Universiteit Utrecht

Mathematisch Instituut

Bachelorscriptie

Kwantielregressie versus OLS: effect van geslacht op reistijd

Rosalien Timmerhuis - 4261178

Eerste begeleider:
prof. dr. Wolter Hassink
Departement Economie

Tweede begeleider:
dr. Martin Bootsma
Departement Wiskunde

9 juni 2017

Samenvatting

Dit onderzoek richt zich op de verandering van het effect van geslacht op de bereidheid tot het aangaan van reistijd in het Nederlandse woon-werkverkeer in de periode van 1985 tot en met 2014. Het verband hiertussen wordt onderzocht met twee verschillende lineaire regressie methoden, namelijk de *Ordinary Least Squares*- en de kwantielregressiemethode. In de modellen worden de leeftijd, loon, hoogst genoten opleidingsniveau en provincie van het woonhuis meegenomen als controlevariabelen. Voor dit onderzoek zijn data van het Arbeidsaanbodpanel van 1985 tot en met 2014 gebruikt, N = 44.333. De verandering over tijd wordt gemeten aan de hand van drie jaarcategorieën van ongeveer tien jaar.

Mannen zijn in elke jaarcategorie bereid langer te reizen dan vrouwen. Tussen 1985-1995 en 1996-2005 is het effect van geslacht sterk toegenomen en tussen 1996-2005 en 2006-2014 weer iets afgenomen. Kwantielregressie blijkt in dit onderzoek van meerwaarde te zijn, aangezien er in lage kwantielen geen effect gemeten is van geslacht en in de hoge kwantielen juist een erg sterk effect. Bij OLS blijken deze verschillen niet.

Inhoudsopgave

1 Inleiding **3**

2 Theoretisch kader regressie-analyse **6**
2.1 Ordinary Least Squares-methode 7
2.2 Kwantielregressie 8
2.2.1 Kwantielen 9
2.2.2 Kwantielregressie methode 11
2.3 Verschillen OLS en kwantielregressie 12

3 Theoretisch kader economisch perspectief **14**

4 Data **17**
4.1 Data bewerking 17
4.2 Data statistieken 20

5 Statistische modellen **25**
5.1 Model OLS 25
5.2 Model kwantielregressie 26

6 Resultaten **27**

7 Conclusie en discussie **30**
7.1 Conclusie 30
7.2 Discussie 31
7.3 Evaluatie onderzoek 31

Referenties **33**

A Appendix: Wiskundige bewijzen **34**

B Appendix: Commando's Stata **37**

C Appendix: Stata output **39**

1 Inleiding

Een regressie wordt uitgevoerd om samenhang in data te vinden tussen (een) afhankelijke en onafhankelijke variabele(n). De verdeling van een afhankelijke variabele kan worden beïnvloed door onafhankelijke variabelen. Na een regressie kan dit geschatte effect worden afgelezen aan de bijbehorende parameter. Er bestaan verschillende methoden om de totstandkoming van de verdeling van de afhankelijke variabele te schatten. Twee methoden die in deze scriptie aan bod komen, zijn de de *Ordinary Least Squares* (OLS) en de kwantielregressie. Bij het schatten met de OLS-methode wordt het gemiddelde van de afhankelijke variabele in alle metingen als uitgangspunt gebruikt. Bij kwantielregressie staat een percentiel naar keuze uit de verdeling van de afhankelijke variabele centraal. Of kwantielregressie van meerwaarde is ten opzichte van OLS, hangt af van de situatie.

In deze scriptie wordt de meerwaarde van kwantielregressie onderzocht aan de hand van een praktijkvoorbeeld. In dit voorbeeld is de afhankelijke variabele de reistijd voor woon-werkverkeer in minuten. Deze verdeling wordt bekeken voor twee verschillende groepen, namelijk voor mannen en voor vrouwen in Nederland. Daarbij worden de verdelingen ook vergeleken over de tijd, en wel van 1985 tot en met 2014. Bij OLS wordt gekeken naar het effect van geslacht op de gemiddelde reistijd van alle respondenten. Bij kwantielregressie kan er gekeken worden naar een specifiek deel van de verdeling, bijvoorbeeld naar de groep individuen met een grote reistijd. Over het verschil in reistijd tussen mannen en vrouwen is nog weinig bekend. Daarbij is er over de afgelopen jaren veel veranderd in het arbeidsmarktgedrag van de beroepsbevolking, met name bij vrouwen, wat effect kan hebben op de bereidheid tot reistijd. De conclusie van dit onderzoek kan gebruikt worden bij het verklaren van het gedrag van man en vrouw in het aangaan van reistijd. Afhankelijk van de conclusie kan dit eventueel bijdragen aan het verklaren van het verschil in arbeidsmarktgedrag tussen mannen en vrouwen.

Uit onderzoek (Vlasblom et al., 2016, p. 24) blijkt dat tussen 1986 en 2012 een flinke verschuiving van de arbeidsverdeling in huishoudens heeft plaatsgevonden. Waar in 1986 in de helft van de huishoudens de vrouw niet werkte en de man voltijds, is dat in 2012 nog maar 19%. Van origine is de rolverdeling zodanig dat de man werkt en de vrouw het huishouden regelt en de zorgtaken op zich neemt. In de afgelopen jaren is de arbeidsparticipatie bij vrouwen gestegen, in tegenstelling tot mannen. Waar in 1986 85% van de mannen en 44% van de vrouwen werkten, is dat in 2014 respectievelijk 83% en 70%, zowel deeltijd als voltijd banen meegerekend (van Echtelt et al., 2016 p.19). Aangezien het arbeidsmarktgedrag meer naar elkaar is toegegroeid, zou dit kunnen betekenen dat de bereidheid van man en vrouw tot het accepteren van een zekere reistijd ook meer naar elkaar is

toegegroeid. Vrouwen werken echter nog steeds veel vaker in deeltijd dan mannen (Vlasblom, 2014, p. 24) en ze zorgen op meer vaste doordeweekse dagen voor kinderen (Souren, 2006). Dit kan betekenen dat een moeder nog steeds graag dicht bij de school van de kinderen wil werken in verband met opvang na school, met een kortere reistijd als gevolg.

Het doel van dit onderzoek is inzicht krijgen of het effect van geslacht op reistijd in Nederland is veranderd door de jaren heen, door middel van empirisch onderzoek aan de hand van het arbeidsaanbodpanel (AAP). Het arbeidsaanbodpanel is opgezet om diverse aspecten van de arbeidssituatie van werkenden en niet-werkenden in Nederland in kaart te brengen (van Echtelt et al., 2016, p.5). Aan de hand van deze data wordt de volgende onderzoeksvraag getoetst:

Is het effect van geslacht op de reistijd voor het woon-werkverkeer in Nederland veranderd in de periode van 1985 tot en met 2014?

De vragenlijsten van het arbeidsaanbodpanel zijn afgenomen vanaf 1985 en vervolgens in de even jaren daarna. Aangezien een deel van de respondenten meermaals heeft meegedaan aan het onderzoek, heeft de dataset een panelstructuur. De enquêteresultaten van het arbeidsaanbodpanel van de periode 1985 tot en met 2014 zijn verkregen via DANS (*Data Archiving and Networked Services*). DANS is het Nederlands instituut voor toegang tot digitale onderzoeksgegevens. Met behulp van Stata wordt het effect van geslacht op de reistijd geschat, zowel door middel van OLS als kwantielregressie. Er zullen verschillende jaarcategorieën met elkaar vergeleken worden om uitspraken te kunnen doen over een verandering door de jaren heen. Door de resultaten van de twee schattingsmethodes te vergelijken, kan worden ontdekt of kwantielregressie hier van toegevoegde waarde is.

Naast de variabele geslacht bestaan er meer variabelen die van invloed kunnen zijn op de bereidheid tot reistijd, hiervoor kan worden gecorrigeerd door middel van controlevariabelen. Uit onderzoek (Molnár, 2004) blijkt dat mensen met een hoger opleidingsniveau over het algemeen verder naar hun werk reizen, daarbij is het opleidingsniveau van vrouwen in de afgelopen jaren toegenomen. Uit een ander onderzoek (Geurs en Ritsema van Eck, 2000, p. 45-65) blijkt dat de bereikbaarheid van arbeidsplaatsen binnen 45 minuten reistijd per auto en openbaar vervoer in de randstad veel groter is dan in andere provincies. Deze conclusies zouden kunnen duiden op een verschil in de bereidheid tot reistijd van de beroepsbevolking per opleidingsniveau en per provincie in Nederland.

De structuur van deze scriptie is als volgt: ten eerste er worden de gebruikte regressiemethoden en de wiskunde daarachter besproken, in dit onderdeel

worden onder andere de verschillen tussen OLS en kwantielregressie behandeld. Vervolgens worden verschillende bestaande theorieën rondom effecten op reistijd van het woon-werkverkeer besproken. Daarna worden de data die voor dit onderzoek relevant zijn gefilterd uit de verkregen data en worden deze data in kaart gebracht. Aan de hand van de voorgaande theoretische secties en de beschikbare data worden de statistische modellen opgesteld voor het empirisch onderzoek. Vervolgens worden de regressies uitgevoerd op de overgebleven data, hieruit volgen resultaten. Deze resultaten worden geanalyseerd, waaruit een conclusie wordt gevormd. In de discussie worden de resultaten geïnterpreteerd en wordt het onderzoek geëvalueerd. Ten slotte bevat deze scriptie nog drie appendices met respectievelijk wiskundige bewijzen, Stata in- en output.

2 Theoretisch kader regressie-analyse

Binnen de statistiek wordt met het woord regressie de eventuele samenhang tussen variabelen bedoeld. Regressie-analyse wordt gebruikt om verbanden op te sporen in data en hier dan bijvoorbeeld voorspellingen over te doen. De relatie tussen een afhankelijke, ook wel kritische, variabele en een onafhankelijke variabele, de predictor, wordt bekeken. Het is mogelijk om meerdere onafhankelijke variabelen mee te nemen, dit wordt multivariabele regressie-analyse genoemd. In grote lijnen ziet een regressiemodel er als volgt uit:

$$E(Y|X) = g(X, \beta)$$

Hierin is Y de afhankelijke variabele, X de matrix die de onafhankelijke variabele(n) bevat, β de parametervector en de functie g is afhankelijk van het gebruikte model. De parametervector β bestaat uit coëfficiënten die voor elke onafhankelijke variabele aangeeft wat het effect van deze variabele is op de afhankelijke variabele Y. Bij een regressie gaat de interesse uit naar de geschatte waardes van vector β, deze schatter wordt aangegeven met $\hat{\beta}$. Een model zal over het algemeen niet volledig overeen komen met de werkelijke waarde van Y, dit wordt als volgt genoteerd:

$$Y = g(X, \beta) + \epsilon$$

Hierin geeft ϵ de errorterm weer, het verschil tussen de werkelijke waarde van de kritische variabele en de geschatte waarde volgens het regressiemodel.

Er bestaan veel verschillende methoden voor regressie-analyse. Zoals in de inleiding al is beschreven, wordt in deze scriptie geschat door middel van de lineaire OLS- en kwantielregressie-methode. De vorm van een lineair regressiemodel is:

$$Y = X\beta + \epsilon$$

Hierin is Y een $(n \times 1)$ vector, X de $(n \times (k+1))$ matrix van k onafhankelijke variabelen en n observaties en β een $((k+1) \times 1)$ vector. De eerste kolom van matrix X bestaat slechts uit 1'en. Zo ontstaat na de vermenigvuldiging van matrix X met vector β onder andere de coëfficiënt β_0, dit is een constante. Een regressie-methode is lineair op het moment dat de hierboven genoemde functie $g(X, \beta)$ linear is. Lineair refereert naar de de relatie tussen de parameters van β die geschat worden en de afhankelijke variabele Y. Dit betekent bijvoorbeeld dat $Y = e^X\beta + \epsilon$ linear is, maar $Y = e^\beta X + \epsilon$ niet. Vandaar is het mogelijk om in een lineair regressiemodel logaritmes of kwadraten te nemen van de onafhankelijke variabelen.

Bij kwantielregressie wordt op meerdere plaatsen van de verdeling van de afhankelijke variabele naar de de totstandkoming van de afhankelijke variabele gekeken. Echter hoeft deze informatie op verschillende plekken van

de verdeling niet per se van toegevoegde waarde zijn, bijvoorbeeld omdat de coëfficiënten nagenoeg gelijk blijven over de verdeling. Daarom vergelijken we dit met de meeste gebruikte vorm van regressie: de OLS-methode. In deze sectie worden specificaties over beide methoden besproken, zoals de totstandkoming van de schatter $\hat{\beta}$ en worden de verschillen tussen beide methoden behandeld.

2.1 Ordinary Least Squares-methode

De bekendste lineaire regressie methode is de *Ordinary Least Squares*-methode (OLS). Deze methode valt onder de kleinste-kwadraten methodes. Zoals de naam al doet vermoeden, worden bij deze methoden kwadraten van de afwijkingen van de meetwaarden ten opzichte van het model geminimaliseerd. Kleinste-kwadraten methodes hoeven niet lineair te zijn, OLS is wel een lineaire methode. De vorm van een OLS-model is:

$$Y = X\beta + \epsilon$$

Bij de OLS-methode wordt de som van de gekwadrateerde afwijkingen van de meetwaarden ten opzichte van de geschatte lineaire functie geminimaliseerd, dit wordt de doelstellingsfunctie genoemd:

$$\min \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2$$

Equivalent aan deze doelstellingsfunctie is:

$$\min E(Y - X\hat{\beta})^2$$

De minimalisatie gaat over de gehele verdeling van Y, waar de afwijking op elk deel van de verdeling even zwaar meetelt. Er wordt dus eigenlijk gezocht naar het gemiddelde. Het gemiddelde van Y, μ, is namelijk gelijk aan de c waarvoor het gekwadrateerde verschil met Y het kleinst is (Davino et al., 2013, p. 2), zie Appendix A voor het bewijs.

$$\mu = \arg\min_{c \in \mathbb{R}} E(Y - c)^2 \tag{1}$$

In deze functie is *arg min* gebruikt, omdat het argument gezocht wordt waarvoor de gekwadrateerde afwijkingen samen het kleinst zijn. Bij OLS-regressie wordt dus het effect van de onafhankelijke variabelen op het gemiddelde van de verdeling van de afhankelijke variabele geschat.

Als het model een correcte beschrijving van de werkelijkheid is, moet de verwachte waarde van de schatter gelijk zijn aan de werkelijke waarde van β, dus: $E(\hat{\beta}|X) = \beta$, dit wordt consistentie genoemd. Dit kan herschreven worden tot de volgende functie (zie Appendix A):

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

waarin X' de getransponeerde matrix is van X.

Aannames OLS

Voor OLS bestaan een aantal aannames over de data en het model. Als de onderstaande aannames gelden, zijn de geschatte parameters volgende de Gauss–Markov theorie BLUE, *Best Linear Unbiased Estimator*. Een OLS model waarvan de geschatte parameters BLUE zijn, is de meest efficiënte methode. Als aan deze aannames niet is voldaan, hebben de resultaten van de regressie geen of een verminderde betekenis.

De eerste aanname is al eerder langsgekomen, het regressie model moet lineair zijn in de parameters. Vervolgens moeten er voldoende metingen zijn; er moeten meer observaties zijn dan variabelen om de $k+1$ coëfficiënten te vinden, oftewel $n > k+1$. De errorterm in de regressie moet het conditionele gemiddelde van 0 hebben en is niet afhankelijk van X, dus $E[\epsilon|X] = 0$, dit wordt strikte exogeniteit genoemd. Een lichtere eis dan de strikte exogeniteit wordt aangegeven met zwakke exogeniteit, ofwel $Cov(\epsilon, X) = 0$, waarmee ook consistentie van de schatter bereikt kan worden. Er moet echter dan ook sprake zijn van een oneindig grote dataset. Een andere aanname is dat er geen sprake mag zijn van perfecte multicollineariteit, dit betekent dat er geen lineaire relatie mag bestaan tussen de onafhankelijke variabelen. Hiervan is geen sprake als de rank van matrix X volledig is, oftewel rank$(X) = k+1$. Dit is equivalent aan de uitspraak dat matrix $(X'X)$ inverteerbaar is. Het moet niet het geval zijn dat de afhankelijke variabele de onafhankelijke variabele(n) beïnvloeden. Dit komt omdat de causale relatie bekeken wordt in een regressiemodel en niet de correlatie tussen twee variabelen. Ten slotte moet de onzekerheid van het model overal gelijk en onafhankelijk van X zijn, ofwel $Var(\epsilon|X) = \sigma_\epsilon^2 I_n$, beter bekend als i.i.d. (*independent and identically distributed*). Hierin is I_n de identiteitsmatrix en σ_ϵ^2 een parameter die de variantie aangeeft van de errorterm in een observatie. Deze aanname kan opgesplitst worden in twee kleinere aannames; namelijk homoskedasticiteit en autocorrelatie. Er moet sprake zijn van homoskedasticiteit van de errortermen; dit impliceert dat de variantie van de errortermen in elke observatie hetzelfde is conditioneel op de onafhankelijke variabelen, er geldt $E(\epsilon_i^2|X) = \sigma_\epsilon^2$. Tussen de errortermen mag geen sprake zijn van autocorrelatie, dit betekent dat er geen correlatie mag bestaan tussen verschillende errortermen, oftewel: $Cov(\epsilon_i \epsilon_j) = 0$ als $i \neq j$. In paneldata is er vaak sprake van autocorrelatie, omdat individuen herhaaldelijk terugkomen en er dan een correlatie is met de individu zelf.

2.2 Kwantielregressie

In 1760 schreef Boskovic een beginnende theorie over mediaan-regressie, zijn theorie was echter niet volledig. Pas sinds de ontdekking van de Simplexmethode voor optimaliseringsproblemen kon er een volledige theorie geschre-

ven worden. Deze volledige theorie over kwantielregressie werd voor het eerst beschreven door Koenker en Bassett (1978). Kwantielregressie is een techniek die het $q-$de kwantiel schat van de afhankelijke variabele Y conditioneel op onafhankelijke variabelen X. Hierbij kan het effect van de onafhankelijke variabelen op de afhankelijke variabele geschat worden op verschillende plaatsen, ook wel kwantielen, van de verdeling. Kwantielregressie kan zowel lineair als non-lineair zijn. Zoals al eerder genoemd is, refereert lineair naar de relatie tussen de parameters van β die geschat worden en de afhankelijke variabele Y. Een multivariabel lineair kwantielregressie model ziet er als volgt uit:

$$\xi_{\mathrm{q}}(y|X) = X\beta(q) + \epsilon$$

Hierin is ξ_{q} het q-kwantiel. Het doel is weer om de schatter van β, namelijk $\hat{\beta}$, te vinden, deze geeft namelijk de coëfficiënten van de onafhankelijke variabelen aan.

2.2.1 Kwantielen

Een kwantiel is een getal dat binnen de statistiek gebruikt wordt om een dataset te verdelen. Voor het getal $q \in (0, 1)$, is het q-kwantiel ξ_{q} het getal zodanig dat het deel q van de gegevens kleiner of gelijk is aan ξ_{q} en deel $1-q$ groter of gelijk is aan ξ_{q}. Een meer gebruikte term is percentiel, percentielen zijn kwantielen die bij een percentage behoren. De mediaan, het midden van een geordende dataset, is bijvoorbeeld gelijk aan het 50^e percentiel, dit wordt genoteerd met $q = 0.5$, oftewel $\xi_{0.5}$

Gegeven is een continue random variabele Y met een kansdichtsheidsfunctie $f_{\mathrm{Y}}(y)$ en een verdelingsfunctie $F_{\mathrm{Y}}(y)$. De waarde $F_{\mathrm{Y}}(y)$ van variabele Y in het punt y, is dus de cumulatieve kans op waarden van Y kleiner dan of gelijk aan y, $F_{\mathrm{Y}}(y) \in [0, 1]$. Het is bekend dat:

$$F_{\mathrm{Y}}(y) = \int_{-\infty}^{y} f_{\mathrm{Y}}(t)\, dt = P(Y \leq y)$$

Hier wordt de kans bepaald dat Y kleiner of gelijk is aan y, de geordende dataset wordt gesplitst in een deel kleiner of gelijk aan y en een deel groter dan y. Het q-kwantiel ξ_{q} met $q \in (0, 1)$ splitst de dataset in deel q kleiner of gelijk aan ξ_{q} en deel $1 - q$ groter of gelijk aan ξ_{q}. Vandaar wordt het q-kwantiel ξ_{q} als volgt gedefinieerd:

$$\xi_{\mathrm{q}}(Y) = \inf \{y \in \mathbb{R} : F_{\mathrm{Y}}(y) \geq q\}$$

Hier wordt het infimum gebruikt, de grootste ondergrens, in plaats van het minimum. Hiervoor is gekozen omdat omdat $F_{\mathrm{Y}}(y)$ een stijgende en rechtscontinue functie is. Als er gekeken wordt naar een punt y waarvoor $F_{\mathrm{Y}}(y)$ niet links-continu is, bestaat er geen minimum, maar wel een infimum.

De verdelingsfunctie van Y in punt ξ_{q}, geeft dus het deel aan wat kleiner of gelijk is aan waarde ξ_{q}, namelijk q.

$$F_{\mathrm{Y}}(\xi_{\mathrm{q}}(Y)) = q$$

Net als OLS-regressie is kwantielregressie een minimaliseringsprobleem, het verschil tussen de verkregen dataset en het model moet namelijk zo klein mogelijk worden. De mediaan, $\xi_{0.5}$, is een gemakkelijk uitgangspunt als er naar kwantielen gekeken wordt, omdat deze precies in het midden van de geordende dataset ligt. Daarom starten we met het definiëren van de mediaan, zodat dit kan worden uitgebreid naar een algemene vorm.
De mediaan is gedefinieerd als de geminimaliseerde som van de absolute residuen. Het residu is $Y - c$, oftewel het verschil tussen de waarde van de random variabele en een element $c \in \mathbb{R}$. Er kan aangenomen worden dat:

$$mediaan = \xi_{0.5} = \arg\min_{c \in \mathbb{R}} E\,|Y - c| \tag{3}$$

We willen echter het kwantiel kunnen bepalen voor elke $q \in (0, 1)$, daarom wordt de functie $\rho_{\mathrm{q}}(.)$ geïntroduceerd. Deze functie zorgt ervoor dat voor elke waarde van c groter dan y (de negatieve residuen) vermenigvuldigd wordt met $(1 - q)$ en kleiner dan y (de positieve residuen) vermenigvuldigd wordt met q (Angrist & Pischke, 2013, p. 271). Noem $r = y - c$ het residu, dan is $\rho(r)$ als volgt gedefinieerd:

$$\begin{aligned} \rho_{\mathrm{q}}(r) &= ((1-q)I(r \leq 0) + q\ I(r\ \ > 0))|r| \\ \rho_{\mathrm{q}}(y-c) &= ((1-q)I(y \leq c) + q\ I(y\ \ > c))|y-c| \end{aligned} \tag{4}$$

Hierin is $I(.)$ de indicatorfunctie:

$$I(y \leq c) = \begin{cases} 0 & \text{Als } y - c > 0 \\ 1 & \text{Als } y - c \leq 0 \end{cases} \qquad I(y > c) = \begin{cases} 0 & \text{Als } y - c \leq 0 \\ 1 & \text{Als } y - c > 0 \end{cases}$$

Om voor elk kwantiel de waarde te berekenen, worden vergelijking (3) en (4) samengevoegd, hieruit volgt de doelfunctie van kwantielregressie. Het bewijs van deze functie staat beschreven in Appendix A.

$$\xi_{\mathrm{q}} = \arg\min_{c \in \mathbb{R}} E[\rho_{\mathrm{q}}(Y - c)] \tag{5}$$

Zoals in formule (5) te zien is, is het bepalen van ξ_{q} dus een minimaliseringsprobleem. Om dit minimaliseringsprobleem uit te werken, worden eerst de verwachte waarde uit formule (5) en de $\rho-$functie uitgeschreven. De residuen zodanig dat $c \geq y$ worden vermenigvuldigd met $(1 - q)$ en de residuen zodanig dat $c \leq y$ worden vermenigvuldigd met q. Hierdoor tellen de residuen in de buurt van het gewenste kwantiel meer mee. Dit leidt tot de volgende notatie:

$$\xi_{\mathrm{q}} = \arg\min_{c \in \mathbb{R}} \left((1-q) \int_{-\infty}^{c} (c - y) f_Y(y) dy + q \int_{c}^{\infty} (y - c) f_Y(y) dy \right)$$

Om het minimum te bepalen zal deze functie gedifferentieerd en gelijk gesteld aan 0 worden. In Apendix A zijn de tussenstappen uitgewerkt, de uiteindelijk de uitkomst is dat:

$$\int_{-\infty}^{c} f(y)\, dy = q$$

Zoals al eerder is benoemd, is het doel bij het uitvoeren van regressies om de coëfficiënten van β te schatten. Aangezien een uitdrukking voor $\hat{\beta}$ wordt bepaald, die voor elke q geldt, starten we met beredeneren met formule (5). De tussenstappen zijn te lezen in Appendix A, uiteindelijk volgt hieruit (Angrist & Pischke, 2009, p.271):

$$\hat{\beta}(q, Y, X) = \arg\min_{\beta} \sum_{i=1}^{n} (\rho_q(y_i - x_i'\beta)) \tag{6}$$

Een aantal eigenschappen van de kwantielen en de schatters worden nu besproken, de bewijzen zijn te lezen in Appendix A:

1. Een kwantiel is invariant voor een monotone transformatie $h(x)$. Dat wil zeggen dat kwantielen van de getransformeerde random variabeles gelijk zijn aan de getransformeerde kwantielen van de originele random variabele.
$$h(\xi_{\mathrm{q}}(X)) = \xi_{\mathrm{q}}(h(X)) \tag{7}$$

2. Bij het vermenigvuldigen van de afhankelijke variabele met een getal $a > 0$, zal de schatter zich als volgt gedragen:
$$\hat{\beta}(q, aY, X) = a\hat{\beta}(q, Y, X) \tag{8}$$
$$\hat{\beta}(q, -aY, X) = -a\hat{\beta}(1 - q, Y, X)$$

2.2.2 Kwantielregressie methode

Aannames kwantielregressie

Ook bij kwantielregressie bestaan er aannames om de kwaliteit van de schatting te waarborgen. Er moeten net als bij OLS voldoende observaties zijn om coëfficiënten voor de variabelen te vinden, oftewel $n > k + 1$. Homoskedasticiteit van de errorterm is geen vereiste (Davino, 2013, p. 35), maar heteroskedasticiteit kan wel negatieve gevolgen hebben. Bij het uitvoeren van een lineaire regressie wordt een lijn gecreëerd, bij kwantielregressie kan per kwantiel een lijn gecreëerd worden. Als er sprake is van heteroskedasticiteit van de errorterm lopen deze lijnen niet meer evenwijdig. Er ontstaan divergerende lijnen bij toenemende heteroskedasticiteit in de loop van de verdeling en convergerende lijnen als de variantie in het eerste deel van het verdeling hoger is dan aan het einde (Davino, 2013, p. 138). Een gevolg

hiervan is dat de geschatte lijnen kunnen snijden, waardoor de natuurlijke ordening van de kwantielen verdwijnt. Als dit niet gewenst is, moeten de errortermen i.i.d. zijn. Bij kwantielregressie wordt geen specifieke verdeling vereist voor de errortermen.

Zowel in kwantielregressie als OLS model kan gebruik gemaakt worden van dummy-variabelen. Dit wordt uitgelegd aan de hand van een simpel (kwantielregressie)voorbeeld. Voor de onafhankelijke variabele 'geslacht' wordt een 1 gebruikt als het individu een man is en een 0 in het geval dat het individu een vrouw is. Voor de man en vrouw ontstaan dan respectievelijk de volgende vergelijkingen:

$$\hat{y_q} = \hat{\beta}_0(q) + \hat{\beta}_1(q) \cdot 1 = \hat{\beta}_0(q) + \hat{\beta}_1(q) \ (man)$$

$$\hat{y_q} = \hat{\beta}_0(q) + \hat{\beta}_1(q) \cdot 0 = \hat{\beta}_0(q) \qquad (vrouw)$$

Dit komt bijvoorbeeld van pas als men het effect van geslacht wil meten. Neem aan dat 'man' heeft waarde 1 en 'vrouw' waarde 0. Het geschatte effect van man zijn ten opzicht van de vrouw is dan af te lezen aan $\hat{\beta}_1(q)$.

2.3 Verschillen OLS en kwantielregressie

Om te beslissen of kwantielregressie van meerwaarde is boven OLS, is het overzichtelijk om de belangrijkste verschillen te onderscheiden, dit zal in deze subsectie gebeuren.

- Het belangrijkste verschil tussen OLS- en kwantielregressie is dat OLS het gemiddelde van de afhankelijke variabele als uitgangspunt neemt en kwantielregressie de mediaan (of een ander kwantiel op de verdeling waar de interesse naar uit gaat).

- Een voordeel van kwantielregressie boven OLS is dat het een meer robuuste methode is, hiermee wordt bedoeld dat uitschieters in de data een minder grote verstoring tot gevolg hebben. Uitschieters hebben namelijk pas invloed op de kleinste en grootste kwantielen. OLS is gevoeliger voor afwijkingen, omdat hier de gekwadrateerde afwijkingen worden geminimaliseerd. Aangezien de afwijking van een uitschieter even zwaar mee telt als een afwijking ergens in het midden van de verdeling, zal de geschatte lijn door de minimalisatie meer richting de uitschieter bewegen dan wenselijk is.

- Bij kwantielregressie is het mogelijk de focus te leggen op bepaalde delen van de verdeling, de regressie wordt namelijk uitgevoerd op een percentiel naar keuze. Dit betekent dat er gekeken kan worden naar een bepaalde groep respondenten, vaak gaat de interesse uit naar de groepen die passen bij de extreme kwantielen. Binnen het onderwerp

van deze scriptie zou dat bijvoorbeeld betekenen dat er specifiek gekeken wordt naar de groep mensen met lange reistijden.

- Volgens de Gauss–Markov theorie is OLS de meest efficiënte lineaire manier om te schatten, mits er voldaan is aan de aannames uit sectie 2.1. Met meest efficiënt wordt bedoeld dat het de methode de kleinste variantie heeft. In het geval dat de errorterm normaal verdeeld is, is OLS ook de beste methode in vergelijking met niet-lineaire methodes (Koenker & Bassett, 1978, p. 36). De mate van efficiëntie bij niet-lineaire kwantielregressie is hoger als de data anders verdeeld is, denk aan Laplace- of Cauchy-verdelingen.

Het is dus afhankelijk van de dataset en naar waar de interesse uitgaat welke schattingsmethode meer geschikt is.

3 Theoretisch kader economisch perspectief

Bij het maken van een keuze wordt een afweging gemaakt; de kosten en baten van de te nemen keuze worden tegen elkaar afgewogen. De keuze in deze scriptie is het wel of niet aangaan van reistijd voor het werk van een bepaalde duur. In veel economische modellen wordt een baan gekarakteriseerd met alleen de variabele van het loon (van den Berg & Gorter, 1996, p. 2). Een reistijd wordt afgelegd om een baan te verkrijgen, daarom staat loon aan de baten kant van de afweging. Hier tegenover, aan de kosten kant, staat de tijdsinvestering van de reistijd. Kortom, in dit economisch model wordt het besluit om reistijd aan te gaan gemaakt aan de hand van loon/reistijd-combinaties (van den Berg & Gorter, 1996, p. 3).

Reistijd bij man en vrouw

De reistijd naar het werk hangt uiteraard samen met de afstand tussen huis en werk. Hoe groter de radius van het zoekgebied naar een baan, hoe langer de reistijd, maar hoe meer potentiële banen om uit te kiezen. Hoe meer banen er zijn om uit te kiezen, hoe groter de kans is om een baan te vinden waarvan de baten hoger zijn. Hieruit volgt dat hoe langer de reistijd is, hoe meer kans er is op een goed betaalde baan. Tussen mannen en vrouwen vanaf 23 jaar is er sprake van een inkomensongelijkheid, ook als er gecorrigeerd wordt voor zaken als opleidingsniveau en leeftijd. Deze ongelijkheid is over de jaren 1993 tot 2002 afgenomen (Plantenga & Sjoerdsma, 2002). Volgens de uitkomsten van dit onderzoek hebben vrouwen een minder gunstige loon/reistijd-combinatie dan mannen.

Van origine is de rolverdeling zodanig dat de man werkt en de vrouw het huishouden regelt en de zorgtaken op zich neemt. Een nadeel van reizen voor werk is dat er in die reistijd geen zorg- of huishoudtaken uitgevoerd kunnen worden. Uit onderzoek is gebleken (Souren, 2006) dat het arbeidsmarktgedrag van werkenden samenhangt met de zorgtaken die zij thuis op zich nemen (Souren, 2006), een alternatief voor het verzorgen van kinderen is namelijk betaald werk. Mannen en vrouwen van 35–44 jaar vormen de grootste groep met kinderen en daarom ook de grootste groep met zorgtaken. De groep van 50–54 jaar bevat de meeste personen die zorgen voor een (zieke) naaste, twee derde hiervan is vrouw (Souren, 2006). Bij deze zorg voor naasten bestaat geen relatie met de arbeidsdeelname, echter kan dit wel meespelen bij de afweging om liever dichtbij huis te werken. Aan de hand van voorgaande literatuur wordt de volgende hypothese opgesteld:

Hypothese 1: Mannen zijn bereid langer te reizen in het woon-werkverkeer dan vrouwen.

Veranderingen door de tijd heen

Het tijdsbestek waarnaar deze scriptie onderzoek doet, is van 1985 tot en met 2014. In die jaren hebben er veranderingen plaatsgevonden, die invloed hebben op de gemaakte afweging bij het aangaan van reistijd. Ten eerste is het opleidingsniveau van vrouwen is flink gestegen (Emancipatiemonitor, 2016). Uit onderzoek (Molnár, 2004) volgt dat mensen met een hoger opleidingsniveau over het algemeen verder naar hun werk reizen. Werkenden met een hbo of universitaire achtergrond wonen gemiddeld 6 kilometer verder van hun werk dan mensen met vbo/mavo als hoogst genoten opleiding (Molnár, 2004). Doordat vrouwen door de jaren heen steeds beter zijn opgeleid, zijn de kosten van het verzorgen van kinderen toegenomen (Bovenberg en Haan, 2016). Een alternatief voor het verzorgen van kinderen is namelijk betaald werk, die optie wordt door de betere opleiding en dus de hogere lonen steeds aantrekkelijker. Dit heeft als gevolg dat het aandeel van vrouwen in het het woon-werkverkeer gestegen is. In 2004 is de nieuwe 'Wet Basisvoorziening Kinderopvang' (WBK) ingevoerd, ouders ontvangen van de overheid een tegemoetkoming in de kosten voor de kinderopvang. Uit financiële overwegingen is het combineren van zorg en arbeid makkelijker geworden (Plantenga & Sjoerdsma, 2002).

Een andere verandering is de verbetering van de infrastructuur in Nederland. Door de geografische ligging van het woonhuis, kan het zijn dat iemand hoe dan ook vastzit aan een langere reistijd, aangezien er niet veel werkgelegenheid is in de buurt van het huis. In opdracht van het RIVM is onderzoek gedaan naar de bereikbaarheid arbeidsplaatsen binnen 45 minuten reistijd per auto en openbaar vervoer. Dit is in kaart gebracht voor 1995 en er is een voorspelling gedaan voor het jaar 2020. De bereikbaarheid van arbeidsplaatsen binnen 45 minuten is in de Randstad het grootst (Geurs en Ritsema van Eck, 2000, p. 45-65). De grootste verwachte groei van het aantal arbeidsplaatsen is ook te zien in de Randstad. Door voorspelde verbeteringen van het (spoor)wegennetwerk zal de bereikbaarheid arbeidsplaatsen over heel Nederland stijgen. Rondom grote steden wordt echter een verslechtering van de bereikbaarheid voorspeld. Investeringen in het spoorwegennetwerk zullen vooral terecht komen in de randstad. Een verbeterde infrastructuur heeft als gevolg dat reistijden korter worden. Naar aanleiding van de bovenstaande literatuur wordt de volgende hypothese gesteld:

Hypothese 2: Het effect van geslacht op reistijd is kleiner geworden in de periode van 1985 tot en met 2014.

Meerwaarde kwantielregressie

In sectie 2.3 zijn de verschillen tussen OLS en kwantielregressie besproken. Hieruit volgde dat OLS gevoeliger is voor uitschieters in de data. Doordat reistijd niet negatief kan zijn, zullen er slechts uitschieters naar boven voorkomen. Door de uitschieters naar boven, maar niet naar onder, zal de bij schattingen aan de hand van het gemiddelde de waarde iets hoger liggen dan wenselijk, bij kwantielregressie zal dit effect hooguit merkbaar zijn in de hoge kwantielen. Een ander verschil is dat kwantielregressie de mogelijkheid geeft om naar een bepaald deel van de verdeling te kijken. Aan de hand van de verdeling van de reistijd kunnen groepen gemaakt worden, namelijk een groep mensen met een korte reistijd of mensen met juist een lange reistijd. Het is mogelijk dat in de groep met lange reistijd variabelen een ander effect hebben dan in de groep mensen met korte reistijd. Bij het schatten met behulp van OLS wordt hier geen onderscheid in gemaakt, bij kwantielregressie is dat wel mogelijk. Ten slotte wordt in deze scriptie een effect over de tijd onderzocht. Het is aannemelijk dat de verdeling over de jaren zal veranderen. Doordat bij kwantielregressie per gewenst percentiel een regressie uitgevoerd wordt, wordt de veranderde verdeling duidelijk. OLS maskeert deze verandering meer, doordat er naar het gemiddelde van de verdeling gekeken wordt. Naar aanleiding van de bovenstaande verschillen wordt de volgende hypothese gesteld voor desbetreffende situatie:

Hypothese 3: Kwantielregressie levert een meerwaarde aan informatie op ten opzichte van OLS-regressie.

Zowel in het onderzoek van Molnár (2014) als Bakker, Jorritsma, en Olde Kalter (2010) is gekozen voor reisafstand als afhankelijke variabele. Reisafstand en reistijd zijn twee samenhangende variabelen, maar bevatten niet dezelfde informatie. Aan een afstand valt niet af te lezen of de reiziger te maken krijgt met files, het missen van een overstap of andere voorbeelden van verminderd comfort. Toch zijn dit zaken die een werknemer meeneemt in de afweging bij het kiezen voor een baan. In de literatuur worden meermaals alleen reizen met het openbaar vervoer en de auto meegenomen, omdat dit samen het grootste deel van de reizen vormt. Om een compleet beeld te schetsen van de verandering van het effect van geslacht, zullen alle vervoerswijzen moeten worden meegenomen.

4 Data

In deze sectie wordt behandeld welke data verkregen zijn en welke bewerkingen daarop zijn uitgevoerd, zodat slechts de data die voor dit onderzoek relevant zijn overblijven. Vervolgens worden de data statistieken van de overgebleven data gepresenteerd. Alle databewerkingen zijn uitgevoerd met Stata, in Appendix B en C is respectievelijk de in- en uitvoer te vinden.

De dataset die wordt gebruikt is afkomstig uit het arbeidsaanbodpanel (AAP). Dit panel is in 1985 opgezet om diverse aspecten van de arbeidssituatie van werkenden en niet-werkenden in Nederland in kaart te brengen. Belangrijke thema's daarbij zijn arbeidsmobiliteit, scholing, zoekgedrag naar een (andere) baan, arbeidsduur, meningen over arbeid en huidige baan. Na 1985 is dit onderzoek in elk even jaar herhaald. Het onderzoek is een panelonderzoek, waarbij (sommige) respondenten meerdere jaren achter elkaar deelgenomen hebben. Dit maakt het mogelijk na te gaan in hoeverre gebeurtenissen een effect hebben op latere gebeurtenissen. In totaal worden per meting ongeveer 4500 respondenten in 2300 huishoudens bevraagd (Vlasblom et al., 2014, p. 5).

4.1 Data bewerking

De complete dataset van DANS bevatte 74.444 observaties. Niet alle respondenten zijn geschikt voor dit onderzoek en niet alle waargenomen variabelen zijn relevant voor deze onderzoeksvraag. Uit deze data is al eens een meer compacte dataset gecreëerd [1], deze zal gebruikt worden voor het onderzoek. Hier zijn nog een aantal bewerkingen aan toegevoegd. Na alle bewerkingen is er een bruikbare dataset overgebleven met 44.333 observaties.

In dit onderzoek wordt het effect van geslacht op de reistijd door de jaren heen onderzocht. Reistijd zal de afhankelijke variabele zijn en geslacht een onafhankelijke variabele. Zoals in sectie 3 beschreven is, hebben meer variabelen invloed op de reistijd, waardoor het model gecorrigeerd zal worden door de volgende controlevariabelen: loon, leeftijd, hoogst genoten opleiding en provincie van het woonhuis van de respondent. Controlevariabelen zijn variabelen waar in principe niet de aandacht naar uit gaat, maar wel effect hebben op de afhankelijke variabele. Door het toevoegen van controlevariabelen wordt dit effect afzonderlijk gemeten. Zonder het toevoegen van de controlevariabelen bestaat de invloed natuurlijk ook, maar is terug te vinden in de errorterm of in de parameters van een andere variabele. Aangezien men wil dat $E(\epsilon|X) = 0$, is dit onwenselijk en zorgt het toevoegen van controlevariabelen dus voor een beter model. In deze scriptie wordt een effect

[1] Deze dataset heeft Sander Schenk, student aan de UU, onder leiding van prof. dr. Hassink in een mastercursus samengesteld.

over de tijd onderzocht, vandaar worden er drie tijdscategorieën gemaakt van ongeveer tien jaar. Het eerste tijdvak bevat de data van de enquêtes van 1985 tot en met 1995, de tweede van 1996 tot en met 2005 en de laatste van 2006 tot en met 2014.

De afhankelijke variabele in de modellen is reistijd in minuten van een enkele reis woon-werkverkeer. Alleen werkenden hebben te maken met woon-werkverkeer, dus in de overgebleven dataset zijn alleen de werkende respondenten behouden. Binnen de werkenden bestaat er een groep thuiswerkers met een reistijd gelijk aan 0 minuten, deze individuen zijn ook gewoon werkenden en doen dus mee aan het onderzoek. Reistijden groter dan 300 minuten zijn uit de data verwijderd, aangezien een enkele reistijd groter dan vijf uur niet aannemelijk is.

Het effect van geslacht is in een regressiemodel te vinden bij de coëfficiënt van de variabele 'geslacht'. De variabele is namelijk een dummy-variabele, oftewel 'geslacht' $\in \{0, 1\}$. De waarde van de variabele is 0 als de respondent een vrouw is en waarde 1 als de respondent een man is. Daarom is het mogelijk om aan de coëfficiënt aflezen hoeveel minuten een man meer of minder onderweg is dan een vrouw (zie sectie 2.2.2.).

Controlevariabelen
Volgens het model van de loon/reistijd-combinaties heeft loon effect op de bereidheid tot reistijd, vandaar dat loon een controlevariabele is. In het model is gekozen voor het log(loon), hiermee wordt de logaritme aan van het gemiddelde netto loon per maand, uitgedrukt in euro. Door het log(loon) te nemen wordt de ratio tussen de lonen meegenomen in de regressie in plaats van het absolute verschil. In het model wordt het netto loon gebruikt, omdat dit het loon is wat de werknemer ontvangt en waar dus de afweging besproken in sectie 3 mee wordt gemaakt. Op 1 januari 2002 is de euro ingevoerd in Nederland. Dit betekent dat ingevulde lonen in de enquêtes van 1985 tot en met 2000 ingevuld zijn in gulden (er is namelijk geen enquête afgenomen in 2001). Aangezien de euro het huidige betaalmiddel is, is er gekozen om de lonen uit de jaren tot en met 2000 te delen door 2,20371. Dit is de vastgestelde koers volgens de Rijksoverheid. Door de jaren heen is er sprake van, waardoor het lijkt dat de lonen stijgen. In dit model wordt er geen rekening gehouden met inflatie, aangezien er wordt aangenomen dat de lonen van alle respondenten nagenoeg op dezelfde manier meegroeien met de inflatie.

Ook leeftijd kan effect hebben op de bereidheid tot het aangaan van reistijd. Sectie 3 laat zien dat leeftijd samenhangt met de hoeveelheid zorgtaken (bijvoorbeeld door de komst van een kind), wat vervolgens samenhangt met de bereidheid tot reistijd. De schommeling in hoeveelheid zorgtaken doet

vermoeden dat het effect van leeftijd niet lineair is, daarom is er gekozen om weer dummy-variabelen te creëren. Er zijn zes groepen gemaakt, namelijk 16-24, 25-34, 35-44, 45-54, 55-64 en 65-68 jaar oud.

Uit onderzoek (Molnár, 2004) bleek dat hoger opgeleide werkenden verder weg woonden van hun werk. Dit doet vermoeden dat opleidingsniveau van invloed is op de bereidheid tot reistijd, daarom zal opleidingsniveau meegenomen worden als controlevariabele in het model. Om het hoogst genoten opleidingsniveau aan te geven wordt de Standaard Onderwijsindeling (SOI) uit 1978, bestaande uit zeven niveaus, gebruikt. Niveau 1 van de SOI bestaat uit onderwijs aan kleuters en niveau 2 is het basisonderwijs. Deze twee niveaus zijn samengevoegd tot groep 1, aangezien er in Nederland nagenoeg geen mensen zijn, die als hoogst genoten opleiding de kleuterklassen van het basisonderwijs hebben. Groep 2 is vmbo/lbo/mavo, groep 3 mbo/havo/vwo en groep 4 hbo. Niveau 6 bestaat uit een WO bachelor en niveau 7 WO master en opleidingen na het doctoraal examen. Deze twee groepen zijn ook samengevoegd tot groep 5, zodat er een grotere, dus meer representatieve, groep overblijft en er is aangenomen dat het verschil in banen tussen groep 6 en groep 7 klein is.

Door een hogere werkgelegenheid en meer infrastructuurverbeteringen in de randstad, is het mogelijk dat de provincie waar een respondent woont van invloed kan zijn op de reistijd. Bewoners uit de stad Amsterdam zijn niet gerekend tot de provincie Noord-Holland. Zij hebben een eigen groep, omdat er mogelijk een verschil is tussen het wonen in de hoofdstad of in wonen Noord-Holland, in verband met baanaanbod- en infrastructuurverschillen binnen een stad en (net) erbuiten (Geurs en Ritsema van Eck, 2000, p. 45-65).

Data selectie

Op het moment dat een respondent een vraag waaruit variabelewaarden voor het model gevormd worden niet heeft ingevuld, een reistijd groter dan 300 minuten ingevuld heeft of niet-werkend is, is de betreffende respondent uit de dataset verwijderd. Data verwijderen kan grote gevolgen hebben. Vandaar dat alle verwijderde data zijn onderzocht of de gegevens die wel ingevuld zijn, overeen komen met de data die overgebleven is, zie tabel 8 in Appendix C. Het blijkt dat de verwijderde groep respondenten naar verhouding meer vrouwen bevat, lager opgeleid is en een kortere reistijd heeft. Dit hangt samen met het gegeven dat meer vrouwen niet-werkend en lager opgeleid zijn (Vlasblom et al., 2014), en daarbij dat lager opgeleiden vaak dichterbij hun werk wonen. Aangezien in de overgebleven data elke educatie-groep nog vertegenwoordigd is en het percentage vrouwen alsnog 44,7% is, is de verwijdering van deze data geoorloofd.

4.2 Data statistieken

In tabel 1 zijn data statistieken van de gehele dataset weergegeven, daar is onder andere per geslacht af te lezen hoe de samenstelling van de groep participanten is. De leeftijds-, educatie- en provinciegroepen geven aan welk deel van de data binnen die categorie valt, hier is een standaard deviatie dus niet van belang.

Tabel 1: Data statistieken van de gehele dataset: 1985-2014

	Vrouwen		Mannen	
	Gem.	Stand. dev. [2]	Gem.	Stand. dev. [2]
Reistijd (min./enk. reis [3])	20,51	18,02	25,59	22,02
Loon (€/maand)	844,27	755,69	1265,32	1023,60
Logloon	6,34	0,98	6,84	0,81
16-24 jaar	0,141		0,098	
25-34 jaar	0,235		0,221	
35-44 jaar	0,280		0,283	
45-54 jaar	0,245		0,266	
55-64 jaar	0,097		0,130	
65-68 jaar	0,0014		0,0021	
Groningen	0,037		0,039	
Friesland	0,035		0,038	
Drenthe	0,032		0,030	
Overijssel	0,071		0,075	
Flevoland	0,030		0,029	
Gelderland	0,106		0,108	
Utrecht	0,066		0,061	
N-Holland	0,137		0,126	
Z-Holland	0,160		0,161	
Zeeland	0,036		0,039	
N-Brabant	0,138		0,139	
Limburg	0,063		0,071	
Amsterdam	0,088		0,084	
Basisonderwijs	0,038		0,058	
Vmbo/lbo/mavo	0,260		0,290	
Mbo/havo/vwo	0,393		0,348	
Hbo	0,236		0,213	
WO	0,073		0,092	

N = 44.333

[2] De leeftijds-, provincie- en educatievariabelen zijn dummy-variabelen, hierdoor geeft een standaard deviatie bij die variabelen geen relevante informatie.

[3] Reistijd in minuten voor een enkele reis

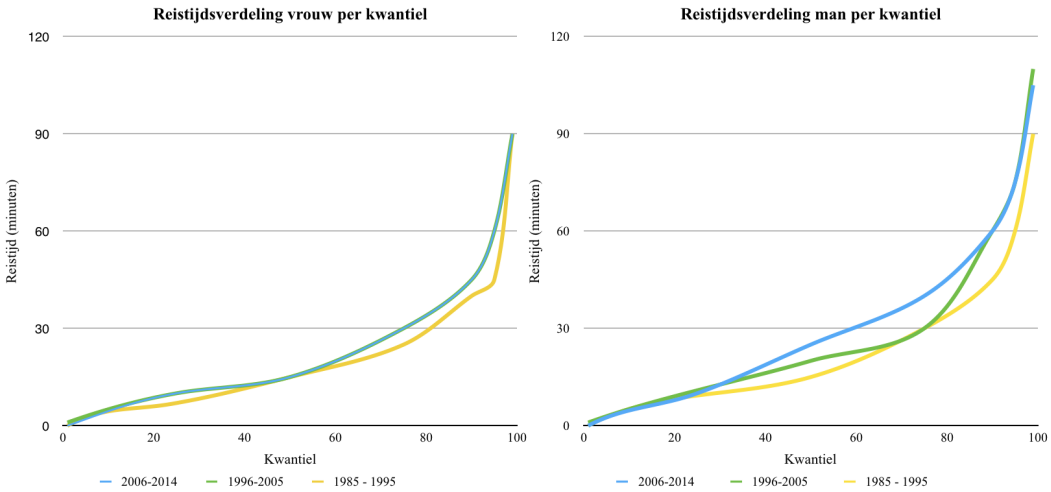
Reistijd

In tabel 2 staan de reistijden genoteerd zowel gemiddeld als per kwantiel. De reistijden per kwantiel zijn grafisch weergegeven in figuur 1. Daarnaast zijn in figuur 2 en 3 de verdelingen van reistijden afgebeeld.

Tabel 2: Reistijd (minuten) per kwantiel en gemiddeld

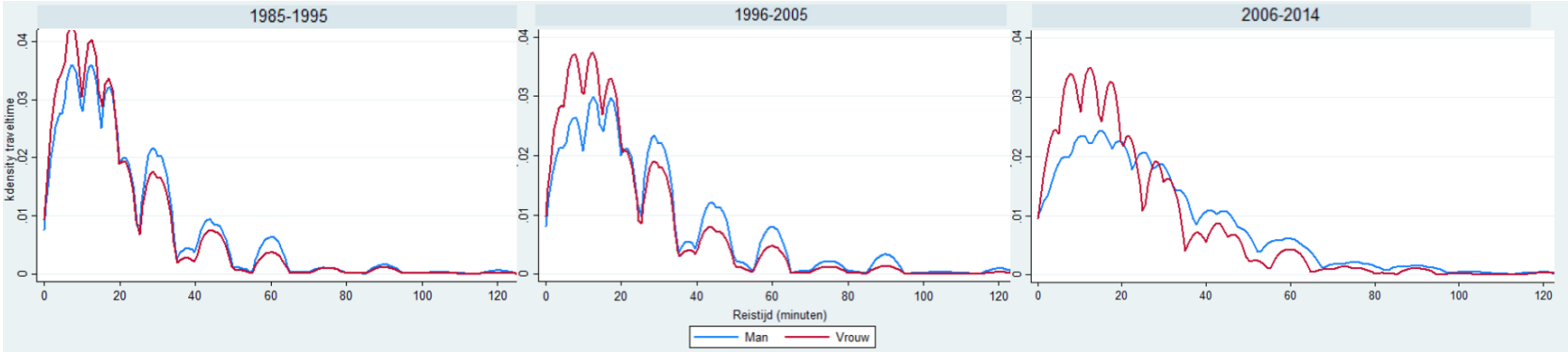
	1985-1995		1996-2005		2006-2014	
	Vrouw	Man	Vrouw	Man	Vrouw	Man
$\xi_{0,05}$	3	3	3	3	2	3
$\xi_{0,25}$	7	10	10	10	10	10
$\xi_{0,50}$	15	15	15	20	15	25
$\xi_{0,75}$	25	30	30	30	30	40
$\xi_{0,90}$	40	45	45	60	45	60
$\xi_{0,95}$	45	60	60	75	60	75
Gem.	18,5	21,8	20,4	26,5	21,7	28,4

Figuur 1: Reistijd (minuten) per kwantiel

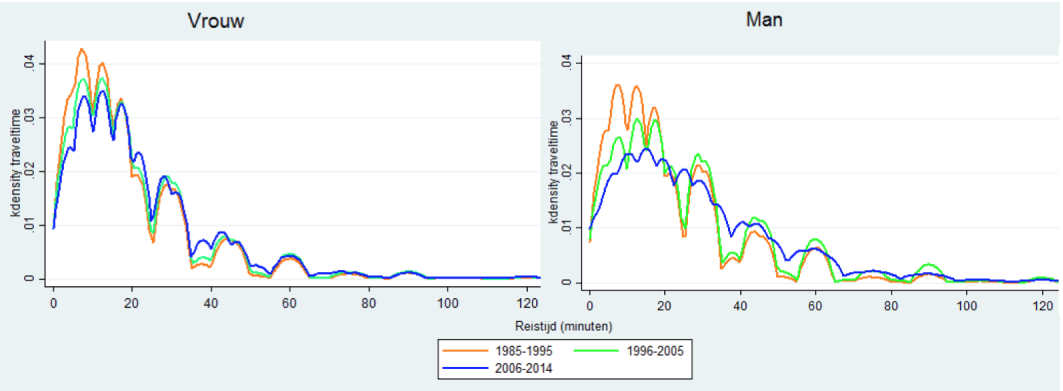


Allereerst is te zien dat mannen langer reizen dan vrouwen. Zowel mannen als vrouwen zijn in de loop der jaren langer gaan reizen, deze groei is sterker in de hogere kwantielen. Het verschil tussen jaarcategorie 2 en 3 is klein en bij mannen duidelijker aanwezig dan bij vrouwen. De verwachting wordt gewekt dat het effect van geslacht in hogere kwantielen sterker aanwezig is, doordat de grafiek bij mannen steiler is bij de hogere kwantielen. Aangezien beide gele lijnen een gelijkwaardig verloop hebben, de blauwe en groene lijn bij de vrouwen hoger, maar nagenoeg op elkaar liggen en bij mannen niet, doet dit vermoeden dat het effect van geslacht toegenomen is door de jaren

Figuur 2: Verdeling van de reistijd per tijdvak



Figuur 3: Verdeling van reistijd per geslacht



heen. Kwantielregressie lijkt van meerwaarde aangezien de gemiddelden hoger liggen dan de mediaan en de grafiek in de hoge kwantielen een sterke stijging laat zien.

In alle drie de grafieken van figuur 2 wordt duidelijk dat meer vrouwen kortere reistijd hebben dan mannen. Tot ongeveer twintig minuten blijft de lijn van de vrouw namelijk boven de lijn van de man, na twintig minuten draait dit om, vanaf dat moment zijn er meer mannen die de betreffende reistijd afleggen. Dit omslagpunt lijkt zich door de jaren heen iets hoger te worden. Ook is de grafiek is in de loop van der jaren breder en lager geworden, dit betekent dat mensen gemiddeld genomen langer zijn gaan reizen naar het werk.

In figuur 3 zijn dezelfde data gebruikt als in figuur 2, hier worden de verschillen per jaarcategorieën binnen hetzelfde geslacht duidelijk. Bij de mannen liggen de drie lijnen verder uit elkaar, wat betekent dat het gedrag bij mannen door de jaren heen sterker veranderd is dan bij vrouwen. In beide grafieken is te zien dat de kortste reistijden het meest in de eerste jaarcategorie voorkwamen, gevolgd door de tweede en daarna de derde. Rond de dertig minuten ligt het omslagpunt, vanaf hier is volgorde precies andersom. In de loop der jaren zijn er dus minder mensen met een korte reistijd en meer mensen met een lange reistijd.

Zowel in figuur 2 als 3 is duidelijk te zien aan de pieken dat de respondenten hun reistijd afgerond hebben op kwartieren. Vandaar het grillig verloop van de grafieken. Aangezien deze afronding zowel naar boven als naar onder gedaan wordt, zal dit geen groot effect hebben op de resultaten.

Veranderingen over de tijd

In tabel 3, 4 en 5 zijn opnieuw data statistieken vermeld, maar nu afzonderlijk voor elke jaarcategorie. Hierin zijn niet alle gegevens gepresenteerd, slechts de reistijd, (log)lonen en de opleidingsniveau's van de participanten. Hier is voor gekozen omdat er bij deze variabelen duidelijke veranderingen over de jaren heen aanwezig zijn, bij de weggelaten variabelen was dit niet geval. Uit tabel 3, 4 en 5 wordt duidelijk dat de gemiddelde reistijden toenemen en dat deze toename bij mannen, zowel relatief als absoluut, iets sterker is dan bij vrouwen. Ook is het opleidingsniveau van zowel man als vrouw sterk gestegen, in dit geval is de stijging bij vrouwen relatief sterker dan bij mannen. Toch blijven er meer mannen hoog opgeleid (groep 4 en 5), dit verschil is in de derde jaarcategorie echter niet groot meer. De lonen zijn ook sterk gestegen over de afgelopen jaren, dit wordt grotendeels veroorzaakt door inflatie. De lonen van de vrouwen zijn relatief iets sterker gestegen, maar zijn nog steeds lager dan het loon van mannen. Dit valt (deels) te verklaren doordat meer vrouwen in deeltijd werken (Vlasblom et al., 2014, p. 24).

Tabel 3: Data statistieken 1985-1995

	Vrouwen		Mannen	
	Gem.	Stand. dev.	Gem.	Stand. dev.
Reistijd (min./enk. reis) [4]	18,54	17,14	21,83	19,04
Loon (€/maand)	292,15	212,70	524,04	236,88
Logloon	5,48	0,69	6,18	0,42
Basisonderwijs	0,074		0,103	
Vmbo/lbo/mavo	0,355		0,359	
Mbo/havo/vwo	0,388		0,332	
Hbo	0,158		0,165	
WO	0,024		0,041	

N=12.869

Tabel 4: Data statistieken 1996-2005

	Vrouwen		Mannen	
	Gem.	Stand. dev.	Gem.	Stand. dev.
Reistijd (min./enk. reis)[4]	20,39	17,75	26,53	23,43
Loon (€/maand)	711,63	578,98	1188,68	910,84
Logloon	6,24	0,90	6,86	0,67
Basisonderwijs	0,031		0,043	
Vmbo/lbo/mavo	0,294		0,318	
Mbo/havo/vwo	0,374		0,334	
Hbo	0,239		0,212	
WO	0,061		0,093	

N= 14.416

Tabel 5: Data statistieken 2006-2014

	Vrouwen		Mannen	
	Gem.	Stand. dev.	Gem.	Stand. dev.
Reistijd (min./enk. reis)[4]	21,71	18,60	28,24	22,77
Loon (€/maand)	1258,33	830,27	2035,14	1035,00
Logloon	6,91	0,78	7,45	0,72
Basisonderwijs	0,023		0,029	
Vmbo/lbo/mavo	0,180		0,198	
Mbo/havo/vwo	0,410		0,376	
Hbo	0,277		0,259	
WO	0,110		0,137	

N=17.048

5 Statistische modellen

In deze sectie worden de modellen voor de kwantiel- en OLS-regressie gepresenteerd. Ze zijn gemaakt aan de hand van de besproken variabelen in sectie 4. Daarnaast worden per regressiemodel specificaties gegeven.

5.1 Model OLS

Het lineaire regressiemodel ziet er als volgt uit:

$$reistijd = \beta_0 + \beta_1 man + \beta_2 log(loon) + \beta_3 leeftijdscategorie \\ \beta_4 educatie + \beta_5 provincie + \epsilon_1 \quad (9)$$

Aannames OLS
Vervolgens is het belangrijk om te onderzoeken of de data en het model voldoen aan de aannames uit sectie 2.1, waar aan voldaan moet zijn om de resultaten van de regressie van waarde te laten zijn.

In formule (9) wordt duidelijk dat er sprake is van een lineair regressiemodel. De regressies worden per tijdvak uitgevoerd, daardoor zijn er drie regressies met, op chronologische volgorde, 12.869, 14.416 en 17.048 metingen. Aan de aanname $n > k + 1 \Leftrightarrow n > 6$ is duidelijk voldaan. Daarbij is vereist dat er sprake is van strikte exogeniteit, geen perfecte multicollineariteit en de verdeling van de errortermen onafhankelijk en gelijk verdeeld (i.i.d.) zijn. Door de combinatie van de Centrale Limiet Stelling, de grote dataset en de aanname dat de errorterm i.i.d. is, hoeft er geen normaliteit vereist te worden.

Een aantal van deze aannames zijn getest in Stata. De multicollineariteit wordt onderzocht door middel van de *Variance Inflation Factor*. Uit de VIF-test (Appendix C, figuur 5) is gebleken dat in jaarcategorie 1 sprake is van multicollineariteit tussen de leeftijdsgroepen en in jaarcategorie 3 tussen de educatiegroepen. De strikte exogeniteit kan niet worden getest. In figuur 6 zijn de de histogrammen van de residuen afgebeeld. Door de enorme pieken bij 0, mag aangenomen worden dat er in ieder geval sprake is van $E(\epsilon) = 0$. De homoskedasticiteit van de errortermen wordt getest met de White-test (Appendix C, figuur 7), hieruit blijkt dat in elke tijdscategorie sprake is van homoskedasticiteit van de errorterm.

[4] Reistijd in minuten voor een enkele reis

5.2 Model kwantielregressie

In het model voor de kwantielregressie worden dezelfde variabelen gebruikt als in het OLS-model, zodat het kwantielregressiemodel er als volgt uitziet:

$$\begin{aligned} reistijd_q = \beta_0(q) + \beta_1(q)man + \beta_2(q)log(loon) + \\ \beta_3(q)leeftijdscategorie + \beta_4(q)educatie + \beta_5(q)provincie + \epsilon_2 \end{aligned} \quad (10)$$

De totstandkoming van de afhankelijke variabele zal op meerdere kwantielen bekeken worden, om reizigersgroepen met verschillende reistijden met elkaar te vergelijken. De kwantielen waarvoor de reistijd geschat gaat zullen variëren tussen $q = 0.05$ en $q = 0.95$. Zo wordt een goed beeld gevormd over de gehele verdeling, maar zullen de uitschieters de waardes niet onwenselijk beïnvloeden.

Aan de aanname voor voldoende metingen is natuurlijk ook voor dit model voldaan, zie sectie 5.1. De heteroskedasticiteit van de errorterm wordt onderzocht in verband met het snijden van de geschatte lijnen en het verdwijnen van de natuurlijke ordening van de kwantielen. De resultaten hiervan zijn te zien in figuur 8, Appendix C. Het blijkt dat in jaarcategorie 1 sprake is van heteroskedasticiteit, die in de kleine kwantielen groter is dan in de hoge kwantielen. In jaarcategorie 2 en 3 is er sprake van homoskedasticiteit.

6 Resultaten

In deze sectie zullen de resultaten van de regressies vermeld en besproken worden. Hieronder in tabel 6 en 7 staan respectievelijk de waardes van de coëfficiënten voor de variabele 'geslacht' van de OLS- en kwantielregressie. Aan die coëfficiënt, β_1, valt af te lezen wat de extra reistijd in minuten is voor een man ten opzichte van een vrouw. De coëfficiënten van de overige variabelen zijn te vinden in Appendix C.

Vervolgens wordt het significantieniveau van de gevonden coëfficiënten bekeken. Om te bepalen of de gevonden coëfficiënten significant zijn, worden de waarden vergeleken met de nulhypothese; namelijk dat er geen effect is tussen de onafhankelijke en afhankelijke variabelen. In het geval van de nulhypothese is β_1 gelijk aan de nulvector. Aan de hand van een tweezijdige p-waarde van het testen van de nulhypothese is af te lezen of het gevonden effect van geslacht significant is. Het significantieniveau wordt overzichtelijk weergegeven met sterren; met drie sterren wordt aangegeven dat de coëfficiënt voldoet aan een significantieniveau met alfa-waarde van 0.01, oftewel $p < 0.01$, twee sterren duidt op alfa-waarde 0.05 en één ster op 0.10.

Tabel 6: Man-coëfficiënten OLS-regressie

Tijdvak	β_1	Std. err.
1985 – 1995	0,591	0,391
1996 – 2005	3,795 ***	0,375
2006 – 2014	3,164 ***	0,336

Tabel 7: Man-coëfficiënten kwantielregressie

	1985-1995		1996-2005		2006-2014	
Kwantiel	β_1	Std. err.	β_1	Std. err.	β_1	Std. err.
$q = 0.05$	0,266	0,283	-.398	0,274	-0,260	0,291
$q = 0.19$	0,151	0,331	.605 **	0,276	0,620 ***	0,219
$q = 0.25$	-0,0035	0,266	1,146 ***	.271	0,885 ***	0,275
$q = 0.50$	0,329	0,3599	2,416 ***	.353	2,783 ***	0,357
$q = 0.75$	0,953	0,591	4,997 ***	.560	4,789 ***	0,512
$q = 0.90$	1,958 *	1,102	8,673 ***	1.004	7,037 ***	0,862
$q = 0.95$	2,816 *	1,581	12,557 ***	1.530	9,334 ***	1,372

Beschrijving resultaten
Zowel uit de resultaten van OLS (tabel 6) als kwantielregressie (tabel 7) volgt dat er in de eerste jaarcategorie geen significant verschil aanwezig is bij de geslachtscoëfficiënten. Dit betekent dat in de periode van 1985 tot en met 1995 er geen aanwijsbaar verschil is in het gedrag van mannen en vrouwen in het aangaan tot reistijd dat te wijten is aan geslacht.

In tijdvak 2 en 3 is er zowel bij OLS als kwantielregressie wel sprake van significantie van de geslachtscoëfficiënt. Bij kwantielregressie geldt dit echter niet in de lage kwantielen. Bij het 19^e kwantiel bevindt zich het omslagpunt, waar jaarcategorie 2 en 3 minstens een significantieniveau hebben van $\alpha = 0.05$. Dit betekent dat in de groep mensen met een relatief korte reistijd geen aanwijsbaar verschil aanwezig is door een verschil in geslacht, maar vanaf een reistijd langer dan ongeveer 10 minuten wel.

In jaarcategorie 2 en 3 geldt: hoe hoger de kwantielen worden, hoe hoger de coëfficiënt wordt. Dit duidt erop dat hoe langer de reistijden worden, hoe groter het verschil tussen man en vrouw wordt. In jaarcategorie 2 zijn de verschillen tussen man en vrouw echter groter dan in jaarcategorie 3. Dit duidt erop dat het verschil tussen man en vrouw eerst groter is geworden en daarna weer iets kleiner.

Significantie tussen de tijdvakken
Er is af te lezen in tabel 6 en 7 dat de waarden van de geslachtscoëfficiënt in tijdvak 2 hoger zijn dan in tijdvak 3. Deze coëfficiënten zijn significant ten opzichte van de nulhypothese in de eigen regressie. Om uitspraken te kunnen doen over de daling moet worden bepaald of de coëfficiënten ook significant zijn ten opzichte van elkaar. Voor de volledigheid wordt tijdvak 1 ook meegenomen. Het is namelijk mogelijk dat een variabele niet significant is in de eigen regressie, maar wel significant ten opzichte van andere regressies. De nulhypothese voor deze test is als volgt (UCLA: Statistical Consulting Group):

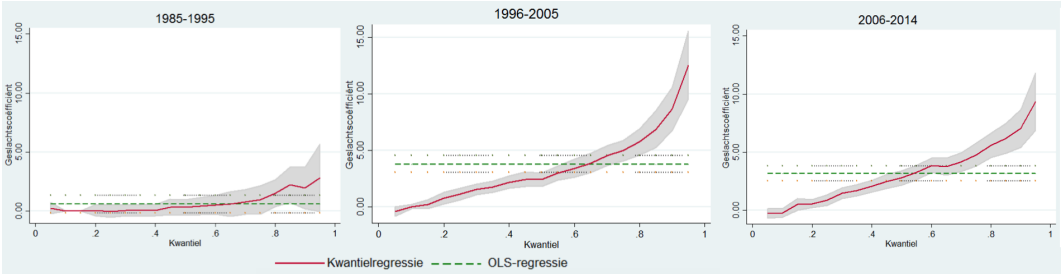
$$H_0 : \beta_{1,categorie1} = \beta_{1,categorie2} = \beta_{1,categorie3}$$

In Appendix B en C is te lezen wat de Stata in- en output van deze F-test is (figuur 9). Het blijkt dat alle drie de geslachtscoëfficiënten significant zijn ten opzichte van elkaar met een $p = 0,000$.

Verschillen OLS en kwantielregressie

In figuur 4 zijn de geslachtscoëfficiënten per jaarcategorie en per regressiemethode afgebeeld. Voor beide methoden is ook het 95% betrouwbaarheidsinterval van de coëfficiënten weergegeven. In tijdvak 1 zit relatief een grote groei in de coëfficiënten naarmate de kwantielen groter worden, maar door het lage significantieniveau blijft de OLS-waarde binnen het significantiegebied van de kwantielregressie.
In het 2^e en 3^e tijdvak stijgen de coëfficiënten veel sterker naarmate de kwantielen groter worden en raken de betrouwbaarheidsintervallen elkaar niet. Aangezien er een groot verschil zit tussen de coëfficiënten uit de lage en hoge kwantielen, die een bepaalde groepen werkenden vertegenwoordigen, levert kwantielregressie meer informatie dan OLS.
Zoals het gemiddelde het uitgangspunt is bij OLS is de mediaan dat voor kwantielregressie. Daarom worden de mediaan-regressiewaarden nog afzonderlijk vergeleken met de OLS-waarden. In tijdvak 1 zijn met beide regressievormen de coëfficiënten niet significant. De coëfficiënten van de mediaanregressie zijn lager dan de OLS-regressie coëfficiënten, wat verklaard kan worden door uitschieters in de dataset die bij kwantielregressie minder invloed hebben. Bij OLS neemt de coëfficiënt in tijdvak 3 af ten opzichte van tijdvak 2. Bij de mediaanregressie is sprake van het tegenovergestelde.

Figuur 4: Geslachtscoëfficiënten per regressiemethode per jaarcategorie



7 Conclusie en discussie

7.1 Conclusie

In deze scriptie is gezocht naar een antwoord op de vraag: 'Is het effect van geslacht op de reistijd voor het woon-werkverkeer in Nederland veranderd in de periode van 1985 tot en met 2014?' Hiervoor is een empirisch onderzoek uitgevoerd door middel van OLS en kwantielregressie aan de hand van het arbeidsaanbodpanel.

In elke jaarcategorie bleek dat mannen vaker langer reizen naar hun werk dan vrouwen. Hypothese 1 kan dus aangenomen worden; mannen zijn bereid langer te reizen in het woon-werkverkeer dan vrouwen.

Uit de resultaten is gebleken dat het verschil in reistijd tussen man en vrouw in de periode van 1985 tot 1995 niet is toe te schrijven aan het verschil in geslacht. Ook in de lage kwantielen, oftewel de groep werkenden met een reistijd tot ongeveer tien minuten, kan het verschil tussen man en vrouw niet verklaard worden door het verschil in geslacht. In de periode van 1996 tot en met 2014 is vanaf een reistijd van ongeveer tien minuten het verschil in reistijd (gedeeltelijk) te verklaren door geslacht. Hoe langer de reistijd is van de werkenden, hoe sterker het effect van geslacht is. In de periode van 1996 tot 2005 is het effect van geslacht sterker dan in de periode van 2006 tot en met 2014. Hypothese 2 wordt dus verworpen; het effect van geslacht op reistijd is namelijk niet steeds kleiner geworden.

Ten slotte wordt er antwoord gegeven op de onderzoeksvraag. Het effect van geslacht op reistijd is sterk gestegen tussen de tijdvakken 1985-1995 en 1996-2005, maar dit effect is weer iets afgenomen tussen de tijdvakken 1996-2005 en 2006-2014.

In deze scriptie werd ook onderzoek gedaan naar de meerwaarde van kwantielregressie boven OLS voor dit specifieke praktijkvoorbeeld. In deze situatie was kwantielregressie van toegevoegde waarde, aangezien er een groot verschil zit tussen de coëfficiënten uit de lage en hoge kwantielen. Deze verschillen tussen reizigersgroepen bleven onzichtbaar bij OLS. De resultaten hebben ten slotte uitgewezen dat OLS gevoeliger is voor uitschieters dan kwantielregressie, doordat de geslachtscoëfficiënten bij OLS hoger waren dan die van de mediaanregressie. Hypothese 3 wordt dus aangenomen; in dit praktijkvoorbeeld levert kwantielregressie een meerwaarde aan informatie ten opzichte van OLS.

7.2 Discussie

Voor deze scriptie zijn de resultaten van het arbeidsaanbodpanel van 1985 tot en met 2014 gebruikt, hieruit zijn 43.333 enquêtes gebruikt, ingevuld door Nederlanders van 16 tot en met 66 jaar die representatief zijn voor de Nederlandse beroepsbevolking (Vlasblom et al., 2014, p. 12). Op basis hiervan kan gesteld worden dat de resultaten van dit onderzoek generaliseerbaar en dus extern valide zijn.

De algehele stijging van de reistijd kan verklaard worden door een verbeterde infrastructuur. Een gevolg van een verbeterde infrastructuur is een verhoogd comfort, waardoor een langere reis als minder vervelend ervaren wordt.

In de de jaren 1985-1995 viel het verschil tussen man en vrouw niet te verklaren door het verschil in geslacht. Wel blijkt uit de resultaten dat in alle jaren een duidelijk verband bestaat tussen opleidingsniveau en de bereidheid tot reistijd (Appendix C), namelijk hoe lager het opleidingsniveau, hoe korter de reistijd. Aan het verschil tussen mannen en vrouw ligt dus mogelijk wel ten grondslag dat vrouwen in deze periode over het algemeen nog een beduidend lager opleidingsniveau hadden.

Tussen tijdvak 1 en tijdvak 2 is het effect van geslacht sterk toegenomen. Dit betekent dat puur het verschillende geslacht meer effect heeft gekregen óf dat er sprake is van een eigenschap die bij veel mannen óf vrouwen voorkomt niet is meegenomen in het model. De eerste optie is het meest interessant, want dat zou betekenen dat vrouwen een intrinsieke reden hebben waardoor ze minder bereid zijn om ver te reizen. De tweede optie kan echter ook waar zijn. In dit onderzoek is de werksector van de respondent niet meegenomen in het model, maar dit kan wel meespelen bij het verklaren van dit resultaat. Meer vrouwen hebben in deze periode een baan gekregen, echter werkten veel vrouwen in vrouwensectoren bijvoorbeeld zorg en onderwijs. Het is mogelijk dat er een verband bestaat tussen sectoren en reistijd. Tussen tijdvak 2 en tijdvak 3 nam het effect van geslacht weer iets af. Sinds 2005 is er een daling te zien in het verschil in richtingkeuze op (hoge)scholen en universiteiten (Emancipatiemonitor, 2016). Dit zou dus als gevolg hebben dat het effect van geslacht weer is afgenomen tussen tijdvak 2 en tijdvak 3. Deze theorie is niet onderzocht in dit onderzoek, maar zou bij vervolgonderzoek interessant kunnen zijn.

7.3 Evaluatie onderzoek

Een beperking van dit onderzoek is het minimaal gebruik van controlevariabelen. De regressiemodellen waren specifieker geweest als er meer controleva-

riabelen toegevoegd waren, variabelen die in vervolgonderzoek meegenomen zouden kunnen worden zijn voltijd of deeltijd werken, de werksector, mantelzorgen en (aantal) kinderen. Ook is er data verwijderd in dit onderzoek. Zoals in tabel 9 af te lezen is, is de groep respondenten van de verwijderde data niet geheel gelijk aan de behouden groep respondenten. De gemiddelde reistijd is een stuk lager, wat waarschijnlijk samenhangt met het feit dat er meer vrouwen in die groep zitten. Ook is het opleidingsniveau is een stuk lager, wat ook weer samenhangt met de grote hoeveelheid vrouwen in de groep. Advies voor herhaalonderzoek is om individuen die een vraag, waaruit een variabele gevormd wordt, niet ingevuld hebben niet te verwijderen uit de data. Ook deze data kan gebruikt worden door deze variabele niet meegenomen te laten worden bij de berekening van het onderdeel waar die variabele in voorkomt. De ingevulde waardes van de individu op de andere variabelen worden in andere onderdelen van de berekening wel meegenomen.

Mogelijk is de aanname wat betreft de afwezigheid van autocorrelatie te sterk vanwege de paneldata structuur van de gebruikte data. Dit betekent dat de regressie minder efficiënt is dan aangenomen is. Uit de test voor multicollineariteit bleek dat er in jaarcategorie 1 sprake is van multicollineariteit tussen de leeftijdsgroepen en in jaarcategorie 3 tussen de educatiegroepen. Dit heeft een algemene verminderde efficiëntie in jaarcategorie 1 en 3 als gevolg, maar aangezien dit onderzoek gefocust was op het effect van geslacht, heeft de multicollineariteit geen invloed op de resultaten.

Een andere suggestie voor vervolgonderzoek is dat er meer gebruik gemaakt wordt van de panelstructuur. Door de panelstructuur kunnen veranderingen over tijd beter in kaart gebracht worden, omdat steeds dezelfde respondenten gevolgd worden. De belangrijkste modellen zijn *Fixed Effects* (FE), bij interesse in een variabele die varieert over tijd, en *Random Effects* (RE), voor als de niet-tijdsvariërende predictoren ook meegenomen moeten worden. Zowel OLS als kwantielregressie kunnen gebruik maken van FE en RE.

Referenties

[1] Angrist, J., & Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press

[2] Bakker, P., & Jorritsma, P., & Olde Kalter, M. (2010). *Woon-werkverkeer als drijvende kracht achter groei automobiliteit* Roermond: Colloquium Vervoersplanologisch Speurwerk

[3] Berg, G. van der, & Gorter, C. (1996). *Job Search and Commuting Time.* Amsterdam, Nederland: Vrije Universiteit

[4] Bovenberg, L., & Haan, F. (2016). *Economieonderwijs.* Meppel, Nederland: Ten Brink. pagina 29 en 71

[5] Davino, C., & Furno, M., & Vistocco, D. (2013). *Quantile Regression.* Somerset: John Wiley & Sons

[6] Geurs, K., & Ritsema van Eck, J. (2000). *Effecten van een compacte verstedelijkings- variant op mobiliteit, bereikbaarheid, CO2-emissie en geluid.* Bilthoven, Nederland: RIVM. pagina 45-65

[7] Koenders, P., & Deursen, C. van,. (2008). Reizen naar en voor het werk en verzuim in de banksector. *TBV, 16 (4).* pagina 143

[8] Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46 (1)*, pagina 33-50

[9] Molnár, H. (2004). Woon-werkverkeer. *Sociaal-economische trends, 4e kwartaal 2004.* pagina 47

[10] Plantenga, J., & Sjoerdsma, S. (2002). *The Gender Pay Gap in the Netherlands.* Utrecht.

[11] Portegijs, W., & Brakel, M. van den. (2016). *Emancipatiemonitor 2016.* Den Haag, Nederland: SCP en CBS. pagina 43 - 44

[12] Souren, M. (2006). Zorgtaken en arbeidsparticipatie. *Sociaal-economische trends, 3e kwartaal.* pagina 32-35

[13] UCLA: Statistical Consulting Group. Geraadpleegd van: https://stats.idre.ucla.edu/stata/faq/how-can-i-compare-regression-coefficients-across-3-or-more-groups/

[14] Vlasblom, J., & Voogd-Hamelink, M. de, & Echtelt, P. van, (2014). *Aanbod van Arbeid 2014.* Den Haag: Sociaal en Cultureel Planbureau

[15] Voogd-Hamelink, M. de, & Echtelt, P. van, & Croezen, S., & Vlasblom, J., (2016). *Aanbod van Arbeid 2016.* Den Haag: Sociaal en Cultureel Planbureau

A Appendix: Wiskundige bewijzen

(1) Bepaling μ

Te bewijzen: $\mu = \arg\min_{c\in\mathbb{R}} E(Y-c)^2$

$$\mu = \arg\min_{c\in\mathbb{R}} \left(\int_{-\infty}^{\infty} (y-c)^2 f_Y(y)dy \right)$$

$$\mu = \arg\min_{c\in\mathbb{R}} \left(\int_{-\infty}^{\infty} y^2 f_Y(y)dy - 2c \int_{-\infty}^{\infty} y f_Y(y)dy + \int_{-\infty}^{\infty} c^2 f_Y(y)dy \right)$$

Dit is een minimaliseringsprobleem, daarom wordt de functie gedifferentieerd naar c en de afgeleide gelijk gesteld aan 0 om het minimum te bepalen. Aangezien in de eerste integraal geen c voorkomt, valt deze term weg.

$$\frac{d}{dc}\left(-2c \int_{-\infty}^{\infty} y f_Y(y)dy \right) + \frac{d}{dc}\left(c^2 \int_{-\infty}^{\infty} f_Y(y)dy \right) = 0$$

$$-2 \int_{-\infty}^{\infty} y f_Y(y)dy + 2c \int_{-\infty}^{\infty} f_Y(y)dy = 0 \Leftrightarrow 2E(Y) = c \int_{-\infty}^{\infty}$$

Hieruit volgt $E(Y) = c = \mu$

(2) Schatter $\hat{\beta}$ voor OLS

De verwachte waarde van de schatter moet gelijk zijn aan de werkelijk waarde van β, dus: $E(\hat{\beta}|X) = \beta$

$E(\hat{\beta}|X) = E((X'X)^{-1}X'Y|X)$ Substitueer $Y = X\beta + \epsilon$

$= E((X'X)^{-1}X'(X\beta + \epsilon)|X)$

$= E((X'X)^{-1}(X'X)\beta|X) + E((X'X)^{-1}X'\epsilon|X)$

$= \beta + (X'X)^{-1}X'E(\epsilon|X)$ Vanwege strikte exogeniteit geldt:

$= \beta + (X'X)^{-1}X'0 = \beta$

(3) en (5) Bepaling ξ_q

Te bewijzen: $\xi_{\mathrm{q}} = \arg\min_{c\in\mathbb{R}} E[\rho_{\mathrm{q}}(X-c)]$

Bewijs:

$$\arg\min_{c\in\mathbb{R}} \left((1-q) \int_{-\infty}^{c} (c-x) f_X(x)dx + q \int_{c}^{\infty} (x-c) f_X(x)dx \right)$$

Aangezien dit een minimaliseringsprobleem is, wordt de functie gedifferentieerd naar c en de afgeleide gelijk gesteld aan 0 om het minimum te bepalen.

$$(1-q) \cdot \frac{d}{dc} \int_{-\infty}^{c} (x-c) f(x)\, dx + q \cdot \frac{d}{dc} \int_{c}^{\infty} (c-x) f(x)\, dx = 0$$

$$(1-q)\int_{-\infty}^{c} f(x)\,dx - q\cdot\int_{c}^{\infty} f(x)\,dx = 0 \Leftrightarrow (1-q)\int_{-\infty}^{c} f(x)\,dx = q\int_{c}^{\infty} f(x)\,dx$$

$$(1-q)\int_{-\infty}^{c} f(x)\,dx = q\left(1-\int_{-\infty}^{c} f(x)\,dx\right) \Leftrightarrow (1-q+q)\int_{-\infty}^{c} f(x)\,dx = q$$

$$\int_{-\infty}^{c} f(x)\,dx = q$$

Het bewijs van formule (3) volgt vanzelf door $q = \frac{1}{2}$ in te vullen. Dan ontstaat er namelijk

$$\int_{-\infty}^{c} f(x)\,dx = 0.5$$

Dit betekent dat $\xi_{0.5}$ gelijk is aan c.

(6) Schatter $\hat{\beta}$ voor kwantielregressie

Gegeven functie (5):

$$\xi_{\mathrm{q}}(y|X) = \arg\min_{c\in\mathbb{R}} E[\rho_{\mathrm{q}}(Y-c)] = X\hat{\beta}(q)$$

Introduceer functie d die de totale ρ-afwijkingen weergeeft.

$$d(y,\hat{y}) = \sum_{i=1}^{n} \rho_q(y-\hat{y})$$

$$d(y,\hat{y}(\beta)) = \sum_{i=1}^{n} \rho_q(y-\hat{y}(\beta))$$

$$d(y,\hat{y}(\beta)) = \sum_{i=1}^{n} \rho_q(y-X\hat{\beta})$$

Voor de kleinste totale afwijking moet het argument β bepaald worden waarvoor de afwijking minimaal is.

$$\hat{\beta}(q,y,X) = \arg\min_{\beta} \sum_{i=1}^{n} (\rho_q(y_i - X\beta))$$

(7) Monotone transformatie kwantiel

Te bewijzen: $h(\xi_{\mathrm{q}}(Y)) = \xi_{\mathrm{q}}(h(Y))$
Bewijs: $\xi_{\mathrm{q}}(Y) = \inf\{y \in \mathbb{R} : F_{\mathrm{Y}}(y) \geq q\}$. De cumulatieve verdelingsfunctie $F(Y)$is monotone stijgende functie en h is een monotone functie. Een monotone functie is dalend óf stijgend, dit betekent dat deze functie de orde bewaart. Bij het bepalen van $\xi_q(Y)|X)$ is de volgorde in de geordende dataset van belang en aangezien de volgorde niet veranderd bij een monotone transformatie geldt $h(\xi_{\mathrm{q}}(Y)) = \xi_{\mathrm{q}}(h(Y))$.

(8) Schaalverandering afhankelijke variabele

Te bewijzen: $\hat{\beta}(q, aY, X) = a\hat{\beta}(q, Y, X)$
Bewijs:

$$\hat{\beta}(q, y, X) = \arg\min_{\beta} \sum_{i=1}^{n} (\rho_q(y_i - X\beta))$$

$$\hat{\beta}(q, y, X) = \arg\min_{\beta} \sum_{i:y_i > x_i\beta} q|y_i - x_i\beta| + \sum_{i:y_i < x_i\beta} (1-q)|y_i - x_i\beta|$$

$$\hat{\beta}(q, ay, X) = \arg\min_{\beta} \sum_{i:ay_i > x_i\beta} q|ay_i - x_i\beta| + \sum_{i:ay_i < x_i\beta} (1-q)|ay_i - x_i\beta|$$

Na vermenigvuldiging van Y met $a > 0$, blijft de ordening, dat alle positieve errortermen $ay_i > x_i\beta$ vermenigvuldigt worden met q en de negatieve errortermen $ay_i < x_i\beta$ met $(1-q)$, gelijk.

$$\hat{\beta}(q, y, X) = \arg\min_{a\beta} \sum_{i:ay_i > ax_i\beta} q|ay_i - ax_i\beta| + \sum_{i:ay_i < ax_i\beta} (1-q)|ay_i - ax_i\beta|$$

Door β te vermenigvuldigen met a, wordt de vergroting van Y gecompenseerd. De gehele errorterm is nu echter a keer zo groot als bij $\hat{\beta}(q, Y, X)$. Dit betekent dat het effect van variabelen ook a keer te sterk gemeten wordt, wat betekent dat $\hat{\beta}$ ook a keer te groot geschat wordt.

$$\hat{\beta}(q, aY, X) = a\hat{\beta}(q, Y, X)$$

Te bewijzen: $\hat{\beta}(q, -aY, X) = -a\hat{\beta}(1-q, Y, X)$
Bewijs: Dit bewijs lijkt erg op het voorgaande bewijs. Slechts het afwijkende deel van het bewijs wordt behandeld. Als $y_i > x_i\beta$, dan $-ay_i < x_i\beta$. Dus de (zonder absoluut strepen) negatieve residuen worden nu vermenigvuldigd met q en de positieve residuen met $(1-q)$. Zoals in sectie 2.2.1 lezen is, hoort dat precies andersom. Dit gegeven gecombineerd tot het bewijs hierboven leidt tot:

$$\hat{\beta}(q, -aY, X) = -a\hat{\beta}(1-q, Y, X)$$

B Appendix: Commando's Stata

In deze Appendix zijn de commando's in Stata genoteerd. De namen van de variabelen wijzen voor zich, over het algemeen is de Engelse vertaling gebruikt. Met 'male' wordt de geslachtsdummy. De provincie-, opleidings- en leeftijdsdummy's zijn respectievelijk dp_i, dE_i en dL_i genoemd, waarbij de volgorde van nummering precies gelijk loopt met de volgorde uit tabel 1.

Sectie 4.1: Data bewerking

Creëren jaarcategorieën: gen yearcat = 1 if year >= 1985 & year <= 1995
replace yearcat = 2 if year >= 1996 & year <=2005
replace yearcat = 3 if year >= 2006 & year <= 2014
Creëer log(loon)-variabele: generate logearning = log(earning)
Omrekenen gulden: replace earning=earning/2.20371 if year<2001
Creëren provincie dummy's: tabulate province, gen(dp)
Creëren leeftijdsdummy's:tabulate agecat, gen(dL)
Creëren educatie dummy's:tabulate educ, gen(dE)

Sectie 4.2: Data statistieken

Tabel 1: summ traveltime earning logearning dL* dp* dE* if male==g met $g \in \{0, 1\}$
Tabel 2, 3 en 4: summ traveltime earning logearning if male==g & yearcat ==i met $g \in \{0, 1\}$ en $i \in \{1, 2, 3\}$
Tabel 5: summ traveltime if male==g & yearcat==i, detail met $g \in \{0, 1\}$ en $i \in \{1, 2, 3\}$
Figuur 2: twoway kdensity traveltime if male==1 || kdensity traveltime if male==0|| by(yearcat)
Figuur 3: twoway kdensity traveltime if yearcat==i & male==g || kdensity traveltime if yearcat==i & male= ==g || kdensity traveltime if yearcat==i & male==g met $g \in \{0, 1\}$ en $i \in \{1, 2, 3\}$
Verwijderde data: summ if traveltime==. | earning==. |traveltime> 300| province==.| educ==.

Sectie 5: Statistische modellen

Geen multicollineariteit: Voer de regressie uit en vervolgens: vif
Strikte exogeniteit: Voer de regressie uit en vervolgens: predict e, residual histogram e
White's test: Voer de regressie uit en vervolgens: estat imtest
Heteroskedasticiteit kwantielregressie: sqreg traveltime male logearning dp* dE* dL* if yearcat==3, q(0.25 0.40 0.5 0.75 0.85 0.9)
test[q40]male=[q25]male test[q90]male=[q85]male, accum

Sectie 6: Resultaten

Tabel 6: reg traveltime male logearning dL* dp* dE* if yearcat== i
met $i \in \{1, 2, 3\}$
Tabel 7: qreg traveltime male logearning dL* dp* dE* if yearcat==i, quantile(q)
met $i \in \{1, 2, 3\}$ en $q \in \{0.05, 0.19, 0.25, 0.5, 0.75, 0.90, 0.95\}$
Figuur 3: ssc install grqreg, replace
version 10.1: qreg traveltime male logearning dp* dL* dE* if yearcat==i
met $i \in \{1, 2, 3\}$
grqreg male, ci olsci ols
Tabel 8: qreg traveltime male logearning dL* dp* if yearcat==i, quantile(q)
met $i \in \{1, 2, 3\}$ en $q \in \{0.05, 0.25, 0.5, 0.75, 0.90, 0.95\}$
Significantie jaarcategorieën:generate year1 = 0
generate year2 = 0
replace year1 = 1 if yearcat==1
replace year2 = 1 if yearcat==2
generate year1ma = year1*male
generate year2ma = year2*male
(q)reg traveltime year1 year2 male year1ma year2ma logearning dp* dL* dE*
test year1ma year2ma

Appendix C: Stata output

ssc install outreg2
Voer de regressie uit:
outreg2 using outreg.doc, replace
shellout outreg.doc

C Appendix: Stata output

In tabel 8 staan statistieken van de verwijderde dataset. Bij de gemiddelde reistijd is het commando 'reistijd'>300 niet meegenomen, aangezien deze ontzettend veel invloed had op de gemiddelde reistijd.

Tabel 8: Data statistieken van de gehele (verworpen) dataset

	Verworpen data		Overgebleven data	
	Gem.	Stand. dev. [5]	Gem.	Stand. dev. [5]
Reistijd (min./enk. reis) [6]	12,94	20,73	23,32	20,48
Loon (€/maand)	1081,39	745,58	1076,93	937,17
Percentage man (%)	38,7		55,3	
16-24 jaar	0,103		0,118	
25-34 jaar	0,177		0,227	
35-44 jaar	0,223		0,282	
45-54 jaar	0,212		0,256	
55-64 jaar	0,266		0,115	
65-68 jaar	0,018		0,017	
Groningen	0,042		0,038	
Friesland	0,045		0037	
Drenthe	0,034		0,031	
Overijssel	0,075		0,073	
Flevoland	0,106		0,030	
Gelderland	0,058		0,107	
Utrecht	0,097		0,063	
N-Holland	0,145		0,131	
Z-Holland	0,050		0,161	
Zeeland	0,141		0,038	
N-Brabant	0,079		0,138	
Limburg	0,104		0,067	
Amsterdam	0,023		0,085	
Basisonderwijs	0,142		0,049	
Vmbo/lbo/mavo	0,409		0,276	
Mbo/havo/vwo	0,299		0,368	
Hbo	0,115		0,223	
WO	0,032		0,083	

N = respectievelijk 30.111 en 44.333

[5] De leeftijds-, provincie- en educatievariabelen zijn dummy-variabelen, hierdoor geeft een standaard deviatie bij die variabelen geen relevante informatie.

[6] Reistijd in minuten voor een enkele reis

Aannames OLS en kwantielregressie

Hieronder staan de resultaten van de uitgevoerde testen om te concluderen of aan de aannames was voldaan. De afbeeldingen zijn op chronologische volgorde geplaatst.

De multicollineariteit wordt onderzocht met Stata door middel van de *Variance Inflation Factor* [7]. Het is namelijk niet de bedoeling dat er een lineaire relatie bestaat tussen onafhankelijke variabelen. De VIF-test neemt zowel directe als indirecte multicollineariteit mee. Van indirecte multicollineariteit is sprake als bijvoorbeeld Groningen en Friesland samen gecorreleerd zijn met Drenthe, maar Groningen en Drenthe niet per se. Uit de VIF-test is gebleken dat in jaarcategorie 1 sprake is van multicollineariteit tussen de leeftijdsgroepen en in jaarcategorie 3 tussen de educatiegroepen. Dit heeft een verminderde efficiëntie in jaarcategorie 1 en 3 als gevolg.

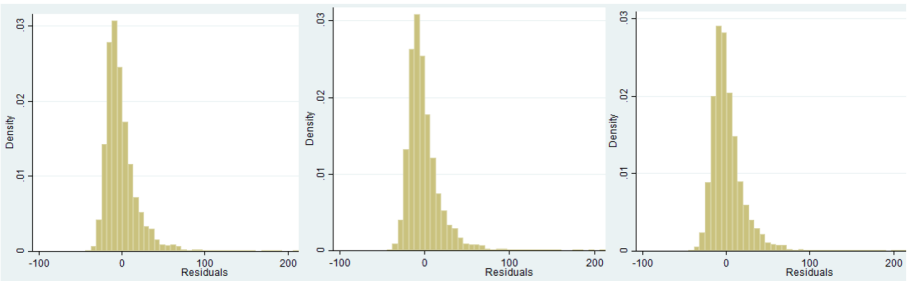
Figuur 5: Testresultaat multicollineariteit

Variable	VIF	1/VIF
dL2	934.19	0.001070
dL3	869.38	0.001150
dL4	652.84	0.001532
dL1	580.86	0.001722
dL5	218.25	0.004582
dp9	8.89	0.112489
dp11	8.73	0.114555
dp6	7.86	0.127248
dp8	7.83	0.127637
dE3	7.58	0.131956
dE4	7.45	0.134309
dp13	7.39	0.135278
dp12	5.76	0.173667
dp4	5.15	0.194345
dp7	4.87	0.205445
dE5	4.79	0.208860
dp2	3.58	0.279055
dp1	3.55	0.281419
dE2	3.49	0.286922
dp10	3.06	0.326979
dp3	2.80	0.357352
logearning	1.62	0.617880
male	1.45	0.688704
Mean VIF	145.71	

Variable	VIF	1/VIF
dp9	4.61	0.216823
dp11	4.47	0.223473
dp8	4.21	0.237541
dp13	3.96	0.252386
dE3	3.81	0.262449
dE4	3.80	0.263225
dp6	3.68	0.271622
dL3	3.59	0.278462
dL4	3.36	0.297472
dL2	3.11	0.321601
dE5	3.05	0.327462
dp4	3.01	0.331927
dp7	2.64	0.378107
dp12	2.57	0.388498
dp10	2.14	0.467941
dp1	2.12	0.472034
dL5	1.98	0.504156
dp2	1.97	0.506396
dp3	1.90	0.526955
dE2	1.54	0.649271
logearning	1.40	0.713674
male	1.20	0.834943
dL6	1.00	0.996736
Mean VIF	2.83	

Variable	VIF	1/VIF
dE4	407.65	0.002453
dE5	335.62	0.002980
dE3	262.19	0.003814
dE6	186.10	0.005374
dE2	44.19	0.022628
dp9	5.54	0.180438
dp8	4.68	0.213787
dp11	4.56	0.219088
dp6	3.73	0.268099
dL4	3.49	0.286823
dL3	3.13	0.319765
dp4	3.09	0.323269
dp12	2.87	0.348704
dp7	2.86	0.349812
dL5	2.84	0.351956
dL2	2.54	0.394414
dp13	2.27	0.439908
dp10	2.12	0.472704
dp5	2.08	0.479872
dp2	2.02	0.493942
dp1	1.99	0.501524
logearning	1.93	0.517686
male	1.21	0.824939
dL6	1.04	0.960173
Mean VIF	53.74	

Figuur 6: Histogram errortermen



[7] Vanaf een VIF-waarde van 10 wordt gesproken van multicollineariteit.

De homoskedasticiteit van de errortermen impliceert dat de onzekerheid van het model overal gelijk is, $Var(\epsilon_1|X) = \sigma_\epsilon^2 I_n$. Dit wordt getest met de White-test, die de hypothese test dat de variantie van de residuen homogeen is. Bij alle drie de tijdscategorieën is sprake van een p-waarde van 0.000, dus er is sprake van homoskedasticiteit van de errorterm.

Figuur 7: Testresultaat homoskedasticiteit errorterm

Source	chi2	df	p
Heteroskedasticity	340.65	179	0.0000
Skewness	105.81	23	0.0000
Kurtosis	12.35	1	0.0004
Total	458.81	203	0.0000

Source	chi2	df	p
Heteroskedasticity	536.47	178	0.0000
Skewness	167.64	23	0.0000
Kurtosis	22.55	1	0.0000
Total	726.66	202	0.0000

Source	chi2	df	p
Heteroskedasticity	711.53	205	0.0000
Skewness	285.94	24	0.0000
Kurtosis	47.00	1	0.0000
Total	1044.47	230	0.0000

Figuur 8: Test homoskedasticiteit van kwantielregressie

```
test[q90]male=[q85]male, accum

( 1)  - [q25]male + [q40]male = 0
( 2)  - [q40]male + [q50]male = 0
( 3)  - [q50]male + [q75]male = 0
( 4)  - [q75]male + [q85]male = 0
( 5)  - [q85]male + [q90]male = 0

       F(  5, 12845) =    1.51
            Prob > F =    0.1830
```

```
test[q90]male=[q85]male, accum

( 1)  - [q25]male + [q40]male = 0
( 2)  - [q40]male + [q50]male = 0
( 3)  - [q50]male + [q75]male = 0
( 4)  - [q75]male + [q85]male = 0
( 5)  - [q85]male + [q90]male = 0

       F(  5, 14392) =   19.80
            Prob > F =    0.0000
```

```
test[q90]male=[q85]male, accum

( 1)  - [q25]male + [q40]male = 0
( 2)  - [q40]male + [q50]male = 0
( 3)  - [q50]male + [q75]male = 0
( 4)  - [q75]male + [q85]male = 0
( 5)  - [q85]male + [q90]male = 0

       F(  5, 17023) =   32.93
            Prob > F =    0.0000
```

Significantie geslachtscoëfficiënt t.o.v. elkaar

Figuur 9: Respectievelijk OLS, 50^e percentiel en 90^e percentiel

```
test year1ma year2ma

( 1)  year1ma = 0
( 2)  year2ma = 0

       F(  2, 44304) =   37.26
            Prob > F =    0.0000
```

```
test year1ma year2ma

( 1)  year1ma = 0
( 2)  year2ma = 0

       F(  2, 44304) =   32.93
            Prob > F =    0.0000
```

```
test year1ma year2ma

( 1)  year1ma = 0
( 2)  year2ma = 0

       F(  2, 44304) =   13.81
            Prob > F =    0.0000
```

Resultaten OLS regressies

Met 'male' wordt de geslachtscoëfficiënt aangeduid en met 'logearning' de logaritme van het netto maandloon. Voor de provincie-, opleidings- en leeftijdsvariabele zijn dummy's aangemaakt, respectievelijk dp_i, dE_i en dL_i, waarbij de volgorde van nummering precies gelijk loopt met de volgorde uit tabel 1. De sterrenaanduiding werkt op dezelfde manier als in deze scriptie gebruikt is.

Figuur 10: OLS regressies per jaarcategorie

VARIABLES	traveltime	VARIABLES	Traveltime	VARIABLES	traveltime
male	0.591	male	3.795***	male	3.164***
	(0.391)		(0.375)		(0.336)
logearning	4.440***	logearning	4.043***	logearning	6.475***
	(0.314)		(0.241)		(0.266)
dp1	-8.484***	dp1	-3.875***	dp1	-1.305
	(1.500)		(1.265)		(1.184)
dp2	-8.525***	dp2	-4.154***	dp2	0.800
	(1.496)		(1.309)		(1.175)
dp3	-9.004***	dp3	-3.614***	o.dp3	-
	(1.588)		(1.337)		
dp4	-9.265***	dp4	-3.960***	dp4	-0.420
	(1.415)		(1.124)		(1.015)
o.dp5	-	o.dp5	-	dp5	0.989
					(1.160)
dp6	-5.869***	dp6	-3.061***	dp6	0.351
	(1.359)		(1.076)		(0.976)
dp7	-3.768***	dp7	-0.393	dp7	1.078
	(1.426)		(1.170)		(1.040)
dp8	-3.458**	dp8	-0.0904	dp8	2.273**
	(1.359)		(1.050)		(0.941)
dp9	-4.760***	dp9	-2.359**	dp9	0.869
	(1.347)		(1.035)		(0.921)
dp10	-8.413***	dp10	-6.050***	dp10	2.115*
	(1.549)		(1.259)		(1.151)
dp11	-5.767***	dp11	-5.011***	dp11	-0.844
	(1.349)		(1.041)		(0.944)
dp12	-8.051***	dp12	-5.339***	dp12	-1.229
	(1.397)		(1.176)		(1.035)
dp13	-3.110**	dp13	-1.802*	dp13	1.829
	(1.366)		(1.068)		(1.126)
dL1	9.812	o.dL1	-	o.dL1	-
	(10.34)				
dL2	8.123	dL2	-1.466**	dL2	-1.201*
	(10.34)		(0.708)		(0.677)
dL3	7.653	dL3	-2.195***	dL3	-3.337***
	(10.34)		(0.685)		(0.632)
dL4	6.732	dL4	-1.766**	dL4	-3.739***
	(10.34)		(0.705)		(0.621)
dL5	5.508	dL5	-3.932***	dL5	-3.218***
	(10.36)		(0.859)		(0.660)
o.dL6	-	dL6	-6.585	dL6	-9.911***
			(14.48)		(2.368)
dE2	-6.027***	dE2	-10.51***	dE2	-11.61*
	(1.016)		(1.109)		(6.374)
dE3	-6.449***	dE3	-8.979***	dE3	-11.96*
	(0.905)		(0.720)		(6.310)
dE4	-4.543***	dE4	-6.745***	dE4	
	(0.900)		(0.695)		(6.306)
dE5	-1.676*	dE5	-3.936***	dE5	-7.784
	(0.935)		(0.714)		(6.310)
o.dE6	-	o.dE6	-	dE6	-4.379
					(6.320)
Constant	-3.478	Constant	6.217***	Constant	-11.71*
	(10.57)		(2.026)		(6.538)
Observations	12,869	Observations	14,416	Observations	17,048
R-squared	0.059	R-squared	0.082	R-squared	0.105

Resultaten kwantielregressie

Figuur 11: Resultaten kwantielregressie $q = 0,1$

VARIABLES	(jaarcat. 1) traveltime	VARIABLES	(jaarcat. 2) traveltime	VARIABLES	(jaarcat. 3) traveltime
male	0	male	0.0320	male	-0.220
	(0.0799)		(0.145)		(0.248)
logearning	-0	logearning	0.657***	logearning	1.892***
	(0.0642)		(0.0935)		(0.197)
dp1	-0	dp1	-0.434	dp1	-1.955**
	(0.187)		(0.387)		(0.817)
dp2	-1***	dp2	-1.674***	dp2	-1.342*
	(0.186)		(0.409)		(0.812)
dp3	-0	dp3	-0.189	dp3	-1.686**
	(0.214)		(0.422)		(0.833)
dp4	-0	dp4	-0.814***	dp4	-2.453***
	(0.157)		(0.314)		(0.683)
dp5	-0	dp5	-0.581	dp5	-2.439***
	(0.279)		(0.415)		(0.800)
dp6	-0	dp6	-0.469	dp6	-2.042***
	(0.134)		(0.286)		(0.649)
dp7	-0	dp7	-0.169	dp7	-1.339*
	(0.160)		(0.335)		(0.697)
dp8	-0	dp8	-0.0673	dp8	-0.799
	(0.134)		(0.272)		(0.620)
dp9	-0	dp9	-0.283	dp9	-0.780
	(0.129)		(0.264)		(0.603)
dp10	-0	dp10	-0.655*	dp10	-1.515*
	(0.203)		(0.384)		(0.793)
dp11	-0	dp11	-0.407	dp11	-1.743***
	(0.130)		(0.267)		(0.624)
dp12	-0	dp12	-0.134	dp12	-1.566**
	(0.150)		(0.341)		(0.699)
o.dp13	-	o.dp13	-	o.dp13	-
dL1	5**	dL1	-5.581	dL1	6.198***
	(2.112)		(5.619)		(1.750)
dL2	5**	dL2	-5.661	dL2	5.873***
	(2.111)		(5.616)		(1.740)
dL3	5**	dL3	-5.880	dL3	5.331***
	(2.111)		(5.615)		(1.731)
dL4	5**	dL4	-5.609	dL4	4.681***
	(2.111)		(5.616)		(1.728)
dL5	3	dL5	-6.284	dL5	4.890***
	(2.114)		(5.619)		(1.735)
o.dL6	-	o.dL6	-	o.dL6	-
dE2	0	dE2	-1.474***	dE2	-3.726
	(0.207)		(0.430)		(4.712)
dE3	0	dE3	-1.188***	dE3	-4.098
	(0.185)		(0.280)		(4.665)
dE4	0	dE4	-1.038***	dE4	-3.746
	(0.184)		(0.270)		(4.662)
dE5	-0	dE5	-0.885***	dE5	-3.884
	(0.191)		(0.277)		(4.664)
o.dE6	-	o.dE6	-	dE6	-3.996
					(4.672)
Constant	-0	Constant	7.729	Constant	-8.045
	(2.152)		(5.653)		(5.138)
Observations	12,869	Observations	14,416	Observations	17,048

Figuur 12: Resultaten kwantielregressie $q = 0,5$

VARIABLES	(jaarcat. 1) traveltime	VARIABLES	(jaarcat. 2) traveltime	VARIABLES	(jaarcat. 3) traveltime
male	0.329	male	2.415***	male	2.783***
	(0.359)		(0.353)		(0.357)
logearning	3.349***	logearning	3.246***	logearning	5.011***
	(0.288)		(0.227)		(0.283)
dp1	-5.200***	dp1	-2.517***	dp1	-2.185*
	(0.839)		(0.938)		(1.174)
dp2	-6.268***	dp2	-2.462**	dp2	-0.482
	(0.836)		(0.992)		(1.167)
dp3	-4.703***	dp3	-2.347**	dp3	-2.895**
	(0.964)		(1.023)		(1.197)
dp4	-5.082***	dp4	-2.855***	dp4	-2.938***
	(0.706)		(0.762)		(0.981)
dp5	3.856***	dp5	-0.856	dp5	-1.790
	(1.254)		(1.005)		(1.150)
dp6	-4.112***	dp6	-1.661**	dp6	-1.342
	(0.604)		(0.695)		(0.933)
dp7	-2.462***	dp7	-0.390	dp7	-0.641
	(0.721)		(0.813)		(1.002)
dp8	-1.755***	dp8	0.807	dp8	0.574
	(0.604)		(0.660)		(0.891)
dp9	-2.215***	dp9	-0.673	dp9	0.00432
	(0.582)		(0.640)		(0.867)
dp10	-4.565***	dp10	-4.487***	dp10	-1.554
	(0.912)		(0.931)		(1.139)
dp11	-3.063***	dp11	-2.915***	dp11	-2.322***
	(0.585)		(0.646)		(0.896)
dp12	-3.857***	dp12	-2.975***	dp12	-3.151***
	(0.675)		(0.827)		(1.004)
o.dp13	-	o.dp13	-	o.dp13	-
dL1	9.266	dL1	3.046	dL1	5.380**
	(9.493)		(13.63)		(2.515)
dL2	8.306	dL2	2.524	dL2	6.143**
	(9.488)		(13.62)		(2.500)
dL3	7.628	dL3	2.182	dL3	4.033
	(9.488)		(13.62)		(2.487)
dL4	7.449	dL4	2.763	dL4	3.934
	(9.490)		(13.62)		(2.483)
dL5	6.561	dL5	0.606	dL5	4.247*
	(9.505)		(13.63)		(2.493)
o.dL6	-	o.dL6	-	o.dL6	-
dE2	-5.787***	dE2	-9.044***	dE2	-8.984
	(0.933)		(1.044)		(6.770)
dE3	-5.531***	dE3	-7.881***	dE3	-10.05
	(0.831)		(0.678)		(6.703)
dE4	-3.973***	dE4	-5.644***	dE4	-8.197
	(0.826)		(0.654)		(6.698)
dE5	-2.324***	dE5	-3.984***	dE5	-5.811
	(0.858)		(0.672)		(6.702)
o.dE6	-	o.dE6	-	dE6	-2.671
					(6.713)
Constant	-4.767	Constant	0.860	Constant	-12.71*
	(9.676)		(13.71)		(7.383)
Observations	12,869	Observations	14,416	Observations	17,048

Figuur 13: Resultaten kwantielregressie $q = 0,9$

VARIABLES	(jaarcat. 1) traveltime	VARIABLES	(jaarcat. 2) traveltime	VARIABLES	(jaarcat. 3) traveltime
male	1.958*	male	8.673***	male	7.037***
	(1.102)		(1.004)		(0.862)
logearning	7.151***	logearning	6.753***	logearning	10.71***
	(0.885)		(0.645)		(0.683)
dp1	-12.38***	dp1	-7.082***	dp1	-8.498***
	(2.573)		(2.671)		(2.837)
dp2	-7.124***	dp2	-4.195	dp2	-1.344
	(2.564)		(2.823)		(2.819)
dp3	-13.04***	dp3	0.606	dp3	-0.323
	(2.956)		(2.912)		(2.890)
dp4	-13.99***	dp4	-3.600*	dp4	-1.500
	(2.166)		(2.169)		(2.369)
dp5	8.991**	dp5	7.364**	dp5	1.737
	(3.848)		(2.861)		(2.778)
dp6	-3.105*	dp6	-2.985	dp6	-4.156*
	(1.853)		(1.977)		(2.254)
dp7	-0.0743	dp7	2.966	dp7	-0.349
	(2.210)		(2.313)		(2.421)
dp8	-0.103	dp8	5.478***	dp8	0.573
	(1.854)		(1.879)		(2.153)
dp9	-1.951	dp9	-2.528	dp9	-3.350
	(1.784)		(1.822)		(2.094)
dp10	-12.04***	dp10	-8.523***	dp10	3.742
	(2.798)		(2.650)		(2.752)
dp11	-5.920***	dp11	-7.645***	dp11	-5.475**
	(1.794)		(1.840)		(2.164)
dp12	-13.42***	dp12	-8.250***	dp12	-6.431***
	(2.069)		(2.354)		(2.425)
o.dp13	-	o.dp13	-	o.dp13	-
dL1	12.46	dL1	24.33	dL1	8.623
	(29.12)		(38.78)		(6.076)
dL2	9.022	dL2	19.71	dL2	3.425
	(29.11)		(38.76)		(6.040)
dL3	8.281	dL3	17.69	dL3	-1.337
	(29.11)		(38.76)		(6.008)
dL4	8.177	dL4	17.89	dL4	-1.458
	(29.11)		(38.76)		(5.998)
dL5	7.245	dL5	16.49	dL5	0.936
	(29.16)		(38.78)		(6.021)
o.dL6	-	o.dL6	-	o.dL6	-
dE2	-10.72***	dE2	-25.34***	dE2	-44.73***
	(2.861)		(2.970)		(16.36)
dE3	-11.73***	dE3	-20.40***	dE3	-42.09***
	(2.549)		(1.929)		(16.19)
dE4	-9.763***	dE4	-17.76***	dE4	-39.93**
	(2.533)		(1.861)		(16.18)
dE5	-2.426	dE5	-10.69***	dE5	-34.28**
	(2.632)		(1.912)		(16.19)
o.dE6	-	o.dE6	-	dE6	-25.74
					(16.22)
Constant	2.847	Constant	-1.809	Constant	7.501
	(29.68)		(39.02)		(17.84)
Observations	12,869	Observations	14,416	Observations	17,048