Stockholms universitet

Regressionsanalys av golfares puttförmåga

Daniel Jacobsén

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2021:18
Matematisk statistik
September 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2021:18**
http://www.math.su.se/matstat

Regressionsanalys av golfares puttförmåga

Daniel Jacobsén*

September 2021

Sammanfattning

Syftet med denna rapport är att analysera vilka variabler som kan vara av betydelse för en golfspelares puttförmåga. I datainsamlingen har data från professionella golfare använts, mer specifikt från de högst rankade spelarna på PGA-touren.

I detta arbete används multipel linjär regression, där feltermerna antas vara normalfördelade och responsvariabeln är medelvärdet på intjänade slag på greenen per runda (betecknat i rapporten som Average). En signifikansnivå (α) på 0.05 väljs och för att finna vilka förklaringsvariabler som är signifikanta utförs Backward elimination. Denna metod resulterar i en modell med följande signifikanta variabler; procent av greener som spelaren träffar på rätt antal slag, antal år på PGA-touren, världsranking och amerikansk medborgarskap. Estimatet för variabeln amerikansk medborgarskap är positivt medan de tre andra får negativa estimat enligt modellen.

Denna rapport reserverar sig för vilka slutsatser som kan dras från den, med tanke på det relativt låga R^2-värdet i den slutgiltiga modellen, den begränsade data som har använts (endast data från de senaste åren) och att den enbart är baserad på professionella golfare. När det kommer till andra kategorier av golfspelare är det inte säkert att samma faktorer är signifikanta.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: daniel.jacobsen87@gmail.com. Handledare: Taras Bodnar, Tony Johansson.

Regressionsanalys av golfares puttförmåga

Daniel Jacobsén

15 augusti 2021

Abstract

The English title of this thesis is 'A regression analysis of golfers putting ability'. The purpose of this thesis is to analyze which variables might be of importance for golfers putting ability. In the data collection, data from professional golfers have been used, more specifically from the top ranked players of the PGA-tour.

In this work we are assuming a multiple linear regression with normal-distributed errors where the dependent variable is strokes gained per round (named in report as Average). In order to find the significant (significance level (α) is set to 0.05) explanatory variables, backwards elimination has been made. This method results in a model with significant variables; percentage of green in regulation, years at the PGA-tour, world ranking and nationality of USA. The estimate for nationality USA is positive while the other three are negative.

This thesis makes reservations of the conclusions one might draw from it, considering the relatively low R-squared number in the final model, the limited data that has been used (only the last few years) and that it is only based on professional golfers. When it comes to less skilled golfers, or even professional golfers on other tours, it is not determined whether the same factors are significant.

Förord

Detta arbete utgör ett kandidatexamensarbete på 15 hp på den Matematiska Instutionen vid Stockholms universitet. Till min hjälp har jag haft Taras Bodnar och Tony Johansson som handledare, till vilka jag vill rikta ett stort tack. För idé och rådgivning vill jag dessutom tacka Daniel Bennedich, golfentusiast och expert på PGA-touren.

Innehåll

1 Introduktion **5**
1.1 Problembeskrivning 5
1.2 Existerande forskning 6
1.3 Mått på puttförmåga 6

2 Teori **8**
2.1 Regressionsanalys 8
2.1.1 Enkel linjär regression 8
2.1.2 Multipel linjär regression 8
2.1.3 Parameterskattning och t-test 8
2.1.4 Breusch-Pagan test 9
2.1.5 Villkor 10
2.2 Förklaringsgrad 10
2.3 Stegvis variabelselektion 11

3 Data **13**

4 Resultat **15**
4.1 Bearbetning av data 15
4.2 Enkel linjär regression på datamaterialet 15
4.3 Multipel linjär regression på datamaterialet 16
4.4 Variabelselektion 16
4.5 Kontroll av villkor 17

5 Slutsatser **20**

6 Litteraturförteckning **22**

7 Appendix A **24**

1 Introduktion

Som bakgrund till arbetet bör nämnas att i fotspåren av Covid-19 är trycket på Sveriges golfbanor år 2020 högre än någonsin, med en 42 procentig ökning av spelade ronder jämfört med året innan enligt Svenska Golfförbundet (2021). Därför är det extra intressant att undersöka golfen och vissa av dess grunder. I egenskap av glad amatör finner författaren det fascinerande att analysera detta spel.

Då puttning utgör drygt 40 procent av slagen enligt Golf Digest (2017) tycks det vara ett lämpligt slag att undersöka. Ett putterslag är ett slag med en klubba som kallas putter där golfbollen enbart rullar och inte flyger, vilket gör den lämplig att använda på greenen för att sänka bollen i hål. Greenen är det utrymme närmast hålet i golf som är extra kortklippt och därför lämplig att använda puttern på.

Det kan därför vara av intresse att undersöka vad som gör vissa spelare till bra puttare och med tanke på den data som finns tillgänglig kommer denna rapport att uteslutande baseras på data från spelare på den (amerikanska) PGA-touren. PGA-touren är den högsta amerikanska proffsgolftouren, där de bästa golfspelarna i världen spelar. Denna tour utspelar sig nästan uteslutande i USA, ett land med en stark tradition inom golfen.

I en golftävling spelas, så gott som alltid, 4 rundor där varje runda består av 18 hål. Varje hål har ett förbestämt par med värdemängd 3, 4 och 5. Om du spelar ett hål på det förbestämda paret heter det att du får ett par, om du spelar på ett slag mindre får du en birdie och på ett slag mer, en bogey. Målet är att slå så få slag som möjligt.

För att kunna undersöka puttförmåga hos golfare behövs ett bra mått på denna. I denna rapport kommer intjänade slag per golfrunda att användas, för detaljer se kap 1.3.

1.1 Problembeskrivning

Denna rapport avser att svara på frågan vilka egenskaper (mer formellt variabler) som är signifikanta för en golfspelares puttförmåga. Data har inhämtats från PGA-touren där de bästa golfarna spelar, där man kan anta att de spelare som puttar bäst också finns. Med tanke på detta är det rimligt att reservera sig i slutsaterna från rapporten att faktorer som är signifikanta bland proffspelare inte nödvändigtvis är signifikanta för golfare på en lägre nivå, exempelvis bland rekreationsgolfare.

I syfte att lösa problemet som beskrivits kommer multipel linjär regression att användas. Denna metod väljs för att det är en robust metod som också är tolkningsbar, parameterarna som modellen får fram är väldigt intuitiva att tolka. Stegvis variabelselektion används i arbetet i syfte att finna den modell som förklarar puttförmågan bäst.

1.2 Existerande forskning

I Fearing m fl (2011) påvisas att antalet intjänade slag på green är ett bättre mått på puttförmåga än medelvärdet av antalet puttslag per hål. Detta på grund av att den senare är beroende av hur bollen har spelats in på green. Samma källa visar att en golfspelare tenderar att tjäna fler slag när denne puttar för par än när denne puttar för birdie. Vidare visar källan också att en spelare tycks putta bättre då den inte längre har en reell chans att vinna tävlingen, ett tecken på att nerver kan tänkas påverka puttförmågan.

I Ryan (2021) undersöks vilka faktorer som påverkar en golfspelares världsranking och i denna tycks antalet intjänade slag på greenen resulera i en bättre ranking. Variabeln placerar sig dock storleksmässigt i mitten av 120 variabler.

I Carnahan (2002) har man undersökt amatörgolfare och hur de presterar på olika sorters greener. I denna kunde påvisas en signifikant skillnad i puttförmåga avseende puttar som lutar nedför kontra puttar som lutar uppför, där golfare i den senare kategorin presterade bättre. Däremot fanns ingen signifikant skillnad i spelarens puttande när greenen lutade till höger eller vänster.

1.3 Mått på puttförmåga

Ett rimligt mått på puttförmåga är hur många puttslag en golfare i snitt gör på en tävling. Det finns dock en problematik i detta, som illustreras i Fearing m fl (2011), att det kan vara avhängigt hur pass bra bollen spelas in på green. Exempelvis kan en golfare vara bra på att spela in bollen nära hålet på green och genom det klarar sig på ett mindre antal slag på green än medelsnittet, utan att för den skull nödvändigtvis vara en bättre puttare. Därför används i denna rapport medelvärdet på intjänade slag på greenen per runda, betecknat som *Average*, som ett mått på en golfares puttförmåga. Det innebär att ett högre värde betyder bättre puttförmåga. Detta har enligt PGA (2021) viktats mot hur många slag en spelare i snitt bör göra från ett visst avstånd, vilket gör måttet lämpligt som val av responsvaribel.

2 Teori

2.1 Regressionsanalys

Inom regressionsanalys används linjära regressionsmodeller för att undersöka om en responsvariabel beror på, eller i vilket fall huvudsakligen förklaras av, en eller flera förklaringsvariabler. Definitionen och modellerna i 2.1.1 och 2.1.2 är hämtade från Sundberg (2020).

2.1.1 Enkel linjär regression

Modellen definieras som

$$y_i = \alpha + \beta x_i + \epsilon_i \,,\ i = 1, 2, \ldots, N,$$

där y_i är responsvariabel (dvs mått på puttförmåga), x_i är förklaringsvariabel, α är interceptet, β är lutningskoefficienten, ϵ_i är oberoende normalfördelad och i är de olika observationerna.

2.1.2 Multipel linjär regression

Modellen definieras som

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon_i \,,\ i = 1, 2, ..., N,$$

där y_i är responsvariabel , $x_{i1}, x_{i2}, \ldots, x_{im}$ är förklaringsvariablerna, α är interceptet, $\beta_1, \beta_2, \ldots, \beta_n$ är lutningskoefficienterna, ϵ_i är oberoende normalfördelad, $\sigma^2 = Var(\epsilon_i)$ och i är de olika observationerna.

2.1.3 Parameterskattning och t-test

Vi definierar komponenterna i den multipla linjära regression som

$$Y = (y_1, y_2 \ldots y_N)^T, \beta = (\alpha, \beta_1, \ldots \beta_m)^T$$

samt

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{m1} \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{1N} & \ldots & x_{mN} \end{pmatrix}$$

Enligt Sundberg (2020) skattas β i en multipel linjär regression utav

$$\hat{\beta} = S^{-1} X^T Y \qquad \hat{\beta} \sim N(\beta, \sigma^2 S^{-1})$$

där $S = X^T X$. Då σ^2 i detta fall är okänt och måste skattas är ett t-test lämpligt att användas vid ett hypotestest av en parameter. I denna rapport utförs följande hypotestest för att undersöka om en parameter är signifikant skild från noll. Vi testar nollhypotesen

$$H_0 : \beta_j = 0$$

mot

$$H_1 : \beta_j \neq 0.$$

En skattning av den j:te parametern i regressionen ($\hat{\beta}_j$) kommer att ha väntevärde σ_j och varians $(\sigma^2 S^{-1})_{jj}$. Enligt Gut (2009) är en komponent ($\hat{\beta}_j$) av en multivariat fördelning ($\hat{\beta}$) också normalfördelad. Då denne kan antas vara en normalfördelning med okänd varians kan vi med hjälp av ett t-test skapa teststatistikan

$$T = \frac{\beta_j}{\hat{\sigma}\sqrt{S_{jj}^{-1}}}.$$

Vi definierar m som antalet parametrar i β och α som signifikansnivå. Då kan nollhypotesen förkastas om

$$|T| \geq t_{\alpha/2}(N - m - 1)$$

där $t_{\alpha/2}(N - m - 1)$ är t-fördelningens kvantil. Att nollhypotesen förkastas innebär att parametern β_j är signifikant skiljd från noll och att tillhörande variabel x_j har signifikant inverkan på responsvariabeln. Testets tillhörande p-värde är

$$P(t(N - m - 1) \geq |T|)$$

och är det p-värde som används i detta arbete om inget annat nämns. I utrycket är $t(N - m - 1)$ t-fördelningen med $N - m - 1$ frihetsgrader.

2.1.4 Breusch-Pagan test

Ett test för att undersöka om heteroskedasticitet råder, det vill säga att variansen av feltermerna i modellen inte är konstant. Vi skriver modellen

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon_i$$

där

$$Var(\epsilon_i) = h(\gamma_0 + \gamma_1 z_{i1} + ... + \gamma_p z_{ip}).$$

Här utgörs nollhypotesen av $0 = \gamma_1 = \gamma_2 = \cdots = \gamma_p$. Det utförs en regression av y med förklaringsvariabler x vilket ger oss $\hat{\epsilon}$. I nästa steg utförs en regression av $\hat{\epsilon}^2$ med z som förklaringsvariabler. Från den senare regressionen fås statistikan $N \cdot R^2$, där N är antalet observationer och R^2 förklaringsgraden för $\hat{\epsilon}^2$-modellen. Enligt Breusch (2021) antas $N \cdot R^2$ vara χ^2-fördelad med p frihetsgrader. Nollhypotesen förkastas om

$$N \cdot R^2 \geq \chi^2_\alpha(p)$$

där högerledet är $\chi^2(p)$-fördelningens α-kvantil. Att nollhypotsen förkastas innebär att heteroskedasticitet antas råda. Tillhörande p-värde utgörs av

$$P(\chi^2(p) \geq N \cdot R^2).$$

2.1.5 Villkor

Det finns vissa förutsättningar som är grundläggande vid utförande av linjär regression. Linjäritet är en sådan, där det förutsätts att det finns linjäritet mellan responsvaribeln och förklarande variabler.

Något man vill undvika vid linjär regression är att linjäritet mellan förklaringsvaribler existerar, enligt Sundberg (2020). Det kallas för att det då råder multikollinearitet, eller nästan-kollinearitet, och kan resultera i en modell med förklaringsvaribler som egentligen inte är linjär med responsvaribeln eller det motsatta. Detta kan undersökas till exempel med en korrelationsmatris eller beräkning av den så kallade variansinflationsfaktorn (VIF).

Det förutsätts också att residualerna (ϵ_i) är normalfördelade med väntevärde noll samt att de är okorrelerade. Huruvida dessa är normalfördelade kan undersökas med en normalfördelningsplot. Det förutsätts också att residualernas varians är konstant och om så inte är fallet råder heteroskedasticitet. I en sådan situation kan en transformation av data vara lämplig. I syfte att kontrollera för detta utförs ett Breusch-Pagan test.

2.2 Förklaringsgrad

Förklaringsgrad, eller R^2, är det vanligaste anpassningsmåttet i linjära modelleringar och definieras som andelen av den totala variationen som en modell förklarar. Denna är en bra metod för att jämföra olika modeller och beräknas enligt

$$R^2 = \frac{Kvs(regression)}{Kvs(totalt)} = 1 - \frac{Kvs(resudial)}{Kvs(totalt)}$$

där Kvs står för kvadratsumma. Denna har sin värdemängd mellan 0 och 1 där ett högre värde tyder på en bättre modell. Ett problem med förklaringsgraden som anpassningsmått är att varje gång du tillför en variabel så ökar R^2, oavsett om den tillför modellen något eller inte. Det kan istället vara lämpligt att undersöka om $\hat{\sigma}^2$ minskar vilket kan tolkas som att mindre slump återstår i modellen. Detta mått kallas för adjusted R^2 (R^2_{adj}) och defineras enligt Sundberg (2020) som

$$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2_0}$$

där $\hat{\sigma}^2$ utgörs av

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{N - m - 1}.$$

I ovanstående utryck är m antalet förklaringsvariabler i modellen (minus α-variabeln), N antalet observationer och $\hat{y}_i$ är det skattade värdet av y_i. Dessutom är $\hat{\sigma}_0^2$ variansskattningen utan x-variabel i modellen och utgörs av

$$\hat{\sigma}_0^2 = \frac{\sum(y_i - \bar{y})^2}{N-1}$$

där $\bar{y}$ är medelvärdet av y. Sammantaget får vi formeln

$$R_{adj}^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \cdot \frac{N-1}{N-m-1}.$$

2.3 Stegvis variabelselektion

Stegvis variabelselektion är ett samlingsnamn för metoder att från en stor uppsättning förklaringsvariabler välja vilka som skall inkluderas i en regressionsmodell. I dessa ökar eller minskar man antalet varibler tills ett stoppkriterium uppnås. Dessa metoder kräver beräkningsmässigt betydligt mindre arbete än att testa varje möjlig uppsättning variabler som modell.

I **Backward elimination** inkluderar man inledningsvis alla förklaringsvariabler och tar successivt bort den variabel som, på en förvald nivå, inte är signifikant. Om flera varibler inte är signifikanta väljs den med högst p-värde. Detta fortgår tills att enbart signifikanta variabler återstår i modellen.

Forward selection inleds med modellen utan x-variabler, dvs med enbart interceptet. Sedan inkluderas successivt den variabel som är mest signifikant, tills dess att modellen inte kan expanderas med några fler signifikanta varibler. Dessa två metoder behöver inte nödvändigtvis resultera i samma modell enligt Sundberg (2020).

3 Data

Detta avsnitt ämnar att precisera vilken data som har använts i arbetet. Som tidigare nämnts består datan av egenskaper från spelare på PGA-touren, närmare bestämt de $n = 200$ golfspelare som under 2020 hade bäst puttförmåga (för definition se kapitel 1.3). Till detta har sedan världsranking och kompletterande data för dessa spelare hämtats från PGA (2021) och Sportsdata (2021). Observera att den senare källan kräver att man skaffar ett konto på sidan.

De förklaringsvariabler som inhämtats är golfspelarnas rådande världsranking, ålder, vilket år de debuterade på proffstouren, vikt, land, hur ofta denne träffar green på rätt antal slag (också kallat greens in regulation i procent) samt vilket håll golfaren svingar klubban (R för right och L för left). I tabell 1 beskrivs de olika variablerna och i figur 1 tar vi en titt på datan som används.

Inom detta arbetes avgränsning har inga fler variabler inkluderats, varför vi reserverar oss för att det absolut kan finnas fler variabler som har påverkan på en golfares puttförmåga.

Variabelnamn	Beskrivning
Average	Medelvärde på antalet slag spelaren tjänar på green per runda
Världsranking	Ranking anno 2020
Vikt	Spelarens vikt (lbs)
Swing	Vilket håll golfaren svingar klubban (R eller L)
PgaDebut	Årtal då spelaren gjorde sin debut på PGA-touren
Land	Land som spelaren representerar
Ålder	Spelarens ålder (år)
GreenProcent	Andelen greener golfaren träffar på rätt antal slag (procent)

Tabell 1: Variabelbeskrivning

AVERAGE	world_rank	Weight	Swings	PgaDebut	Country	age	green_percent
0.369	179	160	R	1998	USA	50.28337	62.59
0.366	7	175	R	2016	USA	48.53936	68.39
0.365	14	161	R	2010	England	48.76386	66.28
0.363	293	175	R	2010	USA	49.26762	67.98
0.362	366	155	R	2011	USA	49.01574	66.03
0.347	73	185	R	2004	USA	49.88090	64.93
0.340	25	195	R	2007	USA	49.57700	69.05

Figur 1: Dataöversikt

4 Resultat

4.1 Bearbetning av data

Vi inför variablerna:

- Tourår = 2020 - PgaDebut,

samt

- LandUSA = 1 om USA, 0 annars.

Syftet med detta är i första fallet en transformation av data för att lättare kunna tolka denna och med den senare att så pass många världsledande golfspelare är från USA, varvid det kan tänkas vara intressant att undersöka om detta även kan tänkas vara en faktor för puttningsförmåga.

4.2 Enkel linjär regression på datamaterialet

I nästa steg utförs enkel linjär regression på respektive förklaringsvariabel. Syftet med detta är för att undersöka vilka variabler som för sig själv är signifikanta på nivån 0.05 (se P-värdet i tabell 2) samt hur väl de förklarar responsvaribeln (se R^2 i tabell 2).

Variabel	P-värde	R^2
Världsranking	0.00019	0.071
Vikt	0.75	0.00060
SwingsR	0.81	0.0003
TourÅr	0.0051	0.023
LandUSA	0.00016	0.080
Ålder	0.020	0.031
GreenProcent	0.0090	0.047

Tabell 2: Enkel Regression

4.3 Multipel linjär regression på datamaterialet

Ur tabell 3 kan utläsas att svaga korrelationer mellan variblerna generellt råder men att det finns en stark sådan mellan ålder och antal år på touren för en golfare. Vidare kan utläsas i tabell 4 att VIF-värdena på dessa är ca 12. Enligt Sundberg (2020) bör man vid värden över 5 överväga att utesluta en variabel vid regression. Det kan därmed antas att det råder multikollinjäritet mellan TourÅr och Ålder, varvid Ålder utesluts ur modellen på grund av dess högre p-värde.

Detta ger oss grundmodellen

$$\begin{aligned} Average_i = \alpha + \beta_1 Världsranking_i + \beta_2 Vikt_i + \beta_3 TourÅr_i + \beta_4 LandUSA_i \\ + \beta_5 GreenProcent_i + \beta_6 SwingsR_i + \epsilon_i\,,\ i = 1, 2, ..., N. \quad (1) \end{aligned}$$

	Världsranking	Vikt	TourÅr	LandUSA	Ålder	GreenProcent
Average	-0.27	-0.02	-0.15	0.28	-0.17	-0.22
Världsranking	1	-0.04	0.01	-0.25	0.28	-0.23
Vikt		1	-0.02	-0.08	0.02	0.10
TourÅr			1	-0.09	0.79	-0.04
LandUSA				1	-0.21	-0.02
Ålder					1	-0.01
GreenProcent						1

Tabell 3: Korrelationsmatris

Variabel	VIF-Värde
Världsranking	1.26
Vikt	1.04
SwingsR	1.10
TourÅr	11.69
LandUSA	1.10
Ålder	11.86
GreenProcent	1.13

Tabell 4: VIF-Värden

4.4 Variabelselektion

I nästa steg i regressionen utförs variabelselektion medelst Backward elimination och Forward selection (se kapitel 2.3). Båda metoderna ger samma resultat, att variablerna Vikt samt SwingsR bör uteslutas ur modellen. Detta ger oss

$$Average_i = \alpha + \beta_1 Världsranking_i + \beta_2 TourÅr_i + \beta_3 LandUSA_i + \beta_4 GreenProcent_i + \epsilon_i \,,\ i = 1, 2, ..., N \quad (2)$$

som lämplig modell.

Den justerade förklaringsgraden (R^2_{adj}) för (2) beräknas till 0.22 och är densamma som för grundmodellen (1). I ett försök att förenkla modellen genom att ta bort ytterligare en varibel sjunker R^2_{adj} till 0.17 eller lägre, vilket stärker uppfattningen om att ingen mer faktor bör tas bort. I valet mellan modeller (1) och (2) föredras den senare på grund av mindre komplexitet, 4 variabler istället för 6, samt att enbart signifikanta variabler inkluderas i (2).

Estimaten för de olika parametrarna i (2), dvs β_1, β_2 osv, går att utläsas i Tabell 5. Enligt modellen råder ett negativt linjärt förhållande mellan världsranking och medelvärdet på intjänade slag på green per runda. Där ett steg högre på rankingen skulle motsvara 0.001 färre intjänade slag per runda. Vidare är förhållandet mellan antal år på touren och responsvariabeln också det svagt negativt, med -0.01 slag per år.

Att vara av amerikansk nationalitet tycks enligt denna modell vara positivt för en spelares puttande då denna faktor tycks innebära att en spelare i snitt tjänar 0.17 slag per runda på greenen. Slutligen tycks förmågan att träffa greenen på rätt antal slag ha en negativ effekt på intjänade slag på greenen, med cirka 0.06 färre intjänade slag per procent i träffade greener.

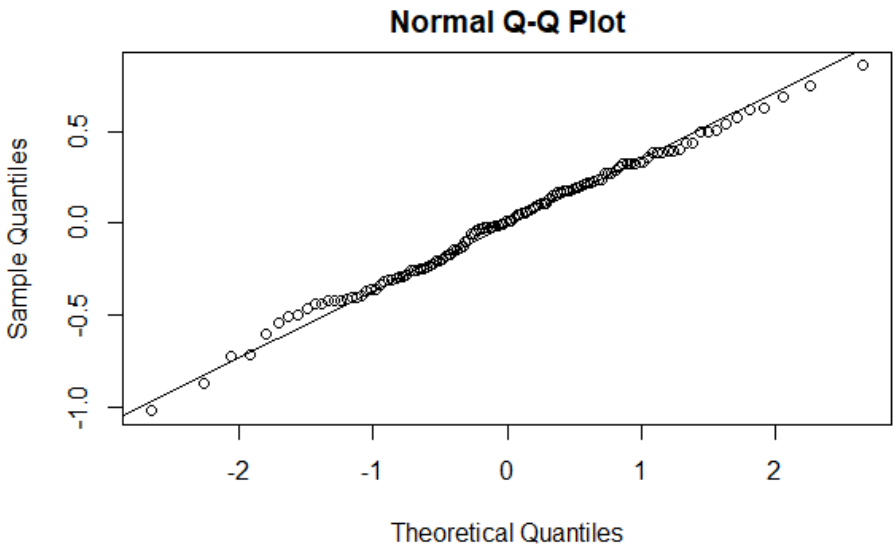
Parameter	Estimat	p-värde
Världsranking	-0.0011	0.000024
TourÅr	-0.010	0.049
LandUSA	0.17	0.0067
GreenProcent	-0.057	0.00026

Tabell 5: Parameterskattningar och dess p-värden

4.5 Kontroll av villkor

I syfte att undersöka om antagandet om normalfördelade residualer håller ritas en normalfördelnings-plot i R (se figur 2). Där kan vi se att punkterna följer den önskade linjen väl vilket påvisar att antagandet bör hålla.

Figur 3 visar att residualerna tycks slumpmässigt fördelade kring noll. Dessutom utförs ett Breusch-Pagan-test av modell (2) som resulterar i ett p-värde på 0.40. Detta indikerar att heteroskedasticitet inte tycks råda och att transformation av data inte är nödvändig.

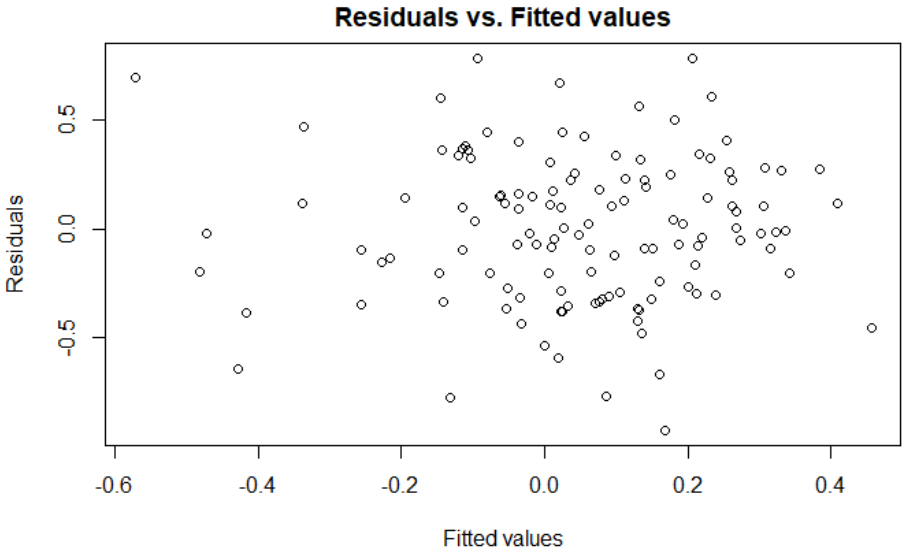


Figur 2: Normalfördelningsplot för modellen

I syfte att undersöka om multikollinearitet råder presenteras i tabell 6 VIF-värdena för modell (2). Dessa är alla långt under 5 och vi kan därmed utesluta att multikollinearitet råder.

Variabel	VIF-Värde
Världsranking	1.10
TourÅr	1.00
LandUSA	1.02
GreenProcent	1.09

Tabell 6: VIF-Värden Modell (2)



Figur 3: Residual-plot för modellen

5 Slutsatser

Från resultatet av den linjära regressionen med intjänade slag på greenen, Average, som responsvariabel, är det lämpligt att hänvisa till estimaten i tabell 5 för att söka dra slutsatser om egenskaper som skulle kunna påverka en golfares puttförmåga.

Esitmatet för förklaringsvariabeln Världsranking är negativt. Enligt modellen tolkas det som att sämre världsranking (det vill säga högre siffra) ger färre intjänade slag på green, vilket således kan tolkas som att sämre världsranking innebär att man är en sämre puttare. Teoretiskt skulle 100 placeringar sämre på världsrankingen motsvara 0.1 extra slag på greenen per runda. I Ryan (2021) beskrivs att även om intjänade slag på green har en positiv korrelation med bättre (lägre siffra) världsranking, så placeras denna storleksmässigt i mitten av hela 120 variabler. Detta stämmer ganska bra överens med resultaten i denna rapport.

Variabeln TourÅr får ett negativt estimat, vilket kan tolkas som att ju längre en golfare spelat på PGA-touren desto sämre puttar denne. Detta kan låta ologiskt, erfarenhet borde ge en bättre puttare, men andra faktorer kan tänkas ligga bakom och en annan möjlighet är att beroendet inte är linjärt.

Från samma tabell kan utläsas att estimatet för GreenProcent, dvs på hur många procent av hålen en golfare träffar greenen på rätt antal slag, är negativt. Detta skulle enligt modellen innebära att en golfare som träffar greenen exempelvis 10 % oftare puttar i genomsnitt 0.6 extra slag per runda. Vad detta skulle kunna bero på besvarar inte denna rapport men kanske har golfaren tränat mer på inspel än på puttning och det kan antas rimligt att man har sin styrka i en del av spelet och inte i alla delar.

Golfare med amerikanskt medborgarskap verkar enligt modellen tjäna 0.17 slag per runda på greenen. Vad detta kan bero på omfattas inte av detta arbetes avgränsningar.

Viktigt att poängtera är att försiktighet bör iakttas när slutsatser dras från denna rapport av tre olika anledningar. Dels är det viktigt att beakta den relativt låga förklaringsgraden som den slutliga regressionsmodellen har, där R^2_{adj} beräknats till 0.22. Det tycks som att det finns en stor slumpfaktor, alternativt variabler som inte inkluderats i datan för detta arbete, som påverkar en golfares puttförmåga. Därför rekommenderar författaren till detta arbete att fortsatt forskning i ämnet görs för att söka ytterligare variabler som kan tänkas påverka en golfares puttförmåga. Vidare spänner datan i detta arbete över ett begränsat tidsintervall, andra resultat hade eventuellt uppnåtts om ett större intervall inkluderats. Slutligen härstammar datan enbart från proffspelare, vilket kanske inte alltid är applicerbart på andra kategorier av golfspelare.

6 Litteraturförteckning

Breusch (2021): Breusch–Pagan test, https://en.wikipedia.org/wiki/Breusch-Pagan_test. Hämtad augusti 2021.

Carnahan (2002): Carnahan, J.V, Experimental Study of Effects of Distance, Slope, and Break on Putting Performance for Active Golfers, Science and Golf IV: Proceedings of the World Scientific Congress on Golf, 113-126. Routledge, London.

Fearing D., Acimovic J och Graves S. (2011): How to Catch a Tiger, Understanding Putting Performance on the PGA TOUR. Journal of Quantitative Analysis in Sports, Vol 7.

Golf Digest (2017): How many putts does the average golfer make? https://www.golfdigest.com/story/how-many-putts-does-the-average-golfer-make-new-data-shows-you-need-more-time-on-the-practice-greenand-the-range. Hämtad april 2021.

Gut (2009): Gut, Allan. 2009. An Intermediate Course in Probability, sida 121. Second Edition, Springer.

PGA (2021): SG-Putting, https://www.pgatour.com/stats.html. Hämtad januari 2021.

Ryan (2021): Ryan, Shane, 2021. The golf stats that matter most on the PGA Tour. I: Golfdigest 11.1.2021.

Sportsdata (2021): Historical sports data, https://sportsdata.io/pga-golf-api. Hämtad i mars 2021.

Svenska Golfförbundet (2021): Statstik. https://golf.se/om-golfsverige/analyser-och-statistik/statistik/. Hämtad mars 2021.

Sundberg (2020): Sundberg, Rolf, 2020. Linjära Statistiska Modeller, Kompendium Augusti 2020. Stockholm: Matematiska Instutionen.

7 Appendix A

```
Call:
lm(formula = AVERAGE ~ world_rank + Years_played + Land + green_p
    data = df_PlayerData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.92185 -0.26675 -0.00945  0.22668  0.78229 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.043340   1.036390   3.901 0.000158 ***
world_rank    -0.001062   0.000259  -4.102 7.51e-05 ***
Years_played  -0.010201   0.005127  -1.990 0.048887 *  
Land           0.174508   0.063209   2.761 0.006672 ** 
green_percent -0.057138   0.015176  -3.765 0.000259 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figur 4: R-Utskrift över parameterskattningar