Erasmus Universiteit Rotterdam

Erasmus School of Economics

Bacherlorscriptie
Econometrie en Operationele Research

Het voorspellen van de marktwaarden van voetbalspelers met behulp van een ordered logit model

Auteur:
Nishant Ramsaroep
343920

Supervisor:
Dr. Andreas Alfons

30 juni 2014

Samenvatting

In dit scriptieonderzoek worden de marktwaarden van de voetbalspelers onderzocht in 8 verschillende Europese competities. De acht competities zijn Eredivisie, Jupiler Pro League, Bundesliga(Dui), Bundesliga(Oos), Ligue 1, Premier League, Serie A en de Primera Division. Tijdens dit onderzoek worden modellen gemaakt die de marktwaarden van voetbalspelers voorspellen waarbij er in de verschillende modellen onderscheid wordt gemaakt tussen drie categorieën namelijk de lage,midden en hoge categorie. Dit onderscheid wordt gemaakt omdat tijdens het onderzoek van het werkcollege is gebleken dat de marktwaarden van alle voetbalspelers niet te omvatten zijn in één model. We vergelijken de voorspellingen van de 3 categorie modellen met het algeheel model om te onderzoeken of de marktwaarden van voetbalspelers beter voorspelt kunnen worden door modellen, die onderscheid maken in wat voor categorie een speler valt, dan één algeheel model dat geen onderscheid maakt. De conclusie die uit dit onderzoek getrokken kan worden is dat de voorspellingen van de marktwaarde van voetbalspelers beter zijn als er drie categorie modellen worden gebruikt mits de voorspelde categorieën kloppen.

Inhoudsopgave

1. Inleiding

In Juni 2003 kocht de Russische oliemagnaat Roman Abramovich de Engelse voetbalclub Chelsea FC. Voor de overname was Chelsea F.C. nog een Premier League club die aan het vechten was tegen degradatie maar door de financiële hulp van Abramovich veranderde dit al snel . Sinds de overname heeft Roman Abramovic al meer dan twee miljard pond in de club geïnvesteerd en dit resulteerde in twaalf prijzen voor de club. Dit fenomeen, een rijke investeerder die een club koopt, zien we laatste jaren steeds vaker in de Europese competities. Manchester City F.C. werd in 2008 gekocht door de Abu Dhabi United Group, Paris Saint-Germain werd in 2011 gekocht door de Qatar Sports Investments en AS Monaco F.C werd in 2011 gekocht door de Russische miljardair Dimitri Rybolovlev. Zelfs in Nederland gebeurt dit. Zo werd Vitesse in 2010 gekocht door de Georgische ondernemer Merab Zjordania.
Experts geloven dat deze overnames voor een grote verschuiving hebben gezorgd op de transfermarkt voor voetballers. De overnames hebben ervoor gezorgd dat er tegenwoordig veel meer geld omgaat in het voetbal en dit heeft ervoor gezorgd dat er steeds hogere bedragen voor spelers worden betaald. Zo zijn acht van de tien hoogste transferbedragen betaald tussen 2009 en nu. De transferwaarde van een speler wordt bepaalt door zijn marktwaarde. Indien het mogelijk zou zijn om de juiste marktwaarde van een voetbalspeler te bepalen zouden verschillende partijen hier profijt van kunnen hebben:

Ten eerste, de voetbalclubs. Als een voetbalclub de waarde van een voetbalspeler kan bepalen, kan hij dit gebruiken wanneer er biedingen op deze speler komen tijdens de transferperiode. Zo kunnen zij ervoor kiezen om de speler te verkopen als het bedrag dat geboden wordt hoger ligt dan zijn daadwerkelijke marktwaarde. Clubs kunnen deze informatie ook gebruiken als zij een speler willen kopen door een bedrag te bieden dat lager ligt dan zijn marktwaarde.

Ten tweede, de voetbalspelers zelf. Als een speler zijn marktwaarde weet kan hij dit ook in zijn voordeel gebruiken. Wanneer hij wordt verkocht tegen een prijs die onder zijn marktwaarde ligt kan hij deze informatie meenemen in de contractonderhandelingen met zijn nieuwe club om bijvoorbeeld een hoger salaris te eisen.

Tijdens het werkcollege (Mutsaerts, Ramsaroep Monster,2014) is er onderzoek gedaan naar de marktwaarden van de voetbalspelers in 8 verschillende Europese competities, de Nederlandse Eredivisie, de Belgische Jupiler Pro League, de Duitse Bundesliga, de Oostenrijkse Bundesliga, de Franse Ligue 1, de Engelse Premier League, de Italiaanse Serie A en de Spaanse Primera Division. Tijdens dit onderzoek zijn er drie modellen gemaakt. Één model die de marktwaarden van alle voetbalspelers bepaalt, één model die alleen de marktwaarden van de keepers bepaalt en één model die alleen die marktwaarden van de veldspelers bepaalt. Het probleem van dit onderzoek is dat de gehele dataset van marktwaarden niet lineair is en dat de extreem lage en extreem hoge marktwaarden dit veroorzaakte. Dit scriptieonderzoek probeert dit probleem te verhelpen.

In dit scriptieonderzoek worden er modellen gemaakt die de marktwaarden van voetbalspelers in 8 verschillende Europese competities voorspellen waarbij er in de verschillende modellen onderscheid wordt gemaakt tussen drie categorieën namelijk, slechte, normale en goede spelers. Hierbij wordt er aangenomen dat de slechtste spelers, de laagste marktwaarden hebben en de beste spelers, de hoogste marktwaarden hebben. In dit onderzoek worden deze modellen vanaf nu aangeduid als categorie modellen Dit onderscheid wordt gemaakt omdat tijdens het onderzoek van het werkcollege is gebleken dat de marktwaarden van alle voetbalspelers niet te omvatten zijn in één model.
De gemaakte modellen zullen worden getest doordat de modellen geschat worden met 75% van

de data, de in-sample data, en de overige 25% van de data, de out-of-sample data, wordt gebruikt om te analyseren hoe goed de modellen daadwerkelijk werken, iets wat niet werd gedaan in het werkcollege. Zo kunnen we zien of de gemaakte modellen betere of juist slechtere voorspellingen geven dan één groot model voor alle voetbalspelers.

Tijdens dit onderzoek worden er bepaalde vragen onderzocht en beantwoord. De hoofdvraag luidt als volgt:
Kan de marktwaarden van voetbalspelers beter bepaalt worden door modellen, die onderscheid maken in wat voor categorie een speler valt, dan één algeheel model dat geen onderscheid maakt ?
Ik maak in drie categorie modellen, voor iedere categorie één model,en die ga ik uiteindelijk met elkaar vergelijken. Daarbij ontstaat de volgende deelvraag:
Zijn er verschillende verklarende variabelen voor de drie verschillende modellen en is er een significant verschil in partieel effect tussen variabelen die overeen komen in de modellen?
Als laatst wordt individueel onderzocht hoe ieder categorie model werkt ten opzicht van het algeheel model als ik voor een speler een categorie heb voorspelt. Het kan zo zijn dat één van de categorie modellen betere voorspellingen geeft dan het algeheel model terwijl een ander categorie model slechtere voorspellingen geeft. Er wordt onderzocht hoe ieder specifiek categorie model werkt ten opzichte van het algemeen model als ik de categorie van een speler heb voorspelt:
Kan het voorspellen van de categorie van een speler ons inzicht geven in welke model we moeten gebruiken?

Dit scriptieonderzoek zal beginnen met een literatuuranalyse over eerdere studies die getracht hebben de marktwaarden van voetbalspelers te voorspellen. Daarna zal er aandacht besteed worden aan de data die is gebruikt voor dit onderzoek. De methodologie zal inzicht geven over hoe de modellen tot stand zijn gekomen en de gebruikte technieken. Ten slotte zullen de resultaten besproken om aan de hand van de resultaten conclusies te trekken om de onderzoeksvragen te beantwoorden.

2. Literatuuranalyse

In dit scriptieonderzoek worden de marktwaarden van de voetbalspelers onderzocht in 8 verschillende Europese competities, de Nederlandse Eredivisie, de Belgische Jupiler Pro League, de Duitse Bundesliga, de Oostenrijkse Bundesliga, de Franse Ligue 1, de Engelse Premier League, de Italiaanse Serie A en de Spaanse Primera Division. Een voetbalspeler is een profvoetballer wanneer hij betaald krijgt om te voetballen. Deze spelers staan onder contract bij een club voor een bepaalde tijd. Tweemaal per jaar is er een transferperiode in Europa. Dit houdt in dat spelers door een andere club gekocht kunnen worden en op die manier van werkgever verwisselen. Het transferbedrag is niet hetzelfde als de marktwaarde. De marktwaarde is een fictief geschat bedrag die is bepaald door een panel van deskundigen. Het transferbedrag is het bedrag dat een club betaalt voor een speler om hem te contracteren.
Er zijn verchillende studies die onderzoek hebben gedaan naar de transferbedragen die voor spelers zijn betaald. Éen van de eerste onderzoeken was die van Carmichael en Thomas(1993). Zij analyseerde de tranferbedragen in de Premier League voor het seizoen 1990/1991. Zij gebruikte de OLS methode. Zij corrigeerde de selectiebias door Heckman twee-staps procedure toe te passen. De belangrijkste variabelen die invloed hadden op de transferbedragen waren volgens hun de leeftijd van de speler, het aantal gespeelde wedstrijden in het afgelopen seizoen en de kenmerken van de verkopende en kopende club.

Reilly en Witt(1995) analyseerde net zoals Carmichael en Thomas de transferbedragen die voor spelers werden betaald. In hun onderzoek namen ze een huidskleur dummie op om de invloed van racisme te onderzoeken. De conclusie van dit onderzoek is dat er geen significant verschil bestaat tussen blanke en gekleurde spelers. Een mogelijke misspecificatie in het model van Reilly en Witt is het onderbreken van de variable $Leeftijd^2$. In het volgend hoofdstuk wordt hierop ingegaan.
In 2000 hebben Dobson,Gerrard en Howe het oorspronkelijke onderzoek van Carmichael en Thomas uit 1993 overgedaan. De spelers die hun gebruikte voor hun onderzoek waren geen Premier League spelers maar spelers die op amateur niveau speelden. De resultaten van dit onderzoek waren dat de betaalde transfersommen voor amateur voetballers gebaseerd waren op dezelfde variabelen en factoren als de variabele die Carmichael en Thomas vonden.

3. Data

3.1. De gehele dataset

De data waarmee is gewerkt is verkregen door mijn supervisor, Dr. Andreas Alfons. Hij verzamelde de data van de Duitse website www.transfermarket.de met behulp van een zelfgeschreven programma. Deze website bevat een grote selectie van voetbalstatistieken van alle professionele voetballers. De dataset waarmee is gewerkt bestaat uit voetbalspelers die op 1 maart 2014 onder contract stonden bij een voetbalclub uit 1 van de 8 geselecteerde competities. De 8 geselecteerde competities zijn de Nederlandse Eredivisie, de Belgische Jupiler Pro League, de Duitse Bundesliga, de Oostenrijkse Bundesliga, de Franse Ligue 1, de Engelse Premier League, de Italiaanse Serie A en de Spaanse Primera Division. De dataset bestaat uit 3782 spelers. De dataset van een speler bevat de spelersstatistieken van alle competities waarin de speler heeft gespeeld vanaf het begin van zijn carrière tot maart 2014. De afhankelijke, te verklaren, variabele is de marktwaarden van de spelers in euro. Twaalf spelers, waarvan www.transfermarket.de geen marktwaarde heeft, worden uit de sample verwijderd. De uiteindelijke dataset bestaat uit 3770 voetbalspelers met ongeveer 300 verschillende variabelen die je kan onderverdelen in 3 verschillende categorieën: Persoonlijke statistieken, Prestatie statistieken en Achtergrond statistieken.

3.2. Persoonlijke statistieken

De persoonlijke statistieken beschrijven de kenmerken van een speler zoals zijn lengte, of een speler rechtsbenig, linksbenig of tweebenig is, en zijn leeftijd. $Leeftijd^2$ hebben we in het werkcollege ook gecreëerd. Deze twee leeftijdsvariabelen maken het mogelijk dat de marktwaarde van een speler stijgt tot een bepaalde leeftijd en na die leeftijd daalt. Er zijn ook dummies gecreëerd voor de continent van een speler, dit werd gedaan door middel van het filteren van de nationaliteit door middel van een bepaalde lijst van landen per continent. Het maken van dummies voor de continent van een spelers werd geprefereerd boven het creëren van dummies voor afzonderlijke landen aangezien dit voor veel minder dummies zorgde. Deze continent dummies zijn handmatig toegevoegd en is ontstaan uit de oorspronkelijke data. Bij het schatten van de modellen is de continent dummie Europa de basiscategorie en daarom niet opgenomen in het model. Tenslotte hebben we ook dummie variabelen toegevoegd om de positie van de speler in het veld te bepalen. Hierbij is er onderscheid gemaakt tussen vier posities: keepers, verdedigers, middenvelders en aanvallers. Bij het schatten van de modellen zijn de keepers de basiscategorie en daarom niet opgenomen in het model.

3.3. Prestatie statistieken

De prestatiestatistieken zijn van erg groot belang om de waarde van een speler te voorspellen omdat deze statistieken laten zien hoe goed of slecht een speler heeft gepresteerd in zijn voetbalcarrière. De prestatie statistieken omvatten variabelen die de speler kan beïnvloeden in het veld, zoals het aantal gescoorde doelpunten, het aantal gele kaarten en het aantal gegeven assists. Daarnaast zijn er ook variabelen die de activiteit van een speler weergeven zoals het aantal wedstrijden die een speler heeft gespeeld of het aantal minuten dat een speler heeft gespeeld. Van al deze variabelen hebben we de jaarlijkse prestaties. Ervaring zou ook een rol kunnen spelen om de marktwaardes van voetbalspelers te voorspellen en daarom hebben we een variabele gecreëerd die het totaal aantal gespeelde minuten weergeeft.

3.4. Achtergrond statistieken

Achtergrond statistieken zijn statistieken die te maken hebben met de club waar de speler onder contract staat of competities waarin een speler ooit gespeeld heeft. Zo zijn er competitie dummies gemaakt die aangeven of een speler ooit in een bepaalde competitie heeft gespeeld zoals het wereld kampioenschap voor voetbal of de Champions League en dummy variabelen die aangeven in welk van de 8 competities de speler nu actief is. Een andere achtergrondstatistiek is de variabele stadiongrootte, deze werd gemaakt met behulp van www.worldstadiumdatabase.com Een andere dummie werd toegevoegd voor spelers die vorig jaar kampioen van hun competitie werden. Tenslotte zijn er ook variabelen gemaakt die informatie geven met betrekking tot hun dienstverband bij de club zoals het aantal maanden dat een speler al bij een club zit of het aantal maanden totdat zijn contract verloopt.

Naast deze variabelen zijn er ook nog andere variabelen die invloed op de marktwaarde van voetbalspelers kunnen hebben. De blessuregevoeligheid zou bijvoorbeeld een rol kunnen spelen of het aantal keer dat een speler een tegenstander heeft gebeten.
In tabel 7 in de appendix vindt men een tabel met alle beschikbare variabelen, een korte beschrijving van deze variabelen en mijn verwachting of deze variabelen een significante bijdrage leveren aan de totstandkoming van de marktwaarden van de voetbalspelers.

4. Methodologie

In dit hoofdstuk zal eerst de onderzoeksopzet beschreven worden. Daarna worden de methodes beschreven die zijn gebruikt om de modellen te schatten en de methodes die zijn gebruikt om de modellen met elkaar te vergelijken.

4.1. Onderzoeksopzet

In het werkcollege hebben we onderzocht of we de marktwaarde van voetbalspelers konden voorspellen. Een plot van de marktwaardes van de voetbalspelers volgde een exponentieel verband en daarom hebben we gekozen om in ons werkcollege deze marktwaardes te transformeren door de logaritme van de marktwaarde te nemen. In Figuur 2 en 3 in de appendix vindt men de plot van de marktwaarden en een plot van de log marktwaarden van de voetbalspelers.
We hebben een model gemaakt dat een lineair verband weergeeft tussen de onafhankelijke variabelen en de log markwaarden. We zagen echter dat, voor de extreem lage en extreem hoge marktwaarden in de dataset, dit lineair model niet juist was gespecificeerd. Dit kan men zien aan de hand van een plot van de log markwaardes en een QQ-plot van de residuen. Dit probleem

hebben wij echter niet verholpen hetgeen ook te zien was aan bepaalde uitkomsten van het model.

Het idee van dit scriptieonderzoek is om dit probleem voor de lage en hoge waarnemingen te verhelpen. Dit wordt gedaan door de gehele dataset aan spelers te splitsen in drie categorieën namelijk spelers met een lage marktwaarde in de **lage categorie**, spelers met een hoge marktwaarde in de **hoge categorie**, en de spelers daartussen in de **midden categorie**.
Om te bepalen in welke categorie een voetbalspeler valt zijn er grenswaarden die bepalen of een speler in de lage, midden of hoge categorie valt. Deze grenswaarden werden in eerste instantie verkregen uit de QQ-plot van de residuen van het algemeen model van het werkcollege. Aan de QQ-plot van de residuen is te zien tot en vanaf welke waarnemingen het algemeen model niet juist was gespecificeerd. In Figuur 4 in de appendix vindt men de QQ-plot van de residuen van het algemeen model van het werkcollege. De spelers opslitsen in drie categorieën aan de hand van afwijkingen in de QQ-plot van de residuen gaf echter te weinig observaties voor de lage en de hoge categorie. Er is daarom gekozen om met een bepaald percentage te werken. Nadat ook bleek dat de laagste 10% en hoogste 10% van de marktwaarden onvoldoende waarnemingen gaven, is ervoor gekozen om de laagste 20% van de marktwaarden tot de lage categorie te beschouwen en de hoogste 20% van de marktwaarden tot de hoge categorie te beschouwen. Dit leverde de volgende grenswaardes op:

	Log marktwaarde
Lage categorie	3,22 - 5,99
Midden categorie	6,10 - 8,39
Hoge categorie	8,41 - 11,70

Tabel 1: *De grenswaarde die bepalen in welke categorie een speler zit*

Vervolgens zijn er vier modellen geschat met behulp van de Ordinary Least Squares methode. Voor ieder categorie één model die de marktwaarde voorspelt maar ook één algeheel model, een model dat de marktwaarde van alle voetbalspelers voorspelt. Dit algeheel model is gemaakt om de resultaten van de drie categorie modellen te kunnen vergelijken met het algeheel model. Voor elke categorie is er een random subsample gekozen van 75% van de data om de modellen te schatten. Bij het schatten van het algeheel model is er gebruik gemaakt van alle drie de random subsamples die gebruikt zijn bij het schatten van de drie categorie modellen, dit is gedaan omdat het algeheel model op deze manier met dezelfde data is geschat als de drie categorie modellen.De overige 25% van de data is gebruikt om de modellen te vergelijken door middel van out-of-sample voorspellingen.
Om één van de drie categorie modellen te gebruiken moet men natuurlijk weten in welke categorie een speler zit. Dit wordt gedaan met behulp van een ordered logit model. Het ordered logit model is geschat met dezelfde 75% data waarvan de marktwaarden en het type speler bekend zijn. In het ordered logit model krijgen de spelers in de lage categorie een 1, spelers in de midden categorie een 2 en spelers in de hoge categorie een 3.
Na het schatten van het ordered logit model is de 25% out-of-sample data gebruikt om te voorspellen in welke categorie een speler valt. Nadat iedere speler van de 25% out-of-sample data een voorspelde categorie heeft, wordt zijn marktwaarde voorspelt met behulp van het model dat bij zijn categorie past. Het algeheel model wordt ook gebruikt om de marktwaarde van de 25% out-of-sample data te voorspellen zodat deze voorspelling vergeleken kan worden met de

voorspelling van de categorie modellen.

Y_m	Y_c	X variabelen											
3.5	2												75% van de data
4.6	3												
4.2	3												
....	...												
1.98	1												
onbekend	onbekend												25% van de data

Figuur 1: *Onderzoeksopzet*

Samengevat ziet het scriptieonderzoek eruit zoals bovenstaand figuur. Er worden vier modellen geschat met behulp van Ym, de marktwaarden van de 75% in-sample data. Vervolgens schat ik een ordered logit model met de behulp van Yc, de categorie van de spelers van de 75% in-sample data. Vervolgens wordt voor de 25% out-of-sample data eerst hun categorie voorspelt met behulp van het ordered logit model.Aan de hand van de voorspelde categorieen wordt het bijbehorende categorie model gebruikt om de marktwaarde van de spelers te voorspellen.

4.2. Ordinary Least Squares

Om de relatie tussen de onafhankelijke variabelen en de log markwaarde te beschrijven, is gebruik gemaakt van de Ordinary Least Squares (OLS) methode. OLS is een methode voor het schatten van de onbekende parameters in een lineair regressie model. Er is vanuit gegaan dat de response variabele, de log markwaarde van een voetbalspeler, een lineaire vergelijking is van de regressors, de onafhankelijke variabelen:

$$Y = X\beta + \varepsilon. \tag{1}$$

In deze vergelijking is Y een nx1 vector van de afhankelijke variabele, de log marktwaarde van voetbalspelers in euro's. X is een vector nxk matrix met op iedere rij alle waardes van alle k variabelen van één bepaalde speler en ε is een nx1 vector van onobserveerbare storingstermen. n is het aantal waarnemingen in de dataset oftewel het aantal spelers.
Het belangrijkste idee achter deze methode is dat het de onbekende parameter β schat met b door een rechte lijn door de dataset te trekken zodanig dat de som van het kwadraat van de verticale afstanden, de residuen, geminimaliseerd wordt. Hieronder ziet men de de formule van de som van de gekwadrateerde residuen.

$$S(b) = \sum e_i^2 = e'e = (Y - Xb)'(Y - Xb) = Y'Y - Y'Xb - b'X'Y + b'X'Xb. \tag{2}$$

Om de OLS schatter,b, te bepalen, minimaliseren we de som van de gekwadrateerde residuen. Dit wordt gedaan door de afgeleide naar b gelijk te stellen aan nul. Dit geeft het volgende resultaat:

$$b = (X'X)^{-1}X'Y. \tag{3}$$

De zojuist beschreven OLS methode is gebruikt bij het schatten van de 3 categorie modellen en het algeheel model. Bij het construeren van de modellen hebben we gebruik gemaakt van de forward selection methode Dit houdt in dat ik begonnen ben met leeg model zonder variabelen en stap voor stap variabelen toevoegen aan het model. Een variabele wordt in het model opgenomen als deze significant blijkt te zijn. Een variabele is significant als blijkt dat deze variabele significant verschilt van nul. In dit onderzoek is gekozen voor een significantieniveau van 5%. Bij het maken

van de modellen zijn alleen variabelen bekeken van 6 jaar geleden tot en met het heden omdat variabele voor dit tijdperk irrelevant worden geacht. De gecreëerde parameters geven een relatie tussen de verklarende variabelen en de logaritmes van de marktwaarden.
Nadat de modellen zijn geschat, wordt er onderzocht of de 3 belangrijkste aannames van de OLS methode gelden. Het gaat om de volgende 3 aannames:

- Homoskedasticiteit
- Lineair model
- Residuen zijn normaal verdeeld met gemiddelde nul

Indien deze aannames niet gelden is het belangrijk om hiermee rekening te houden bij het trekken van conclusies omdat de resultaten onnauwkeurig kunnen zijn. De aannames worden hieronder beschreven, samen met de gevolgen als de aanname niet geldt.

4.2.1 De 3 aannames

Homoskedasticiteit
Er is sprake van homoskedasticiteit als de residuen een constante variantie hebben over alle observaties.

$$E(\varepsilon_i^2) = \sigma^2. \tag{4}$$

Door naar de spreidingsdiagram van de residuen tegen de gefitte waarde te kijken is te zien of er sprake is van homoskedasticiteit. Als de residuen van het model binnen 2 denkbeeldige bandbreedtes vallen, is er sprake van homoskedasticiteit. De statistische toets om na te gaan of er sprake is van homoskedasticiteit is de Breusch-Pagan toets. In het geval dat deze aanname niet geldt, heteroskedasticiteit, zijn de standaard fouten van de OLS incorrect en is de OLS methode niet langer efficiënt. Indien dit het geval is worden er, in plaats van de normale standaardfouten, de White standaardfouten gebruikt. Deze White standaardfouten corrigeren voor heteroskedasticiteit.

Lineair model
Om de OLS techniek toe te passen moet er een lineaire relatie zijn tussen de te verklaren variabele en de verklarende variabelen. Als de aanname niet geldt en er zijn weinig verklarende variabelen zal de OLS schatting onzuiver zijn en als de aanname niet geldt wanneer er veel verklarende variabelen zijn zal de OLS schatting niet efficiënt zijn. Een spreidingsdiagram van de residuen tegen de gefitte waarden laat zien of er een lineaire relatie bestaan. Als de residuen binnen 2 bandbreedtes vallen, is er sprake van een lineair verband. Indien dit niet het geval is, zou een andere techniek gebruikt moeten worden om de log marktwaarde van de voetbalspelers te voorspellen.

Residuen zijn normaal verdeeld met gemiddelde nul
Deze aanname impliceert dat de residuen een normale verdeling volgen en een gemiddelde nul hebben.Samen met de aanname van homoskedasticiteit zijn deze aannames te formuleren als:

$$\varepsilon \sim N(0, \sigma^2). \tag{5}$$

Indien dit niet het geval is zal de OLS techniek niet langer efficiënt zijn. Een histogram van de residuen laat zien of de residuen normaal verdeeld zijn. Twee belangrijke indicatoren die aangeven of de residuen normaal verdeeld zijn, zijn de skewness en de kurtosis van de residuen. Deze zouden respectievelijk nul en drie moeten zijn in het geval van normaal verdeelde residuen. De statistische toets om na te gaan of er sprake is van normaal verdeelde residuen is de Jarque-Bera toets.

4.3. Ordered logit model

Het ordered logit model wordt gebruikt wanneer een afhankelijke variabele een gelimiteerd aantal waardes kan aannemen en deze waardes zijn geordend. De afhankelijke variable, log marktwaarde, is opgedeeld in 3 categorieen, laag, midden en hoog en daarom wordt er een ordered logit model gebruikt. Het ordered logit model schat een latente variabele, Y^*, middels een lineair regressie model. Om deze latente variable Y^* vervolgens te koppelen aan één van de mogelijke geordende waarden gebruikt het model grenswaarden die de voorspelde Y^* indelen in één van de geordende categorieën.

Het model schat latente variabele Y^* middels een linear regressie model:

$$Y_i^* = X_i'\beta + \varepsilon_i. \tag{6}$$

Om deze latente variabelen te koppelen aan de geordende categorieën gebruikt het model grenswaarden:

$$Y_i = 1 \text{ if } -\infty < Y_i^* \leq \alpha_1$$
$$Y_i = 2 \text{ if } \alpha_1 < Y_i^* \leq \alpha_2$$
$$Y_i = 3 \text{ if } \alpha_2 < Y_i^* \leq \infty$$

Om ervoor te zorgen dat de uitkomste in Y_i geordend zijn geldt het volgende:

$$\alpha_1 < \alpha_2$$

4.4. Voorspellingen

Nadat de modellen zijn geschat worden er voorspellingen gedaan om zo inzicht te krijgen in welke methode beter is om de marktwaarde van voetbalspelers te voorspellen. De in-sample fit van de 3 categorie modellen en de in-sample fit van het algemeen model worden met elkaar vergeleken. Dit wordt gedaan om te onderzoeken of het opslitsen van de spelers in 3 categorien tot betere voorspelingen leidt zonder rekening te hoeven houden met het ordered logit model. Normaal gesproken wordt de R^2 van modellen met elkaar vergeleken om na te gaan welk model beter is. De R^2 geeft aan hoeveel procent van de totale variantie van de afhankelijke variabele wordt verklaard door het model. In dit geval kan dat niet omdat de samples van de drie categorie modellen verschillend zijn vergeleken met de sample van het algheel model.

Om de in-sample fit van de drie categoriemodellen te vergelijken met de in-sample fit van het algeheel model wordt er een vector gemaakt met de gefitte waarde van alle drie categorie modellen tesamen. Het aantal waarnemingen in deze vector is precies gelijk aan het aantal waarnemingen van het algeheel model. Vervolgens wordt de correlatie tussen de gefitte waarden van de drie modellen en de echte waarden vergeleken met de correlatie tussen de gefitte waarden van het algeheel model en de echte waarden. Hetzelfde wordt gedaan met de gekwadrateerde correlatie, de gekwadrateerde correlatie van de gefitte waarden van de drie modellen en de echte waarde wordt vergeleken met de gekwadrateerde correlatie van de gefitte waarde van het algeel model en de echte waarden,dit is de R^2 van het algeheel model. De methode met de beste voorspellingen geeft een hogere (gekwadrateerde) correlatie.

De out-of-sample voorspellingen van de 3 categorie modellen en de out-of-sample voorspellingen van het algemeen model worden ook met elkaar vergeleken. Dit wordt gedaan om de invloed van het ordered logit model te onderzoeken. Het vergelijken van de voorspellingen wordt op dezelfde wijze gedaan als voor de in-sample fit.

Om de individuele voorspellingen van de categorie modellen te onderzoeken wordt als eerst de

categorieën van de 25% out-of-sample data bepaalt. Vervolgens wordt voor elke categorie de marktwaarde van de spelers op 2 manieren geschat:
1) door het bijbehorende categorie model
2) door het algemeen model
Deze twee voorspellingen worden met elkaar vergelijken door de twee Root Mean Squared Errors(RMSE) van deze voorspellingen met elkaar te vergelijken. DE RMSE is een maatstaf voor het verschil tussen de, door het model, voorspelde waarden en de echte waarden. Een lagere RMSE is het model met betere voorspellingen.

5. Resultaten

In dit hoofdstuk worden de resultaten besproken van het onderzoek. In totaal zijn er in vijf modellen gemaakt in dit onderzoek: één ordererd logit model om te bepalen in welke categorie een speler valt, drie categorie modellen voor de drie verschillende categorieën om de marktwaarden te bepalen en één model om de marktwaarden van alle spelers te bepalen. Dit model voor alle spelers is gemaakt zodat de resultaten van de 3 aparte modellen te vergelijken zijn met één algemeen model.

5.1. Ordered logit model

Zoals beschreven is het ordered logit model is geschat met behulp van een random sup-sample van 75% van de data van alle spelers. Nadat de latente variabele middels een lineair regressie model is geschat, konden de spelers in categorieën worden ingedeeld met behulp van de gegeven limietpunten.

$$\begin{aligned} Y_i^* = {} & \widehat{\beta_1} age + \widehat{\beta_2} age^2 + \widehat{\beta_3} assists\_2013 + \widehat{\beta_4} bundesliga\_oos + \widehat{\beta_5} CL + \widehat{\beta_6} EL \\ & + \widehat{\beta_7} eredivisie + \widehat{\beta_8} issamer + \widehat{\beta_9} jupiler + + \widehat{\beta_{10}} matches\_2010 + \widehat{\beta_{11}} matches\_2011 \\ & + \widehat{\beta_{12}} matches\_2012 + \widehat{\beta_{13}} matches\_2013 + \widehat{\beta_{14}} ligue1 + \widehat{\beta_{15}} sommin + \widehat{\beta_{16}} stadiongrootte \\ & + \widehat{\beta_{17}} waschamply + \widehat{\beta_{18}} assists\_2012 + \widehat{\beta_{19}} bundesliga\_duit + \widehat{\beta_{20}} isstriker \end{aligned} \tag{7}$$

$$\begin{gathered} Y_i = 1 \text{ if } -\infty < Y_i^* \leq 22,08 \\ Y_i = 2 \text{ if } 22,08 < Y_i^* \leq 28,26 \\ Y_i = 3 \text{ if } 28,26 < Y_i^* \leq \infty \end{gathered}$$

In Figuur 9 in de appendix kunt u de coefficienten van de geschatte β vinden

Interpretatie van de variabelen
Age en age^2 zijn beide significant zijn. Het verschil is dat age een positieve coëfficiënt heeft en age^2 een negatieve coëfficiënt. Dit houdt dus in dat naarmate een speler ouder wordt de kans groter is dat hij in een hogere categorie valt maar dat dit alleen geldt tot een bepaalde leeftijd. Het negatieve verband van age^2 zorgt er voor dat, vanaf een bepaalde leeftijd, de kans op een lagere categorie groter wordt naarmate de speler ouder wordt. Dit is logisch omdat spelers na een bepaalde leeftijd over hun top zijn en dus slechter gaan presteren.
Niet alle competitie dummies zitten in het ordered logit model . De competities Serie A en Primera Division zitten niet in het model. Dit houdt in dat de kans voor een bepaalde categorie gelijk is voor de Premier League, Serie A en Primera Division.(Premier League is de basis competitie die altijd buiten het model wordt gelaten) De overige competities die wel in het model zitten hebben een negatieve coëfficiënt. Dit houdt in dat wanneer men in 1 van deze competities speelt de kans

groter is dat de speler in een lagere categorie valt dan wanneer de speler in 1 van de 3 zojuist vermelde competities zit die niet in het model zitten.
Daarnaast geeft het spelen in één van de twee Europese competities een grotere kans om in een hogere categorie te komen. Dit is te verklaren omdat de CL de hoogste Europese competitie is en de EL de een na hoogste Europese competitie. De beste spelers zullen dus spelen in de CL en EL.
Het aantal gespeelde wedstrijden is ook van invloed om te bepalen in welke categorie een speler zit en deze variabelen hebben een positieve coëfficiënt.Dit betekent hoe meer wedstrijden je hebt gespeeld, hoe groter de kans is om in een hogere categorie terecht te komen. Dit is logisch omdat de betere spelers meer wedstrijden spelen dan minder goede spelers.
Het aantal gegeven assists van de afgelopen 2 jaar en de variabele die aangeeft of een speler een spits is, zijn ook significant. Dit geeft dus aan dat wanneer een speler een spits is de kans groter is dat hij in een hogere categorie zit dan een speler die een keeper is. Daarnaast geven verdedigers over het algemeen minder assists dan middenvelders en spitsen dus zou je ook kunnen zeggen dat wanneer je een verdediger bent de kans groter is dat je in een lagere categorie valt.
Stadiongroote is ook van significante invloed en heeft een positieve coefficient dus hoe groter het stadion, hoe groter de kans is dat een speler in een hoge categorie terecht komt. Dit is te verklaren doordat over het algemeen de beste clubs de grootste stadions hebben en dat deze beste clubs ook de beste spelers hebben.
Daarnaast zijn ook de ervaringsvariabele sommin en de dummie variabele die aangeeft of een speler vorig seizoen kampioen is geworden positief significant.

De Likelihoof Ratio is een toets die toetst of de opgenomen variabelen in het model significant zijn.De nul hypothese van de LR toets is: De variabelen zijn niet significant. De Likelihood ratio statistic is 2854. Deze statistic is chi kwadraat verdeeld onder de nullhypothese De kans van op de nullhypothese is 0,00 dus we verwerpen de nullhypothese:
De variabelen in het model zijn wel significant.

5.2. de 4 OLS modellen

Het maken van de 4 OLS modellen is gedaan met 75% van de data van de lage spelers voor het lage model, 75% van de data van de midden spelers voor het midden model, 75% van de data van de hoge spelers voor het hoge model en de drie subsamples van de 3 categorie modellen voor het algehele model. Dit resulteerde in het volgende:

Model voor de lage spelers

$$\begin{aligned}\log(y_i) = {} & \widehat{\beta_1} + \widehat{\beta_2} monts\_to\_go + \widehat{\beta_3} age + \widehat{\beta_4} age^2 + \widehat{\beta_5} matches\_2012 + \widehat{\beta_6} matches\_2013 \\ & + \widehat{\beta_7} minutes\_2012 + \widehat{\beta_8} minutes\_2013 + \widehat{\beta_9} stadiongrootte + \widehat{\beta_10} sommin + \widehat{\beta_{11}} waschamply \\ & + \widehat{\beta_{12}} EL + \widehat{\beta_{13}} bundesliga\_duit + \widehat{\beta_{14}} eredivisie + \widehat{\beta_{15}} bundesliga\_oos + \widehat{\beta_{16}} serie\_A\end{aligned} \tag{8}$$

Model voor de midden spelers

$$\begin{aligned}\log(y_i) = {} & \widehat{\beta_1} + \widehat{\beta_2} monts\_at\_club + \widehat{\beta_3} monts\_to\_go + \widehat{\beta_4} age + \widehat{\beta_5} age^2 + \widehat{\beta_6} bundesliga\_oos + \widehat{\beta_7} CL \\ & + \widehat{\beta_8} EL + \widehat{\beta_9} eredivisie + \widehat{\beta_10} goals\_2013 + \widehat{\beta_{11}} stadiongrootte + \widehat{\beta_{12}} jupiler \\ & + \widehat{\beta_{13}} ligue1 + \widehat{\beta_{14}} matches\_2010 + \widehat{\beta_{15}} matches\_2011 + \widehat{\beta_{16}} matches\_2012 + \widehat{\beta_{17}} matches\_2013 \\ & + \widehat{\beta_{18}} bundesliga\_duit + \widehat{\beta_{19}} subon\_2013 + \widehat{\beta_{20}} waschamply + \widehat{\beta_{21}} WK + \widehat{\beta_{22}} serie\_A + \widehat{\beta_{23}} isdef \\ & + \widehat{\beta_{24}} international\_competition + \widehat{\beta_{25}} issamer \end{aligned} \tag{9}$$

Model voor de hoge spelers

$$\begin{aligned}\log(y_i) = {} & \widehat{\beta_1} + \widehat{\beta_2} monts\_to\_go + \widehat{\beta_3} bundesliga\_oos + \widehat{\beta_4} goals\_2012 + \widehat{\beta_5} goals\_2013 + \widehat{\beta_6} CL \\ & + \widehat{\beta_7} stadiongrootte + \widehat{\beta_8} assists\_2013 + \widehat{\beta_9} bundesliga\_duit + \widehat{\beta_{10}} eredivisie + \widehat{\beta_{11}} issamer \\ & + \widehat{\beta_{12}} jupiler + \widehat{\beta_{13}} matches\_2011 + \widehat{\beta_{14}} matches\_2012 + \widehat{\beta_{15}} matches\_2013 + \widehat{\beta_{16}} minutes\_2013 \\ & + \widehat{\beta_{17}} sommin + \widehat{\beta_{18}} waschamply + \widehat{\beta_{19}} age + \widehat{\beta_{20}} age^2 + \widehat{\beta_{21}} primera\_division + \widehat{\beta_{22}} ligue1 \end{aligned} \tag{10}$$

Het algeheel model

$$\begin{aligned}\log(y_i) = {} & \widehat{\beta_1} + \widehat{\beta_2} monts\_at\_club + \widehat{\beta_3} monts\_to\_go + \widehat{\beta_4} age + \widehat{\beta_5} age^2 + \widehat{\beta_6} bundesliga\_oos \\ & + \widehat{\beta_7} eredivisie + \widehat{\beta_8} EL + \widehat{\beta_9} EK + \widehat{\beta_{10}} in\_last\_year + \widehat{\beta_{11}} international\_competition \\ & + \widehat{\beta_{12}} waschamply + \widehat{\beta_{13}} stadiongrootte + \widehat{\beta_{14}} jupiler + \widehat{\beta_{15}} ligue1 + \widehat{\beta_{16}} matches\_2010 \\ & + \widehat{\beta_{17}} matches\_2011 + \widehat{\beta_{18}} matches\_2012 + \widehat{\beta_{19}} matches\_2013 + \widehat{\beta_{20}} assists\_2013 \\ & + \widehat{\beta_{21}} bundesliga\_duit + \widehat{\beta_{22}} CL + \widehat{\beta_{23}} goals\_2013 + \widehat{\beta_{24}} issamer + \widehat{\beta_{25}} primera\_division \\ & + \widehat{\beta_{26}} serie\_A + \widehat{\beta_{27}} sommin \end{aligned} \tag{11}$$

In Figuur 5,6,7 en 8 in de appendix kunt u de coefficienten van de geschatte β vinden van elk model

5.2.1 Overeenkomsten en verschillen in de 3 categorie modellen

Als we de variabelen in de 3 categorie modellen met elkaar vergelijken zien we opvallende overeenkomsten en verschillen:
Net als in het ordered logit model zijn age en age^2 beide significant signifcant. De positieve coëfficiënt van age en de negatieve coëfficiënt van age^2 maken het mogelijk dat de marktwaarde van een speler stijgt tot een bepaalde leeftijd en na die leeftijd daalt.
De variabelen CL en EL zijn, mits significant, altijd positief. Dit is te verklaren omdat de CL de hoogste Europese competitie is en de EL de een na hoogste Europese competitie. De beste spelers zullen dus spelen in de CL en EL.
Het aantal maanden tot het contract afloopt is voor alle drie de modellen positief significant. Een verklaring zou kunnen zijn dat de sterspelers vaak contracten krijgen aangeboden met een lange looptijd.
Voor het de lage categorie model zijn de Duitse Bundesliga, Oostenrijkse Bundesliga , Eredivisie en de Serie A allemaal negatief significant. Dit betekend dat de marktwaarde van spelers in de

lage categorie en spelend in één van deze competities lager ligt dan de marktwaarde van spelers in de lage competitie die spelen in de Premier League, de basis categorie. Op dezelfde manier zijn de significante competities van de andere modellen te interpreteren. Voor het hoge categorie model zien we dat assists2013 positief significant is terwijl deze variabele niet significant is in de andere twee categorie modellen. Dit betekent dat het aantal assists in 2013 alleen invloed heeft op de marktwaarde van spelers in de hoge categorie. Op dezelfde manier zijn de variabelen te interpreteren die in maar één van de categorie modellen voorkomt zoals Subon2013,WK, Het aantal gespeelde wedstrijden in voorgaande jaren is ook significant en in de meeste gevallen zien we een patroon in deze variabelen: de recentere jaren hebben een grotere coëfficiënt dus die hebben meer invloed op de marktwaarde van de spelers dan het aantal gespeelde wedstrijden van een aantal jaar geleden. In het hoge categorie model is dit echter niet het geval: matches2013 heeft een kleinere coëfficiënt dan matches2012. Een verklaring hiervoor zou kunnen zijn dat minutes2013 ook in dit model zit en die gecorreleerd met matches2013 is.

5.2.2 Toetsen van de 3 aannames

Homoskedasticiteit
Zoals eerder is beschreven wordt er getoetst of de residuen van de 4 modellen een constante variantie hebben over alle observaties middels een Breasch Pagan toets. De nullhypothese in de Breusch Pagan toets is : Er is sprake van homoskedasticiteit De toetsgrootheid is chi kwadraat verdeeld onder de nullhypothese In de figuur 10 in de appendix zet men voor alle 4 de modellen de toetsgrootheid en de kans op de nullhypothese. Iedere model heeft heteroskedasticiteit behalve het model voor de hoge spelers. Voor de modellen met heteroskedasticiteit gebruiken we in plaats van de normale standardfouten, de White standard fouten. De parameterschattingen blijven echter wel gelijk.

Lineair model
Zoals eerder is beschreven worden de spreidingsdiagrammen van de residuen tegen de gefitte waarden van de modellen geinterpreteerd om na te gaan of er een linear verband zichtbaar is.
Allereerst lijken de drie categorie modellen een patroon te hebben door de diagonale lijnen. Dit wordt echter veroorzaakt doordat veel spelers één dezelfde marktwaarde hebben. Daarnaast lijken de 3 afzonderlijke modellen niet binnen 2 bandbreedtes te vallen, ze lijken eerder op een ruit maar dit komt door de gekozen cut off punten. Er is daarom gekozen om de residuen van de drie aparte modellen in één spreidingsdiagram weer te geven. Als we naar die spreidingsdiagram kijken zien we dat de residuen binnen 2 bandbreedes blijven en daarom gaan we ervan uit dat er een lineair verband is.

Residuen zijn normaal verdeeld met gemiddelde nul
Om te bekijken of de residuen normaal zijn verdeeld en een gemiddelde van nul hebben bekijken we de histogram van de residuen en vergelijken deze met de normale verdeling. Dit houdt in dat idealiter de kurtosis 3 is en de skewness 0. De beste manier om dit te onderzoeken is met behulp van de Jarque Bera toets.
De nullhypothese in de Jarque Bera toets is : De residuen zijn normaal verdeeld De toetsgrootheid is chi kwadraat verdeeld onder de nullhypothese. In de tabel hieronder ziet u voor ieder model het gemiddelde van de residuen, de JB statistic en de kans op normaal verdeelde residuen

De nullhyopthese wordt alleen in het algeheel model verworpen. Het opsplitsen van het hele model heeft er dus voor gezorgd dat de residuen nu wel normaal verdeeld zijn in de drie categorie modellen. Dit was niet het geval voor het model voor alle spelers

	Gemiddelde	JB statistic	P-waarde
Lage model	0,00	4,76	0,09
Midden model	0,00	2,19	0,33
Hoge model	0,00	1,92	0,38
Algeheel model	0,00	111,80	0,0

Tabel 2: *Tabel met de gemiddelde van de residuen, JB Statistic en P-waarde voor ieder model*

In Figuur 13, 14, 15,en 16 in de appendix kunt men de histogrammen van de residuen van alle vier de modellen vinden.

5.3. In-sample fit

Nadat de 4 modellen zijn geschat worden ze met elkaar vergeleken. Eerst wordt de in sample fit van de modellen vergeleken. Er wordt nog geen rekening gehouden met het ordered logit model omdat de categorieën bekend zijn voor de in-sample data dus kan er worden gezien hoe de drie categorie modellen werken ten opzichte van het algeheel model Om de resultaten van de 3 aparte modellen te vergelijken met het algemeen model worden de gefitte waarden van de 2 methodes vergeleken zoals eerder beschreven.
In de tabel hieronder is de correlatiematrix van de gefitte waarden van de 3 specifieke modellen ,de gefitte waarden van het algemeen model en de echte waarden weergegeven:

Correlatie	3modellen	algeheel	Echte waarden
3modellen	1	NVT	0,953
algeheel	NVT	1	0,895
Echtewaarden	0,953	0,895	1

Tabel 3: *De correlatiematrix van de gefitte waarden van de 3 specifieke modellen ,de gefitte waarden van het algemeen model en de echte waarden*

De correlatie tussen de echte waarden en de gefitte waarden van de 3 specifieke modellen is groter is dan de correlatie tussen de echte waarden en de gefitte waarden van het algemeen model. Vervolgens worden ook de gekwadrateerde correlaties met elkaar vergeleken:
De gekwadrateerde correlatie van het algemeen model is **0.800221**(Dit is de R-kwadraat van het algeheel model)
De gekwadrateerde correlatie van de 3 modellen is **0.909855**

Uit deze resultten blijkt dat de marktwaarde van de spelers beter te verklaren is door drie aparte modellen te maken voor elke categorie dan wanneer de marktwaarde van de spelers verklaard wordt met een algeheel model.
Ook als we de 2 spreidingsdiagrammen van de gefitte waarden van de drie modellen tegen de echte waarden en de gefitte waarde van het algeheel model tegen de echte waarden analyseren volgt dit resultaat Er is te zien dat de puntenwolk van de 3 modellen dichterbij de lijn, die de perfecte voorspellingen weergeeft, ligt dan de puntenwolk van het algeheel model. In Figuur 17 en 18 in de appendix kunt men de spreidingsdiagrammen vinden.

5.4. Out-of-sample voorspellingen

Na de in-sample fit vergelijkingen worden de out of sample voorspellingen met elkaar vergeleken. Hierbij is van te voren nog niet duidelijk welke categorie de spelers hebben. Het ordered logit model wordt eerst gebruikt om de 25% out-of-sample data eerst een categorie toe te kennen.

In tabel 4 is de hitrate tabel van het ordered logit model weergegeven: Over het algemeen

		Voorspelling			
		Laag	Midden	Hoog	
Waargenomen	Laag	0,10	0,12	0,00	0,22
	Midden	0,01	0,43	0,13	0,57
	Hoog	0,00	0,03	0,18	0,21
		0,11	0,58	0,31	1
		Hitrate = 0,71			

Tabel 4: *Hitrate tabel van het Ordered logit model*

heeft het model een hit rate van 71%. Dit houdt in dat het model in 71% van de gevallen de juiste categorie voorspelt. Daarnaast zijn er een paar dingen die opvallen in deze resultaten: Het model maakt vaker de foute voorspelling dat een speler in een hogere categorie zit dan zijn daadwerkelijke categorie dan de foute voorspelling dat een speler in een lagere categorie zit:
Wanneer de echte categorie van de spelers laag is, voorspelt het ordered logit model in 54% van de gevallen de midden categorie.
Wanneer de echte categorie van een speler midden is voorspelt het ordered logit model in 22% van de gevallen een hoge categorie en in minder dan 2% van de gevallen een lage categorie.
Wanneer de echte categorie van een speler hoog is voorspelt het ordered logit model in 17% van de gevallen een midden categorie.

Voor de individuele categorieën maakt het ordered logit model de beste voorspellingen voor de lage categorie:
Wanneer het model een lage categorie voorspelt is dit in 91% van de gevallen juist.
Wanneer het model een midden categorie voorspelt is dit in 74% van de gevallen juist
Wanneer het model een hoge categorie voorspelt is dit in 58% van de gevallen juist

Nadat iedere speler met behulp van het ordered logit model een categorie heeft gekregen, wordt voor iedere categorie de marktwaarden voorspelt met behulp van het bijbehorende model. De marktwaarden van de 25% out-of-sample data wordt vervolgens ook voorspelt door het algehele model zodat we de voorspellingen met elkaar kunnen vergelijken. Deze voorspellingen vergelijk ik op dezelfde manier als de in-sample voorspellingen. In de tabel hieronder is de correlatiematrix van de gefitte waarden van de 3 specifieke modellen ,de gefitte waarden van het algemeen model en de echte waarden weergegeven:

De correlatie tussen de echte waarden en de gefitte waarden van de 3 categorie modellen is in dit geval lager dan de correlatie van de echte waarden en de gefitte waarden van het algemeen model.

Vervolgens worden ook de gekwadrateerde correlaties met elkaar vergeleken:

Correlatie	3modellen	algeheel	Echte waarden
3modellen	1	NVT	0,705
algeheel	NVT	1	0,869
Echtewaarden	0,705	0,869	1

Tabel 5: *De correlatiematrix van de gefitte waarden van de 3 specifieke modellen ,de gefitte waarden van het algemeen model en de echte waarden*

De gekwadrateerde correlatie van het algemeen model is **0.755**
De gekwadrateerde correlatie van de drie modellen is **0.50**

Uit deze resultten blijkt dat het algeheel model betere voorspellingen geeft dan een model dat onderscheid maakt tussen drie categorieën waarvan de categorie bepaling wordt gedaan middels een ordered logit model.

Op het eerste gezicht lijkt dit resultaat in strijd met de eerdere bevindingen waar bleek dat het model opsplitsen in drie categorieën betere voorspellingen gaf. Er worden nu een lagere (gekwadrateerde) correlatie gevonden omdat het ordered logit model eerst moet schatten in wat voor categorie een speler valt. Als dit niet juist wordt voorspelt komt zo een speler in het verkeerde model wat zorgt voor grotere fouten en dus een kleinere correlatie. De hitrate van het ordered logit model was 71% en geeft dus niet altijd de juiste voorspelling.

5.5. Voorspellingen van de individuele modellen

De drie modellen hebben een hogere (gekwadrateerde) correlatie dan het algemeen model als duidelijk is in wat voor categorie een speler valt maar de 3 modellen hebben een lagere (gekwadrateerde) correlatie dan het algemeen model als de categorie van de speler onbekend is en deze eerst voorspelt wordt door het ordered logit model.
We zagen al eerder dat de hitrate voor de verschillende categorieën sterk verschilt. Nu wordt onderzocht of het ordered logit model inzichten kan geven in welk model het best gebruikt kan worden als we de categorie van een speler hebben voorspelt. Dit wordt gedaan door met het ordered logit model de categorie van een speler te bepalen. Vervolgens wordt de marktwaarde van deze categorie spelers bepaalt door:
1) de marktwaarde te laten voorspellen door het categorie model dat bij deze categorie hoort
2) de marktwaarde te laten voorspellen door het algemeen model
Deze voorspellingen worden met elkaar vergeleken zoals beschreven in de Methodologie. In de tabel hieronder zien de resultaten. De indices in de eerste kolom geven aan wat volgens het ordered logit model de categorie van de speler is. Vervolgens voorspellen we de marktwaarde van deze categorie met zijn bijbehorend model en ook met het algeheel model. De voorspelling die de laagste Root Mean Squared Error heeft, is de beste voorspelling.

De root mean squared error is alleen lager voor het lage model. Dit was te verwachten omdat het ordered logit model voor de lage categorie de spelers 91% van de gevallen goed voorspelt. Voor de andere categorien was dit een stuk lager.
Aan de hand van deze resultaten kunnen de voorspellingen van het ordered logit model inzicht geven in het model dat gekozen moet worden om de beste voorspellingen te krijgen:
Indien het ordered logit model een lage categorie voorspelt, worden de beste voorspelingen verkregen door het lage categorie model.

	Bijbehorend model RMSE	Algeheel model RMSE
Laag	0,728	0,746
Midden	0,824	0,690
Hoog	0,656	0,611

Tabel 6: *De RMSE van de voorspelde marktwaarden van de 3 voorspelde categorieën van het bijbehorende model en het algeheel model*

Indien het ordered logit model een midden of hoge categorie voorspelt, worden de beste voorspelingen verkregen door het algeheel model.

6. Conclusie

In dit onderzoek zijn drie modellen geschat om de marktwaarde van voetbalspelers te bepalen. Ieder model bepaalt de marktwaarde van een bepaalde categorie waar de spelers in vallen, de zogenoemde lage,midden en hoge categorie. Deze modellen zijn gemaakt omdat tijdens het werkcollege bleek dat de marktwaarde van de voetbalspelers niet te omvatten is in één model. Er is geprobeerd om met deze drie categorie modellen het algeheel model van het werkcollege te verbeteren. De hoofdvraag van het onderzoek luidde alsvolgt:
Kan de marktwaarden van voetbalspelers beter bepaald worden door modellen die onderscheid maken in verschillende categorieén van spelers dan één algeheel model dat geen onderscheid maakt ?
Er is geen eenduidig antwoord op deze vraag. De in-sample voorspellingen van de drie categorie modellen geven betere resultaten dan de in-sample voorspellingen van het algeheel model. Je zou dus kunnen zeggen dat de drie categorie modellen beter werken. In de praktijk weten we natuurlijk niet in welk categorie een speler valt en om daarachter te komen is er een ordered logit model geschat die voor iedere speler zijn categorie voorspelt. Dit ordered logit model heeft een hitrate van 71% en voorspelt dus in 29% van de gevallen een verkeerde categorie voor een speler. Als we de out-of-sample voorspellingen met elkaar vergelijken, waarbij we dus eerst het ordered logit model gebruiken om de categorie van een speler te bepalen, zien we dat de voorspellingen door het algeheel model beter zijn dan de voorspellingen van de 3 categorie modellen.
Bij het beantwoorden van de hoofdvraag zijn we erachter gekomen dat het ordered logit model ervoor zorgt dat we voor de out-of-sample data geen betere voorspellingen krijgen. De conclusie die uit dit onderzoek getrokken kan worden is dat de voorspellingen van de marktwaarde van voetbalspelers beter zijn als er drie categorie modellen worden gebruikt mits de voorspelde categorieën kloppen.

De eerste deelvraag is:
Zijn er verschillende verklarende variabelen voor de drie verschillende modellen en is er een significant verschil in partieel effect tussen variabelen die overeen komen in de modellen?
Het antwoord op deze vraag is dat er verschillende verklarende variabelen zijn voor de modellen. De variabelen zijn in het vorig hoofdstuk besproken.

In de hitrate tabel van het ordered logit model is te zien dat de hitrate per categorie verschilt en daarom is er ook onderzocht of we bepaalde voorspellingen van het ordered logit model in ons

voordeel kunnen gebruiken. De deelvraag die hierbij onstond was:
Kan het voorspellen van de type speler ons inzicht geven in welke model we moeten gebruiken?
Aan de hand van voorspellingen voor de individuele categorieën door het desbetreffende categorie model en het algeheel model is dit onderzocht. We zagen dat de hitrate voor de lage categorie relatief hoog was vergeleken met de andere modellen en dat was ook te zien aan de resultaten. Alleen wanneer het ordered logit model een lage type speler voorspelt werkt het categorie model beter dan het algeheel model. Als het ordered logit model een midden of hoge categorie voor een speler voorspelt zijn de voorspellingen van het algeheel model beter dan de categorie modellen voor deze spelers.

Met dit resultaat blijkt dat het ordered logit model in het voordeel gebruikt kan worden:
Indien het ordered logit model een lage categorie voorspelt, worden de beste voorspelingen verkregen door het lage categorie model.
Indien het ordered logit model een midden of hoge categorie voorspelt, worden de beste voorspelingen verkregen door het algeheel model.

Als voetbalfan heb ik uiteraard met veel plezier aan dit scriptieonderzoek gewerkt. Nadat de euforie had plaatsgemaakt voor het harde werken merkte ik dat er veel tijd en energie in het onderzoek gestoken zou moeten worden om dit scriptie onderzoek met een voldaan gevoel af te sluiten vanwege alle tegenslagen gedurende het onderzoek. Ik wil graag mijn supervisor Andreas Alfons bedanken. In tijden waarin ik het licht aan het eind van de tunnel niet meer kon zien door de tegenvallende resultaten, was hij er om mij te motiveren. Daarnaast wil ik ook mijn studiegenoten Annemarijn Mutsaerts en Sander Monster bedanken voor hun inzet tijdens het werkcollege en over de ideeen die daarbij zijn onstaan die ik tijdens dit onderzoek heb kunnen gebruiken.

Referenties

[1] The Telegraph,2013] Top 20 most expensive transfer fees of all time

[2.] Independent,2013] Roman Abramovichs 10-year Chelsea anniversary: What did he ever do for us? (2009).

[3.] Bleacher Report,2014] How Much Is Barcelonas Lionel Messi Worth Based on Form in 2014?

[4.] F Carmichael and D Thomas, 1993] Bargaining in the transfer market: Theory and Evidence Applied Economics V25

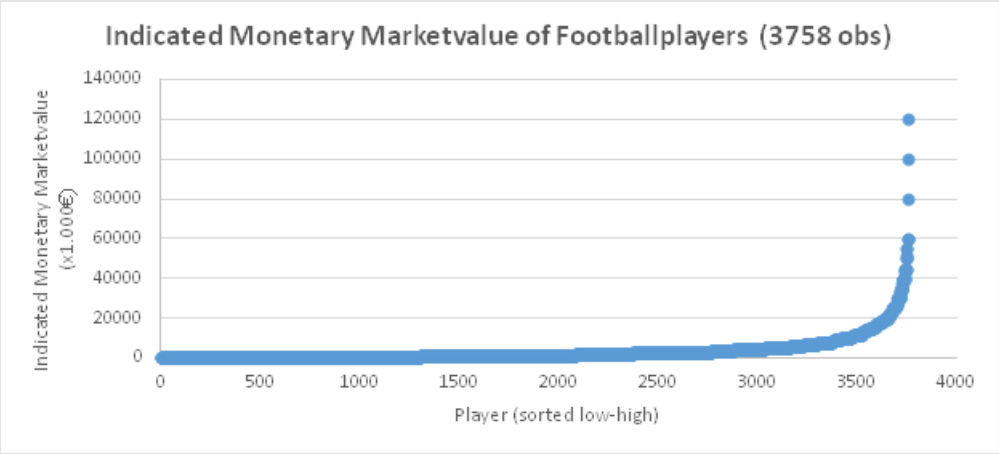
[5.] B Reilly and R Witt, 1995] English league transfer prices: Is there a radical dimension? Applied economics letters V2

[6.] C Heij, P de Boer, P H Franses, T Kloek and H K van Dijk , 2004] Econometric Methods with Applications in Business and Economics

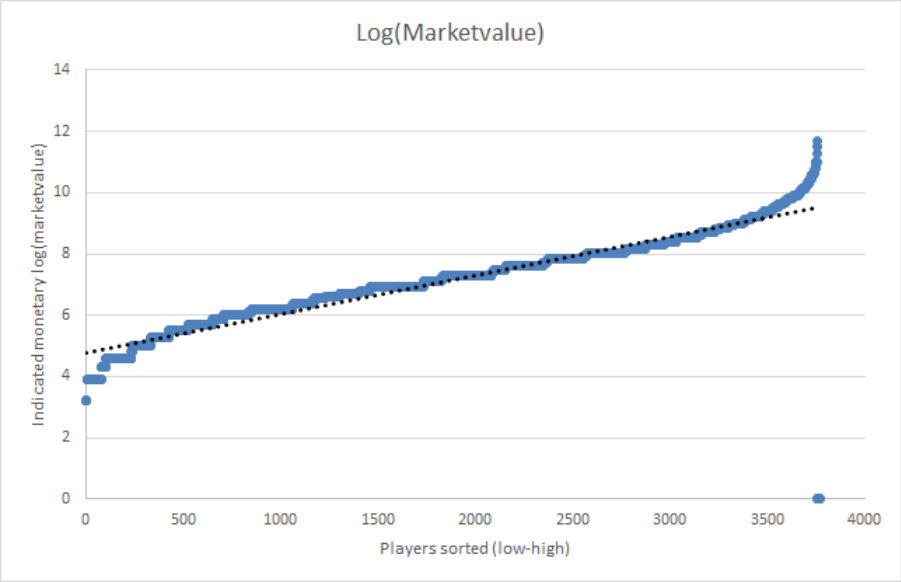
[6.] P H Franses and R Paap, 2001] Quantitative Models in Marketing Research, Cambridge: Cambridge University Press

[6.] S Monster , N Ramsaroep and A Mutsaerts 2014] Estimating the market value of football players - An analysis over the eight main football leagues in Europe.

A. Appendix



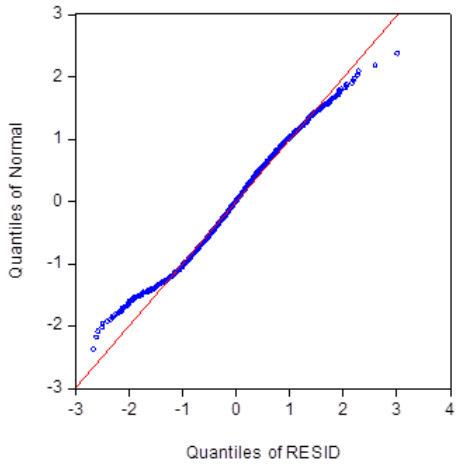
Figuur 2: *Plot van de marktwaarden*



Figuur 3: *Plot van de log marktwaarden*

Variabelen	Beschrijving	Verwachting
age	Leeftijd van de speler	ja
age2	Gekwadrateerde leeftijd van de speler	yes
assists_x	Het aantal gegeven assists van de speler in jaar x	ja
bundeslige_ger	dummie variabele die 1 is als de speler in de Duitse Bundesliga speelt	ja
bunderliga_aus	dummie variabele die 1 is als de speler in de Oostenrijkse Bundesliga speelt	ja
canuseboth	dummie variabele die 1 is als de speler tweebenig is	nee
canuseleft	dummie variabele die 1 is als de speler linksbenig is	nee
canuseright	dummie variabele die 1 is als de speler rechtsbenig is	nee
cl	dummie variabele die 1 is als de speler ooit in de Champions League heeft gespeeld	ja
ek	dummie variabele die 1 is als de speler ooit op een EK heeft gespeeld	ja
el	dummie variabele die 1 is als de speler ooit in de Europa League heeft gespeeld	ja
eredivisie	dummie variabele die 1 is als de speler in de Nederlandse Eredivisie speelt	ja
goals_x	Het aantal gescoorde goals van de speler in jaar x	ja
goalsconceded_x	Het aantal tegengoals van de speler in jaar x	ja
height	De lengte van de speler	ja
inlastyear	dummie variabele die 1 is als de speler in zijn laatste contract jaar zit	ja
international_comp	dummie variabele die 1 is als de speler ooit in een internationale landencompetitie heeft gespeeld exclusief het EK en WK	nee
isafrica	dummie variabele die 1 is als de speler uit Afrika komt	nee
isasia	dummie variabele die 1 is als de speler uit Azië komt	nee
isdef	dummie variabele die 1 is als de speler een verdediger is	ja
iseurope	dummie variabele die 1 is als de speler uit Europa komt	nee
isgoal	dummie variabele die 1 is als de speler een keeper is	ja
ismid	dummie variabele die 1 is als de speler een middenvelder is	ja
isnamer	dummie variabele die 1 is als de speler uit Noord-Amerika komt	nee
isocean	dummie variabele die 1 is als de speler uit Oceanië komt	nee
issamer	dummie variabele die 1 is als de speler uit Zuid-Amerika komt	nee
isstriker	dummie variabele die 1 is als de speler een aanvaller is	ja
jupiler	dummie variabele die 1 is als de speler in de Belgische Jupiler Pro League speelt	ja
ligue1	dummie variabele die 1 is als de speler in de Franse Ligue 1 speelt	ja
matches_x	Het aantal gespeelde wedstijden van de speler in jaar x	nee
minutes_x	Het aantal gespeelde minuten van de speler in jaar x	ja
months_at_club	Het aantal maanden dat een speler bij zijn club speelt	ja
months_to_go	Het aantal maanden dat een speler nog heeft tot zijn contract verloopt	ja
owngoals_x	Het aantal eigen doelputen van de speler in jaar x	nee
premier_leugue	dummie variabele die 1 is als de speler in de Engelse Premier League speelt	ja
primera_division	dummie variabele die 1 is als de speler in de Spaanse Primera Division speelt	ja
red_x	Het aantal directe rode kaarten ontvangen van de speler in jaar x	nee
serie_a	dummie variabele die 1 is als de speler in de Italiaanse Serie A speelt	ja
sommin	Het totaal aantal gespeelde minuten van een speler in heel zijn voetbalcarrière	ja
stadiumsize	De stadiongrootte van de club van de speler	ja
suboff_x	Het aantal keer dat een speler is gewisseld in jaar x	nee
subon_x	Het aantal keer dat een speler mocht invallen in jaar x	nee
tonill_x	Het aantal keer dat een keeper geen tegengoals kreeg in jaar x	ja
waschamply	dummie variabele die 1 is als de speler vorig jaar landskampioen was	nee
wk	dummie variabele die 1 is als de speler ooit op een WK heeft gespeeld	nee
yellow_x	Het aantal gele kaarten ontvangen van de speler in jaar x	nee
yellowred_x	Het aantal rode kaarten ontvangen als gevolg van twee gele kaarten van de speler in jaar x	nee

Tabel 7: *Alle beschikbare variabelen met een korte beschrijving en mijn verwachting met betrekking tot hun significantie*



Figuur 4: *QQ-plot van de residuen van het algemeen model van het werkcollege*

Dependent Variable: LOG_MARKETVALUE_1000_1
Method: Least Squares
Date: 06/25/14 Time: 06:22
Sample: 1 3770 IF _ALL75=1
Included observations: 2627
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
_MONTHS_AT_CLUB	0.001785	0.000448	3.984844	0.0001
_MONTHS_TO_GO	0.014730	0.001526	9.650078	0.0000
AGE	0.568796	0.045287	12.55985	0.0000
AGE2	-0.011673	0.000863	-13.51908	0.0000
BUNDESLIGA_OOS	-1.440525	0.055568	-25.92343	0.0000
EREDIVISIE	-0.960314	0.050729	-18.93017	0.0000
EL	0.278881	0.026654	10.46317	0.0000
EK	0.141418	0.052221	2.708082	0.0068
INLASTYEAR	0.173814	0.041414	4.196950	0.0000
INTERNATIONAL_COMP	0.114960	0.042452	2.708019	0.0068
WASCHAMPLY	0.287632	0.062706	4.586985	0.0000
STADIONGROOTTE	9.31E-06	1.00E-06	9.269685	0.0000
JUPILER	-0.769604	0.049187	-15.64640	0.0000
LIGUE1	-0.366511	0.043882	-8.352249	0.0000
MATCHES2010	0.005959	0.001150	5.181045	0.0000
MATCHES2011	0.007664	0.001147	6.680339	0.0000
MATCHES2012	0.015594	0.001299	12.00707	0.0000
MATCHES2013	0.031336	0.001723	18.18321	0.0000
ASSISTS2013	0.028068	0.006426	4.367677	0.0000
BUNDESLIGA_DUIT	-0.576087	0.051538	-11.17801	0.0000
CL	0.313063	0.033012	9.483443	0.0000
GOALS2013	0.018776	0.004144	4.531277	0.0000
ISSAMER	0.265007	0.044872	5.905881	0.0000
PRIMERA_DIV	-0.244453	0.045141	-5.415299	0.0000
SERIE_A	-0.161128	0.053146	-3.031820	0.0025
SOMMIN	6.26E-06	2.73E-06	2.291928	0.0220
C	-1.436023	0.589613	-2.435535	0.0149

R-squared	0.800221	Mean dependent var	7.277486
Adjusted R-squared	0.798223	S.D. dependent var	1.408348
S.E. of regression	0.632624	Akaike info criterion	1.932343
Sum squared resid	1040.554	Schwarz criterion	1.992712
Log likelihood	-2511.133	Hannan-Quinn criter.	1.954205
F-statistic	400.5529	Durbin-Watson stat	1.864473
Prob(F-statistic)	0.000000		

Figuur 5: *De parameterschattingen van het algeheel model*

Dependent Variable: LOG_MARKETVALUE_1000_1
Method: Least Squares
Date: 06/25/14 Time: 06:24
Sample: 1 3770 IF _LOW75=1
Included observations: 526
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
_MONTHS_TO_GO	0.005810	0.001940	2.993961	0.0029
AGE	0.447235	0.045386	9.854110	0.0000
AGE2	-0.008313	0.000889	-9.346357	0.0000
MATCHES2012	0.019141	0.006357	3.010735	0.0027
MATCHES2013	0.047712	0.007937	6.011683	0.0000
MINUTES2012	-0.000182	7.52E-05	-2.419546	0.0159
MINUTES2013	-0.000336	0.000101	-3.318754	0.0010
STADIONGROOTTE	5.89E-06	1.66E-06	3.554672	0.0004
SOMMIN	1.70E-05	5.04E-06	3.377371	0.0008
WASCHAMPLY	0.294833	0.106701	2.763174	0.0059
EL	0.209171	0.057287	3.651285	0.0003
BUNDESLIGA_DUIT	-0.443314	0.077100	-5.749876	0.0000
EREDIVISIE	-0.284320	0.056985	-4.989382	0.0000
BUNDESLIGA_OOS	-0.289958	0.069556	-4.168724	0.0000
SERIE_A	-0.176151	0.086766	-2.030176	0.0429
C	-0.906212	0.584803	-1.549602	0.1219

R-squared	0.474435	Mean dependent var	5.291052
Adjusted R-squared	0.458977	S.D. dependent var	0.599098
S.E. of regression	0.440662	Akaike info criterion	1.228869
Sum squared resid	99.03336	Schwarz criterion	1.358612
Log likelihood	-307.1926	Hannan-Quinn criter.	1.279669
F-statistic	30.69222	Durbin-Watson stat	1.943128
Prob(F-statistic)	0.000000		

Figuur 6: *De parameterschattingen van het lage categorie model*

Dependent Variable: LOG_MARKETVALUE_1000_1
Method: Least Squares
Date: 06/25/14 Time: 06:23
Sample: 1 3770 IF _MID75=1
Included observations: 1534
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
_MONTHS_AT_CLUB	0.001111	0.000427	2.602859	0.0093
_MONTHS_TO_GO	0.006282	0.000989	6.351879	0.0000
AGE	0.301484	0.034099	8.841425	0.0000
AGE2	-0.006458	0.000631	-10.23421	0.0000
BUNDESLIGA_OOS	-1.014789	0.051526	-19.69477	0.0000
CL	0.158655	0.029851	5.314918	0.0000
EL	0.178476	0.024270	7.353651	0.0000
EREDIVISIE	-0.643954	0.043711	-14.73192	0.0000
GOALS2013	0.016471	0.005087	3.238059	0.0012
STADIONGROOTTE	6.77E-06	9.91E-07	6.838296	0.0000
JUPILER	-0.555919	0.041625	-13.35537	0.0000
LIGUE1	-0.207053	0.033237	-6.229585	0.0000
MATCHES2010	0.004609	0.000977	4.717645	0.0000
MATCHES2011	0.004377	0.001008	4.341154	0.0000
MATCHES2012	0.008636	0.001170	7.382286	0.0000
MATCHES2013	0.021775	0.001667	13.06032	0.0000
BUNDESLIGA_DUIT	-0.261823	0.039745	-6.587513	0.0000
SUBON2013	-0.013581	0.003152	-4.308485	0.0000
WASCHAMPLY	0.145553	0.080410	1.810146	0.0705
WK	0.127123	0.051676	2.460016	0.0140
SERIE_A	-0.118177	0.047205	-2.503487	0.0124
ISDEF	-0.062695	0.026618	-2.355335	0.0186
INTERNATIONAL_COMP	0.091149	0.040225	2.266003	0.0236
ISSAMER	0.089594	0.042988	2.084154	0.0373
C	2.814133	0.453853	6.200544	0.0000

R-squared	0.534950	Mean dependent var	7.224488
Adjusted R-squared	0.527553	S.D. dependent var	0.647468
S.E. of regression	0.445035	Akaike info criterion	1.234837
Sum squared resid	298.8671	Schwarz criterion	1.321793
Log likelihood	-922.1196	Hannan-Quinn criter.	1.267195
F-statistic	72.32542	Durbin-Watson stat	1.702196
Prob(F-statistic)	0.000000		

Figuur 7: *De parameterschattingen van het midden categorie model*

Dependent Variable: LOG_MARKETVALUE_1000_1
Method: Least Squares
Date: 06/06/14 Time: 21:30
Sample: 1 3770 IF _HIGH75=1
Included observations: 581

Variable	Coefficient	Std. Error	t-Statistic	Prob.
_MONTHS_TO_GO	0.006598	0.001048	6.297739	0.0000
BUNDESLIGA_OOS	-0.987790	0.209797	-4.708323	0.0000
GOALS2012	0.007625	0.002947	2.587064	0.0099
GOALS2013	0.014270	0.004627	3.083819	0.0021
CL	0.301530	0.037763	7.984726	0.0000
STADIONGROOTTE	4.44E-06	9.65E-07	4.605386	0.0000
ASSISTS2013	0.016720	0.005858	2.854451	0.0045
BUNDESLIGA_DUIT	-0.224095	0.045924	-4.879679	0.0000
EREDIVISIE	-0.681591	0.087906	-7.753628	0.0000
ISSAMER	0.175074	0.044606	3.924869	0.0001
JUPILER	-0.557344	0.118666	-4.696731	0.0000
MATCHES2011	0.005487	0.001393	3.938678	0.0001
MATCHES2012	0.009854	0.001811	5.440677	0.0000
MATCHES2013	-0.013209	0.004993	-2.645575	0.0084
MINUTES2013	0.000233	5.35E-05	4.352106	0.0000
SOMMIN	1.62E-05	3.32E-06	4.863380	0.0000
WASCHAMPLY	0.297008	0.050623	5.867103	0.0000
AGE	0.195178	0.065140	2.996272	0.0029
AGE2	-0.005345	0.001317	-4.058858	0.0001
PRIMERA_DIV	-0.139061	0.048843	-2.847078	0.0046
LIGUE1	-0.108800	0.054358	-2.001567	0.0458
C	5.990214	0.806101	7.431092	0.0000

R-squared	0.669755	Mean dependent var	9.196751
Adjusted R-squared	0.657348	S.D. dependent var	0.642183
S.E. of regression	0.375911	Akaike info criterion	0.918202
Sum squared resid	78.99183	Schwarz criterion	1.083477
Log likelihood	-244.7377	Hannan-Quinn criter.	0.982633
F-statistic	53.98481	Durbin-Watson stat	1.778065
Prob(F-statistic)	0.000000		

Figuur 8: *De parameterschattingen van het hoge categorie model*

Dependent Variable: _ORDVAL
Method: ML - Ordered Logit (Quadratic hill climbing)
Date: 06/24/14 Time: 18:02
Sample: 1 3770 IF _ALL75=1
Included observations: 2819
Number of ordered indicator values: 3
Convergence achieved after 8 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AGE	1.752985	0.143130	12.24748	0.0000
AGE2	-0.037356	0.002802	-13.32963	0.0000
ASSISTS2013	0.144331	0.030049	4.803130	0.0000
BUNDESLIGA_OOS	-3.954374	0.250350	-15.79538	0.0000
CL	0.768365	0.134676	5.705294	0.0000
EL	0.752322	0.112898	6.663735	0.0000
EREDIVISIE	-2.109755	0.187530	-11.25022	0.0000
ISSAMER	0.891699	0.179400	4.970454	0.0000
JUPILER	-1.602045	0.192206	-8.335055	0.0000
MATCHES2010	0.012441	0.004698	2.647964	0.0081
MATCHES2011	0.017992	0.004685	3.840624	0.0001
MATCHES2012	0.039770	0.005541	7.176848	0.0000
MATCHES2013	0.095077	0.007254	13.10720	0.0000
LIGUE1	-0.531520	0.154616	-3.437676	0.0006
SOMMIN	6.37E-05	1.18E-05	5.424392	0.0000
STADIONGROOTTE	2.84E-05	3.86E-06	7.359344	0.0000
WASCHAMPLY	0.880749	0.265690	3.314955	0.0009
ASSISTS2012	0.047372	0.021476	2.205781	0.0274
BUNDESLIGA_DUIT	-1.368566	0.168574	-8.118513	0.0000
ISSTRIKER	0.314991	0.127504	2.470436	0.0135
Limit Points				
LIMIT_2:C(21)	22.07931	1.809790	12.19993	0.0000
LIMIT_3:C(22)	28.26309	1.868635	15.12499	0.0000

Pseudo R-squared	0.515917	Akaike info criterion	0.965819
Schwarz criterion	1.012208	Log likelihood	-1339.322
Hannan-Quinn criter.	0.982558	Restr. log likelihood	-2766.721
LR statistic	2854.799	Avg. log likelihood	-0.475105
Prob(LR statistic)	0.000000		

Figuur 9: *De parameterschattingen van het Ordered Logit model*

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	17.09810	Prob. F(26,2600)	0.0000
Obs*R-squared	383.5818	Prob. Chi-Square(26)	0.0000
Scaled explained SS	562.9916	Prob. Chi-Square(26)	0.0000

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	3.738352	Prob. F(15,510)	0.0000
Obs*R-squared	52.10544	Prob. Chi-Square(15)	0.0000
Scaled explained SS	48.07862	Prob. Chi-Square(15)	0.0000

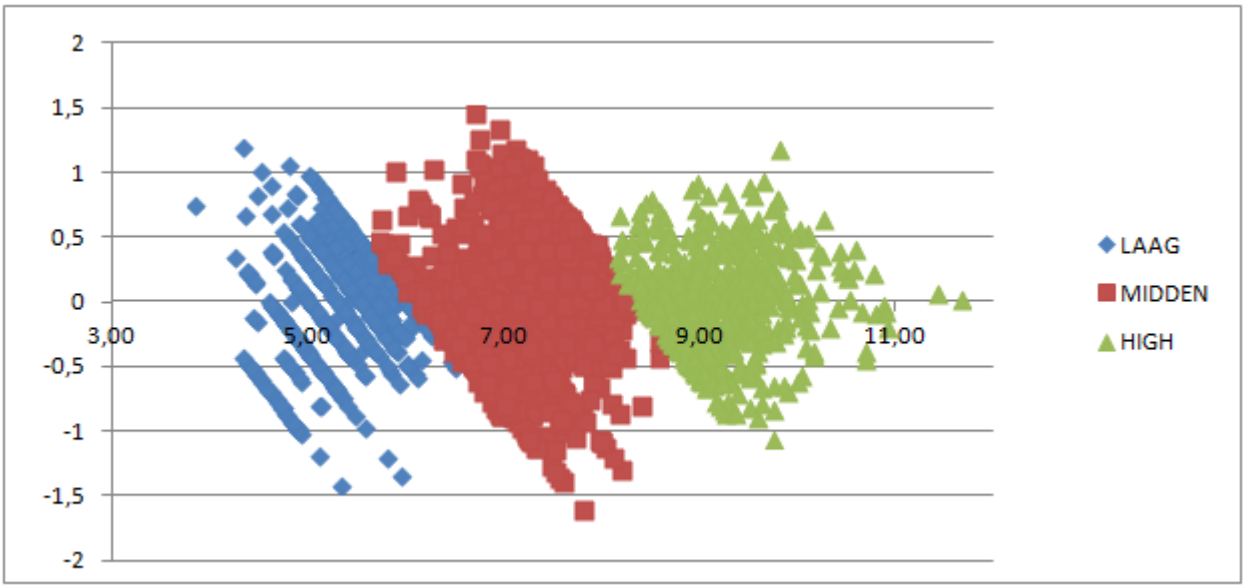
Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	4.230267	Prob. F(24,1509)	0.0000
Obs*R-squared	96.70224	Prob. Chi-Square(24)	0.0000
Scaled explained SS	96.26342	Prob. Chi-Square(24)	0.0000

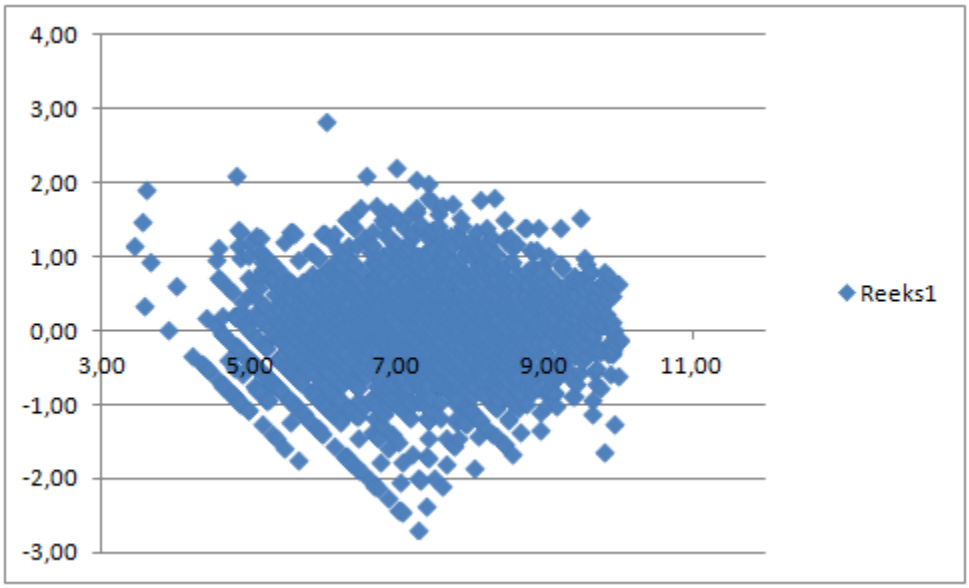
Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	1.382625	Prob. F(21,559)	0.1193
Obs*R-squared	28.68776	Prob. Chi-Square(21)	0.1217
Scaled explained SS	23.25814	Prob. Chi-Square(21)	0.3304

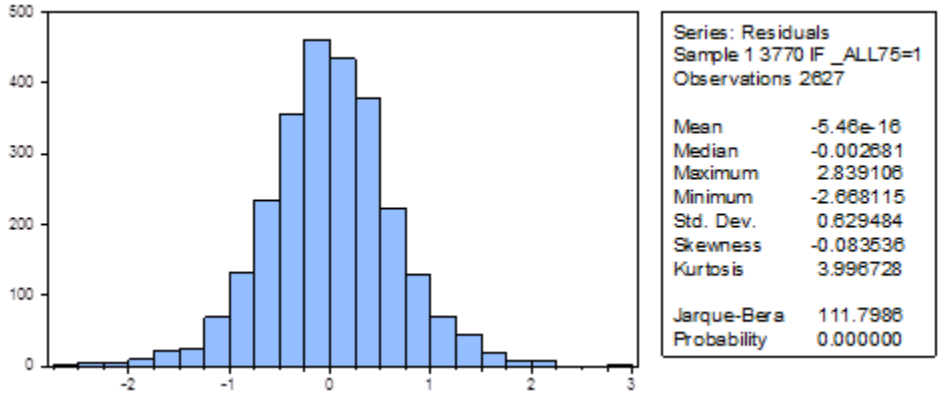
Figuur 10: *Uitkomsten van de Breusch-Pagan toeten voor het algehele model, het lage categorie model, het midden categorie model en het hoge categorie model*



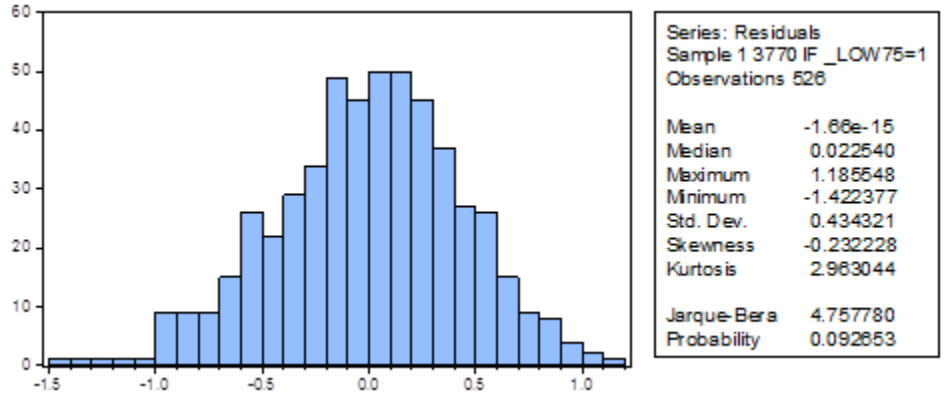
Figuur 11: *Spreidingsdiagram van de gefitte waarden van de drie categorie modellen tegen de residuen*



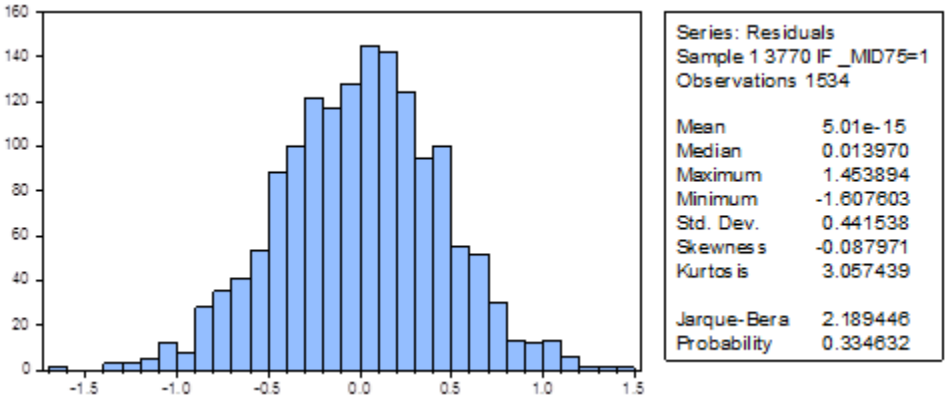
Figuur 12: *Spreidingsdiagram van de gefitte waarden van het algeheel model tegen de residuen*



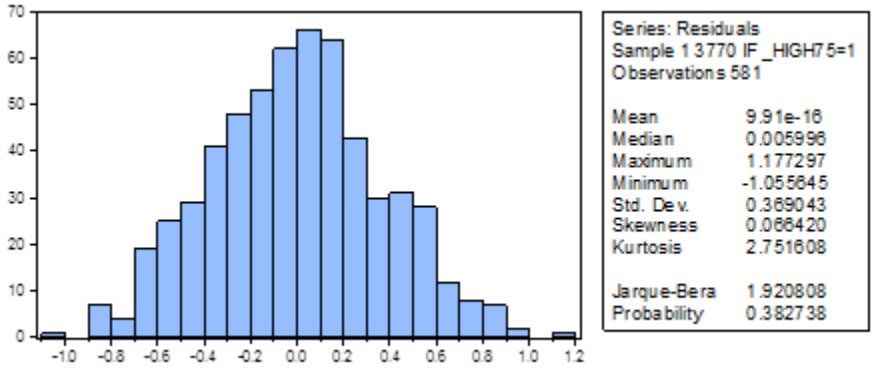
Figuur 13: *Histogram van de residuen van het algehele model*



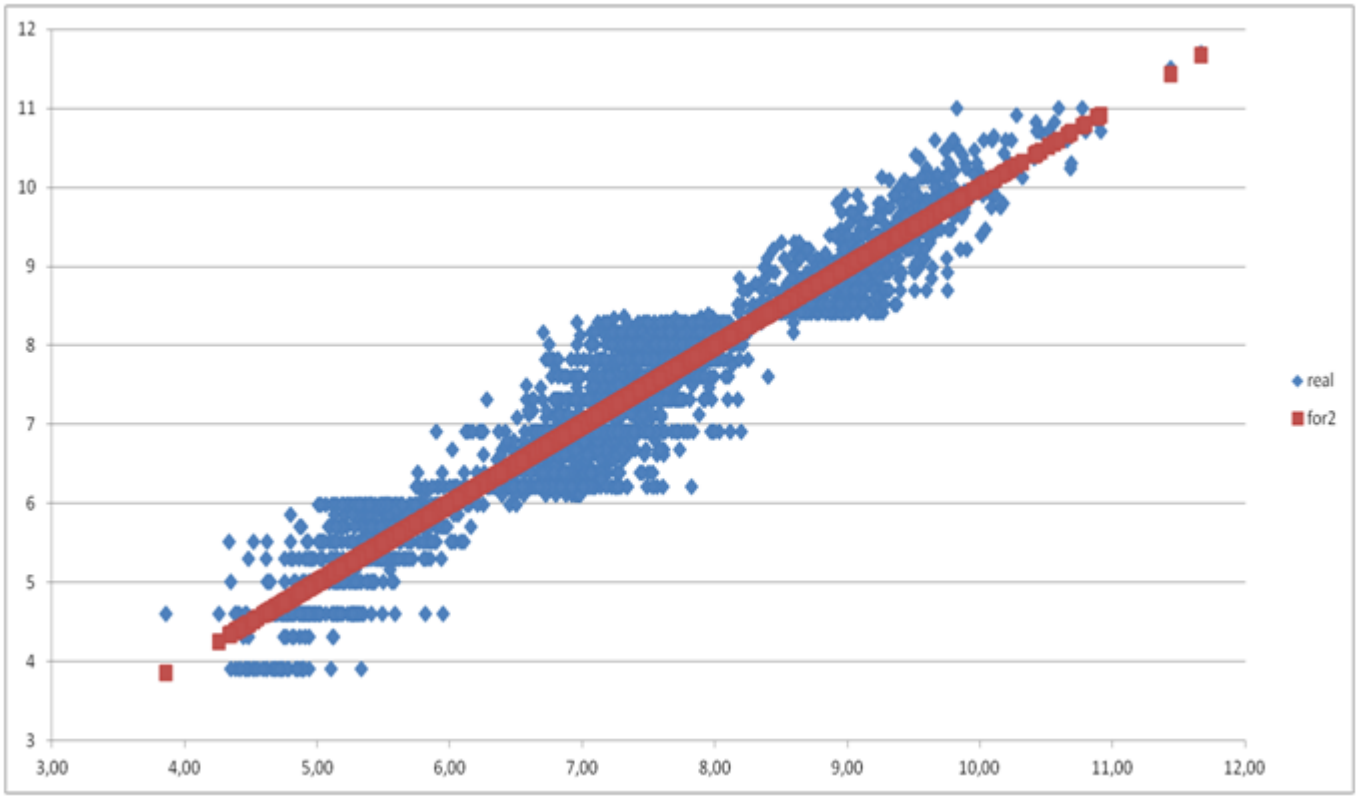
Figuur 14: *Histogram van de residuen van het lage categorie model*



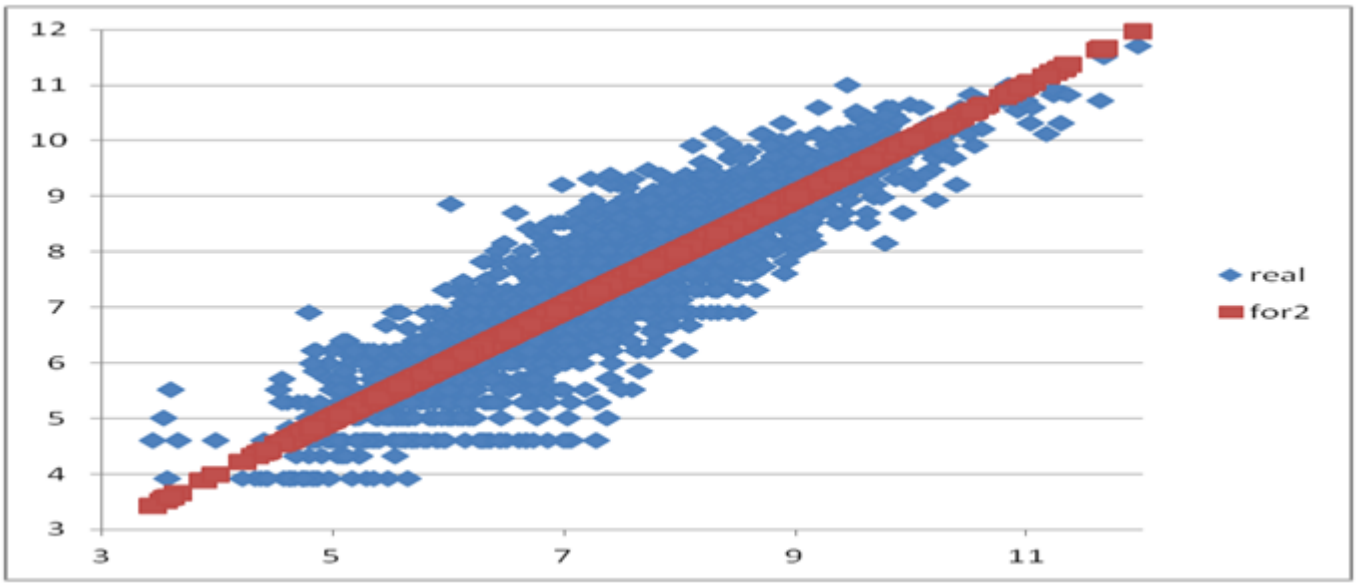
Figuur 15: *Histogram van de residuen van het midden categorie model*



Figuur 16: *Histogram van de residuen van het hoge categorie model*



Figuur 17: *Spreidingsdiagram van de gefitte waarden van de drie categorie modellen tegen de echte waarden*



Figuur 18: *Spreidingsdiagram van de gefitte waarden van het algeheel model tegen de echte waarden*