period	3.17	2.83	3.51	< 0.0001
Match/tävling	2.79	2.39	3.18	< 0.0001

Inledningsvis börjar vi med att titta på en av skattningarna som verkar påverka poängsnittet mest, nämligen könet. Från Tabell 6 ser vi att denna skattades till 14.81, vilket innebär att männen överlag har 14.81 högre snittpoäng än vad kvinnorna har. Detta kan bero på att herrbowlare (i allmänhet) har en bättre fysisk kapacitet som i sin tur bidrar till en högre hastighet och rotation i bowlingklotet.

Vidare hade vi variablerna *Ålder* och *Ålder*^{2} som fick skattningarna 1.43 och -0.018 respektive. Med andra ord kan vi hävda att poängsnittet i allmänhet ökar upp till att man är cirka 39 år gammal, då man tydligen är som bäst. Därefter sjunker poängsnittet, vilket (som tidigare nämnt) troligtvis beror på att den fysiska kapaciteten blir sämre.

Variabeln log_e*(Antal serier)*, det vill säga det (naturliga) logaritmiska värdet av parametern *antalet spelade serier*, blev uppskattad till 2.76. Det betyder att snittpoängen ökar med 2.76 då man spelar $e \approx 2.72$ gånger fler serier. Att poängsnittet ökar då man spelar fler serier är inte föga förvånande. Desto fler serier man spelar, desto mer 'tränad' blir man och kan därför prestera bättre. Om vi fortsättningsvis jämför parameterskattningen från Tabell 6 (2.76) med motsvarande värde från Tabell 4 (26.14), så kan man vid första anblick bli förvånad över den stora skillnaden. Detta beror på att då vi undersökte hur antalet serier påverkar snittpoängen i Avsnitt 4.2.1, så visar skattningen (26.14) hur snittpoängen i allmänhet ökar *mellan* individer desto fler serier man spelar. I det här fallet visar parameterskattningen (2.76) hur snittpoängen ökar *inom* varje individ då man spelar fler antal serier.

Något som även bör uppmärksammas är att olika bowlingspelare har spelat olika antal serier under året. Tidigare antog vi att varje person måste spelat minst 5 serier på våren och på hösten samt 5 serier i form av match- och tävlingsspel. Trots detta, finns det personer som har spelat betydligt fler serier än vissa andra. Detta kan blir ett problem eftersom variansen för varje individ måste vara densamma för att modellen ska hålla. Om vi för enkelhetens skull låter

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \alpha_i \sim N(0, \sigma_\alpha^2), \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2),$$

där Y_{ij} är poängen individ i fick under serie j, så ges variansen för denna av

$$\mathbb{V}\text{ar}(Y_{ij}) = \sigma_\alpha^2 + \sigma_\epsilon^2,$$

där σ_α^2 och σ_ϵ^2 är sinsemellan oberoende. Om antalet observationer skulle vara detsamma i varje stickprov, så att vi hade k stickprov med n observationer i varje, hade stickprovsmedelvärdena antagit fördelningen

$$\overline{Y}_{i\cdot} \sim N\left(\mu, \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}\right). \tag{4.8}$$

I vårt fall har vi alltid k stickprov, men antalet observationer varierar mellan varje stickprov. Detta är en konsekvens av att vi hade olika antal spelade serier för olika individer. Eftersom $\mathbb{V}\text{ar}(\overline{Y}_{i\cdot}) = \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}$, så minskar variansen då antalet spelade serier ökar, vilket i sin tur (givetvis) minskar standardavvikelsen. Då vi även menar på att alla individer ska anta samma varians, kan man ifrågasätta vår slutmodell. Å ena sidan, kan man tänka sig att σ_α^2 är mycket större än σ_ϵ^2/n, vilket innebär att effekten av σ_ϵ^2/n inte bidrar till någon större skillnad i variansen. Å andra sidan, kan detta vara en av faktorerna till att modellen inte uppfyllde normalfördelningsantagandet. Hursomhelst, bör man ha i åtanke att varje stickprov inte behöver anta samma varians.

Slutligen har vi de kategoriska variablerna *Period* och *Match/tävling*. Det vill säga, påverkandet av ifall speltillfället var på hösten eller våren samt om det var en match eller tävling. De två effekterna skattades till 3.17 (vår = 1 och höst = 0) respektive 2.79 (tävling = 1, match = 0). Den naturliga tolkningen är då att man i allmänhet spelar bättre på våren och tävlingar. Att man spelar bättre under början på året kan bero på den enkla anledningen att bowlingsäsongen inte pågår under sommaren och att man då är 'otränad' inför hösten, vilket resulterar i ett lägre resultat under höstperioden. Därefter hinner man träna upp formen inför vårsäsongen, som i allmänhet verka öka snittpoängen med 3.17. Att man spelar bättre på tävlingar kan bero på att man under turneringar spelar på samma bana en hel serie (europeisk spelsätt), medan man på matcher spelar mellan två olika banor under en serier (amerikanskt spelsätt). Den intresserade kan läsa *Blå Boken, §C2 punkt 1 & 2*. Hursomhelst, när bowlare då spelar på en och samma bana, får hen fler möjligheter att hitta en 'bättre spellinje'. Tävlingar brukar även i allmänhet omfatta fler serier än matcher (där matcher är av fyra till antal), vilket ger bowlingspelaren ytterligare möjligheter att hitta en bättre spellinje. De bidragande faktorerna som nämns ovan är antagligen bara ett fåtal utav flera orsaker som påverkar att man spelar bättre på tävlingar respektive på våren.

5 Slutsats

Syftet med detta arbete var att undersöka vilka faktorer som påverkar en bowlingspelares snittpoäng. Till vår hjälp hade vi en datasamling som vara hämtad från BITS [8], där vi primärt var intresserade över två olika typer av snittpoäng:

(i) Snittpoängen på hösten respektive våren som spelats i form av både matcher och tävlingar under år 2016.

(ii) Snittpoängen under år 2016 oavsett spelform och årstid.

Inledningsvis analyserade vi (ii), där vi till en början visade att responsvariabeln *snittpoäng* inte förhöll sig linjärt mot de kontinuerliga variablerna *antal serier* och *ålder*. Istället verkade det finnas ett logaritmiskt samband mellan variablerna snittpoängen och antal serier samtidigt som en anpassning av ett andragradspolynom verkade lämpligt mellan variablerna snittpoängen och ålder. Med det delresultatet kunde vi utföra en multipel linjär regression mellan responsvariabeln och de transformerade/anpassade förklarande variablerna tillsammans med den kategoriska variabeln *kön*.

Vidare ville vi även analysera (ii), för att se hur poängsnittet påverkades beroende på när ett speltillfälle inträffade, samt om det var i form av en match eller tävling. Eftersom vi hade 4 beroende observationer för varje licensierade bowlingspelare, använde vi oss av en blandad modell där de slumpmässiga effekternas (det vill säga varje individ) kovariansmatris antog en viss struktur - vilket även lyckligtvis löste problemet med beroendet. Genom att analysera vilken av strukturerna som verkade lämpligast, kunde vi först och främst visa att observationerna inom varje individ var beroende av varandra. Vi fick slutligen även fram parameterskattningar för samtliga förklarande variabler, där var och en av dessa var signifikanta på 5%-nivån. När man utför regression i form av en blandad modell måste man uppfylla antagandet om att residualerna ska följa en normalfördelning med en konstant varians kring noll. Vi såg från Figur 5 & 6 att residualerna inte var normalfördelade, vilket gör att slutresultatet kan ifrågasättas. Varför vi inte uppfyllde detta antagandet kan bero på att vi har en maxgräns för snittpoängen vilket gör att poängen inte kan varierar hur mycket som helst. Det kunde även bero på att 'lågpoängssnittare' har en mindre varians än vad 'högpoängssnittare' har. En ytterligare anledning till att normalfördelningsantagandet inte var uppfyllt, kan vara för att olika individer har spelat olika antal serier. Konsekvensen av detta blir att alla individer inte antar samma varians.

Hursomhelst så visade sig framför allt att männen i allmänhet har högre poängsnitt än kvinnorna, vilket antagligen beror på den fördel i form av fysisk kapacitet som herrar besitter. Vidare märkte vi även att desto fler serier man spelar, desto högre poängsnitt har en bowlingspelare. Poängsnittet ökar dock inte linjärt, utan det finns istället ett logaritmiskt samband mellan hur poängen ökar i takt

med antal spelade serier. Att man spelar bättre beroende på antal spelade serier bör beror på den enkla anledning att desto mer man utövar en sport, desto bättre blir man. Vi såg dessutom att poängsnittet ökar upp till att man ungefär är 39 år gammal, och minskar därefter. Detta delresultat är logiskt då den fysiska kapaciteten blir begränsad vid en viss ålder. Slutligen såg vi även att bowlingspelare i allmänhet blir bättre på våren och att snittpoängen ökar då man spelar en tävling. Anledningen till att man spelar bättre på våren kan bero på ett antal faktorer, men gissningsvis är bowlare otränad på hösten och får istället en bättre spelform under våren. Att man spelar bättre på tävlingar än på matcher, kan ha sin grund i att man (i allmänhet) spelar på samma bana en hel serie på tävlingar medan man spelar på två olika banor under matcher. Detta ger bowlingspelarna fler möjligheter att finna en bättre spellinje på tävlingar.

6 Diskussion

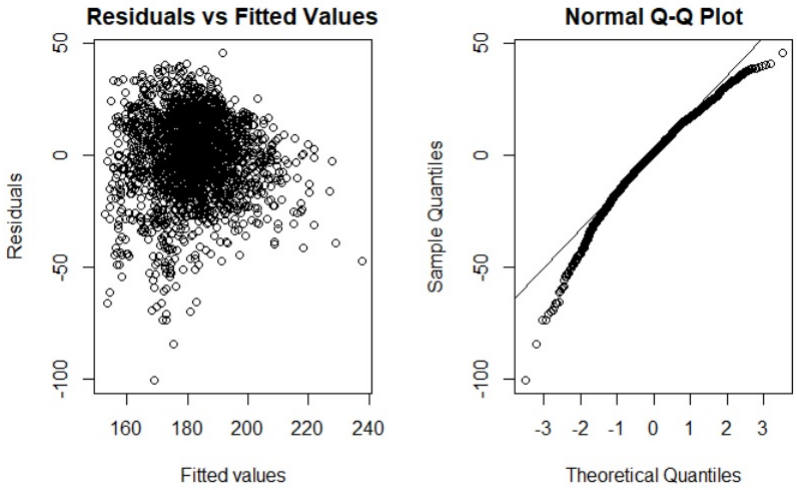
Vi kunde alltså säga att en manlig bowlare runt 39 år som i grund har spelat relativ många serier, bör spela en tävling på våren för att få ett så högst resultat som möjligt. Med tanke på hur få i antal de förklarande variablerna är, så kan slutresultatet anses som någorlunda bra. För att erhålla en bättre modell, hade fler förklarande variabler varit av stort intresse. En av de viktigaste faktorerna som påverkar ett poängsnitt är *antalet år* en person i fråga har utövat sporten. Eftersom bowling är en så kallad 'precisionssport', där siktet är avgörande för resultatet, brukar flera år av utövande ge lovande resultat. Det skulle även vara intressant att ta del av data på individnivå, som exempelvis bowlingklotets *hastighet* och *rotation* när den träffar käglorna. Vi kommer inte gå in på detalj varför dessa två faktorerna påverkar poängsnittet, men i allmänhet brukar högre fart och rotation på bowlingklotet göra så att man får ner fler antal käglor. Vidare kunde man även använt sig av förutsättningarna som fanns i bowlinghallarna. Då menar vi exempelvis vilket *oljemaskin* samt vilken typ av *olja* som de olika bowlinghallarna använde sig av. Slutligen nämner vi den faktorn som nästintill påverkar poängsnittet mest, nämligen vilken *oljeprofil* som matchen/tävlingen spelades på. Vi kommer inte gå in i detalj på vad som menas med en oljeprofil, men när man talar om sportbowling så är det oftast oljeprofilen som har störst inverkan på snittpoängen. Man kan nämligen mäta *svårighetsgraden* på bowlingbanorna med hjälp av något som kallas för *ratio*. För mer information om oljeprofiler, se [12].

Punkterna ovan är ett antal av många faktorer som också påverkar ett poängsnitt. Vidare såg vi även att problemet med att varje licensierad bowlingspelare hade fyra beroende observationer, löstes smidigt med en blandad modell. Vi noterade dock från Figur 5 & 6, att residualerna inte helt och hållet var normalfördelade, vilket gör att man kan ifrågasätta slutmodellen. För att lösa detta problem kunde man istället kanske använt sig av en generaliserad blandad modell, som tillåter att residualerna antar en annan fördelning än normalfördelningen [4].

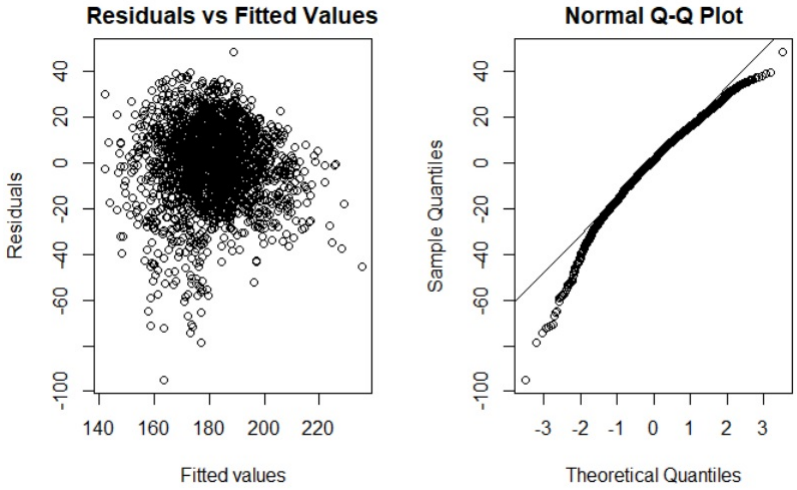
Man kunde dessutom utfört en djupare analys för tidsaspekten. Detta nämndes inte i avsnittet som förklarar datasamlingen, men från första början var vi givna det precisa datumet varje speltillfälle inträffade. Hade vi istället använt oss av detta och gjort ett utförligare analys över tid, kunde vi kanske erhållt en bättre slutmodell. Denna modell skulle måhända fortfarande förklarats med hjälp av en blandad modell, men med en annorlunda kovariansstruktur, som exempelvis AR(1) strukturen som antar en homogen varians och korrelation som minskar exponentiellt mot 0, desto längre tid det är mellan mätningarna. Vidare kunde man då istället försökt visa att de *skalade* residualerna är normalfördelade [5].

7 Appendix

Följande appendix behandlar residualer och tillhörande QQ-plots från Avsnitt 4.2.1.



Figur 9: Residualer och tillhörande QQ-plot för modell (4.2)



Figur 10: Residualer och tillhörande QQ-plot för modell (4.3)

Referenser

[1] A. Agresti, *Categorical Data Analysis, Second Edition*, (2002)

[2] A. Galecki T. Burzykowski, *Linear Mixed-Effects Models Using R*, Springer (2013)

[3] A. Loy, H. Hofman D. Cook, *Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners*, (2017)

[4] C. Anderson, J. Verkuilen T. Johnson, *Applied Generalized Linear Mixed Models: Continuous and Discrete Data*, (2013)

[5] H. Brown & R. Prescott, *Applied Mixed Models in Medicine, Second Edition* (2006)

[6] L. Marsh D. Cormier, *Spline Regression Models*, (2002)

[7] R. Sundberg, *Lineära Statistiska Modeller*, Kompendium (2016)

[8] BITS:
`bits.swebowl.se`

[9] Blå Boken:
`http://www.swebowl.se/globalassets/svenska-bowlingforbundet/2-forbundsinfo/bla-boken/alla`

[10] Svenska Bowlingförbundet: Parabowling
`http://www.swebowl.se/globalassets/svenska-bowlingforbundet-funktionsnedsatta/regel-vad-a`

[11] Wikipedia: Allmänt om bowling
`https://en.wikipedia.org/wiki/Ten-pin_bowling`

[12] Oljeprofiler:
`http://www.kegel.net/wpa/2016/2/24/5-must-know-things-about-oil-pattern-taper`