Bilag 1 - Beregning af omkostningsækvivalenter

Bilaget indeholder den tekniske beregning af omkostningsækvivalenterne til brug for benchmarkingen 2012.

FORSYNINGSSEKRETARIATET OKTOBER 2011

INDLEDNING 3

MINDSTE KVADRATERS METODE 3

Sammenhæng mellem omkostninger og costdriver 5

Normalfordeling af fejlled 5

Heteroskedasticitet 6

Multikollinaritet 7

Ekstreme observationer og observationer med stor indflydelse 7

Transformationer 9

Valg af den bedste model 10

Datakvalitet 11

BESTEMMELSE AF OMKOSTNINGSÆKVIVALENTER FOR VANDFORSYNINGER 12

Boringer 12

Vandværk 15

Trykforøgere 18

Rentvandsledning 21

Stik 25

Kunder 28

BESTEMMELSE AF OMKOSTNINGSÆKVIVALENTER FOR SPILDEVAND 31

Ledning 31

Pumper 35

Åbne Bassiner 40

Lukkede bassiner 42

Renseanlæg 44

Kunder 49

Indledning

Dette bilag har en relativ teknisk karakter, og er primært en gennemgang af den statistiske metode, der ligger til grund for Forsyningssekretariatets beregning af omkostningsækvivalenterne til brug for benchmarkingen af vand- og spildevandsforsyningerne.

Det er vigtigt, at resultaterne af benchmarkingen er så præcise som mulig. Ulempen ved at sigte mod størst mulig præcision er dog, at de metoder, der anvendes, kan være ret komplicerede. Forsyningssekretariatet har vurderet, at præcision i denne sammenhæng vægter højest.

I det følgende beskrives først den metode, Forsyningssekretariatet har brugt til at beregne omkostningsækvivalenterne. Dernæst er der en detaljeret gennemgang af beregningen af de seks omkostningsækvivalenter for hhv. vand- og spildevandsforsyningerne.

Dette bilag blev sendt i høring hos branchen den 18. marts 2011. Forsyningssekretariatet har modtaget høringssvar fra DANVA og FVD samt fra en række forsyninger.

Høringssvarene er indarbejdet i bilaget, hvor det har været relevant. Derudover er Forsyningssekretariatets bemærkninger til høringssvarene vedlagt i bilag 5 og 6.

Flere forsyninger har afgivet høringssvar som relaterer sig til individuelle forhold i forsyningen og ikke til den generelle ækvivalentberegning. Disse høringssvar har Forsyningssekretariatet ikke kommenteret. I stedet vil de blive indarbejdet i forbindelse med forsyningernes individuelle udkast til afgørelser.

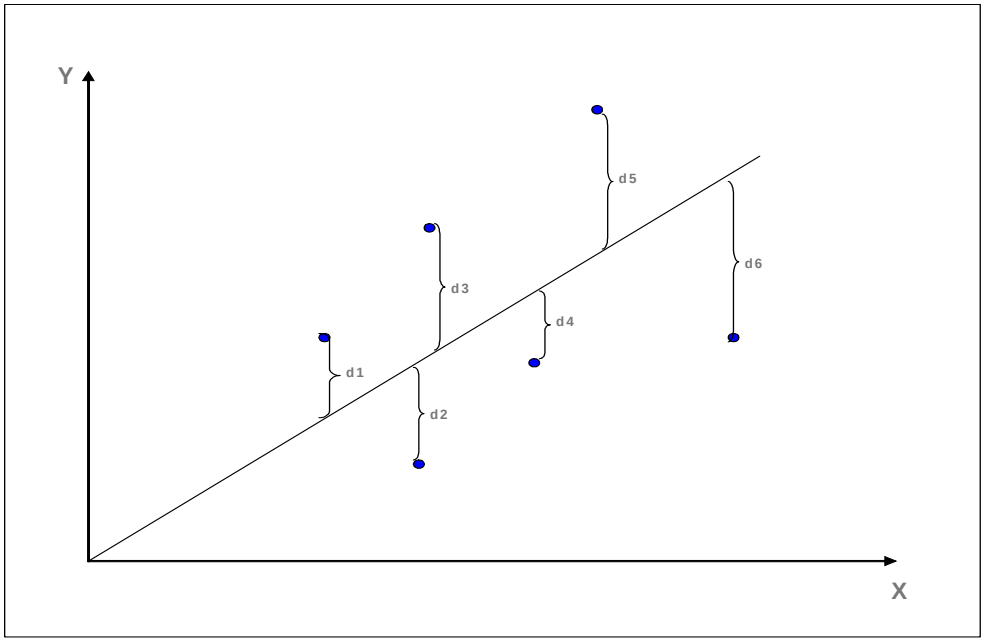
Mindste kvadraters metode

Forsyningssekretariatet har benyttet regressionsanalyse til at beregne omkostningsækvivalenterne for de costdrivere, som er blevet defineret for vand- og spildevandsforsyningerne.

For at sikre de mest retvisende resultater af regressionsanalysen har det været nødvendigt, at foretage en kvalitetssikring af det data Forsyningssekretariatet har modtaget. Det betyder, at Forsyningssekretariatet har fjernet forsyninger helt fra datasættet, hvis der har været store mangler i den enkelte indberetning.

I regressionsanalysen bliver der benyttet mindste kvadraters metode. Metoden består i at fastlægge den linje, der minimerer summen af kvadratet på den lodrette afstand til linjen jf. figur 1.

Figur 1: Mindste kvadraters metode



Figur 1 illustrerer seks forsyningers indberetning af omkostninger og en tilhørende underliggende faktor, f.eks. udpumpet vandmængde. De er illusteret med de runde punkter. Mindste kvadraters metode placerer linjen, hvor summen af de kvadrerede afstande angivet med d1-d6 er så lille som muligt. Dette bevirker, at jo mere en observation adskiller sig fra resten af observationerne, jo mindre vægtning får denne observation. Mindste kvadraters metode giver således et mere sandsynligt bud på den korrekte sammenhæng i data, end et simpelt gennemsnit ville gøre.

Mindste kvadraters metode sikrer således, at modellen angiver den sammenhæng, der er den mest sandsynlige, mellem responsvariablen (omkostningerne), angivet med et Y, og de forklarende variable (f.eks. løftehøjde eller kilometer byledning), angivet med et X.

En forudsætning for at benytte mindste kvadraters metode er, at modellen er lineær, og at de forskelle - kaldet fejlled - der er mellem modellens forudsagte værdier af omkostningerne og den enkelte observations faktiske værdi, er normalfordelte. At fejlleddene er normalfordelte kræver, at de har ens spredning og en middelværdi på 0.

Spredningen af en gruppe observationer, f.eks. forsyningernes indberetning af omkostninger forbundet med en given costdriver, udtrykker, hvor stor forskel der er på observationerne.

Den generelle model, som Forsyningssekretariatet vil opstille for hver enkel omkostningsækvivalent, bliver af formen:

$$(1) \quad Y = B_0 + B_1X_1 + ... + B_NX_N$$

Med mindste kvadraters metode estimeres således den værdi af B'erne, der angiver den mest sandsynlige sammenhæng mellem Y og samtlige X'er.

De enkelte B–værdier angiver hældningen af den tendenslinje, der fastsættes. Det vil sige, en given B-værdi angiver ændringen i omkostningerne (Y) ved en ændring på 1 i den forklarende variabel (X).

Sammenhæng mellem omkostninger og costdriver

For at være sikker på at modellen kan bruges, kontrolleres der for om der er en god sammenhæng mellem omkostningerne og de underliggende forhold for hver costdriver, f.eks. løftehøjde og vandmængde for boringer. Der kontrolleres således for, om der er signifikans i modellen. Signifikans i en regressionsanalyse betyder, at det ikke kan afvises, at der er en sammenhæng mellem variablen (Y) og de pågældende forklarende variable (X).

Forsyningssekretariatet har valgt at tage udgangspunkt i et 5 pct. signifikansniveau. Det betyder, at hvis sandsynligheden for at et B = 0 er større end 5 pct., vil det blive afvist, at der er en sammenhæng mellem den tilknyttede forklarende variable (X) og responsvariablen (Y). Det vil sige, i de tilfælde, afvises det, at den forklarende variabel er med til at beskrive de samlede omkostninger forbundet med en costdriver.

Signifikansen beregnes på baggrund af den fundne værdi af et B, og den spredning der er tilknyttet denne variabel. Hvis spredningen er høj i forhold til størrelsen af B, svækker det signifikansen.

Signifikansen af en variabel angives med både en t-værdi og en p-værdi. Både t-værdien og p-værdien fortæller, hvad sandsynligheden er for at B=0. Selve t-værdien skal ligge udenfor intervallet -/+ 1,96 for, at det kan afvises at B=0 og t-værdien kan desuden omregnes til en p-værdi. Denne p-værdi angiver sandsynligheden for at B=0 og skal således være mindre end 0,05 (5 pct.) for, at det kan accepteres, at den pågældende variabel indgår i modellen.

Værdierne fremgår af resultattabellerne i gennemgangen af hver enkel omkostningsækvivalent.

Normalfordeling af fejlled

Mindste kvadraters metode forudsætter, at fejlledene er normalfordelte med en middelværdi på nul.

Det kan kontrolleres om fejlleddene er normalfordelte ved at plotte fejlleddene mod de observerede værdier af de enkelte X'er. I praksis benyttes et mål, der kaldes de standardiserede fejlled, som i sidste ende

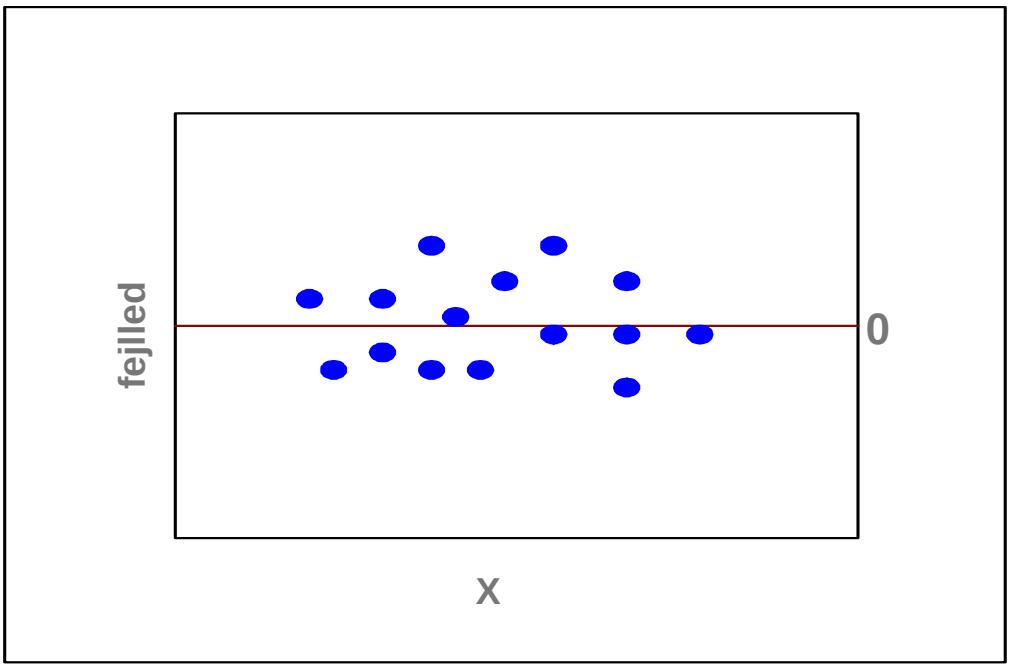
udtrykker det samme som de faktiske fejlled, men derudover har et par praktiske fordele med hensyn til kontrol af spredningen.

For at betingelsen om normalfordelte fejlled er opfyldt, skal fejlledsplottene vise en tilfældig fordeling omkring 0 jf. figur 2. Yderligere skal 95 % af observationernes standardiserede fejlled ligge inden for spændet [-2;2].

Hvis antagelsen om normalfordeling af fejlleddene ikke er opfyldt, kan der opstå problemer med, at modellen ikke er robust og dermed, at beregningen af modellens signifikans og B-værdier er mere usikker.

Det undersøges om betingelsen er opfyldt for hver af de beregnede omkostningsækvivalenter.

Figur 2: Normalfordelte fejlled



Heteroskedasticitet

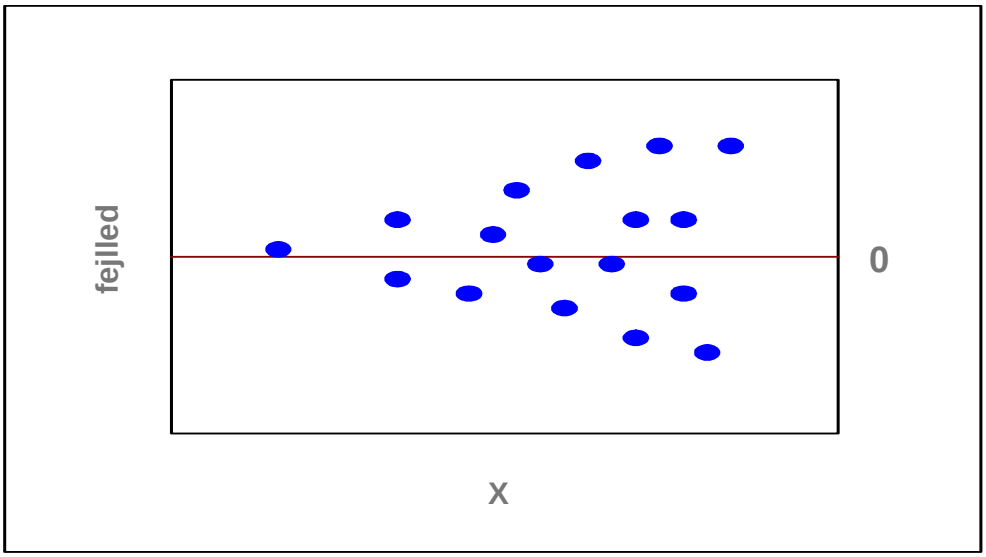
Heteroskedasticitet betyder, at spredningen af fejlleddene ikke er ens og dermed, at normalfordelingsbetingelsen ikke er fuldstændig opfyldt. I et datasæt med store forskelle i observationernes størrelse kan der lettere opstå heteroskedasticitet. Det skyldes, at de store observationer alt andet lige kan variere mere fra tendenslinien på grund af deres størrelse.

Heteroskedasticitet bliver ofte afbilledet i en trompetform jf. figur 3, men der kan også være andre typer af heteroskadasticitet. Derfor skal man være opmærksom på alle afvigelser fra en tilfældig fordeling af fejlleddene.

Hvis antagelsen om ens spredning af fejlleddene ikke er opfyldt, kan der opstå problemer med, at modellen ikke er robust og dermed, at beregningen

af modellens signifikans og B-værdier er mere usikker. Det er dog muligt at korrigere for dette ved at beregne en korrigeret spredning og t-værdi. Den korrigerede og mere robuste spredning samt t-værdi tager hensyn til den øgede usikkerhed, som er en konsekvens af heteroskedasticiteten.

Figur 3: Heteroskadasticitet



Ved beregningen af hver enkel omkostningsækvivalent bliver der kontrolleret for om betingelserne vedrørende normalfordeling af fejlled og heteroskedasticitet er opfyldt for modellerne.

Multikollinaritet

Der kan opstå yderligere problemer, hvis de X'er, der indgår i modellen, er korrelerede. F.eks. er X_1 og X_2 korrelerede hvis X_2 stiger når X_1 stiger. Det betyder, at X_1 og X_2 beskriver den samme variation for Y. Betegnelsen for dette problem er multikollinaritet, og det kan give forkerte beregninger af B-værdierne og have betydning for beregningen af signifikans.

Det kan være svært at vurdere, hvornår multikollinaritet kan være et problem. Umiddelbart betegnes en korrelation på over 0,7 mellem to parametre for høj, og Forsyningssekretariatet ligger sig derfor i udgangspunktet op af dette niveau.

Ved beregningen af hver enkel omkostningsækvivalent bliver der kontrolleret for om betingelserne vedrørende multikollinaritet er opfyldt for modellerne.

Ekstreme observationer og observationer med stor indflydelse

Enkelte observationer i et datasæt kan have en uhensigtsmæssig stor påvirkning af bestemmelsen af B-værdierne. Det er derfor nødvendigt at overveje, hvordan sådanne observationer skal behandles.

Forsyningssekretariatet har vurderet, at Cook's afstand er et passende mål til at identificere observationer med stor indflydelse på analysen. Cook's afstand siger noget om, hvor meget B-værdierne ændres, hvis en given observation fjernes fra analysen. Det vil sige, hver eneste observation får beregnet en Cook's afstand.

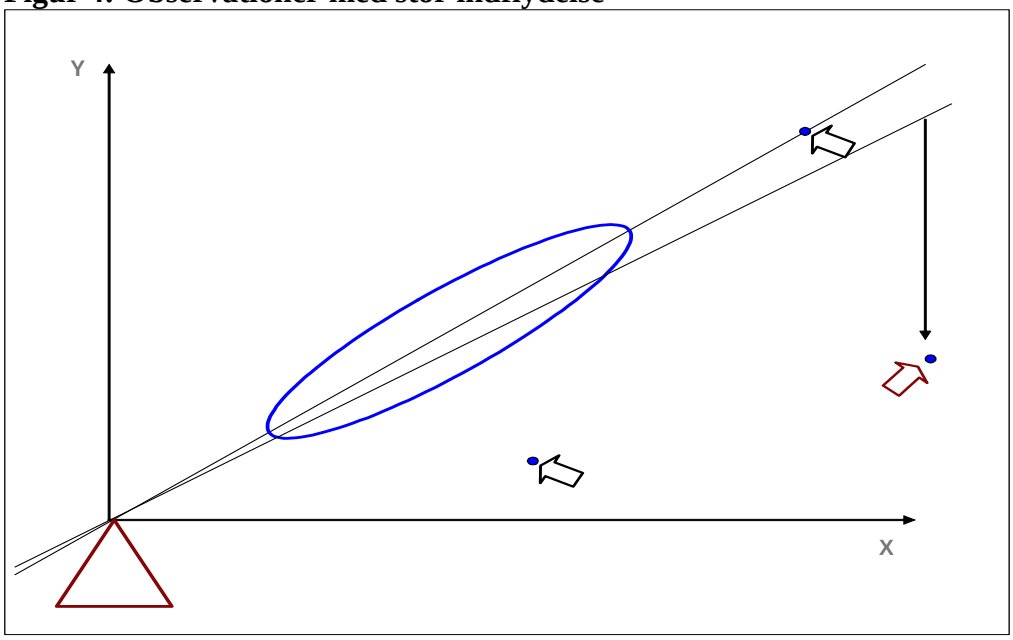
Der er ikke nogen faste statistiske regler for, hvornår en observations Cook's afstand er så stor, at observationen bør fjernes fra datasættet. Forsyningssekretariatet anvender en metode, der forslår, at observationer som falder udenfor 50 pct.-fraktilen i en F-fordeling med (p,n-p)[1] frihedsgrader bør fjernes fra datasættet. Dette svarer til en Cook's afstand på ca. 0,5. Denne metode er i overensstemmelse med statistisk praksis og anvendes også som standard i det statistikprogram, der benyttes til regressionskørslerne[2].

Figur 4 illustrerer, hvordan ekstreme observationer kan have indflydelse på modellens resultater. De ekstreme observationer er i figuren de observationer, der ikke ligger inden i den blå oval. Figuren illustrerer, hvordan punktet ved den højrevendte pil kan påvirke linjen uhensigtsmæssigt meget, da den vil trække tendenslinien ned mod sig. Det skyldes, at en observation kan påvirke regressionsanalysen på samme måde som en vægtstang og dermed få stor betydning for resultaterne af regressionsanalysen.

Samtidig kan der også være tilfældet at andre punkter, der måske skiller sig en del ud fra mængden, ikke har særlig stor indflydelse på resultaterne, jf. de to punkter i figur 4 med venstrevendte pile. I forhold til vægtstangen i nulpunktet, ved den røde trekant, vil de ikke påvirke tendenslinien i nævneværdig grad, og har derfor ikke nogen væsentlig betydning for analysens resultater.

[1] Her er n antal observationer og p er antal parametre.
[2] Jf. standardbetingelser for Cook's D i statistikprogrammet "R", samt Gross, Jürgen, Linear Regression, Springer 2003.

Figur 4: Observationer med stor indflydelse



Ved beregningen af hver enkel omkostningsækvivalent kontrolleres der for, om der er ekstreme observationer, der har en uhensigtsmæssig indflydelse på resultaterne, og derfor bør fjernes fra datasættet.

Transformationer

Hvis det viser sig, at data ikke er lineært, kan man løse problemet ved hjælp af forskellige transformationer af data. De mest udbredte transformationer er at tage logaritmen, kvadratroden eller andre potenser til variablene i ligningen.

Problemet med andre sammenhænge i data end lineære er, at mindste kvadraters metode kun kan bruges på et lineært datasæt. Derfor er det nødvendigt at vælge en transformation, der skaber en lineær sammenhæng i forhold til den forventede sammenhæng i det oprindelige datasæt.

Et eksempel kan være, at det forventes, at en produktion kan beskrives ved en klassisk Cobb-Douglas funktion af formen:

$$(2) \quad Y = B_0 X_1^{B1} X_2^{B2}$$

I dette tilfælde er det nødvendigt at lave en såkaldt "log-log"-transformation. Det vil sige, at logaritmen tages på begge sider af lighedstegnet. Det giver følgende lineære form:

$$(3) \quad \log Y = \log B_0 + B_1 \log X_1 + B_2 \log X_2$$

I denne form er det muligt at estimere B-værdierne ved hjælp af mindste kvadraters metode.

Årsagen til at den logaritmiske transformation er populær skyldes, at valget af omkostningsfunktion bliver bestemt af data. I den logaritmiske model vil størrelsen af B-værdien afgøre, hvorvidt der er tale om stordriftsfordele eller -ulemper.

For at sikre at "log-log"-transformationen er den rigtige kan man kontrollere fejlledsplottene. Det vil sige, modelkontrollen afslører, hvorvidt der er tale om log-normaltfordelt data. Dette kan ses ved at fejlledene i den transformerede model er normalfordelte.

Da hele formålet med beregningen af omkostningsækvivalenterne drejer sig om at bestemme forholdet mellem Y og X'erne, er det nødvendigt at tilbagetransformere ligning (3). Tilbagetransformeringen er ikke så enkel som det umiddelbart kunne se ud til. Det skyldes, at hvis "log-log"-transformationen giver en normalfordeling af fejlleddene, så kan det ikke være tilfældet i parametrenes oprindelige form. Det giver en skævhed, når den estimerede model tilbagetransformeres, og det er derfor nødvendigt at benytte en korrektionsfaktor, der justerer for dette. Korrektionsfaktoren kan bestemmes til[3]:

$$(4) \quad KF = \exp(\sigma^2/2)$$

Her betegner "exp" den naturlige eksponential funktion og σ er spredningen af fejlleddet.

Valg af den bedste model

Forsyningssekretariatet har lagt sig op af den statistiske værdi, der betegnes "R^2" til at bestemme den model, der beskriver data bedst. R^2 er et udtryk for, hvor godt en given model beskriver det data, der er lagt til grund for beregningen af modellen.

Værdien af R^2 ligger mellem 0 og 1, hvor 1 angiver en perfekt beskrivelse af modellen og 0 det modsatte. R^2 kan ikke entydigt bruges som et mål for, hvorvidt en model er god eller dårlig. Derfor er det stadig nødvendigt, at foretage en generel vurdering i fastlæggelsen af den funktionelle form i de konkrete tilfælde.

Forsyningssekretariatet har også vægtet det højt at medtage så stor en andel af observationerne som muligt i grundlaget for modellerne.

[3]Hvis man i stedet foretager en transformation af data således at variablene Y og X f.eks. er opløftet i en potens skal der foretages en anden korrektion: $KF = \sigma^2$. Kilde: Don M. Miller, *Reducing Transformation Bias in Curve Fitting*.

Endelig har Forsyningssekretariatet også lagt vægt på, at vælge den funktionelle form der passer bedst til data.

Datakvalitet

DANVA og FVD har anført, at forsyningernes kontoplaner ikke har været indrettet til at fordele data på det ønskede detaljeringsniveau, hvilket har medført at mange forsyninger har været nødsaget til at fordele en meget stor del af deres omkostninger ud fra skøn. Det bekymrer DANVA og FVD, at forsyningerne har brugt forskellige skøn til at fordele de omkostninger som ikke var direkte henførbare til en given costdriver.

Overordnet set, er det Forsyningssekretariatets vurdering, at forsyningerne selv bedst ved, hvordan den skønsmæssige fordeling af de øvrige omkostninger i den enkelte forsyning bør foretages. Derfor er det ikke umiddelbart til skade for datakvaliteten at skønnene er fordelt på baggrund af forskellige principper. Der kan naturligvis være fejlkilder i et datasæt som dette, f.eks. tastefejl og fejlskøn. Et af formålene med at Forsyningssekretariatet har benyttet regressionsanalysen er for at kunne tage hensyn til den slags datafejl, under en forholdsvis ustringent antagelse om at disse fejl er tilnærmelsesvis normalfordelte omkring 0, hvorved det i stor udstrækning ikke har nogen betydning for analysens resultater. Derudover indebærer regressionsanalysen, at observationer, der adskiller sig meget fra resten af observationerne, får mindre vægt i beregningen af B-værdierne.

Forsyningssekretariatet har foretaget en række følsomhedsanalyser af omkostningsækvivalenterne for at teste deres robusthed og herunder også datakvaliteten.

Endelig overvejer Forsyningssekretariatet, på opfordringer fra branchen, at revidere ækvivalentberegningerne næste år gældende for benchmarkberegningen der skal indgå i prisloftet 2013.

DANVA og FVD kommenterer ligeledes i deres høringssvar i relation til datakvalitet, at det har skabt uklarhed om datakvaliteten at definitionen af drift og vedligehold kontra investeringer ikke stemmer overens med princippet i forsyningernes årsregnskaber.

Den usikkerhed der eventuelt vil være i datagrundlaget som følge af dette bliver der ligeledes taget højde for ved hjælp af de følsomhedsanalyser, som Forsyningssekretariatet har foretaget på resultaterne, jf. afsnittet vedr. følsomhedsanalyse i papiret ”Resultatorienteret benchmarking af vand- og spildevandforsyningerne”.

Bestemmelse af omkostningsækvivalenter for vandforsyninger

Forsyningssekretariatet har beregnet omkostningsækvivalenter for de 6 costdrivere:

- Boringer
- Vandværker
- Trykforøgere
- Rentvandsledning
- Stik
- Kunder

Nedenfor vil beregningen af de enkelte omkostningsækvivalenter blive beskrevet.

Boringer

Omkostningsækvivalenten for boringer forventes at afhænge af de tre variable: antallet af boringer, den samlede løftehøjde for alle boringer samt den samlede oppumpede vandmængde fra samtlige boringer en forsyning har.

En indledende kontrol viser, at der er en høj korrelation mellem antallet af boringer og den oppumpede vandmængde på ca. 0,97. Det betyder, at en stor del (ca. 97 pct.) af sammenhængen, mellem omkostninger og henholdsvis antal boringer og den oppumpede vandmængde, er identisk. Forsyningssekretariatet vurderer derfor, at det ikke er relevant at benytte en model, der både inkluderer antal boringer og den oppumpede vandmængde.

Forsyningssekretariatet vurderer desuden, at den oppumpede vandmængde er det mest intuitive og præcise mål til at beskrive driftsomkostningerne forbundet med boringer, og har derfor valgt at bruge denne variabel i modellen.

Det er således muligt at opstille en model for omkostningsækvivalenten for boringer. Forsyningssekretariatet har en forventning om, at omkostningsækvivalenten for boringer kan opstilles som en omkostningsfunktion af formen:

$$(5) \quad Y = B_0 X_1^{B1} X_2^{B2}$$

Hvor Y er omkostningerne forbundet med boringerne, X_1 er løftehøjde og X_2 er den oppumpede vandmængde. B_0, B_1 og B_2 vil blive bestemt ved hjælp af mindste kvadraters metode som beskrevet i afsnittet "*Mindste kvadraters metode*" ovenfor. Det vil sige, at der benyttes et matematisk

skøn, som maksimerer sandsynligheden for, at det er de rigtige værdier af B'erne, der bliver fundet i modellen. Forsyningernes indberetninger benyttes til dette.

Formen på den valgte omkostningsfunktion i ligning (5) kaldes en Cobb Douglas form. Intuitionen bag denne form er, at løftehøjden og den oppumpede mængde afhænger multiplikativt af hinanden. Det betyder, at den samlede løftehøjde har betydning for ændringen i driftsomkostningerne for boringer, når der pumpes en ekstra m^3 vand op af boringerne.

Transformation

For at benytte mindste kvadraters metode til at finde B_0, B_1 og B_2 skal den valgte form af omkostningsfunktionen i ligning (5) transformeres til en lineær form. Dette gøres ved at tage logaritmen til begge sider af ligning (5), hvilket giver den lineære sammenhæng udtrykt ved ligning (6), jf. afsnittet vedr. transformationer ovenfor.

$$(6) \quad \log Y = \log B_0 + B_1 \log X_1 + B_2 \log X_2$$

Resultater

Den første regressionsanalyse viser, at B_0(de faste omkostninger) ikke er signifikant. Det vil sige, at B_0 ikke er med til at beskrive omkostningerne forbundet med drift af boringer i væsentlig grad, og B_0 bør derfor sættes til nul. Det giver den endelige model:

$$(7) \quad \log Y = B_1 \log X_1 + B_2 \log X_2$$

De endelige resultater af regressionsanalysen viser, at der er god signifikans for både løftehøjden og den oppumpede vandmængde. Yderligere findes B_1 = 0,19539 og B_2 = 0,86155 jf. tabel 1.

Tabel 1: Regressionsanalysens resultater, Boringer

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Log(løftehøjde)	0,19539	0,06383	3.061	0,00260
Log(oppumpet vandmængde)	0,86155	0,02592	33.234	<2e-16
Antal observationer: 155 Justeret R^2= 0,99				

Kontrol af modellen

Det skal kontrolleres, om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen.

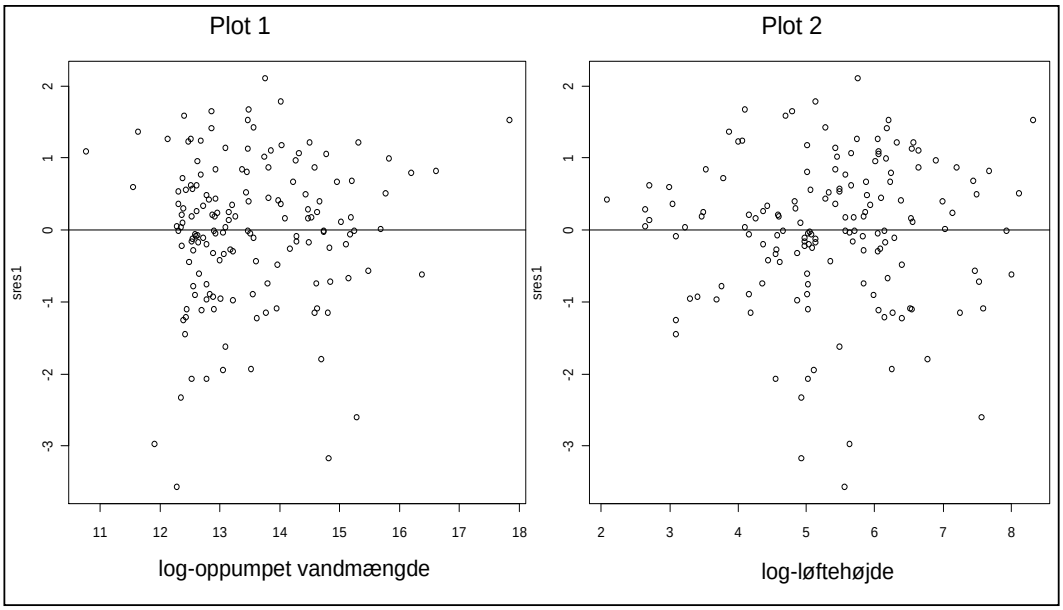
Betingelserne kontrolleres for begge variable: løftehøjde og oppumpet vandmængde. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af henholdsvis løftehøjde og oppumpet vandmængde.

I figur 5 nedenfor viser plot 2 de standardiserede fejlled plottet imod de observerede værdier af løftehøjden. Plottet viser en umiddelbar tilfældig fordeling af fejlleddene omkring 0, hvilket betyder at normalfordelingsbetingelsen er opfyldt.

Yderligere lader der ikke til at være tegn på heteroskedaticitet. Der er nogle få observationer, der falder uden for intervallet [-2;2], men det er umiddelbart under 5 % af observationerne, og dermed hvad der kan forventes, jf. plot 2 i figur 5.

Plot 1 i figur 5 viser stort set det samme billede som plot 2 blot for den oppumpede vandmængde. Det vil sige, betingelserne vedrørende normalfordelte fejlled og heteroskedasticitet vurderes at være opfyldt for modellen.

Figur 5: Fejlledsplots for boringer



Der er ikke umiddelbart nogen grund til at forvente multikollinaritet i modellen. Korrelationen mellem løftehøjden og den oppumpede mængde kontrolleres dog alligevel og giver en korrelation på 0,56 hvilket er acceptabelt.

Der er ikke nogen enkelte observationer, der har stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for boring

Den endelige model for omkostningsækvivalenten er således beskrevet ved den oprindelige foreslåede model i ligning (5). Som nævnt i det indledende kapitel, er det nødvendigt at foretage en afsluttende korrektion når modellen tilbagetransformeres. Den endelig omkostningsækvivalent for boringer udgøres dermed af løftehøjden af boringerne, den oppumpede vandmængde og en ekstra korrektionsfaktor.

Korrektionsfaktoren beregnes som KF = $\exp(\sigma^2/2)$ = 1,42786. Det vil sige, den endelige model kan derefter opstilles:

$$(8) \quad Y = 1{,}428X_1^{0{,}195}X_2^{0{,}862}$$

Forsyningerne får således forklaret omkostningerne i forbindelse med boringer ud fra, hvor højt de løfter vandet (i meter), og hvor mange m^3 vand de pumper op af deres boringer.

Vandværk

Omkostningerne forbundet med vandværk forventes at kunne afhænge både af vandværkernes samlede kapacitet samt den samlede udpumpede vandmængde. En indledende kontrol har dog vist, at der er en høj korrelation mellem kapacitet og udpumpet vandmængde på 0,97. Det er derfor ikke relevant at benytte en model, der både inkluderer kapacitet og udpumpet vandmængde. Forsyningssekretariatet har vurderet, at den udpumpede vandmængde er den variabel, som beskriver omkostningerne mest præcist.

Det er således muligt at opstille en model for omkostningsækvivalenten for boringer. Forsyningssekretariatet har en forventning om, at omkostningsækvivalenten for vandværker kan opstilles som en omkostningsfunktion af formen:

$$(9) \quad Y = B_0X_1^{B1}$$

Hvor Y er driftsomkostningerne forbundet med vandværker og X_1 er den udpumpede vandmængde angivet i m^3.

For at benytte mindste kvadraters metode til at finde B_0 og B_1, skal den valgte form af omkostningsfunktionen i ligning (9) transformeres til en lineær form. Dette gøres ved at tage logaritmen til begge sider af ligning (9), hvilket giver den lineære sammenhæng udtrykt ved ligning (10), jf. afsnittet vedr. transformationer ovenfor.

Den logaritmiske transformation har den fordel, at modellen viser, om der er stigende eller aftagende skalaafkast i data, givet den forudsætning at data

faktisk er logaritmisk normalfordelt. Det vil sige, at den transformerede model skal opfylde de ovenfor nævnte betingelser om normalfordeling og signifikans.

Den foreslåede model har derfor formen:

$$(10) \quad \log Y = \log B_0 + B_1 \log X_1$$

Resultater

Den første regressionsanalyse viser, at B_0 ikke er signifikant. Det vil sige, at B_0 ikke er med til at beskrive omkostningerne forbundet med drift af vandværker i væsentlig grad. B_0 bør derfor sættes til nul. Den endelige model bliver derfor:

$$(11) \quad \log Y = B_1 \log X_1$$

De endelige resultater af regressionsanalysen viser, at der er god signifikans i modellen, hvilket betyder, at det ikke kan afvises, at den udpumpede vandmængde har betydning for omkostningerne ved at drive vandværket. Yderligere findes en B_1-værdi på 1,02791 jf. tabel 2.

Tabel 2: Regressionsanalysens resultater, Vandværker

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Log (udpumpet mængde)	1,02791	0,00405(0,06261)	253,8(16,4)	<2e-16(<0,01)
Antal observationer: 159 Justeret R^2= 0,9975				

Kontrol af model

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen.

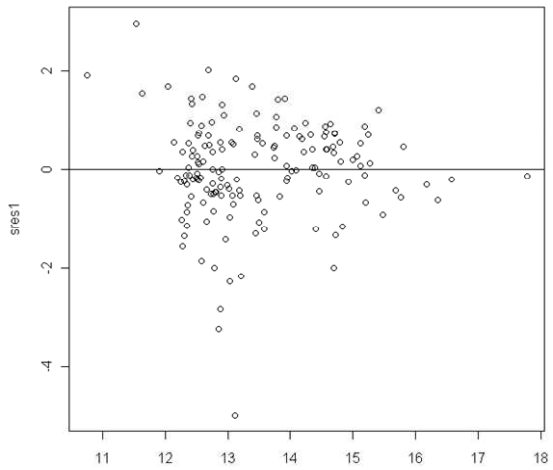
Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af den samlede udpumpede vandmængde.

Dette viser umiddelbart en tilfældig fordeling omkring 0 jf. figur 6. Ifølge plottet kan der være en smule heteroskedasticitet i modellen. For at tage hensyn til dette og dermed sikre, at signifikansen i modellen er bevaret beregnes den robuste spredning samt t- og p-værdi.

Den robuste spredning beregnes til 0,06261. Den robuste spredning samt tilhørende t-værdier og p-værdier er angivet i parenteser i tabel 2. På baggrund af de robuste resultater er konklusionen, idet p-værdien stadig er

under 0,01, at heteroskedasticiteten ikke påvirker signifikansen i modellen i nogen betydende grad.

Figur 6: Fejlledsplot for vandværker



Der er ikke nogen enkelte observationer, der har stor indflydelse på resultatet. Dette måles med Cook's afstand som beskrevet i afsnittet "*Ekstreme observationer og observationer med stor indflydelse***"** ovenfor.

Endelig omkostningsækvivalent for vandværk

Den endelige model for omkostningsækvivalenten er således beskrevet af den foreslåede model i ligning (9).

Som nævnt i det indledende kapitel og som det også var tilfældet vedrørende boringer, er det nødvendigt at foretage en afsluttende korrektion når modellen tilbagetransformeres:

$$(12) \quad Y = KF^{*}X_1^{B1}$$

Hvis B_1 er større end 1, betyder det, at modellen udviser et aftagende skalaafkast forbundet med at drive vandværk. Det vil sige, det bliver dyrere pr. m^3 vand jo mere vand forsyningen samlet set pumper ud af deres vandværker. Den endelige omkostningsækvivalent bliver:

$$(13) \quad Y=1{,}27X_1^{1{,}028}$$

Forsyningerne får således forklaret driftsomkostningerne i forbindelse med vandværker ud fra, hvor mange m^3 vand de pumper ud fra deres boringer. Da B_1 er større end 1 viser data, at der er stordriftsulemper forbundet med

driften af vandværker. Det vil sige, at omkostningerne pr. m^3 vand stiger jo flere m^3 forsyningerne udpumper fra deres vandværker.

DANVA og FVD har i deres høringssvar bemærket, at stordriftsulemper ved store vandværker ikke er intuitiv, men kan skyldes at de store forsyninger ofte har flere kilometer råvandsledning, mere overkapacitet og bedre servicemål på drikkevand- og forsyningssikkerhed. Modellen tager således hensyn til, at disse forhold ikke direkte indgår i modellen for vandværker.

Trykforøgere

Omkostningerne i forbindelse med de forskellige kategorier af trykforøgere forventes at afhænge lineært af antallet af trykforøgere i de forskellige kategorier. Den første model opstilles derfor som de samlede omkostninger til trykforøgere beskrevet ved summen af antallet af trykforøgere i de fem angivne kategorier.

$$(14) \quad Y = B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5$$

X_1 angiver antallet af trykforøgere i intervallet 0-50 m^3/t, X_2 er trykforøgere i intervallet 50-100 m^3/t, X_3 er trykforøgere i intervallet 100-200 m^3/t, X_4 er antallet af trykforøgere i intervallet 200-600 m^3/t og X_5 er trykforøgere i intervallet 600-max m^3/t.

Resultater

Den første regressionsanalyse viser et par problemer. Det vil sige, at estimatet for kategorien 50-100 m^3/t (B_2) bliver negativ jf. tabel 3. Dette giver ikke intuitiv mening, da omkostningerne til at drive og vedligeholde en trykforøger ikke kan være negativ. Yderligere er denne variabel ikke signifikant, jf. en p-værdi på 0,529. Derfor ligges kategorierne 0-50 m^3/t og 50-100 m^3/t i stedet sammen til én kategori: 0-100 m^3/t.

Tabel 3: Første regressionskørsel, Trykforøgere

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
0-50 m^3/t	32.547	10.840	3,003	0,003279
50-100 m^3/t	-22.256	35.252	-0,631	0,529066
100-200 m^3/t	255.639	43.588	5,865	4,33e-08
200-600 m^3/t	130.557	50.517	2,584	0,010994
600-max m^3/t	371.110	108.944	3,406	0,000905
Antal observationer: 120 Justeret R^2= 0,59				

Den efterfølgende regressionskørsel viser, at der stadig er problemer i modellen, da B_4 er mindre end B_3. Det betyder, at det er dyrere at drive og vedligeholde en mindre pumpe (100-200 m^3/t) end en større pumpe (200-600 m^3/t), jf. tabel 4. Dette skyldes sandsynligvis, at antallet af observationer i de enkelte kategorier er relativt begrænset. Forsyningssekretariatet vurderer ikke, at de pågældende B-værdier giver intuitiv mening, da elforbruget, alt andet lige, er højere jo flere m^3 vand pumpen trykker igennem. Begge estimater for de to kategorier er dog signifikante.

De to kategorier (100-200 m^3/t) og (200-600 m^3/t) lægges i stedet sammen til én og regressionsanalysen foretages igen. Resultaterne af denne model kan ses i tabel 5.

Tabel 4: Anden regressionskørsel, Trykforøgere

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
0-100 m^3/t	21.564	6.503	3.316	0.00122
100-200 m^3/t	243.711	42.664	5.712	8.62e-08
200-600 m^3/t	125.930	50.513	2.493	0.01406
600-max m^3/t	359.275	108.819	3.302	0.00128
Antal observationer: 120 Justeret R^2= 0,59				

Tabel 5: Regressionsanalysens resultater, Trykforøgere

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
0-100 m^3/t	53.204	6.800	7,824	2,96e-12
100-600 m^3/t	125.224	23.676	5,289	6,08e-07
600-max m^3/t	411.776	86.774	4,745	6.14e-06
Antal observationer: 116 Justeret R^2= 0,61				

Ingen af parametrene er længere insignifikante og sammenhængen mellem omkostningerne på trykforøgerkategorierne (B-værdierne) er nu også intuitiv og derfor fortsættes med yderligere kontrol af denne model.

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved, at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

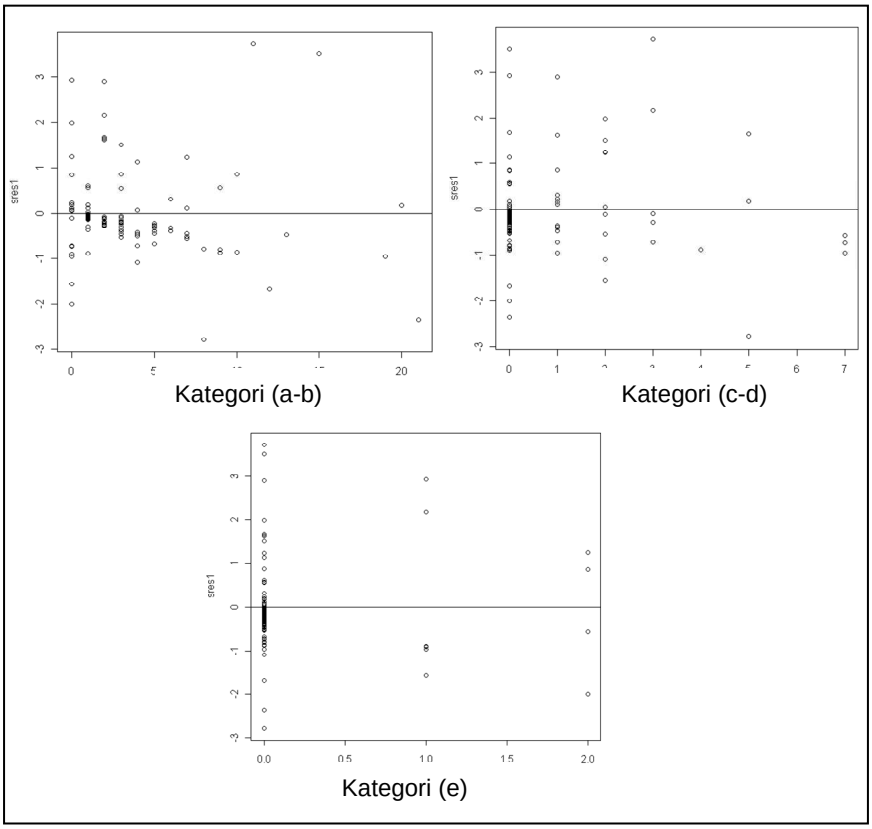
Kontrol af model

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen.

Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af de tre kategorier af trykforøgere.

Der er ikke nogen umiddelbare problemer i plottene, bortset fra, at der er relativt få observationer især i den sidste kategori, hvilket kan medføre, at modellens resultater mister noget robusthed.

Figur 7: Fejlledsplot for trykforøgere



Det har været nødvendigt at fjerne enkelte observationer med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for trykforøgere

Den endelige model til at beskrive omkostningerne forbundet med at drive og vedligeholde trykforøgere reduceres til at indeholde tre kategorier jf. tabel 5 og har følgende udseende:

$$(15)\quad Y = 53.204(X_1 + X_2) + 125.224(X_3 + X_4) + 411.776X_5$$

Dermed angiver modellen, at en trykforøger i intervallet 0-100m^3/t koster 53.204 kr., en trykforøger i intervallet 100-600m^3/t koster 125.224 kr. at drive og en trykforøger i kategorien 600-max m^3/t en koster 411.776 kr. at drive.

Rentvandsledning

Omkostningerne i forbindelse med drift af rentvandsledning forventes i udgangspunktet både at kunne afhænge af ledningernes længde (meter) samt deres volumen (m^3) fordelt på de fire zoner land, by, city og indre city.

Forsyningssekretariatet har imidlertid vurderet, at meter ledning er den variabel, som beskriver omkostningerne mest præcist, da det er fremgået af indberetningerne, at der har været større usikkerhed omkring volumen af forsyningernes ledningsnet og der derfor er foretaget flere skøn, som er forskellige i forsyningernes opgørelser.

Forsyningssekretariatet har derfor besluttet at benytte meter ledning som beskrivende variabel for driftsomkostningerne forbundet med ledning. Det fremgår også af resultatet af den valgte model nedenfor, at R^2 er høj og dermed, at data er godt beskrevet af modellen.

Forsyningssekretariatet forventer, at driftsomkostningerne forbundet med ledning afhænger lineært af længden af rentvandsledning i de forskellige zoner. Den første model opstilles derfor som de samlede omkostninger til rentvandsledning beskrevet af summen af længden af rentvandsledning i de fire angivne zoner.

$$(16) \quad Y = B_1X_1+B_2X_2+B_3X_3+B_4X_4$$

Y angiver de samlede driftsomkostninger forbundet med ledning, X_1 angiver antal meter ledning i landzone, X_2 angiver antal meter ledning i byzone, X_3 angiver antal meter ledning i cityzone og X_4 angiver antal meter ledning i indre cityzone.

Resultater

De indledende regressionsanalyser gør det klart, at det er nødvendigt at slå zonerne land og by sammen til én kategori og yderligere at slå zonerne indre city og city sammen. Det skyldes, at landzonen og indre cityzonen bliver insignifikante efter der er fjernet outliere, det vil sige, p-værdien er større end 0,05, jf. Tabel 6.

Tabel 6: Første regressionskørsel, Rentvandsledning

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Land	-0,6097	0,9033	-0,675	0,501
By	13,2525	1,2105	10,948	< 2e-16
City	52,5318	7,9811	6,582	6,88e-10
Indre city	159,4753	17263,6571	0,092	0,927
Antal observationer: 155 Justeret R^2= 0,82				

Det betyder, at der opstilles en ny model, hvor de pågældende kategorier er slået sammen:

$$(17) \quad Y = B_1(X_1+X_2)+B_2(X_3+X_4)$$

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

Tabel 7: Regressionsanalysens resultater, Rentvandsledning

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Land + By	6,0435	0,4453(0,75)	13,57(8,058)	<2e-16 (0,0001)
City + Indre city	52,3825	2,7095(2,92)	19,33(17,939)	<2e-16 (0,0001)
Antal observationer: 164 Justeret R^2= 0,86				

DANVA og FVD samt flere forsyninger har i deres høringssvar kommenteret på, at zonerne er reduceret til "kun 2", samt at springet i B-værdiens størrelse mellem Land+By-zonen og City+Indre City-zonen umiddelbart virker meget stort.

Forsyningssekretariatet har derfor gennemført en række yderlige regressionsanalyser for rentvandsledninger for at teste, hvor robuste B-værdiernes størrelse er for de forskellige zoner.

For at teste dette er datasættet splitttet op i forskellige dele, hvorefter regressionsanalysen er kørt på de forskellige datasæt.

Datasættet er således blevet opdelt i 3 nye datasæt:

- Forsyninger som kun har ledning i land- og/eller byzone
- Forsyninger som har ledning i cityzone, men ikke i indre city-zone (de fleste af disse forsyninger har også ledning i land og by)
- Forsyninger som har ledning i cityzone (dvs. inklusiv forsyninger som har ledning i både cityzone og indre city-zone er medtaget her)

Denne opsplitning er foretaget, for at kontrollere om resultaterne påvirkes af, hvilke zoner forsyningerne har ledning i. Herunder om der er en tendens til at være forskel i fordelingen af omkostninger på ledninger alt efter størrelsen af forsyningerne[4].

Resultatet af disse regressionskørsler fremgår af tabel 8 nedenfor. En streg angiver, at det ikke har været muligt at beregne en værdi for den pågældende faktor. Resultaterne i kolonnen "Alle forsyninger" fremgår også af tabel 7.

Tabel 8: Resultat (B-værdier) for regressionsanalyser for forskellige sammensætninger af datasættet for rentvandsledning.

Variabel	Forsyninger der kun har ledning i land og/eller by	Forsyninger der har ledning i city, men ikke i indre city	Forsyninger der har ledning i city	Alle forsyninger
Land + By	7,44	5,46	5,99	6,04
City	-	80,01	-	-
City + Indre City	-	-	52,52	52,38

Det fremgår af tabellen, at for forsyninger der kun har land og by findes en omkostningsækvivalent der er 1,4 kr. højere, end den der er angivet som gældende i Forsyningssekretariatets beregninger for "Alle forsyninger".

For kategorien "Forsyninger der har ledning i city" ses et næsten identisk billede af forholdet imellem omkostningerne forbundet med Land+By og City+Indre City som i kategorien "Alle forsyninger", hvilket peger på, at dette forhold er, hvad der gør sig gældende i branchen. Dog er parameteren for Land+By en smule højere i "Alle forsyninger"-kategorien, hvilket sandsynligvis skyldes, at forsyningerne i kategorien "Forsyninger der kun har ledning i land og/eller by" trækker en smule op for denne parameter og tilsvarende sænkes City+Indre City marginalt.

[4] De mindre selskaber har hovedsageligt kun ledning i land- og byzone, mens de større selskaber også har ledning i city- og indre city-zone.

I kategorien ” Forsyninger der har ledning i city, men ikke i indre city” ses et nogenlunde tilsvarende billede af Land+By-parameteren i forhold til kategorien ”Alle forsyninger”. Til gengæld sker der et noget kraftigere spring op til city-ledninger, hvilket igen bekræfter tendensen til, at der er stor forskel imellem omkostningerne i land og by og omkostningerne i city.

Alt i alt vurderer Forsyningssekretariatet ikke, at den relative store forskel mellem omkostningen på at drive ledning i Land+By-zonen og City+Indre City-zonen er misvisende og skyldes fejl i data eller manglende datakvalitet.

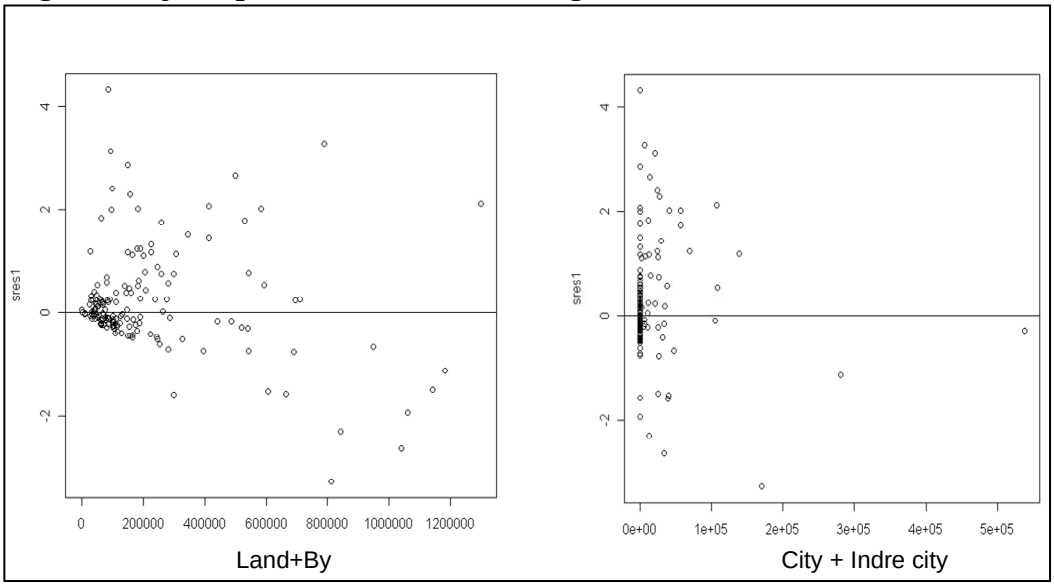
Forsyningssekretariatet konkluderer på baggrund af ovenstående, at resultatet i tabel 7 er retvisende og at koefficienternes størrelse hverken skyldes datamaterialets sammensætning eller eventuelle usikkerheder i den måde som virksomhederne har fordelt deres omkostninger på. Der fortsættes derfor med den generelle kontrol af modellen i ligning (17).

Kontrol af model

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af de to kategorier af ledning.

Fejlledsplottene i figur 8 nedenfor viser umiddelbart en tilfældig fordeling omkring 0. Der kan være en smule heteroskedasticitet i modellen, især for Land+By-zonen. For at tage hensyn til dette, og for at sikre at signifikansen i modellen er bevaret, beregnes de robuste spredninger samt t- og p-værdier.

Figur 8: Fejlledsplot for rentvandsledning



Den robuste spredning beregnes til 0,75 og 2,92 for hhv. Land+By og City+Indre City. De robuste spredninger samt tilhørende t-værdier og p-værdier er angivet i parenteser i tabel 7. På baggrund af de robuste resultater

er konklusionen, at heteroskedasticiteten ikke påvirker signifikansen i modellen i nogen betydende grad.

Multikollinaritet kontrolleres i modellen ved at se på korrelationen mellem de to parametre Land+By og City+Indre City. Korrelationen beregnes til 0,34, og dermed er der ikke nogen umiddelbare problemer med multikollinaritet i modellen.

Der er ikke nogen enkelte observationer, der har stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for rentvandsledning

Den endelige model til at beskrive omkostningerne forbundet med at drive og vedligeholde rentvandsledning reduceres til at indeholde to kategorier jf. tabel 7 og har følgende samlede endelige udseende.

$$(18)\quad Y = 6{,}043(X_1 + X_2) + 52{,}38(X_3 + X_4)$$

Det betyder, at for hver meter rentvandsledning i land- eller byzone er der forbundet en driftsomkostning på 6,04 kr. Ligeledes er der for hver meter rentvandsledning i city- eller indre cityzone forbundet en driftsomkostning på 52,38 kr.

Stik

Beregningen af omkostningsækvivalenten for stik kræver en mere kompliceret beregning end for de andre costdrivere for vandforsyninger. Det har været nødvendigt at tage forsyninger med stik i indre cityzonen ud af datasættet og i første omgang kun beregne driftsomkostninger forbundet med stik i zonerne land, by og city på baggrund af regressionsanalysen.

Dette har været nødvendigt, da variationen i omkostningerne for de forsyninger, som har indre cityzone, har adskilt sig meget fra de øvrige forsyninger. Det betyder, at de alle har haft en høj Cook's D afstand, og derfor er blevet identificeret som outliere i modellen. Det vil sige, ingen af forsyningerne med indre city-zone kunne indgå i den samlede regressionsanalyse. Den følgende model opstilles derfor til at beskrive omkostningerne forbundet med stik:

$$(19)\quad Y = B_1X_1 + B_2X_2 + B_3X_3$$

Y angiver de samlede driftsomkostninger forbundet med stik, X_1 angiver antal stik i landzone, X_2 angiver antallet af stik i byzone og X_3 angiver antallet af stik i cityzone.

Resultater

Resultaterne af regressionsanalysen viser følgende resultater efter der er fjernet outliere.

Tabel 9: Første regressionskørsel, Stik

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Land	145,89	49,17	2,967	0,00351
By	179,23	20,94	8,561	1,42e-14
City	513,38	119,81	4,285	3,30e-05
Antal observationer: 147 Justeret R^2= 0,72				

Spredningen for B-værdierne for stik i zonerne land og by overlapper hinanden. Det vil sige, det kan ikke afvises, at B-værdierne for stik i de to zoner er lig hinanden. Forsyningssekretariatet vurderer derfor, at de to zoner land og by lægges sammen som det også er tilfældet på rentvandsledning.

Modellen kommer således til at se ud på følgende måde:

$$(20) \quad Y = B_1(X_1+X_2)+B_3X_3$$

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger). Resultaterne af regressionsanalysen kan ses i tabel 10 nedenfor.

Tabel 10: Regressionsanalysens resultater, Stik

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Land + By	170,97	14,01(21)	12,201(8,095)	<2e-16 (0,0001)
City	530,29	115,24(144)	4,602(3,680)	8,98e-06 (0,0001)
Antal observationer: 149 Justeret R^2= 0,72				

Der er dog stadig tegn på, at indre cityzonen skiller sig væsentligt ud fra de andre rent omkostningsmæssigt, og derfor beregnes en omkostning for indre city ved hjælp af en gennemsnitsberegning som beskrevet på næste side under afsnittet ”Beregning af omkostningen for indre cityzone”.

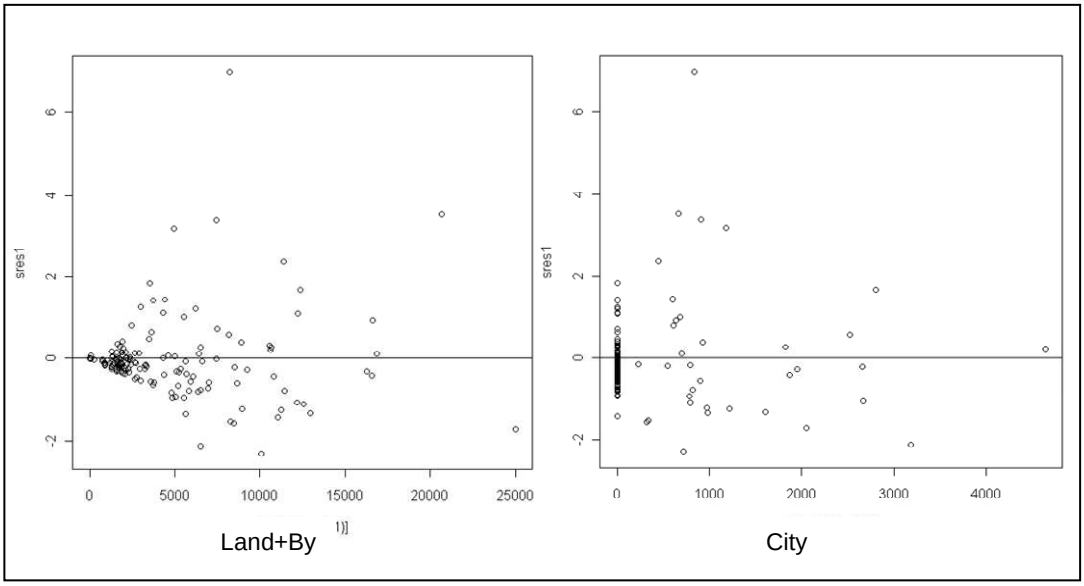
Kontrol af modellen (for zonerne land+by og city)

Først kontrolleres der for om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen som indeholder de to kategorier land+by og city.

Der er to plots der skal kontrolleres, jf. figur 9. Plottet til venstre i figur 9 viser de standardiserede fejlled plottet imod de observerede værdier af antallet af stik i land+by-zone.

Normalfordelingsbetingelsen er ikke fuldt opfyldt, da plottet viser tegn på heteroskedasticitet. Dermed skal der beregnes robuste spredninger samt t- og p-værdier for denne variabel. Resultaterne af denne beregning er angivet i parenteserne i tabel 10.

Figur 9: Fejlledsplot for stik



Plottet til højre i figur 9 viser fejlleddene plottet mod observationerne af antal stik i cityzone. Dette plot viser ikke umiddelbart tegn på heteroskedasticitet og normalfordelingsbetingelsen er derudover opfyldt. For god ordens skyld er den robuste spredning, t-værdi og p-værdi også angivet for stik i cityzonen i parenteser i tabel 10.

Det har været nødvendigt at fjerne enkelte observationer med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Multikollinaritet kontrolleres i modellen ved at se på korrelationen mellem de to parametre, der indgår i modellen for stik. Korrelationen beregnes til 0,64 og ligger dermed lige under grænsen på 0,70. Forsyningssekretariatet

vurderer ikke, at korrelationen skader modellens resultater, og resultaterne accepteres derfor.

Beregning af omkostningen for indre cityzone

For at beregne omkostningsækvivalenten for stik i indre city-zonen beregnes først værdien af de stik, som forsyningerne med indre city-zone har i de øvrige to kategorier (land+by og city). Denne værdi beregnes på baggrund af resultaterne af modellen i ligning (20) som fremgår i tabel 10.

Når denne værdi er fastlagt beregnes en værdi for de omkostninger forsyningerne har med at drive deres indre city stik. Dette gøres ved at tage forskellen på modellens forudsagte værdi og forsyningens samlede indberettede omkostninger til at drive deres indre city stik.

På baggrund af hver forsynings beregnede omkostninger til stik i indre city zone beregnes en gennemsnitlig udgift til stik i indre cityzonen, jf. tabel 11 nedenfor.

Tabel 11: Stik i indre city

Omkostninger i alt	Værdi af øvrige zoner	Forskel	Antal stik	Gennemsnit
286.650	612.722	-326.072	32	-10.190
13.315.913	7.494.149	5.821.764	512	11.371
7.440.430	1.242.921	6.197.509	2.305	2.689
6.508.839	3.157.045	3.351.794	2.975	1.127
13.687.317	8.648.766	5.038.551	6.134	821
30.594.609	9.164.602	21.430.007	8.347	2.567
		Gns. omkostning pr. stik		1.398

Endelig omkostningsækvivalent for stik

Det betyder, at den endelige omkostningsækvivalent til at beskrive omkostningerne forbundet med at drive og vedligeholde stik kan opstilles således:

$$(21) \quad Y = 171(X_1+X_2) + 530X_3 + 1.398X_4$$

Fortolkningen af modellen er, at et stik i landzone (X_1) eller byzone (X_2) har en omkostningsværdi på 171 kr. Tilsvarende har et stik i cityzone (X_3) eller indre city-zone (X_4) en omkostningsværdi på hhv. 530 kr. og 1.398 kr.

Kunder

Forsyningerne har indberettet samlede omkostninger forbundet med kundehåndtering samt antallet af henholdsvis målere og husstande, som forsyningen forsyner med rent vand.

Omkostningerne forbundet med kundehåndtering forventes at afhænge af antallet af enten målere eller husstande, da disse størrelser giver en indikation af, hvor mange kunder forsyningerne har i deres forsyningsområde. Indledende analyser viser, at der ikke er væsentlig forskel på at bruge målere eller husstande. Forsyningssekretariatet vurderer derfor, at det er mest retvisende at benytte antallet af målere, da forsyningernes indberetning af målere er mere præcis end antallet af husstande.

Forsyningssekretariatet har således en formodning om at omkostningsækvivalenten for kunder kan opstilles som en omkostningsfunktion af formen:

$$(22) \quad Y = B_1 X_1$$

Y angiver samlede omkostninger forbundet med kunder og X_1 angiver antallet af målere. B_1 bliver dernæst estimeret ved hjælp af mindste kvadraters metode.

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger) .

Resultater

Resultaterne af regressionsanalysen viser god signifikans i modellen med en p-værdi på under 2e-16. Dette tyder på, at sammenhængen i data er meget tydelig. Yderligere findes B-værdien til 145,329, jf. tabel 12.

Tabel 12: Regressionsanalysens resultater, Kunder

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Målere	145,329	6,668 (10)	21,79 (14)	<2e-16
Antal observationer: 169 Justeret R^2= 0,73				

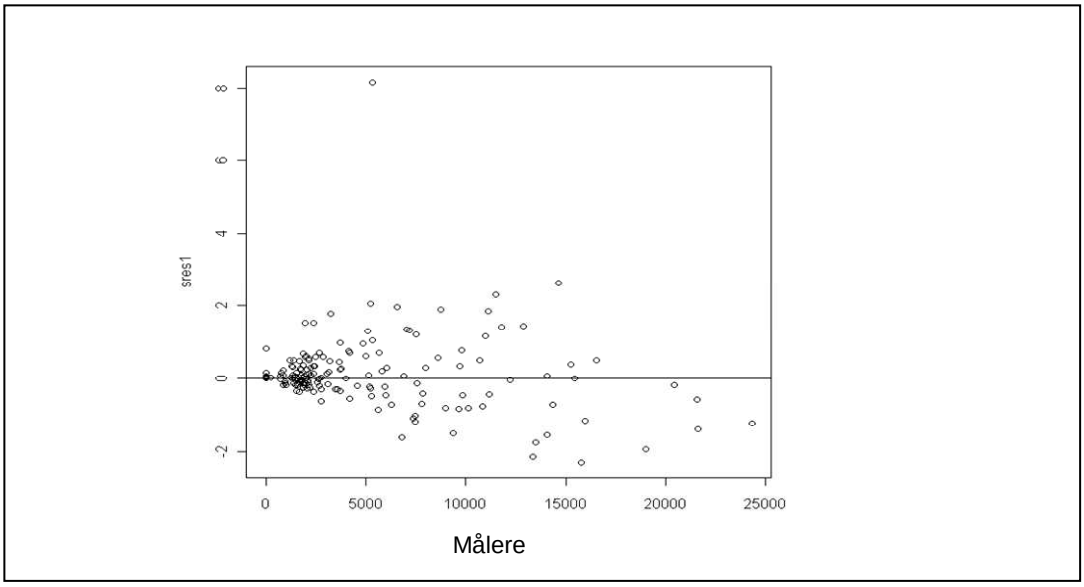
Kontrol af model

Kontrol af om modellens fejlled er normalfordelte med middelværdi 0 udføres ved at undersøge fejlledsplot for variablen X_1.

Der er ikke umiddelbart noget problem i fejlledsplottet, jf. figur 10 nedenfor. Der er dog en smule tegn på heteroskedasticitet, hvilket der kontrolleres for med beregning af den robuste spredning og ny beregnet t-værdi. De korrigerede værdier for spredning og t-værdi fremgår af parenteserne i tabel

12. Disse værdier viser, at heteroskedasticiteten ikke er væsentlig i forhold til signifikansen.

Figur 10: Fejlledsplot for kunder



Det har været nødvendigt at fjerne enkelte observationer med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for kunder

Den endelige omkostningsækvivalent for kunder kan dermed opstilles med følgende udtryk:

$$(23)\quad Y=145{,}3X_1$$

Dermed kan forsyningerne få forklaret omkostninger i forbindelse med kundehåndtering på baggrund af, hvor mange målere en given forsyning har.

Bestemmelse af omkostningsækvivalenter for Spildevand

Forsyningssekretariatet har beregnet en omkostningsækvivalent for hver af de seks costdrivere:

- Ledning
- Pumper
- Åbne bassiner
- Lukkede bassiner
- Renseanlæg
- Kunder

Nedenfor vil beregningen af de enkelte omkostningsækvivalenter blive gennemgået.

Ledning

Omkostningerne i forbindelse med drift af ledning forventes i udgangspunktet både at kunne afhænge af ledningernes længde (meter) samt deres volumen (m^3) fordelt på de fire zoner; land, by, city og indre city.

Forsyningssekretariatet har imidlertid vurderet, at antal meter ledning er den variabel, som beskriver omkostningerne mest præcist, da det er fremgået af indberetningerne, at der har været større usikkerhed omkring opgørelsen af volumen af forsyningernes ledningsnet og derfor, at der er foretaget flere skøn, som er forskellige i forsyningernes opgørelser.

Forsyningssekretariatet har derfor besluttet at benytte ledning målt i meter som beskrivende variabel for driftsomkostningerne forbundet med ledning. Det fremgår også af resultatet af den valgte model nedenfor, at determinationskoefficienten R^2 er høj og dermed, at data er godt beskrevet af modellen.

Forsyningssekretariatet forventer yderligere, at omkostningerne forbundet med drift af ledningsnettet afhænger lineært af antallet af meter ledning i de 4 zoner. Dermed kan den første model for omkostningsækvivalenten forbundet med ledning opstilles som:

$$(24) \quad Y = B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$$

Hvor Y er omkostningerne forbundet med drift af det samlede antal meter ledning, X_1 antal meter i landzonen, X_2 antal meter i byzonen, X_3 antal meter i cityzonen og X_4 antal meter i indre cityzonen.

Resultater

Resultaterne af regressionsanalysen for den lineære model beskrevet ved ligning (24) ovenfor, viser imidlertid, at det ikke er muligt at beskrive data ved hjælp af alle fire forklarende variable.

Det skyldes, at variablen land (X_1) bliver insignifikant i modellen. Det vil sige, at land-kategorien ifølge modellen ikke - isoleret set - bidrager til at forklare de samlede driftsomkostninger forbundet med ledningsnettet jf. tabel 13 nedenfor. Derudover viser analysen, at alle observationer, som har ledning i indre cityzone, er outliere. Det vil sige, disse observationer har en uhensigtsmæssig indflydelse på resultaterne, når kategorien indre city indgår som en selvstændig kategori i modellen.

DANVA har i deres høringssvar anført, at Forsyningssekretariatet har en stram fortolkning af signifikanskriteriet, når en p-værdi på 0,0777 medfører at kategorierne land- og byzone lægges sammen.

I tabel 13 nedenfor er samtlige observationer med indre city ledning som sagt outliere. Forsyningssekretariatet har derfor også lavet en regressionskørsel hvor city og indre city er lagt sammen samtidig med at land- og by-kategorierne holdes adskilt. Resultatet af denne regressionskørsel viser, at land- og by-kategorien får en B-værdi på henholdsvis 5,7 og 5,3. Forsyningssekretariatet fastholder derfor, at kategorierne land og by bør lægges sammen til én samlet kategori.

Tabel 13: Første regressionskørsel, Ledning

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Land	4,487	2,512	1,787	0,0777
By	6,653	1,623	4,099	9,70e-05
City	75,997	13,648	5,568	3,17e-07
Antal observationer: 83 Justeret R^2= 0,8368				

Derfor slås Land- og By-kategorierne samt City- og Indre City-kategorierne sammen, således at omkostningerne bliver beskrevet ved følgende model:

$$(25) \quad Y = B_1(X_1+X_2) + B_2(X_3+X_4)$$

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette

udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

De endelige resultater af regressionsanalysen viser, at begge de forklarende variable Land+By og City+Indre City er signifikante. Det vil sige, at begge variable har betydning for de omkostninger, der er forbundet med at drive ledningsnettet. Derudover er R^2 = 0,8359, hvilket betyder at 84 pct. af omkostninger forbundet med drift af ledning er beskrevet ved den lineære regressionsmodel opstillet i ligning (25).

B-værdierne bliver hhv. B_1 = 5,4725 og B_2 = 93,5338, jf. tabel 14 nedenfor.

Tabel 14: Resultat af regressionsanalysen for ledning:

Variabel	B-værdi	Spredning	t-værdi	Pr(>|t|)
Land+By	5,4725	0,5125	10,580	2e-16
City+Indre City	93,5338	10,5958	8,827	1,01e-13
Antal observationer: 89 Justeret R^2= 0,8359				

DANVA og FVD samt flere forsyninger har i deres høringssvar kommenteret på, at zonerne er reduceret til "kun 2", samt at springet i B-værdiens størrelse mellem Land+By-zonen og City+Indre City-zonen umiddelbart virker meget stor.

Forsyningssekretariatet har derfor gennemført en række yderlige regressionsanalyser for rentvandsledninger for at teste, hvor robuste B-værdiernes størrelse er for de forskellige zoner.

For at teste dette, er datasættet splitttet op i forskellige dele, hvorefter regressionsanalysen er kørt på de forskellige datasæt.

Datasættet er således blevet opdelt i 3 nye datasæt:

- Forsyninger som kun har ledning i land- og/eller byzone
- Forsyninger som har ledning i cityzone, men ikke i indre city-zone (de fleste af disse forsyninger har også ledning i land og by)
- Forsyninger som har ledning i cityzone (dvs. forsyninger som har ledning i både cityzone og indre city-zone er medtaget her)

Denne opsplitning er foretaget, for at kontrollere om resultaterne påvirkes af hvilke zoner forsyningerne har ledning i. Herunder om der er en tendens til

at være forskel i fordelingen af omkostninger på ledninger alt efter størrelsen af forsyningerne[5].

Resultatet af disse regressionskørsler fremgår af tabel 15 nedenfor. En streg angiver, at det ikke har været muligt at beregne en værdi for den pågældende faktor.

Tabel 15: Resultat (B-værdier) for regressionsanalyser for forskellige sammensætninger af datasættet for ledning.

Variabel	Forsyninger der kun har ledning i land og/eller by	Forsyninger der har ledning i city, men ikke i indre city	Forsyninger der har ledning i city	Alle forsyninger
Land + By	5,54	5,68	5,08	5,47
City	-	81,47		
City + Indre City	-		88,26	93,53

Resultaterne viser, at B-værdien for Land+By-zonen ikke ændrer sig væsentligt, uanset hvordan data sættes sammen. Prisen for at drive en meter ledning i Land+By-zone er således 5,54 kr., hvis værdien kun estimeres på baggrund af forsyninger, der har ledning i land og/eller by. Det vil sige 7 øre mere, end hvis værdien beregnes på baggrund af alle forsyningers observationer (5,47 kr.).

Springet mellem Land+By-zonen og City+Indre City-zonen er markant for alle data-sammensætninger.

Forsyningssekretariatet konkluderer på baggrund af ovenstående, at resultatet i tabel 14 er retvisende og at koefficienternes størrelse hverken skyldes datamaterialets sammensætning eller eventuelle usikkerheder i den måde som virksomhederne har fordelt deres omkostninger på. Der fortsættes derfor med den generelle kontrol af modellen i ligning (25).

Kontrol af model

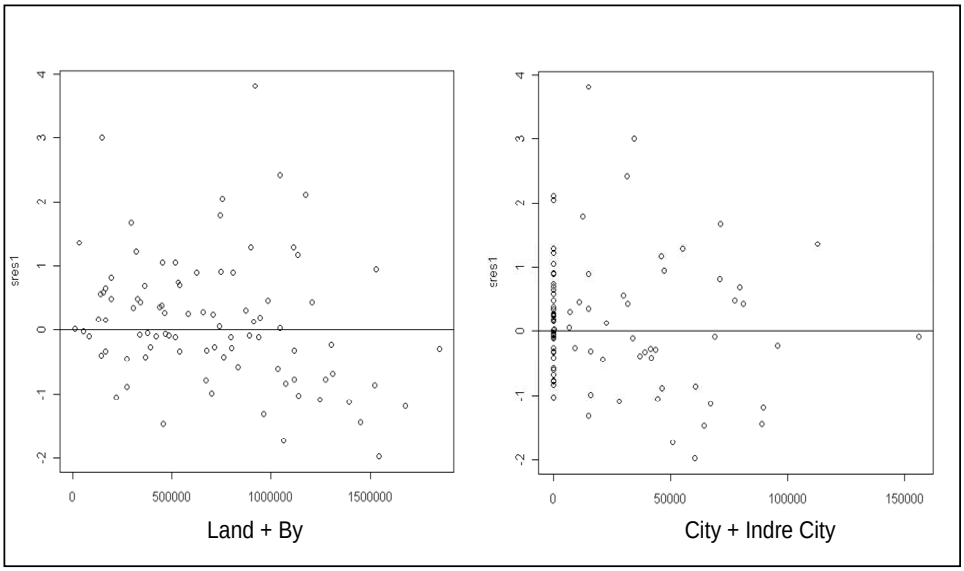
Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater, er opfyldt for modellen. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af de to kategorier af ledning.

De to plots i figur 11 nedenfor viser umiddelbart en tilfældig fordeling af de standardiserede fejlled omkring 0. Normalitetsbetingelsen vurderes

[5] De mindre selskaber har hovedsageligt kun ledning i land- og byzone, mens de større selskaber også har ledning i city- og indre city-zone.

derfor at være opfyldt, ligesom der ikke vurderes at være nævneværdig heteroskedasticitet i modellen.

Figur 11: Fejlledsplot for ledning



Multikollinaritet kontrolleres i modellen ved at se på korrelationen mellem de to parametre, der indgår i modellen for ledning. Korrelationen beregnes til 0,33 og ligger dermed et pænt stykke under grænsen på 0,70. Forsyningssekretariatet vurderer ikke at korrelationen skader modellens resultater og accepterer derfor resultaterne.

Det har været nødvendigt at fjerne enkelte observationer med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for ledning

På baggrund af ovenstående ser den endelige model for ledning således ud:

$$(26) \quad Y = 5{,}47(X_1+X_2) + 93{,}53(X_3 + X_4)$$

Pumper

Forsyningerne har indberettet de samlede driftsomkostninger forbundet med drift af pumper samt antallet af pumper fordelt på kategorierne:

Kategori	Størrelse
a	0-10 l/s
b	11 - 100 l/s
c	101 - 300 l/s
d	301 - 600 l/s
e	601 - 1000 l/s
f	over 1000 l/s

Omkostningerne forbundet med drift af ledningsnettet forventes at afhænge af antallet af pumper i de 6 kategorier. Den første model for omkostningsækvivalenten forbundet med pumper opstilles som:

$$(27) \quad Y = B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6$$

Hvor Y er omkostningerne forbundet med drift af det samlede antal pumper, X_1 antallet af kategori a pumper, X_2 antallet af kategori b pumper, X_3 antallet af kategori c pumper, X_4 antallet af kategori d pumper, X_5 antallet af kategori e pumper og X_6 antallet af kategori f pumper.

Resultater

Resultaterne af regressionsanalysen for den lineære model beskrevet ved ligning (27), viser imidlertid at kategorierne c, d, e og f bliver insignifikante isoleret set når alle 6 kategorier indgår i modellen, jf. tabel 16 nedenfor.

Tabel 16: Første regressionskørsel for pumper:

Variabel	B-værdi	Spredning	t-værdi	p-værdi
kategori a	8.274	1.752	4,722	9,52e-06
kategori b	16.623	2.673	6,220	1,99e-0,8
Kategori c	48.323	26.635	1,814	0,0733
Kategori d	188.298	97.357	1,934	0,0566
Kategori e	117.449	244.567	0,480	0,6323
kategori f	788.024	403.398	1,953	0,0542
Antal observationer: 83 Justeret R^2: 0,83				

Forsyningssekretariatet har derfor slået flere af kategorierne sammen på følgende måde:

Kategori	Størrelse
a	0-10 l/s
b	11 - 100 l/s
c-d	101-600
e-f	601 - max l/s

I det udkast som Forsyningssekretariatet udsendte i marts måned var pumpekategorierne slået sammen således at kategorierne hed: a, b-c og d-f. På baggrund af høringssvarene fra bl.a. DANVA og flere af forsyningerne, har Forsyningssekretariatet i stedet forsøgt at adskille kategorierne yderligere, således at kapacitetsspændet i de enkelte kategorier bliver mindre, og derved øge sammenligningsgrundlaget mellem pumper indenfor samme kategori.

For at opdele pumperne i de i tabellen nævnte kategorier foretages beregningen af omkostningsækvivalenten for pumper på tilsvarende måde som beregningen af stik for vandforsyninger, jf. afsnittet vedr. stik ovenfor. Det vil sige, at det har været nødvendigt at tage forsyninger med pumper i kategori e og/eller f ud af datasættet og i første omgang kun beregne driftsomkostningerne forbundet med pumper i kategorierne a, b, c og d på baggrund af regressionsanalysen.

Forsyninger med pumper i kategori e og/eller f er taget ud af datasættet, da det ikke er muligt at estimere B-værdier for disse kategorier. Det skyldes dels at der er meget få forsyninger som har kategori e og f pumper, samt at en stor andel af disse forsyninger bliver identificeret som outliere i modellen, og derfor alligevel ryger ud af beregningen.

Det vil sige den lineære regressionsmodel ser i stedet således ud:

$$(28) \quad Y = B_1(X_1) + B_2(X_2) + B_3(X_3 + X_4)$$

Hvor Y er omkostningerne forbundet med drift af det samlede antal pumper, X_1 er antallet af kategori a pumper, X_2 er antallet af kategori b pumper og $X_3 + X_4$ er antallet af kategori c+d pumper. B-værdierne vil blive bestemt med mindste kvadraters metode.

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

De endelige resultater af regressionsanalysen for den lineære model beskrevet ved ligning (28) viser, at de 3 forklarende variable - a pumper, b pumper og c+d pumper - alle er signifikante. Det vil sige, at alle tre

variable har betydning for de omkostninger, der er forbundet med at drive pumper.

B-værdierne bliver hhv. B_1 = 7.983 og B_2 = 17.432 og B_3 = 67.697, jf. tabel 17 nedenfor.

Tabel 17: Resultat af regressionsanalysen for pumper:

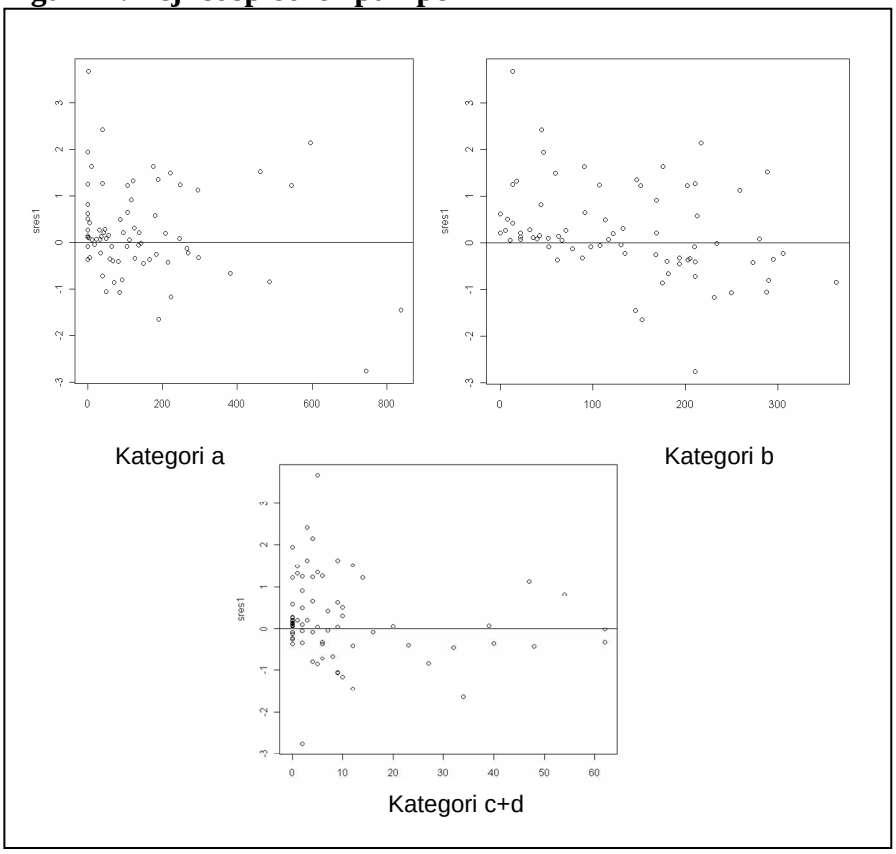
Variabel	B-værdi	Spredning	t-værdi	p-værdi
kategori a	7.983	1.825	4,374	4,10e-05
kategori (b)	17.432	2.924	5,963	8,78e-08
kategori (c + d)	67.697	21.029	3,219	0,00194
Antal observationer: 74 Justeret R^2: 0,81				

Kontrol af modellen (for kategorierne a, b, c+d)

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af de 3 kategorier af pumper.

De tre plots i figur 12 nedenfor viser umiddelbart en tilfældig fordeling omkring 0. Normalitetsbetingelsen vurderes derfor at være opfyldt, ligesom der ikke vurderes at være heteroskedasticitet i modellen.

Figur 12: Fejlledsplot for pumper



Det har været nødvendigt at fjerne en enkelt observation med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Beregning af omkostningen for pumpekategorien e+f

For at beregne omkostningsækvivalenten for pumper i kategorien e+f beregnes først værdien af de pumper, som forsyningerne med pumper i kategorierne e og f har i de øvrige pumpekategorier (a, b og c+d). Denne værdi beregnes på baggrund af resultaterne af modellen i ligning (28) som fremgår i tabel 17.

Når denne værdi er fastlagt beregnes en værdi for de omkostninger forsyningerne har med at drive deres pumper i kategori e+f. Dette gøres ved at tage forskellen på modellens forudsagte værdi og forsyningens samlede indberettede omkostninger til at drive deres pumper.

På baggrund af hver forsynings beregnede omkostninger til pumper i kategori e+f beregnes en gennemsnitlig udgift til pumper i kategorien e+f, jf. tabel 18 nedenfor.

Tabel 18: Residualberegning for kategori e-f pumper

Omkostninger i alt i kr.	Værdi af øvrige pumper i kr.	Forskel i kr.	Antal e+f pumper	Gennemsnit i kr.
1.842.312	1.361.533	480.779	1	480.779
4.237.413	5.205.158	-967.745	1	-967.745
8.534.096	4.808.377	3.725.719	1	3.725.719
6.110.207	3.011.870	3.098.337	2	1.549.169
10.061.998	13.111.610	-3.049.612	2	-1.524.806
10.225.176	4.263.856	5.961.320	5	1.192.264
61.317.808	6.681.248	54.636.560	16	3.414.785
1.614.919	933.025	681.894	1	681.894
5.167.090	4.867.334	299.756	1	299.756
7.357.052	3.848.106	3.508.946	1	3.508.946
5.042.861	4.476.647	566.214	2	283.107
2.554.931	3.862.220	-1.307.289	4	-326.822
4.566.605	5.372.188	-805.583	3	-268.528
11.020.409	9.708.508	1.311.901	3	437.300
8.172.047	6.526.898	1.645.149	4	411.287
2.711.290	473.879	2.237.411	11	203.401
4.634.909	3.906.631	728.278	8	91.035
	82.419.088			13.191.541
Gns. omkostning pr. pumpe				**775.973 kr.**

Endelig omkostningsækvivalent for pumper

På baggrund af ovenstående ser den endelige model for pumper således ud:

$$(29) \quad Y = 7.983X_1 + 17.432\ X_2 + 67.697(X_3 + X_4) + 775.973(X_5 + X_6)$$

Fortolkningen af modellen er, at en kategori a pumpe har en omkostningsværdi på 7.983 kr., en kategori b pumpe har en omkostningsværdi på 17.432 kr., en kategori c+d pumpe har en omkostningsværdi på 67.697 kr. og e+f pumpe har en omkostningsværdi 775.973 kr.

Åbne Bassiner

Omkostningerne forbundet med drift af åbne bassiner forventes at afhænge af antallet af åbne bassiner samt den samlede størrelse af bassinerne.

En indledende kontrol af modellen viser, at der er en høj korrelation mellem antallet af bassiner og den samlede størrelse af bassinerne på 0,84. Det betyder, at en stor del af sammenhængen mellem omkostninger og henholdsvis antal bassiner og størrelsen af bassiner er identisk. Dermed er det ikke relevant at benytte en model, som inkluderer både antal og størrelse. Yderligere viser modellen, at begge variable ikke er signifikante, når de indgår sammen i modellen.

Forsyningssekretariatet har vurderet, at antal bassiner er den bedste af de to variable til at beskrive driftsomkostningerne forbundet med åbne bassiner. Det skyldes primært, at R^2 er lidt højere i modellen, hvor antal indgår som forklarende variabel i forhold til modellen, hvor størrelsen af bassiner indgår som forklarende variabel. Derudover vurderer Forsyningssekretariatet, at indberetningerne vedrørende antal bassiner er mere præcis end det er tilfældet for størrelsen af bassinerne.

Dermed kan modellen for omkostningsækvivalenten forbundet med åbne bassiner opstilles som:

$$(30) \quad Y = B_1 X_1$$

Hvor Y er omkostningerne forbundet med drift af de åbne bassiner og X_1 er antallet af bassiner.

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

Resultater

De endelige resultater af regressionsanalysen viser, at den forklarende variabel 'antal bassiner' er signifikant, dvs. at variablen som forventet har betydning for driftsomkostningerne forbundet med drift af de åbne bassiner.

B-værdien findes til 13.523, jf. tabel 19 nedenfor:

Tabel 19: Resultat af regressionsanalysen for åbne bassiner:

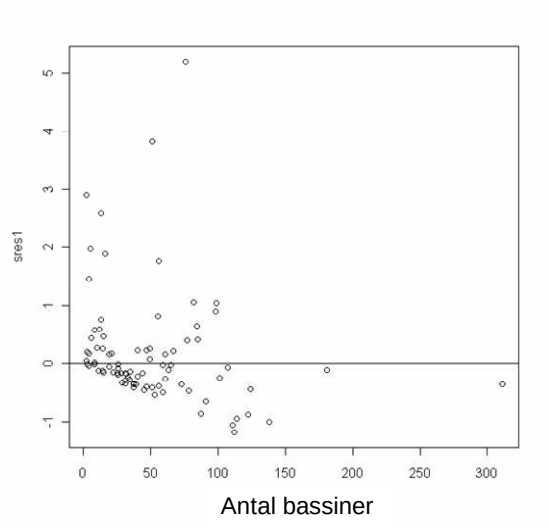
Variabel	B-værdi	Spredning	t-værdi	p-værdi
Antal bassiner	13.523	1.905 (2.117)	7,098 (6,39)	3,56e-10 (0,0001)
Antal observationer: 87 Justeret R^2: 0,36				

Kontrol af model

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af antal bassiner.

Plottet i figur 13 nedenfor viser umiddelbart en tilfældig fordeling omkring 0. Plottet viser dog en tendens til heteroskedasticitet i modellen. For at kontrollere for om heteroskedasticiteten har betydning for signifikansen i modellen beregnes en robust t-værdi. Den robuste t-værdi er 6,39, hvilket stadig medfører at variablen er signifikant med en p-værdi mindre end 0,0001, jf. tabel 19. Normalitetsbetingelsen vurderes derfor at være opfyldt, ligesom der ikke vurderes at være væsentlig heteroskedasticitet i modellen.

Figur 13: Fejlledsplot for åbne bassiner



Der er ikke nogen enkelte observationer, der har stor indflydelse på resultatet. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for åbne bassiner

På baggrund af ovenstående ser den endelige model for åbne bassiner således ud:

$$(31) \quad Y = 13.523X_1$$

Lukkede bassiner

Omkostningerne forbundet med drift af lukkede bassiner forventes at afhænge af antallet af åbne bassiner samt den samlede størrelse af bassinerne.

En indledende kontrol af regressionsmodellen, hvor både antal og størrelse indgår viser, at begge variable ikke er signifikante når de indgår i modellen.

Forsyningssekretariatet har vurderet, at størrelsen af de lukkede bassiner opgjort i m^3 er den bedste af de to variable til at beskrive driftsomkostningerne forbundet med lukkede bassiner. Det skyldes primært, at R^2 er væsentlig højere i modellen, hvor størrelse indgår som forklarende variabel, i forhold til modellen hvor antallet af bassiner indgår som forklarende variabel.

Dermed kan modellen for omkostningsækvivalenten forbundet med lukkede bassiner opstilles som:

$$(32) \quad Y = B_1 X_1$$

Hvor Y er omkostningerne forbundet med drift af de lukkede bassiner og X_1 er den samlede størrelse af bassinerne angivet i m^3.

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

Resultater

De endelige resultater af regressionsanalysen viser, at den forklarende variabel 'samlede størrelse af bassiner' er signifikant, dvs. at variablen som forventet har betydning for driftsomkostningerne forbundet med drift af de lukkede bassiner. B-værdien findes til 24,06, jf. tabel 20 nedenfor:

Tabel 20: Resultat af regressionsanalysen for lukkede bassiner:

Variabel	B-værdi	Spredning	t-værdi	p-værdi
Størrelse af bassiner	24,06	2,98 (4,31)	8,075 (5,58)	9,34e-12 (0,0001)
Antal observationer: 76 Justeret R^2: 0,46				

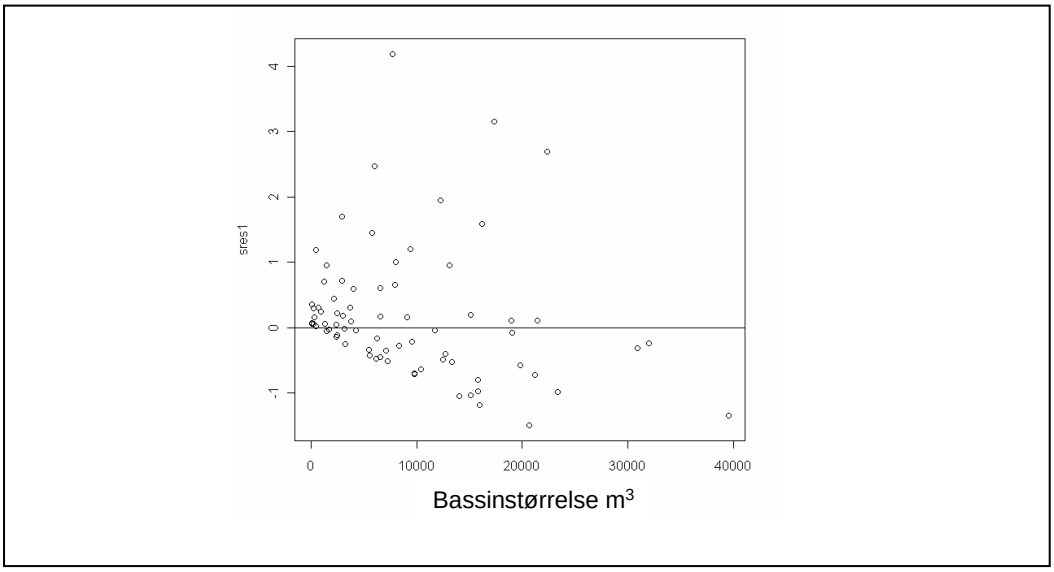
Kontrol af model

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af størrelsen af bassiner.

Plottet i figur 14 nedenfor viser umiddelbart en tilfældig fordeling omkring 0. Plottet viser dog en tendens til heteroskedasticitet i modellen. For at kontrollere om heteroskedasticiteten har betydning for signifikansen i modellen beregnes en robust t-værdi.

Den robuste t-værdi beregnes til 5,58, hvilket stadig medfører, at variablen er signifikant med en p-værdi mindre end 0,0001. Normalitetsbetingelsen vurderes derfor at være opfyldt, ligesom det vurderes, at heteroskedasticiteten i modellen ikke påvirker resultatet.

Figur 14: Fejlledsplot for lukkede bassiner



Det har været nødvendigt at fjerne enkelte observationer med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for lukkede bassiner

På baggrund af ovenstående ser den endelige model for lukkede bassiner således ud:

$$(33) \quad Y = 24{,}06X_1$$

Renseanlæg

Forsyningerne har indberettet omkostninger for hvert af deres renseanlæg. Derudover har forsyningerne indberettet en række oplysninger vedr. type, kapacitet, belastning, slamhåndtering og udløbskvalitet.

Omkostningerne forbundet med drift af renseanlæg kan afhænge af flere af de indberettede faktorer. Forsyningssekretariatet har dog vurderet, ved at

teste forskellige modeller, hvor faktorerne indgår som forklarende variable på forskellig vis, at belastningen på renseanlægget målt i PE er den faktor som bedst beskriver driftsomkostningerne på det enkelte renseanlæg.

Årsagen til at Forsyningssekretariatet kun benytter en model der indeholder PE er, at p-niveauet og C/N-forholdet er fundet insignifikante i de indledende analyser. Yderligere er faktorerne vedrørende renseanlæggets type og slambehandlingsmetode ikke blevet inddraget i modellen, da det ikke har været muligt at inddrage disse faktorer på en fornuftig måde.

De indledende undersøgelser viste at det kun er muligt at opdele efter typerne "c-slam" eller "ikke c-slam" og efter rensningstyperne "mere end MBNK" og "mindre end MBNK". Forsyningssekretariatet har efterfølgende undersøgt om der var en sammenhæng imellem omkostningerne og hvilken slambehandlingsmetode eller rensningsmetode, der foretages på det pågældende anlæg. Undersøgelsen tyder på, at der ikke er nogen sammenhæng i data imellem disse forhold. Derfor er disse variable udeladt af omkostningsækvivalenten for renseanlæg.

På baggrund af høringssvarene har Forsyningssekretariatet fundet det rimeligt at ændre opgørelsen af omkostningsækvivalenten for renseanlæg, mht. at inkludere kapaciteten af renseanlæggene.

Dette skyldes to forhold nævnt i høringssvarene. Det første er, at der kan være stor usikkerhed omkring præcisionen af målingerne af belastningen i PE. Denne kan svinge med op til 20 pct. i hver retning. Derudover er det også nævnt, at der kan være større faste omkostninger ved at have stor kapacitet, og en større omkostning ved at værket er i drift på trods af den mindre faktiske belastning.

Forsyningssekretariatet har derfor vurderet, at det er mest rimeligt at konstruere et nyt mål, som skal indgå i omkostningsækvivalenten for renseanlæg. Dette mål bliver konstrueret som gennemsnittet af kapaciteten og belastningsgraden målt i PE. Dermed kan omkostningsækvivalenten opstilles som følger.

$$(34) \quad Y = B_1 X_1$$

Hvor Y er omkostningerne forbundet med drift af renseanlægget og X_1 er gennemsnittet af anlæggets belastning og kapacitet målt i PE.

En indledende kontrol af modellen viser, at der er en tendens til skalafordele i datasættet. Det vil sige, at jo større belastning og kapacitet der er på renseanlægget jo lavere bliver enhedsomkostningerne.

Da modellen (34) ovenfor er lineær, vil den ikke tage højde for de formodede skalafordele. Forsyningssekretariatet foretager derfor en transformation af funktionen for at tilpasse modellen til data.

Den valgte transformation af funktionen ser således ud:

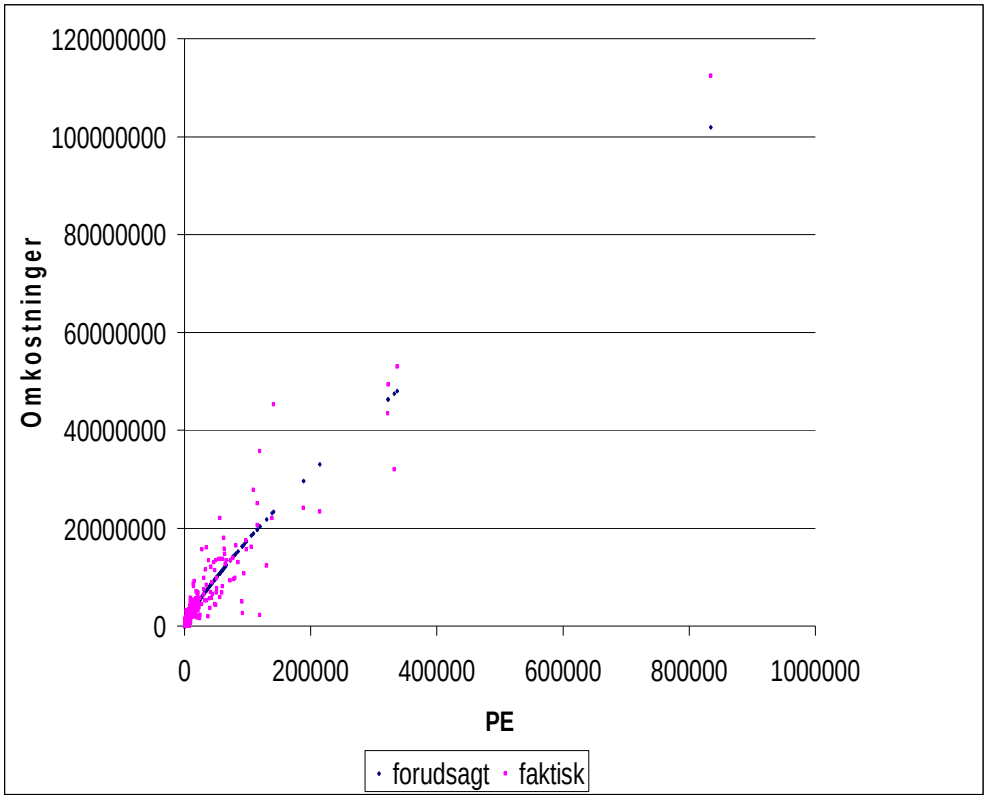
$$(35) \quad Y^{0,40} = B_1 X_1^{\,0,33}$$

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

Forsyningssekretariatet har valgt ovenstående funktion på baggrund af en række test af andre mulige transformationer. Forsyningssekretariatet vurderer, at den valgte transformation beskriver data bedst. Modellen har også den fordel, at det ikke har været nødvendigt at fjerne nogen observationer fra datagrundlaget.

Figur 15 nedenfor viser et tilpasningsplot. De mørkeblå punkter angiver modellens forudsagte omkostninger og de røde punkter angiver de faktiske omkostninger. Omkostningerne er plottet imod gennemsnittet af anlæggets belastning og kapacitet i PE.

Figur 15: Tilpasningsplot af de forudsagte og de faktiske omkostninger



Resultater

De endelige resultater af regressionsanalysen viser, at den forklarende variabel 'antal PE' er meget signifikant. Det vil sige, at variablen som forventet har betydning for driftsomkostningerne forbundet med drift af renseanlæg.

B-værdien findes til 17,27 jf. tabellen nedenfor.

Tabel 21: Resultat af regressionsanalysen for renseanlæg:

Variabel	B-værdi	Spredning	t-værdi	p-værdi
Antal PE	17,27	0,14 (0,38)	123 (50,4)	2e-16 (0,0001)
Antal observationer: 586 Justeret R^2: 0,96				

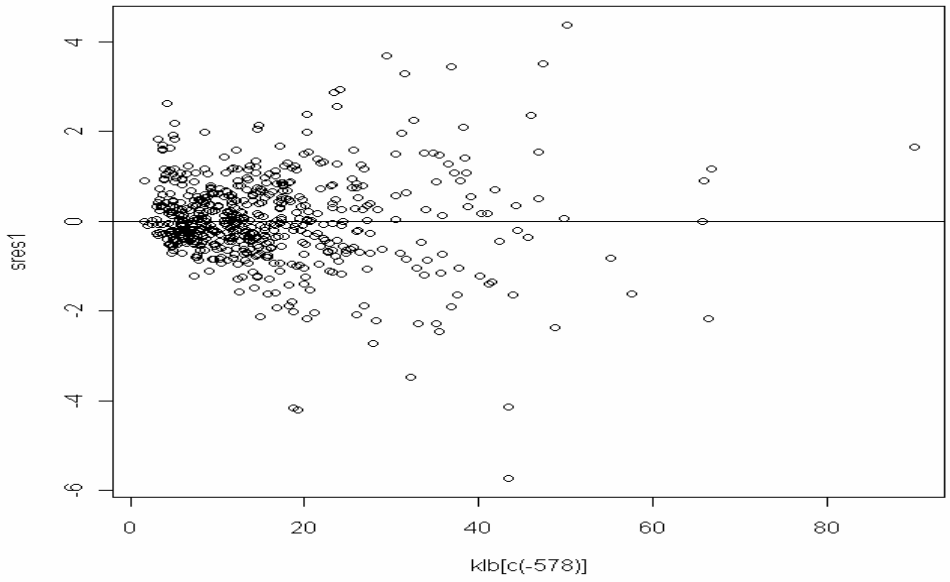
Modelkontrol

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen. Kontrollen foretages ved at plotte de standardiserede fejlled mod de observerede værdier af gennemsnittet af renseanlæggets belastning og kapacitet i PE.

Plottet i figur 16 nedenfor viser umiddelbart en tilfældig fordeling omkring 0. Plottet viser dog en tendens til at der heteroskedasticitet i modellen. Der beregnes derfor en robust t-værdi, for at kontrollere for om heteroskedasticiteten har betydning for signifikansen i modellen.

Den robuste t-værdi beregnes til 50,4, hvilket stadig medfører at variablen er meget signifikant med en p-værdi mindre end 0,0001. Normalitetsbetingelse vurderes derfor at være opfyldt, ligesom det vurderes, at heteroskedasticitet i modellen ikke påvirker resultatet.

Figur 16: Fejlledsplot for renseanlæg



Der er ikke nogen enkelte observationer der har stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelige omkostningsækvivalent for renseanlæg

På baggrund af ovenstående ser den endelige model for renseanlæg således ud:

$$(36) \quad Y = B_1^{2,5} X_1^{0,83} + KF$$

KF er korrektionsfaktoren som er det kvadrerede fejlled, dvs. $46{,}9^2$ = 2.195 kr. Når B_1 og KF indsættes ser modellen således ud:

$$(37) \quad Y = 1.239 X_1^{0,83} + 2.195$$

Kunder

Forsyningerne har indberettet samlede omkostninger forbundet med kundehåndtering, samt antallet af henholdsvis målere og husstande som forsyningen aftager spildevand fra.

Omkostningerne forbundet med kundehåndtering forventes at afhænge af antallet af enten målere eller husstande, da disse størrelser giver en indikation af, hvor mange kunder forsyningerne har i deres forsyningsområde. Indledende analyser viser, at der ikke er væsentlig forskel på at bruge målere eller husstande.

Forsyningssekretariatet har dog valgt at tage udgangspunkt i antal målere som udtryk for antallet af kunder. Det skyldes, at det har været vanskeligt for rigtig mange forsyninger at opgøre antal husstande i forsyningens forsyningsområde. Forsyningssekretariatet har derfor vurderet at datamaterialet for målere er mest retvisende.

Forsyningssekretariatet har således en formodning om, at omkostningsækvivalenten for kunder kan opstilles som en omkostningsfunktion af formen:

$$(38) \quad Y=B_1X_1$$

Hvor Y er omkostningerne forbundet med kundehåndtering og X_1 er antallet af målere. B_1 bestemmes ved hjælp af mindste kvadraters metode.

Der er ikke medtaget faste omkostninger i modellen, da regressionsanalysens resultater viser, at data bliver væsentligt dårligere beskrevet af modellen, hvis der inkluderes faste omkostninger. Dette udtrykkes ved at R^2 bliver væsentlig lavere, hvis der inkluderes et B_0 (faste omkostninger).

Resultater

De endelige resultater af regressionsanalysen viser, at den forklarende variabel 'antal målere' er signifikant, dvs. at variablen som forventet har betydning for driftsomkostningerne forbundet med kundehåndtering.

B-værdien findes til 124,29 jf. tabellen nedenfor.

Tabel 22: Resultat af regressionsanalysen for kunder:

Variabel	B-værdi	Spredning	t-værdi	p-værdi
Antal målere	124,29	7,23 (10,79)	17,2 (11,51)	2e-16 (0,0001)
Antal observationer: 81 Justeret R^2: 0,78				

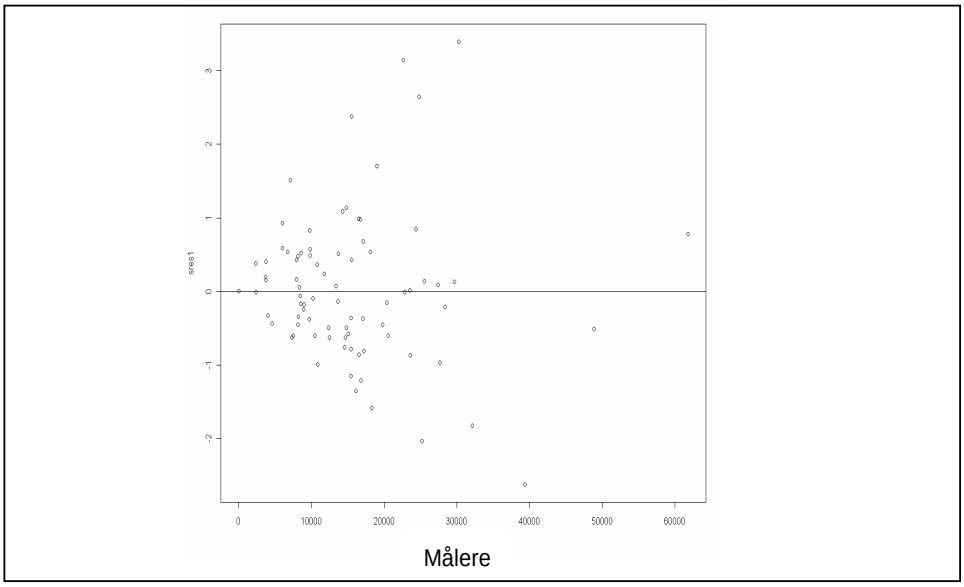
Modelkontrol

Det skal kontrolleres om betingelserne for at bruge regressionsanalysens resultater er opfyldt for modellen.

Plottet i figur 17 nedenfor viser umiddelbart en tilfældig fordeling omkring 0. Plottet viser dog en tendens til, at der er heteroskedasticitet i modellen. For at kontrollere om heteroskedasticiteten har betydning for signifikansen i modellen beregnes en robust t-værdi.

Den robuste t-værdi beregnes til 11,51, hvilket stadig medfører at variablen er meget signifikant med en p-værdi mindre end 0,0001. Normalitetsbetingelsen vurderes derfor at være opfyldt, ligesom det vurderes at heteroskedasticitet i modellen ikke påvirker resultatet.

Figur 17: Fejlledsplot for kunder



Det har været nødvendigt at fjerne enkelte observationer med stor indflydelse på B–værdierne. Dette måles med Cook's afstand som beskrevet i afsnittet *"Ekstreme observationer og observationer med stor indflydelse"* ovenfor.

Endelig omkostningsækvivalent for kunder

På baggrund af ovenstående ser den endelige model for omkostninger forbundet med kundehåndtering således ud:

$$(39) \quad Y = 124{,}3X_1$$