LABORATORY FOR INDUSTRIAL MATHEMATICS EINDHOVEN

Endinet

Regressie-analyse Energiekamer

Laboratory for Industrial Mathematics Eindhoven
Postbus 513
5600 MB Eindhoven
tel.: 040 247 4875
fax: 040 244 2489
e-mail: lime@tue.nl
WWW: http://www.lime.tue.nl

A. Di Bucchianico / B.J. Janssen 21 april 2010

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

INLEIDING

Eén maal in de drie jaar wordt door de Energiekamer de methode gewijzigd waarmee de tarieven bepaald worden voor de netwerkbeheerders. Begin 2010 heeft de Energiekamer in een ontwerpbesluit laten weten aansluitdichtheid als kostendriver mee te nemen. Via een regressie-analyse meent de Energiekamer voldoende onderbouwd te hebben dat aansluitdichtheid een relevante kostenfactor is. Endinet twijfelt aan de juistheid van de door de Energiekamer gebezigde statistische analyse. De opdracht voor LIME (een onderdeel van de Technische Universiteit Eindhoven, faculteit Wiskunde en Informatica) is een grondige regressie-analyse uit te voeren en deze kritisch te vergelijken met de door de Energiekamer uitgevoerde statistische analyse.

ANALYSE

De regressie-analyse is uitgevoerd op de dataset die de Energiekamer heeft aangeleverd in het databestand "`102382 Regressieanalyse aansluitdichtheid voor ontwerp NE5R.xls`". De dataset bevat gegevens van 9 netwerkbeheerders. De namen van de netwerkbeheerders staan in Tabel 1. De nummers in deze tabel worden verder in dit rapport gebruikt in de grafieken.

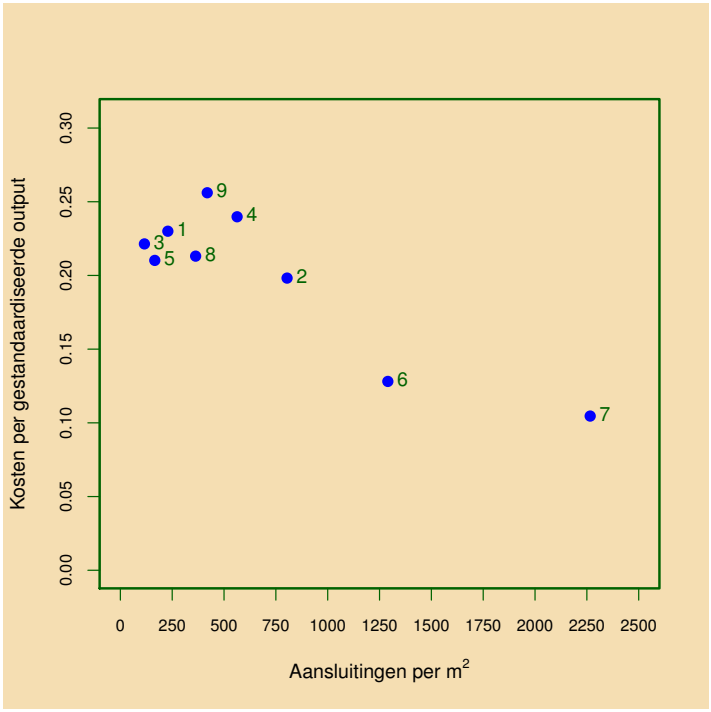
	Naam
1	N.V. Continuon Netbeheer
2	Netbeheerder Centraal Overijssel B.V.
3	DELTA Netwerkbedrijf B.V.
4	ENECO Netbeheer B.V. (STEDIN)
5	Essent Netwerk B.V.
6	NRE Netwerk B.V.
7	ONS netbeheer
8	RENDO Netbeheer B.V.
9	Westland Energie Infrastructuur B.V.

Tabel 1: Namen van netwerkbeheerders

Om te onderzoeken of aansluitdichtheid een kostendriver is, heeft de Energiekamer een regressie-analyse uitgevoerd met aansluitingen per m^2 als onafhankelijke variabele en gestandaardiseerde kosten per eenheid output als afhankelijke variabele (responsvariabele). In principe hoeft het kleine aantal waarnemingen geen probleem te zijn voor een dergelijke analyse, mits men de juiste diagnostische toetsen uitvoert.

Analyse volledige data

Alvorens de regressie-analyse uit te voeren, is het goed eerst de ruwe data grafisch weer te geven. In Figuur 1 is te zien dat netwerkbeheerders 6 en 7 qua aantal aansluitingen per m^2 sterk afwijken van de overige netwerkbeheerders.
We gaan nu eerst een lineair regressiemodel (OLS = Ordinary Least Squares) maken van de data met aansluitingen per m^2 als onafhankelijke variabele en gestandaardiseerde kosten per



Figuur 1: Strooidiagram van de door de Energiekamer gebruikte data

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
aansluiting_perm2	1	0.0161	0.0161	24.911	0.002
Residuals	7	0.0045	0.0006		

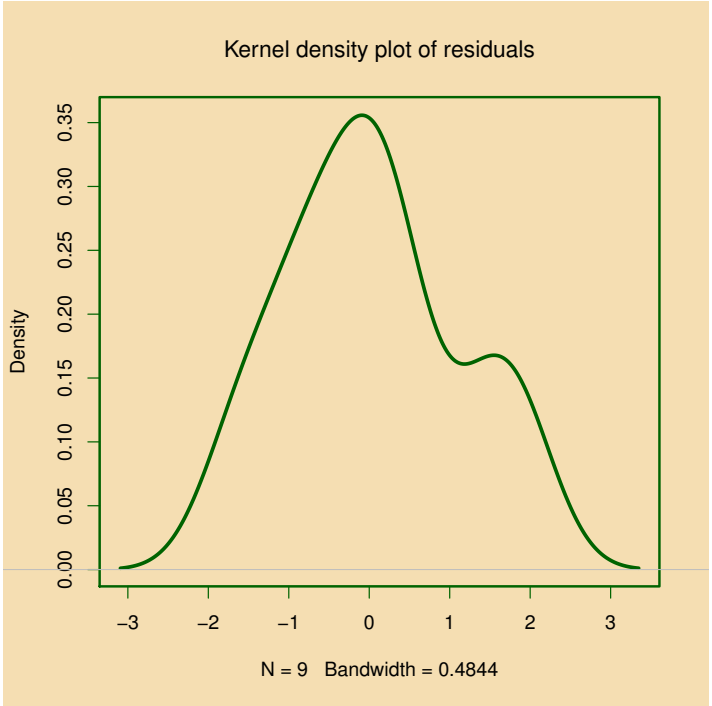
Residual standard error: 0.025 on 7 degrees of freedom
Multiple R-squared: 0.781, Adjusted R-squared: 0.749
F-statistic: 24.911 on 1 and 7 degrees of freedom, p-value: 0.00

Tabel 2: ANOVA-tabel voor lineair regressiemodel

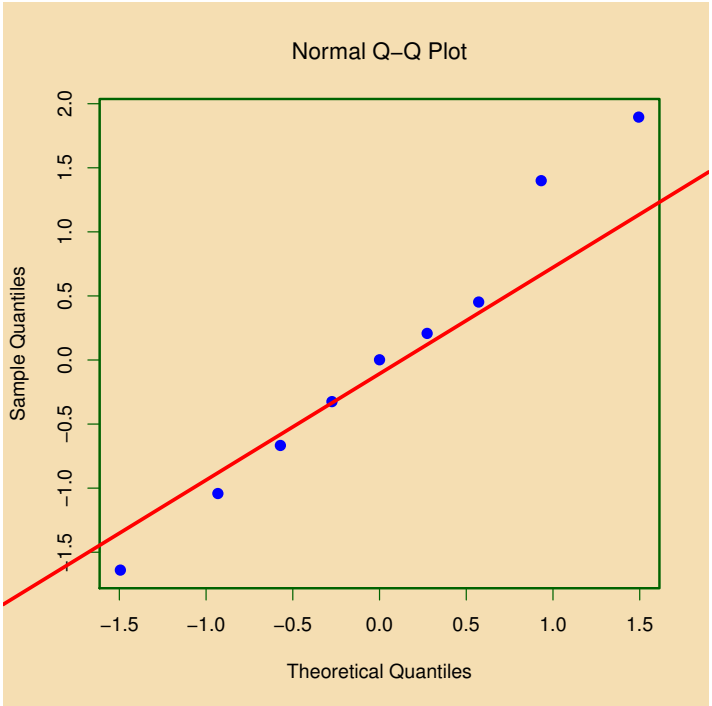
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.244731	0.012312	19.9	0.000
aansluiting_perm2	-0.000065	0.000013	-5.0	0.002

Tabel 3: Tabel van regressiecoëfficiënten van lineair regressiemodel

eenheid output als afhankelijke variabele. De resultaten van deze analyse zijn weergegeven in de Tabellen 2 en 3. Uit Tabel 2 zien we o.a. dat het regressiemodel significant is (p-waarde voor de helling is kleiner dan 0,05) en dat er ongeveer 75% van de spreiding in de data verklaard wordt. Tabel 3 levert zowel schattingen voor de regressiecoëfficiënten (intercept en helling) als een maat voor de nauwkeurigheid van deze schattingen (standaardafwijkingen). Het is echter onjuist deze gegevens zonder nadere inspectie te gebruiken. Een regressie-analyse is gebaseerd op veronderstellingen m.b.t. waarnemingen (onafhankelijkheid van waarnemingen, normaliteit van de waarnemingen, gelijkheid van spreiding over het hele gebied). Verder kan OLS verkeerde uitkomsten geven indien er geëxtrapoleerd of geïnterpoleerd wordt. Om

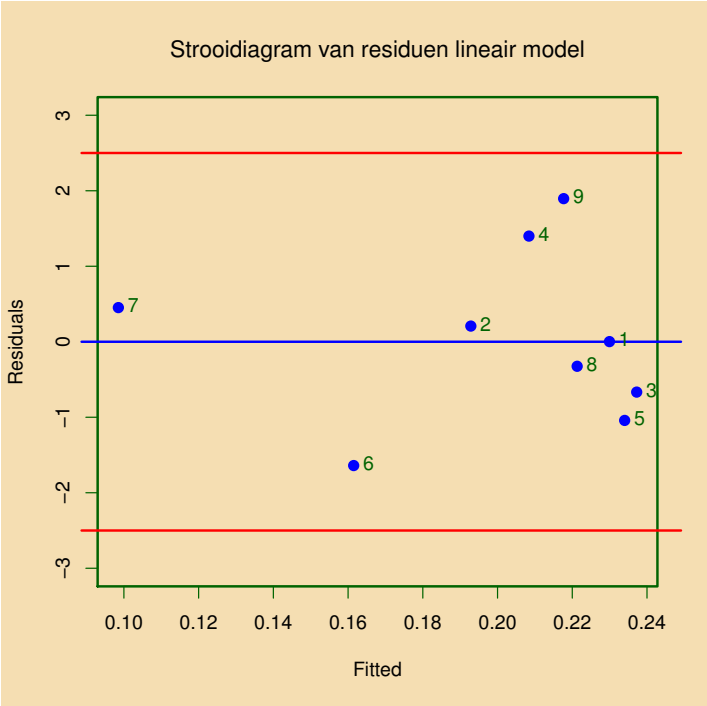


Figuur 2: Density plot van residuen van lineair regressiemodel



Figuur 3: Normal probability plot van lineair regressiemodel

hier inzicht in te krijgen is het noodzakelijk (en daarom ook een standaardprocedure in de



Figuur 4: Strooidiagram van residuen lineair regressiemodel

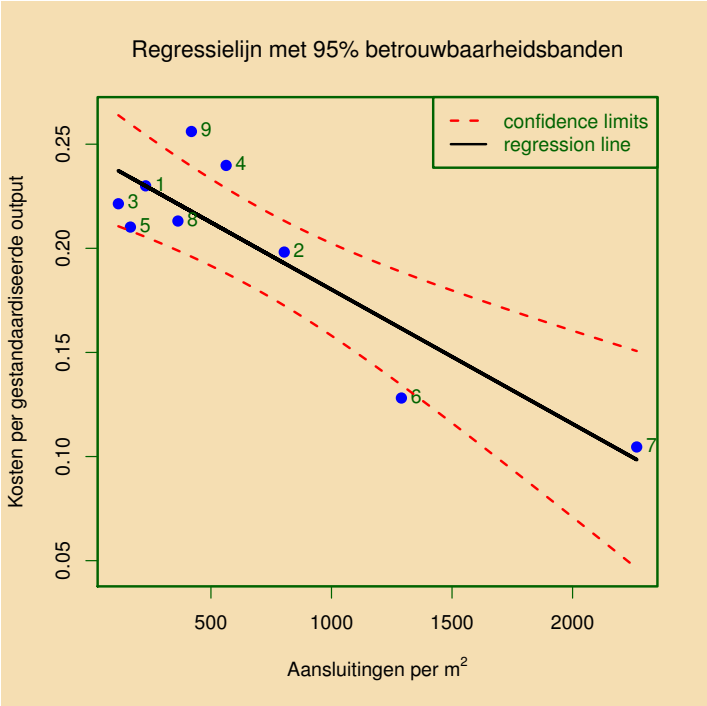
toegepaste statistiek) elk regressiemodel te onderwerpen aan een aantal diagnostische toetsen (zowel grafisch als via statistische kentallen / toetsen). Zo zijn alle uitspraken over kansen en significantie gebaseerd op de aanname van normaliteit (kansverdeling). Grafische toetsen op normaliteit zijn te zien in de Figuren 2 (moet zoveel mogelijk een mooie klokkromme zijn) en 3 (de datapunten moeten zoveel mogelijk op de rechte lijn liggen). In beide grafieken zien we afwijkingen. Op zich is dat nog geen probleem, want bij kleine aantallen waarnemingen zullen er altijd afwijkingen te zien zijn. Ook in Figuur 4 (een grafiek van gestudentiseerde residuen, d.w.z. gestandaardiseerde verschillen tussen datapunten en de regressielijnen) zien we geen grote afwijkingen omdat alle waarden binnen de gebruikelijke $\pm 2{,}5$ grenzen liggen. Om objectief normaliteit te toetsen is voor de volledigheid de bekende toets van Shapiro-Wilk uitgevoerd (zie bijv. [DS98] of [MP92]). Zoals te verwachten, gezien het bovenstaande, geeft

```
  Shapiro-Wilk normality test

data:  residuals linear model
W = 0.978    p-value = 0.953
```

Tabel 4: Uitkomst van de toets van Shapiro-Wilk

de toets van Shapiro-Wilk geen significante afwijking van normaliteit aan (p-waarde is groter dan 0,05).
In Figuur 5 is het gefitte lineaire regressiemodel getekend met daarbij 95% betrouwbaarheidsgrenzen. Hierbij valt al meteen op dat liefst 3 van de 9 netwerkbeheerders niet binnen de 95% betrouwhaarheidsgrenzen liggen. Vooral netwerkbeheerder 9 heeft een verrassend hoge waarde. Verder valt op dat bij de netwerkbedrijven 6 en 7 de betrouwbaarheid van het re-



Figuur 5: Strooidiagram van lineair regressiemodel

gressiemodel veel kleiner is dan bij de overige netwerkbedrijven. Dit komt door de relatief grote afstand qua aansluitingen per m^2 van deze netwerkbedrijven ten opzichte van de punten van de andere netwerkbedrijven. Verder valt op dat de regressielijn vrijwel precies door het datapunt van netwerkbeheerder 7 gaat. Dat is een bekend verschijnsel in OLS wat bekend staat onder het fenomeen hefboompunt. Een OLS regressielijn neigt altijd sterk naar een ver verwijderde (qua waarde van de onafhankelijke variabele(n)) waarneming. Het gevolg is dat de regressielijn (i.h.b. de helling) sterk beïnvloed wordt door een dergelijke waarneming. Dit verschijnsel kan leiden tot onterechte conclusies en dient daarom altijd onderzocht te worden. In de statistiek gebruikt men de "leverage" grootheid om potentieel invloedrijke punten te onderzoeken (zie bijv. de standaardwerken [CW94], [DS98] en [MP92]). Daarnaast zijn er grootheden die aangeven of zulke waarnemingen ook werkelijk invloed hebben op bijvoorbeeld de modelschattingen (DFFITS), op de vector van regressiecoëfficiënten (Cook's D) of op een indivuele regressiecoëfficiënt (DFBETAS, hier nemen we alleen de helling omdat de intercept niet van belang is).

Observation	Leverage	DFFITS	DFBETAS	Cook's D
7	0.75	0.79	0.73	0.35

Tabel 5: Waarnemingen die de regressielijn (potentieel) te veel beinvloeden

Uit Tabellen 5 en 6 zien we inderdaad dat waarneming 7 een potentieel invloedrijk punt is met bijna significante waarden voor de invloedsmaten Cook's D en DFFITS en een significante waarde voor de DFBETAS van de helling. Hierbij dient opgemerkt te worden dat de grenswaarden in Tabel 6 geen exacte waarden zijn maar benaderingen die gelden voor

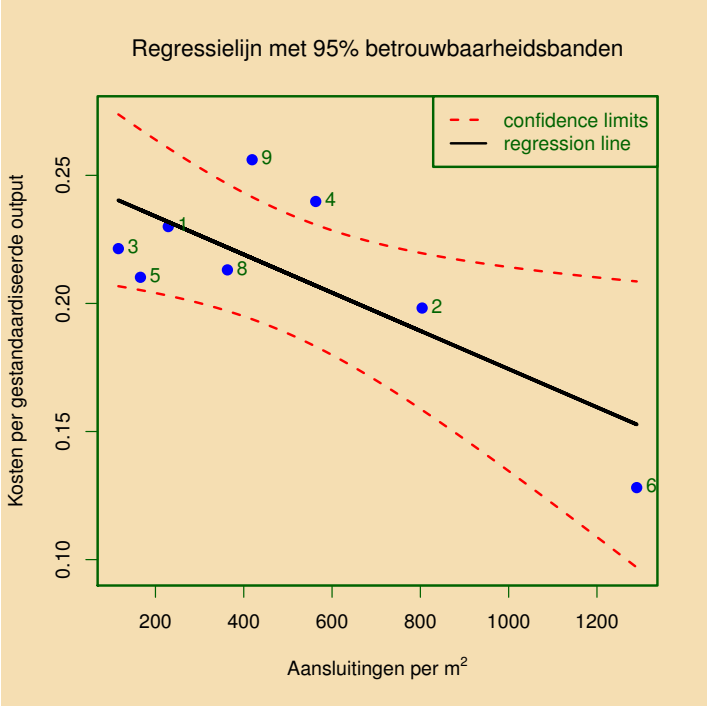
Leverage	DFFITS	DFBETAS	Cook's D Test
0.44	0.94	0.67	0.44

Tabel 6: Grenswaarden van grootheden voor regressiediagnostiek voor dataset met 9 waarnemingen

grote datasets. De conclusie is dat onnauwkeurigheden in de bepaling van de gestandaardiseerde kosten per eenheid output voor deze netwerkbeheerder grote gevolgen hebben voor de regressielijn.
De Energiekamer heeft ook een kwadratisch en een loglineair regressiemodel gefit. Deze modellen hebben dezelfde problemen als het lineaire model dat in bovenstaande analyse onderzocht is. Het is daarom niet zinvol om deze modellen met een informatiecriterium te vergelijken. Overigens is het AIC (Akaike Informatie Criterium) gebruikelijker dan het door de Energiekamer gebruikte Schwartz Bayesiaanse Criterium (ook wel BIC = Bayesiaans Informatie Criterium genoemd).

Regressiemodel zonder netwerkbeheerder 7

We herhalen nu de analyse als we netwerkbeheerder 7 weghalen (zie Figuur 6). De reden hiervoor is niet alleen het feit dat deze netwerkbeheerder in belangrijke mate de helling bepaalt (hefboompunt), maar ook dat deze netwerkbeheerder niet meer actief is. Uit Tabellen 7 en



Figuur 6: Strooidiagram van lineair regressiemodel zonder netwerkbeheerder 7

8 zien we dat de regressie nog steeds significant is (de p-waarde is nog steeds kleiner dan 0,05, maar hij is nu wel veel groter geworden) en dat de helling weliswaar groter is geworden maar

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
aansluiting_perm2	1	0.0060	0.0060	8.204	0.029
Residuals	6	0.0044	0.0007		

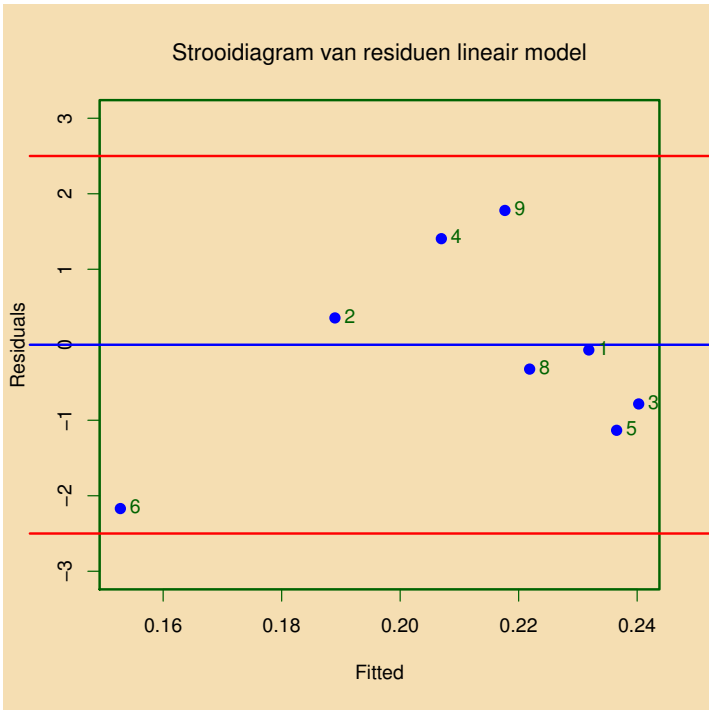
Residual standard error: 0.027 on 6 degrees of freedom
Multiple R-squared: 0.578, Adjusted R-squared: 0.507
F-statistic: 8.204 on 1 and 6 degrees of freedom, p-value: 0.03

Tabel 7: ANOVA-tabel voor lineair model zonder netwerkbeheerder 7

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.248906	0.016007	15.5	0.000
aansluiting_perm2	-0.000075	0.000026	-2.9	0.029

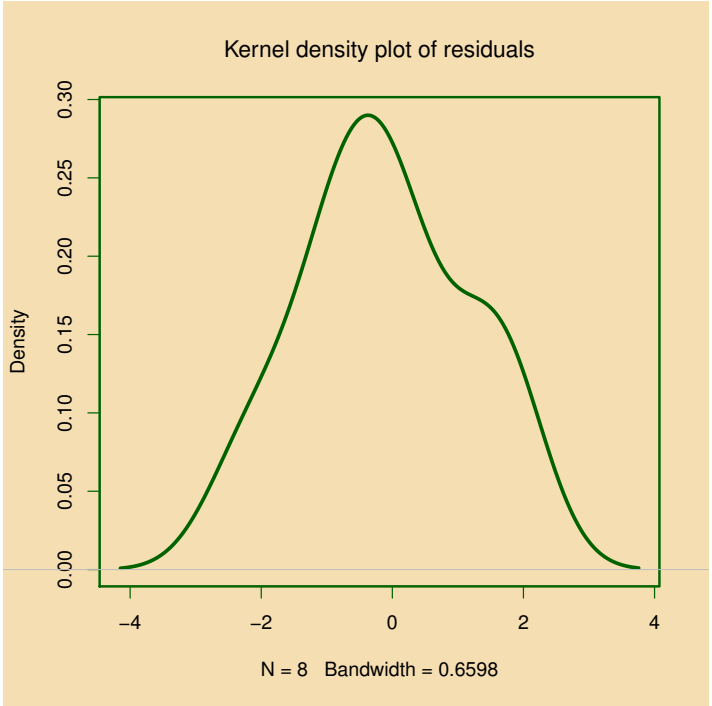
Tabel 8: Tabel van regressiecoëfficiënten van lineair model zonder netwerkbeheerder 7

ook duidelijk onnauwkeuriger (de standaardafwijking is verdubbeld). Ook valt op dat het regressiemodel nog slechts 51% van de spreiding in de data verklaard wordt. Net als in de analyse gebaseerd op de volledige data, vallen de relatief hoge waarden voor netwerkbeheerders 4 en 9 op (zie Figuur 6). De grafische toetsen op normaliteit (Figuren 7, en 8, 9) tonen net

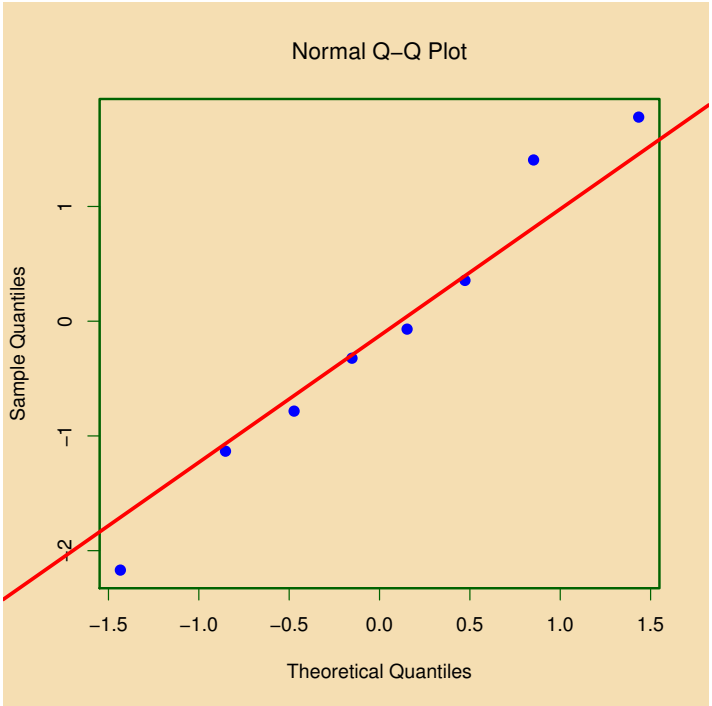


Figuur 7: Strooidiagram van residuen lineair regressiemodel zonder netwerkbeheerder 7

als in het geval van de analyse op de volledige dataset geen grote problemen. Om objectief normaliteit te toetsen is voor de volledigheid weer de toets van Shapiro-Wilk uitgevoerd. Ook hier is geen probleem te constateren, omdat de p-waarde weer groter dan 0,05 is.
De diagnostiek gebaseerd op invloedsmaten (zie Tabel 10) geeft aan dat netwerkbeheerder



Figuur 8: Density plot van residuen lineair regressiemodel zonder netwerkbeheerder 7



Figuur 9: Normal probability plot van lineair regressiemodel zonder netwerkbeheerder 7

6 een significant hefboompunt is (de grenswaarden zijn nu gebaseerd op 8 waarnemingen en

```
  Shapiro-Wilk normality test

data:  residuals linear model
W = 0.977    p-value = 0.948
```

Tabel 9: Uitkomst van de toets van Shapiro-Wilk

Observation	Leverage	DFFITS	DFBETAS	Cook's D
6	0.71	-3.42	-3.11	3.62

Tabel 10: Waarnemingen die de regressielijn (potentieel) te veel beinvloeden (zonder netwerkbeheerder 7)

Leverage	DFFITS	DFBETAS	Cook's D Test
0.50	1.00	0.71	0.50

Tabel 11: Grenswaarden van grootheden voor regressiediagnostiek voor dataset met 8 waarnemingen

zijn te vinden in Tabel 11). In feite wordt de helling nu vrijwel volledig bepaald door de waarde van netwerkbeheerder 6. Dit betekent dat een onjuiste of onnauwkeurige bepaling van de gestandaardiseerde kosten per eenheid output van netwerkbeheerder 6 grote gevolgen heeft voor de helling. M.a.w., het is de waarde van netwerkbeheerder 6 die bepaalt of de gestandaardiseerde kosten per eenheid output een kostendriver is. Dit is vanzelfsprekend een zeer ongewenste situatie.

CONCLUSIES

Het lineaire regressiemodel toont ernstige tekortkomingen vanwege de uitzonderlijke posities van twee netwerkbedrijven (hefboomeffect). De aansluitdichtheden van deze netwerkbedrijven wijken sterk af van de andere netwerkbedrijven en beïnvloeden daardoor in te grote mate de helling van de regressielijn. De uitzonderlijke positie van deze twee netwerkbedrijven in combinatie met het geringe aantal datapunten maakt in feite het fitten van elk OLS regressiemodel onmogelijk. De situatie is nog ernstiger als het datapunt van de niet meer bestaande netwerkbeheerder 7 weggelaten wordt. Dan bepaalt de waarde van de gestandaardiseerde kosten per eenheid output van netwerkbeheerder 6 volledig of er al dan niet sprake van zou zijn dat aansluitdichtheid correleert met gestandaardiseerde kosten per eenheid output. Tenslotte dient opgemerkt te worden dat een regressiemodel nooit een causaal verband kan valideren (zo is er een statistisch zeer significant verband tussen het aantal ooievaars en het aantal geboorten; de causale verklaring ligt in het verband tussen het aantal schoorstenen van huizen en het aantal gezinnen). Eventuele statistische verbanden tussen aansluitdichtheid en gestandaardiseerde kosten per eenheid output leveren dus geen causale verklaring op. M.a.w., op deze manier kan men nooit aantonen dat de gestandaardiseerde kosten per eenheid output verklaard worden door de aansluitdichtheid.

GERAADPLEEGDE BRONNEN

Energiekamer, Ontwerp-methodebesluit vijfde reguleringsperiode regionale netbeheerders elektriciteit, geraadpleegd op 19 april 2010 als document
"`http://www.energiekamer.nl/images/Ontwerp-methodebesluit_tcm7-135457.pdf`"

Excelbestand "`102382 Regressieanalyse aansluitdichtheid voor ontwerp NE5R.xls`", geraadpleegd op 19 april 2010 op `www.energiekamer.nl`

Referenties

[CW94] Cook, R.D. en S. Weisberg: *An Introduction to Regression Graphics*. John Wiley & Sons Inc., New York, 1994.

[DS98] Draper, N.R. en H. Smith: *Applied Regression Analysis*. John Wiley & Sons Inc., New York, derde uitgave, 1998.

[MP92] Montgomery, D.C. en E.A. Peck: *Introduction to Linear Regression Analysis*. John Wiley & Sons Inc., New York, tweede uitgave, 1992.