Universiteit Utrecht

Mathematisch Instituut

Bachelorscriptie

# Kwantielregressie: optimalisatiemethode en economische toepassing

Steppe Röttgering - 5548594

Eerste begeleider:
prof. dr. Wolter Hassink
Departement Economie

Tweede begeleider:
dr. Cristian Spitoni
Departement Wiskunde

12 juni 2018

## Samenvatting

Dit onderzoek richt zich allereerst op de optimalisatiemethode die ten grondslag ligt aan de mediaanregressie schatter, dit is een specifiek geval van de kwantielregressie schatter. Deze schatter is een minimaliseringsprobleem waarvoor we een lineair programmeringsprobleem (LP) kunnen opstellen. Met behulp van de simplexmethode worden twee bivariate voorbeelden analytisch opgelost. Met behulp van het softwarepakket Stata worden de oplossingen gecontroleerd. Het blijkt dat de simplexmethode dezelfde oplossing genereerd als Stata.

Daarnaast richt dit onderzoek zich op een economische toepassing van de kwantielregressie. De kwantielregressie wordt gebruikt om op verschillende plekken in de verdeling het effect te analyseren van de onafhankelijke variabelen op de afhankelijke variabele. De dataset WoOn 2015 wordt geanalyseerd voor de WOZ-waarde. De controlevariabelen zijn het perceeloppervlak, het aantal kamers en of de woning een tuin bezit. We analyseren het OLS model en het kwantielregressie model voor bovenstaande variabelen. Het blijkt dat de kwantielregressie een meerwaarde is ten opzichte van OLS, omdat de kwantielregressie per kwantiel informatie verschaft en deze informatie niet uit OLS verkregen kan worden.

# Inhoudsopgave

# 1 Inleiding

In de statistiek wordt met het begrip regressie een bepaalde samenhang tussen verschillende variabelen bedoeld. In een regressievergelijking kan het effect van de onafhankelijke variabelen op de afhankelijke variabele worden verklaard. Aan de hand van een databestand kan er een schatting worden gemaakt voor dit effect. Er zijn verschillende methoden om het effect van de onafhankelijke variabele op de afhankelijke variabele te schatten. Merk op dat er bij verschillende methodes verschillende schatters horen. De meest gebruikte methode om een verdeling te schatten is de OLS methode. Bij de OLS methode wordt het effect van de onafhankelijke variabelen op het gemiddelde van de verdeling van de afhankelijke variabele geschat. Om consistentie van de OLS-schatter te waarborgen moet $E(\hat{\beta}|X) = \beta$ gelden. De OLS-schatter is [4]

$$\hat{\beta} = (X'X)^{-1}X'y$$

In deze scriptie zal de nadruk liggen op twee andere schatters, de mediaanregressie schatter en de kwantielregressie schatter. Met de kwantielregressie kan men het effect van de onafhankelijke variabelen op een plek in de verdeling voor de afhankelijke variabele schatten. Merk op dat bij de mediaanregressie deze plek de mediaan is. Aangezien de mediaanregressie een specifiek geval is van de kwantielregressie, wordt alleen de theorie voor de kwantielregressie behandeld. Het doel van deze scriptie is om de optimalisatiemethode toe te lichten die ten grondslag ligt aan de mediaanregressie schatter.

Vanuit de literatuur is bekend dat een optimalisatiemethode ten grondslag ligt aan de kwantielregressie schatter. Deze optimalisatietechniek is de simplexmethode. We stellen een kwantielregressie vergelijking op populatieniveau op en herleiden hieruit de kwantielregressie schatter. In de literatuur wordt het lineairprogrammeringsprobleem (LP) voor de kwantielregressie schatter gegeven [3], maar het LP wordt niet duidelijk verklaard vanuit de simplexmethode. In deze scriptie zullen wij de optimalisatiemethode voor dit LP verder toelichten. Om de optimalisatiemethode centraal te stellen, behandelen wij het LP voor de mediaan regressie. We analyseren twee bivariate voorbeelden op steekproef niveau met betrekking tot de mediaan regressie. De parameters $\beta_0$ en $\beta_1$ zullen geschat worden. Het LP wordt grafisch weergegeven om meer begrip te geven aan de situatie. Tot slot wordt het softwarepakket Stata gebruikt ter controle.

De mediaanregressie schatter is verklaard en de kwantielregressie schatter is gegeven, maar wat is de meerwaarde van deze schatters ten opzichte van de OLS schatter. Om de meerwaarde aan te tonen wordt in Hoofdstuk 4 de dataset WoOn 2015 [2] geanalyseerd. Als afhankelijke variabele wordt de

WOZ-waarde gekozen, omdat het effect van de onafhankelijke variabelen op de WOZ-waarde bij de verschillende plekken in de verdeling mogelijk anders is dan het effect bij het gemiddelde. De verschillende effecten van de onafhankelijke variabelen op de WOZ-waarde worden gegeven.

De structuur van de scriptie is als volgt. In Hoofdstuk 2 (Theoretisch kader) wordt het begrip regressievergelijking verder toegelicht. Vervolgens introduceren wij het kwantielregressie model en de kwantielregressie schatter op populatieniveau wordt gegeven. Daarna wordt een optimalisatietechniek (simplexmethode) om een LP op te lossen uitgelegd. Het theoretisch kader wordt besloten door een gegeneraliseerd LP die equivalent is met de kwantielregressie schatter. In Hoofdstuk 3 worden twee bivariate voorbeelden voor de mediaanregressie schatter op steekproefniveau uitgewerkt met de simplexmethode. Voor een beter begrip worden de voorbeelden grafisch weergegeven. Ter controle worden de voorbeelden met Stata uitgewerkt. In Hoofdstuk 4 wordt de meerwaarde van de kwantielregressie laten zien voor de dataset WoOn 2015. De WOZ-waarde wordt geanalyseerd met behulp van de verschillende schatters. Tot slot worden in Hoofdstuk 5 de conclusies getrokken. Het effect van de simplexmethode op de mediaanregressie schatter wordt benoemd en we blikken terug op de meerwaarde van de kwantielregressie ten opzichte van de WOZ-waarde.

# 2 Theoretisch Kader

Dit hoofdstuk begint met het toelichten van het begrip regressievergelijking. Het model op populatieniveau wordt gegeven, de parameters worden verklaard en verschillende methoden om een regressievergelijking te analyseren worden benoemd. In sectie 2.2 zal het begrip kwantielregressie worden uitgewerkt. Het kwantielregressie model en de bijbehorende schatter worden gegeven en verklaard. In sectie 2.3 wordt de simplexmethode beschreven. Het lineair programmeringsprobleem en de oplosmethode worden behandeld. In sectie 2.4 wordt het LP voor de kwantielregressie schatter gegeven. Hiermee wordt het hoofdstuk afgerond en is er genoeg theorie om de voorbeelden in hoofdstuk 3 uit te werken en de toepassing in hoofdstuk 4 te verklaren.

## 2.1 Regressievergelijking

Zoals gezegd wordt met het begrip regressie een bepaalde samenhang tussen verschillende variabelen bedoeld. Een regressievergelijking bestaat uit twee typen variabelen, de afhankelijke en de onafhankelijke. Om het effect van de onafhankelijke variabelen op de afhankelijke variabelen te schatten worden controlevariabelen ingevoerd. Bij een regressie-analyse zoeken wij verschillende verbanden tussen de variabelen. We krijgen het volgende regressiemodel op populatieniveau [4]:

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

Hier is $Y$ de afhankelijke variabele, $X$ de matrix die de onafhankelijke variabele(n) weergeeft, $\beta$ de parametervector en de functie $g$ is afhankelijk van het gebruikte model. De parametervector $\beta$ is de coëfficiëntenvector voor de onafhankelijke variabele(n) $X$, die het effect op de afhankelijke variabele $Y$ weergeeft. Het doel bij een regressie-analyse is om de geschatte waardes voor de parametervector $\beta$ te bepalen, de schatter die hierbij hoort wordt genoteerd als $\hat{\beta}$. In het algemeen zal een regressiemodel niet de werkelijke waarde van $Y$ weergeven:

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

Hierbij is $u$ de foutterm, een variabele die niet opgenomen is in $X$, maar wel de afhankelijke variabele $Y$ beïnvloedt. Merk op dat er verschillende vormen bestaan van regressievergelijkingen; er bestaan lineaire en niet-lineaire regressievergelijkingen. Als een model lineair is, betekent dit dat de verhouding tussen de geschatte parameters en de afhankelijke variabele lineair is. Hierdoor kan men in een lineair regressiemodel ook kwadraten van de onafhankelijke variabelen nemen of logaritmes toepassen. Een lineair regressiemodel op populatieniveau ziet er als volgt uit:

$$Y = X\beta + u$$

Hier is $Y$ een $n$-dimensionale kolomvector, $X$ de $(n \times (k+1))$ matrix van $k$ onafhankelijke variabelen en $n$ observaties, $\beta$ een $(k+1)$-dimensionale kolomvector en $u$ een $n$-dimensionale kolomvector. Omdat de eerste kolom van de matrix $X$ uit enen bestaat, verkrijgen we de constante $\beta_0$ na vermenigvuldiging tussen de matrix $X$ en de parametervector $\beta$.

Het is bekend dat er verschillende methoden zijn om een regressievergelijking te analyseren. In Hoofdstuk 4 zullen wij met behulp van de OLS-methode en de kwantielregressie een model analyseren. Bij de OLS-methode gaan we uit van het gemiddelde van de verdeling voor de afhankelijke variabele. In deze methode minimaliseren we de gekwadrateerde afwijkingen tussen de verwachte en de geschatte waarden met betrekking tot $\beta$. Om de consistentie van de schatter waar te borgen moet aan verschillende voorwaardes voldaan zijn, deze voorwaardes worden toegelicht in sectie 4.2.1. De kwanielregressie wordt nauwkeuriger beschreven in de volgende sectie.

## 2.2 Kwantielregressie

Bij een regressie analyse wordt vaak het gemiddelde van de verdeling voor de afhankelijke variabele bekenen, bijvoorbeeld met de OLS methode (zie vorige sectie). Merk op dat het ook interessant kan zijn om verschillende plekken (kwantielen) in een verdeling te analyseren. De kwantielregressie is een manier om verschillende delen in de populatieverdeling te analyseren. Het doel van deze sectie is om de kwantielregressie methode verder toe te lichten. We beginnen met het definiëren van het kwantiel. Het kwantielregressie model wordt gegeven en tot slot wordt een algemene formule gegeven voor de kwantielregressie schatter.

De kwantielregressie methode is ontwikkeld door het samenvoegen van verschillende methoden. Het begon bij Boskovic (1760), die een theorie over de mediaan-regressie schreef. Koenker en Bassett [7] gingen vervolgens door op de mediaan-regressie van Boskovic. Met de theorie van de simplexmethode (sectie 2.3), een optimalisatietheorie over minimaliseringsproblemen, konden Koenker en Bassett hun theorie completeren.

Zoals gezegd is de kwantielregressie een methode om op verschillende plekken in de verdeling (kwantielen) de afhankelijke variabele $Y$ te analyseren. We beginnen met een definitie voor het q-de kwantiel $\xi_q$. Een kwantiel is een getal in de statistiek die de verdeling voor $Y$ in verschillende delen verdeeld. Het getal $q \in (0,1)$ geeft aan welk deel van de verdeling wordt bekeken. Kwantielen die bij een percentage horen heten percentielen. Het $50^e$ percentieel heet de mediaan en wordt genoteerd met $\xi_{0.5}$. Voor de mediaanregressie gaan we bij de kwantielregressie uit van de mediaan.

Vanuit de statistiek is bekend dat bij een continue random variabele $Y$ een kansdichtsheidsfunctie $f_Y(y)$ en een verdelingsfunctie $F_Y(y)$ horen. De bijbehorende waarde voor $F_Y(y)$ is de cumulatieve kans op de waarden van $Y$ kleiner dan of gelijk aan $y$, ofwel:

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)\, dt = P(Y \leq y)$$

We gebruiken nu de definitie gegeven in [4]. Voor $q \in (0, 1)$ wordt $\xi_q$ als volgt gedefinieerd:

$$\xi_q(Y) = \inf \{y \in \mathbb{R} : F_Y(y) \geq q\}$$

De verdelingsfunctie $F_Y(y)$ is een monotoon stijgende functie die rechtscontinu is. Voor een punt $y$ waarvoor $F_Y(y)$ niet links-continu is, is er geen minimum. Wij kunnen echter wel het infimum, de grootste ondergrens, bepalen. Vandaar dat wij het infimum nemen voor dit punt.

Om uiteindelijk de schatter $\hat{\beta}(q)$ te bepalen, moeten wij $\xi_q$ een specifiekere waarde meegeven. Voor het bewijs is het handig om eerst een speciaal geval te bekijken. Hiermee kunnen wij dan $\forall q \in (0, 1)$ de algemene vorm voor $\xi_q(Y)$ weergeven. Bekijk het speciale geval $\xi_{0.5}$ en definieer dit als volgt (het bewijs staat in sectie 6.1):

$$\xi_{0.5} = \arg\min_{\beta \in \mathbb{R}} E\,|Y - X\beta| \tag{1}$$

Nu kunnen wij $\xi_q$ veralgemeniseren. We volgen hierbij de methode van pagina 823 [4]. Om onderscheid te maken in de verschillende delen in de verdeling, introduceren we de indicatorfunctie $I(.)$, die als volgt wordt gedefinieerd:

$$I(Y \geq X\beta) = \begin{cases} 1 & \text{Als } Y \geq X\beta \\ 0 & \text{Als } Y < X\beta \end{cases} \qquad I(Y < X\beta) = \begin{cases} 1 & \text{Als } Y < X\beta \\ 0 & \text{Als } Y \geq X\beta \end{cases}$$

Verwerk nu de indicatorfunctie $I(.)$ in (1), dan ontstaat er $\forall q \in (0, 1)$ een algemene uitdrukking (zie sectie 6.2 voor het bewijs):

$$\xi_q = \arg\min_{\beta \in \mathbb{R}} q \cdot E\,(Y - X\beta) \cdot I(Y \geq X\beta) + (1-q) \cdot E\,(X\beta - Y) \cdot I(Y < X\beta) \tag{2}$$

Dit is een minimaliseringsprobleem met betrekking tot $\beta$ voor de verschillende delen in de verdeling. Definieer de functie

$$p_q(Y) = q \cdot I(Y \geq X\beta) + (1 - q) \cdot I(Y < X\beta) \tag{3}$$

die de verschillende delen van de verdeling weergeeft. Hierdoor kunnen wij (2) schrijven als [4]:

$$\xi_q = \arg\min_{\beta \in \mathbb{R}} E\,[p_q(Y - X\beta)] \tag{4}$$

Met de theorie van de regressie-analyse volgt dat een lineaire kwantielregressie-vergelijking op populatieniveau er als volgt uit ziet:

$$\xi_{\mathrm{q}}(Y|X) = X\beta(q) + u$$

Het doel bij een regressie analyse is altijd om de schatter $\hat{\beta}(q)$ te bepalen, hierdoor vergaar je informatie over de coëfficiënten van de onafhankelijke variabelen. Nu volgt de gewenste formule voor de kwantielregressie schatter op steekproef niveau [5]:

$$\hat{\beta}(q) = \arg\min_{\beta} \sum_{i=1}^{n} (\rho_q(y_i - x_i'\beta)) \tag{5}$$

In hoofdstuk 4 zullen wij uitspraken doen over de geschatte waardes van de parameters. Voor de betrouwbaarheid van deze uitspraken is het van belang dat de kwaliteit van de kwantielregressie schatter $\hat{\beta}(q)$ gewaarborgd wordt. Hierom zijn er twee aannames opgesteld voor de kwantielregressie (zie pagina 139 [3]):

1. Er zijn meer observaties dan variabelen in het model ($n > k + 1$)
2. Geen heteroskedasticiteit ($Var(u_i) = \sigma^2$, met $\sigma$ een constante)

De eerste aanname kan gemakkelijk getest worden en spreekt voor zich. De tweede aanname verdient enige uitleg. Heteroskedasticiteit van de foutterm kan namelijk nadelige gevolgen voor de schatting van de coëfficiënten hebben. Zowel $\beta_0(q)$ als de andere coëfficiënten $\beta(q)$ voor de variabelen veranderen. We weten dat er voor ieder kwantiel een lijn gecreëerd kan worden. Bij heteroskedasticiteit van de foutterm kunnen de lijnen van de verschillende kwantielen elkaar kruisen. Hierdoor verandert de natuurlijke ordening van de kwantielen.

## 2.3 Simplexmethode

In de vorige sectie hebben wij het begrip simplexmethode geïntroduceerd, maar geen uitgebreide toelichting meegegeven. Het is bekend dat Koenker en Bassett [7] de mediaan-regressie van Boskovic (1760) aanvulde met de simplexmethode, waardoor de kwantielregressie techniek werd ontwikkeld. In deze sectie wordt het doel van de methode uitgelegd, het LP wordt gegeven en de oplosmethode voor een LP wordt nauwkeurig uitgelegd. In deze sectie gebruiken wij de notatie van de colleges van dhr. J.A. Hoogeveen [6]. Merk op dat de variabele $x$ uit deze sectie niet dezelfde $x$ is als in de vorige sectie.

De simplexmethode is een methode waarbij een lineair programmeringsprobleem wordt geoptimaliseerd. Hierbij gaan wij uit van een doelstellingsfunctie $cx$ en nevenvoorwaarden $Ax \leq b$ met $x \geq 0$. De simplex methode is een

itererend proces, waarbij wij beginnen in een oplossing die aan de nevenvoorwaarden voldoet, dit noemen wij een toegelaten basisoplossing (TBO). Wij vinden aan de hand van iteraties een steeds betere oplossing. Wanneer een iteratie op een gegeven moment niks beters oplevert, dan stopt het proces en hebben wij een optimale oplossing gevonden. Wij kunnen ook vaststellen dat het probleem niet optimaal oplosbaar is.

Nu wordt de simplexmethode stap voor stap uitgelegd. Begin altijd met het minimaliseren/maximaliseren van de doelstellingsfunctie $z = cx$ met coëfficiëntenvector $c = (c_1, \ldots, c_n)'$ en beslissingsvariabelen $x = (x_1, \ldots, x_n)'$. Dit heet het primale probleem. Stel dat wij de doelstellingsfunctie $z$ willen minimaliseren (het maximaliseren van de doelstellingsfunctie gaat op dezelfde manier, hier kom ik later op terug). Uiteraard horen er voorwaarden bij een doelstellingsfunctie. Deze voorwaarden noteren wij met $Ax \leq b$ en $x \geq 0$, waarin $A$ een $(m \times n)$ matrix is die coëfficiënten voor de variabelen $x$ weergeeft. Hierbij is $m$ het aantal vergelijkingen en $n$ het aantal variabelen. Wij kunnen met deze gegevens het toegelaten gebied voor de oplossingen als volgt definiëren:

$$X = \{x \in \mathbb{R}^n | Ax \leq b, b \geq 0\}$$

Hier is $a^i$ de $i'$de rij en $a_j$ de $j'$de kolom van de matrix $A$. Wij zijn altijd geïnteresseerd in de optimale TBO. Een punt $y$ is een TBO als aan de volgende voorwaarden voldaan is:

1. Als $y$ voldoet aan de beperkingen van het gebied $X$
2. Minstens $(n - m)$ van de coördinaten van $y$ de waarde nul hebben
3. De bijbehorende basismatrix $B = (a_{B_1}, \ldots, a_{B_m})$ is inverteerbaar

Merk op dat wij hier een nieuwe basismatrix $B$ introduceren. De matrix $A$ bestaat uit een gedeelte $B$ en $N$. De basismatrix $B$ bevat de kolommen uit $A$ voor de basisvariabelen $x_B = (x_{B_1}, \ldots, x_{B_m})'$. De matrix $N$ bevat de overige kolommen uit $A$. Op dezelfde wijze wordt de coëfficiëntenvector $c$ opgesplitst in $c_B = (c_{B_1}, \ldots, c_{B_m})'$ en $c_N$. Wij definiëren $R$ als de indexverzameling met alle indices $j \in \{1, \ldots, n\}$ waarvoor geldt $x_j \in x_N$.

Met deze algemene informatie kunnen wij nu eindelijk interessante objecten definiëren en vanuit daar een procedure gaan uiteenzetten om de optimale TBO te vinden. Wij gebruiken voor het oplossen van een dergelijk LP het simplex tableau. Dit tableau karakteriseert zich door zijn eenvoud en simpele notatie. In het simplex tableau staat veel informatie om de optimale TBO te bepalen. De belangrijkste termen worden nu besproken. We definiëren $y_j = B^{-1}a_j$ als de vector die hoort bij $x_j$, we hebben $B^{-1}b$ als rechterkantvector, de coëfficiënt van $x_j$ in de doelstellingsfunctie is $z_j = c_B y_j$

en de waarde van de doelstellingsfunctie $z$ in de huidige TBO noteren we met $z_0 = c_B B^{-1}$.

De procedure van de simplexmethode werkt nu als volgt. Het volgende primale lineaire programmeringsprobleem wordt opgesteld:

$$\text{(P) min } z = cx \text{ o.d.v. } Ax \leq b,\ x \geq 0$$

Om handig met dit LP te kunnen rekenen, moeten we $Ax = b$ verkrijgen. Dit doen we door spelingsvariabelen $s_i$ toe te voegen aan de vergelijkingen. Een TBO vinden we nu door de spelingsvariabelen als basisvariabelen te nemen. Als alle termen op de nulde rij van het tableau nu kleiner of gelijk aan 0 zijn, ofwel $(z_j - c_j) \leq 0$, dan hebben we een optimale TBO gevonden en zijn we klaar. Dan is er dus geen enkele variabele die we kunnen verhogen, zonder dat het een verslechtering van de huidige waarde van de TBO oplevert. Helaas is dit bijna nooit het geval.

We bevinden ons weer in de oorspronkelijke situatie; we stoppen de spelingsvariabelen in de basis. Maar nu geldt niet $\forall j \in \{1, \ldots, n\}$ dat $(z_j - c_j) \leq 0$. We moeten nu gaan itereren. Bepaal de variabele $x_k$ waarvoor $(z_j - c_j)$ maximaal is, breng deze $x_k$ in de basis. Om te kijken welke basisvariabele $x_{B_r}$ nu uit de basis gaat, gebruiken we de ratioregel:

$$r \leftarrow argmin\{\frac{(B^{-1}b)_i}{y_{ik}} | y_{ik} > 0\}$$

Update nu het tableau door middel van pivoteren op $y_{rk}$. Als nu $\forall j \in \{1, \ldots, n\}$ geldt dat $(z_j - c_j) \leq 0$, dan is de TBO optimaal. Als $(z_j - c_j) \geq 0$, dan herhalen we het gehele bovenstaande proces. Merk op: bij een maximalisatieprobleem hebben we een optimale TBO als $(z_j - c_j) \geq 0$, breng dan de variabale $x_k$ in de basis die minimale $(z_j - c_j)$ heeft.

Er zijn echter een aantal kanttekeningen bij de simplexmethode. We beginnen bij de veronderstelling dat $Ax \leq b$. Als hier de rechterkant $b < 0$ is, moeten we de rij met $-1$ vermenigvuldigen en een kunstmatige variabele $k_i$ toevoegen. Deze kunstmatige variabele kiezen we dan als basisvariabele. Om vervolgens weer een TBO te creëren moeten we de kunstmatige variabelen uit de basis halen (zie sectie 3.1.2). Ook kan het voorkomen dat bij de variabele $x_k$ waarvoor $(z_k - c_k)$ als enige groter of gelijk nul is, dat $y_{ik} < 0$ is. Hierdoor kunnen we niet pivoteren op dit element en hebben we te maken met een onbegrensd minimum. Er bestaan nog een aantal kanttekeningen bij de simplexmethode, maar om de voorbeelden in sectie 3.1.2 en 3.2.2 op te lossen, hoeven we daar geen aandacht aan te besteden.

## 2.4 Lineair programmeringsprobleem voor schatter $\hat{\beta}(q)$

Gegeven dat de schatter $\hat{\beta}(q)$ een minimaliseringsprobleem is, kunnen wij met de theorie in sectie 2.3 dit probleem oplossen met de simplexmethode.

Hiervoor moeten wij eerst het minimaliseringsprobleem schrijven als een LP. We minimaliseren het residu voor alle kwantielen, dit wordt de doelstellingsfunctie. Het residu $\hat{u}_i$ is het verschil tussen de werkelijke waarde $y_i$ en de gefitte waarde van $\hat{y}_i$ [4].

$$\hat{u}_i = y_i - \hat{y}_i$$

Merk op dat er bij een kwantielregressie geen eenduidige schatter is, omdat we de populatie analyseren over verschillende kwantielen. Vandaar dat we $\hat{u}_i$ moeten opsplitsen in twee delen:

$$\hat{u}_i = \begin{cases} v_i & \text{als } \hat{u}_i \cdot I(\hat{u}_i \geq 0) \\ w_i & \text{als } \hat{u}_i \cdot I(\hat{u}_i < 0) \end{cases}$$

De voorwaarden voor het LP worden verkregen uit de regressievergelijking. De schatter $\hat{\beta}(q)$ is equivalent met het volgende primale lineaire programmeringsprobleem [5]:

$$\begin{gathered} \text{(P) min } q \cdot v_i + (1-q) \cdot w_i \\ \text{o.d.v. } y_i - x_i'\beta = v_i - w_i \\ \beta, y_i \in \mathbb{R}^n \text{ , } x_i' \in \mathbb{R}^n \text{ x } \mathbb{R}^m, \, v_i, w_i \geq 0 \end{gathered}$$

In het volgende hoofdstuk zullen wij dit LP specificeren voor de mediaan regressie en oplossen voor een bivariaat model.

# 3 Resultaten

Het doel van deze sectie is om het LP voor de mediaanregressie schatter op te lossen met behulp van de simplexmethode. Het LP voor de mediaanregressie schatter ziet er als volgt uit [8]:

$$\begin{gathered}\text{(P) min } \sum_{i=1}^{n} \hat{u}_i \\ \text{o.d.v. } y_i - x_i'\beta = \sum_{i=1}^{n} \hat{u}_i \\ \beta, y_i \in \mathbb{R}^n \text{ , } x_i' \in \mathbb{R}^n \text{ x } \mathbb{R}^m, \hat{u}_i \geq 0\end{gathered}$$

In deze sectie lossen wij twee bivariate voorbeelden op. In sectie 3.1 wordt het LP voor voorbeeld 1 gegeven. Om men meer begrip mee te geven wordt in sectie 3.1.1 het LP grafisch weergegeven en opgelost. In sectie 3.1.2 wordt het LP analytisch opgelost met de simplexmethode. In sectie 3.1.3 wordt de gevonden oplossing gecontroleerd met het softwarepakket Stata. In sectie 3.2 herhalen wij dit proces voor voorbeeld 2.

## 3.1 Voorbeeld 1

Wij verkrijgen het volgende LP voor het bivariate voorbeeld 1 [3]:

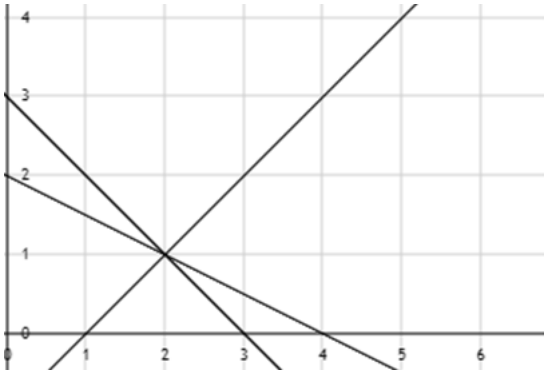
$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 \geq 1 \\ \beta_0 + \beta_1 + u_2 \geq 3 \\ \beta_0 + 2\beta_1 + u_3 \geq 4 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3 \geq 0\end{gathered}$$

De getallen zijn zo gekozen, om het rekenwerk bij de simplexmethode te vereenvoudigen. De nadruk moet namelijk niet liggen op het rekenwerk, maar op de oplosmethode. De volgende voorwaarden zijn van kracht:

$$\begin{gathered}\beta_0 - \beta_1 \leq 1 \\ \beta_0 + \beta_1 \leq 3 \\ \beta_0 + 2\beta_1 \leq 4\end{gathered}$$

### 3.1.1 Grafische methode

Voorbeeld 1 kunnen wij eenvoudig oplossen met de grafische methode. Zet de bovenstaande voorwaarden in een $\beta_0 - \beta_1-$grafiek (zie Figuur 1). Merk op dat de drie lijnen elkaar in één punt snijden. Lees het punt af waarin de lijnen elkaar snijden. Omdat deze drie lijnen elkaar snijden in het punt $(\hat{\beta}_0, \hat{\beta}_1) = (2, 1)$, is dit de oplossing van ons LP.




Figuur 1: Snijdende lijnen voorbeeld 1, $\beta_0$ op x-as, $\beta_1$ op y-as

### 3.1.2 Simplexmethode

Een analytische manier om dit LP op te lossen is met behulp van de simplexmethode (sectie 2.3). We bekijken het LP voor voorbeeld 1 [3]:

$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 \geq 1 \\ \beta_0 + \beta_1 + u_2 \geq 3 \\ \beta_0 + 2\beta_1 + u_3 \geq 4 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3 \geq 0\end{gathered}$$

Merk op dat wij te maken hebben met de nevenvoorwaarden $Ax \geq b$. Om dit in de gebruikelijke $Ax \leq b$ te zetten, voegen wij de spelingsvariabelen $s_i$ (voor $i = 1, 2, 3$) met een negatieve coëfficiënt toe. Dit levert het volgende LP op:

$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 - s_1 = 1 \\ \beta_0 + \beta_1 + u_2 - s_2 = 3 \\ \beta_0 + 2\beta_1 + u_3 - s_3 = 4 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3, s_1, s_2, s_3 \geq 0\end{gathered}$$

Merk op dat wij nu geen basis kunnen creëren met de eenheidsvectoren voor de spelingsvariabelen $s_i$. Voeg daarom de kunstmatige variabelen $k_i$ voor $i = 1, 2, 3$ toe. Het volgende LP ontstaat:

$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 - s_1 + k_1 = 1 \\ \beta_0 + \beta_1 + u_2 - s_2 + k_2 = 3 \\ \beta_0 + 2\beta_1 + u_3 - s_3 + k_3 = 4 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3, s_1, s_2, s_3, k_1, k_2, k_3 \geq 0\end{gathered}$$

Merk op dat wij dit LP kunnen oplossen met de simplexmethode. Zet allereerst de waarden in het simplextableau. Neem de kunstmatige variabelen $k_i$ als basisvariabelen, corrigeer daarvoor met een hulprij $z_1$. De oorspronkelijke doelstellingsfunctie $z$ wordt daaronder weergegeven. Verder worden de waardes uit het bovenstaande LP letterlijk overgenomen:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | 0 |
| z | 0 | 0 | -1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $k_1$ | 1 | -1 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $k_2$ | 1 | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 3 |
| $k_3$ | 1 | 2 | 0 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 4 |

Wij kunnen echter nog geen iteratie met dit tableau uitvoeren, omdat er bij de basisvariabelen $k_i$ geen eenheidsvectoren staan. Creëer een nieuwe doelstellingsfunctie $z_2$. Tel hiervoor alle rijen in het tableau bij elkaar op. Onder onze basisvariabelen $k_i$ staan nu wel de eenheidsvectoren:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 3 | 2 | 0 | 0 | 0 | -1 | -1 | -1 | 0 | 0 | 0 | 8 |
| $k_1$ | 1 | -1 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $k_2$ | 1 | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 3 |
| $k_3$ | 1 | 2 | 0 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 4 |

Merk op dat wij nu met ons iteratieproces kunnen starten, omdat er positieve waarden staan op de nulde rij ($z_2$). Het levert het meest op om $\beta_0$ in de basis te brengen. Aan de hand van de ratiotest (sectie 2.3) gaat $k_1$ uit de basis. We pivoteren nu op de kolom van $\beta_0$:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 0 | 5 | -3 | 0 | 0 | 2 | -1 | -1 | -3 | 0 | 0 | 5 |
| $\beta_0$ | 1 | -1 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $k_2$ | 0 | 2 | -1 | 1 | 0 | 1 | -1 | 0 | -1 | 1 | 0 | 2 |
| $k_3$ | 0 | 3 | -1 | 0 | 1 | 1 | 0 | -1 | -1 | 0 | 1 | 3 |

Het levert nu het meest op als we $\beta_1$ in de basis brengen. Met behulp van de ratioregel moet $k_2$ of $k_3$ uit de basis. Omdat beide dezelfde ratio hebben, mogen we kiezen. Merk op dat het rekenwerk eenvoudiger is als we $k_2$ uit de basis halen:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 0 | 0 | -1/2 | -5/2 | 0 | -1/2 | 3/2 | -1 | -1/2 | -5/2 | 0 | 0 |
| $\beta_0$ | 1 | 0 | 1/2 | 1/2 | 0 | -1/2 | -1/2 | 0 | 1/2 | 1/2 | 0 | 2 |
| $\beta_1$ | 0 | 1 | -1/2 | 1/2 | 0 | 1/2 | -1/2 | 0 | -1/2 | 1/2 | 0 | 1 |
| $k_3$ | 0 | 0 | 1/2 | -3/2 | 1 | -1/2 | 3/2 | -1 | 1/2 | -3/2 | 1 | 0 |

Alleen onder $s_2$ staat nu nog een positief getal op de nulde rij, dus $s_2$ moet in de basis. Met behulp van de ratioregel gaat $k_3$ nu als nog uit de basis:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 0 | 0 | -1 | -1 | -1 | 0 | 0 | 0 | -1 | -1 | -1 | 0 |
| $\beta_0$ | 1 | 0 | 2/3 | 0 | 1/3 | -2/3 | 0 | -1/3 | 2/3 | 0 | 1/3 | 2 |
| $\beta_1$ | 0 | 1 | -1/3 | 0 | 1/3 | 1/3 | 0 | -1/3 | -1/3 | 0 | 1/3 | 1 |
| $s_2$ | 0 | 0 | 1/3 | -1 | 2/3 | -1/3 | 1 | -2/3 | 1/3 | -1 | 2/3 | 0 |

De TBO is: $(\beta_0, \beta_1, u_1, u_2, u_3, s_1, s_2, s_3) = (2, 1, 0, 0, 0, 0, 0, 0)$. Deze is niet optimaal, want er zijn meer termen gelijk aan nul op de nulde rij, dan er basisvariabelen zijn. Bij de keuze om $k_2$ in plaats van $k_3$ uit de basis te halen, had $s_1$ of $s_3$ nu in de basis gezeten. De waarden van $\beta_0$ en $\beta_1$ zijn wel optimaal.

### 3.1.3 Kwantielregressie met Stata

De controle methode is aan de hand van het softwarepakket Stata. We maken van ons LP een fictieve dataset in Stata. Vul daarom de gegevens voor voorbeeld 1 in Stata (Appendix B: Stata commando's). De afhankelijke variabele is de vector $y = (1, 3, 4)'$ en we hebben de volgende matrix:

$$X = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$$

Met het commando *qreg* verkrijgen we de waardes voor $\hat{\beta}_0$ en $\hat{\beta}_1$ voor de kwantielregressie. Het kwantielregressie commando werkt echter niet (zie Figuur 2). Daarom proberen wij het commando *bsqreg*, dit is het commando om de bootstrap methode uit te voeren (zie Figuur 2). De bootstrap methode is een alternatieve methode voor de kwantielregressie [4].

```
. qreg y x
note: weighted least squares perfect fit

Iteration  1: sum of abs. weighted deviations =          0
convergence not achieved.
VCE computation failed; try a different bandwidth or bsqreg
r(498);

. bsqreg y x
(fitting base model)
convergence not achieved.
convergence not achieved
r(430);
```

Figuur 2: Stata output voorbeeld 1

Dit commando werkt echter ook niet (zie Figuur 2). Volgens Stata is de

oplossing eenduidig en uniek en kan er geen oplossing gevonden worden. We vermoeden dat dit komt doordat het LP voor voorbeeld 1 één snijpunt heeft (zie Figuur 1). Wij zullen daarom voorbeeld 2 introduceren, waar de drie lijnen elkaar niet in hetzelfde punt snijden.

## 3.2 Voorbeeld 2

Met voorbeeld 1 kon geen kwantielregressie gedaan worden in Stata (zie sectie 3.1.3). Om het vermoeden uit de vorige sectie te onderzoeken, introduceren wij voorbeeld 2. In voorbeeld 2 snijden de drie lijnen elkaar niet in hetzelfde punt. Bekijk voorbeeld 1, maar pas voorbeeld 1 gedeeltelijk aan. Bij de tweede vergelijking verandert het $\leq$ teken in een $\geq$ teken. En in de derde vergelijking wordt de 4 een 5. Door deze aanpassing snijden de lijnen elkaar niet meer in één punt. Wij verkrijgen het volgende LP voor het bivariate voorbeeld 2 [3]:

$$\begin{gathered}
\text{(P) Min } u_1 + u_2 + u_3 \\
\text{o.d.v.} \\
\beta_0 - \beta_1 + u_1 \geq 1 \\
\beta_0 + \beta_1 + u_2 \geq 3 \\
\beta_0 + 2\beta_1 + u_3 \geq 5 \\
\beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3 \geq 0
\end{gathered}$$

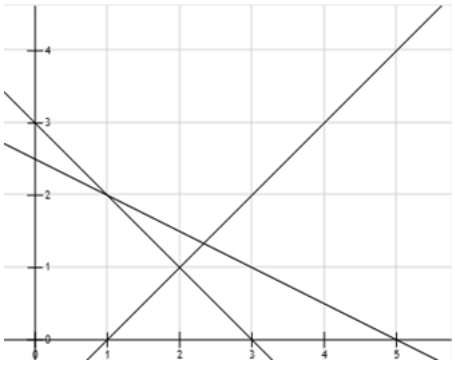
Vanuit dit LP kunnen de volgende voorwaarden worden opgesteld:

$$\begin{gathered}
\beta_0 - \beta_1 \leq 1 \\
\beta_0 + \beta_1 \geq 3 \\
\beta_0 + 2\beta_1 \leq 5
\end{gathered}$$

We gaan nu in sectie 3.2.1 het bovenstaande LP met de grafische methode oplossen. In sectie 3.2.2 lossen we het LP voor voorbeeld 2 analytisch op met de simplexmethode en in sectie 3.2.3 wordt de gevonden oplossing gecontroleerd met het softwarepakket Stata.

### 3.2.1 Grafische methode

Wij kunnen voorbeeld 2 oplossen met de grafische methode. Daarom beginnen wij met de bovenstaande voorwaarden in een $\beta_0 - \beta_1-$grafiek te plotten (zie Figuur 3). Merk op dat de drie lijnen elkaar niet in één punt snijden. Het toegelaten gebied van deze lijnen wordt weergegeven door het gebied tussen de hoekpunten $(2, 1)$, $(2\frac{1}{3}, 1\frac{1}{3})$ en $(1, 2)$. De oplossing is het punt $(\hat{\beta}_0, \hat{\beta}_1) = (2\frac{1}{3}, 1\frac{1}{3})$, omdat dit punt het meest oplevert.



Figuur 3: Snijdende lijnen voorbeeld 2, $\beta_0$ op x-as, $\beta_1$ op y-as

### 3.2.2 Simplexmethode

We kunnen dit LP ook analytisch oplossen met behulp van de simplexmethode (sectie 2.2). Wij bekijken het LP voor voorbeeld 2 [3]:

$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 \geq 1 \\ \beta_0 + \beta_1 + u_2 \geq 3 \\ \beta_0 + 2\beta_1 + u_3 \geq 5 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3 \geq 0\end{gathered}$$

Merk op dat wij te maken hebben met de nevenvoorwaarden $Ax \geq b$. Om dit in de gebruikelijke $Ax \leq b$ te zetten, voegen we de spelingsvariabelen $s_i$ (voor $i = 1, 2, 3$) met een negatieve coëfficiënt toe. Dit levert het volgende LP op:

$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 - s_1 = 1 \\ \beta_0 + \beta_1 + u_2 - s_2 = 3 \\ \beta_0 + 2\beta_1 + u_3 - s_3 = 5 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3, s_1, s_2, s_3 \geq 0\end{gathered}$$

Merk op dat wij nu geen basis kunnen creëren met de eenheidsvectoren voor de spelingsvariabelen $s_i$. Daarom moeten wij kunstmatige variabele $k_i$ voor $i = 1, 2, 3$ toevoegen. Het volgende LP ontstaat:

$$\begin{gathered}\text{(P) Min } u_1 + u_2 + u_3 \\ \text{o.d.v.} \\ \beta_0 - \beta_1 + u_1 - s_1 + k_1 = 1 \\ \beta_0 + \beta_1 + u_2 - s_2 + k_2 = 3 \\ \beta_0 + 2\beta_1 + u_3 - s_3 + k_3 = 5 \\ \beta_0, \beta_1 \in \mathbb{R} \ \& \ u_1, u_2, u_3, s_1, s_2, s_3, k_1, k_2, k_3 \geq 0\end{gathered}$$

Merk op dat wij dit LP kunnen oplossen met de simplexmethode. Zet allereerst de waarden in het simplextableau. Neem de kunstmatige variabelen $k_i$ als basisvariabelen, corrigeer daarvoor met een hulprij $z_1$. De oorspronkelijke doelstellingsfunctie $z$ wordt daaronder weergegeven. Verder worden de waardes uit het bovenstaande LP letterlijk overgenomen:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | 0 |
| z | 0 | 0 | -1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $k_1$ | 1 | -1 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $k_2$ | 1 | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 3 |
| $k_3$ | 1 | 2 | 0 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 5 |

Wij kunnen echter nog geen iteratie met dit tableau uitvoeren, omdat er bij de basisvariabelen $k_i$ geen eenheidsvectoren staan. Creëer een nieuwe doelstellingsfunctie $z_2$. Tel hiervoor alle rijen in het tableau bij elkaar op. Onder onze basisvariabelen $k_i$ staan nu de eenheidsvectoren:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 3 | 2 | 0 | 0 | 0 | -1 | -1 | -1 | 0 | 0 | 0 | 9 |
| $k_1$ | 1 | -1 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $k_2$ | 1 | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 3 |
| $k_3$ | 1 | 2 | 0 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 5 |

Merk op dat wij nu met ons iteratieproces kunnen starten, omdat er positieve waarden staan op de nulde rij. Het levert het meest op om $\beta_0$ in de basis te brengen. Aan de hand van de ratiotest (sectie 2.3) gaat $k_1$ uit de basis. Wij pivoteren nu op de kolom van $\beta_0$:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 0 | 5 | -3 | 0 | 0 | 2 | -1 | -1 | -3 | 0 | 0 | 6 |
| $\beta_0$ | 1 | -1 | 1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 1 |
| $k_2$ | 0 | 2 | -1 | 1 | 0 | 1 | -1 | 0 | -1 | 1 | 0 | 2 |
| $k_3$ | 0 | 3 | -1 | 0 | 1 | 1 | 0 | -1 | -1 | 0 | 1 | 4 |

Nu levert het het meest op als wij $\beta_1$ in de basis brengen. Met behulp van de ratioregel moet $k_2$ uit de basis. Nu pivoteren wij op de kolom van $\beta_0$:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 0 | 0 | -1/2 | -5/2 | 0 | -1/2 | 3/2 | -1 | -1/2 | -5/2 | 0 | 1 |
| $\beta_0$ | 1 | 0 | 1/2 | 1/2 | 0 | -1/2 | -1/2 | 0 | 1/2 | 1/2 | 0 | 2 |
| $\beta_1$ | 0 | 1 | -1/2 | 1/2 | 0 | 1/2 | -1/2 | 0 | -1/2 | 1/2 | 0 | 1 |
| $k_3$ | 0 | 0 | 1/2 | -3/2 | 1 | -1/2 | 3/2 | -1 | 1/2 | -3/2 | 1 | 1 |

Alleen onder $s_2$ staat nu nog een positief getal op de nulde rij, dus $s_2$ moet in de basis. Met behulp van de ratioregel gaat $k_3$ nu alsnog uit de basis:

| | $\beta_0$ | $\beta_1$ | $u_1$ | $u_2$ | $u_3$ | $s_1$ | $s_2$ | $s_3$ | $k_1$ | $k_2$ | $k_3$ | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $z_2$ | 0 | 0 | -1 | -1 | -1 | 0 | 0 | 0 | -1 | -1 | -1 | 0 |
| $\beta_0$ | 1 | 0 | 2/3 | 0 | 1/3 | -2/3 | 0 | -1/3 | 2/3 | 0 | 1/3 | 7/3 |
| $\beta_1$ | 0 | 1 | -1/3 | 0 | 1/3 | 1/3 | 0 | -1/3 | -1/3 | 0 | 1/3 | 4/3 |
| $s_2$ | 0 | 0 | 1/3 | -1 | 2/3 | -1/3 | 1 | -2/3 | 1/3 | -1 | 2/3 | 2/3 |

De TBO is: $(\beta_0, \beta_1, u_1, u_2, u_3, s_1, s_2, s_3) = (2\frac{1}{3}, 1\frac{1}{3}, 0, 0, 0, 0, \frac{2}{3}, 0)$. Deze oplossing is niet optimaal, want er zijn meer termen gelijk aan nul op de nulde rij, dan er basisvariabelen zijn. De waarden van $\beta_0$ en $\beta_1$ zijn wel optimaal.

### 3.2.3 Kwantielregressie met Stata

De controle methode is met behulp van het softwarepakket Stata. We kunnen in Stata met het commando *qreg* of *bsqreg* $\hat{\beta}_0$ en $\hat{\beta}_1$ bepalen (zie sectie 3.1.3). Creëer daarom een fictieve dataset voor voorbeeld 2 in Stata (zie Appendix B: Stata input). De afhankelijke variabele is de vector $y = (1, 3, 5)'$ en we hebben de volgende matrix:

$$X = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$$

Het verschil met voorbeeld 1 is dat de lijnen elkaar nu niet in één punt snijden (zie Figuur 3). Door de fictieve dataset te bootstrappen (zie Figuur 4), kunnen wij de coëfficiënten voor $\hat{\beta}_0$ en $\hat{\beta}_1$ aflezen.

```
. bsqreg y x
(fitting base model)

Bootstrap replications (20)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
x...........x....x....

Median regression, bootstrap(20) SEs                Number of obs =          3
  Raw sum of deviations        2 (about 3)
  Min sum of deviations .3333333                   Pseudo R2     =     0.8333

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    1.333333   .3940345     3.38   0.183    -3.673349    6.340016
       _cons |    2.333333   .4748037     4.91   0.128     -3.69962    8.366287
------------------------------------------------------------------------------
```

Figuur 4: Stata output voorbeeld 2

De oplossing is $(\hat{\beta}_0, \hat{\beta}_1) = (2\frac{1}{3}, 1\frac{1}{3})$, wat overeenkomt met de oplossingen in sectie 3.2.1 en 3.2.2.

# 4 Toepassing: regressie-analyse op WoOn 2015

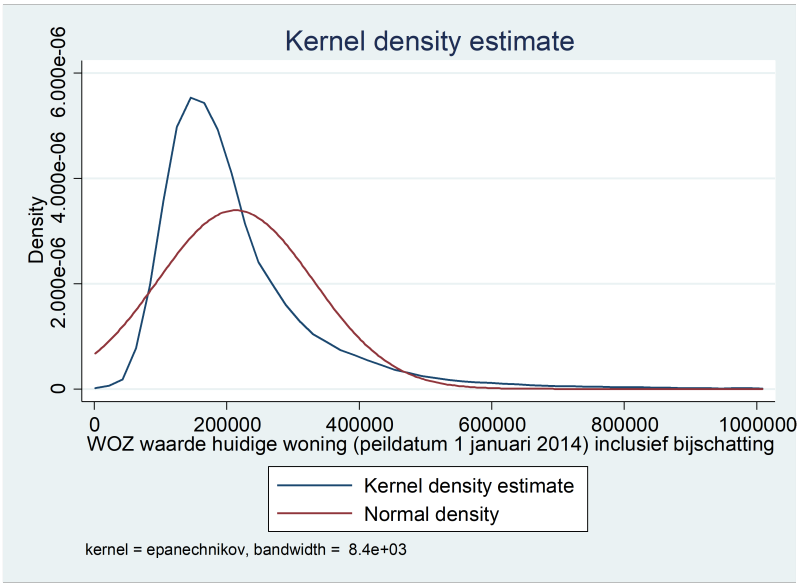
In sectie 3.2.3 is voorbeeld 2 in Stata geanalyseerd met de kwantielregressie. Om de meerwaarde van de kwantielregressie schatter ten opzichte van de OLS schatter te achterhalen, analyseren wij in dit hoofdstuk de dataset WoOn 2015. Eerst wordt de dataset WoOn 2015 [2] beschreven. In sectie 4.2 gaan wij een lineair regressie model en een kwantielregressie model opstellen voor de afhankelijke variabele, de WOZ-waarde. In sectie 4.3 worden de resultaten van de verschillende regressies vergeleken. De effecten van de onafhankelijke variabelen op de WOZ-waarde worden beschreven in verschillende kwantielen en voor het gemiddelde.

## 4.1 Data

Wij gebruiken de dataset WoOn 2015 [2]. Het WoOn 2015 is door BZK ontwikkeld met ondersteuning van het CBS. Dit WoonOnderzoek wordt elke drie jaar afgenomen onder een groot aantal huishoudens. In totaal zijn er 62668 huishoudens geënqueteerd. *Het WoOn doet onderzoek naar de woonkwaliteit en woonbehoefte ter ondersteuning van het regeringsbeleid op het gebied van wonen. In het WoON komt informatie over huishoudenssituatie, huidige en gewenste woonsituatie, woonlasten en inkomens samen* [10]. Vervolgens worden de gegevens van o.a. de Gemeentelijke Basis Administratie-GBA en de Belastingdiest toegevoegd aan de resultaten.

## 4.2 Statistische modellen

Om de meerwaarde van de kwantielregressie ten opzichte van de OLS methode aan te tonen, worden er in deze sectie twee statistische modellen opgesteld. Het is bekend dat een model bestaat uit een afhankelijke variabele en meerdere onafhankelijke variabelen. We kiezen als afhankelijke variabele de WOZ-waarde (in Stata: WOZwaarde). De afkorting WOZ staat voor wet waardering onroerende zaken. De WOZ-waarde is de waarde van de woning en wordt ieder jaar door de gemeente vastgesteld [10]. De WOZ-waarde wordt door de gemeente gebruikt voor gemeentelijke belastingen, heffingen (denk aan onroerendezaakbelasting, rioolheffing en waterschapslasten) en het speelt een belangrijke rol in erf- of schenkbelasting. In Figuur 5 wordt de Kernel density van de WOZ-waarde weergegeven. Merk op dat wij de waardes van de WOZ-waarde die groter dan een miljoen zijn niet hebben meegenomen. De hoge uitschieters kunnen namelijk het gemiddelde van de WOZ-waarde negatief beïnvloeden.




Figuur 5: Kernel density WOZ-waarde

De WOZ-waarde hangt af van de ligging van de woning, de kenmerken van de woning en het perceeloppervlak. In de dataset WoOn 2015 zitten de variabelen perceeloppervlak en een aantal kenmerken. De ligging van de woning is niet verwerkt in de dataset. Wij kiezen daarom als onafhankelijke variabelen het perceeloppervlak (in Stata: gebruiksopp), hoeveel kamers de woning heeft (in Stata: Kamers) en of de woning een tuin heeft (in Stata: tuin1). Het gemiddelde en de mediaan van de WOZ-waarde en de verschillende controlevariabelen staan in Tabel 1.

| Variabele | Gemiddelde | Mediaan |
|---|---|---|
| WOZ-waarde (in euro's) | 213620 | 183000 |
| Gebruiksoppervlakte (in $m^2$) | 123 | 110 |
| Aantal kamers | 4,36 | 4 |
| Tuin | 0,71 | 1 |

Tabel 1: N=54936, gemiddelde en mediaan voor variabelen

Weet dat er 62668 respondenten zijn, maar slechts 55095 respondenten hebben ingevuld of zij wel of geen tuin hebben. Doordat wij de waardes voor de WOZ-waarde boven één miljoen euro uit de regressie halen, houden wij 54936 respondenten over. We hebben nu voldoende informatie om het lineaire regressie model en het kwantielregressie model op te stellen:

**Lineair regressie model**

$$WOZwaarde = \beta_0 + \beta_1 gebruiksopp + \beta_2 Kamers + \beta_3 tuin1 + u_1 \quad (6)$$

**Kwantielregressie model**

$$WOZwaarde_q = \beta_0 + \beta_1(q) gebruiksopp + \beta_2(q) Kamers + \beta_3(q) tuin1 + u_2 \quad (7)$$

De kwantielen waarvoor de WOZ-waarde geschat gaat worden (zie sectie 4.3), zullen variëren tussen $q = 0.05$ en $q = 0.95$ (zie Tabel 2). Zo krijgen wij een goed beeld van de gehele verdeling en zullen de uitschieters de schatting niet ongewenst beïnvloeden. We behandelen in de volgende sectie het gebruikelijke significantieniveau van vijf procent [9].

## 4.3 Resultaten

Om een statistisch model te analyseren moet aan de aannames voor het model voldaan zijn. Daarom beginnen wij met het checken voor de zuiverheid van (6) en (7). Vervolgens zullen wij een economische intepretatie geven aan de geschatte waardes voor de parameters. En tot slot worden de coëfficiënten voor de mediaan-regressie vergeleken met de coëfficiënten voor het OLS model.

**Aannames OLS**
Het is belangrijk om te onderzoeken of het model voldoet aan de aannames voor de methode, anders zijn de resultaten niet betrouwbaar. We gebruiken de OLS-methode om het lineaire regressiemodel te analyseren. Het is bekend dat de volgende aannames moeten gelden voor een algemeen OLS model [9]:

1. De parameters in het populatiemodel zijn lineair
2. Er zijn meer observaties dan variabelen in het model ($n > k + 1$)
3. Alle variabelen zijn onafhankelijk van de foutterm ($E(u|X) = 0$)
4. Geen perfecte multicollineariteit tussen de variabelen
5. Er is geen autocorrelatie ($Corr(u_i, u_j) = 0$)
6. Geen heteroskedasticiteit ($Var(u_i) = \sigma^2$, met $\sigma$ een constante)

Merk op dat aanname vijf over de autocorrelatie niet opgaat voor ons model, aangezien ons model geen panel data is. Gegeven (6) is het duidelijk dat de parameters in het model lineair zijn. Aan de tweede aanname is ook voldaan ($n = 54936 > 4 = k + 1$). Door de grote dataset in combinatie met de Centrale Limiet Stelling [7], hoeft er geen normaliteit van de erroterm vereist te worden. De derde aanname over strikte exogeniteit is lastig

te testen. Wij kunnen de aanname wel aannemen, omdat er hoge pieken zijn bij nul (Appendix B: Figuur 6). Multicollineariteit tussen variabelen wordt in Stata uitgesloten, omdat Stata zelf de variabele die multicollineariteit veroorzaakt weg laat. Ook aan de $R^2$ (zie tabel 2) zien we dat er geen sprake is van multicollineariteit (Appendix B: Figuur 7 en Figuur 8, toont nog twee methodes). De homoskedasticiteit van de fouttermen wordt getest met de White-test (Appendix B: Figuur 9 en 10). De fouttermen zijn allen heteroskedastisch, daarom zullen wij met robuuste standaardfouten werken. Hierdoor zijn de standaardfouten groter dan met een normale regressie, de coëfficiënten voor de variabelen blijven hetzelde. Een alternatieve test voor de heteroskedasticiteit is de Breusch Pagan test (Appendix B: Figuur 11), dit levert uiteraard dezelfde conclusie op.

**Aannames kwantielregressie**
Om ook de betrouwbaarheid van (7) te waarborgen, moeten wij de aannames uit sectie 2.2 testen. Aan de eerste aanname over voldoende observaties ($n = 54936 > 4 = k+1$) is natuurlijk voldaan. Aan de tweede aanname over de heteroskedasticiteit van de foutterm is ook voldaan (Appendix B: Figuur 12). Hierdoor snijden de geschatte lijnen elkaar niet en blijft de natuurlijke ordening van de kwantielen behouden.

**Economische interpretatie**
Als aan de aannames voor (6) en (7) is voldaan en de geschatte coëfficiënten (zie tabel 2) zijn significant, dan kunnen we een betrouwbare economische interpretatie geven over het effect van de afhankelijke variabele op de WOZ-waarde. Uit Stata volgt dat alle geschatte coëfficiënten F-waarde 0.000 hebben, waardoor het significantieniveau van vijf procent behaald is.
De economische interpretatie van tabel 2 is als volgt. Als wij bijvoorbeeld kijken naar de coëfficiënten voor de mediaan regressie $\xi_{0.50}$, dan is de constante $\hat{\beta}_0$ gelijk aan 21833 euro. Als het perceeloppervlak met één vierkante meter stijgt, dan stijgt de WOZ-waarde in het $50^e$ kwantiel met 1209 euro. Als de woning één kamer meer heeft, dan stijgt de WOZ-waarde in het $50^e$ kwantiel met 6186 euro. En als de woning een tuin heeft, dan stijgt de WOZ-waarde in het $50^e$ kwantiel met 5179 euro. Op dezelde manier kunnen wij voor de andere geschatte coëfficiënten in tabel 2 een uitspraak doen over het effect van de onafhankelijke variabelen op de WOZ-waarde.

Het verschil in het effect van de onafhankelijke variabele op de WOZ-waarde in de verschillende kwantielen is groot. Daarom beschrijven wij nu voor iedere parameter de verandering naar mate de kwantielen groter worden. De constante $\hat{\beta}_0$ heeft een grote invloed in een laag kwantiel. Deze invloed neemt af naar mate wij in een hoger kwantiel komen. Behalve in een extreem hoog kwantiel, daar heeft $\hat{\beta}_0$ de hoogste waarde. Merk op dat dit een dalparaboolachtig figuur is. De invloed van de parameter $\hat{\beta}_1$ voor het

| | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $R^2$ |
|---|---|---|---|---|---|
| $\xi_{0,05}$ | 27400 | 272 | 9139 | 15418 | 0.1460 |
| | (1265) | (7,31) | (333) | (1050) | |
| $\xi_{0,25}$ | 26367 | 763 | 7701 | 11305 | 0.2216 |
| | (936) | (5,41) | (247) | (777) | |
| $\xi_{0,50}$ | 21833 | 1209 | 6186 | 5179 | 0.2840 |
| | (987) | (5,71) | (260) | (819) | |
| $\xi_{0,75}$ | 19100 | 1650 | 6800 | -4150 | 0.3418 |
| | (1349) | (7,80) | (355) | (1119) | |
| $\xi_{0,95}$ | 46573 | 2285 | 13830 | -40254 | 0.3701 |
| | (4136) | (23,9) | (1090) | (3433) | |
| OLS | 35488 | 796 | 16501 | 9759 | 0.3983 |
| | (1136) | (6,57) | (299) | (943) | |

Tabel 2: Gegeven N = 54936, parameters voor (6) en (7) en Std. err. (.)

perceeloppervlak stijgt naar mate wij in een hoger kwantiel komen (van 279 euro in $\xi_{0.05}$ naar 2285 euro in $\xi_{0.95}$). De parameter $\hat{\beta}_2$, voor het aantal kamers in de woning, vertoont ongeveer hetzelfde gedrag als de constante $\hat{\beta}_0$. Voor het wel of geen tuin hebben (parameter $\hat{\beta}_3$) daalt de invloed op de WOZ-waarde als wij in een hoger kwantiel komen.

Tot slot worden de resultaten voor de kwantielregressie en OLS vergeleken. Omdat de kwantielregressie uitgaat van de mediaan, vergelijken wij de geschatte coëfficiënten van de mediaan regressie ($\xi_{0,50}$) met de geschatte coëfficiënten voor OLS. De constante $\hat{\beta}_0$ is lager bij de mediaan-regressie dan bij OLS. De parameter $\hat{\beta}_1$ is hoger bij de mediaan-regressie dan OLS waarde. De OLS waarde voor parameter $\hat{\beta}_2$ is bijna drie keer zo groot als voor $\xi_{0.50}$. Merk op dat de OLS waarde van $\hat{\beta}_2$ groter is dan alle verschillende waardes in de kwantielen. Voor parameter $\hat{\beta}_3$ is de OLS waarde weer hoger dan de mediaan-regressie. De OLS waarde is dus alleen lager voor de invloed van het perceeloppervlak op de WOZ-waarde ten opzichte van de mediaan regressie.

# 5 Conclusie en discussie

In dit hoofdstuk wordt mijn onderzoek afgerond. In sectie 5.1 worden conclusies besproken over de oplosmethode voor de mediaan regressie schatter. Ook wordt de meerwaarde van de kwantielregressie boven de OLS methode besproken. In sectie 5.2 worden vermoedens uitgesproken, een vervolg onderzoek geopperd en bijzondere gevallen toegelicht.

## 5.1 Conclusie

Het doel van mijn onderzoek was om de optimalisatiemethode toe te lichten die ten grondslag ligt aan de mediaanregressie schatter. In sectie 3.1.2 en in sectie 3.2.2 hebben wij het LP voor voorbeeld 1 en voorbeeld 2 opgelost met behulp van de simplexmethode. De optimalisatietechniek van de simplexmethode ligt dus ten grondslag aan de afleiding van de mediaanregressie schatter. In voorbeeld 1 kon het LP niet met Stata worden opgelost, in voorbeeld 2 wel.

In hoofdstuk 4 werd de dataset WoOn 2015 geanalyseerd met behulp van de schatters voor OLS en de kwantielregressie. Aan de hand van de Kernal density (Figuur 5) van de WOZ-waarde kunnen wij opmaken dat er verschillende groepen gemaakt kunnen worden in de verdeling. Er is een groep met een lage WOZ-waarde en een hoge WOZ-waarde. In sectie 2.2 is bekend geworden dat wij verschillende delen van de verdeling kunnen analyseren met de kwantielregressie methode. Het is mogelijk dat in de groep met lage WOZ-waardes (lage kwantielen) de effecten van de onafhankelijke variabelen anders zullen zijn dan bij hoge WOZ-waardes (hoge kwantielen). In tabel 2 zien wij dat voor de verschillende kwantielen de effecten van de onafhankelijke variabelen op de WOZ-waarde verschillend zijn. We hebben dus meer informatie tot onze beschikking met de kwantielregressie. Hieruit kunnen wij concluderen dat de kwantielregressie voor dit model een meerwaarde is ten opzichte van OLS.

## 5.2 Discussie

Zoals gezegd hebben wij het LP voor de mediaanregressie schatter opgelost met de simplexmethode. Wij hebben twee bivariate voorbeelden bekeken. In voorbeeld 1 konden wij geen controle met Stata doen, want de oplossing was eenduidig en uniek. Het vermoeden was dat dit kwam doordat het LP slechts één snijpunt had. In voorbeeld 2 werkten wij daarom met een LP met drie snijpunten. In sectie 3.2.3 konden wij wél in Stata de gewenste controle uitvoeren. Het vermoeden is daarom bevestigd en ik denk daarom dat wij een kwantielregressie model pas in Stata kunnen toepassen als het LP meer dan één snijpunt heeft. Ik denk dat dit te maken kan hebben met

perfecte multicollineariteit, maar daar is vervolg onderzoek voor nodig.

Het is bekend dat voor ons model de kwantielregressie een meerwaarde is ten opzichte van OLS. Wij kunnen uitspraken doen over de effecten van onafhankelijke variabele op de WOZ-waarde en in het bijzonder wat de WOZ-waarde van de woning hoger maakt in een laag en hoog kwantiel. Het bezitten van een tuin in een laag kwantiel zorgt bijvoorbeeld voor een positieve invloed op de WOZ-waarde. Maar in een hoog kwantiel heeft een tuin een negatieve invloed; er wordt vermoedelijk vanuit gegaan dat woningen met een hoge WOZ-waarde al een tuin bezitten. Daarom zijn in een hoog kwantiel het perceeloppervlak en het aantal kamers meer van positieve invloed op de WOZ-waarde. Verder is een bijzonder gegeven in tabel 2 dat de OLS waarde voor de parameter voor het aantal kamers veel hoger ligt dan in alle kwantielen. Dit komt denk ik doordat in tabel 1 het gemiddelde aantal kamers ook al hoger is dan voor de mediaanregressie.

Voor een vervolgonderzoek kan dus gekeken worden of perfecte multicollineariteit iets te maken heeft met de fout in Stata voor ons voorbeeld. Ook het aantal observaties zou vergroot kunnen worden. Het is namelijk bekend dat voor grote datasets de kwantielregressie in Stata wel werkt (zie sectie 4.3). Daarnaast zou er onderzoek gedaan kunnen worden naar andere itererende methodes, de interior point method schijnt sneller te werken voor grote datasets in Stata dan de kwantielregressie methode.

# Referenties


[1] Angrist, J., & Pischke, J. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press

[2] BZK/CBS, WoON 2015

[3] Davino, C., & Furno, M., & Vistocco, D. (2014). *Quantile Regression.* Somerset: John Wiley & Sons

[4] Hassink, W.H.J. (2017). *Econometrics*, Chapter 1, 2, 3, 9 & 22

[5] He, X., & Wang, H. Y. (2011). *A Short course on Quantile Regression*

[6] Hoogeveen, J.A. (2018). *Optimalisering (en Complexiteit).* Geraadpleegd van http://www.cs.uu.nl/docs/vakken/opt/

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

[8] Portnoy, S., & Koenker, R. (1997). Statistical Science. *The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error versus Absolute-Error Estimators*

[9] Solomons, A. (2016). *Econometrics.* Geraadpleegd van https://www.uu.blackboard.com/bbcswebdav/pid-2508189-dt-content-rid-73946172/courses/REBO-2016-2-ECB2METRIE-V/L3-20161128-short.pdf

[10] [WOZ-waarde aanvullende informatie]. (2018). Geraadpleegd van https://www.hypotheker.nl/begrippenlijst/wet-en-regelgeving/woz-waarde/

# 6 Appendix A: Wiskundige bewijzen

## 6.1 Bepaling $\xi_{0.5}$

**Te bewijzen:** $\xi_{0.5} = \arg\min_{\beta\in\mathbb{R}} E|Y - X\beta|$
**Bewijs:** Met behulp van de definitie voor de verwachte waarde $E(.)$ geldt:

$$E|Y - X\beta| = \int_{\beta}^{b}(y - X\beta)f(y)\,dy + \int_{a}^{\beta}(X\beta - y)f(y)\,dy \tag{8}$$

Omdat dit probleem een minimaliseringsprobleem voorstelt, moeten wij de verschillende afgeleiden nemen naar $\beta$. Dit doen wij aan de hand van de Leipnitz regel [4].

$$\frac{d}{d\beta}\int_{\beta}^{b}(y - X\beta)f(y)\,dy = -\int_{\beta}^{b}f(y)\,dy \tag{9}$$

$$\frac{d}{d\beta}\int_{a}^{\beta}(X\beta - y)f(y)\,dy = \int_{a}^{\beta}f(y)\,dy \tag{10}$$

Wij kunnen nu het minimaliseringsprobleem

$$\frac{d}{d\beta}E|Y - X\beta| = 0 \tag{11}$$

oplossen door de formules (9) en (10) in (8) te stoppen. Dit geeft ons het gewenste resultaat:

$$\int_{a}^{\beta}f(y)\,dy = 0.5 \tag{12}$$

## 6.2 Bepaling $\xi_q$

**Te bewijzen:** $\xi_q = \arg\min_{\beta\in\mathbb{R}} E\left[p_q(Y - X\beta)\right]$
**Bewijs:** Wij beginnen met het definiëren van de indicatorfunctie $I(.)$:

$$I(Y \geq X\beta) = \begin{cases} 1 & \text{Als } Y \geq X\beta \\ 0 & \text{Als } Y < X\beta \end{cases} \qquad I(Y < X\beta) = \begin{cases} 1 & \text{Als } Y < X\beta \\ 0 & \text{Als } Y \geq X\beta \end{cases}$$

Vanuit hier vinden wij met behulp van het te bewijzen bij sectie 2.2 het volgende resultaat

$$\xi_q = \arg\min_{\beta\in\mathbb{R}} q \cdot E\left(Y - X\beta\right) \cdot I(Y \geq X\beta) + (1-q) \cdot E\left(X\beta - Y\right) \cdot I(Y < X\beta) \tag{13}$$

En nu kunnen wij met behulp van de definitie voor de verwachte waarde $E(.)$ gemakkelijk vinden dat:

$$q \cdot E\left(Y - X\beta\right) \cdot I(Y \geq X\beta) + (1-q) \cdot E\left(X\beta - Y\right) \cdot I(Y < X\beta) =$$

$$q \int_{\beta}^{b} (y - X\beta) f(y)\, dy + (1 - q) \int_{a}^{\beta} (X\beta - y) f(y)\, dy$$

(14)

Los nu weer het minimaliseringsprobleem

$$q \cdot E\left(Y - X\beta\right) \cdot I(Y \geq X\beta) + (1 - q) \cdot E\left(X\beta - Y\right) \cdot I(Y < X\beta) = 0 \quad (15)$$

op door middel van de afgeleiden naar $\beta$ te nemen voor de verschillende integralen in vergelijking (14). Dit levert het gewenste resultaat op:

$$\int_{a}^{\beta} f(y)\, dy = q \tag{16}$$

Door simpelweg $p_q(Y) = q \cdot I(Y \geq X\beta) + (1 - q) \cdot I(Y < X\beta)$ te implementeren, krijgen we de algemene formule

$$\xi_q = \arg\min_{\beta \in \mathbb{R}} E\left[p_q(Y - X\beta)\right] \tag{17}$$

### 6.3 Schatter $\hat{\beta}(q)$ voor kwantielregressie

Gegeven functie (17) op populatieniveau:

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

Introduceer functie $d$ die de totale $\rho$-afwijkingen weergeeft.

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

Dan volgt na invullen van gegevens voor $\hat{y}(\beta)$

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

Voor de kleinste totale afwijking moet het argument $\beta$ bepaald worden waarvoor de afwijking minimaal is, ofwel dan is de kwantielregressie schatter op steekproefniveau

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

# 7 Appendix B: Stata

In deze sectie behandelen wij de Stata commando's en de Stata output.

## 7.1 Stata commando's

In deze sectie wordt per sectie behandelt welke Stata commando's er zijn ingevoerd om de gewenste figuren en tabellen te krijgen voor de verschillende secties. Na elke komma en elke nieuwe stap volgt een *enter*.

### Sectie 3.1.3

Het model voor voorbeeld 1 wordt in Stata ingevoerd:

1. Vul in: input y x
2. Vul achtereenvolgens in: 1 -1, 3 1, 4 2, end, list
3. Vul in: qreg, bsqreg (zie Figuur 2)

### Sectie 3.2.3

Het model voor voorbeeld 2 wordt in Stata ingevoerd:

1. Vul in: input y x
2. Vul achtereenvolgens in: 1 -1, 3 1, 5 2, end, list
3. Vul in: bsqreg (zie Figuur 4)

### Sectie 4.2.1

In sectie 4.2.1 moeten de aannames worden getest voor het OLS model.
Strikte exogeniteit:

1. Vul in: reg WOZwaarde gebruiksopp Kamers tuin1
2. Vul in: predict e, residual
3. Vul in: histogram e (zie Figuur 6)

Multicollineariteit:

1. Vul in: reg WOZwaarde gebruiksopp Kamers tuin1
2. Vul in: vif (zie Figuur 7)
3. Vul in corr WOZwaarde gebruiksopp Kamers tuin1 (zie Figuur 8)

Heteroskedasticiteit:

1. Vul in: reg WOZwaarde gebruiksopp Kamers tuin1
2. Vul in: estat hettest (Figuur 9) óf estat imtest (Figuur 10)

**Sectie 4.2.2**

In sectie 4.2.2 moeten de aannames worden getest voor het kwantielregressie model. In Stata is dat alleen de aanname voor heteroskedasticiteit. Heteroskedasticiteit kwantielregressie: sqreg WOZwaarde gebruiksopp Kamers tuin1, q(0.05,0.25,0.50,0.75,0.95) (zie Figuur 12)

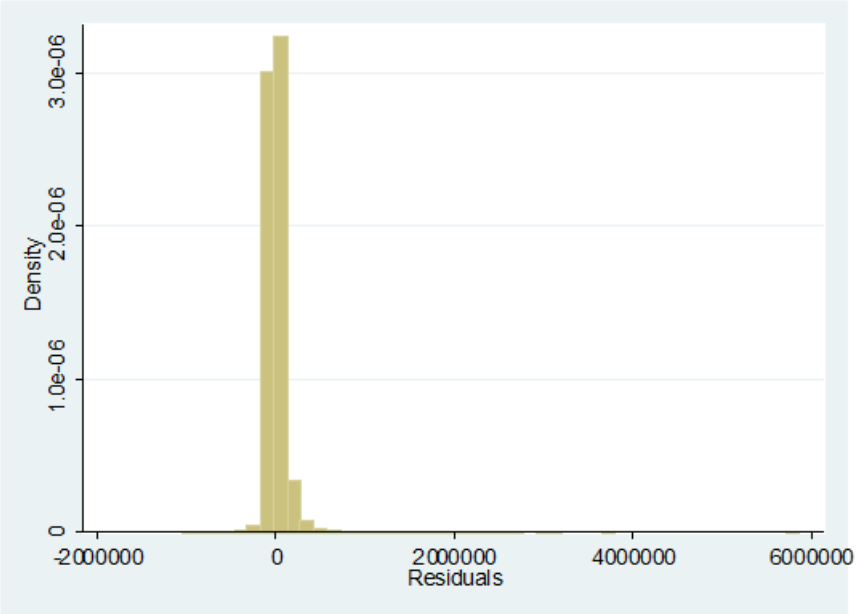
**Sectie 4.3**

In sectie 4.3 wordt Tabel 1 verkegen met de volgende Stata commando's:

1. qreg WOZwaarde gebruiksopp Kamers tuin1, q(0.05)
2. qreg WOZwaarde gebruiksopp Kamers tuin1, q(0.25)
3. qreg WOZwaarde gebruiksopp Kamers tuin1, q(0.50)
4. qreg WOZwaarde gebruiksopp Kamers tuin1, q(0.75)
5. qreg WOZwaarde gebruiksopp Kamers tuin1, q(0.95)
6. reg WOZwaarde gebruiksopp Kamers tuin1

## 7.2 Stata output

De onderstaande figuren zijn de output voor de commando's in sectie 8.1.




Figuur 6: Histogram voor strikte exogeniteit

```
. vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| Kamers | 1.55 | 0.647071 |
| gebruiksopp | 1.36 | 0.736897 |
| tuin1 | 1.25 | 0.798349 |
| Mean VIF | 1.39 | |

Figuur 7: Multicollineariteit

```
. corr WOZwaarde gebruiksopp Kamers tuin1
(obs=55,095)
```

| | WOZwaa~e | gebrui~p | Kamers | tuin1 |
|---|---|---|---|---|
| WOZwaarde | 1.0000 | | | |
| gebruiksopp | 0.5625 | 1.0000 | | |
| Kamers | 0.4339 | 0.5077 | 1.0000 | |
| tuin1 | 0.2332 | 0.2902 | 0.4426 | 1.0000 |

Figuur 8: Correlatie tussen variabelen

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of WOZwaarde

         chi2(1)      = 83229.48
         Prob > chi2  =   0.0000
```

Figuur 9: Test 1 heteroskedasticiteit OLS

```
. estat imtest

Cameron & Trivedi's decomposition of IM-test
```

| Source | chi2 | df | p |
|---|---|---|---|
| Heteroskedasticity | 1166.43 | 8 | 0.0000 |
| Skewness | -2524892.99 | 3 | 1.0000 |
| Kurtosis | -2.52e+18 | 1 | 1.0000 |
| Total | -2.52e+18 | 12 | 1.0000 |

Figuur 10: Test 2 heteroskedasticiteit OLS

```
. predict uhat, resid
(7,573 missing values generated)

. gen uhat2=uhat^2
(7,573 missing values generated)

. reg uhat2 gebruiksopp Kamers tuin1
```

| Source | SS | df | MS | Number of obs | = | 55,095 |
|---|---|---|---|---|---|---|
| | | | | F(3, 55091) | = | 305.01 |
| Model | 3.7150e+25 | 3 | 1.2383e+25 | Prob > F | = | 0.0000 |
| Residual | 2.2367e+27 | 55,091 | 4.0600e+22 | R-squared | = | 0.0163 |
| | | | | Adj R-squared | = | 0.0163 |
| Total | 2.2738e+27 | 55,094 | 4.1272e+22 | Root MSE | = | 2.0e+11 |

| uhat2 | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| gebruiksopp | 3.91e+08 | 1.42e+07 | 27.58 | 0.000 | 3.64e+08 | 4.19e+08 |
| Kamers | -7.01e+08 | 6.64e+08 | -1.05 | 0.292 | -2.00e+09 | 6.01e+08 |
| tuin1 | -1.63e+10 | 2.11e+09 | -7.75 | 0.000 | -2.05e+10 | -1.22e+10 |
| _cons | -1.96e+10 | 2.52e+09 | -7.78 | 0.000 | -2.46e+10 | -1.47e+10 |

Figuur 11: Breusch Pagan test voor heteroskedasticiteit OLS

```
. sqreg WOZwaarde gebruiksopp Kamers tuin1, q(0.05,0.25,0.50,0.75,0.95)
(fitting base model)

Bootstrap replications (20)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................
```

Simultaneous quantile regression — Number of obs = 55,095
bootstrap(20) SEs

| | |
|---|---|
| .05 Pseudo R2 = | 0.1441 |
| .25 Pseudo R2 = | 0.2195 |
| .50 Pseudo R2 = | 0.2814 |
| .75 Pseudo R2 = | 0.3384 |
| .95 Pseudo R2 = | 0.3709 |

| WOZwaarde | Coef. | Bootstrap Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| q05 | | | | | | |
| gebruiksopp | 286.5528 | 21.80351 | 13.14 | 0.000 | 243.8178 | 329.2879 |
| Kamers | 9027.748 | 522.846 | 17.27 | 0.000 | 8002.966 | 10052.53 |
| tuin1 | 14994.13 | 670.1296 | 22.37 | 0.000 | 13680.67 | 16307.59 |
| _cons | 26846.32 | 1239.791 | 21.65 | 0.000 | 24416.32 | 29276.32 |
| q25 | | | | | | |
| gebruiksopp | 784.4228 | 16.09142 | 48.75 | 0.000 | 752.8835 | 815.9621 |
| Kamers | 7522.949 | 352.1159 | 21.36 | 0.000 | 6832.799 | 8213.098 |
| tuin1 | 10778.86 | 543.5559 | 19.83 | 0.000 | 9713.486 | 11844.23 |
| _cons | 25443.67 | 1032.64 | 24.64 | 0.000 | 23419.69 | 27467.65 |
| q50 | | | | | | |
| gebruiksopp | 1235.294 | 18.388 | 67.18 | 0.000 | 1199.254 | 1271.335 |
| Kamers | 6000 | 361.7269 | 16.59 | 0.000 | 5291.013 | 6708.987 |
| tuin1 | 4470.588 | 814.8663 | 5.49 | 0.000 | 2873.444 | 6067.732 |
| _cons | 20529.41 | 918.3696 | 22.35 | 0.000 | 18729.4 | 22329.42 |
| q75 | | | | | | |
| gebruiksopp | 1710.432 | 21.57047 | 79.30 | 0.000 | 1668.153 | 1752.71 |
| Kamers | 6354.317 | 537.9811 | 11.81 | 0.000 | 5299.87 | 7408.763 |
| tuin1 | -5638.489 | 564.036 | -10.00 | 0.000 | -6744.004 | -4532.975 |
| _cons | 15915.47 | 1501.63 | 10.60 | 0.000 | 12972.26 | 18858.67 |
| q95 | | | | | | |
| gebruiksopp | 2534.884 | 53.52897 | 47.36 | 0.000 | 2429.967 | 2639.801 |
| Kamers | 12593.02 | 1404.962 | 8.96 | 0.000 | 9839.287 | 15346.76 |
| tuin1 | -47779.07 | 3438.208 | -13.90 | 0.000 | -54517.98 | -41040.16 |
| _cons | 34406.98 | 2330.195 | 14.77 | 0.000 | 29839.78 | 38974.18 |

Figuur 12: Test heteroskedasticiteit kwantiel