LUNDS UNIVERSITET

Den historiskt bästa Formel-1 föraren

En studie om den bästa Formel-1 föraren mellan åren 1950-2022

Croneborg, Claes
Sjöberg, Viktor

Handledare: Yvette Baurne, Jonas Wallin

Kandidatuppsats i statistik, 15 hp

Ekonomihögskolan vid Lunds universitet

Statistiska instutitionen

13 januari 2023

Abstract

It is difficult to definitively say who the greatest Formula One driver is. It is a subjective assessment and can depend on a variety of factors, including the driver's skill, their results and accomplishments, their team and car, and the era in which they competed. Some of the most successful and highly regarded Formula One drivers include Michael Schumacher, Ayrton Senna, Lewis Hamilton, Alain Prost, and Juan Manuel Fangio. All of these drivers have achieved a high level of success and are considered among the greatest in the history of the sport. Ultimately, the question of who the greatest Formula One driver is can be a matter of personal opinion, and can vary depending on the criteria used to evaluate them.

Formula One is a highly competitive sport, with drivers consistently vying for the title of the greatest of all time. In Formula One the car has a significant role in a driver's success and failure respectively. In this study, we used Bayesian multilevel beta regression to analyze a dataset of Formula One based of every races and driver across all years, in order to determine the greatest driver in the history of the sport. The dataset included a variety of performance metrics, such as position, year, what car they drove. We found that Jim Clark was the top performer based on our analysis. Our results suggest that Jim Clark was the most consistently successful driver over a 3 years period as well as the highest career peak of all.

Keywords: *Formula One, Multilevel-Model, Beta-Regression, Bayesian Estimation, Monte-Carlo simulation, Markov Chain, Statistics.*

Notation

f - Förare
t - Team/stall/konstruktör
s - Säsong/år
r - Race/tävling
N - Antal förare
p - Slutposition i racet
π - Relativ slutposition $[0, 1]$
π_t - Transformerad relativ slutposition $(0, 1)$
μ - Medelvärde
σ^2 - Varians
$\Gamma(\cdot)$ - Gamma-funktionen
α - Signifikansnivå, vilket genomgående sätts till 5% om inget annat anges.
β - Parameter
$\mathcal{N}$ - Normalfördelning
$\mathcal{S}$ - Tillståndsrum (State pace)
ϕ - Beta-regressionens precisionparameter
$\perp\!\!\!\perp$ - oberoende alltså $X \perp\!\!\!\perp Y \rightarrow X$ och Y är oberoende.
$s.v.$ - Förkortning *Slumpvariabel*
Ω - Utfallsrum
ε - Residual/felterm
θ - Okänd parameter
η - Linjära kombinationen
$\hat{R}$ - Rhat

Innehåll

1 Inledning

En frekvent återkommande fråga och diskussion från såväl idrottare/atleter som dess publik och intresserade är vem av utövarna som varit eller är bäst genom tiderna. Ett fåtal idrotter är något enklare att göra ett uttalande om, men inom merparten av idrotterna blir denna fråga snabbt mycket komplex. Ett definitivt svar på vem den bästa fotbolls- eller hockeyspelaren är mycket en individuell referens men där ett fåtal utövare ofta kommer på tal mer än andra.

Till skillnad från dessa är den högsta nivån inom motorsport, närmare bestämt Formel-1, en sport vars utövare och dess resultat till största del grundar sig i bilen och dess kapacitet. Tidigare studie uppskattar att hela 86% av resultatet beror på bilen vilket ur förarens prestationsmässiga synpunkt står utanför atletens kontroll (Bell m. fl., 2016).

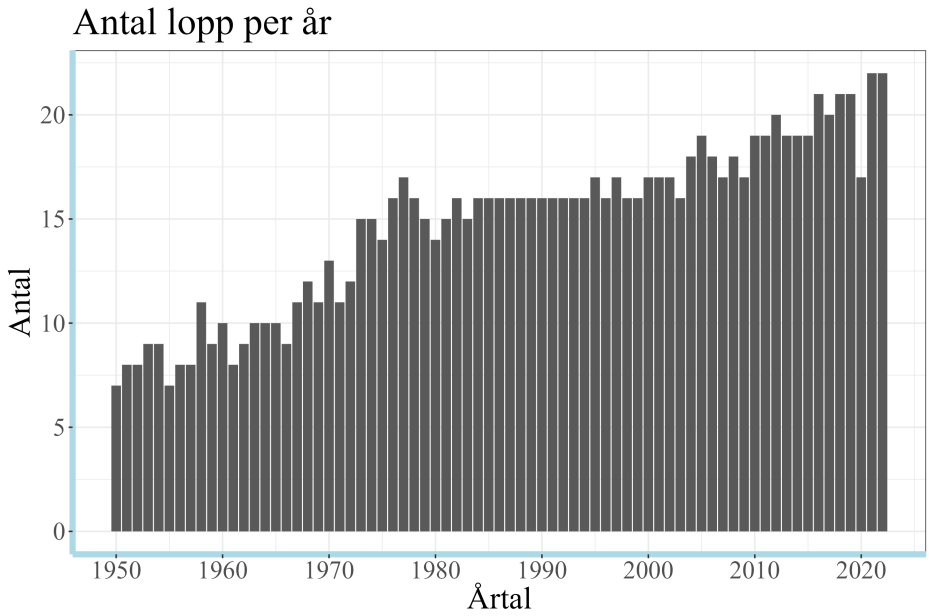
Uppsatsen är i huvudsak indelad i tre delar där del ett består av det inledande kapitlet som är till för att läsaren skall få en överblick av sporten och introduktion av frågeställningen. Del två utgörs av kapitel 2, beskrivning av datamaterialet och dess nödvändiga bearbetning, samt metod och presentation av modell i kapitel 3. Uppsatsen avslutas sedan med presentation av resultatet och diskussion kring detta i kapitel 4 & 5.

1.1 Bakgrund

Formula One, mer känt som Formel-1 är en av världens populäraste sporter med över 400 miljoner tittare under året 2021 (F1, 2022). Från dess början år 1950 tills uppsatsen skrivits har 772 förare från 41 olika nationer tävlat minst en gång i något av de 1079 Grand Prix som under historien anordnats. Idag består Formel-1 av tio olika stall-lag-konstruktörer och 20 förare. Båda har varierat under åren sen Formel-1 startade. Några race bestod av uppemot 50 startande där några av stallen kunde ha fem förare och andra två.

Tävlingshelgen består av träningar, kval på fredag och lördag för att på söndagen avsluta med tävlingen/racet. Startpositionen i racet baseras på kvalet, där förarna under en given tid skall sätta det snabbaste varvet som de kan. Föraren med det snabbaste varvet startar först och föraren med långsammast tid startat sist. Racet består istället av många varv med en minimumdistans om 305 km (Monaco 260 km). Vid längre banor som *"Circuit de Spa-Francorchamps"* (Belgien) körs därmed 44 varv jämfört med kortare banor som *"Circuit de Monaco"* som är 77 varv. Under en säsong har det historiskt körts mellan 7-24 lopp, där antal lopp ökat med åren illustrerat i Figur 1.1 (FIA, 2022).

Till skillnad från de flesta andra motorsports-kategorier, exempelvis; Nascar, Indycar eller långloppsracing även kallade spec-series, är Formel-1 mycket friare i den mening



Figur 1.1: Antal lopp per säsong.

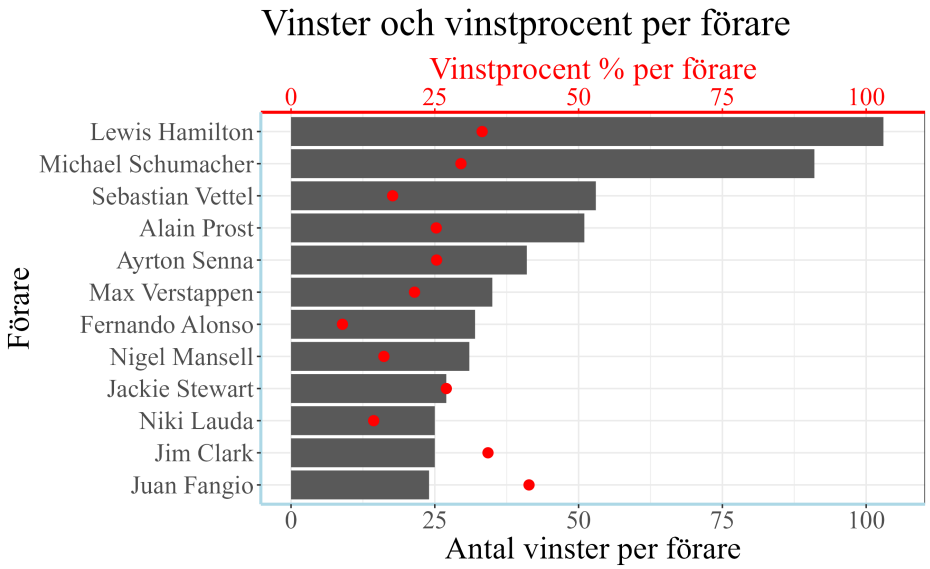
att utveckling och design av bilarna sker från grunden givet ett visst tekniskt reglemente (Complex, 2020). Ganska naturligt leder detta inte enbart till att bilarna i jämförelse med spec-series skiljer sig åt desto mer, utan även att det ligger till grund för den tekniska utveckling som sker både inom men också utanför Formel-1 (Jaques, 2016). En stark bakomliggande faktor till dessa skillnader är organisationernas olika uppbyggnader. Något/några av stallen har 300 anställda och en budget på ungefärligt en miljard kronor/år, till andra som istället har närmare 1500 anställda och en budget om fyra miljarder kronor/år (Pretorius, 2021). Givetvis har detta en direkt påverkan på bilen och dess kapacitet, således också förarens möjlighet till framgång alternativt misslyckande.

I Formel-1 tävlar inte bara förarna mot varandra utan också stallen. Varje stall får poäng efter varje lopp baserat på hur deras förare presterade. I slutet av säsongen krönas konstruktörsmästaren som är den konstruktör med flest poäng. Detta betyder att det inte bara är en bragd för en förare att vinna utan också för stallen som har den bästa kombinationen av förare och bil. Stallen vill ha den bästa föraren de kan få, och förare vill hitta de stall med den bästa bilen. Detta leder till att förare byter stall när de presterar bättre respektive sämre.

I Formel-1 ser vi förare komma och gå, oftast då de inte presterar i enlighet med vad stallet förväntade sig, de skadar sig eller andra orsaker som gör att förare inte har möjligheten att fortsätta köra i Formel-1. Detta gäller också för konstruktörer dock med andra anledningar till att de inte forsätter, ofta av ekonomiska skäl.

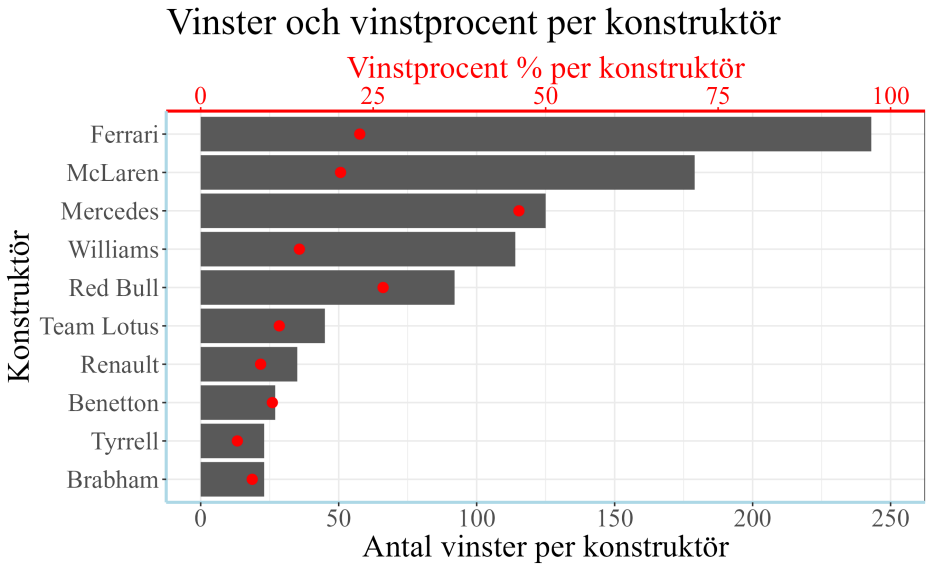
1.2 Syfte och frågeställningar

Även om experter och gamla förare har stor expertis inom ämnet är svaret på vem den bästa föraren högst subjektiv. Ett snabbt 'statistiskt' svar skulle kunna uttryckas i hur många segrar en viss förare tagit, alternativt hur många segrar i relation till hur många lopp denne ställt upp i. Enligt Figur 1.2 skulle svaret lyda att Hamilton tagit flest segrar men Fangio har högst vinstprocent.



Figur 1.2: Topp tolv förare.

I Figur 1.3 ser vi istället att de bästa konstruktörerna skulle vara Ferrari och Mercedes med samma argument.



Figur 1.3: Topp tio konstruktörer.

Att uttrycka den bästa föraren enbart baserat på antalet segrar eller vinstprocent blir snabbt orättvist då bilens kapacitet är den största påverkan på förarens möjlighet till vinst. Dessutom består dagens Formel-1 säsonger av många fler race än tidigare år. Syftet med denna uppsats är att försöka komma fram till ett mer välgrundat uttalande om den bästa föraren genom tiderna. Detta genom att med statistiska metoder lösgöra bilens kapacitet från förarens egna prestation och skicklighet som leder till det slutliga tävlingsresultatet. Frågeställningen för studien formuleras enligt:

1. Vilka är de 15 bästa förarna baserat på förarnas karriärstopp (peak)?
2. Vilka är de 15 bästa förarna baserat på förarnas bästa tre-år-i-följd period?

Denna studie lägger störst vikt vid fråga två och den bästa föraren genom historien kommer att utpekas grundat på detta mått. Kariärstopp kommer också att presenteras i avsikt att ge en mer nyanserad bild men också möjligöra för egen tolkning. Kariärstopp eller peak är den säsongen som en förare presterade som allra bäst. Sist kommer resultat för konstruktörer att presenteras, detta för att visa hur mycket av en prestation som grundar sig i bilens kapacitet. Även här kommer en karriärstopp och bästa 3-års period presenteras.

1.3 Tidigare studier

Även om antalet studier vilka försökt komma till ett uttalande om den bästa föraren genom tiderna inte är av större mängd ställs alla inför samma problem, just att lösgöra bilens kapacitet från förarens prestation. Med varierande metoder och tillvägagångssätt är resultaten inte alltid samma, även om namn som såväl förare som experter nämner och studier oftast fått fram, varit relativt lika.

Studierna skiljer sig alla från varandra i hur utfallsvariabeln sattes. Bell m. fl. (2016) och f1metrics (2019) har båda valt att sätta utfallsvariabeln på samma sätt som poängsystemet för världsmästerskapet i Formel-1, det vill säga att en slutposition ger ett visst antal poäng. Under åren har däremot poängsystemet förändrats flertalet gånger där det exempelvis under perioden 1991-2002 enbart var de sex bäst placerade förarna som fick poängen 10-6-4-3-2-1. Dagens poängsystem vilket implementerades år 2010 utgörs istället av att de tio bäst placerade förarna får poängen 25-18-15-12-10-8-6-4-2-1 (F1, 2022).

För att kunna jämföra förarna över tid behövde båda ovannämnda studier ha ett enhetligt poängsystem för alla förare över alla år. Bell m. fl. (2016) vars studie undersökte den bästa föraren mellan åren 1950-2014 baserade därför utfallsvariabeln med utgångspunkt i poängsystemet använt under tidsperioden 1991-2002. f1metrics (2019) undersökte istället den bästa föraren åren 1950-2019 i enlighet med dagens poängsystem. Till skillnad från den poängbaserade utfallsvariabeln använder Eichenberger och Stadelmann (2009) den faktiska slutpositionen som utfallsvariabel. Kesteren och Bergkamps (2022) utfallsvariabel grundar sig också i slutpositionen men omvandlas till en kvot av andel slagna motståndare som kom i mål, alltså förstaplats = 1 och sistaplats av de som kom i mål = 0. Till skillnad från de tidigarenämnda studierna är den sistnämnda begränsad till åren 2014-2021.

Utöver att utfallsvariablerna skiljer sig åt gör även metoderna det. Eichenberger och Stadelmann (2009) undersöker med hjälp av en 1-nivås regressionsmodell som med hjälp av dummyvariabler försöker kontrollera för konstruktörskapaciteten. Bell m. fl. (2016) och Kesteren och Bergkamp (2022) använder istället en multilevel-level modell då de hävdar att metoden visat sig effektiv för att kunna särskilja individens prestation från grupperingens.

2 Data

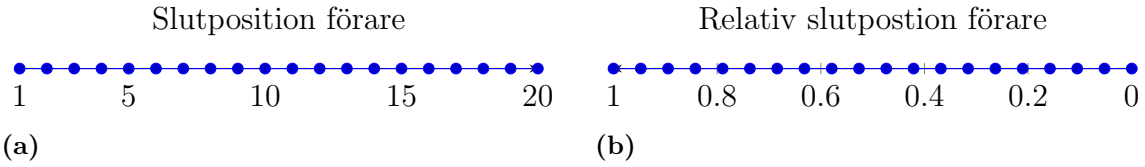
På Formel-1´s egna hemsida finns resultatlistor ända från dess början vilket bör anses vara fullt pålitligt. Resultaten finns dessutom samlade och kan hämtas från Ergast API (Newell, 2022). I tabell 2.1 finns en kortare förklaring över varje enskild datafil som

Datafiler	**Beskrivning**
Circtuis	Information om unika banor
Constructors	Information om unika stall/konstruktör/team
Drivers	Information om unika förare
Races	Information om unika race
Results	Information om samtliga race-resultat
Status	Information om anledning till utkörning/brytning/motorhaveri

Tabell 2.1: Förklaring av datamaterial.

denna studie baseras på. I ett första steg kommer förare som inte fått en slutposition på grund av utkörning, motorhaveri eller annan orsak tas bort från datamaterialet. Kvar blir 14 323 st unika observationer vilka representerar slutpositionerna från de 1079 Grand Prix som hittills arrangerats.

Givet att vi sedermera skall kunna genomföra undersökningen behöver data inhämtas om vad förare f i teamet t fick som slutposition p i ett race under året s. För varje enskilt race kommer även variabeln N ange hur många förare som kom i mål i just det racet. Därefter kommer slutpositionen p likt Kesteren och Bergkamp (2022) transformeras om till den relativa slutpositionen π. π är således ett relativt mått av andelen slagna förare och illustreras i Figur 2.1b.



Figur 2.1: Omräkning till relativ slutposition π_t.

Utöver den relativa slutpositionen kommer även två dummyvariabler v och b införas. v beskriver om racet gick under våta förhållanden, och b om racet arrangerats på en stadsbana. Dummyvariablerna antar enbart 0, eller 1 om det antingen var regn eller stadsbana i respektive fall. Summering av den bearbetade och omräknade datan presenteras i tabell 2.2.

Bearbetad data	**Beskrivning**
f	Förare
t	Konstruktör/Team
p	Slutposition i race
N	Antal som kom i mål vid ett givet race
π	Relativ slutposition (Förstaplats = 1, Sistaplats = 0)
v	Dummyvariabel om vått race
b	Dummyvariabel om stadsbana
s	Årtal

Tabell 2.2: Bearbetat datamaterial.

3 Metod och modellering

Denna studie lägger vikt på hur man på bästa sätt analyserar resultat från data med starkt grupperade inre strukturer. Mer specifikt är varje resultat grupperat i förare, som i sin tur är grupperade inom olika stall/konstruktörer. En metod som visat sig effektiv i att hantera grupperade data är den så kallade multilevel-modellen (Buxton, 2008). Metoden används då man undersöker huruvida individen påverkas av gruppen som den är grupperad i, och i så fall hur stor denna påverkan är. Ett vanligt exempel är om goda studieresultat beror på individens egen studiemotivation och intellektuella kapacitet, eller skolan individen studerar på. Som vad Phillips (2014) kom fram till, att 86% av resultatet beror på bilens kapacitet inser vi snabbt att det finns starka beroendestrukturer. Gruppbehörigheten, det vill säga till vilket stall en förare tillhör, har därför en mycket stark påverkan på utfallet. På liknande sätt som man undersöker studieresultat kommer denna studie försöka lösgöra förarens egna prestation och skicklighet från bilens kapacitet. Studien baseras på empirisk data såväl som simuleringar baserade på inhämtade data. Simuleringarna är utförda i programmet R (statistikprogrammet R).

3.1 Multilevel modelling

En multilevel-modells huvudsakliga syfte är modellering av grupperade data, alltså när en individ/datapunkt är uppbyggd av två eller fler nivåer. Här skiljer man på två olika typer av multilevel-modeller, nämligen om studien är av longitudinell eller hierarkisk struktur. Longitudinella studier används då man inhämtar data från samma individ (eller grupp av individer) över tid. Exempelvis studieresultat från samma person när denne gick i lågstadiet, högstadiet och gymnasium (Goldstein, 2011c). En hierarkisk struktur skulle istället innebära att vi inhämtar studieresultat från olika elever på olika skolor. Eleven (nivå-1) är då grupperad i skolan (nivå-2) (Fox, 2015).

I denna studie inhämtas data från samma population där resultaten är grupperade i stallen de olika förarna tillhör. Resultaten är därför av hierarkisk struktur där nivå-1 är förarna och nivå-2 stallen. Med anledning av att varje stall har flera förare som ibland byter gruppbehörighet är strukturen däremot inte längre strikt hierarkisk (Bell m. fl., 2016). Detta öppnar möjligheten att utnyttja korsklassificering som för modellen ger utökad information om hur en förare presterat i relation till en annan förare, som presterat i relation till en annan etcetera. Med den utökade information har skattningen av parametrar visat sig bli alltmer precisa vid utnyttjandet av denna korsrelation (Goldstein, 2011a). I grupperade datastruktur har vi till skillnad från en vanlig modell ett starkt beroende mellan varje observation. Sättet en multilevel-modell hanterar denna beroendestrukturer är genom att tillåta slumpmässiga variationer på både grupp- och populationsnivå. Alltså tillåts förarna att variera inom stallet där vi estimerar ett intercept, eller lutning eller både och på nivå-1 i modellen.

Det finns olika typer av multilevel-modeller, nämligen *Mixed-effect* och *Random-effect* modeller. Anledningen till uppdelningen är vad för variabler som modellen är uppbyggd av, där de olika typerna är *fixed-variabel* och *random-varibel*. *Fixed-variabel* hanteras som en konstant med ett bestämt värde på alla modellnivåer och utgörs ofta av modellparmeterar exempelvis standardavvikelse eller varians. I *Random-variabel* finnes istället en *random-effect* där slumpmässiga variationer tillåts på de olika nivåerna i modellen (Newsom, 2017). Modeller med både *fixed-* och *random-varibel* benämns därför *Mixed-effects-modell*. Om modellen istället bara är uppbyggd på variabler med en *random-effect* kallas denna för *Random-effects-modell*. I motsats till den vanliga modellen där en *random-effect* enbart finnes i feltermen ε, tillåter multilevel-modeller detta även inom grupperingarna på de olika nivåerna. I dess enklaste form formulerar då Goldstein (2011d) en 2-nivås multilevel-modell enligt ekvation 3.1:

$$\begin{aligned} y_{ij} &= \beta_0 + u_{0_i} + \varepsilon_{0ij} \\ var(\varepsilon_{0ij}) &= \sigma^2_{\varepsilon 0}, \qquad \varepsilon_{0ij} \sim \mathcal{N}(0, \sigma^2_{\varepsilon}) \\ var(u_{0_i}) &= \sigma^2_{u_0}, \qquad u_{0j} \sim \mathcal{N}(0, \sigma^2_{u_0}) \end{aligned} \tag{3.1}$$

Av ekvationen 3.1 ser vi att modellen tillåter en *random-effect* i feltermen ε_{0ij} men också variablerna u_{0j} som båda antas vara normalfördelade. För att göra denna modellering krävs antagandena om att residualerna på nivå-1 är oberoende från varandra och konstanta över hela utfallsrummet Ω. Variansen för responsvariabel y_{ij}, alltså individ i i gruppering j, kan då uttryckas i summan av individens varians adderat med grupperingens varians (Goldstein, 2011c). Detta formuleras i ekvation 3.2.

$$var(y_{ij}|\beta_0, u_{0_i}, x_{ij}) = var(\mu_0 + e0_{ij}) = \sigma^2_{\mu 0} + \sigma^2_{e0}, \qquad \text{där } \sigma^2_{\mu 0} \perp\!\!\!\perp \sigma^2_{e0} \tag{3.2}$$

Då antagandena är uppfyllda kan även måttet *Intra-klass-korrelation ICC* bildas som mäter korrelationen inom grupperingarna. Uttryckt annorlunda är då detta en kvot för hur stor variansen inom stallet är i relation till hela populationen, alltså startfältet. Efter lite matematisk bearbetning formuleras då ICC enligt ekvation 3.3.

$$ICC \ = \rho = \frac{\sigma^2_{\mu 0}}{(\sigma^2_{\mu 0} + \sigma^2_{e0})} \tag{3.3}$$

där: $\sigma^2_{\mu 0}$= *Grupperingens varians*, $\qquad \sigma^2_{e0}$ = *Populationens varians.*

Ovannämnda beräkningar och ansatser ligger alla till grund för att kunna göra estimeringar på ett grupperat datamaterial med starka korrelationer. I och med att förarana i samma stall antages sitta i samma material (bil) reder en multilevel-model ut hur mycket av responsvariabeln y_{ij}'s varians beror av föraren. Genom att även utnyttja information om hur en förare presterade i relation till en annan förbättras skattningen ytterligare vid utnyttjandet av korsklassificering. Den oförklarade variansen som återstår i reponsvariabel y_{ij} vilket av modellen inte kan förklaras kan därför beräknas om till ett mått på förarens skicklighet.

3.2 Beta-regression

Linjär regression är en teknik för att hitta ett linjärt samband mellan en responsvariabel (beroende variabeln) och de förklarande variablerna (oberoende variablerna). Då ett linjärt samband existerar kan modellen prediktera värdet av responsvariabeln vid given information om de oberoende variablerna (Faraway, 2004). För att skattningen skall vara tillförlitlig behöver flertalet grundläggande förutsättningar uppfyllas. Främst behöver det linjära sambandet existera; $E(\varepsilon_i) \equiv E(\varepsilon|x_i) = 0$.

Vid $Beta - regression$ uttrycks det linjära sambandet genom en *link-funktionen* som används för att transformera en linjär kombination och dess koefficienter till en skala av sannolikheter. Parametern av intresse har då supporten $[0, 1]$ och kan formuleras enligt ekvation 3.4 (MacKenzie m. fl., 2018):

$$logit(\mu) = log\left(\frac{\mu}{1-\mu}\right) = \eta = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.4}$$

där η är den linjära prediktorn, β en vektor av parametrar vilka ska skattas av X, designmatrisen av kovariaten, samt slumpmässiga felet ε (Douma och Weedon, 2019).

Residualerna (ε), alltså skillnaden mellan utfallet och det predikterade värdet, skall vara: (1) oberoende och normalfördelade; $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ och (2) residualerna är homoskedastiska, det vill säga konstanta över hela utfallsrummet (Knaub, 2007). Notera att det inte innebär att responsvariablen behöver vara normalfördelad utan enbart residualerna ε. När man vid en multipel linjär regression har flera förklarande variabler kan även problem uppstå vid två eller fler förklarande variabler med starka korrelationer sinsemellan. Modellen lider i så fall av multikolinjäritet vilket ger en missvisande skattning då den underminerar den oberoende variabelns statistiska signifikans (Allen, 1997). Modellen behöver i sådant fall justeras genom att ta bort någon av de variabler vilka är starkt beroende. Då alla antagandena är uppfyllda kan man anta att skattaren är väntevärdesriktig, mest effektiv bland väntevärdesriktiga skattare och har en, åtminstone asymptotiskt, normalfördelad maximum-likelihood skattare (Fox, 2015).

Beta-regression är en typ av regression som används när den förväntade värdemängden på responsvariabeln är begränsad till intervallet 0 till 1. Ofta används denna typ när man vill undersöka procentuella mängder exempelvis marknadsandelar, alternativt sannolikheter. Utöver att Beta-regressionen kan hantera kategoriska variabler har modellen även visat sig vara robust då den hanterar outliers och skev data väl (Harrell Jr, 2015). Responsvariabeln i denna studie följer, med anledning av att vara den relativa slutpositionen, en Beta-fördelning där dess täthetsfunktion formuleras enligt ekvation 3.5:

$$f(y; p, q) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\left((1-\mu)\phi\right)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1 \tag{3.5}$$

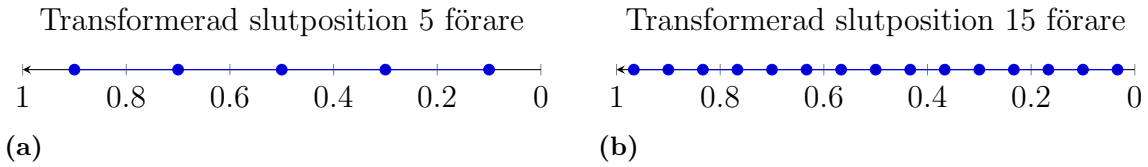
$$\text{där } \mu = \frac{p}{(p+q)} \text{ och } \phi = p + q$$

och $\Gamma(\cdot)$ betecknar gammafunktionen och $E(y) = \mu$ (Ferrari och Cribari-Neto, 2004). ϕ kan ses som en precisionparameter med anledning av att variansen minskar då ϕ ökar när man håller μ konstant och man inser snabbt att ϕ är relaterad till både μ och σ^2 (Cribari-Neto och Zeileis, 2010).

Även om Beta-regressionen anses vara robust och hantera skev data väl har dess största nackdel visat sig vara beräkning vid dess ändpunkter $[0, 1]$. På grund av att precisionsparametern ϕ visat sig vara ej konstant över hela utfallsrummet Ω blir skattaren lidande av högre bias ju närmare responsvariabeln ligger ändpunkterna (Abonazel m. fl., 2022). Biasen gör således att modellen gör systematiska skattningsfel och dessutom generaliserar dåligt på ny data som ger felaktiga prediktioner (Douma och Weedon, 2019). Den relativa slutpositionen π förstår vi snabbt utsätts för detta då responsvariabel både för vinnaren och föraren som kom sist i ett race är 1 repsektive 0. I syfte att motverka och minska biasen i skattaren föreslår Smithson och Verkuilen (2006) därför en transformation av responsvariabel presenterad i ekvation 3.6:

$$y_j = \frac{y_i(N-1) + 0.5}{N}, \qquad y_i \in [0, 1], \quad i \in 1, 2, \cdots, N \tag{3.6}$$

Transformationen av responsvariabeln påverkas av N's storlek, vilket i denna studie representeras av antalet förare som kom i mål ett givet race. Responsvariabeln för vinnaren och den sämst placerade föraren kommer därför variera från race till race beroende på antalet förare som tar målflagg. Illustrerat i Figur ser vi reponsvariabeln för respektive slutposition vid olika N.



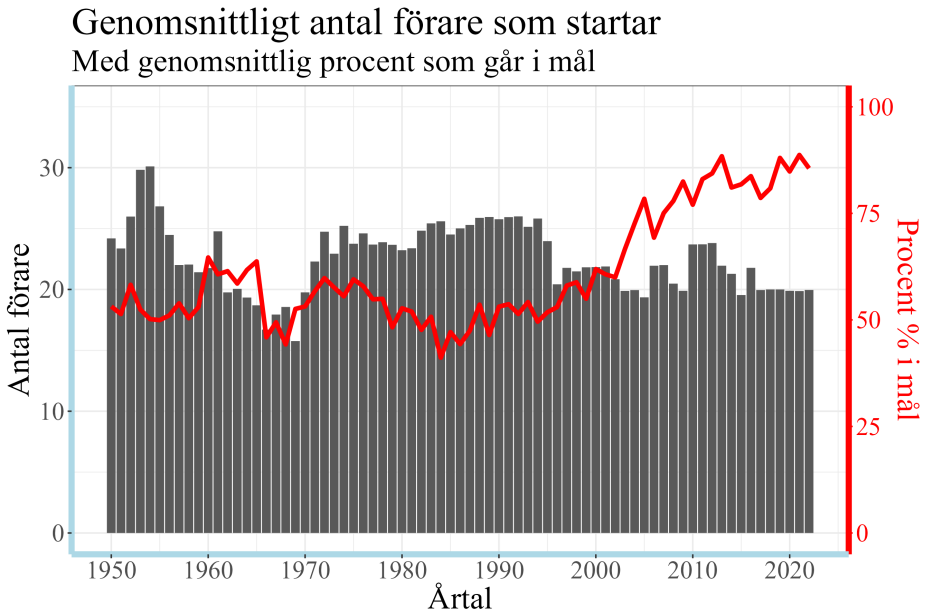
(a) **(b)**

Figur 3.1: Transformerad slutposition vid olika stora N.

Att notera i Figurerna 3.1a & 3.1b är hur nära första- respektive sistaplats tillåts vara ändpunkterna baserat på hur många som kommer i mål. Vid litet N värderas förstaplatsen lägre och sistaplatsen högre, och istället förstaplatsen högre och sistaplatsen lägre vid större N.

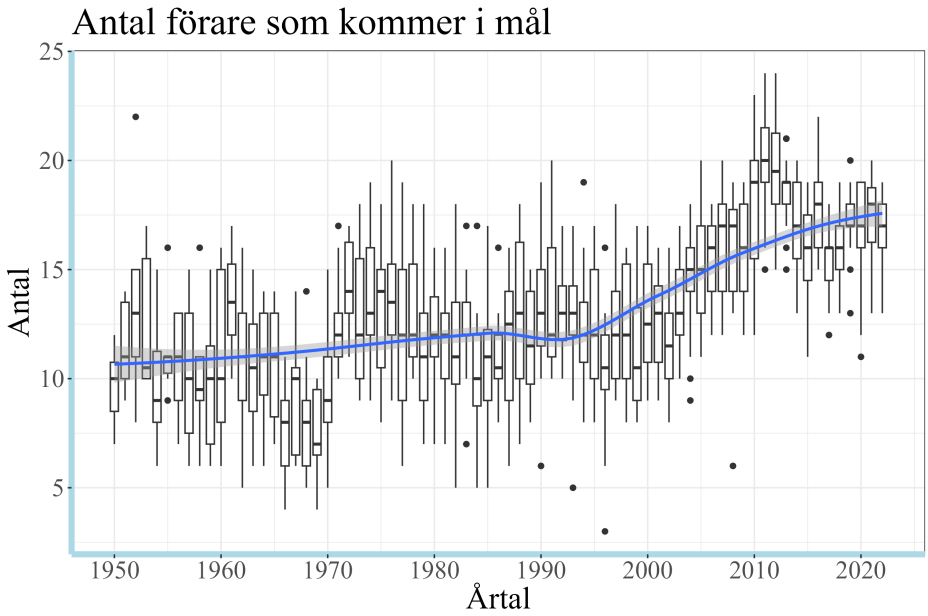
3.2.1 Justering för andel slagna förare

När vi analyserar datamaterialet inses snabbt att det råder stora skillnader mellan racen idag jämförelsevis med hur det varit tidigare år. Antalet förare som startar och andelen av dessa som kom i mål varierar såväl under, som mellan åren. I Figur 3.2 ser vi hur antalet startande i genomsnitt var över 30 st per race under 50-talets topp, jämförelsevis med 16 st under 60-talets botten. Tekniska utvecklingen har även gjort att bilarnas tillförlitlighet förbättrats som påverkar den markanta ökning av andelen startande förare som går i mål. Idag går cirka 85% av de startande förarna i mål till skillnad från när andelen var som lägst under 80-talet och då låg på 45%.



Figur 3.2: Antal startande med procentuell andel som tar målflagg (röd linje).

I Figur 3.3 ser vi, grupperat per år, hur många förare som i genomsnitt kommer i mål per race. Det innebär att N's storlek både varierar mellan varje enskilt race, men framförallt i genomsnitt mellan åren.



Figur 3.3: Antal (st) förare i mål per race grupperat per år.

Under 1950-talet kommer N i genomsnitt vara strax över tio, och istället cirka 17 under 2010-talet. Till följd av detta uppstår i huvudsak två problem:

(a) Responsvariabeln, nämligen relativa slutpositionen utvärderas i genomsnitt olika från år till år med anledning av transformationens grundade i N's storlek.

(b) Skillnaden mellan varje slutposition är större desto lägre N. Om vi då föreställer

oss att två förare inom samma team varje race placerar sig i samma ordning kommer modellen uppfatta att skillnaden stallkamraterna sinsemellan är större desto mindre N. Å andra sidan kommer ICC (3.3) sjunka i och med att variansen inom stallen (σ^2_{e0}) ökar. Konsekvensen av ovannämnda blir att modellen uppfattar den ökade variansen som större skillnad mellan stallkamraterna än vad den faktiskt är.

I och med de kraftiga variationerna i N's storlek mellan åren kan vi dra slutsatsen att förarnas påverkan och prestation kommer att värderas olika. Rankingen premieras av att antalet som kommer i mål är lågt då variansen på nivå-1 (förarna) ökar i takt med att N minskar. För att på så rättvist sätt som möjligt kunna jämföra alla förare samtidigt är därför en alternativ transformation av responsvariabeln nödvändig. Responsvariabel i denna undersökning grundas i en variant av transformationsformeln 3.6 i kombination med hur Bell m. fl. (2016) hanterade förare vilka placerade sig utanför poängplats. Under alla år har det i genomsnitt varit 13 st per race som kommit i mål. N sätts därför till konstanten 13 och innebär att förarna alltid värderas lika. Responsvariabeln för slutpositionerna 1-13 transformeras då enligt:

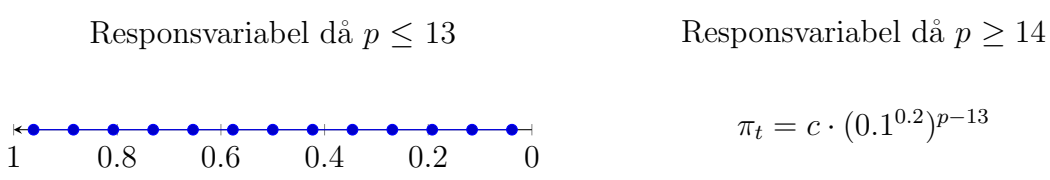
$$y_j = \frac{y_i(13-1)+0.5}{13} \tag{3.7}$$

Ett problem även påpekat av Kesteren och Bergkamp (2022) är att man vid ett ordinarie poängsystem går miste om väldigt mycket data då slutpositioner utanför poängplats utesluts. För att behålla all tillgänglig data kommer responsvariabeln för slutpositionerna $p \geq 14$ transformeras på liknande sätt som Bell m. fl. (2016). Det blir då en andel av transformerade responsvariabeln för slutpositionen 13 (0.038462) enligt ekvation 3.8:

$$c \cdot (0.1^{0.2})^{p-13} \tag{3.8}$$

där $c = 0.038462$ (π_t för slutposition 13) och $p =$ slutposition.

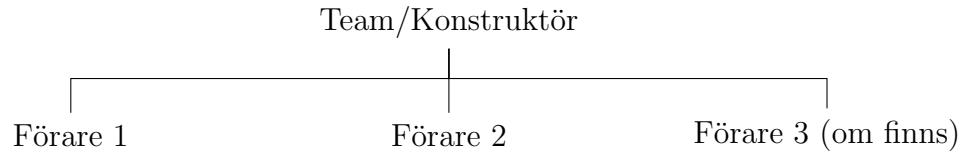
Responsvariabeln följer således ett alternativt poängssytem vilket säkerställer att slutpositionerna utvärderas lika över alla år och benämns π_t. π_t vilken modellen bygger på sammanfattas enligt Figur 3.4.



Figur 3.4: Transformerad responsvaribael π_t

3.3 2-nivås modell

I och med att vi i denna studie undersöker förarna, grupperade i stall/konstruktör bildar detta en 2-nivå modell av hierarkisk karaktär med möjligheten att utnyttja korsklassificering. Föraren nivå-1 och stallet nivå-2 stallet illustrerat enligt Figur 3.5.



Figur 3.5: Översikt modellen

Då den relativa slutpositionen π_t är en andel av slagna förare kommer modellen modelleras utifrån att responsvariabeln tillhör en Beta-fördelning. Detta ger en Beta-regression som tillåter de olika variablerna, förare och stall, att variera på respektive nivå. Att bestämma vilka variabler som ska inkluderas i modellen görs antingen genom statistiska metoder eller generella argument om huruvida variabeln har en påverkan på utfallet av responsvariabeln, eller inte. Istället för att lägga vikt vid att på bästa sätt prediktera slutpositionen av en förare i ett visst race, fokuserar denna studie och därmed modellformulering, istället på att ge så bra skattning av förarskickligheten som möjligt. Utgångspunkten ligger i att få så bra skattning av förarens prestation och skicklighet som möjligt, med så få variabler som möjligt för att undvika multikolinjäritet och överdriven modelkomplexitet. Detta leder till modellformulering enligt ekvation 3.9:

$$
\begin{aligned}
y_{fts} &\sim Beta(\mu_{fts}, \phi) \qquad (3.9)\\
logit(\mu_{fts}) &= \beta_N + \beta_f + \beta_{fs} + \beta_{ts}\\
\beta_f &\sim \mathcal{N}(0, \sigma_f^2)\\
\beta_N &\sim \mathcal{N}(0, \sigma_N^2)\\
\beta_{fs} &\sim \mathcal{N}(0, \sigma_{fs}^2)\\
\beta_{ts} &\sim \mathcal{N}(0, \sigma_{ts}^2)\\
\varepsilon_i &\sim \mathcal{N}(0, \sigma_i^2)
\end{aligned}
$$

Av modellformuleringen förstås att föraren har två intercept β_f och β_{fs} till skillnad från stallet som bara har β_{ts}. Modellen estimerar den övergripande förarkoefficienten β_f vilken kan ses som genomsnittlig skicklighet sett över förarens alla aktiva säsonger. Därefter fungerar koefficienten β_{fs} som en korrektion vilken justerar för årliga förändringar från den skattade förarkoefficientens β_f. Exempelvis om en förare med tiden blivit allt skickligare eller kanske istället varit i en årslång formsvacka. Stallet har istället bara en koefficient β_{ts} som varierar från år till år. Detta med anledning av att konstuktörsfördelen tenderar att kunna variera mycket kraftigare än förarens

skicklighet från år till år. Modellen skulle exempelvis överestimera Mercedes-bilens konstruktörsfördel år 2022, som var första gången på åtta år där stallet inte vann konstruktörsmästerkapet. Konsekvensen blir då att förarens skicklighet underestimeras. Slutligen har vi koefficienten β_N där N representerar antalet förare som kom i mål. Anledningen till införandet av parametern β_N är för att koefficienten skall fånga upp reponsvariabelns variation beroende på hur många som kommer i mål. Mer specifikt fångar koefficienten upp information om hur responsvariablernas varians beror på hur många förare som kommer i mål. Utifrån det som nämnts i avsnitt 3.1 & 3.2 är innebörden att de skattade koefficienterna, β_f summerat med β_{fs}, kan ses som ett mått på förarnas skicklighet på en log-odds kvotskala (Kesteren och Bergkamp, 2022). Rankingen denna studie beräknar representeras därmed av koefficienterna $\beta_f + \beta_{fs}$ för respektive förare och säsong.

3.4 Utökning av grundmodellen

Enligt tidigare undersökning har förarens betydelse ökat under våta förhållanden (Bell m. fl., 2016). På grund av detta kommer modellen utökas med hjälp av dummyvariabeln v presenterad i tabell 2.2 vilket leder till modellformuleringen av förarens skicklighet enligt ekvation 3.10.

$$\beta_f = \begin{cases} \gamma_{0f} + \gamma_{1f} \cdot v & \text{om } v = 1 \\ \gamma_{0f} & \text{om } v = 0 \end{cases} \tag{3.10}$$

$$\beta_t = \begin{cases} \gamma_{0t} + \gamma_{1t} \cdot b & \text{om } b = 1 \\ \gamma_{0t} & \text{om } b = 0 \end{cases} \tag{3.11}$$

På liknande sätt har även visats att bilarna passar olika bantyper olika väl där några av bilarna passar högfartsbanor, och andra passar lågfartsbanor bättre. Stadsbanor kännetecknas ofta av lägre genomsnittshastighet med snäva kurvor, och istället hög genomsnittshastighet med längre svepande kurvor på permanenta banor (James, 2017). I och med detta kommer den utökade modellen för stallet innebära en dummyvariabel med 0, eller 1 om stadsbana formulerad enligt ekvation 3.11. Syftet är att undersöka huruvida skattningen av parametrarna förbättras med hjälp av den utökade information. En summering av modellernas olika variabler beskrivs i tabell 3.1.
Hur de olika modellerna utvärderas diskuteras i avsnitt 3.7.

Variabler	Beskrivning
π_t - Responsvariabel	Transformerad relativ slutposition
f - Förare	Namn på förare
fs - Förare-År	Förare f givet ett specifikt år s
ts - Team-år	Stall/konstruktör/team t givet ett specifikt år s
s - År	Året som racet ägde rum
N - Antal	Förare som tog målflagg
v - Väder	Dummyvariabel med 0 om torrt, 1 om regn
b - Bantyp	Dummyvariabel med 0 om en permanent bana, 1 om stadsbana

Tabell 3.1: Variabler i modellen.

3.5 Bayesiansk estimering

Den Bayesianska skattningsmetoden har funnits under en mycket lång tid men har ofta, tills på senare tid, blivit negligerad i förmån till den frekventistiska metoden (Hackenberger, 2019). Den huvudsakliga skillnaden mellan metoderna blir hantering av skattningens osäkerhet. I frekventistiska ansatsen antas den skattade parametern/variabeln ha ett fast men okänt värde. Man hittar därför ett värde på en parameter vilket gör den insamlade datan mest sannolik givet att den följer en viss fördelning. Till skillnad från den frekventistiska använder istället den Bayesianska ansatsen information om fördelningen innan skattningen av parametern görs. Här etableras först *apriori- och aposteriori- förrdelningarna* där förstnämnda är vår initiala uppfattning om slumpvariabelns (*s.v.*) fördelningen. *Aposteriori* är istället beviset, med andra ord det empiriska resultatet, av slumpvariabelns fördelning efter inhämtande av data (Spiegelhalter, 2020).

Generella formeln för 'Bayes-Theorem' formuleras enligt ekvation 3.12.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.12}$$

Av ovan formulering inses snabbt att den Bayesianska ansaten grundar sig i en betingad sannolikhet. Att händelse A sker givet att händelse B har skett. Vid en frekventistisk ansats är detta information vilket inte måste innehavas för att göra en parameterskattning. Ett alternativt sätt att se på den Bayesianska ansatsen är hur vår uppfattning om att händelse A kommer ske, *förändras* med kunskapen om att händelse B har skett. Skattningen kan därmed ses som en form av medelvärde av initiala uppfattningen i kombination med det insamlade beviset av *s.v.*'s fördelning vilket Fan (2016) i det kontinuerliga fallet formulerar om till:

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) \times f_{\Theta}(\theta) \tag{3.13}$$

Aposteriori sannolikhetsfördelning $\propto$ Likelihood $\times$ Apriori sannolikhetsfördelning

Likelihoodfunktionen i ekvation 3.13, även kallad 'Bayes-factor', beskriver hur sannolikt det är att observera den inhämtande datan givet vissa modellparametrar. Här

förstår vi hur *aposteriori*-fördelningen blir ett viktat medelvärde av *apriori*-fördelningen multiplicerat med sannolikheten för observationen. När uttrycket sedan integreras över alla parametervärden ges *aposteriori*-fördelningen utifrån vilka modellparametrarna som kommer att skattas (Lambert, 2018).

3.6 Monte-Carlo och Markov Kedja

Många gånger undersöks ett samband med många parametrar där den anlalytiska lösningen fort blir mycket komplext, ibland rentav omöjlig att lösa. En allmänt vedertagen lösning för dessa problem har med tiden blivit att med en numerisk ansats approximera en lösning på det matematiska problemet genom en s.k. *Monte-Carlo simulering* (Sortino m. fl., 2010). I huvudsak är en Monte-Carlo simulering en metod som baserat på slumpgenererade tal som approximerar en lösning på ett komplext och ibland omöjligt matematisk problem. Estimeringen av parametern grundar sig därför på genomsnittet av iterationer. Det finns många typer av *Monte-carlo* metoder men den som kommer att användas är *Monte-carlo integration*, som är en teknik för numerisk integration med slumptal. Det är en speciell Monte Carlo-metod som numeriskt beräknar en bestämd integral. (Harrison, 2010).

Slumptalsgenerering bygger på flera Markovkedjor. En Markovkedja är en slags stokastisk process där det enda som påverkar framtiden är det nuvarande tillståndet (Tim, 2018). Egenskaperna vilka behöver uppfyllas för att anses vara en Markovkedja är att historien och framtiden skall vara oberoende, givet att det nuvarande tillståndet är känt (Schön, Wallin och Wikström, 2017). Det här innebär med andra ord att man antar att den stokastiska processen är stationär. Alltså förändras inte sannolikheterna i Markovkedjan, att gå från ett tillstånd till ett annat, över tid. För att förtydliga detta kan vi uttrycka det i matematisk form enligt: vi låter $X_n = X_{n n=0}^{\infty} = X_0, X_1, X_2, ...$ vara en sekvens av slumpvariabler *s.v.* och $\mathcal{S}$ vara tillståndsrummet som slumpvariablerna kan anta värden i. En Markovkedja bildar då en sekvens av slumpvariabler X_n som antar värden i tillståndsrummet $\mathcal{S}$ (Chalmers, 2015). Om Markov-egenskaperna, enligt tidigare förklarat, uppfylls kan man sammanfatta det med formeln (3.14):

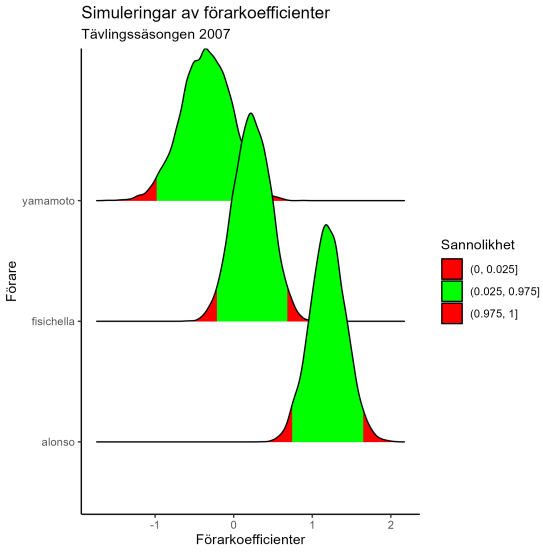
$$P(X_n = s_n | X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, ..., X_0 = s_0) = P(X_n = s_n | X_{n-1} = s_{n-1}) \tag{3.14}$$

Markov-kedjan kan med beskrivningen enligt ovan ses som en stokastisk process via Monte-Carlo-simulering.

Resultatet kommer vara varje given förares transformerade slutposition, baserat på *aprioriförde*lningen såväl som *aposterioriförde*lningen. Simuleringen genomförs med hjälp av R-paketet **brms** (Bürkner, 2017). Brms (Bayesian Regression Models using Stan) är ett paket för R som gör det möjligt att använda Bayesianska regresionsmodeller. Paketet gör det enkelt att utföra komplexa multivariata regressioner med korrelerade observationer från olika fördelningar. För att inte gå alltför djupt in på ämnet måste det ändå noteras att Markov-processen kan användas med olika samplingsalgoritmer. Några av de mer väletablerade är Gibbs, Metropolis-Hastings eller som BRMS-paketet använder, *No-U-Turn samplern* vilket är en utökning av *Hamiltonian-sampling* (Bürkner, 2017).

No-U-Turn sampler är en teknik för att effektivt generera stöd från en målfördelning inom Monte Carlo-simulering. Det är särskilt användbart när modellen som studeras är högdimensionellt eller har komplexa samband mellan observationer, eftersom det automatiskt anpassar sig till dessa egenskaper. Till skillnad från *Hamiltonian-sampling* behöver man inte bestämmas några parametrar själv utan det gör *No-U-Turn sampler* automatiskt (Hoffman, Gelman m. fl., 2014).

Kredibilitetsintervall



Figur 3.6: Exempel från genomförd simulering av förarkoefficienten.

I den Bayesianska ansatsen använder vi kredibilitetsintervall vilket inte skall förväxlas med ett konfidensintervall. Kredibilitetsintervall är ett intervall där integralens gränser (intervallet) bestäms av att funktionsvärdet av *aposteriori-fördelningens* integral skall uppgå till önskad sannolikhet (ofta 95%). Att finna dessa gränser görs enklast genom att hitta kvantilerna för önskat kredibilitetsintervall (Goldstein, 2011b). Med den inbyggda funktionen *quantile* i programmet R kan vi beräkna gränserna med en given sannolikhet (Hyndman och Fan, 1996). Vi har valt ett kredibilitetsintervall om [2.5% 97.5%] där medelvärdet av simuleringarna vilka faller inom intervallet kommer utgöra vår skattning av respektive förares skicklighet. Detta illustreras av det grönmarkerade området i Figur 3.6.

3.7 Modellvalidering

När vi bygger en regressionsmodell i syfte att antingen prediktera eller undersöka samband mellan variabler är det av största vikt att validera modellen. *Korsvalidering* är en statistisk metod för att undersöka hur väl modellen passar datan. Här finns olika tekniken grundat i korsvalidering exempelvis *K-fold* och *Leave-one-out (LOO)*-metoden. Den senare innebär att en datapunkt (eller kluster) hålles ute från modellen för att sedan prediktera dess värde baserat på resterande datapunkter (Berrar, 2018). Den numeriska beräkningen ger då ett mått på hur väl modellen predikterar den datapunkt som hålles ute, summerat över alla datapunkter i. Denna studie använder LOO-metoden som på det generella sättet har den matematiska formuleringen:

$$\frac{1}{n}\sum_{i} L(y_i - \hat{f}_{\lambda}^{-i}(x_i)) \tag{3.15}$$

där $\hat{f}_i^{-i}(x_i)$ är passningen av x_i då denna hållits ute (Goldstein, 2011e).

Med utgångspunkt i LOO-metoden finns olika mått där man med R-paketet **brms** beräknar expected log pointwise predictive density (Bürkner, 2022).

$$elpd_{loo} = \sum_{i=1}^{n} \log\, p(y_i|y_{-i}) \tag{3.16}$$

$$\text{där } p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i}) \tag{3.17}$$

Expected log pointwise predictive density, ELPD, är enligt ekvation 3.16 den *Leave-one-out* predikterade fördelningen, givet datamaterialet utan datapunkten i (Vehtari, Gelman och Gabry, 2016). Då vi jämför flera modeller är det den med högst ELPD som passar datan bäst och den vi väljer att gå vidare med. I kontexten av denna studie jämför vi då huruvida dummyvariablerna om väder och bantyp tillför information till modellen så att dess prediktionsförmåga förbättras.

I en Markovkedja är $\hat{R}$ (*R-hat*) en skattning på hur tillförlitligt estimatet i en simulering är. $\hat{R}$ är ett mått på hur nära den simulerade fördelningen, är den verkliga stationära fördelningen. Den samplade fördelningen ska konvergera mot den faktiska fördelningen. Då $\hat{R}$ är nära 1 betyder det att den simulerade stationära fördelningen är nära den faktiska stationära fördelningen. Om $\hat{R}$ istället är mycket ovan 1 betyder det istället att den simulerade stationära fördelningen inte konvergerar till den faktiska stationära fördelningen. $\hat{R}$ kan anta värden mellan $1 \leq \hat{R} < \infty$, att avgöra om $\hat{R}$ -värdet är för högt beror på hur modellen ser ut men generellt sätt ska värden över 1.1 inte accepteras, ifall en rimlig anledning varför inte ges. Varje variabel i modellen får ett $\hat{R}$ -värde och ifall samtliga värden är nära 1 betyder det att modellen är tillförlitlig. (Brooks m. fl., 2011).

I många fall kan det vara missvisande att uteslutande titta på den numeriska summeringen av regressionsmodellen. Som komplement till den numeriska summeringen kommer vi även göra en grafisk valideringen vilket av vissa anses vara en av statistikernas viktigaste redskap inom tillämpad statistik (Gabry m. fl., 2018). För en Bayesiansk modell finns ett flertal olika grafiska valideringsmetoder bland annat 'posterior predictive checks', 'prior predictive checks' eller 'mixed checks' där först- och sistnämnda bäst lämpar sig för hierarkiska modeller (Gelman, Hwang och Vehtari, 2013). Genom att studera plottarna ser vi hur väl modellens predikterade värden överenstämmer med de verkliga, vilka bör passa någorlunda väl. Trace-plots är ytterligare ett sätt för att validera modellen. Man vill att Markov-kedjan undersökt alla möjliga värden. I plottarna skall det därför inte kunna anas ett mönster i iterationerna men även att de rört sig över alla möjliga värden.

4 Resultat

I detta avsnitt presenteras resultatet från modellen och dess simuleringar i programmet *R*. Simuleringarna baseras på de 14 323 st observationer av förare som kommit i mål. I simuleringenarna har 4 markov-kedjor med 3500 iterationer använts för att undersöka hela utfallsrummet Ω.

4.1 Modellval

Inledningsvis jämförs de olika modellerna genom *LOO*-korsvalidering för att undersöka vilken som presterar bäst. Resultatet av korsvalideringen för de fyra modellerna hittas i tabell 4.1.

Modell	ELPD LOO	ΔELPD	ΔSE	SE ELPD LOO
Väder + Bantyp	6401	0	0	98.473
Bantyp	6395	−5.209	4.092	98.300
Väder	6366	−34.282	7.924	98.037
Basic	6364	−36.346	8.886	97.868

Tabell 4.1: LOO resultat race.

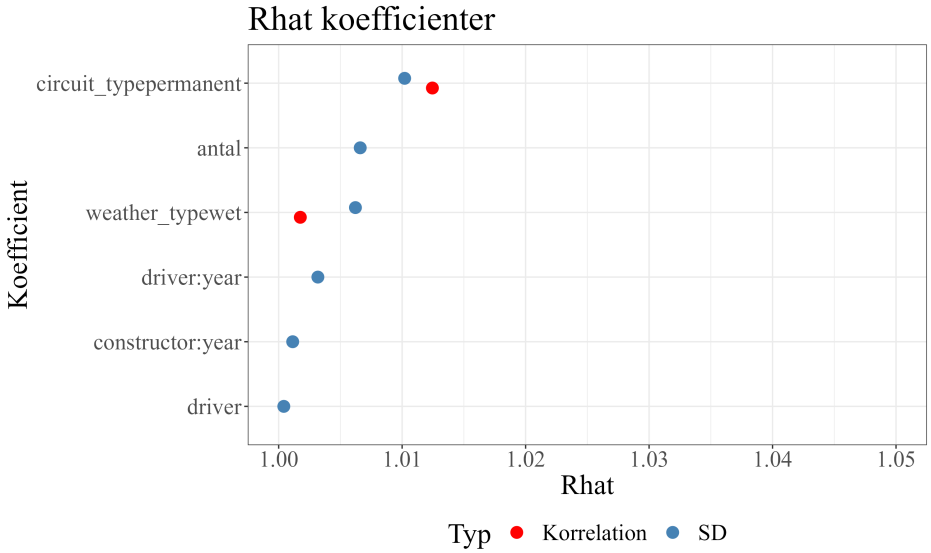
Vi kan se att modellen med högst ELPD inkluderar variablerna *Väder* och *Bantyp*. Informationen modellen får om vilken bantyp och väderförhållanden förbättrar alltså skattningen och dess prediktionsförmåga. Hädanefter kommer därför modelldiagnostiken i tabell 4.2 och den ranking som presenteras enbart grunda sig i den utökade modellen med information om väder och bantyp.

Parameter	Rhat	mean	sd	2.5%	97.5%
β_N	1.00	0.85	0.14	0.61	1.17
β_{ts}	1.00	0.52	0.03	0.47	0.58
β_f	1.00	0.49	0.03	0.44	0.54
β_{fs}	1.01	0.25	0.02	0.21	0.28
$\gamma_{ot} + \gamma_{1t} \cdot b$	1.01	0.19	0.03	0.13	0.26
$\gamma_{of} + \gamma_{1f} \cdot v$	1.00	0.15	0.04	0.05	0.22
$\gamma_{ot} + \gamma_{1t} \cdot b$	1.02	0.58	0.18	0.24	0.93
$\gamma_{of} + \gamma_{1f} \cdot v$	1.00	-0.06	0.17	-0.38	0.26
ϕ	1.00	5.35	0.07	5.21	5.49

Random effects	
σ^2	0.04
τ_{00}	0.04
ICC	0.55
N_{antal}	22
N_{driver}	559
N_{year}	73
$N_{constructor}$	138
Observationer	14323
R^2	0.631

Tabell 4.2: Summeringsstatistik bästa modell.

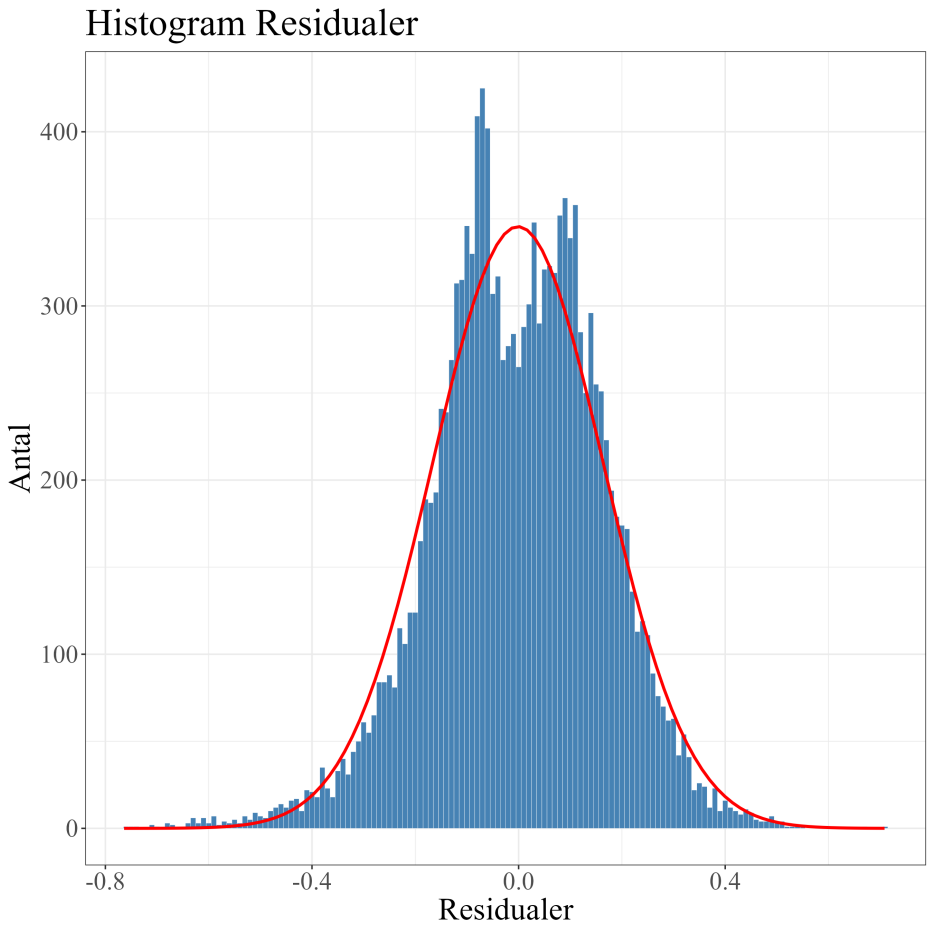
Summeringsstatistiken visar att $\hat{R}$ ligger inom acceptansnivån (≤ 1.1) för alla variabler i modellen. Även om det inte är en garanti ger det oss en mycket god indikation på att simuleringarna har konvergerat. Resultaten kan därmed anses pålitliga med en acceptabel osäkerhet. I Figur 4.1 illustreras $\hat{R}$ för respektive koefficient i simulering-



Figur 4.1: $\hat{R}$ för respektive variabler.

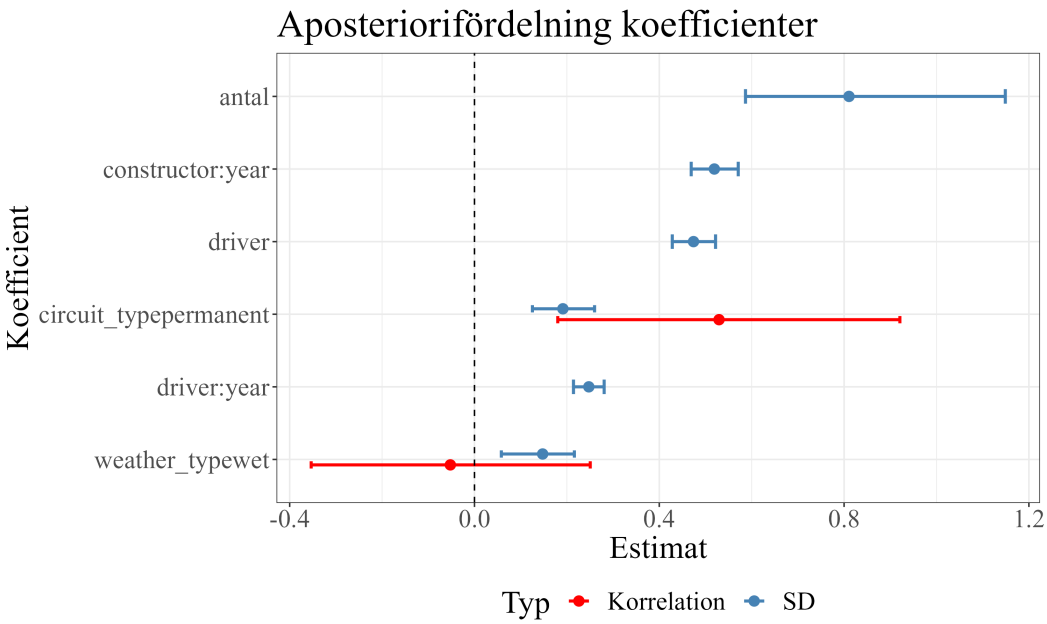
arna. Figur A.3 i Bilaga A illustrerar istället ett $\hat{R}$ för varje unik $\beta_f, \beta_{fs}, \beta_{ts}, \beta_N$. Av båda visualiseringar kan uttydas att alla har ett $\hat{R} \leq 1.05$ och de flesta ligger istället mycket nära 1. Kedjorna har därför, av visualiseringen att döma, konvergerat väl.

Vid modellering genom regression krävs även för att kunna validera modellen, diskuterat i avsnitt 3.2, att residualerna är oberoende och normalfördelade med väntevärdet 0. Dessutom att variansen är konstant över hela utfallsrummet, alltså $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.



Figur 4.2: Histogram av residualer.

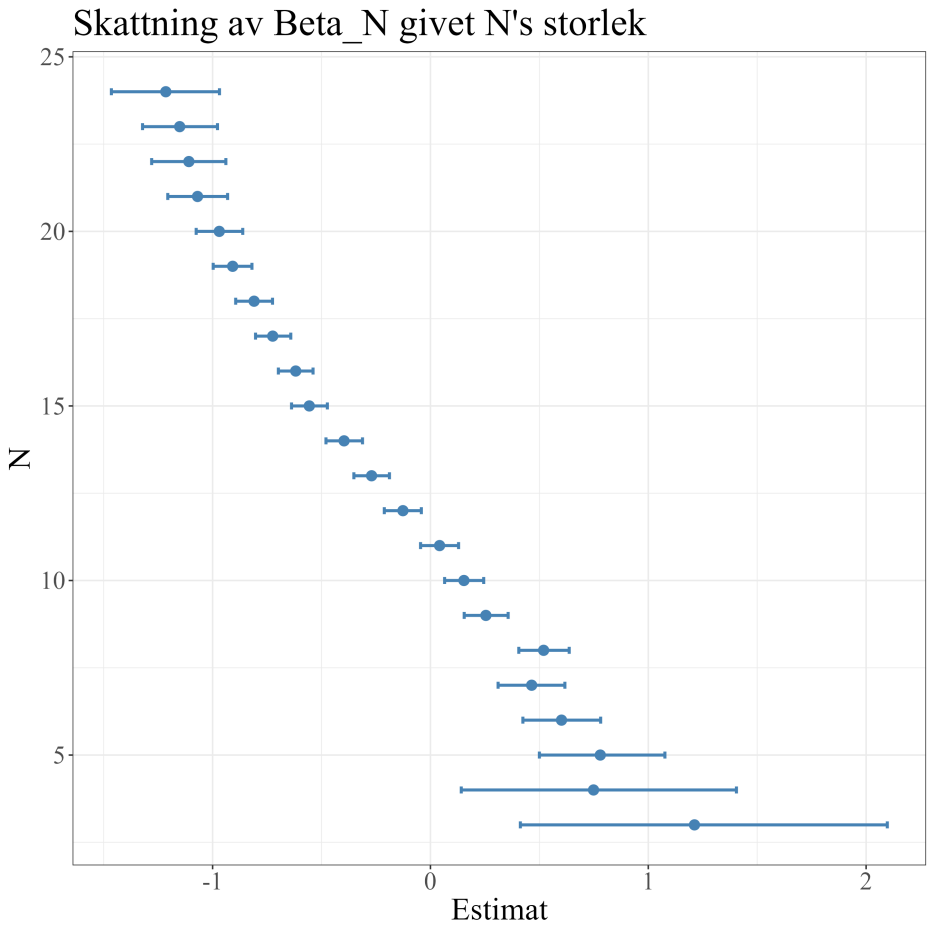
I Figur 4.2 visualiseras modellens residualer mot den teoretiska normalfördelningen. Även om mindre avvikelser gentemot den teoretiska fördelningen finns, så är $E(\varepsilon_i|x_i) = 0$, och antagandet om ett linjärt samband anses vara uppfyllt. Däremot kan vi i Figur A.2 i Bilaga A se att de största avvikelserna tenderar att uppstå vid låga värden av responsvariabel y_i. Vi hittar även posterior-predicitive check plottar i Bilaga A Figurer A.4 - A.7, som visualiserar modellens predikterade värde av responsvariabeln, gentemot det observerade värdet. Här kan vi se att modellens prediktionsförmåga är beronde av antalet observationer, per förare, för att skattningen skall vara god. Många förare under 60-talet startade enbart ett, eller ett fåtal tävlingar, och modellen får därför svårigheter i att prediktera responsvariabeln, därmed också skatta koefficienterna. Att poängtera är däremot att modellens skattningarna av de högpresterande förarna, under alla år, till synes är bra.



Figur 4.3: Aposteriori-fördelning av modellparametrar. Med 95% kredibilitetsintervall.

Figur 4.3 visar istället den *aposteriori-fördelningen* för modellparametrarna. Punkterna representerar det skattade medelvärdet, och linjerna ett 95% kredibilitetsintervall. Ganska snabbt ser vi att variabeln antal, har en störst påverkan på modellens skattningar. Detta på grund utav det, i jämförelsevis, höga värde kombinerat med större osäkerhet i form av bredare kredibilitetsintervall, i skattningen av koefficenten β_N. Enligt Figur A.1 i Bilaga A finner vi trace-plottar av simuleringarna av markov-kedjorna. Kedjorna i simuleringen ser tämligen slumpartade ut då ett mönster eller trend inte går att uttydas.

I avsnitt 3.2.1 diskuterades om de svårigheter som uppstår vid evaluering av förare under olika tidsperioder. I syfte att försöka justera för detta infördes därför koefficienten β_N. Vid ytterligare undersökning, av skattningen av koefficienten, kan vi i Figur 4.4 se ett mycket tydligt mönster. Skattningen av β_N minskar i takt med att N ökar. Här illustreras även ett 95% kredibilitetsintervall, där vi således på 5%-nivån kan statistiskt säkerställa att antalet som går i mål har en påverkan för modellen, och dess skattning av de olika koefficienterna. Syftet med införandet av koefficienten är därmed uppfyllt.

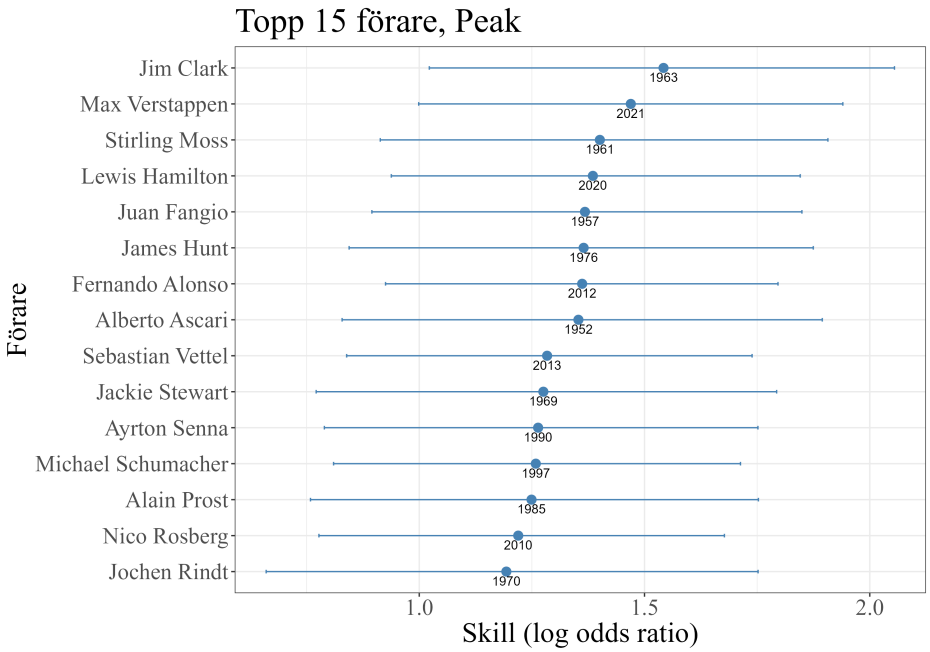


Figur 4.4: β_N för olika N. 95% kredibilitetsintervall.

4.2 Förarskicklighet

Förarskickligheten utvärderas här i de två olika måtten karriärstopp och bästa tre-årsperiod. Rankningarna presenteras då av koefficienterna $\beta_f + \beta_{fs}$.

4.2.1 Kariärstopp



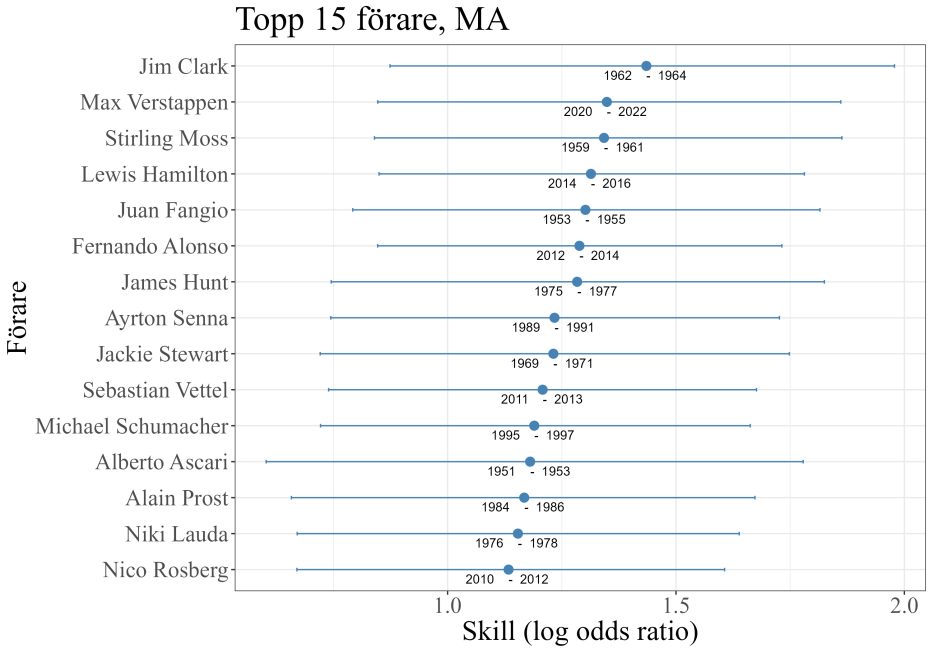
Figur 4.5: Topp-15 förarskicklighet, peak per förare. 95% kredibilitetsintervall.

Förare	VM-titel | år
Clark	Ja
M. Verstappen	Ja
Moss	Nej
Hamilton	Ja
Fangio	Ja
Hunt	Ja
Alonso	Nej
Ascari	Ja
Vettel	Ja
Stewart	Ja
Senna	Ja
M. Schumacher	Nej
Prost	Ja
N.Rosberg	Nej
Rindt	Ja

Tabell 4.3: Topp-15, tagen VM-titel året då föraren körde som bäst.

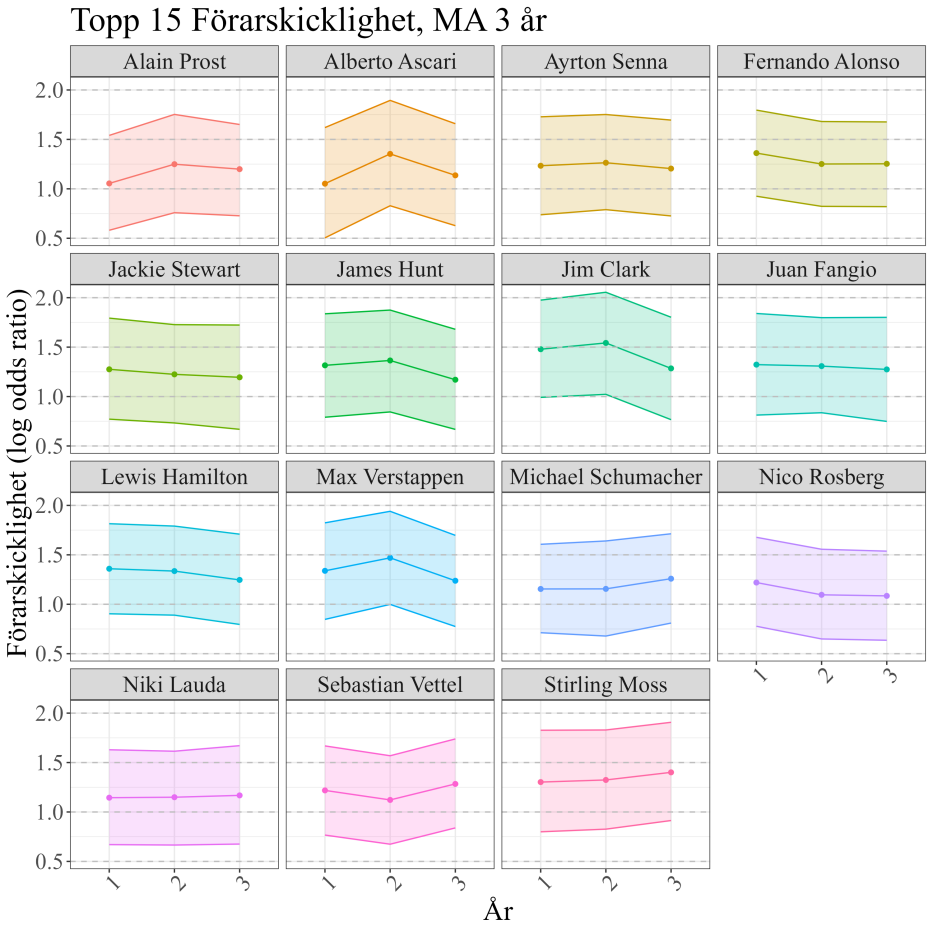
Figur 4.5 visar de 15 bästa estimaten per varje förare. Clark har exempelvis mer än ett år där hans ranking skattades tillräckligt bra för att få flertal platser på listan. I Figur 4.5 ser vi punktskattningen av estimaten, vilket årtal, men även kredibilitetsintervallet av skattningen. Tabell 4.3, med VM-titel | år, visar istället ifall en förare vann mästerskapet året de körde, enligt modellen, som allra bäst. Resultatet har lösgjort förarens skicklighet från bilens kapacitet där den högsta förarskickligheten inte är en garanti för att man ska ha vunnit världsmästerskapet. Då vi har överlappande kredibilitetsintervall förarna sinsemellan kan vi inte statistiskt särskilja förarna, även om de ger en god indikation.

4.2.2 Förarskicklighet Moving average



Figur 4.6: Topp-15 bästa 3-års intervall. 95% kredibilitetsintervall.

Figur 4.6 och 4.7 visar topp 15 förare med högst förarskicklighet under tre år samt hur de presterade under respektive år. Återigen är Clark i topp och vår modell skattar att han är bäst vilket ger svar på den andra frågeställningen som presenterades i avsnitt 1.2. Michael Schumacher och Lewis Hamilton är de två förare med flest antal världsmästartitlar (7) men enligt modellen endast elfte respektive fjärde bäst. Trots en lägre skattad förarskicklighet har de vunnit mest, detta då föraren och bilen tillsammans varit de bästa under de åren. Den enda föraren som inte vunnit ett världsmästerskap men av modellen rangordnas inom topp 15, är Moss. Notera däremot att alla förarna i topp 15 ligger inom varandras kredibilitetsintervall.

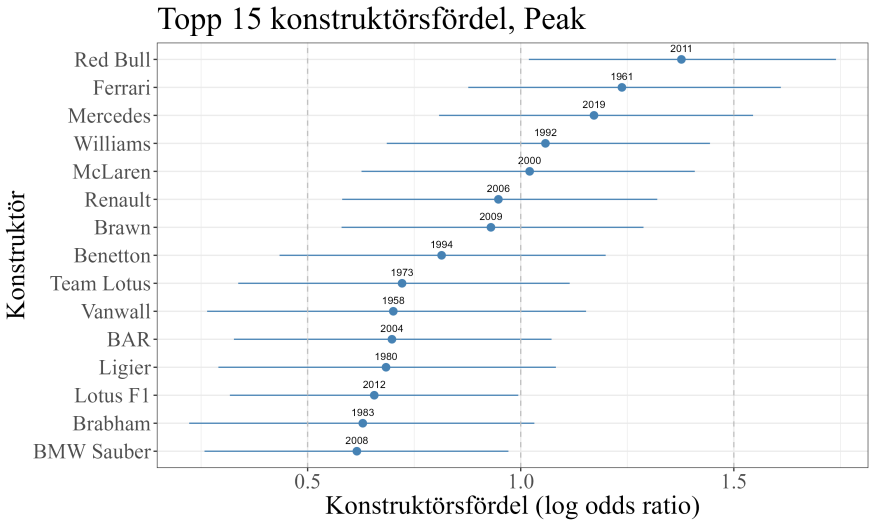


Figur 4.7: Topp-15 med högst förarskicklighet under respektive förares bästa 3-års period. 95% kredibilitetsintervall.

4.3 Konstruktörsfördel

Under uppsatsens gång har det diskuterats om hur bilen har en direkt påverkan på förarens resultat. I modellen representerar därför koefficienten β_{ts} den årliga konstruktörsfördelen, alltså hur bra bilen var i jämförelse med de andra tävlande. Notera här att det inte är ett mått på den snabbaste bilen utan istället konstruktörsfördelen i *relation* till de andra stallen under samma säsong.

4.3.1 Karriärstopp



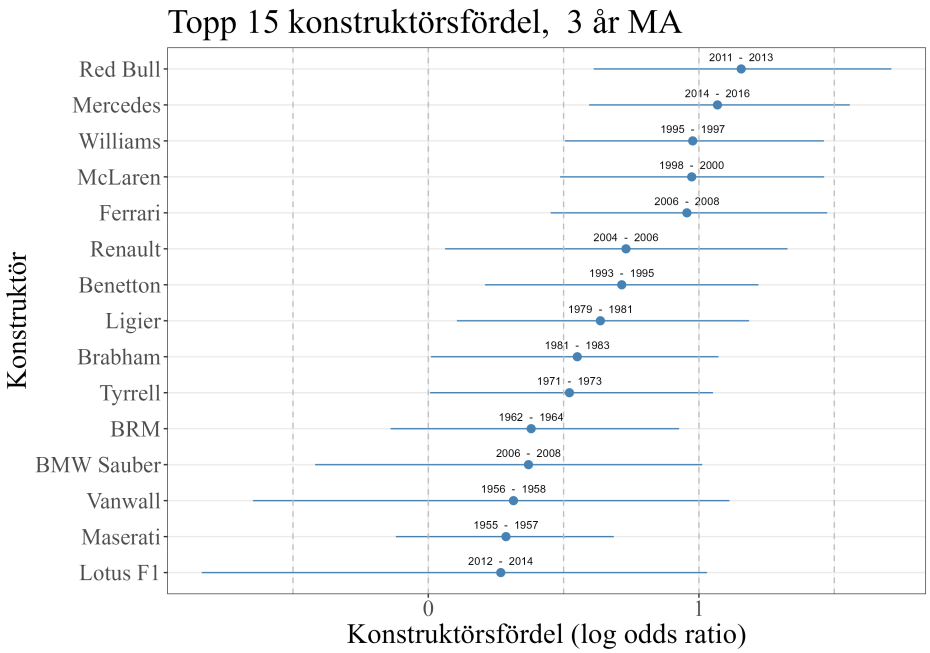
Figur 4.8: Topp-15 konstruktörtörsfördel, peak per stall. 95% kredibilitetsintervall.

Konstruktör	VM-titel | år
Red bull	Ja
Ferrari	Ja
Mercedes	Ja
Williams	Ja
McLaren	Nej
Renault	Ja
Brawn	Ja
Bennetton	Nej
Team Lotus	Ja
Vanwall	Ja
BAR	Nej
Ligier	Nej
Lotus F1	Nej
Brabham	Nej
BMW Sauber	Nej

Tabell 4.4: Topp-15, tagen VM-titel året då stallet var som bäst.

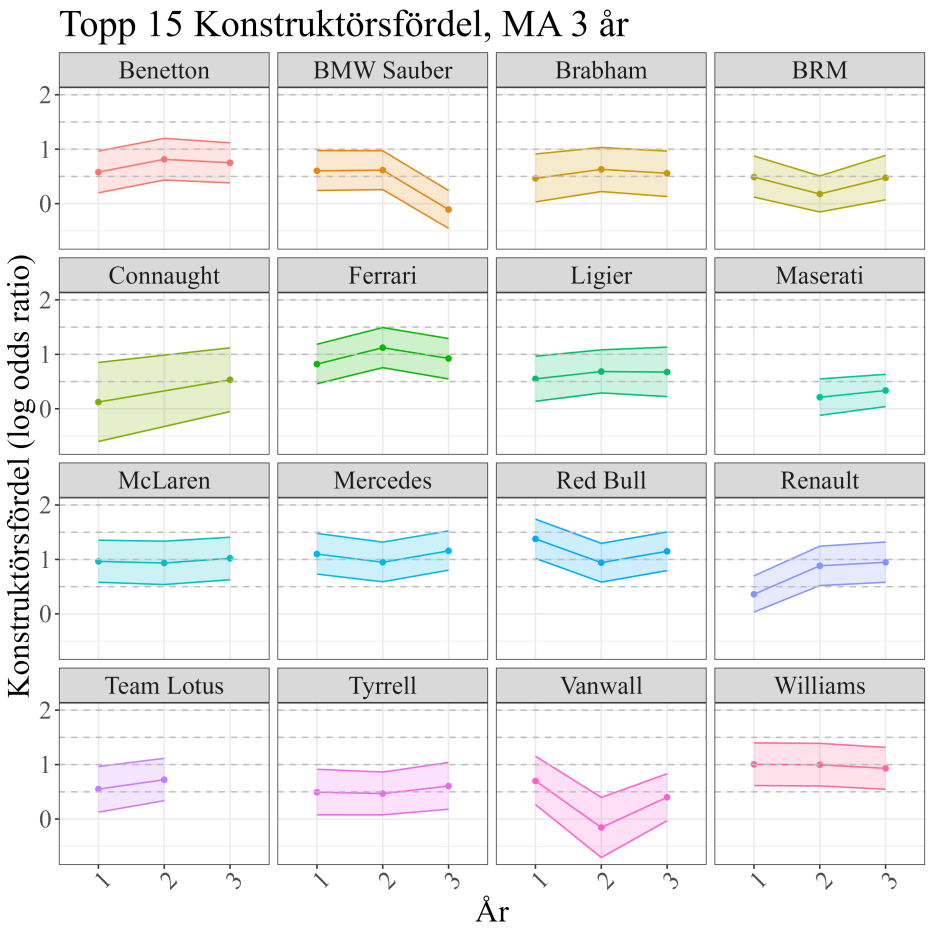
Figur 4.8 visar de 15 bästa konstruktörerna och deras bästa år då konstruktörsfördelen var som högst. I tabell 4.4 ser vi också att McLaren är det stall som haft högst konstruktörsfördel även om de inte vann konstruktörsmästerskapet just det året. Olika anledningar till detta kan finnas, men något av det mer troliga är att även Ferrari under just det året också hade en hög konstruktörsfördel. Även här finner vi att säkerheten i skattningarna inte är så pass hög att det kan statistiskt säkerställas vilket stall som haft det bästa året någonsin. Däremot kan vi med statistiskt säkerhet uttala oss om att RedBull haft en bil som varit mer konkurrenskraftig än Lotus F1 någonsin har haft.

4.3.2 Konstruktörsfördel Moving average



Figur 4.9: Topp-15 bästa 3-års intervall. 95% kredibilitetsintervall.

Av samtliga stall byggs bilen från grunden med stor frihet så länge det följer ett visst tekniskt reglemente. Det tekniska reglementet ändras med jämna mellanrum vilket också gör att konstruktörsfördelen mellan två år kan variera stort. Åren 2014-2021 vann Mercedes samtliga konstruktörsmästerskap där modellen enligt Figur 4.9 visar att den bästa 3-års perioden var 2014-2016. Red bull hade däremot återigen en 3-års period (2011-2013) vars genomsnittliga estimat slår Mercedes. Liksom Mercedes, vann Red-Bull konstruktörsmästerskapet samtliga år under denna 3-års period.



Figur 4.10: Topp-15 med högst konstruktörsfördel under respektive konstruktörs bästa 3-års period. 95% kredibilitetsintervall.

5 Diskussion

Resultatet från modellen som denna studie byggts på visar att Jim Clark är den historisk bästa Formel-1 föraren. Av de 72 lopp han startade i, kom han i mål 44 gånger och av dessa vann han 25. Utöver det att han fortfarande håller några av de mest prestigefyllda rekorden i sporten blev han även tvåfaldig världsmästare åren 1963 och 1965. Det råder ingen tvekan om att han är en av den genom tiderna bästa Formel-1 föraren. Vilket också återspeglas i hur ofta hans namn kommer på tal i denna diskussion. Även om modellen ger en bra indikation på hur bra de olika förarna är, bör resultaten ändå tolkas med en viss försiktighet. Av de 15 bäst placerade förarna kan vi inte statistik säkerställa att deras förarskicklighet är särskilda. Detta på grund av överlappande kredibilitetsintervall i skattningarna av förarkoefficenterna. Däremot kan vi på signifikansnivån $\alpha = 5\%$ statistiskt säkerställa att Jim Clark haft en högsta förarskicklighet (career-peak) som är skild från den 80:de bästa föraren, enligt modellen. Med andra ord kan modellen statistiskt säkerställa att Jim Clark varit bättre än $\approx 86\%$ (80/559) som under historien tävlat i Formel-1. Vidare är resultateten inte felfria där både felkällor samt modell måste diskuteras, identifieras och förbättras.

5.1 Tidigare studier

M-Verstappen och Hamilton blev placerade betydligt högre än andra förare när Kesteren och Bergkamp (2022) gjorde sin analys för den s.k "Hybrid eran" (2014-2021). Vårt resultat som presenterades i det tidigare avsnittet (4) visar att båda dessa förare även är med bland de 15 bästa genom tiderna. Liksom Kesteren och Bergkamp's (2022) studie placerade sig M.Verstappen före Hamilton även i vår analys, men med desto större skillnad. En anledning skulle kunna ligga i att han under de två senaste åren varit stallkamrat med en sedan tidigare duktig förare. På grund av korsklassificering har modellen då mer information om hur M.Verstappen skulle stå sig gentemot andra förare, som då både kan dra upp och ned estimaten. Utöver att resultaten av estimaten skiljer sig åt, gör även resultaten från korsvalidering det.

Kesteren och Bergkamp's (2022) modell som presterade bäst var den enklaste utan information om bantyp och väder, vilket som för denna studie var information som förbättrade skattningarna väsentligt. Detta kan bero på att urvalet för modellen skiljer sig åt och att det historiskt sätt varit större variationer i slutpositioner beroende på väder och bantyp. Kesteren och Bergkamp (2022) påpekade också att det skulle vara svårt att jämföra över hela historien och istället lättare under en viss tidsperiod. Detta med anledning av att bilarna, reglementet och sporten har utvecklats och förändrats genom åren. Kesteren och Bergkamp (2022) undersökte därför en begränsad tidsperiod då förändringarna i det tekniska reglementet och tävlingsformatet var små. Kesteren och Bergkamp (2022) kunde därför ha två koefficienter för både förare och stall, alltså för den genomsnittliga förarskickligheten, och för konstruktörsfördelen med

årlig justering.

I den linjära regressionsmodellen Eichenberger och Stadelmann (2009) genomförde kom man fram till att Fangio, Clark följt av Schumacher är de bästa Formel-1 förarna genom tiderna. Problemet vi ser med att använda en 1-nivås modell är att den verkar hantera komplexa relationer och asymmetriska observationerna sämre än beta regression. Fördelen vid beta-regression blir dessutom att resultaten är lättare att tolka. Bell m. fl. (2016) som likt denna studie genomförde analysen med en multilevel-modell, sett över alla år, kom även de fram till att Fangio var den bästa. Här var Prost näst bästa följt av Clark på tredjeplats.

När vi jämför denna och de tidigare studierna är resultaten snarlika, även om de skiljer sig lite åt. Självklart kommer det att finnas skillnader, framförallt i att denna studie rangordnar Hamilton högre jämförelsevis med tidigare studier. Vi har även M.Verstappen som i tidigare undersökningar inte varit inkluderad, då han ännu inte hade börjat tävla i Formel-1. Studierna skiljer sig även åt i rankingen av främst Michael Schumacher, Ayrton Senna samt Alain Prost som av många experter anses vara bland de absolut bästa genom tiderna, men av resultatet från modellen i denna studie rangordnas i nedre hälften av topp-15. En möjlig orsak till detta är att Michael Schumacher pensionerade sig från Formel-1 för att några år senare återuppta karriären. Efter sin återkomst presterade Schumacher, enligt många experter, inte på samma höga nivå. Vid utnyttjandet av korsklassificering skulle då detta kunna vara en bidragande faktor till Nico Rosberg's höga ranking (som var stallkamrat med Michael Schumacher när han kommit tillbaka) då modellen tror att Michael Schumacher var bättre än vad han kanske faktiskt var de åren efter att han kommit tillbaka. En möjlig orsak till Ayrton Senna och Alain Prost jämförelsevis låga ranking kan vara det faktum att de under en viss period tävlade för samma konstruktör. Även om de presterade på en mycket hög nivå kan det med anledning av beräkningen av ICC gjort att modellen istället uppfattar denna höga prestationsnivå som en hög konstruktörsfördel, så tillvida att individerna inom grupperingen presterar jämt. Tidigare studier har även de breda kredibilitetsintervall, även om en punktskattning om förarens skicklighet kunde ges var det inte statistiskt säkerställt vem som är den bästa genom tiderna. En sak har dock samtliga studier gemensamt, det komplexa problemet att lösgöra föraren från bilen. Här har man försökt med olika metoder och där den som verkar fungera bäst, är olika multilevel-modeller.

5.2 Modellförbättring

Formel-1 handlar inte bara om bilen och föraren, bakom framgångarna ligger även ett helt strategi-team (Motorsport, 2022). Under varje race (enligt dagens regler) måste varje bil byta däck minst en gång. Detta leder till olika strategier för varje stall och förare, vilket inte tas hänsyn till i modellen. Felaktiga strategiska beslut kan därför komma att påverka förarens slutposition och modellens beräkning av förarens skicklighet. Vi har även funnit att antalet förare som kommer i mål har en stor påverkan på modellens estimat av förarkoefficienterna. För att modellen skall passa ännu bättre bör denna justeras sådant att vi kan ha en generell formel för alla förare. Detta skulle göra att responsvariabeln blir mer jämnt fördelad över hela utfallsrummet, Ω, och inte

längre fungera som ett alternativt poängsystem.

Denna studie har kunnat identifiera och beskriva en betydande utmaning för att bedöma förarnas idrottsliga prestationer över flera tidsperioder. Vi har visat att antalet förare som startar och antalet som kommer i mål varierar kraftigt såväl mellan som inom åren. Dessa variationer har av tidigare studier förbisetts vid modellbyggandet av en multilevel-modell vilken då kan ha kommit att påverka bedömningen och resultatet av förarnas/stallets prestationer.

För att hantera dessa utmaningar har vi formulerat en transformation av en förares placering i ett lopp i syfte att korrigera för antalet startande förare i varje race och därmed göra jämförelsen mellan förarna mer rättvis. Transformationen ligger även till grund för möjligheten att värdera alla slutpositioner, utan att gå miste om information för de förare och stall vilka placerade sig utanför poängposition. Modellen ges således mer information (datapunkter) på lika antalet race i jämförelse med tidigare studier vilka undersökt den bästa föraren genom historien. Dessa metodologiska förbättringar utgör en stark grund för vår analys och studiens ökade validitet, vilket hoppas bidra till ett mer heltäckande och pålitligt resultat.

En brytning av ett lopp kan antingen bero på bilen eller föraren, vilket i sin tur leder till färre observationer. Ett förslag för att bibehålla allt fler observationer från populationen (istället för att utesluta de från modellen) är att inkorporera denna information genom att skapa två dummyvariabeler. Den ena variabeln erhåller värdet 1 för ”avbrott pågrund av bilen” och 0 för ”inte avbrott på grund av bilen”. Andra variabeln erhåller värdet 1 för ”avbrott på grund av förare” och 0 för ”inte avbrott på grund av förare”. Modellen får information om huruvida en förare bröt tävlingen eller inte, och i sådant fall på grund av bilen eller föraren. På så vis kan en modell byggas så att förarar-estimaten (koefficienterna) straffas för om det var förarens fel, och istället neutral om det berodde på bilen. Förare som hade en tendens till att krascha, eller på annat sätt behöva avbryta racet skulle då rangordnas sämre, vilket även medför en mer rättvis bedömning av konstruktörsfördelen.

Framgångar i Formel-1 handlar inte bara om loppet i sig då kvalet har en direkt påverkan på utfallet och därmed stor betydelse. För att bli en av de bästa inom sporten krävs därför att en förare kan prestera i både kval och race. Då vi enbart undersöker racet tittar vi således bara på den ena aspekten av sporten och utesluter att undersöka förarens skicklighet i kval. För framtida studier föreslås att göra en liknande modell där responsvariabeln är i form av slutpositionen i kvalet. Förarestimaten för både kval och race kan sedan slås ihop till en sammanslagen ranking baserat på hur viktigt kvalet är för racet. Genom att med marginal-effekten bestämma hur stor betydelse kvalet har för racet kan man sedan vikta ihop de två estimaten, race och kval, till en sammanslagen ranking.

Litteratur

Abonazel, Mohamed R., Zakariya Yahya Algamal, Fuad A. Awwad och Ibrahim M. Taha (2022). 'A new two-parameter estimator for beta regression model: Method, simulation, and application'. I: *Frontiers in Applied Mathematics and Statistics* 7. DOI: 10.3389/fams.2021.780322.

Allen, Michael Patrick (1997). *The problem of multicollinearity*. URL: https://link.springer.com/chapter/10.1007/978-0-585-25657-3_37.

Bell, Andrew, James Smith, Clive E Sabel och Kelvyn Jones (2016). 'Formula for success: multilevel modelling of Formula One driver and constructor performance, 1950–2014'. I: *Journal of Quantitative Analysis in Sports* 12.2, s. 99–112.

Berrar, Daniel (jan. 2018). 'Cross-Validation'. I: ISBN: 9780128096338. DOI: 10.1016/B978-0-12-809633-8.20349-X.

Brooks, Steve, Andrew Gelman, Galin Jones och Xiao-Li Meng (2011). *Handbook of markov chain monte carlo*. CRC press.

Bürkner, Paul-Christian (2017). 'Brms: An R package for Bayesian multilevel models using Stan'. en. I: *J. Stat. Softw.* 80.1. ISSN: 1548-7660. DOI: 10.18637/jss.v080.i01. URL: http://dx.doi.org/10.18637/jss.v080.i01.

Buxton, Richard (2008). 'Statistics: Multilevel Modelling'. I: *Machine Learning Support Centre*. DOI: https://www.statstutor.ac.uk/resources/uploaded/multilevelmodelling.pdf.

Bürkner, Paul-Christian (2022). *Bayesian regression models using 'Stan' [R package brms version 2.18.0]*. URL: https://cran.r-project.org/web/packages/brms/index.html.

– (2017). 'brms: An R Package for Bayesian Multilevel Models Using Stan'. I: *Journal of Statistical Software* 80, 1–28. DOI: 10.18637/jss.v080.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v080i01.

Chalmers (2015). *Lecture 5, Markov chains*. URL: http://www.math.chalmers.se/Stat/Grundutb/CTH/mve051/1516/Lectures/Lecture5.pdf.

Complex (2020). *What is a spec racing series?* URL: https://www.complex.com/sports/2010/04/what-is-a-spec-racing-series#:~:text=What%27s%20a%20spec%20series%3F%20Also%20called%20%22one-make%22%20series%2C,skill%2C%20and%20more%20often%20provides%20some%20exciting%20action..

Cribari-Neto, Francisco och Achim Zeileis (2010). 'Beta regression in R'. I: *Journal of statistical software* 34, s. 1–24.

Douma, Jacob C. och James T. Weedon (2019). 'Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression'. I: *Methods in Ecology and Evolution* 10.9, s. 1412–1430. DOI: https://doi.org/

10.1111/2041-210X.13234. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13234. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13234.

Eichenberger, Reiner och David Stadelmann (2009). 'Who Is The Best Formula 1 Driver? An Economic Approach to Evaluating Talent'. I: *Economic Analysis and Policy* 39.3, s. 389–406. ISSN: 0313-5926. DOI: https://doi.org/10.1016/S0313-5926(09)50035-5. URL: https://www.sciencedirect.com/science/article/pii/S0313592609500355.

F1 (2022). *Formula 1 announces TV, race attendance and digital audience figures for 2021: Formula 1®*. URL: https://www.formula1.com/en/latest/article.formula-1-announces-tv-race-attendance-and-digital-audience-figures-for-2021.1YDpVJIOHGNuok907sWcKW.html.

f1metrics (2019). *Mathematical and statistical insights into Formula 1*. URL: https://f1metrics.wordpress.com/2019/11/22/the-f1metrics-top-100.

Fan, Zhou (2016). *Lecture notes Bayesian Analysis*. URL: https://web.stanford.edu/class/stats200/Lecture20.pdf.

Faraway, Julian J (2004). *Linear models with R*. Chapman och Hall/CRC.

Ferrari, Silvia och Francisco Cribari-Neto (2004). 'Beta regression for modelling rates and proportions'. I: *Journal of Applied Statistics* 31.7, 799–815. DOI: 10.1080/0266476042000214501.

FIA (2022). *2022 Formula One sporting regulations - Fédération Internationale de l'Automobile*. URL: https://www.fia.com/sites/default/files/2022_formula_1_sporting_regulations_-_iss_5_-_2022-03-15.pdf.

Fox, John (2015). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications. ISBN: 9781483321318. URL: https://books.google.se/books?id=3wrwCQAAQBAJ.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt och Andrew Gelman (2018). *Visualization in bayesian workflow*. URL: https://doi.org/10.48550/arXiv.1709.01449.

Gelman, Andrew, Jessica Hwang och Aki Vehtari (2013). *Understanding predictive information criteria for Bayesian models - statistics and computing*. URL: https://link.springer.com/article/10.1007/s11222-013-9416-2.

Goldstein, Harvey (2011a). '1.10 Cross classification and multiple membership structures'. I: *Multilevel statistical models*. 4. utg. Wiley, 9–10.

– (2011b). 'Convergence of MCMC chains'. I: *Multilevel statistical models*. 4. utg. Wiley, 48–49.

– (2011c). *Multilevel statistical models*. Vol. 4. Wiley.

– (2011d). 'Parameter estimation'. I: *Multilevel statistical models*. 4. utg. Wiley, 19–22.

– (2011e). 'Semiparametric smoothing models'. I: *Multilevel statistical models*. Wiley, s. 289.

Hackenberger, Branimir K (2019). *Bayes or not Bayes, is this the question?* URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6406060/.

Harrell Jr, Frank E (2015). 'Regression Modeling Strategies With Applications To Linear Models, Logistic And Ordinal Regression, And Survival Analysis. pdf'. I.

Harrison, Robert L. (2010). *Introduction to monte carlo simulation.* URL: https://pubmed.ncbi.nlm.nih.gov/20733932/.

Hoffman, Matthew D, Andrew Gelman m. fl. (2014). 'The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.' I: *J. Mach. Learn. Res.* 15.1, s. 1593–1623.

Hyndman, Rob J. och Yanan Fan (1996). 'Sample Quantiles in Statistical Packages'. I: *The American Statistician* 50.4, s. 361–365. ISSN: 00031305. URL: http://www.jstor.org/stable/2684934 (hämtad 2022-12-15).

James, Neil (2017). *Comparing formula 1 challenges of race tracks with Street Circuits.* URL: https://bleacherreport.com/articles/2040725-comparing-formula-1-challenges-of-racetracks-with-street-circuits.

Jaques, Taylore (2016). *Formula One Engine Efficiency.* URL: http://large.stanford.edu/courses/2016/ph240/jaques2/#:~:text=What%20%20makes%20the%20F1%20cars%20so%20green,engines%20that%20are%20four%2C%20six%2C%20or%20eight%20cylinders..

Kesteren, Erik-Jan van och Tom Bergkamp (2022). 'Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage'. I: *arXiv preprint arXiv:2203.08489*.

Knaub, James (jan. 2007). 'HETEROSCEDASTICITY AND HOMOSCEDASTICITY'. I: Vol 2, s. 431–432. DOI: 10.4135/9781412952644.n201.

Lambert, Ben (2018). 'Evaluation of model fit and hypothesis testing'. I: *A student's Guide to Bayesian statistics*. Sage, 280–315.

MacKenzie, Darryl I., James D. Nichols, J. Andrew Royle, Kenneth H. Pollock, Larissa L. Bailey och James E. Hines (2018). 'Chapter 3 - Fundamental Principals of Statistical Inference'. I: *Occupancy Estimation and Modeling (Second Edition)*. Utg. av Darryl I. MacKenzie, James D. Nichols, J. Andrew Royle, Kenneth H. Pollock, Larissa L. Bailey och James E. Hines. Second Edition. Boston: Academic Press, s. 71–111. ISBN: 978-0-12-407197-1. DOI: https://doi.org/10.1016/B978-0-12-407197-1.00004-1. URL: https://www.sciencedirect.com/science/article/pii/B9780124071971000041.

Motorsport (2022). *Insider's guide: How F1 race strategy works.* URL: https://www.motorsport.com/f1/news/how-f1-race-strategy-works/6791893/.

Newell, Chris (2022). *Database images.* URL: http://ergast.com/mrd/db/#csv.

Newsom, Jason (2017). *Distinguishing Between Random and Fixed: Variables, Effects, and Coefficients.* (hämtad 2022-10-17). URL: \url{http://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html}.

Phillips, Andrew J. K. (2014). 'Uncovering Formula One driver performances from 1950 to 2013 by adjusting for team and competition effects'. I: *Journal of Quantitative Analysis in Sports* 10.2, s. 261–278. DOI: doi:10.1515/jqas-2013-0031. URL: https://doi.org/10.1515/jqas-2013-0031.

Pretorius, Louis (2021). *How do F1 teams work?* URL: https://onestopracing.com/how-do-f1-teams-work/.

Schön, David, Lisa Wallin och Petter Wikström (2017). *MVE220 financial risk: An introduction to Markov chains and their applications within finance*. URL: http://www.math.chalmers.se/Stat/Grundutb/CTH/mve220/1617/redingprojects16-17/IntroMarkovChainsandApplications.pdf.

Smithson, Michael och Jay Verkuilen (2006). 'A Better Lemon Squeezer? maximum-likelihood regression with beta-distributed dependent variables.' I: *Psychological Methods* 11.1, 54–71. DOI: 10.1037/1082-989x.11.1.54.

Sortino, Frank, Robert van der Meer, Auke Plantinga och Bernardo Kuan (2010). 'Chapter 3 - Beyond the Sortino Ratio'. I: *The Sortino Framework for Constructing Portfolios*. Utg. av Frank Sortino. Boston: Elsevier, s. 23–52. ISBN: 978-0-12-374992-5. DOI: https://doi.org/10.1016/B978-0-12-374992-5.00003-X. URL: https://www.sciencedirect.com/science/article/pii/B978012374992500003X.

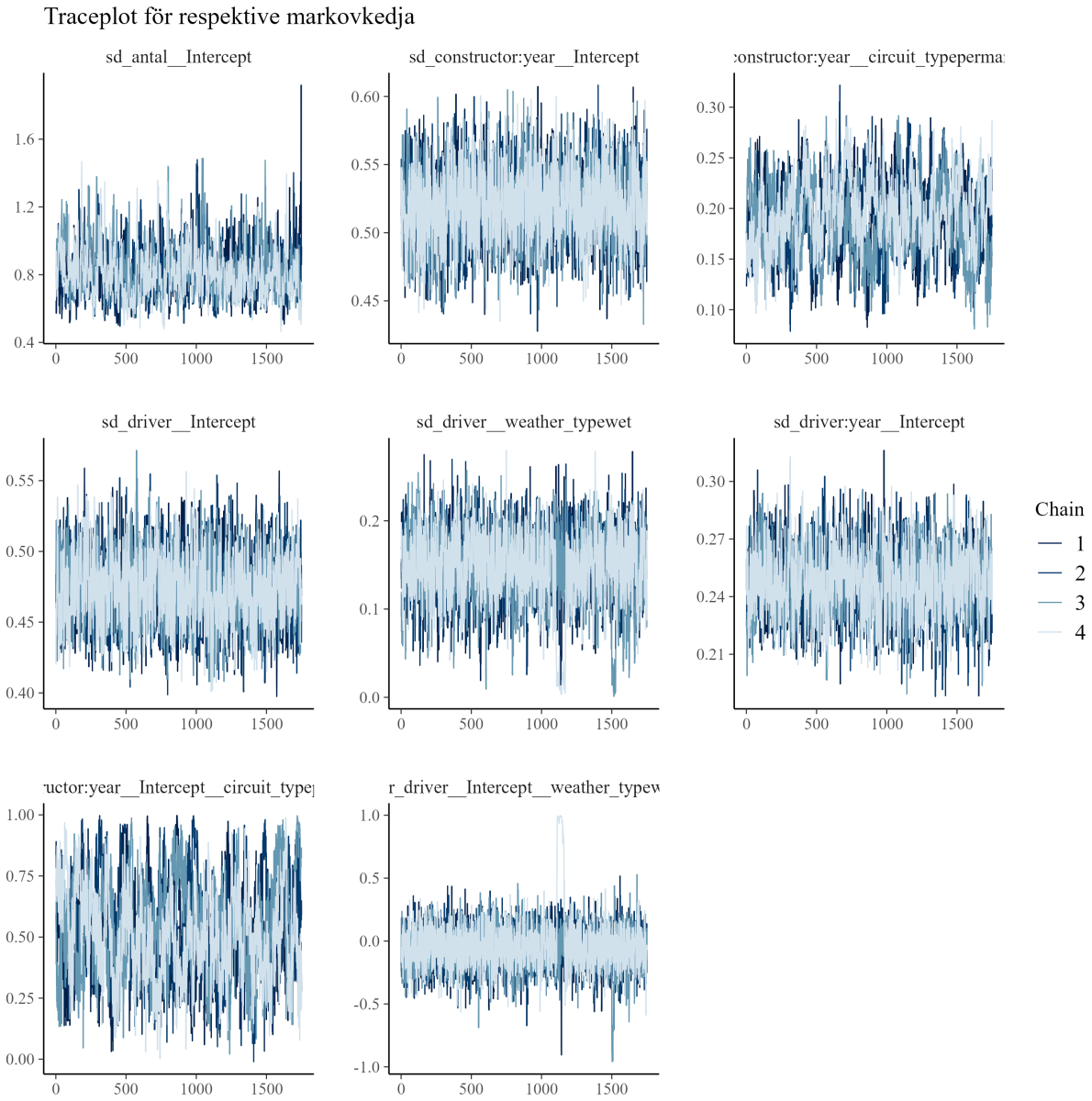
Spiegelhalter, David J. (2020). 'Learning from Experience the Bayesian Way'. I: *The Art of Statistics: Learning From Data*. Pelican, an imprint of Penguin Books, 242–265.

Tim, Johan (2018). *Föreläsning 10, Markovkedjor*. URL: https://users.mai.liu.se/johth11/TAMS79_FO10.pdf.

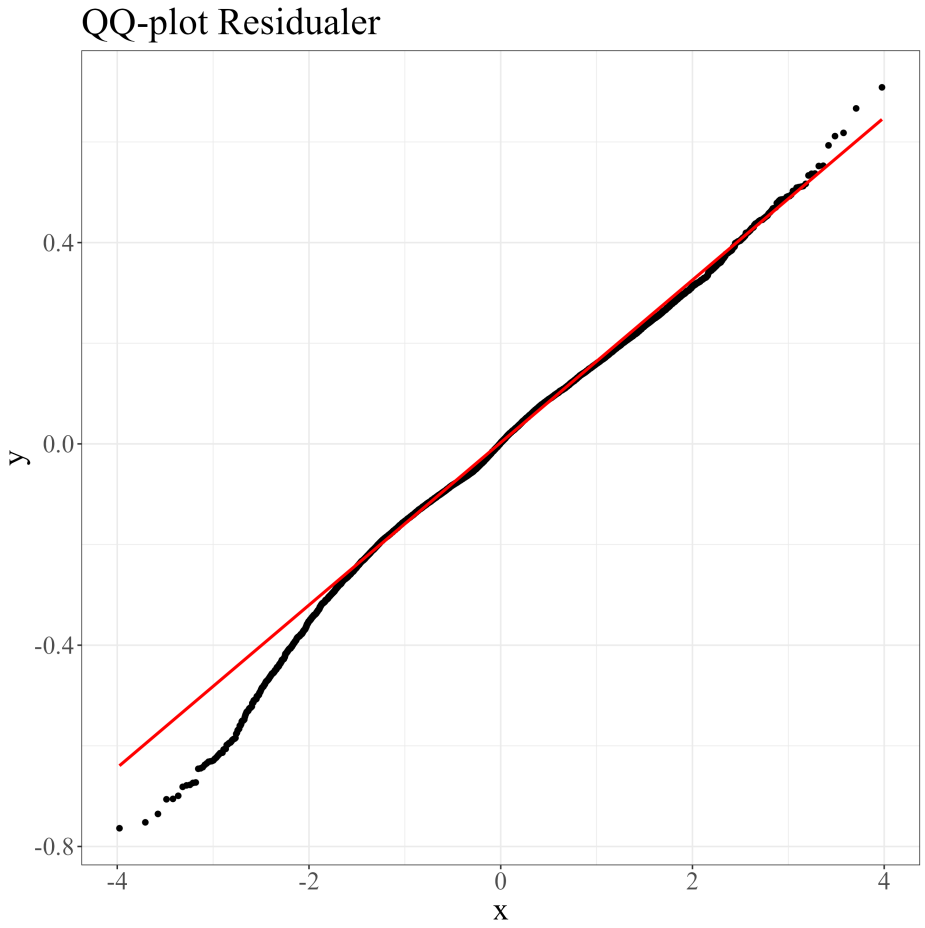
Vehtari, Aki, Andrew Gelman och Jonah Gabry (2016). 'Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC'. I: *Statistics and Computing* 27.5, s. 1413–1432. DOI: 10.1007/s11222-016-9696-4. URL: https://doi.org/10.1007%2Fs11222-016-9696-4.

Bilaga A

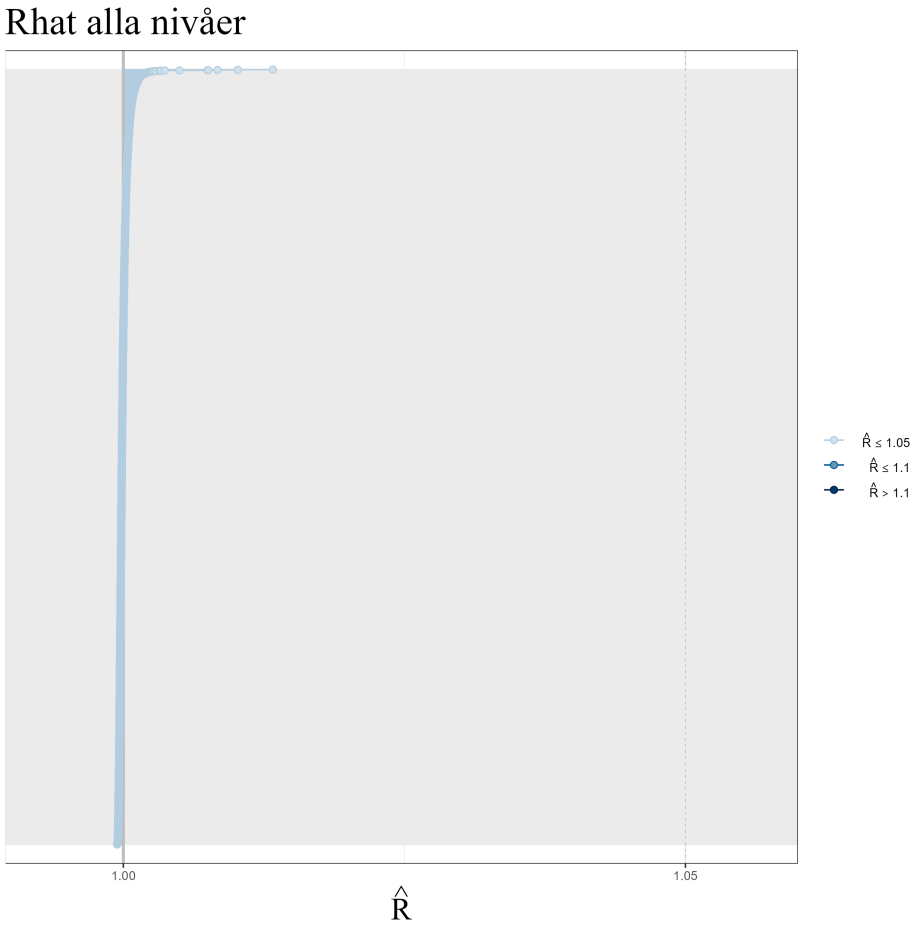
Grafer



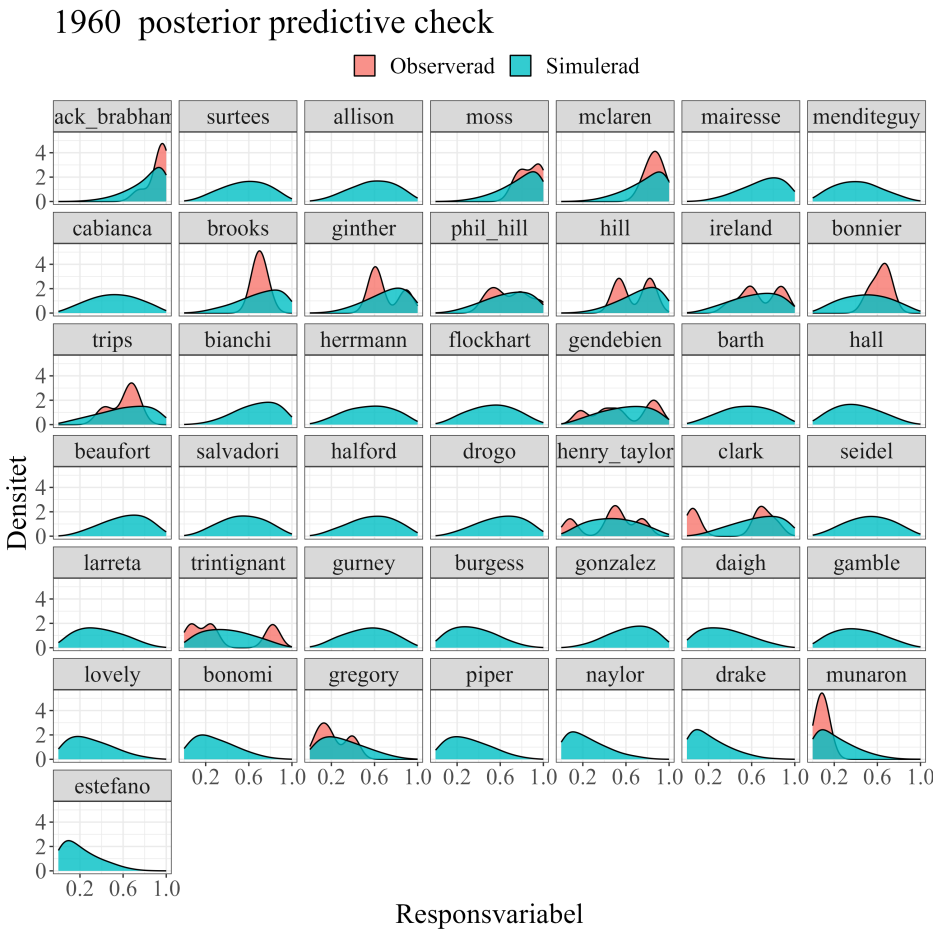
Figur A.1: Markov-kedjor.



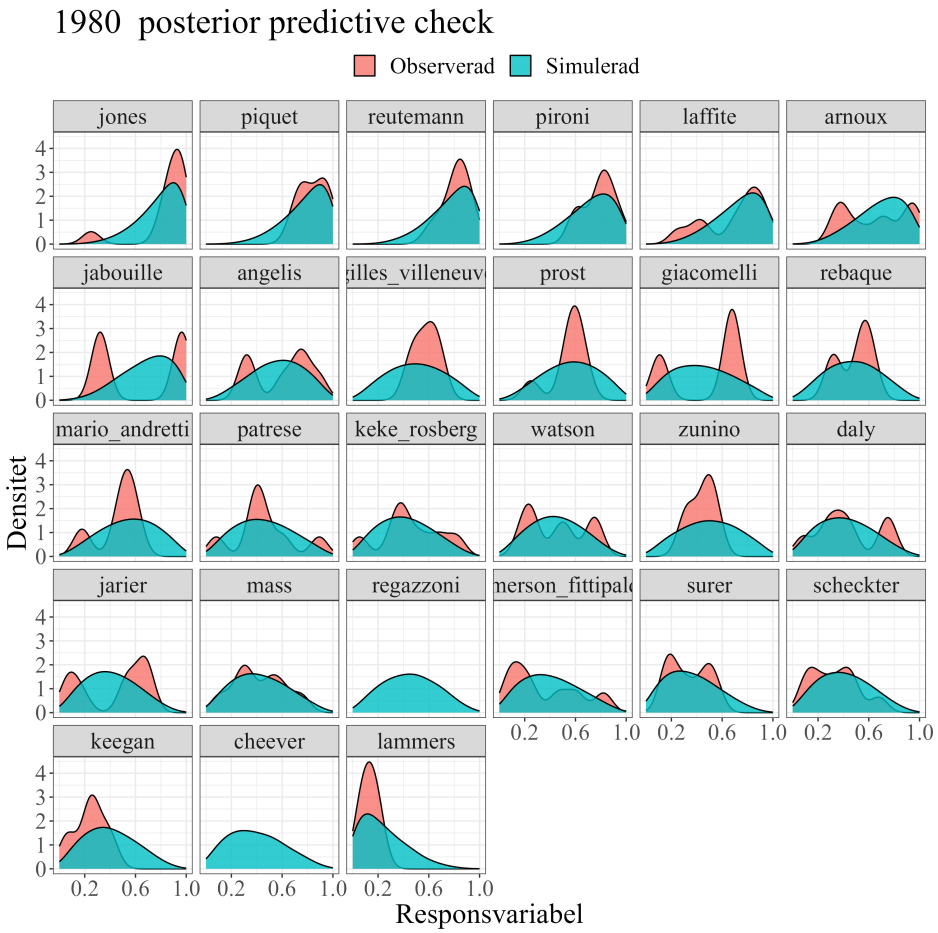
Figur A.2: QQ-plot residualer, bästa modell.



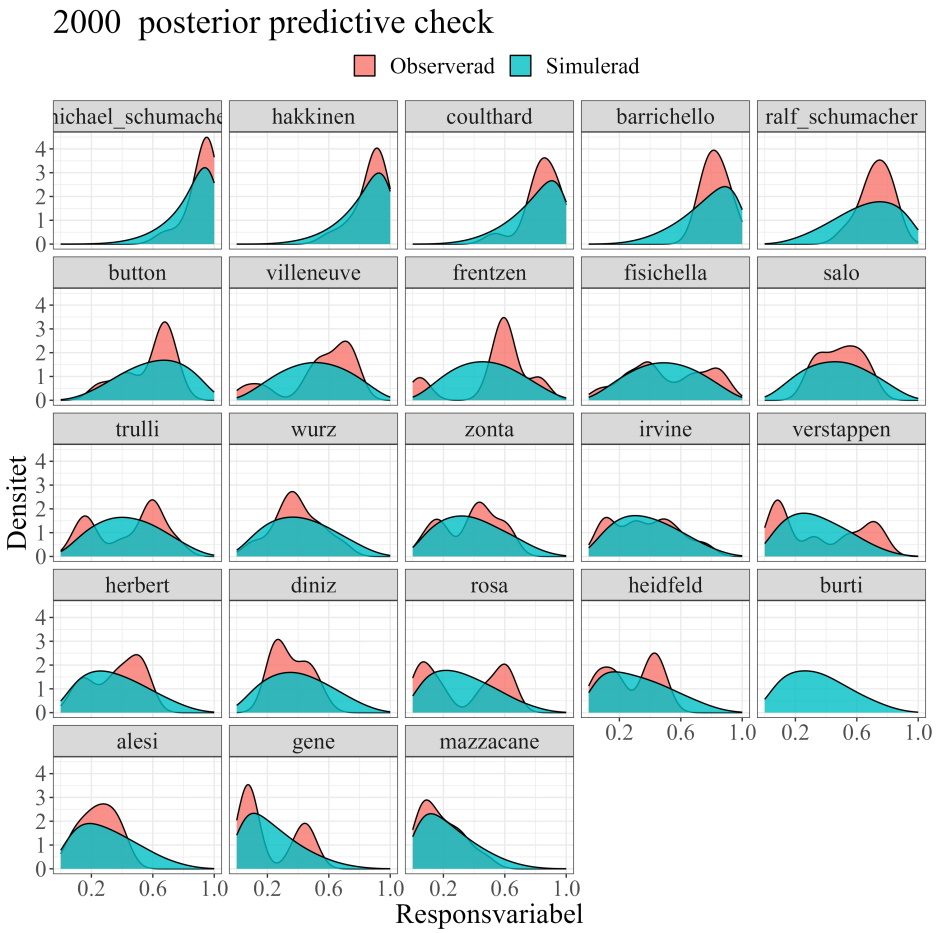
Figur A.3: Rhat alla unika variabler, bästa modell.



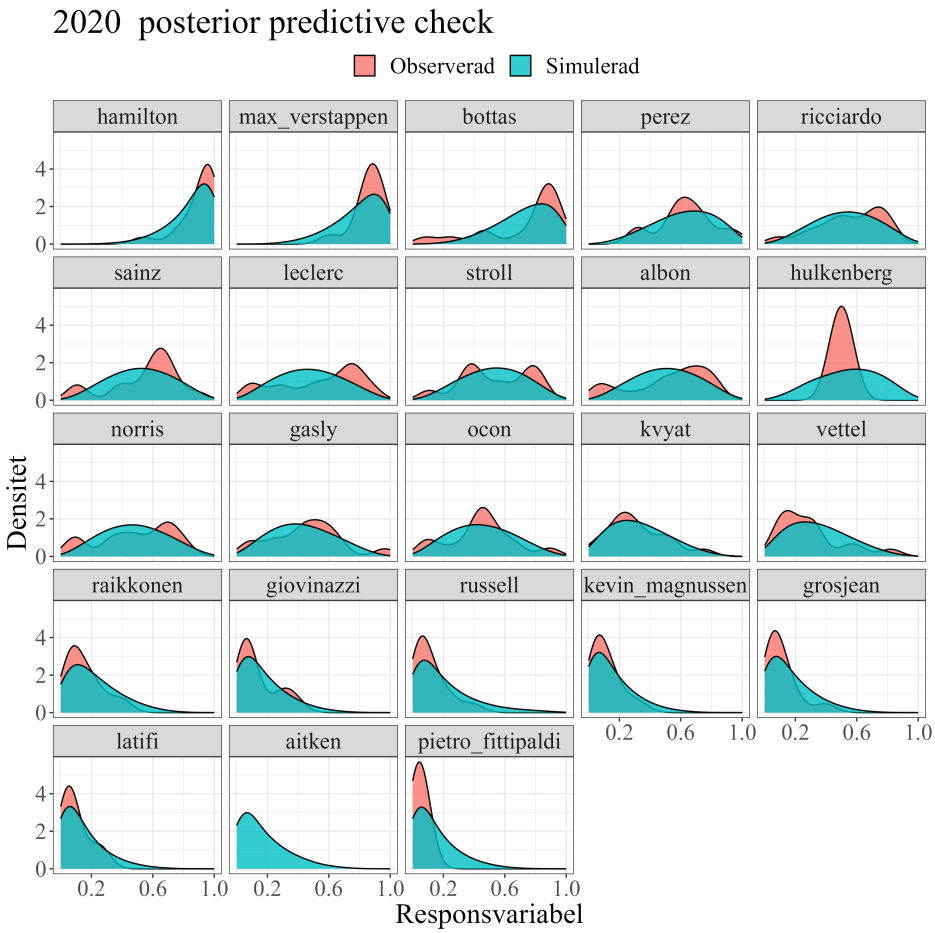
Figur A.4: Posterior predicitive check, år 1960.



Figur A.5: Posterior predicitive check, år 1980.



Figur A.6: Posterior predicitive check, år 2000.



Figur A.7: Posterior predicitive check, år 2020.

Bilaga B

Tabeller

Rank	Förare	År	Est	Lower	Upper	Antal VM-titlar
1	Jim Clark	1963	1.54	1.02	2.05	2
2	Max Verstappen	2021	1.47	1.00	1.94	2
3	Stirling Moss	1961	1.40	0.91	1.91	0
4	Lewis Hamilton	2020	1.39	0.94	1.85	7
5	Juan Manuel Fangio	1957	1.37	0.89	1.85	5
6	James Hunt	1976	1.36	0.84	1.87	1
7	Fernando Alonso	2012	1.36	0.93	1.80	2
8	Alberto Ascari	1952	1.35	0.83	1.89	2
9	Sebastian Vettel	2013	1.28	0.84	1.74	4
10	Jackie Stewart	1969	1.28	0.77	1.79	3
11	Ayrton Senna	1990	1.26	0.79	1.75	3
12	Michael Schumacher	1997	1.26	0.81	1.71	7
13	Alain Prost	1985	1.25	0.76	1.75	4
14	Nico Rosberg	2010	1.22	0.78	1.68	1
15	Jochen Rindt	1970	1.19	0.66	1.75	1
16	Niki Lauda	1978	1.18	0.68	1.67	3
17	Charles Leclerc	2020	1.15	0.66	1.63	0
18	Ronnie Peterson	1978	1.09	0.64	1.56	0
19	Carlos Reutemann	1974	1.08	0.65	1.52	0
20	Nino Farina	1953	1.06	0.55	1.57	1
21	Emerson Fittipaldi	1972	1.05	0.56	1.57	2
22	Jody Scheckter	1974	1.03	0.53	1.52	1
23	Nigel Mansell	1988	1.02	0.50	1.54	1
24	José Froilán González	1952	1.01	0.38	1.62	0
25	Jack Brabham	1960	1.00	0.56	1.47	3

Tabell B.1: Karriärstopp med hänsyn till konstruktörskapacitet. 95% kredibilitetsintervall.

Rank	Förare	År	Est	Lower	Upper	Antal VM-titlar
1	Jim Clark	1962 - 1964	1.46	0.90	2.00	2
2	Max Verstappen	2020 - 2022	1.33	0.84	1.84	2
3	Stirling Moss	1959 - 1961	1.35	0.85	1.85	0
4	Lewis Hamilton	2014 - 2016	1.30	0.85	1.77	7
5	Juan Manuel Fangio	1953 - 1955	1.32	0.81	1.83	5
6	Fernando Alonso	2012 - 2014	1.26	0.83	1.70	1
7	James Hunt	1975 - 1977	1.29	0.76	1.83	1
8	Ayrton Senna	1989 - 1991	1.23	0.74	1.73	3
9	Jackie Stewart	1969 - 1971	1.25	0.74	1.76	3
10	Sebastian Vettel	2011 - 2013	1.18	0.71	1.66	4
11	Michael Schumacher	1995 - 1997	1.19	0.72	1.67	7
12	Alberto Ascari	1951 - 1953	1.20	0.62	1.81	2
13	Alain Prost	1984 - 1986	1.17	0.67	1.68	4
14	Niki Lauda	1976 - 1978	1.15	0.67	1.64	3
15	Nico Rosberg	2010 - 2012	1.12	0.67	1.59	1
16	Ronnie Peterson	1971 - 1973	1.05	0.58	1.52	0
17	Jochen Rindt	1968 - 1970	1.04	0.45	1.63	1
18	Charles Leclerc	2019 - 2021	1.02	0.46	1.56	0
19	Nino Farina	1953 - 1955	0.97	0.39	1.53	1
20	Jody Scheckter	1974 - 1976	0.96	0.44	1.47	1
21	Nigel Mansell	1987 - 1989	0.93	0.41	1.46	1
22	José Froilán González	1951 - 1953	0.93	0.33	1.55	0
23	Emerson Fittipaldi	1972 - 1974	0.92	0.33	1.48	2
24	Jack Brabham	1959 - 1961	0.91	0.40	1.41	3
25	Patrick Depailler	1976 - 1978	0.90	0.40	1.40	0

Tabell B.2: Bästa 3-års period med hänsyn till konstruktörskapacitet. 95% kredibilitetsintervall.