Marktwaarde bepaling voor voetbalspelers van verschillende leeftijdsgroepen uit acht Europese competities

Annemarijn Mutsaerts

Erasmus Universiteit Rotterdam
Erasmus School of Economics

begeleider: **Dr. A. Alfons**

juni 2014

Samenvatting

Dit onderzoek beschrijft de constructie van twee modellen die de marktwaarde van voetbalspelers uit acht verschillende competities verklaart: Champions League, Duitse Bundesliga, Eredivisie, Jupiler Pro League, Ligue 1, Oostenrijkse Bundesliga, Primera Division en Serie A. Het eerste model is voor spelers van 23 jaar en jonger en is gebaseerd op 1201 observaties. Het tweede model is voor spelers van 30 jaar en ouder en is gebaseerd op 585 observaties. Voor het construeren van dit model is de OLS methode gebruikt. Als afhankelijke variabele wordt log(marktwaarde) gebruikt. Deze modellen geven een betere verklaring van de marktwaarde dan eerdere modellen geconstrueerd in het werkcollege. Er geldt dat de twee modellen voor de verschillende leeftijdscategorieën uiteenlopende verklarende variabelen bevatten.

Inhoudsopgave

1. Inleiding

De Braziliaanse voetballer Neymar da Silva Santos, bekend onder de voetbalnaam Neymar, werd in de zomer van 2013 voor 86,2 miljoen euro verkocht aan FC Barcelona (www.nusport.nl, 2014). Dit is een opmerkelijk hoog bedrag voor een 21 jarige voetballer. Zijn marktwaarde wordt op dit moment geschat op 60 miljoen euro (www.transfermarkt.com, 2014).

Ook bij jonge spelers gaat al veel geld om in voetbal. Maar waarom is de ene jonge voetballer zoveel meer waard dan de andere? Er valt bij deze groep spelers namelijk nauwelijks te spreken over ervaring. Hoe is een hele goede jonge speler te herkennen? Als dit al vroeg in de carrière van een speler zou kunnen, zou in deze speler geïnvesteerd kunnen worden. Het herkennen van talent is dus niet allen voor de huidige club van de speler relevant, maar ook voor kopende clubs. Ook in het werkcollege onderzoek kan worden gezien dat leeftijd en $leeftijd^2$ een belangrijke invloed hebben bij het verklaren van de marktwaarde. Om deze reden wordt er in dit onderzoek meer aandacht besteed aan verschillende leeftijdscategorieën in voetbal.

Dit onderzoek is een vervolg van een eerder onderzoek dat is gedaan in het werkcollege. In dit eerdere onderzoek, *Estimating the market value of football players: An analysis over the eight main football leagues in Europe* zijn door Sander Monster, Nishant Ramsaroep en Annemarijn Mutsaerts, drie modellen geconstrueerd die de marktwaarde van respectievelijk alle voetbalspelers, veldspelers en keepers uit acht verschillende Europese competities verklaart. Aan dit werkcollege onderzoek wordt met regelmaat gerefereerd. Dit model wordt in het onderzoek **het algemene veldspelersmodel** genoemd.

Omdat verwacht wordt dat voor jonge spelers andere variabelen van invloed zijn dan voor oude spelers, is er in dit onderzoek voor gekozen om dezelfde database op te splitsen in twee leeftijdscategorieën. De eerste categorie vertegenwoordigd jonge veldspelers van 23 jaar en jonger en de tweede categorie vertegenwoordigd de oude spelers van 30 jaar en ouder. De onderzoeksvragen die in dit verslag naar voren komen luiden als volgt:

1. Welke modellen verklaren het beste de marktwaarde voor oude en jonge spelers?
2. Wat zijn de grootste verschillen tussen de modellen voor jonge en oude spelers?
3. Wat gebeurt er met de variabelen $leeftijd$ en $leeftijd^2$?
4. Zijn ervaringsvariabelen relevant voor jonge en oude spelers?
5. Wat is de leeftijdsverdeling bij de acht verschillende competities?
6. Zijn er competities waarbij jonge respectievelijk oude spelers een hogere marktwaarde hebben?
7. Geeft het opsplitsen van de jonge en oude spelers een betere verklaring dan het algemene model uit het werkcollege?

Allereerst worden deze twee datasets geanalyseerd aan de hand van een Kruskal-Wallis toets. Vervolgens worden twee modellen geconstrueerd die de *log(marktwaarde)* van de jonge en de oude spelers verklaart. Het voornaamste doel van deze twee modellen is dat ze inzicht geven over de

relevante variabelen bij het verklaren van de marktwaarde van de twee verschillende categorieën. Wat zijn relevante statistieken? Bij het vervaardigen van de modellen wordt gebruik gemaakt van de *Ordanary Least Square* methode. Als de modellen zijn geconstrueerd wordt allereerst gekeken naar de overeenkomsten en verschillen tussen deze twee modellen. Daarna wordt gekeken of het model dat is vervaardigd met het opsplitsen van de data tot een beter resultaat heeft geleid dan het model uit het werkcollege.

Verwacht wordt dat er wel degelijk een verschil in verklarende variabelen bestaat tussen oude en jonge spelers. Van de ervaringsvariabelen en variabelen uit het verleden (2010 en eerder) wordt verwacht dat ze vooral voor oude spelers relevant zijn en minder voor jonge spelers. Dit wordt verwacht omdat voor de jonge spelers geldt dat nog niet veel statistieken uit het verleden bekend zijn, omdat ze minder lang meegaan in de voetbalwereld. De variabele leeftijd, die een belangrijke invloed had in het werkcollege onderzoek, heeft waarschijnlijk geen invloed meer door het opsplitsen van de categorieën.

2. Literatuuronderzoek

Voetballers staat al jaren lang in de top tien van bestbetaalde sporters ter wereld (www.sport.infonu.nl). Er is in het verleden veel onderzoek gedaan naar de marktwaarden en transfer prijzen van voetbalspelers. De competitie waar het meest onderzoek naar gedaan is, is de Premier League. Een aantal van deze onderzoeken en de resultaten daarvan worden hieronder kort besproken.

Er zijn enkele onderzoeken bekend over de totstandkoming van de transferbedragen in de Engelse competitie. In het onderzoek van Carmichael en Thomas worden de transferbedragen van voetbalspelers in de Engelse voetbalcompetitie onderzocht voor data uit het seizoen 1990-1991. Zij deden als een van de eerste onderzoek naar dit vraagstuk waarop velen volgden. Zij maakten gebruik van een *two-person bargaining theory*, een onderhandelingstheorie tussen de kopende en de verkopende club (Carmichael en Thomas, 1993).

Reilly en Witt deden in 1995 een vervolgonderzoek naar de transferprijzen in Premier League voetbal. Het grootste verschil met het onderzoek van Carmichael en Thamas is dat zij een ras dummy toevoegden aan hun model. Hun hypothese was dat voor voetballers met een blanke huidskleur hogere transferprijzen werden betaald. Hiermee probeerden zij te onderzoeken of er destijds gediscrimineerd werd in de voetbalwereld. De conclusie van dit onderzoek was dat er geen verschillen bestaan in transferprijzen voor voetbalspelers met een andere huidskleur. Er wordt dus niet gediscrimineerd (Reilly en Witt, 1995).

In 1995 heeft er een belangrijke verandering plaatsgevonden in de voetbalwereld. De voetbalspeler Jean-Marc Bosman speelde voor een Belgische voetbalclub. In 1990 verliep zijn contract bij Club Luik en wilde hij een nieuwe overeenkomst aangaan met een voetbalclub. Club Luik eiste echter een transfersom op al was het contract verlopen. Dit zorgde er uiteindelijk voor dat zijn nieuwe club USL Dunkerque de speler niet opnam in de selectie. De speler ging hiermee naar de rechter, omdat hij vond dat dit in strijd was met het verdrag van Rome. Op 15 december 1995 is er een uitspraak gedaan door het Europees Hof waarin Bosman in het gelijk werd gesteld. Vanaf dat moment was er jurisprudentie omtrend het vragen van een transfersom in geval van verkoop van een speler na afloop van zijn contract. Het incident met Bosman is hiervoor de aanleiding geweest (Simmons, 1997). Voor de onderzoeken naar transferprijzen die data bevatten van voor 1995 geldt dus dat er nog een transfersom mocht worden gevraagd na het aflopen van het contract

van de spelers. Na 1995 kan ook worden geconstateerd dat de contractduur van spelers enorm is toegenomen. Alleen als een speler nog een lopend contract heeft, mag een transfersom worden gevraagd door de huidige club van de speler. In onderzoeken na 1995 is de variabele *contractuntil* daarom ook een erg belangrijke variabele.

In 1997 deed Thomas weer een onderzoek naar de transferprijzen. Dit keer samen met Speight. In dit onderzoek wordt een model geconstrueerd dat de transferprijzen in Engels voetbal voorspelt. Opvallend is dat in dit onderzoek ook de variabele $leeftijd^2$ voorkomt. In de onderzoeken van Carmichael en Thomas en Reilly en Witt, was dit niet het geval. Zij namen alleen $leeftijd$ mee in hun modellen. In het onderzoek werkcollege is ook te zien dat dit een zeer significante variabele is. Door het opsplitsen van de databases in twee categorieën, die gebaseerd zijn op leeftijd, is ook de variabele $leeftijd^2$ niet meer significant (Speight en Thomas, 1997).

In deze drie onderzoeken kwamen steeds drie verschillende categorieën variabelen naar voren: *spelers karakteristieken, kopende club karakteristieken en verkopende club karakteristieken*. In 2000 voegde Dobson, Gerrard en Howe daar een vierde categorie aan toe, namelijk *tijdseffecten*. Zij deden in 1999 onderzoek naar de Engelse voetbal competitie. Het grootste verschil met voorgaande onderzoeken is dat zij in dit onderzoek het onderscheid maakten tussen professioneel voetbal en amateurvoetbal. Ook de variabelen uit dit onderzoek waren op te splitsen in de vier categorieën die hierboven genoemd zijn. Zeer opmerkelijk is dat zij geen bewijs vonden voor verschillen in verklarende variabelen tussen deze twee groepen (Dobson, Gerrard en Howe, 2000).

Frick herhaalde het onderzoek in 2007, maar dit keer niet voor de Engelse Premier League, maar voor de Duitse Bundesliga. hij voegde daarnaast ook een dummy variabele toe, die iets zegt over de continentale afkomst van de speler. De continenten Azië en Noord-Amerika hebben in dit onderzoek een negatief significante invloed op de transfer prijzen in de Duitse Bundesliga, en de continenten Europa en Zuid-Amerika een positieve significante invloed. Daarnaast is in dit onderzoek ook een trendlijn toegevoegd. In de modellen in dit onderzoek is ook te zien dat het continent Zuid-Amerika een significantie invloed heeft. Hier is als basis Europa genomen (Frick, 2007).

In veel onderzoeken die in het verleden zijn gedaan naar prijzen in de voetbalwereld, is te zien dat de afhankelijke variabelen in de meeste onderzoeken *log(transferfee)* is. Over het algemeen geldt ook dat er vier typen variabelen zijn die relevant worden geacht bij het verklaren van de *log(transferfee)*: spelers karakteristieken, kopende club karakteristieken en verkopende club karakteristieken en tijdseffecten. In dit onderzoek is niet de *log(transferfee)* maar de *log(marktwaarde)* gebruikt als afhankelijke variabelen. De modellen komen echter wel veel overheen op het gebied van verklarende variabelen. Wel komen alle vier de typen variabelen terug in dit onderzoek.

3. Data

3.1. Data Constructie

De data zijn verkregen door een automatische scraper toe te passen. Dit is gedaan door de begeleider van dit onderzoek: Dr. A. Alfons. De data zijn afkomstig van de site *www.tranfermarkt.com*. Deze site bevat veel verschillende statistieken van spelers uit Europees voetbal. Voor dit onderzoek zijn de statistieken gebruikt van spelers uit 8 verschillende Europese competities: de Oostenrijkse Bundesliga, de Duitse Bundesliga, de Premier League, de Primera Division, Ligue 1, de Julpiler Pro League, Serie A en de Eredivisie. De voetbal statistieken zijn afkomstig uit maart 2014 en

gaan terug tot het begin van de voetbalcarrière van de desbetreffende voetbalspelers. In totaal zijn de statistieken van 3782 spelers bekend. Deze spelers zijn op te splitsen in keepers en veldspelers.

De afhankelijke variabele in dit onderzoek is de monetaire marktwaarde in euros van een voetbalspeler. Over hoe deze monetaire marktwaarde tot stand komt zegt de site *www.transfermarkt.com* het volgende:

"De Transfermarkt-spelerscijfers worden door onze experts, gegevensonderhouders en moderators bepaald. Bij een klik op een cijfer zie je de specifiekere samenstelling en het stemgedrag van de gebruiker. Bij spelers met minstens 20 stemmen wordt bovendien een deel van het afgegeven cijfer door een standaard-afwijkingsberekening buiten beschouwing gelaten, die te veel van een mening van een andere gebruiker afwijkt. Deze stemmen bepalen dan het eindcijfer."(bron: www.transfermarkt.nl)

Een aantal statistieken is handmatig aangepast als gevolg van verkeerde waarden op de site. Zo was bij een aantal spelers het aantal minuten wat een speler heeft gespeeld negatief. Deze zijn handmatig aangepast door de waarden te controleren op verschillende voetbal sites met vergelijkbare datasets.

Voor dit onderzoek is de dataset gesplitst in 2 groepen; een eerste groep bestaande uit speler van 23 jaar en jonger en een tweede groep bestaande uit spelers van 30 jaar en ouder. Groep 1 heeft 1426 observaties en groep 2 heeft 726 observaties. In het werkcollege onderzoek is geconcludeerd dat er een groot verschil bestaat in verklarende variabelen tussen veldspelers en keepers. Daarom is er in dit onderzoek voor gekozen om alleen de veldspelers te onderzoeken. De modellen in dit onderzoek zijn dus gebaseerd op de statistieken van veldspelers.

Na het verwijderen van de keeper statistieken bestaan de jonge spelers uit 1276 observaties en de oude spelers uit 585 observaties. Om de data vervolgens geschikt te maken voor het creëren van een model zijn de spelers zonder monetaire marktwaarde verwijderd uit de datasets. Na deze eliminatie bestaat groep 1 uit 1201 observaties en groep 2 uit 585 observaties. Alleen voor de jonge spelers geldt dus dat er observaties zijn verwijderd. Dit wordt waarschijnlijk veroorzaakt door het feit dat voor sommige jonge spelers geldt dat ze nog geen marktwaarde hebben omdat ze spelen in een amateurteam, terwijl ze op www.transfermarkt.com wel worden opgenomen in de A-selectie. Deze spelers zijn irrelevant voor het onderzoek, omdat hier alleen de professionele voetbalspelers worden bestudeerd. Er wordt daarom verwacht dat dit **sample selection problem** geen problemen met zich meebrengt.

Om te zorgen voor een juiste specificatie van de data, wordt niet de marktwaarde, maar de log(marktwaarde) gebruikt als afhankelijke variabelen. Door deze aanpassing kan er een model worden gecreëerd op basis van de *Ordinary least square method* (OLS).

Van deze spelers worden meer dan 150 verschillende statistieken gebruikt om het uiteindelijke model te bepalen. Deze statistieken zijn onder te verdelen in 3 verschillende categorieën; persoonlijke statistieken, prestatie statistieken en achtergrond statistieken. Deze variabelen gaan tot 1997 terug, indien ze voor de speler beschikbaar zijn.

3.2. Persoonlijke statistieken

Deze statistieken beschrijven de eigenschappen van de speler. Voorbeelden hiervan zijn: lengte, links- of rechtsvoetig, een dummy voor de competitie waar een speler in speelt, leeftijd en $leeftijd^2$. Voor leeftijd geldt over het algemeen dat spelers meer waard worden naarmate de leeftijd vordert. Dit geldt echter maar tot een bepaalde leeftijd. Daarna daalt de waarde van de voetbalspelers. Dit laatste effect kan worden aangetoond met $leeftijd^2$. Deze effecten zijn interessant om te onderzoeken, omdat in dit onderzoek de groepen zijn gesplitst op leeftijd. De verwachting is dat leeftijd significant is in het model voor de jonge spelers en $leeftijd^2$ significant voor de oude spelers. Er is ook handmatig een dummy toegevoegd voor de continentale afkomst van een speler. Hierbij is gekozen om niet alle landen apart mee te nemen, maar de landen te groeperen op continent. Hiermee kan worden onderzocht of spelers van een bepaald continent een hogere marktwaarde hebben dat andere spelers. Als laatste is ook een dummy toegevoegd voor de veldpositie van de spelers; aanvaller, middenveldspeler of verdediger. Als basis variabele is *isstriker* gebruikt om multicollineariteit tegen te gaan. De statistieken van de keepers zijn, zoals eerder genoemd, verwijderd uit de database.

3.3. Prestatie statistieken

De prestatiestatistieken zijn erg relevant voor het creëren van een model dat de marktwaarde verklaart. Voorbeelden hiervan zijn: het aantal goals gemaakt in een seizoen, het aantal minuten gespeeld in een seizoen, het aantal gespeelde wedstrijden in een seizoen, het aantal rode en gele kaarten in een seizoen, aantal assists. Dit zijn allemaal originele data van de site. Handmatig is daar nog aantoegevoegd het totaal aantal minuten gespeeld in het verleden. Deze ervaringsvariabele zou vooral relevant kunnen zijn bij jonge spelers omdat zij relatief weinig hebben gespeeld.

3.4. Achtergrond statistieken

De achtergrond statistieken geven algemene informatie over eigenschappen van een speler. Er is een dummy voor speciale competities (Champions League, Europa League, Europese Kampioenschap, World Cup) toegevoegd. Daarnaast is er ook een variabele *stadiongrootte* toegevoegd die gerelateerd is aan het aantal zitplaatsen in het stadion van de club waar de voetbalspeler momenteel speelt. Deze variabele is handmatig extra toegevoegd (www.worldstdiumdatabase.com, 20 maart 2014). Daarnaast is er ook een dummy *waschampion* toegevoegd die informatie geeft over of de speler kampioen is geworden in het laatste seizoen. De variabelen *inlastyear* en *contractuntil* geven informatie over de contracten van de voetbal spelers.

Figuur 1 geeft een overzicht van alle variabelen die in dit onderzoek zijn gebruikt.

4. Methodologie

4.1. Analyse van de Datasets

Voordat het model wordt geconstrueerd om de marktwaarde van de jonge en oude spelers te verklaren, wordt eerst de database van beide groepen geanalyseerd. Om onderzoeksvraag 5 te beantwoorden, die gaat over het verschil in verdeling van de leeftijd van de verschillende competities, wordt een **Kruskal-Wallistoets** uitgevoerd. Deze toets analyseert hoe de verdeling bij de verschillende competities is wat betreft leeftijd (Breslow, 1970).

Voor deze toets wordt de originele dataset gebruikt. Dat wil zeggen: de database met alle spelers ongeacht leeftijd. Uit deze database worden de keepers verwijderd, omdat in dit onderzoek alleen veldspelers worden onderzocht. Deze database bevat 3328 observaties uit 8 verschillende competities. Vervolgens wordt op deze database de Kruskal-Wallistoets uitgevoerd.

De **Kruskall-Wallistoets** maakt gebruik van de rangnummers van de data. Omdat in deze database veel spelers met dezelfde leeftijd voorkomen wordt het gemiddelde rangnummer genomen van de data met dezelfde waarde.

Als de toets wordt verworpen kan worden geconcludeerd dat er een verschil bestaat in de leeftijdsverdeling in de acht competities. In het geval van verwerping kan een verdere analyse worden gedaan naar de hoogte van die verschillen. In dit onderzoek wordt de mediaan, de MAD en de IQR bekeken. Deze robuuste statistieken worden gebruikt als schatters voor de schaalparameters. Belangrijk om op te merken is dat deze toetsen *geen aanname doen over de verdeling* van de data.

De mediaan is een goede manier om de robuuste eigenschappen van een database te vergelijken, maar deze zegt niets over de standaardafwijking σ. Daarvoor kan de *mediane absolute deviatie* (MAD) gebruikt worden. De MAD uit competitie j, MAD_j, wordt alsvolgt bepaald: Eerst wordt de gehele database gesplitst in acht aparte databases Y_j, met $j = 1, ..., 8$, voor elke competitie één. Vervolgens wordt de mediaan bepaald, $mediaan(Y_j)$. Van elke observatie X_i uit database Y_j wordt de mediaan afgetrokken en daarvan wordt de absolute waarde genomen. Tot slot wordt daarvan de mediaan genomen. Dit staat weergegeven in formule 1. Dit is de MAD (Pham-Ghia en Hung, 2000).

$$MAD_i = mediaan(|X_i - mediaan(Y_j)|) \tag{1}$$

Daarnaast kan ook worden gekeken naar de *interquantile range* (IQR). De IQR geeft het verschil weer tussen het 25ste percentiel en het 75ste percentiel, oftewel de spreiding van de data. De IQR waarde van de verschillende datasets wordt bepaald door formule 2. Hierin is Q_3 het 75% percentiel en Q_1 het 25% percentiel.

$$IQR = Q_3 - Q_1 \tag{2}$$

Met behulp van de mediaan, de MAD waarde en de IQR waarde kan een schatting worden gedaan van de middelste waarde en de spreiding van de data. Deze robuuste statistieken geven meer inzicht in de verdeling van leeftijd voor de verschillende competities.

4.2. Theoretisch Model

Voor het schatten van *log(marktwaarde)* aan de hand van de twee modellen voor de verschillende leeftijdscategorieën, wordt gebruik gemaakt van een simple regression. In dit onderzoek is gekozen voor de *Ordinairy Least Square* (OLS) methode voor het creëren van de twee modellen. Deze methode kan worden gebruikt, omdat de afhankelijke variabele in deze modellen continu is en zonder restricties na de transformatie van de marktwaarde naar de *log(marktwaarde)*. Deze methode is bruikbaar voor een lineair regressie model. Het lineaire model kan worden geschreven als: $y = X\beta + \varepsilon$. Hierin is Y een vector van de afhankelijke variabele log(marktwaarde), X een matrix met de verklarende variabelen en β een vector met onbekende parameters. Het geschatte model kan worden geschreven als: $y = Xb + e$. Hierin is b een vector van de geschatte waarden van β en e geeft de vector van de residuen weer. Deze kan worden verkregen door: $e = y - Xb$.

Om de *Least Squares estimator* te herleiden wordt gebruik gemaakt van formule 3, de som van de gekwadrateerde residuen.

$$S(b) = \sum e_i^2 = e'e = (y - Xb)'(y - Xb) = y'y - y'Xb - b'X'y + b'X'Xb \tag{3}$$

Hieruit kan de schatter b worden verkregen door de som van de gekwadrateerde residuen (formule 3) te differentiëren. De schatter voor b kan worden geschreven als in formule 4.

$$b = (X'X)^{-1}X'y \tag{4}$$

Het idee hierachter is dat de datapunten een rechte lijn volgen die de som van de verticale afstand tussen de gekwadrateerde residuen en deze lijn minimaliseert (Heij et all, 2004).

De afhankelijke variabele in de twee modellen is de *log(marktwaarde)* in euro van de verschillende voetbalspelers. Deze *log(marktwaarde)* wordt in dit onderzoek y genoemd. Door deze transformatie toe te passen volgt de lijn een betere lineaire fit door de data punten. Intuïtief komt dit neer op een exponentiële stijging als de waarde van de voetbalspelers toeneemt. Er kan aan de hand van figuur 8 en figuur 15 worden aangenomen dat de errortermen ε_i een normale verdeling volgen met een gemiddelde rond nul. X is een matrix van alle verklarende variabelen met een significante invloed en b is een matrix die de bijbehorende parameters weergeeft. Deze kunnen worden verkregen door *backward elemination* toe te passen. Bij deze methode worden insignificant variabelen verwijderd (backward elimination). Hierbij worden telkens de *Schwarz infomation criterion* (SIC) de *Akaike information criterion* (AIC) geminimaliseerd om een zo efficïent mogelijk model te krijgen. De SIC en de AIC worden herleid volgens formule 5 en formule 6.

$$SIC(p) = log(s_p^2) + \frac{plog(n)}{n} \tag{5}$$

$$AIC(p) = log(s_p^2) + \frac{2p}{n} \tag{6}$$

Hierin is p het aantal toegevoegde regresoren in het model, n het aantal observaties en S_p^2 is de maximum likelihood estomator van de errorvariantie in het model met p regresoren. De R^2 representeert de fit van het model en moet dus zo hoog mogelijk zijn. De R^2 wordt herleid volgens formule 7 (Heij et all, 2004).

$$R^2 = \frac{b^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \tag{7}$$

R^2 staat ook wel bekend als de *coefficient of determination*.

De geschatte parameters geven de raletie weer tussen de variabelen en de *log(marktwaarde)*. Deze relatie is alleen aannemelijk als de aannames van de OLS methode gelden. De drie belangrijkste aannames voor de OLS methode in dit onderzoek staan hieronder weergegeven:

1. Homoskedasticiteit; De errortermen hebben dezelfde variantie voor elke observatie, $E[\varepsilon_i^2] = \sigma^2$.
2. Normaliteit; De errortermen moeten normaal verdeeld zijn, $\varepsilon \sim N(0, \sigma^2)$
3. Correcte specificatie; De data moeten een lineaire trend volgen, $Y = \beta X + \varepsilon$.

Als de aannames gelden, dan geeft de OLS methode een *Best Linear Unbiased Estimator* (BLUE) (Heij et all, 2004).

4.3. Aannames voor OLS

Na een robuuste creatie van de twee modellen aan de hand van backward elimination worden de aannames getest en kunnen eventuele aanpassingen worden gedaan om het model zo optimaal mogelijk te maken. Het model dat is geconstrueerd volgens de OLS methode presteert het best als alle aannames gelden. Als dit niet het geval is moet er voorzichtig worden omgegaan met het interpreteren van de resultaten. Alle aannames worden achtereenvolgens gecontroleerd.

1. Homoskedasticiteit
De aanname voor homoskedasticiteit impliceert dat de errortermen een constante variantie hebben voor alle observaties zoals if formule 8.

$$E[\varepsilon|X] = \sigma^2 \tag{8}$$

Er wordt gekeken naar een plot van de residuen om deze aanname te controleren. Daarnaast worden er twee toetsen uitgevoerd om de aanname van homoskedasticiteit te testen; de *Breusch-Pagan toets* en de *White toets*. Als deze toetsen worden verworpen en de aanname dus niet geldt, is er sprake van heteroskedasticiteit. In dat geval kan de OLS methode worden uitgevoerd met White standaard fouten. In geval van heteroskedasticiteit kan de variantie van de schatter b worden geschreven zoals in formule 9 (Heij et all, 2004).

$$var(b) = (X'X)^{-1}(\sum_{i=1}^{n} \sigma_i^2 x_i x_i')(X'X)^{-1} \tag{9}$$

Hierin is x_i een vector van de verklarende variabelen voor alle i observaties en σ_i^2 de variantie van observatie i (Heij et all, 2004). In de meeste gevallen geldt dat de waarde voor σ_i^2 onbekend is. Deze kan worden geschat door de OLS methode en deze wordt weergegeven door e_i. De geschatte variantie van b wordt dan zoals in formule 10.

$$\hat{var}(b) = (X'X)^{-1}(\sum_{i=1}^{n} e_i^2 x_i x_i')(X'X)^{-1} \tag{10}$$

Deze methode voor het schatten van de covariantie matrix van b word de *White estimate* genoemd. De wortel van de diagonaal elementen van $\hat{var}(b)$ worden de *White standaard fouten* genoemd (Heij et all, 2004).

2. Normaliteit
De aanname voor normaliteit impliceert dat de errortermen ε_i normaal en onafhankelijk verdeeld zijn met een gemiddelde nul en constante variantie σ^2. Dit kan worden geschreven als in formule 11.

$$\varepsilon \sim NID(0, \sigma^2) \qquad (i = 1, ..., n) \tag{11}$$

Deze aanname kan worden gecontroleerd aan de hand van een histogram van de residuen. Deze histogram geeft een grafische weergave van de verdeling van de residuen. Daarnaast wordt ook een *Jarque-Bera toets* uitgevoerd. Deze *goodness-of-fit* test kijkt of de *skweness* en de *kurtosis* correspondeert met die van de normale verdeling. Als dit niet het geval is, dan is de OLS methode niet efficiënt.

Daarnaast wordt ook een QQ-plot geconstrueerd. Dit is een verdelingsplot waarmee grafisch twee kansverdelingen met elkaar worden vergeleken door hun kwantielen tegen elkaar te plotten. Als de punten de 45 graden lijn volgen kan worden geconcludeerd dat de residuen normaal verdeeld

zijn.

3. Correcte specificatie
De aanname voor correcte specificatie impliceert dat de data Y_i worden gegenereerd volgens een lineaire trend, zoals in formule 12. Hierin is Y de waarde van de afhankelijke variabele, X een matrix van de verklarende variabelen, β de bijbehorende parameter en ε de errorterm.

$$Y = \beta X + \varepsilon \tag{12}$$

Als dit niet het geval is, is er sprake van misspecificatie en is het model niet bruikbaar. Om hieraan te voldoen wordt een transformatie van de afhankelijke data toegepast. Als afhankelijke variabele wordt de *log(marktwaarde)* gebruikt in plaats van de *marktwaarde*. Als het model is geconstrueerd kan met behulp van een dotplot van de residuen deze aanname worden gecontroleerd. In deze dotplot mag geen sprake zijn van een trend of een patroon. Als dat het geval is, kan worden geconcludeerd dat het model juist gespecificeerd is.

4.4. Verbetering van voorgaande modellen

Dit onderzoek is een vervolg van het werkcollege onderzoek. Hierin is met dezelfde data een model geconstrueerd voor alle veldspelers. Hierin wordt de dataset opgesplitst in twee aparte datasets, één met alle observaties van spelers van 23 jaar en jonger en één met alle observaties van spelers van 30 jaar en ouder. Er wordt gekeken of het opsplitsen van de datasets leidt tot betere modellen of niet. Verschillende manieren komen aan bod.

Met behulp van een *in-sample fit* kunnen de twee nieuwe modellen het beste worden vergeleken met de oude modellen. Allereerst worden de gefitte waarden bepaald voor de jonge spelers met het jonge spelermodel en de gefitte waarden voor de oude spelers met het oude spelers model. Vervolgens worden ook de gefitte waarden van de jonge en oude spelers bepaald met behulp van het algemene veldspelers model uit het werkcollege. Om een te kijken of er een verbetering heeft plaatsgevonden tussen het algemene veldspelermodel en de nieuwe modellen, worden de mean squared errors (MSE) vergeleken. De MSE wordt bepaald door formule 13.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{13}$$

De MSE is het gemiddelde van de errors in het kwadraat. In formule 13 is n het aantal observaties, y_i de waarde van de afhankelijke variabele i en $\hat{Y}_i$ de gefitte waarde van de in-sample fit. De MSE is een goede maatstaaf voor het vergelijken van de modellen. Het is niet mogelijk om de R^2 van de verschillende modellen met elkaar te vergelijken, omdat de modellen uit dit onderzoek zijn gebaseerd op andere data dan het model uit het werkcollege.
Daarnaast wordt ook gekeken naar de toe en/of afname van het aantal verklarende variabelen. Ook wordt er gekeken naar de aannames van de OLS methode zoals: homoskedasticiteit, specificatie en normaliteit.

5. Resultaten

5.1. Kruskal-Wallistoets

Om te concluderen of er een verschil bestaat in de verdelingen van de leeftijd van de verschillende competities, is een **Kruskal-Wallistoets** toegepast. Deze toets wordt gebruikt om de rangorde

scores van meerdere groepen te vergelijken. Er wordt bij deze toets geen aanname gedaan over de verdeling van de verschillende databases. Deze methode wordt gebruikt om onderzoeksvraag 5 te beantwoorden, die gaat over het verschil in verdeling van de leeftijd van de verschillende competities.

Allereerst zijn rangnummers toegekend aan de database met de leeftijden van alle spelers. Vervolgens wordt de Kruskal-wallistoets uitgevoerd. De nulhypothese van deze toets zegt dat de 8 competities dezelfde leeftijdsverdeling hebben. De kritieke waarde KW wordt bepaald door formule 14.Deze mag worden toegepast omdat de steekproef groot genoeg is.

$$KW = \frac{12}{(n(n+1))} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(n+1) \tag{14}$$

Hierin is R_i^2 de som van de rangnummers uit steekproef i in het kwadraat, n_i het aantal observaties in steekproef i en n het totaal aantal observaties. m is het aantal steekproeven; in dit geval acht. De Kruskal-Walliswaarde KW is $X^2(m-1)$ verdeeld.

De database die wordt gebruikt kent acht competities en dus acht verschillende steekproeven. De Kruskal-Walliswaarde is gelijk aan 215.432. Deze waarde is $X^2(7)$ verdeeld. De bijbehorende kritieke waarde is 14.067. Dit betekent dat de nulhypothese wordt verworpen. Hieruit kan worden geconcludeerd dat er en verschil bestaat in de leeftijdsverdeling van verschillende competities.

Om vast te kunnen stellen wat die verschillen in leeftijd behelzen, worden de medianen van de leeftijd van de verschillende competities met elkaar vergeleken. Deze zijn te vinden in figuur 1 in de Appendix.

De competitie met de hoogste mediaan leeftijd is Serie A. Deze competitie wordt in de figuur aangegeven met een gele kleur en heeft een leeftijdsmediaan van 27 jaar. De competitie met de laagste leeftijdsmediaan is de Eredivisie. Deze competitie wordt ook in de figuur met een gele kleur aangegeven en heeft een mediaan leeftijd van 23 jaar. De rest van de competities liggen daartussen. Geconstateerd wordt dat spelers in de Eredivisie, Oostenrijkse Bundesliga en Jupiler Pro League relatief jonge spelers hebben. Een verklaring hiervoor zou kunnen zijn dat deze clubs hoogstwaarschijnlijk niet het geld hebben om dure spelers te kopen en zij leggen daarom de nadruk op de jonge talentvolle spelers. Naarmate de leeftijd van deze spelers vordert, worden zij verkocht aan andere clubs. De competities Primera Division, Premier Leage, Seria A, Bundesliga Duitsland en Ligue 1 geldt dat ze worden gezien als de betere competities van Europa. Opvallend is dat voor deze competities geldt dat de mediaan leeftijd rond de optimale leeftijd van ongeveer 26 ligt.

Met het bepalen van de mediaan is een verdelingsvrije vergelijking gedaan, maar deze zegt niets over de standaardafwijking σ. Daarvoor wordt de mediane absolute deviatie (MAD) gebruikt. De verschillende waarden van de MAD zijn ook te vinden in figuur 1 in de Appendix. In dit tabel is te zien dat de Jupiler League en Ligue 1 een MAD waarde hebben van 4 en de overige competities een MAD waarde van 3. Hieruit kan worden geconcludeerd dat de Jupiler League en Ligue 1 een grotere spreiding hebben in de leeftijden van de spelers dan de overige competities.

Om de spreiding verder te analyseren wordt ook de *Interquantile Range* (IQR) bepaald. Dit wordt gedaan aan de hand van het 25% percentiel en het 75% percentiel. De waarden van de IQR voor alle competities is te vinden in figuur 1 in de Appendix. Hierin is te zien dat de competitie

Ligue 1 een grote spreiding heeft met een IQR waarde van 8. De competitie Premier League daarentegen heeft een hele lage spreiding met een IQR waarde van 5.

Deze verdelingsvrije statistieken geven inzicht in de verdeling van de leeftijd van de spelers in de acht verschillende competities.

5.2. Modellen

Er zijn twee modellen geconstrueerd om de *log(marktwaarde)* van voetbalspelers in Europees voetbal te verklaren. Het eerste model verklaart de *log(marktwaarde)* voor jonge spelers van 23 jaar en jonger en het tweede model voor oude spelers van 30 jaar en ouder. Om multicolineariteit te voorkomen, is bij de afkomst van de spelers *isEurope* en bij de competities van de spelers *Premier_League* als basis genomen. Er zijn drie dummy's die de positie van spelers aangeven: *isdef*, *isstriker* en *ismid*. Hierbij is *isstriker* als basis genomen.

5.2.1 Model voor jonge spelers

Het model voor de jonge spelers ziet eruit zoals in formule 15. De waarden van de parameters zijn te vinden in figuur 2 in de appendix.

$$\begin{aligned} log(y_i) = {} & \hat{\beta_1} + \hat{\beta_2} monts\_at\_club + \hat{\beta_3} months\_to\_go + \hat{\beta_4} assists2013 \\ & + \hat{\beta_5} assists2012 + \hat{\beta_6} bundesliga\_duit + \hat{\beta_7} bundesliga\_oost + \hat{\beta_8} eredivisie \\ & + \hat{\beta_9} jupiler + \hat{\beta_{10}} CL + \hat{\beta_{11}} EL + \hat{\beta_{12}} height \\ & + \hat{\beta_{13}} inlastyear + \hat{\beta_{14}} issamer + \hat{\beta_{15}} matches2013 + \hat{\beta_{16}} matches2012 \\ & + \hat{\beta_{17}} minutes2013 + \hat{\beta_{18}} sommin + \hat{\beta_{19}} stadiongrootte \end{aligned} \tag{15}$$

Om het model te construeren zijn eerst alle beschikbare variabelen aan het model toegevoegd. Vervolgens zijn met behulp van *backward elimination* de insignificante variabelen één voor één verwijderd. Hierbij is na elke verwijdering een afweging gemaakt tussen het optimaliseren van de R^2 en het minimaliseren van de Akaike en Schwarz criteria om zo de fit en de efficiëntie van het model te optimaliseren.

Na het vinden van het meest efficiënte model op basis van backward elimination is het model verder geoptimaliseerd door de aannames van de OLS methode te onderzoeken.

Allereerst is de aanname voor **homoskedasticiteit** getoetst. Dit is gedaan aan de hand van een *Breuch-Pagan toets* en een *White toets*. Beide toetsen werden verworpen op een significantieniveau van 5%. Voor de Breuch-Pagan geldt dat er een p-waarde is van 0.0000. Deze is $X^2(18)$ verdeeld. Voor de White toets geldt dat er een p-waarde is van 0.0000. De uitslag van de *Breuch-Pagan toets* is te vinden in figuur 4 en die van de *White-toets* in figuur 5 in de appendix. Dit impliceert dat er sprake is van heteroskedasticiteit. Om deze reden worden de *white standaardfouten* toegepast.

In figuur 6 is een histogram te zien van de residuen van de data voor het jonge model. Deze lijken **normaal verdeeld**. De verwachting en de skewness zijn bijna gelijk aan nul. De kurtosis is relatief hoog, 3.4, ten opzichte van die van de normale verdeling. Dit is de rede dat de *Jarque-Bera toets*

wordt verworpen. Om dit verder te onderzoeken wordt gekeken naar een QQ-plot van de data. Deze is te vinden in figuur 7 in de appendix. In deze methode worden de kwantielen van twee verschillende kansverdelingen tegen elkaar geplot. Uit deze plot kan worden geconcludeerd dat voor het grootste deel van de data geldt dat ze de normale verdeling volgen. Alleen in de lage waarden is een afwijking te zien. Het is daarom belangrijk om voorzichtig om te gaan met het interpreteren van de hele lage *log(marktwaarde)*.

Uit figuur 8 kan worden geconcludeerd dat er sprake is van **correcte specificatie**. Op de x-as van figuur 8 staan de gesorteerde gefitte waarden. Opvallend in figuur 8 zijn de lineaire lijnen. Dit kan worden verklaard door het feit dat bepaalde waarden van de afhankelijke variabele vaker voorkomen.

In figuur 3 is een correlatiematrix weergegeven van alle variabelen uit het jonge model. Hierin is te zien dat er een sterke correlatie (-0.7588) bestaat tussen de variabelen *inlastyear* en *monthstogo*. Omdat beide variabelen een sterke verklaringskracht hebben worden beide variabelen in het model opgenomen.

Het meest optimale model voor het schatten van de *log(marktwaarde)* van de jonge spelers heeft een R^2 van 0.74 en een Akaike en Schwarz criteria van 0.597 en 0.683 respectievelijk.

5.2.2 Model voor oude spelers

Op dezelfde manier als bij het model voor de jonge spelers is een model voor de oude speler geconstrueerd. Na het toepassen van *backward elimination* ziet het model voor de oude spelers eruit als in formule 16. De waarden van de parameters van formule 16 staan weergegeven in de appendix in figuur 9. Voor dit model geldt dat niet alle coëfficiënten significant zijn op een niveau van 5%. Dit is het geval bij de variabelen *goals2013* en *minutes2011*. Er is ervoor gekozen om deze variabelen toch toe te voegen aan het model. Zoals genoemd in de inleiding is het voornaamste doel van dit onderzoek om een duidelijke interpretatie te geven van de statistieken van de spelers. Bij het weglaten van de variabelen *goal2013* en *minutes2011* is dit niet mogelijk. Dit heeft wel als gevolg dat een voorspelling met dit model iets minder nauwkeurig is.

$$\begin{aligned} log(y_i) = \hat{\beta}_1 &+ \hat{\beta}_2 leeftijd + \hat{\beta}_3 assists2013 + \hat{\beta}_4 bundesliga\_oos + \hat{\beta}_5 bundesliga\_duit \\ &+ \hat{\beta}_6 eredivisie + \hat{\beta}_7 jupiler + \hat{\beta}_8 ligue1 + \hat{\beta}_9 primera\_div + \hat{\beta}_{10} serieA \\ &+ \hat{\beta}_{11} CL + \hat{\beta}_{12} EL + \hat{\beta}_{13} goals2013 + \hat{\beta}_{14} goals2012 + \hat{\beta}_{15} isdef \\ &+ \hat{\beta}_{16} issamer + \hat{\beta}_{17} minutes2013 + \hat{\beta}_{18} minutes2012 + \hat{\beta}_{19} minutes2011 \\ &+ \hat{\beta}_{20} minutes2010 + \hat{\beta}_{21} noyeartogo + \hat{\beta}_{22} sommin \\ &+ \hat{\beta}_{23} stadiongrootte + \hat{\beta}_{24} WK + \hat{\beta}_{25} waschamply \end{aligned} \tag{16}$$

Voor een verdere optimalisatie van het model worden de aannames van OLS bekeken.

Allereerst is de aanname voor **homoskedasticiteit** getoetst. Dit is gedaan aan de hand van

een *Breuch-Pagan toets* en een *White toets*. Voor de Breuch-Pagan geldt dat er een p-waarde is van 0.1388. Deze is $X^2(24)$ verdeeld. Voor de White toets geldt dat er een p-waarde is van 0.0350. De uitslag van de *Breuch-Pagan toets* is te vinden in figuur 11 en die van de *White-toets* in figuur 12 in de appendix.Dit impliceert dat er, in tegenstelling met het jonge model, geen sprake is van heteroskedasticiteit. Ook een grafische weergave van de residuen in figuur 15 in de appendix laat zien dat er in deze data sprake is van homoskedasticiteit. Om deze rede is het niet nodig om gebruik te maken van White standaard fouten. In figuur 15 zijn lineaire lijnen zichtbaar. Dit kan worden verklaard door het feit dat bepaalde waarden van de afhankelijke variabele vaker voorkomen.

In figuur 13 is een histogram van de residuen van de oude spelers te zien. Hieruit kan worden geconcludeerd dat de data **normaal verdeeld** zijn. De verwachting van het gemiddelde en de skweness liggen rond de waarde nul en de Kurtosis ligt rond de waarde drie. Ook de *Jarqua-Bera toets* wordt niet verworpen, wat inhoudt dat de residuen de normale verdeling volgen. Ook een grafische weergave van de kwantielen van de residuen bevestigen deze aanname. Deze is te zien in de QQ-plot in figuur 14 in de appendix. Hierin is te zien dat de geplotte quantielen nauwelijks van de 45 graden lijn afwijken.

Uit plot 15 kan worden geconcludeerd dat er sprake is van **correcte specificatie**. Op de x-as van figuur 15 staan de gesorteerde gefitte waarden van de oude spelers.

In figuur 10 is een correlatiematrix weergegeven van alle variabelen uit het oude model. Hierin zijn geen uitzonderlijk hoge waarden te herkennen. Geen van de variabelen vertonen een sterke correlatie.

Het meest optimale model voor het schatten van de *log(marktwaarde)* van de oude spelers heeft een R^2 van 0.83 en een Akaike en Schwarz criteria van -0.336 en -0.145 respectievelijk.

5.3. Verschillen tussen de modellen

Er zijn enkele verschillen en overeenkomsten tussen het model dat de *log(marktwaarde)* voor de jonge spelers verklaart en het model dat de *log(marktwaarde)* voor de oude spelers verklaart. De modellen voor de jonge en oude spelers hebben een groot aantal overlappende variabelen: *assists2013, bundesliga_oos, bundesliga_duit, eredivisie, jupiler, CL, EL, issamer, minutes2013, sommin en stadiongrootte*. Daarnaast zijn er ook enkele verklarende variabelen die alleen in het jonge spelers model significant zijn (*monthsatclub*, *monthstogo*, *assists2012*, *height*, *inlastyear*, *matches2013* en *machtches2012*) en een aantal verklarende variabelen die alleen in het oude spelers model significant zijn (*age, ligue 1, primera_div, serie_A, goals2013, goals2012, isdef, minutes2012, minutes2011, minutes2010, noyearstogo, wk* en *waschamply*). Hieronder wordt een uitgebreidere beschrijving gegeven van enkele opvallende overeenkomsten en verschillen.

Het aantal verklarende variabelen is voor beide modellen verschillend. Het jonge spelers model heeft 18 verklarende variabelen en het oude spelers model heeft 24 verklarende variabelen. Bij het oude spelers model is te zien dat variabelen uit het verleden wel significant zijn. Het aantal minuten gespeeld gaat in dit model tot het jaar 2010 terug. In het jonge spelers model gaan de variabelen uit voorgaande jaren maar terug tot 2012 (*matches2012* en *assists2012*). Dit is een logisch resultaat, omdat er meer variabelen uit het verleden bekend zijn voor de oudere spelers dan voor de jonge spelers tot 23 jaar, omdat zij niet langer dan een aantal jaar op dit niveau voetballen.

In het oude spelers model is te zien dat alle competities significant zijn. Als basis is hier de Premier League genomen. De parameters moeten dus geïnterpreteerd worden ten opzichte van de Premier League. In figuur 9 is te zien dat alle parameters van de competities negatief zijn ten opzichte van de Premier League. Hiervoor geldt dat de competitie Primera Division het dichtst bij de Premier League ligt en de Oostenrijkse Bundesliga het verst van de Premier League ligt. De volgende conclusie kan hieruit worden getrokken: Oude spelers uit de Premier League hebben relatief de hoogste marktwaarde ten opzichte van spelers uit alle andere competities, indien alle andere verklarende variabelen gelijk blijven. Ook geldt dat oude spelers uit de Oostenrijkse Bundesliga relatief de laagste marktwaarde hebben ten opzichte van de Premier League, als alle andere verklarende variabelen gelijk blijven.

Voor het jonge spelers model geldt dat alleen de Duitse Bundesliga, de Oostenrijkse Bundesliga, de Eredivisie en de Jupiler League significante variabelen zijn ten opzichte van de Premier League. Hieruit kan worden geconcludeerd dat alleen deze competities een significant verschil vertonen in de *log(marktwaarde)* ten opzichte van de Premier League. De bijbehorende parameters zijn te vinden in figuur 2 in de appendix. Hierin is te zijn dat voor alle vier de competities de parameters negatief zijn. Hieruit kan worden geconcludeerd dat voor de jonge spelers geldt dat spelers uit deze vier competities een relatief lagere marktwaarde hebben dat spelers uit de Premier League. Voor de jonge spelers geldt dat de *log(marktwaarde)* van spelers uit de Jupiler League het dichtst bij de *log(marktwaarde)* van spelers uit de Premier League ligt en voor spelers uit de Oostenrijkse Bundesliga het verst weg. De volgende conclusie kan hieruit worden getrokken: Jonge spelers uit de Premier League hebben relatief de hoogste *log(marktwaarde)* ten opzichte van spelers uit de andere vier competities, indien alle andere verklarende variabelen gelijk blijven. Ook geldt dat van de vier competities, oude spelers uit de Oostenrijkse Bundesliga relatief de laagste *log(marktwaarde)* hebben ten opzichte van de Premier League, als alle andere verklarende variabelen gelijk blijven. Omdat niet alle competities significant zijn in het model is het moeilijk een conclusie hieruit te trekken.

De R^2 van de twee modellen wordt tot slot vergeleken. De R^2 van het oude spelers model (0.83) is aanzienlijk hoger dan die van het jonge spelers model (0.74). Dit is te verklaren door het feit dat er meer statistieken uit het verleden beschikbaar zijn voor de oude spelers. De variabelen die zijn gebruikt gaan tot 6 jaar terug. Deze zijn in veel gevallen wel beschikbaar voor de oude spelers en niet voor de jonge spelers. Hierdoor is de marktwaarde van de oude spelers nauwkeuriger te verklaren dan die van de jonge spelers en dit is terug te zien in de hoogte van de R^2.

5.4. Vergelijking met het algemene model

In deze sectie worden de twee modellen vergeleken met het algemene model uit het werkcollege. De output van dit model is te vinden in figuur 16 in de appendix. Voor deze vergelijking wordt het algemene veldspeler model gebruikt, omdat het model voor de jonge en oude spelers ook alleen gebaseerd is op de statistieken van veldspelers en niet van keepers.

Voor beide modellen geldt dat de Champions League *(CL)* en de Europa League *(EL)* verklarende variabelen zijn. Dit is ook terug te zien in het algemene model voor veldspelers in het werkcollege onderzoek. Hieruit kan worden geconcludeerd dat de Champions League en de Europa League significante variabelen zijn ongeacht de leeftijd. Daarnaast is ook te zien dat voor beide modellen geldt dat de coëfficiënt voor *CL* hoger is dan die van *EL*. Dit impliceert dat spelers uit de Chapions League een hogere *log(marktwaarde)* hebben dan spelers uit de Europa League. Dit is een logisch

resultaat, omdat de Champions League een hoger toernooi is dan de Europa League.

De variabelen $leeftijd$ en $leeftijd^2$ waren belangrijke verklarende variabelen in het algemene veldspelermodel uit het werkcollege. In de twee modellen van dit onderzoek geldt dat dat niet het geval is. Voor het jonge spelers model zijn deze twee variabelen beide niet significant. Een verklaring hiervoor is dat veel spelers in deze categorie dezelfde leeftijd hebben, omdat de database is gesplitst op leeftijd. Dit heeft als gevolg dat leeftijd niet meer een significante variabele is. Leeftijd is bij de jonge spelers niet van significante invloed bij de voorspelling op de *log(marktwaarde)*. Bij het oude spelers model geldt dat alleen de variabele *leeftijd* een significantie variabele is met een negatieve coëfficiënt. Omdat hier over het algemeen alleen een daling geldt van de marktwaarde naarmate de speler ouder wordt, is het niet nodig om ook $leeftijd^2$ op te nemen in het model. Het negatieve effect wordt dus alleen verklaard door $leeftijd$.

In het algemene veldspelers model uit het werkcollege is te zien dat de *lengte* van spelers wel van significantie invloed is. Opvallend is dat in de gesplitste modellen van dit onderzoek deze variabele alleen terug te zien is in het jonge spelers model. De coëfficiënten die bij deze variabele hoort is positief. Dit impliceert dat voor jonge spelers geldt dat ze een relatief hogere *log(marktwaarde)* hebben als ze langer zijn.

De continentale afkomst van de voetbalspelers is ook meegenomen bij het construeren van een model. Hierbij is Europa als basis genomen. Opvallend is dat alleen de afkomst Zuid Amerika een significante invloed heeft. Dit geldt voor zowel de jonge als de oude spelers. Dit kan worden verklaard door het feit dat in het continent Zuid Amerika relatief sterke voetballanden gevestigd zijn en vergelijking met andere continenten. Landen als Brazilië, Colombia, Uruguay en Argentinië komen voor in de top tien van de FIFA World ranking (www.fifa.com, mei 2014).

Voor het oude spelers model geldt dat de variabele *isdef* een significante invloed heeft. Dit is een logisch resultaat. Voetbalspelers moeten fit en snel zijn, maar voor oude verdedigers kan dit gecompenseerd worden door ervaring. Op deze manier kunnen verdedigers langer doorspelen ten opzichte van aanvallers. Een voorbeeld hiervan is Jaap Stam, die tot z'n 35ste doorspeelde als verdediger bij Ajax, (www.wikipedia.nl, 12 juni 2014).

De ervaringsvariabele *sommin*, het totaal aantal minuten gespeeld in het verleden, is significant in alle drie de modellen (jonge spelers model, oude spelers model, algemeen veldspeler model). Dit impliceert dat de ervaring van een veldspeler belangrijk is ongeacht de leeftijd. De coëfficiënt die bij deze variabele hoort is in alle drie de gevallen positief. Hieruit kan worden geconcludeerd dat de *log(marktwaarde)* van een speler toeneemt als het totaal aantal minuten gespeeld ook toeneemt indien alle andere variabelen gelijk blijven.

De variabele *stadiongrootte* is significant in alle drie de modellen (jonge spelers model, oude spelers model, algemeen veldspeler model). Deze variabele is een indicator voor de capaciteit van het stadion van de club waar de spelers op dit voor spelen. In alle drie de modellen geldt dat de bijbehorende parameter positief is wat impliceert dat de *log(marktwaarde)* relatief toeneemt als de speler bij een club speelt met een grotere stadioncapaciteit.

5.5. Verbetering ten opzichte van eerder onderzoek

Om de nieuwe modellen van de jonge en oude spelers te vergelijken met het oude veldspelermodel uit het werkcollege, wordt een in-sample fit gebruikt. Allereerst worden de gefitte waarden van de jonge spelers bepaald met het jonge spelers model. Hiervan wordt de MSE bepaald. Deze is gelijk aan 0.3205. Vervolgens worden de gefitte waarden van de jonge spelers bepaald met het algemene veldspelermodel uit het werkcollege. Ook hiervan wordt de MSE bepaald. Deze is gelijk aan 0.7788. Deze waarden zijn ook terug te vinden in figuur 19 en figuur 18. Hieruit kan worden geconcludeerd dat de MSE waarde van de gefitte waarden geconstrueerd met het nieuwe model voor de jonge spelers veel lager is dan de MSE waarde van de gefitte waarden geconstrueerd met het algemene veldspelermodel. Het nieuwe model voor jonge spelers kan dus worden beschouwd als een verbetering van het model uit het werkcollege.

Daarnaast zijn de residuen van het oude spelers model homoskedastisch terwijl de residuen van het algemene veldspelers model heteroskedastisch zijn. Daarnaast geldt ook dat de residuen normaal verdeeld zijn in het oude spelers model. De Jarque-Bera test wordt hier niet verworpen. Dit is wel het geval bij het algemene veldspelers model. Dit impliceert dat de data die zijn gebruikt voor het oude spelers model geschikter zijn een OLS schattings methode dan de data die zijn gebruikt voor het algemene veldspelermodel.

Het aantal verklarende variabelen in het oude veldspelers model (24 variabelen) is minder dan het aantal verklarende variabele dat wordt gebruikt in het algemene veldspelers model (26 variabelen). Hieruit kan worden geconcludeerd dat het model uit dit onderzoek efficiënter is.

Hetzelfde wordt vervolgens gedaan voor de oude spelers. Eerst worden de gefitte waarden van de oude spelers bepaald met het oude spelers model. Hiervan wordt de MSE bepaald. Deze is gelijk aan 0.1958. Vervolgens worden de gefitte waarden van de oude spelers bepaald met het algemene veldspelermodel uit het werkcollege. Ook hiervan wordt de MSE bepaald. De data 2744 tot 3328 worden hiervoor gebruikt, zodat de MSE alleen wordt gebaseerd op spelers van 30 jaar en ouder. Deze is gelijk aan 0.5381. Deze waarden zijn ook terug te vinden in figuur 21 en figuur 20. Hieruit kan worden geconcludeerd dat de MSE waarde van de gefitte waarden geconstrueerd met het nieuwe model voor de oude spelers veel lager is dan de MSE waarde van de gefitte waarden geconstrueerd met het algemene veldspelermodel. Het nieuwe model voor oude spelers kan dus worden beschouwd als een verbetering van het model uit het werkcollege.

Voor dit model geldt bovendien dat er veel minder verklarende variabelen zijn. Het jonge spelers-model bevat 18 verklarende variabelen, terwijl het algemene veldspelers model 26 variabelen bevat.

Voor de dataset van het jonge spelers model geldt dat de residuen een veel lagere kurtosis hebben dan de residuen van de dataset voor het algemene spelers model. De data van de jonge spelers volgt dus beter de normale verdeling dan de data van alle veldspelers. Dit heeft als gevolg dat de dataset van de jonge spelers geschikter is voor een OLS schattings methode van de jonge spelers dan de dataset voor alle spelers voor het algemene veldspelermodel.

Als de QQ plot (figuur 7) van de residuen van het jonge spelers model wordt vergeleken met de QQ plot van de residuen van het algemene veldspelermodel uit het werkcollege onderzoek (figuur 17), is een opvallend verschil te zien. Bij het jonge spelers model is te zien dat de residuen alleen bij de lage waarden afwijken van de 45 graden lijn, terwijl bij het algemene veldspelers model

ook de residuen bij de hoge waarden afwijken. Dit betekent dat bij het jonge spelers model alleen voorzichtig gedaan hoeft te worden met het interpreteren van de lage marktwaarden, terwijl bij het algemene veldspelers model zowel bij de hoge als bij de lage marktwaarden voorzichtig gedaan moet worden met het interpreteren. Dit verschil zou verklaard kunnen worden door het feit dat voor jonge spelers geldt dat de marktwaarden nog geen extreme waarden aannemen, tenzij de statistieken er ook echt naar zijn. Dit impliceert dat het jonge spelers model beter de marktwaarde verklaart voor jonge spelers met een hele hoge marktwaarde, dan het algemene veldspelers model.

Over het algemeen geldt dat het opsplitsen van de datasets is twee verschillende leeftijdscategorieën voor een beter model zorgt dan een model dat is gebaseerd op 1 grote datasets.

6. Conclusie

Het werkcollege onderzoek heeft mijn interesse gewekt voor de beschreven materie. Daaropvolgend ontstond de behoefte om hetgeen geconcludeerd werd in het werkcollege nog nader te onderzoeken. In dit onderzoek zijn twee modellen geconstrueerd die inzicht geven over de log(marktwaarde) van jongere (tot en met 23 jaar) en oudere (vanaf 30 jaar) Europese voetbalspelers. Deze splitsing is gemaakt om een beter inzicht te krijgen van relevante statistieken bij het verklaren van de marktwaarde voor deze twee groepen. Op deze manier wordt geprobeerd om het algemene veldspelermodel uit het werkcollege te verbeteren. De twee uiteindelijke modellen zijn te vinden in formule 15 en 16. Voordat deze modellen zijn geconstrueerd is eerst een analyse gedaan op de datasets met behulp van een Kruskal-Wallistoets. Deze twee modellen zijn vervolgens vergeleken met een algemeen velspeler model uit een eerder onderzoek. De onderzoeksvragen worden achtereenvolgens behandeld.

Welke modellen verklaren het beste de marktwaarde voor oude en jonge spelers? Het model weergegeven in formule 15 geeft de beste verklaring van de marktwaarde voor jonge spelers. De bijbehorende coëfficiënten zijn te vinden in figuur 2. Hierbij is gebruik gemaakt van White standaard fouten om voor de heteroskedasticiteit te compenseren. Het model weergegeven in formule 16 geeft de beste verklaring van de marktwaarde voor de oude spelers. De bijbehorende coëfficiënten zijn te vinden in figuur 9.

Wat zijn de grootste verschillen tussen de modellen voor jonge en oude spelers? De verschillen tussen de twee modellen worden beschreven onder de subtitel: *verschillen tussen de modellen*. Het grootste verschil is dat er bij de jonge spelers nog relatief weinig bekend is over statistieken uit het verleden. Dit resulteert in een grotere onzekerheid bij deze groep. Deze statistieken zijn wel bekend bij voor de oude spelers en daarom heeft deze groep een nauwkeuriger model. Dit is ook terug te zien in de resultaten.

Wat gebeurt er met de variabelen $leeftijd$ en $leeftijd^2$? De variabelen $leeftijd$ en $leeftijd^2$ zijn niet terug te zien in het jonge spelers model. Dit is een logische gevolg, omdat het niet meer nodig is om eerst een stijgend en daarna een dalend effect te verwerken.Voor de oude spelers geldt dat alleen $leeftijd$ terug te zien is in het model. Hier geldt: Hoe ouder de spelers worden des te lager de *log(marktwaarde)*.

Zijn ervaringsvariabelen relevant voor jonge en oude spelers? De ervaringsvariabele *sommin*, het totaal aantal minuten gespeeld, is zowel voor de jonge spelers als voor de oude spelers van significante invloed bij het verklaren van de marktwaarde. Dit is in overeenstemming met de

verwachting.

Wat is de leeftijdsverdeling bij de acht verschillende competities? De leeftijdsverdeling is niet bij elke competitie hetzelfde. Dit is aangetoond met behulp van een Kruskal-Wallistoets. Met een verdere analyse van de mediaan, MAD en IQR is vervolgens aangetoond dat de mediaan leeftijd bij Serie A het hoogst ligt en bij de Eredivisie het laagst. De spreiding is het grootst bij Ligue 1 en het kleinst bij de Premier League.

Zijn er competities waarbij jonge respectievelijk oude spelers een hogere marktwaarde hebben? De coëfficiënten van het oude model wijzen uit dat oude spelers uit de Premier League relatief de hoogste marktwaarde hebben en spelers uit de Oostenrijkse Premier League de laagste marktwaarde. Voor de jonge spelers is het niet mogelijk om hier een uitspraak over te doen, omdat niet alle competities significant zijn in het model.

Geeft het opsplitsen van de jonge en oude spelers een betere verklaren dan het algemene model uit het werkcollege? Zoals wordt beschreven onder de subtitel, *Verbetering ten opzichte van eerder onderzoek*, kan worden geconcludeerd dat de modellen die zijn geconstrueerd door het opsplitsen van de datasets, voor een beter resultaat zorgen dan modellen in eerder onderzoek. Dit resultaat is verkregen door een in-sample fit toe te passen en de MSE's te vergelijken.

Het opsplitsen van de databases is een goede oplossing voor het verbeteren van de modellen van het werkcollege. Met de modellen uit dit onderzoek kan een betere verklaring worden gedaan van de marktwaarde voor jonge en oude voetbalspelers in de acht Europese competities.

Met veel interesse en genoegen heb ik deze scriptie vervaardigd. Mijn dank gaat uit naar de begeleider van dit onderzoek, Dr. A. Alfons.

Referenties

[Breslow, 1970] N. Breslow, (1970), University of Washington A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship

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

[Dobson, Gerrard en Howe, 2000] S. Dobson, B. Gerrard en S. Howe (2000) The determination of transfer fees in English nonleague football Applied Economics, 32:9, 1145-115.

[Frick, 2007] B. Frick (2007) The Football Players Labor Market: Empirical Evidence from the Major European Leagues Scottish Journal of Political Economy, 54 (3)

[Heij et all, 2004] C. Heij, P. de Boer, P. H. Franses, T. Kloekand and H. K. van Dijk (2004). Econometric; Methods with Applications in Business and Economics H2 en H5.

[Monster, Ramsaroep en Mutsaerts, 2014] S. Monster, N. Ramsaroep en A. Mutsaerts (2014) Estimating the market value of football players: An analysis over te eight football leagues in Europe.

[Pham-Ghia en Hung, 2000] T. Pham-Gia en T.L. Hung, (2000) The mean and median absolute deviations

[Reilly en Witt, 1995] B. Reilly en R. Wit (1995) English league transfer prices: Is there a racial dimension? Applied Economics Letters, Volume 2, Issue 7, 1995

[Simmons, 1997] R. Simmons (1997) Implications of the Bosman ruling for football transfer markets Economic Affairs, Volume 17, Issue 3, pages 13âĂŞ18, September 1997

[Speight en Thomas, 1997] A. Speight en D. Thomas (1997) Football league transfers: a comparison of negotiated fees with arbitration settlements Applied Economics Letters, Volume 4, Issue 1, 1997

websites:

[www.nusport.nl, 2014]

[www.fifa.com, 2014]

[www.transfermarkt.com, 2014] Voetbalsite; oorsprong van de data.

[www.worldstdiumdatabase.com.]

[www.wikipedia.nl]

[www.sport.infonu.nl]

A. Appendix

	gemiddelde	mediaan	MAD	IQR
Primera Division	26.13	26	3	6
Ligue 1	25.28	25	4	8
Serie A	26.53	27	3	7
Eredivisie	23.46	23	3	6
Jupiler League	24.37	24	4	7
Premier League	26.45	26	3	5
Bundesliga Oostenrijk	24.16	24	3	6
Bundesliga Duitsland	24.67	25	3	6

Figuur 1: *In dit tabel staan de gemiddelde waarden, de medianen, de MAD en de IQR van de leeftijd van de verschillende competities.*

Dependent Variable: LOG_MV
Method: Least Squares
Date: 06/03/14 Time: 10:20
Sample (adjusted): 3 1270
Included observations: 1100 after adjustments
White heteroskedasticity-consistent standard errors & covariance

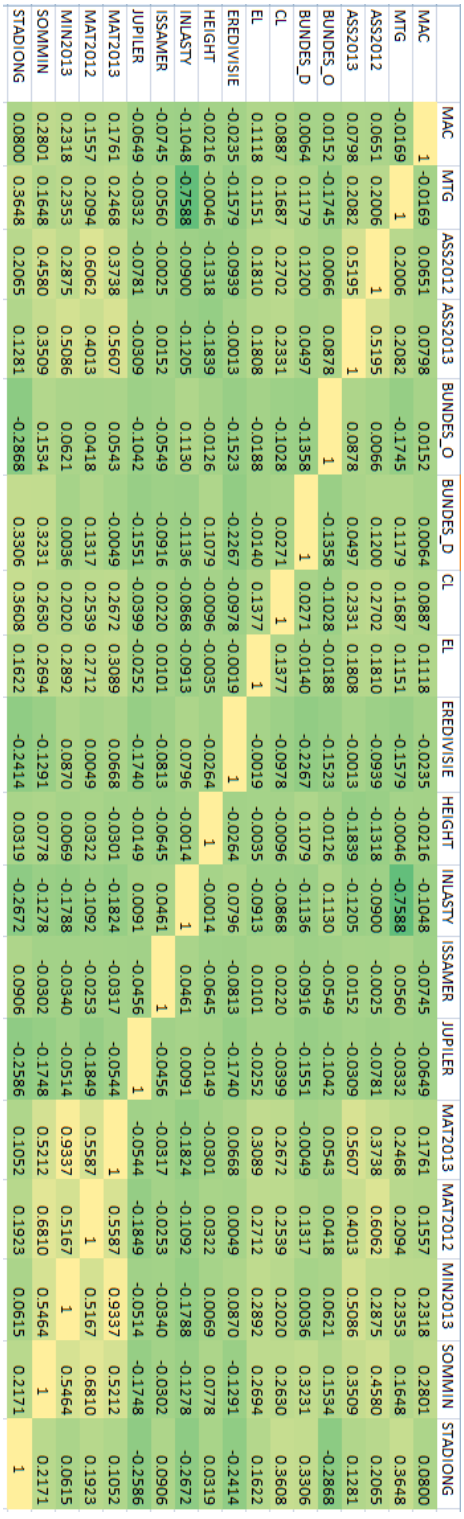
Variable	Coefficient	Std. Error	t-Statistic	Prob.
_MONTHS_AT_CLUB	0.001460	0.000696	2.097700	0.0362
_MONTHS_TO_GO	0.004348	0.001043	4.169854	0.0000
ASSISTS2013	0.015596	0.004971	3.137192	0.0018
ASSISTS2012	0.010391	0.004191	2.479159	0.0133
BUNDESLIGA_DUIT	-0.361342	0.038265	-9.443052	0.0000
BUNDESLIGA_OOS	-0.637805	0.035121	-18.16016	0.0000
EREDIVISIE	-0.382341	0.026841	-14.24465	0.0000
JUPILER	-0.175873	0.031367	-5.607001	0.0000
CL	0.220906	0.031691	6.970500	0.0000
EL	0.120564	0.021575	5.588243	0.0000
HEIGHT	0.004042	0.001601	2.523892	0.0117
INLASTYEAR	0.082999	0.034249	2.423413	0.0155
ISSAMER	0.295157	0.043940	6.717297	0.0000
MATCHES2013	0.022952	0.002846	8.065169	0.0000
MATCHES2012	0.006236	0.001141	5.467482	0.0000
MINUTES2013	-0.000104	3.52E-05	-2.963529	0.0031
SOMMIN	3.04E-05	3.63E-06	8.371596	0.0000
STADIONGROOTTE	4.42E-06	7.80E-07	5.671449	0.0000
C	-2.494777	1.556178	-1.603144	0.1092

R-squared	0.736104	Mean dependent var	3.074474
Adjusted R-squared	0.731710	S.D. dependent var	0.624266
S.E. of regression	0.323349	Akaike info criterion	0.596955
Sum squared resid	113.0237	Schwarz criterion	0.683371
Log likelihood	-309.3250	Hannan-Quinn criter.	0.629647
F-statistic	167.5173	Durbin-Watson stat	1.278042
Prob(F-statistic)	0.000000	Wald F-statistic	187.7484
Prob(Wald F-statistic)	0.000000		

Figuur 2: *In dit tabel is de Eviews output van het model van de jonge spelers te vinden.*

variabele	**omschrijving**
age	Geeft de leeftijd van de speler weer
age^2	De leeftijd^2 van de spelers
assists_x	Het aantal assists van een speler in jaar x
bundesliga_Duit	dummy variabele die 1 is als de speler in de Duitse bundesliga speelt
bundesliga_Oos	dummy variabele die 1 is als de speler in de Oostenrijkse bundesliga speelt
canuseboth	dummy variabele die 1 is als een speler beide voeten kan gebruiken
canuseleft	dummy variabele die 1 is als een speler linksvoetig is
canuseright	dummy variabele die 1 is als een speler rechtsvoetig is
cl	dummy variabele die 1 is als een speler ooit in de Champions League heeft gespeeld
ec	dummy variabele die 1 is als een speler ooit in de Euro Cup heeft gespeeld
el	dummy variabele die 1 is als een speler ooit in de Europa League heeft gespeeld
eredivisie	dummy variabele die 1 is als de speler in de Eredivisie speelt
goals_x	aantal goals gescoord door een speler in jaar x
goalsconceded_x	aantal goals tegengehouden door een speler in jaar x
height	de lengte van een speler
inlastyear	dummy variabele die 1 is als een speler in het laatste jaar van zijn contract zit
international_comp	dummy variabele die 1 is als een speler ooit in een internationale competitie heeft gespeeld met uitzondering van de Euro Cup of World Cup
isafrica	dummy variabele die 1 is als een speler in Afrika is geboren
isasia	dummy variabele die 1 is als een speler in Azie is geboren
isdef	dummy variabele die 1 is als een speler verdediger is
iseurope	dummy variabele die 1 is als een speler in Europa is geboren
isgoal	dummy variabele die 1 is als een speler een keeper is
ismid	dummy variabele die 1 is als een speler een middeveld speler is
isnamer	dummy variabele die 1 is als een speler in Noord Amerika is geboren
isocean	dummy variabele die 1 is als een speler in Oceanie is geboren
issamer	dummy variabele die 1 is als een speler in Zuid Amerika is geboren
isstriker	dummy variabele die 1 is als een speler aanvaller is
jupiler	dummy variabele die 1 is als de speler in de Jupiler League speelt
ligue1	dummy variabele die 1 is als de speler in de Ligue 1 speelt
matches_x	het aantal wedstrijden gespeeld in jaar x
minutes_x	het aantal minuten gespeeld in jaar x
months_at_club	het aantal maanden dat een speler al voor zijn huidige club speelt
months_to_go	het aantal maanden dat het nog duurt tot het contract bij de huidige club afloopt.
owngoals_x	het aantal eigen goals in jaar x
premier_league	dummy variabele die 1 is als de speler in de Premier League speelt
primara_division	dummy variabele die 1 is als de speler in de Primera Division speelt
red_x	aantal rode kaarten ontvangen in jaar x
serie_a	dummy variabele die 1 is als de speler in de Serie A speelt
sommin	totaal aantal minuten gespeeld in de cariere van de speler
stadiumsize	capaciteit van het stadion van de huidige club van de speler
waschamply	dummy variabele die 1 is als een speler kampioen was is het laatste jaar
wk	dummy variabele die 1 is als een speler ooit in de World Cup heeft gespeeld
yellow_x	aantal gele kaarten ontvangen in jaar x
yellowred_x	aantal rode kaarten ontvangen in jaar x door 2 gele kaarten

Tabel 1: *Dit tabel geeft een beschrijving van alle variabelen.*



	MAC	MTG	ASS2012	ASS2013	BUNDES_O	BUNDES_D	CL	EL	EREDIVISIE	HEIGHT	INLASTY	ISSAMER	JUPILER	MAT2013	MAT2012	MIN2013	SOMMIN	STADIONG
MAC	1	-0.0169	0.0651	0.0798	0.0152	0.0064	0.0887	0.1118	-0.0235	-0.0216	-0.1048	-0.0745	-0.0649	0.1761	0.1557	0.2318	0.2801	0.0800
MTG	-0.0169	1	0.2006	0.2082	-0.1745	0.1179	0.1687	0.1151	-0.1579	-0.0046	-0.7588	0.0560	-0.0332	0.2468	0.2094	0.2353	0.1648	0.3648
ASS2012	0.0651	0.2006	1	0.5195	0.0066	0.1200	0.2702	0.1810	-0.0939	-0.1318	-0.0900	-0.0025	-0.0781	0.3738	0.6062	0.2875	0.4580	0.2065
ASS2013	0.0798	0.2082	0.5195	1	0.0878	0.0497	0.2331	0.1808	-0.0013	-0.1839	-0.1205	0.0152	-0.0309	0.5607	0.4013	0.5086	0.3509	0.1281
BUNDES_O	0.0152	-0.1745	0.0066	0.0878	1	-0.1358	-0.1028	-0.0188	-0.1523	-0.0126	0.1130	-0.0549	-0.1042	0.0543	0.0418	0.0621	0.1534	-0.2868
BUNDES_D	0.0064	0.1179	0.1200	0.0497	-0.1358	1	0.0271	-0.0140	-0.2267	0.1079	-0.1136	-0.0916	-0.1551	-0.0049	0.1317	0.0036	0.3231	0.3306
CL	0.0887	0.1687	0.2702	0.2331	-0.1028	0.0271	1	0.1377	-0.0978	-0.0096	-0.0868	0.0220	-0.0399	0.2672	0.2539	0.2020	0.2630	0.3608
EL	0.1118	0.1151	0.1810	0.1808	-0.0188	-0.0140	0.1377	1	-0.0019	-0.0035	-0.0913	0.0101	-0.0252	0.3089	0.2712	0.2892	0.2694	0.1622
EREDIVISIE	-0.0235	-0.1579	-0.0939	-0.0013	-0.1523	-0.2267	-0.0978	-0.0019	1	-0.0264	0.0796	-0.0813	-0.1740	0.0668	0.0049	0.0870	-0.1291	-0.2414
HEIGHT	-0.0216	-0.0046	-0.1318	-0.1839	-0.0126	0.1079	-0.0096	-0.0035	-0.0264	1	-0.0014	-0.0645	-0.0149	-0.0301	0.0322	0.0069	0.0778	0.0319
INLASTY	-0.1048	-0.7588	-0.0900	-0.1205	0.1130	-0.1136	-0.0868	-0.0913	0.0796	-0.0014	1	0.0461	0.0091	-0.1824	-0.1092	-0.1788	-0.1278	-0.2672
ISSAMER	-0.0745	0.0560	-0.0025	0.0152	-0.0549	-0.0916	0.0220	0.0101	-0.0813	-0.0645	0.0461	1	-0.0456	-0.0317	-0.0253	-0.0340	-0.0302	0.0906
JUPILER	-0.0649	-0.0332	-0.0781	-0.0309	-0.1042	-0.1551	-0.0399	-0.0252	-0.1740	-0.0149	0.0091	-0.0456	1	-0.0544	-0.1849	-0.0514	-0.1748	-0.2586
MAT2013	0.1761	0.2468	0.3738	0.5607	0.0543	-0.0049	0.2672	0.3089	0.0668	-0.0301	-0.1824	-0.0317	-0.0544	1	0.5587	0.9337	0.5212	0.1052
MAT2012	0.1557	0.2094	0.6062	0.4013	0.0418	0.1317	0.2539	0.2712	0.0049	0.0322	-0.1092	-0.0253	-0.1849	0.5587	1	0.5167	0.6810	0.1923
MIN2013	0.2318	0.2353	0.2875	0.5086	0.0621	0.0036	0.2020	0.2892	0.0870	0.0069	-0.1788	-0.0340	-0.0514	0.9337	0.5167	1	0.5464	0.0615
SOMMIN	0.2801	0.1648	0.4580	0.3509	0.1534	0.3231	0.2630	0.2694	-0.1291	0.0778	-0.1278	-0.0302	-0.1748	0.5212	0.6810	0.5464	1	0.2171
STADIONG	0.0800	0.3648	0.2065	0.1281	-0.2868	0.3306	0.3608	0.1622	-0.2414	0.0319	-0.2672	0.0906	-0.2586	0.1052	0.1923	0.0615	0.2171	1

Figuur 3: *Dit is een correlatiematrix van variabelen uit het model voor de jonge spelers.*

Heteroskedasticity Test: Breusch-Pagan-Godfrey

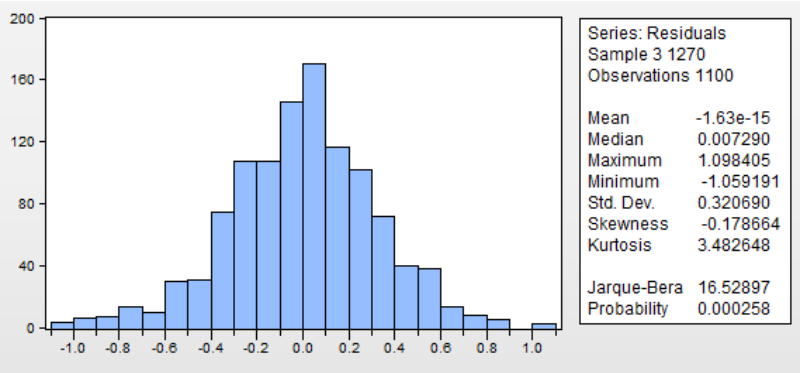
F-statistic	8.678115	Prob. F(18,1081)	0.0000
Obs*R-squared	138.8828	Prob. Chi-Square(18)	0.0000
Scaled explained SS	166.4945	Prob. Chi-Square(18)	0.0000

Figuur 4: *Dit is de uitslag van de Breuch-Pagan toets voor heteoskedasticiteit.*

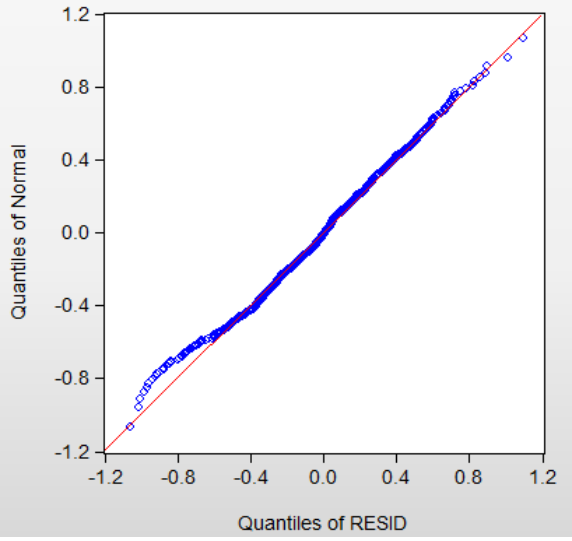
Heteroskedasticity Test: White

F-statistic	2.138126	Prob. F(175,924)	0.0000
Obs*R-squared	317.0530	Prob. Chi-Square(175)	0.0000
Scaled explained SS	380.0870	Prob. Chi-Square(175)	0.0000

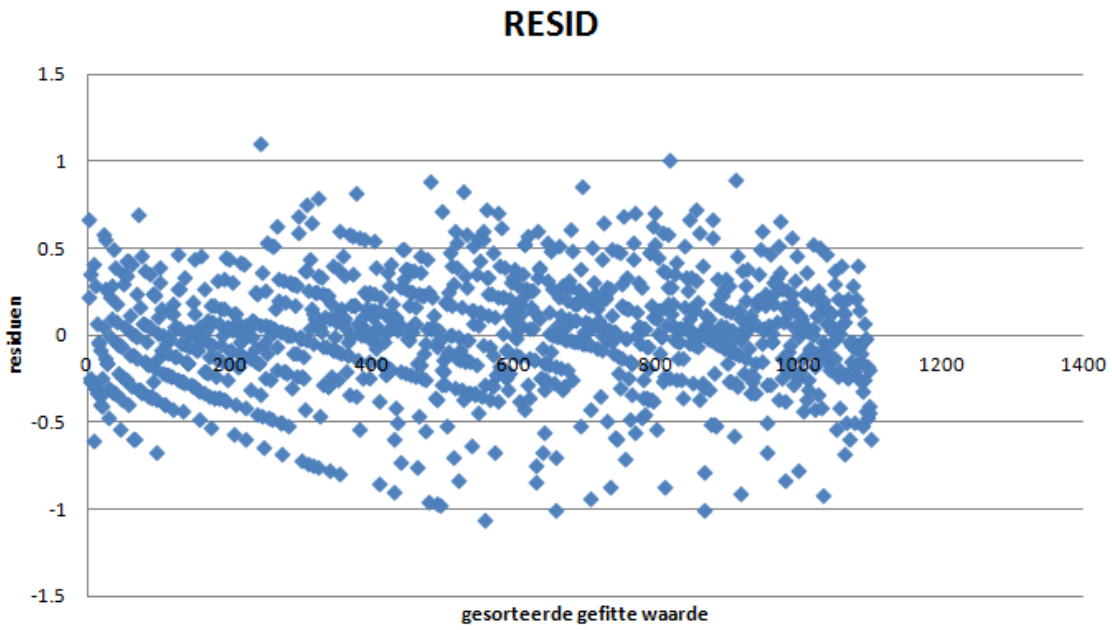
Figuur 5: *Dit is de uitslag van de White toets voor heteoskedasticiteit.*



Figuur 6: *Dit is een histogram van de residuen van de jonge spelers.*



Figuur 7: *Dit is een QQ-plot van de residuen van de jonge spelers.*



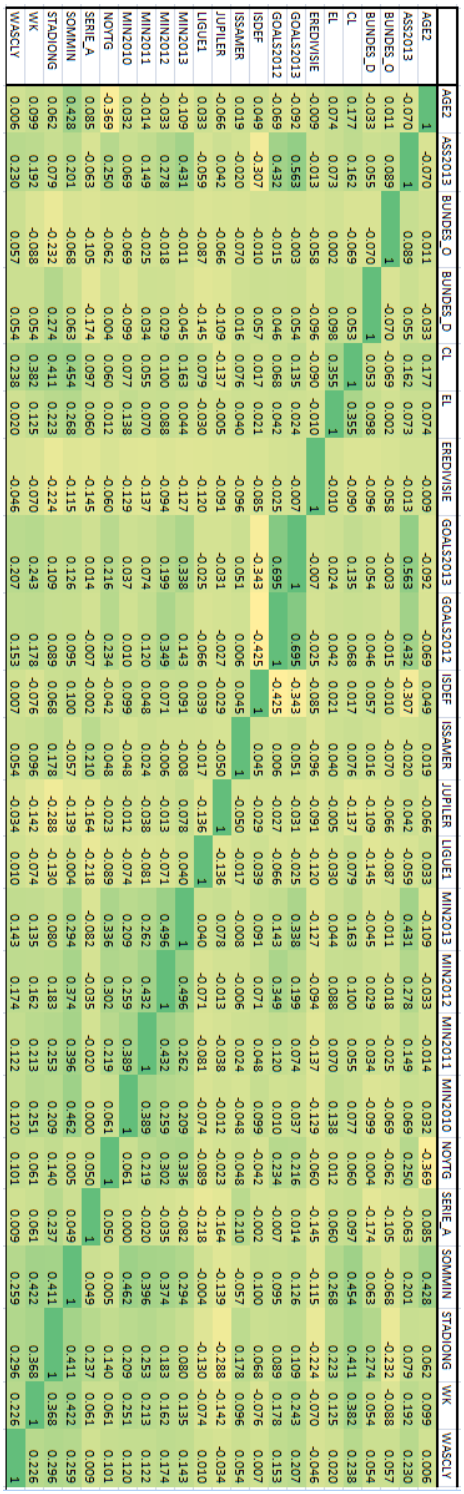
Figuur 8: *Dit is een dotplot van de residuen van de jonge spelers met op de x-as de gesorteerde gefitte waarden.*

Dependent Variable: LOG_MV
Method: Least Squares
Date: 06/03/14 Time: 12:37
Sample (adjusted): 2 585
Included observations: 568 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGE	-0.120171	0.006117	-19.64519	0.0000
ASSISTS2013	0.011690	0.005102	2.291273	0.0223
BUNDESLIGA_OOS	-0.557516	0.050204	-11.10496	0.0000
BUNDESLIGA_DUIT	-0.141404	0.035538	-3.978990	0.0001
EREDIVISIE	-0.296392	0.039806	-7.445895	0.0000
JUPILER	-0.417008	0.037362	-11.16126	0.0000
LIGUE1	-0.188402	0.031367	-6.006322	0.0000
PRIMERA_DIV	-0.066710	0.032587	-2.047110	0.0411
SERIE_A	-0.108411	0.029388	-3.688947	0.0002
CL	0.075028	0.021940	3.419670	0.0007
EL	0.067985	0.018725	3.630816	0.0003
GOALS2013	0.007176	0.004095	1.752268	0.0803
GOALS2012	0.006765	0.002445	2.766981	0.0059
ISDEF	-0.084729	0.020254	-4.183387	0.0000
ISSAMER	0.085370	0.029259	2.917714	0.0037
MINUTES2013	0.000118	1.50E-05	7.853646	0.0000
MINUTES2012	4.28E-05	1.09E-05	3.906154	0.0001
MINUTES2011	1.19E-05	9.65E-06	1.235654	0.2171
MINUTES2010	3.23E-05	9.58E-06	3.368114	0.0008
NOYEARSTOGO	0.042506	0.012675	3.353551	0.0009
SOMMIN	3.93E-06	1.46E-06	2.696683	0.0072
STADIONGROOTTE	4.02E-06	6.42E-07	6.257594	0.0000
WK	0.095816	0.024448	3.919124	0.0001
WASCHAMPLY	0.227644	0.038646	5.890519	0.0000
C	6.298422	0.191016	32.97333	0.0000

R-squared	0.833022	Mean dependent var	3.098156
Adjusted R-squared	0.825642	S.D. dependent var	0.479525
S.E. of regression	0.200231	Akaike info criterion	-0.335671
Sum squared resid	21.77026	Schwarz criterion	-0.144557
Log likelihood	120.3307	Hannan-Quinn criter.	-0.261093
F-statistic	112.8721	Durbin-Watson stat	1.662149
Prob(F-statistic)	0.000000		

Figuur 9: *In dit tabel is de Eviews output van het model van de oude spelers te vinden.*



	AGE2	ASS2013	BUNDES_O	BUNDES_D	CL	EL	EREDIVISIE	GOALS2013	GOALS2012	ISDEF	ISSAMER	JUPILER	LIGUE1	MIN2013	MIN2012	MIN2011	MIN2010	NOYTG	SERIE_A	SOMMIN	STADIONG	WK	WASCLY
AGE2	1	-0.070	0.011	-0.033	0.177	0.074	-0.009	-0.092	-0.069	0.049	0.019	-0.066	0.033	-0.109	-0.033	-0.014	0.032	-0.369	0.085	0.428	0.062	0.099	0.006
ASS2013	-0.070	1	0.089	0.055	0.162	0.073	-0.013	0.563	0.432	-0.307	-0.020	0.042	-0.059	0.431	0.278	0.149	0.069	0.250	-0.063	0.201	0.079	0.192	0.230
BUNDES_O	0.011	0.089	1	-0.070	-0.069	0.002	-0.058	-0.003	-0.015	-0.010	-0.070	-0.066	-0.087	-0.011	-0.018	-0.025	-0.069	-0.062	-0.105	-0.068	-0.232	-0.088	0.057
BUNDES_D	-0.033	0.055	-0.070	1	0.053	0.098	-0.096	0.054	0.046	0.057	0.016	-0.109	-0.145	-0.045	0.029	0.034	-0.099	0.004	-0.174	0.063	0.274	0.054	0.054
CL	0.177	0.162	-0.069	0.053	1	0.355	-0.090	0.135	0.068	0.017	0.076	-0.137	0.079	0.163	0.100	0.055	0.077	0.060	0.097	0.454	0.411	0.382	0.238
EL	0.074	0.073	0.002	0.098	0.355	1	-0.010	0.024	0.042	0.021	0.040	-0.005	-0.030	0.044	0.088	0.070	0.138	0.012	0.060	0.268	0.223	0.125	0.020
EREDIVISIE	-0.009	-0.013	-0.058	-0.096	-0.090	-0.010	1	-0.007	-0.025	-0.085	-0.096	-0.091	-0.120	-0.127	-0.094	-0.137	-0.129	-0.060	-0.145	-0.115	-0.224	-0.070	-0.046
GOALS2013	-0.092	0.563	-0.003	0.054	0.135	0.024	-0.007	1	0.695	-0.343	0.051	-0.031	-0.025	0.338	0.199	0.074	0.037	0.216	0.014	0.126	0.109	0.243	0.207
GOALS2012	-0.069	0.432	-0.015	0.046	0.068	0.042	-0.025	0.695	1	-0.425	0.006	-0.027	-0.066	0.143	0.349	0.120	0.010	0.234	-0.007	0.095	0.089	0.178	0.153
ISDEF	0.049	-0.307	-0.010	0.057	0.017	0.021	-0.085	-0.343	-0.425	1	0.045	-0.029	0.039	0.091	0.071	0.048	0.099	-0.042	-0.002	0.100	0.068	-0.076	0.007
ISSAMER	0.019	-0.020	-0.070	0.016	0.076	0.040	-0.096	0.051	0.006	0.045	1	-0.050	-0.017	-0.008	-0.006	0.024	-0.048	0.048	0.210	-0.057	0.178	0.096	0.054
JUPILER	-0.066	0.042	-0.066	-0.109	-0.137	-0.005	-0.091	-0.031	-0.027	-0.029	-0.050	1	-0.136	0.078	-0.013	-0.038	-0.012	-0.023	-0.164	-0.139	-0.288	-0.142	-0.034
LIGUE1	0.033	-0.059	-0.087	-0.145	0.079	-0.030	-0.120	-0.025	-0.066	0.039	-0.017	-0.136	1	0.040	-0.071	-0.081	-0.074	-0.089	-0.218	-0.004	-0.130	-0.074	0.010
MIN2013	-0.109	0.431	-0.011	-0.045	0.163	0.044	-0.127	0.338	0.143	0.091	-0.008	0.078	0.040	1	0.496	0.262	0.209	0.336	-0.082	0.294	0.080	0.135	0.143
MIN2012	-0.033	0.278	-0.018	0.029	0.100	0.088	-0.094	0.199	0.349	0.071	-0.006	-0.013	-0.071	0.496	1	0.432	0.259	0.302	-0.035	0.374	0.183	0.162	0.174
MIN2011	-0.014	0.149	-0.025	0.034	0.055	0.070	-0.137	0.074	0.120	0.048	0.024	-0.038	-0.081	0.262	0.432	1	0.389	0.219	-0.020	0.396	0.253	0.213	0.122
MIN2010	0.032	0.069	-0.069	-0.099	0.077	0.138	-0.129	0.037	0.010	0.099	-0.048	-0.012	-0.074	0.209	0.259	0.389	1	0.061	0.000	0.462	0.209	0.251	0.120
NOYTG	-0.369	0.250	-0.062	0.004	0.060	0.012	-0.060	0.216	0.234	-0.042	0.048	-0.023	-0.089	0.336	0.302	0.219	0.061	1	0.050	0.005	0.140	0.061	0.101
SERIE_A	0.085	-0.063	-0.105	-0.174	0.097	0.060	-0.145	0.014	-0.007	-0.002	0.210	-0.164	-0.218	-0.082	-0.035	-0.020	0.000	0.050	1	0.049	0.237	0.061	0.009
SOMMIN	0.428	0.201	-0.068	0.063	0.454	0.268	-0.115	0.126	0.095	0.100	-0.057	-0.139	-0.004	0.294	0.374	0.396	0.462	0.005	0.049	1	0.411	0.422	0.259
STADIONG	0.062	0.079	-0.232	0.274	0.411	0.223	-0.224	0.109	0.089	0.068	0.178	-0.288	-0.130	0.080	0.183	0.253	0.209	0.140	0.237	0.411	1	0.368	0.296
WK	0.099	0.192	-0.088	0.054	0.382	0.125	-0.070	0.243	0.178	-0.076	0.096	-0.142	-0.074	0.135	0.162	0.213	0.251	0.061	0.061	0.422	0.368	1	0.226
WASCLY	0.006	0.230	0.057	0.054	0.238	0.020	-0.046	0.207	0.153	0.007	0.054	-0.034	0.010	0.143	0.174	0.122	0.120	0.101	0.009	0.259	0.296	0.226	1

Figuur 10: *Dit is een correlatiematrix van variabelen uit het model voor de oude spelers.*

Heteroskedasticity Test: Breusch-Pagan-Godfrey

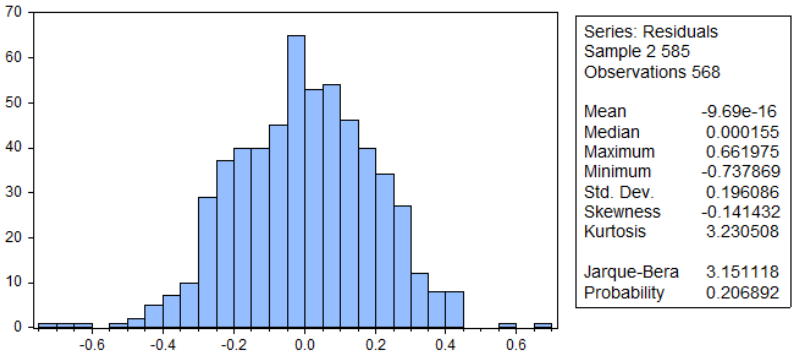
F-statistic	1.330185	Prob. F(24,543)	0.1361
Obs*R-squared	31.53993	Prob. Chi-Square(24)	0.1388
Scaled explained SS	32.69128	Prob. Chi-Square(24)	0.1108

Figuur 11: *Dit is de uitslag van de Breuch-Pagan toets voor heteoskedasticiteit.*

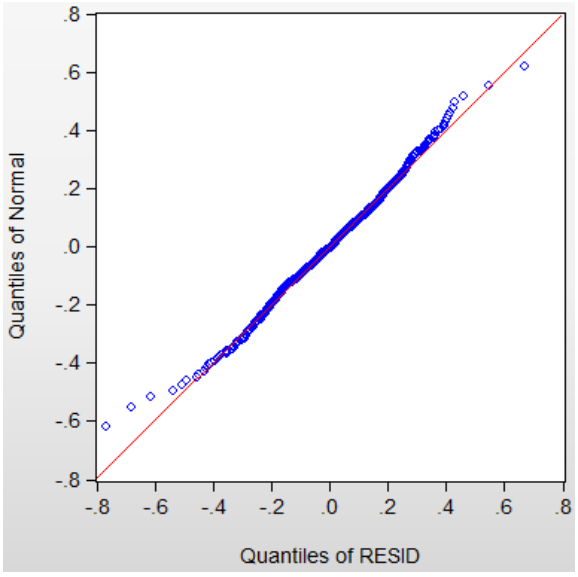
Heteroskedasticity Test: White

F-statistic	1.371450	Prob. F(287,280)	0.0040
Obs*R-squared	331.8976	Prob. Chi-Square(287)	0.0350
Scaled explained SS	344.0133	Prob. Chi-Square(287)	0.0117

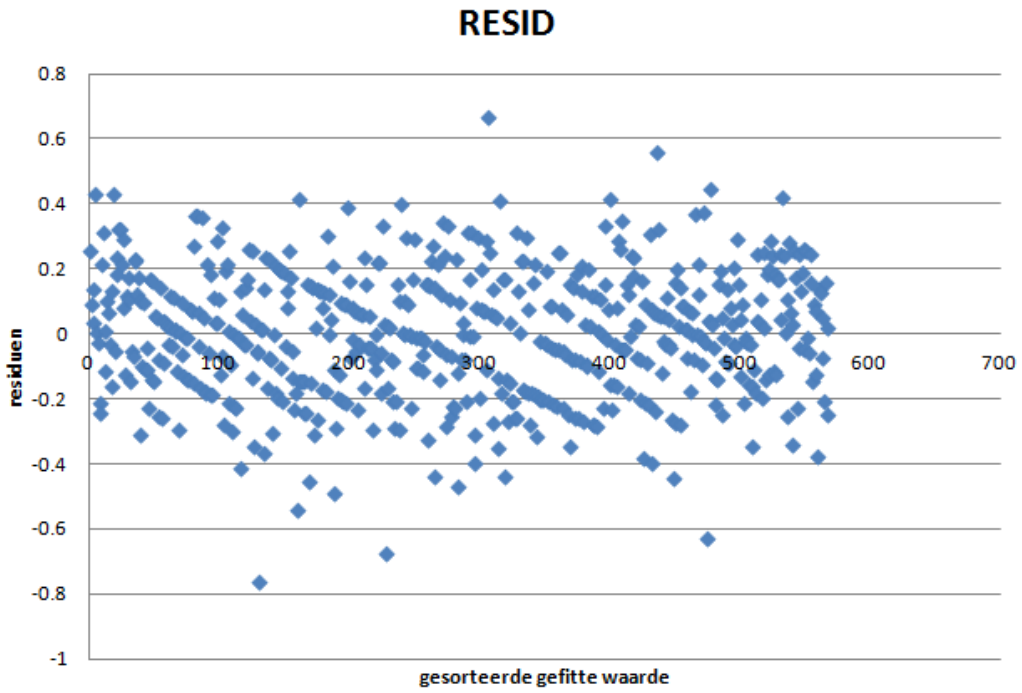
Figuur 12: *Dit is de uitslag van de White toets voor heteoskedasticiteit.*



Figuur 13: *Dit is een histogram van de residuen van de oude spelers.*



Figuur 14: *Dit is een QQ-plot van de residuen van de oude spelers.*



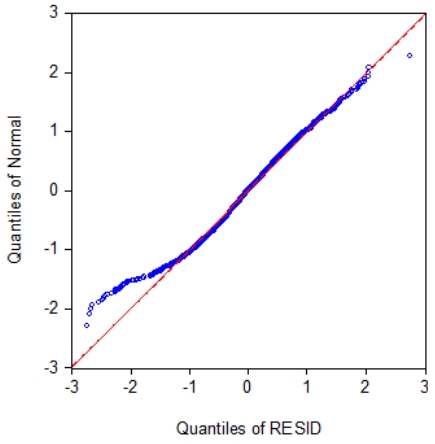
Figuur 15: *Dit is een dotplot van de residuen van de oude spelers met op de x-as de gesorteerde gefitte waarden.*

Dependent Variable: LOG_FIELD_MARKETVALUE
Method: Least Squares
Date: 04/22/14 Time: 01:23
Sample: 1 3770 IF ISGOAL=0
Included observations: 3107

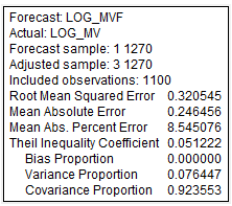
Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGE	0.593640	0.033134	17.91615	0.0000
AGE2	-0.012715	0.000645	-19.72137	0.0000
ASSISTS2012	0.012871	0.004326	2.975250	0.0030
ASSISTS2013	0.022089	0.006617	3.338472	0.0009
BUNDESLIGA_DUIT	-0.618999	0.044483	-13.91546	0.0000
BUNDESLIGA_OOS	-1.545924	0.057719	-26.78374	0.0000
CL	0.319770	0.030737	10.40351	0.0000
EK	0.114346	0.053167	2.150593	0.0316
EL	0.241152	0.025237	9.555599	0.0000
EREDIVISIE	-1.081961	0.048388	-22.35992	0.0000
GOALS2013	0.023272	0.004047	5.750145	0.0000
INLASTYEAR	-0.138634	0.026685	-5.195216	0.0000
ISAFRICA	0.103824	0.039683	2.616336	0.0089
ISSAMER	0.370669	0.040699	9.107542	0.0000
JUPILER	-0.827679	0.052319	-15.81976	0.0000
LENGTECM	0.004093	0.001966	2.081416	0.0375
LIGUE1	-0.396850	0.045763	-8.671879	0.0000
MATCHES2010	0.004379	0.001089	4.022998	0.0001
MATCHES2011	0.007712	0.001083	7.120094	0.0000
MATCHES2012	0.015013	0.001293	11.60764	0.0000
MATCHES2013	0.029914	0.001673	17.87642	0.0000
PRIMERA_DIV	-0.229758	0.046146	-4.978980	0.0000
SERIE_A	-0.137943	0.045813	-3.011030	0.0026
SOMMIN	1.13E-05	2.74E-06	4.126654	0.0000
STADIONGROOTTE	1.07E-05	8.50E-07	12.59782	0.0000
WASCHAMPLY	0.270521	0.054393	4.975265	0.0000
C	-1.705499	0.529442	-3.221315	0.0013

R-squared	0.776693	Mean dependent var	7.406311
Adjusted R-squared	0.774808	S.D. dependent var	1.342688
S.E. of regression	0.637165	Akaike info criterion	1.945076
Sum squared resid	1250.416	Schwarz criterion	1.997576
Log likelihood	-2994.675	Hannan-Quinn criter.	1.963926
F-statistic	412.0256	Durbin-Watson stat	1.779321
Prob(F-statistic)	0.000000		

Figuur 16: *In dit tabel is de Eviews output van het algemene veldspeler model uit het werkcollege te vinden.*

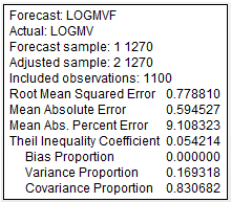


Figuur 17: *Dit is een QQ-plot van de residuen van alle veldspelers.*



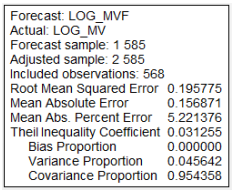
Forecast: LOG_MVF	
Actual: LOG_MV	
Forecast sample: 1 1270	
Adjusted sample: 3 1270	
Included observations: 1100	
Root Mean Squared Error	0.320545
Mean Absolute Error	0.246456
Mean Abs. Percent Error	8.545076
Theil Inequality Coefficient	0.051222
Bias Proportion	0.000000
Variance Proportion	0.076447
Covariance Proportion	0.923553

Figuur 18: *In dit figuur staat de in-sample fit weergegeven voor de jonge spelers met het nieuwe model voor de jonge spelers.*



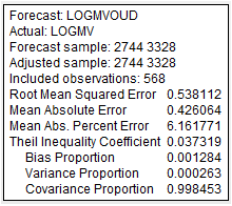
Forecast: LOGMVF	
Actual: LOGMV	
Forecast sample: 1 1270	
Adjusted sample: 2 1270	
Included observations: 1100	
Root Mean Squared Error	0.778810
Mean Absolute Error	0.594527
Mean Abs. Percent Error	9.108323
Theil Inequality Coefficient	0.054214
Bias Proportion	0.000000
Variance Proportion	0.169318
Covariance Proportion	0.830682

Figuur 19: *In dit figuur staat de in-sample fit weergegeven voor de jonge spelers met het oude algemene veldspelermodel uit het werkcollege.*



Forecast: LOG_MVF	
Actual: LOG_MV	
Forecast sample: 1 585	
Adjusted sample: 2 585	
Included observations: 568	
Root Mean Squared Error	0.195775
Mean Absolute Error	0.156871
Mean Abs. Percent Error	5.221376
Theil Inequality Coefficient	0.031255
Bias Proportion	0.000000
Variance Proportion	0.045642
Covariance Proportion	0.954358

Figuur 20: *In dit figuur staat de in-sample fit weergegeven voor de oude spelers met het nieuwe model voor de oude spelers.*



Forecast: LOGMVOUD	
Actual: LOGMV	
Forecast sample: 2744 3328	
Adjusted sample: 2744 3328	
Included observations: 568	
Root Mean Squared Error	0.538112
Mean Absolute Error	0.426064
Mean Abs. Percent Error	6.161771
Theil Inequality Coefficient	0.037319
Bias Proportion	0.001284
Variance Proportion	0.000263
Covariance Proportion	0.998453

Figuur 21: *In dit figuur staat de in-sample fit weergegeven voor de oude spelers met het oude algemene veldspelermodel uit het werkcollege.*