Verbetering van die voorspellingsakkuraatheid van regressiemodelle met minimale aannames

Magderie van der Westhuizen, Giel Hattingh en Hennie Kruger

M.M. van der Westhuizen, J.M. Hattingh en H.A. Kruger, Skool vir Rekenaar-, Statistiese en Wiskundige Wetenskappe, Noordwes-Universiteit, Potchefstroomkampus

Opsomming

Die voorspellingsakkuraatheid van 'n regressiemodel maak in 'n groot mate staat op die toepaslikheid van die modelbouer se aannames. Daarbenewens kan die teenwoordigheid van uitskieters ook tot modelle lei wat onbetroubaar en dus minder robuust is. In hierdie artikel word 'n regressiemodel wat op minimale aannames gebaseer is, bestudeer en uitgebrei in 'n poging om voorspellingsakkuraatheid te verbeter. Die voorgestelde uitbreidings sluit uitskieteropsporing in wat op wiskundige programmeringstegnieke gebaseer is, asook 'n gladstrykingstegniek wat gebruik word om die koers van verandering in die rigting van 'n funksie te beheer. Die voorgestelde modelleringstegnieke word dan op vier welbekende datastelle uit die literatuur toegepas om hul voorspellingsakkuraatheid te illustreer en te evalueer. Die resultate toon dat die twee uitbreidings die voorspellingsvermoë van die oorspronklike minimale-aanname-regressiemodel (soos deur die gemiddelde absolute afwyking gemeet) aansienlik verbeter het. Die resultate vergelyk ook gunstig met ander modelle, soos stuksgewyse lineêre regressiemodelle.

Trefwoorde: lineêre programmering; robuuste modelle; uitskieteropsporing; stuksgewyse lineêre regressie

Abstract

Improving the predictive accuracy of regression models with minimal assumptions

The forecasting accuracy of a regression model relies heavily on the applicability of the assumptions that have been made by the model builder. In addition, the presence of outliers may also lead to models that are not reliable and thus less robust. In this paper a regression model based on minimal assumptions is considered and extended in an effort to improve forecasting accuracy. The proposed extensions include outlier detection that is based on mathematical programming techniques and a smoothing technique that is used to control the rate of change in direction of a function. The suggested modelling techniques are then applied to four well-known data sets from the literature to illustrate and evaluate their forecasting accuracy. The results show that the two extensions have significantly improved the prediction capability of the original minimal assumption regression model (as measured by the mean absolute deviation). The results also compare favourably with those of other models, such as piecewise linear regression models.

This study considers an existing minimal assumption regression model that was proposed in the literature in 1962 (Wagner 1962). Two extensions (outlier detection and smoothing) are added to the model to improve robustness and predictive capability.

The minimal assumption approach requires the fitting of regression functions according to the criteria of the minimal sum of absolute deviations, but without specifying a mathematical form for the functions to be estimated (Wagner 1962). The only restrictive assumptions that are needed are additivity and monotonicity of the individual functions, that is, the regression function expresses the response variable as the sum of single variable functions that are assumed to be monotonically non-increasing or non-decreasing. These are the only assumptions that have to be made and in this sense, the model employs minimal assumptions.

The approach thus assumes an additive regression model of the form

$$y = \sum_{j=1}^{k} f_j(x_j) + error, \tag{1}$$

with y the dependent variable and $x_{j,}\ j = 1, 2, \cdots, k$, the predictor variables. Assume that n observations on the variables y and x_j are available, given by $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ for $i = 1, 2, \cdots, n$. The model now aims to determine estimators of function values $f_j(x_{ij})$, which are abbreviated as f_{ij}, from this data, such that estimates $\hat{y}_i = \sum_j f_j(x_{ij})$ of the response are optimal in the L_1-norm sense.

The estimates f_{ij} are obtained by solving the following linear programme:

$$\text{Minimize} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}), \tag{2}$$

$$\text{subject to} \quad \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \text{for } i = 1, \cdots, n, \tag{3}$$

$$f_{tj} \leq f_{lj}, \text{if } r_{tj} \leq r_{lj}, \text{and}$$
$$f_{tj} = f_{lj}, \text{if } r_{tj} = r_{lj}, \text{for } t, l = 1,2, \ldots, n \text{ with } t \neq l \text{ and } j = 1,2, \ldots, k, \tag{4}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \text{for } i = 1, \cdots, n, \tag{5}$$

where r_{sj} is the rank of x_{sj} in the set $x_{1j}, \ldots, x_{nj}$, and f_{ij} is unrestricted in sign for all i and j.

Not all constraints in (4) and (5) are necessary when the model is implemented, since it is sufficient to impose the inequalities for $r_{lj} = r_{tj} + 1$ and equalities for the others.

Given the model, it is still necessary to decide the direction of monotonicity for each function. One way to approach this problem is to perform a (least squares) multiple linear

regression beforehand and use the signs of the estimated coefficients to estimate whether a function should be restricted to be non-increasing or non-decreasing.

To improve robustness and the predictive capability of the given model, two extensions are added to the model. The first extension is intended to detect possible outliers by implementing mixed integer linear programming techniques. The second extension addresses the potential problem of overfitting by using constrained second derivatives to smooth the functions.

To provide for possible outlier detection, the minimal assumption model was adapted as follows: Constraint (3) was changed to

$$\sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} - \alpha_i = y_i, \qquad \text{for } i = 1, \cdots, n, \text{ (6)}$$

where α_i is an unrestricted slack variable. An additional constraint was added to constrain the absolute value of α_i by Mz_i where M is a large number and z_i is a binary variable. The constraint is formulated as

$$-Mz_i \leq \alpha_i \leq Mz_i, \qquad \text{for } i = 1, \cdots, n. \text{ (7)}$$

In experiments a value of M larger than the span of the y_i values proved sufficient. If z_i is zero, α_i is also constrained to zero and the ith absolute residual contributes to the objective. However, if z_i is one, the optimisation process will choose the ith residual to be zero, since α_i takes up the slack.

To specify the number of data points (outliers) to be omitted, the following constraint was also added to the model:

$$\sum_{i=1}^{n} z_i = p \text{ with } z_i \in \{0,1\} \qquad \text{(8)}$$

In this study the value of p is determined by experimentation.

The second extension to the minimal assumption regression model addresses the problem of possible overfitting of the model. Overfitting takes place when a function fits a data set "too well", which makes the model very sensitive to the behaviour of a specific data set. It is a serious problem, because it may affect the prediction capability of a model and make it less reliable (Hitchcock en Sober 2004). While large data sets may reveal relatively smooth functions $f_j(x_{ij})$, small data sets may show sudden, large fluctuations.

The smoothing technique used is intended to constrain the second derivative of the function, in other words, the rate of change in direction. The slope of a function cannot change more than a specified value and this constrains sudden large fluctuations in the slope.

To implement the smoothing of a function, specific constraints are added to the model. Constrained second derivatives are employed in these constraints and can be described as follows:

Set $f_j(x_j) = f_j$ and consider

$$\frac{\partial f_j}{\partial x_j}|_{x_{i,j}} \approx \frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}},$$

$$\frac{\partial^2 f_j}{\partial x_j^2}|_{x_{i,j}} \approx \frac{\frac{\partial f_j}{\partial x_j}|_{x_{i,j}} - \frac{\partial f_j}{\partial x_j}|_{x_{i-1,j}}}{x_{i,j} - x_{i-1,j}}, \text{and}$$

$$-\beta \leq \frac{\frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}} - \frac{f_j(x_{i,j}) - f_j(x_{i-1,j})}{x_{i,j} - x_{i-1,j}}}{x_{i,j} - x_{i-1,j}} \leq \beta, \quad (9)$$

where $\approx$ denotes an approximation.

The absolute rate of change in direction (the second derivative) is now constrained by the parameter β.

To be able to obtain an alternative mathematical model and to compare the results of the proposed techniques, a piecewise linear regression model is introduced. Piecewise linear regression is a form of regression that allows multiple linear models to be fitted to data for different ranges of x (Ryan en Porth 2007). Breakpoints are the values of x where the slope of the linear function changes. The value of a breakpoint may or may not be known before the analysis, but it is typically unknown and must be estimated. Data sets in this study are modelled either as one linear regression model or as piecewise linear continuous segments, each represented by a linear model. The implementation of the piecewise linear regression models in this research project was also done through the use of a linear programming model.

To illustrate and evaluate the forecasting accuracy of the proposed models and extensions, four well-known data sets were considered. In each case, the mean absolute deviation was used as a measure of performance and was calculated by using a “leave-one-out” jackknife approach.

The results showed that the suggested two extensions to the minimal assumption regression model proved to be successful. The mean absolute deviation was considerably reduced after implementation of the extensions. In some cases it was possible to reduce the mean absolute deviation further by introducing piecewise linear regression models. These improvements were, however, only marginally better than the proposed extended models, and it seems reasonable to draw the conclusion that there are cases where the minimal assumption regression model, and the extensions thereof, will yield better results than other models.

Keywords: linear programming; robust models; outlier detection; piecewise linear regression

1. Inleiding

In 'n poging om die robuustheid en voorspellingsvermoë van regressiemodelle te verbeter, implementeer hierdie studie twee uitbreidings (uitskietopsporing en gladstryking) tot bestaande minimale-aanname-regressiemodel wat in die literatuur voorgestel is (Wagner 1962).

Die minimale-aanname-benadering vereis die passing van regressiefunksies volgens die kriterium van die minimum som van absolute afwykings, maar sonder om 'n wiskundige vorm te spesifiseer vir die funksies wat beraam moet word (Wagner 1962). Die enigste beperkende aannames wat benodig word, is additiwiteit en monotonisiteit van die individuele funksies; met ander woorde, die regressiefunksie druk die responsveranderlike uit as die som van enkelveranderlike funksies waarvan daar aangeneem word dat hulle monotoon nietoenemend of nie-afnemend is. Dit is die enigste aannames wat gemaak hoef te word en in hierdie sin gebruik die model minimale aannames.

Die benadering neem dus 'n additiewe regressiemodel aan van die vorm

$$y = \sum_{j=1}^{k} f_j(x_j) + fout, \tag{1}$$

met y die afhanklike veranderlike en $x_{j,}\, j = 1, 2, \cdots, k$, die voorspellerveranderlikes. Veronderstel dat n waarnemings van die veranderlikes y en x_j beskikbaar is, gegee deur $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ vir $i = 1, 2, \cdots, n$. Die model beoog nou om beramers van funksiewaardes $f_j(x_{ij})$, wat as f_{ij} afgekort word, vanuit hierdie data te bepaal, sodat beramings $\hat{y}_i = \sum_j f_j(x_{ij})$ van die respons optimaal in die L_1-norm-sin is.

Die beramings f_{ij} word verkry deur die volgende lineêre program op te los:

$$\text{Minimaliseer} \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}), \tag{2}$$

onderhewig aan

$$\sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \text{vir } i = 1, \cdots, n, \tag{3}$$

$$f_{tj} \leq f_{lj}, \text{indien } r_{tj} \leq r_{lj}, \text{en} \\ f_{tj} = f_{lj}, \text{indien } r_{tj} = r_{lj}, \text{vir } t, l = 1,2, \ldots, n, \text{met } t \neq l \text{ en } j = 1,2, \ldots, k, \tag{4}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \text{vir } i = 1, \cdots, n, \tag{5}$$

(6)

waar r_{sj} die rangnommer van x_{sj} in die stel, $x_{1j}, \ldots, x_{nj}$, is en f_{ij} onbeperk in teken is vir alle i en j.

In hierdie navorsing word die f_{ij}-waardes wat verkry is, gebruik om vir elke *j*, oor die gebied waaroor die waarneming strek, 'n stuksgewyse lineêre funksie te konstrueer. Hierdie funksie word dan gebruik om beramings van f_{ij} te verkry deur lineêre interpolasie of ekstrapolasie te doen vir daardie x_{ij}-waardes wat nie in die gegewe data voorgekom het nie.

Die eerste uitbreiding se doel is om moontlike uitskieters op te spoor en is geïmplementeer deur gemengde heeltallige lineêre programmeringstegnieke te gebruik. Beperking (3) is verander na

$$\sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} - \alpha_i = y_i, \qquad \text{vir } i = 1, \cdots, n, \text{ (7)}$$

waar α_i 'n onbeperkte spelingveranderlike is. 'n Bykomende beperking is bygevoeg om die absolute waarde van α_i deur Mz_i te beperk waar M 'n groot getal en z_i 'n binêre veranderlike is. Die beperking word geformuleer as

$$-Mz_i \leq \alpha_i \leq Mz_i, \qquad \text{vir } i = 1, \cdots, n. \text{ (8)}$$

In eksperimente is 'n waarde van M wat groter as die spanwydte van die y_i-waardes is, as voldoende bewys. Indien z_i nul is, is α_i ook tot nul beperk en die i-de absolute residu dra tot die doelfunksie by. Indien z_i egter 1 is, sal die optimeringsproses die i-de residu as nul kies, aangesien α_i die speling opneem.

Ten einde die aantal datapunte (uitskieters) te spesifiseer wat weggelaat moet word, is die volgende beperking ook tot die model gevoeg:

$$\sum_{i=1}^{n} z_i = p, \text{met } z_i \in \{0,1\}. \qquad \text{(9)}$$

In hierdie studie word die waarde van p deur eksperimentering bepaal.

Die tweede uitbreiding tot die bestaande model takel die probleem van moontlike oormatige passing, en 'n gladstrykingstegniek word gebruik om die tweede afgeleide van 'n funksie, met ander woorde, die rigtingsveranderingskoers, te beperk. Hierdie beperkinge is soos volg geformuleer:

$$-\beta \leq \frac{\frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}} - \frac{f_j(x_{i,j}) - f_j(x_{i-1,j})}{x_{i,j} - x_{i-1,j}}}{x_{i,j} - x_{i-1,j}} \leq \beta \qquad \text{(10)}$$

Die absolute rigtingsveranderingskoers (die tweede afgeleide) word nou deur die parameter β beperk.

Ten einde 'n alternatiewe wiskundige model te verkry en die resultate van die voorgestelde tegnieke te vergelyk word 'n stuksgewyse lineêre regressiemodel ingevoer. Datastelle in

hierdie studie word óf as een lineêre regressiemodel gemodelleer, óf as stuksgewyse lineêre deurlopende segmente waarvan elkeen deur 'n lineêre model verteenwoordig word. Die implementering van die stuksgewyse lineêre regressiemodelle in hierdie navorsingsprojek is ook gedoen deur van 'n lineêre programmeringsmodel gebruik te maak.

Vier bekende datastelle is beskou om die voorspellingsakkuraatheid van die voorgestelde modelle en uitbreidings te illustreer en te evalueer. In elke geval is die gemiddelde absolute afwyking as 'n prestasiemaatstaf gebruik en is dit bereken deur van 'n "laat -een-weg"-uitsnitbenadering gebruik te maak.

Die volgende moet vir elke datastel bepaal word voordat die modelle opgelos kan word:

1. *Rigting van monotonisiteit*. Dit kan gedoen word deur 'n gewone kleinste kwadrate meervoudige lineêre regressie uit te voer en dan die tekens van die beraamde koëffisiënte te gebruik om die rigting te beraam.

2. *'n Waarde vir p* (aantal uitskieters wat weggelaat moet word). Een wyse waarop 'n waarde vir *p* gespesifiseer kan word, is om die model herhaaldelik op te los deur met *p*=0 te begin en dan elke keer wanneer die model opgelos is, *p* met 1 te inkrementeer. 'n "Beste" waarde vir *p* kan nou bepaal word, gebaseer op die veranderingskoers in die doelfunksiewaarde. Alternatiewelik kan 'n waarde vir *p* geselekteer word om byvoorbeeld 10–20% van die datapunte te elimineer.

3. *'n Waarde vir* β (veranderingskoers in 'n funksie). Hierdie waarde kan ook eksperimenteel bepaal word deur die model herhaaldelik vir verskillende waardes van β op te los en die β-waarde te kies wat die kleinste gemiddelde absolute afwyking oplewer.

Die resultate het getoon dat die voorgestelde twee uitbreidings tot die minimale-aanname-regressiemodel suksesvol bewys is. Nadat die uitbreidings geïmplementeer is, is die gemiddelde absolute afwyking aansienlik gereduseer. In sommige gevalle was dit moontlik om die gemiddelde absolute afwyking verder te reduseer deur stuksgewyse lineêre regressiemodelle in te voer. Hierdie verbeteringe was egter net marginaal beter as die voorgestelde uitgebreide modelle en dit is redelik om tot die gevolgtrekking te kom dat daar gevalle is waar die minimale-aanname-regressiemodel, asook die uitbreidings daarvan, beter resultate as ander modelle sal lewer.

2. Agtergrond

Die suksesse of mislukkings wat deur bestuurders in die sakewêreld ervaar word, is grootliks afhanklik van die gehalte van die besluite wat hulle neem. Die gehalte van 'n besluit is in groot mate op die evaluasie en interpretasie van data gebaseer. 'n Goeie besluit is een wat op logika gebaseer is, al die beskikbare data in ag neem en in baie gevalle 'n kwantitatiewe benadering toepas. Een van die gewildste en waardevolste tegnieke wat aan hierdie vereistes voldoen, is regressieanalise. Die doel hiervan is om die verwantskap tussen verskillende veranderlikes te verstaan en die waarde van een veranderlike vooruit te beraam, gebaseer op waarnemings van die ander. Resultate kan dan gebruik word om die

besluitnemingsproses te lei en om bestuurders in staat te stel om meer gepaste en ingeligte besluite te neem.

Die klassieke lineêre regressiemodel word soos volg voorgestel:

$$y = \mathbf{X}\beta + \varepsilon, \tag{1.1}$$

waar y 'n $n \times 1$-responsvektor van waargenome waardes is, $\mathbf{X}$ 'n $n \times k$-gegewe matriks van waarnemings van voorspellers (regressors) is, waar elke kolomvektor met 'n voorspeller ooreenkom, β 'n $k \times 1$-vektor van onbekende parameters is en ε 'n $n \times 1$-vektor van (stogastiese) foute, ε_i, is.

Daar word aangeneem dat die foutterme onafhanklik verspreide kontinue stogastiese veranderlikes is, met $E(\varepsilon_i) = 0$ en $Var(\varepsilon_i) = \sigma^2 > 0$. β kan beraam word deur die kleinstekwadrate-foutkriterium te gebruik.

In die geval waar nielineêre meervoudige regressiemodelle oorweeg moet word, is dit dikwels moeilik om te besluit wat die vorm van die nielineariteit is. Die metode waarmee in hierdie artikel geëksperimenteer word, het die voordeel dat dit outomaties 'n "goeie" model soek.

Die sukses van 'n regressiemodel maak in 'n groot mate staat op die aannames wat deur die modelbouer gemaak word. Daar is 'n groot aantal literatuurbronne wat in groot besonderhede oor hierdie aannames handel, insluitend die niestogastiese en ongekorreleerde aard van onafhanklike veranderlikes, asook die normale verspreiding van foutveranderlikes, en die toereikende aard van die regressiefunksie. Die onderliggende aannames word deur Bowerman e.a. (Bowerman, O'Connell en Koehler 2005) soos volg aangegee:

- *Onafhanklikheidsaanname.* Enige een van die foutveranderlikes, ε_{i1}, is statisties onafhanklik van enige ander ε_{i2}.
- *Normaliteitsaanname*. Die foutterme het 'n normale verspreiding, gegewe enige kombinasie van waardes vir voorspellers.
- Die foutterme het gemiddeldes wat gelyk is aan nul.
- *Konstantevariansie-aanname*. Die foutterme het konstante variansies wat nie van die voorspellers se kombinasie van waardes afhanklik is nie.

'n Geval wat die robuustheid van 'n regressiemodel kan beïnvloed, is die moontlike teenwoordigheid van uitskieters in die data. Uitskieters kan gedefinieer word as waarnemings wat nie dieselfde model as die res van die data volg nie (Hoeting, Raftery en Madigan 1996), en die kuns van robuuste regressie is daarin geleë om beramers te ontwerp wat nie sterk deur uitskieters beïnvloed word nie (Rousseeuw en Leroy 2003). Die opsporing is dikwels kompleks en een van die faktore wat dit veroorsaak, is die moontlike teenwoordigheid van sogenaamde maskeringsprobleme, waarin sommige uitskieters in 'n datastel die teenwoordigheid van ander uitskieters kan verberg.

Ten einde die bogenoemde probleemareas aan te pak sal hierdie studie die bestaande minimale-aanname-regressiemodel (Wagner 1962) gebruik en sekere uitbreidings daarby voeg om die model se robuustheid en voorspellingsvermoë te verbeter. Die uitbreidings word geïmplementeer deur die gebruik van lineêre en gemengde heeltallige lineêre programmeringstegnieke en sluit gelyktydige uitskieteropsporing en gladstrykingstegnieke in. Om die resultate van die voorgestelde uitbreidings te vergelyk sal 'n stuksgewyse lineêre regressiemodel oorweeg word as alternatiewe metode om nielineêre regressiefunksies te beraam.

Die modelle wat in hierdie navorsingsprojek beskou word, is nou verwant aan die sogenaamde veralgemeende additiewe modelle. Hierdie modelle 'n uitbreiding van die klas van algemene lineêre modelle en word volgens Hastie en Tibshirani (1986) soos volg beskryf.

Die tradisionele lineêre regressiemodel (1.1) word vervang deur die som van gladde funksies wat voorgestel kan word deur

$$g_0^{-1}\big(E(Y)\big) = \alpha + f_1(X_1) + \cdots + f_k(X_k) + \varepsilon \tag{1.2}$$

met $E(\varepsilon) = 0$ en $Var(\varepsilon) = \sigma^2$. Die f_{ij}-funksies is ongespesifiseerd en kan beraam word volgens sekere algoritmes (sien byvoorbeeld Hastie en Tibshirani vir 'n bespreking van hierdie algoritmes). Die leser wat verder in hierdie tipe modelle geïnteresseerd is, word verwys na Rousseeuw en Leroy, Hastie en Tibshirani (1968), Hastie en Tibshirani (1987) en Hastie en Tibshirani (1990).

Die res van die artikel is soos volg gestruktureer: afdeling 3 bied 'n kort inleiding tot Wagner (1962) se minimale-aanname-regressiemodel aan, terwyl afdeling 4 die voorgestelde uitbreidings wat robuustheid moet hanteer, in besonderhede bespreek. In afdeling 5 word die stuksgewyse lineêre regressiemodel geformuleer en in afdeling 6 eksperimentele resultate met die voorgestelde modelle aangebied. In afdeling 7 word die studie met algemene opmerkings afgesluit.

3. 'n Minimale-aanname-regressiemodel

In 1962 het Harvey M. Wagner 'n benadering gepubliseer wat gepaste regressiefunksies vereis volgens die kriteria van die minimum som van absolute afwykings, maar sonder om 'n wiskundige vorm te spesifiseer vir die funksies wat beraam moet word (Wagner 1962). Die enigste beperkende aannames wat benodig word, is additiwiteit en monotonisiteit van die individuele funksies, met ander woorde, die regressiefunksie druk die responsveranderlike uit as die som van enkelveranderlike funksies waarvan daar aangeneem word dat hulle monotoon nietoenemend of nie-afnemend is. Dit is die enigste aannames wat gemaak hoef te word en in hierdie sin gebruik die model minimale aannames.

Wagner veronderstel dus 'n additiewe regressiemodel van die vorm

$$y = \sum_{j=1}^{k} f_j(x_j) + fout, \tag{2.1}$$

met y die afhanklike veranderlike en $x_j, j = 1, 2, \cdots, k$, die voorspellerveranderlikes. Veronderstel dat n waarnemings van die veranderlikes, y en x_j, beskikbaar is, gegee deur $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ vir $i = 1, 2, \cdots, n$. Wagner se model beoog nou om beramers van funksiewaardes, $f_j(x_{ij})$, wat as f_{ij} afgekort word, vanuit hierdie data te bepaal, sodat beramings, $\hat{y}_i = \sum_j f_j(x_{ij})$, van die respons optimaal in the L_1-norm-sin is.

Die beramings, f_{ij}, word verkry deur die volgende lineêre program op te los:

$$\text{Minimaliseer} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}), \tag{2.2}$$

$$\text{onderhewig aan} \quad \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \text{vir } i = 1, \cdots, n, \tag{2.3}$$

$$f_{tj} \leq f_{lj}, \text{indien } r_{tj} \leq r_{lj}, \text{en} \tag{2.4}$$

$$f_{tj} = f_{lj}, \text{indien } r_{tj} = r_{lj}, \text{vir } t, l = 1,2, \ldots, n, \text{met } t \neq l \text{ en } j = 1,2, \ldots, k, \tag{2.5}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \text{vir } i = 1, \cdots, n, \tag{2.6}$$

waar r_{tj} die rangnommer van x_{tj} in die stel, $x_{1j}, \ldots, x_{nj}$, is en f_{ij} onbeperk in teken is vir alle i en j.

(Let daarop dat nie alle beperkinge in (2.4) en (2.5) nodig is wanneer die model geïmplementeer word nie, aangesien dit voldoende is om die ongelykhede vir $r_{lj} = r_{tj} + 1$ voor te skryf en gelykhede vir die ander.)

'n Gedetailleerde bespreking van die minimale-aanname-regressiemodel kan in Wagner (1962) gevind word.

Let daarop dat dit nog steeds nodig is om op die rigting van monotonisiteit vir elke funksie te besluit. Een wyse waarop hierdie probleem benader kan word, is om voor die tyd 'n (kleinstekwadrate) meervoudige lineêre regressie uit te voer en die tekens van die beraamde koëffisiënte te gebruik om te beraam of 'n funksie tot nietoenemend of nie-afnemend beperk moet word.

4. Modelontwikkeling

Hierdie afdeling beskryf twee uitbreidings wat tot die minimale-aanname-regressiemodel (deur Wagner voorgestel) gemaak is en in afdeling 3 aangebied is. Die eerste uitbreiding (afdeling 4.1) is bedoel om moontlike uitskieters op te spoor deur gemengde heeltallige

lineêre programmeringstegnieke (MHLP-tegnieke) te implementeer. Die tweede uitbreiding (Afdeling 4.2) takel die potensiële probleem van oormatige passing van die model deur beperkte tweede afgeleides te gebruik om die funksies glad te stryk. Hierdie uitbreidings word by die model ingesluit om die model se robuustheid te verbeter.

4.1 Identifikasie van uitskieters

Daar is verskillende wyses waarop uitskieters opgespoor kan word (sien byvoorbeeld Hoeting, Raftery en Madigan), maar in hierdie studie is MHLP-tegnieke geïmplementeer om die probleem van moontlike uitskieters te takel. Hierdie benadering is analoog aan die een wat in Hattingh e.a. (Hattingh, Kruger en Du Plessis 2005) beskryf word. Die MHLP-tegnieke is in Wagner se model geïnkorporeer en die aangepaste model word soos volg uiteengesit:

$$\text{Minimaliseer} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}), \tag{3.1}$$

$$\text{onderhewig aan} \quad \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} - \alpha_i = y_i, \qquad \text{vir } i = 1, \cdots, n, \tag{3.2}$$

$$f_{tj} \leq f_{lj}, \qquad \text{indien } r_{lj} = r_{tj} + 1, \text{en} \tag{3.3}$$

$$f_{tj} = f_{lj}, \text{indien } r_{tj} = r_{lj} \text{ vir } t, l = 1,2, \ldots, n \text{ met } t \neq l \text{ en } j = 1,2, \ldots, k,$$

$$-Mz_i \leq \alpha_i \leq Mz_i, \qquad \text{vir } i = 1, \cdots, n, \tag{3.4}$$

$$\sum_{i=1}^{n} z_i = p, \tag{3.5}$$

$$z_i \in \{0,1\}, \qquad \text{vir } i = 1, \cdots, n, \tag{3.6}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \qquad \text{vir } i = 1, \cdots, n, \tag{3.7}$$

$$f_{ij} \text{ en } \alpha_i \text{ is onbeperk in teken vir alle } i \text{ en } j. \tag{3.8}$$

Die veranderlike, α_i (3.2), is 'n onbeperkte veranderlike wat die speling tussen $\sum_{j=1}^{k} f_{ij}$ en y_i opneem, en daardeur ε_{1i} en ε_{2i} toelaat om nul vir daardie i te wees. Die absolute waarde van α_i (3.5) word deur Mz_i beperk waar M 'n groot getal en z_i 'n binêre veranderlike is. In eksperimente is 'n waarde van M groter as die spanwydte van die y_i-waardes as voldoende bewys. Indien z_i nul is, is α_i ook tot nul beperk en dra die i-de absolute residu tot die

doelfunksie by, maar indien z_i 1 is, sal die optimaliseringsproses die i-de residu as nul kies, aangesien α_i die speling opneem. Op hierdie wyse word die absolute residu vir datapunt i uit die model en die doelfunksie weggelaat. Die stel datapunte wat uit die model weggelaat sal word, is daardie punte wat die grootste afname in die doelfunksie sal veroorsaak wanneer dit weggelaat word. Die veranderlike p (3.6) spesifiseer die aantal datapunte wat uit die model weggelaat sal word.

Een wyse waarop 'n waarde vir p gespesifiseer kan word, is die volgende: die model word vir $p = 0$ opgelos (geen datapunte word weggelaat nie) en die waarde van die doelfunksie word vasgelê. Die model word dan herhaaldelik opgelos en die waarde van p word elke keer wanneer die model opgelos word, met 1 geïnkrementeer, terwyl die ooreenstemmende doelfunksiewaardes ook vasgelê word. Die verskillende waardes van p word dan teen die relevante vasgelegde doelfunksiewaardes gestip om waar te neem op watter wyse die doelwaarde verander het. Waarde vir p wat op die veranderingskoers in die doelfunksiewaarde gebaseer is, kan nou bepaal word. Wanneer p klein is en daar uitskieters in die data aanwesig is, kan betreklik groot veranderinge in die doelfunksie vir toenemende p-waardes verwag word, terwyl 'n klein veranderingskoers sal aandui dat 'n gepaste aantal uitskieters (waarde van p) geïdentifiseer is. 'n Ander benadering sou wees om p te selekteer teen 'n waarde wat ongeveer 10–20% van die datapunte elimineer.

4.2 Gladstryking

'n Probleem wat uit die voorgestelde model in afdeling 4.1 mag opduik, is oormatige passing. Oormatige passing vind plaas wanneer 'n funksie "te goed" in 'n datastel pas, wat die model baie sensitief vir die gedrag van 'n spesifieke datastel maak. Dit is 'n ernstige probleem, aangesien dit die voorspellingsvermoë van 'n model kan beïnvloed en dit minder betroubaar maak (Hitchcock en Sober 2004). Dit is ook denkbaar dat sekere datastelle kan lei tot funksies wat onverwagse groot fluktuasies maak, en ter wille van robuustheid moet gepoog word om hierdie tipe gedrag te vermy. As gevolg hiervan word daar dan in hierdie werk 'n gladstrykingstegniek oorweeg.

Die doel van die gladstrykingstegniek wat in hierdie studie gebruik word, is om die tweede afgeleide van die funksie, met ander woorde, die rigtingveranderingskoers te beperk. Die helling van 'n funksie kan beperk word om nie meer as 'n gespesifiseerde waarde te verander nie en dit beperk onverwagse groot fluktuasies in die helling.

Ten einde die gladstryking van 'n funksie te implementeer word spesifieke beperkinge tot die model bygevoeg. Beperkte tweede afgeleides word in hierdie beperkinge gebruik en kan soos volg beskryf word:

Stel $f_j(x_j) = f_j$ en beskou

$$\frac{\partial f_j}{\partial x_j}|_{x_{i,j}} \approx \frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}},$$

$$\frac{\partial^2 f_j}{\partial x_j^2}|_{x_{i,j}} \approx \frac{\frac{\partial f_j}{\partial x_j}|_{x_{i,j}} - \frac{\partial f_j}{\partial x_j}|_{x_{i-1,j}}}{x_{i,j} - x_{i-1,j}}, \text{en}$$

$$-\beta \leq \frac{\frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}} - \frac{f_j(x_{i,j}) - f_j(x_{i-1,j})}{x_{i,j} - x_{i-1,j}}}{x_{i,j} - x_{i-1,j}} \leq \beta, \qquad (3.9)$$

waar ≈ 'n benadering aandui.

Die absolute rigtingveranderingskoers (die tweede afgeleide) word nou deur die parameter β beperk. Daar bestaan sekerlik verskillende metodes om die parameter β te bepaal, maar in hierdie ondersoek is β dieselfde vir alle j en is dit soos volg deur eksperimentering beraam:

Stap 1 Kies 'n lae aanvangswaarde vir β.

Stap 2 Los die model op.

Stap 3 Voer 'n "laat-een-weg"-uitsnit-kruisgeldigheidsbepaling[i] uit.

Stap 4 Bereken en lê die gemiddelde absolute afwyking vas om te bepaal hoe goed die model die data verklaar.

Stap 5 Inkrementeer die waarde van β en herhaal die proses deur by stap 2 te begin.
Die proses word herhaal totdat β 'n voorafbepaalde maksimum afsnywaarde bereik.

Stap 6 Selekteer die β-waarde wat tot die laagste gemiddelde absolute afwyking gelei het.

In die studie is eksperimenteel vasgestel dat 'n β-waarde tussen 1 (aanvangswaarde in stap 1) en 400 (voorafbepaalde maksimum afsnywaarde in stap 5) die funksies voldoende sal gladstryk. Inkrementering van 25 is tussen die aanvangswaarde en maksimum afsnywaarde gebruik.

Ten slotte moet daarop gelet word dat 'n model meer robuust gemaak word deur uitskieters weg te laat, maar dat wanneer te veel datapunte weggelaat word, die datastel te klein kan word om die verwantskap tussen die afhanklike veranderlike en die voorspellerveranderlikes te beraam. Dit mag daarom wys wees om dit gedurende die implementering te kontroleer.

Deur die funksies glad te stryk kan die model verhinder word om die data oormatig te pas, maar wanneer 'n funksie te veel gladgestryk word, is dit moontlik dat slegs 'n algemene neiging aangedui word en voorspellingsakkuraatheid gekompromitteer word.

Die volgende afdeling bevat 'n alternatief vir die spesifisering van wiskundige vorms vir die funksies, $f_j(x_{ij})$. Hierdie model sal ook vir vergelykende doeleindes gebruik word wanneer die resultate wat deur die minimale-aanname-regressiemodel en die voorgestelde robuuste uitbreidings verkry is, geëvalueer word.

5. Stuksgewyse lineêre regressie

Ten einde 'n alternatiewe wiskundige model te verkry en die resultate van die voorgestelde tegnieke te vergelyk, word 'n stuksgewyse lineêre regressiemodel ingevoer. Stuksgewyse lineêre regressie is 'n vorm van regressie wat meervoudige lineêre modelle in sta at stel om by data gepas te word vir verskillende reikwydtes van x (Ryan en Porth 2007). Breekpunte is die waardes van x waar die helling van die lineêre funksie verander. Die waarde van 'n breekpunt kan voor die analise óf bekend óf onbekend wees, maar dit is tipies onbekend en moet beraam word. Datastelle in hierdie studie word óf as een lineêre regressiemodel óf as stuksgewyse lineêre kontinue segmente gemodelleer wat elk 'n lineêre model verteenwoordig word.

'n Model wat twee breekpunte gebruik en dus drie lineêre modelle van die vorm, $y = a + bx$, lewer, word vervolgens geïllustreer. In hierdie model verteenwoordig Q_{1j} en Q_{2j} die twee breekpunte wat gekies is om die 33ste and 66ste persentiele te wees. Die veranderlikes, a_{sj} en b_{sj} $(s = 1, 2, 3 \text{ en } j = 1, \cdots, k)$, is die koëffisiënte van die verskillende lineêre modelle.

$$\text{Minimaliseer} \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}), \tag{4.1}$$

$$\text{onderhewig aan} \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \qquad \text{vir } i = 1, \cdots, n, \tag{4.2}$$

$$f_{ij} = \begin{cases} a_{1j} + b_{1j}x_{ij}, & \text{indien } x_{ij} < Q_{1j}, \quad (4.3) \\ a_{2j} + b_{2j}x_{ij}, & \text{indien } Q_{1j} \leq x_{ij} < Q_{2j}, \quad (4.4) \\ a_{3j} + b_{3j}x_{ij}, & \text{indien } Q_{2j} \leq x_{ij}, \quad (4.5) \end{cases}$$

$$\text{vir } j = 1, \ldots, k,$$

$$a_{1j} + b_{1j}Q_{1j} = a_{2j} + b_{2j}Q_{1j}, \qquad \text{vir } j = 1, \cdots, k, \tag{4.6}$$

$$a_{2j} + b_{2j}Q_{2j} = a_{3j} + b_{3j}Q_{2j}, \qquad \text{vir } j = 1, \cdots, k, \tag{4.7}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \qquad \text{vir } i = 1, \cdots, n, \tag{4.8}$$

$$a_{1j}, a_{2j}, a_{3j}, b_{1j}, b_{2j}, b_{3j} \text{ onbeperk vir } j = 1, \ldots, k.$$

Die doel van hierdie model is om stuksgewyse lineêre modelle gelyktydig met die (additiewe) regressiemodel toe te pas. Uitskieters kan uit hierdie model weggelaat word op dieselfde wyse as wat in afdeling 4.1 beskryf is.

6. Empiriese eksperimente en resultate

Vier bekende datastelle is beskou om die voorgestelde modelle en hul voorspellingsakkuraatheid te illustreer en te evalueer. In elke geval is die gemiddelde absolute afwyking as 'n prestasiemaatstaf gebruik. 'n "Laat-een-weg"-uitsnitbenadering (Efron en Gong 1983) is gebruik om die gemiddelde absolute afwyking te bereken; dit behels die volgende stappe:

- Skrap punte, x_i, een op 'n keer, uit die datastel.
- Herbereken die voorspellingsreël op die basis van die oorblywende $n-1$-punte;
- Kyk hoe goed die herberekende reël die geskrapte punt voorspel.
- Bereken die gemiddelde van die absolute foutvoorspellings oor alle n-skrappings (die gemiddelde absolute afwyking).

Ten einde die resultate van die voorgestelde uitbreidings op Wagner se minimale-aanname-model te evalueer word die resultate van addisionele regressiemodelle ook aangebied vir elke geval wat oorweeg word. Hierdie modelle sluit die gewone L_1-norm- en L_2-norm-regressiemodelle in, sowel as 'n variasie (een en twee breekpunte met datapunte wat geskrap is) van die stuksgewyse lineêre regressiemodel wat in afdeling 5 beskryf is.

Die programmering en ontleding vir al die modelle is in C++ en CPLEX (10.1) gedoen deur van ILOG Concert Technology (Ilog 2006) gebruik te maak.

6.1 Stapelverlies

Die stapelverlies-datastel (stack loss data set) is 'n welbekende datastel wat deur etlike skrywers (Hoeting e.a. 1996, Brownlee 1965, Steel en Uys 2007) ondersoek is. Die data word gebruik om die verwantskap tussen die persentasies van onveranderde ammoniak te ondersoek wat in 21 dae uit 'n aanleg ontsnap. Die volgende drie verklarende veranderlikes word gebruik:

x_1: lugvloei wat die bedryfstempo van die aanleg meet
x_2: inlaattemperatuur van koelwater wat deur die spirale in die toring sirkuleer
x_3: 'n waarde wat proporsioneel tot die konsentrasie van suur in die toring is.

Om die rigting van monotonisiteit vir elke veranderlike te bepaal is 'n meervoudige regressieanalise uitgevoer wat tot positiewe koëffisiënte vir x_1 en x_2 en 'n negatiewe koëffisiënt vir x_3 gelei het. Beide f_1 en f_2 is daarom as monotone nie-afnemende funksies beperk, terwyl f_3 as 'n monotone nietoenemende funksie gespesifiseer is.

Tabel 1 bevat die stapelverliesdata, $y_i, x_{i1}, x_{i2}, \text{en } x_{i3}$ is die afhanklike en onafhanklike veranderlikes, respektiewelik, terwyl $f_1(x_{i1}), f_2(x_{i2}) \text{ en } f_3(x_{i3})$ die funksiewaardes is wat bepaal word deur Wagner se model (sien afdeling 3) op te los. Die voorspellingswaarde, $\hat{y}$, word deur $\hat{y}_i = \sum_{j=1}^{3} \hat{f}_{ij}$ bepaal. Die absolute residue word in die laaste kolom aangedui.

Tabel 1. Stapelverliesdata met funksiewaardes en residue

i	y_i	x_{i1}	$f_1(x_{i1})$	x_{i2}	$f_2(x_{i2})$	x_{i3}	$f_3(x_{i3})$	$\hat{y}_i$	$\lvert y_i - \hat{y}_i \rvert$
1	42	80	16	27	12	89	14	42	0
2	37	80	16	27	12	88	14	42	5
3	37	75	16	25	12	90	9	37	0
4	28	62	2	24	12	87	14	28	0
5	18	62	2	22	2	87	14	18	0
6	18	62	2	23	2	87	14	18	0
7	19	62	2	24	12	93	6	20	1
8	20	62	2	24	12	93	6	20	0
9	15	58	0	23	2	87	14	16	1
10	14	58	0	18	0	80	14	14	0
11	14	58	0	18	0	89	14	14	0
12	13	58	0	17	-1	88	14	13	0
13	11	58	0	17	-1	82	14	13	2
14	12	58	0	19	1	93	6	7	5
15	8	50	-7	18	0	89	14	7	1
16	7	50	-7	18	0	86	14	7	0
17	8	50	-7	19	1	72	14	8	0
18	8	50	-7	19	1	79	14	8	0
19	9	50	-7	20	2	80	14	9	0
20	15	56	-1	20	2	82	14	15	0
21	15	70	7	20	2	91	6	15	0

Voorspellingsakkuraatheid

Beskou tabel 2 hier onder, wat die vergelyking vir die L_2-norm-regressie, die L_1-norm-regressie en die oorspronklike minimale-aanname-regressiemodel wat deur Wagner bekendgestel is, aandui. Geen uitbreidings (uitskieteropsporing en gladstryking) is geïmplementeer nie en uit die gemiddelde absolute afwyking kan daarop gelet word dat die oorspronklike Wagner-model beter as die L_2-norm-regressieresultaat presteer het. Die gemiddelde absolute afwyking van die L_1-norm-regressie is slegs 0.042 (of ongeveer 2%) minder as die gemiddelde absolute afwyking van Wagner se model.

Tabel 2. Stapelverliesdata: Voorspellingsakkuraatheid vir modelle wat nie die voorgestelde uitbreidings implementeer nie

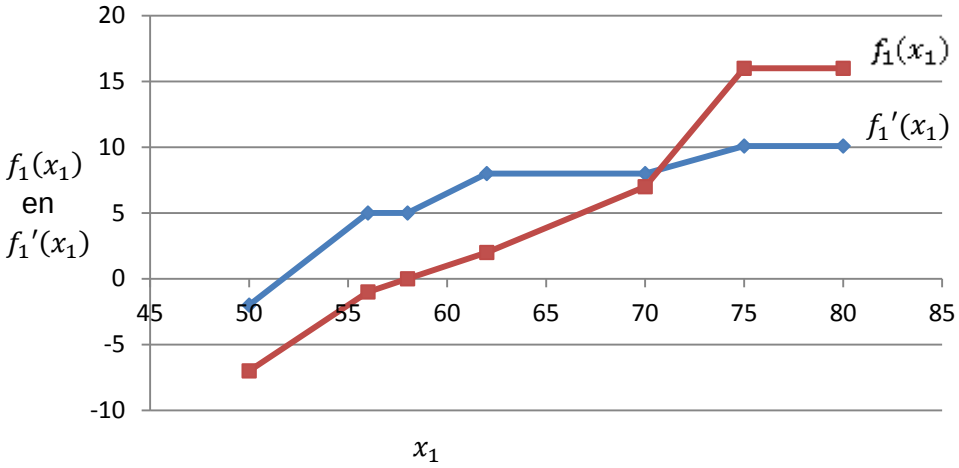
Model	Gemiddelde absolute afwyking
L_2-norm-regressie	2.887
L_1-norm-regressie	2.035
Oorspronklike minimale-aanname-regressiemodel (Ilog 2006)	2.077

Tabel 3 dui die voorspellingsakkuraatheidresultate van die voorgestelde uitbreidings aan en die stuksgewyse lineêre modelle vir die stapelverliesdatastel. Daar is vir verskillende waardes van β geëksperimenteer met die aantal punte wat weggelaat kan word'n Rooster vir waardes van β en p is dus gebruik om te bepaal dat $\beta = 50$ en $p = 2$ die beste resultate sal lewer. Uit tabel 3 is dit duidelik dat die invoering van die twee voorgestelde uitbreidings die voorspellingsakkuraatheid van die oorspronklike minimale-aanname-model verbeter het. Die weglating van twee datapunte en die gebruik van 'n gladstrykingsfaktor van $\beta = 50$ laat die gemiddelde absolute afwyking met 41% afneem, van 2.077 (die oorspronklike minimale-aanname-model in tabel 2) tot 1.220. In twee gevalle het die stuksgewyse lineêre regressiemodelle ook beter resultate as die oorspronklike minimale-aanname-model gelewer.

Tabel 3. Stapelverliesdata: Voorspellingsakkuraatheid vir die uitgebreide modelle

Model	Gemiddelde absolute afwyking
Wagner se model met twee datapunte wat weggelaat word	2.194
Wagner se model met twee datapunte wat weggelaat word en 'n gladstrykingsfaktor van $\beta = 50$	1.220
L_1-norm-regressie met twee datapunte wat weggelaat word en geen breekpunt nie	1.394
Stuksgewyse L_1-norm-regressie met twee datapunte wat weggelaat word en een breekpunt	1.882
Stuksgewyse L_1-norm-regressie met twee datapunte wat weggelaat word en twee breekpunte	2.150

Ten einde die vorm van die funksies met 'n gladstrykingsfaktor van $\beta = 50$ te illustreer, toon figuur 1 die oorspronklike funksie, $f_1(x_1)$, sowel as die gladgestrykte funksie, $f_1'(x_1)$. Die ander funksies, wat nie hier getoon word nie, volg op dieselfde wyse.



Figuur 1. Verandering in funksie ná gladstryking met $\beta = 50$

6.2 Skotse heuwelwedloopdata

In die tweede voorbeeld word die Skotse heuwelwedloopdatastel beskou (Atkinson 1986). Dieselfde proses as die een wat met die stapelverliesdata gevolg is, is hier gevolg om die resultate te verkry. Die Skotse heuwelwedloopdatastel word gebruik om die verwantskap tussen die rekordwentye van 35 heuwelwedlope in Skotland na te vors, met die volgende twee verklarende veranderlikes:

1. x_1: afstand wat in myle gedek is
2. x_2: hoogte wat gedurende die wedloop geklim is.

Voorspellingsakkuraatheid

Tabel 4 dui die gemiddelde absolute fout aan vir die modelle wat nie die voorgestelde uitbreidings implementeer nie, terwyl tabel 5 die resultate van die uitgebreide modelle aanbied.

Tabel 4. Skotse heuwelwedloopdata: Voorspellingsakkuraatheid vir modelle wat nie die voorgestelde uitbreidings implementeer nie

Model	Gemiddelde absolute afwyking
L_2-norm-regressie	9.367
L_1-norm-regressie	8.211
Oorspronklike minimale-aanname-regressiemodel (Ilog 2006)	8.927

Tabel 5. Skotse heuwelwedloopdata: Voorspellingsakkuraatheid vir die uitgebreide modelle

Model	Gemiddelde absolute afwyking
Wagner se model met vier datapunte wat weggelaat word	8.469
Wagner se model met vier datapunte wat weggelaat word en 'n gladstrykingsfaktor van $\beta = 10$	3.921
L_1-norm-regressie met vier datapunte wat weggelaat word en geen breekpunt nie	4.253
Stuksgewyse L_1-norm-regressie met vier datapunte wat weggelaat word en een breekpunt	3.559
Stuksgewyse L_1-norm-regressie met vier datapunte wat weggelaat word en twee breekpunte	4.280

Die effek wat bereik word wanneer die voorgestelde uitbreidings ingesluit word, is betekenisvol. Deur 'n gladstrykingsfaktor van $\beta = 10$ in te voer en vier datapunte (uitskieters) weg te laat, neem die gemiddelde absolute afwyking van 8.927 (die oorspronklike model in tabel 4) af na 3.921 – 'n vermindering van 56%. Die stuksgewyse L_1-norm-regressiemodel met een breekpunt het egter die uitgebreide minimale-aanname-model met ongeveer 9% oortref.

6.3 Brandstofdata

Die derde datastel wat in die empiriese eksperiment gebruik is, bestaan uit data in verband met brandstofverbruik in verskillende state van die VSA (Weisberg 2005). Die verwantskap tussen die responsveranderlike, brandstofverbruik in gallonne per persoon, en die vier verklarende veranderlikes word vir 48 state geëvalueer.

Die vier verklarende veranderlikes vir elke staat is:

1. x_1: 1972 se hoeveelheid belasting per gallon, in sente gemeet
2. x_2: 1972 se per capita-inkomste in duisende dollars
3. x_3: 1971 se duisende myle van die vernaamste hoofweg
4. x_4: die persentasie van die bevolking met 'n rybewys.

Voorspellingsakkuraatheid

Tabel 6 dui die resultate aan wat verkry is deur die verskillende modelle op te los sonder dat die voorgestelde uitbreidings geïmplementeer is. Tabel 7 verskaf die besonderhede van die modeluitbreidings se resultate.

Tabel 6. Brandstofdata: Voorspellingsakkuraatheid vir modelle wat nie die voorgestelde uitbreidings implementeer nie

Model	Gemiddelde absolute afwyking
L_2-norm-regressie	54.532
L_1-norm-regressie	49.466
Oorspronklike minimale-aanname-regressiemodel (Ilog 2006)	52.726

Tabel 7. Brandstofdata: Voorspellingsakkuraatheid vir die uitgebreide modelle

Model	Gemiddelde absolute afwyking
Wagner se model met ses datapunte wat weggelaat word	33.639
Wagner se model met ses datapunte wat weggelaat word en 'n gladstrykingsfaktor van $\beta = 1$	37.425
L_1-norm-regressie met ses datapunte wat weggelaat word en geen breekpunt nie	35.426
Stuksgewyse L_1-norm-regressie met ses datapunte wat weggelaat word en een breekpunt	33.293
Stuksgewyse L_1-norm-regressie met ses datapunte wat weggelaat word en twee breekpunte	44.320

Uit tabel 7 is dit duidelik dat die skrapping van ses datapunte (uitskieters) die gemiddelde absolute afwyking aansienlik verklein het. Vir Wagner se oorspronklike minimale-aanname-regressiemodel is die gemiddelde absolute afwyking met 36.2% verklein deur ses datapunte weg te laat (van 52.726 in tabel 6 tot 33.639). Die invoering van die tweede uitbreiding (gladstryking) het egter nie die gemiddelde absolute afwyking verder verbeter nie, alhoewel dit nogtans, op 37.425, aansienlik beter as die oorspronklike waarde van 52.726 is. Die stuksgewyse L_1-norm-regressiemodel met een breekpunt, wat ook ses datapunte weglaat, het 'n gemiddelde absolute afwyking van 33.293 gelewer, wat net marginaal beter is as die uitgebreide Wagner-model se gemiddelde absolute afwyking van 33.639.

6.4 Bruto nasionale produk (BNP)

Die vierde datastel is uit'n studie verkry wat deur Roux (1994) uitgevoer is. In hierdie studie is 'n regressiemodel beskou wat die bruto nasionale produk (BNP) met 10 faktore vir 43 lande in verband bring. As gevolg van die betreklik groot aantal verklarende veranderlikes is daar besluit om slegs sewe van die 10 oorspronklike veranderlikes te kies. Hierdie besluit is op 'n studie gebaseer wat uitgevoer is deur Hattingh e.a. (2005), wat bewyse gevind het dat drie van die 10 veranderlikes moontlik oortollig kan wees. Die sewe geselekteerde verklarende veranderlikes is:

1. x_1: netto uitvoer per capita
2. x_2: verandering in inflasie
3. x_3: landbou as 'n persentasie van die bruto huishoudelike produk (BHP)
4. x_4: politieke situasie
5. x_5: gemiddelde ongeletterdheid van die bevolking in die land
6. x_6: groei in lewensverwagting van die inwoners van die land
7. x_7: groei in die bevolking van die land.

Voorspellingsakkuraatheid

Tabelle 8 en 9 dui die resultate aan vir die modelle sonder die voorgestelde uitbreidings en vir die uitgebreide modelle, onderskeidelik.

Tabel 8. BNP-data: Voorspellingsakkuraatheid vir modelle wat nie die voorgestelde uitbreidings implementeer nie

Model	Gemiddelde absolute afwyking
L_2-norm-regressie	4038.425
L_1-norm-regressie	3215.116
Oorspronklike minimale-aanname-regressiemodel (Ilog 2006)	3282.140

Tabel 9. BNP-data: Voorspellingsakkuraatheid vir die uitgebreide modelle

Model	Gemiddelde absolute afwyking
Wagner se model met vier datapunte wat weggelaat word	2764.669
Wagner se model met vier datapunte wat weggelaat word en 'n gladstrykingsfaktor van $\beta = 200$	1820.180
L_1-norm-regressie met vier datapunte wat weggelaat word en geen breekpunt nie	2360.977
Stuksgewyse L_1-norm-regressie met vier datapunte wat weggelaat word en een breekpunt	2782.106
Stuksgewyse L_1-norm-regressie met vier datapunte wat weggelaat word en twee breekpunte	1713.885

Die resultate vir die BNP-data toon weer eens aan dat die vooruitskattingsakkuraatheid aansienlik sal verbeter deur die twee voorgestelde uitbreidings tot Wagner se oorspronklike minimale-aanname-model in te voer. Die gemiddelde absolute afwyking vir die oorspronklike model is 3282.14 (tabel 8), en die byvoeging van 'n gladstrykingsfaktor $\beta = 200$ en weglating van vier datapunte het 'n gemiddelde absolute afwyking van 1820.18 tot gevolg gehad – 'n vermindering van 45%. Dit was, soos in vorige gevalle, weer moontlik om die gemiddelde

absolute afwyking verder te reduseer met ongeveer 6% deur 'n stuksgewyse L_1-norm-regressiemodel met twee breekpunte te gebruik en ook vier datapunte weg te laat.

Ten einde die bevindinge op te som, kan verklaar word dat die voorgestelde twee uitbreidings tot die oorspronklike minimale-aanname-regressiemodel wat deur Wagner bekendgestel is, suksesvol bewys is. In al die gevalle wat beskou is, is die gemiddelde absolute afwyking aansienlik gereduseer deur die implementering van die twee uitbreidings (die weglating van uitskieters en gladstryking). In een geval het die invoering van die tweede uitbreiding (gladstryking) nie die mate van voorspellingsakkuraatheid verder verbeter nie. In drie van die vier gevalle was dit moontlik om die gemiddelde absolute afwyking verder te reduseer deur stuksgewyse lineêre modelle in te voer. Hierdie verbeteringe was egter net marginaal beter as die voorgestelde uitgebreide modelle. Indien die resultate wat verkry is, beskou word, blyk dit redelik te wees om tot die gevolgtrekking te kom dat daar gevalle is waar die minimale-aanname-regressiemodel, en die uitbreidings daarvan, beter resultate as ander modelle sal oplewer.

Die bevindinge van hierdie studie het verskeie geleenthede tot verdere navorsing geskep. Ander benaderings, waar nie slegs op die verwagte waarde van die responsveranderlike gesteun word nie, byvoorbeeld regressiekwantiele, kan ook bestudeer word. In hierdie werk word nie veel melding gemaak van interaksieterme in die modelle nie, maar die model kan maklik uitgebrei word om ook interaksieterme in te sluit, soos ook aangedui in die oorspronklike werk van Wagner. As verdere studie kan simulasies ook gedoen word om die voorgestelde model oor 'n wyer klas situasies te toets en te vergelyk met an der modelle, soos die veralgemeende additiewe model waarna in afdeling 2 verwys is. Ten slotte kan die werk wat hier aangebied word, ook aansluiting vind by ander werk en navorsing in die literatuur. Byvoorbeeld, sogenaamde gepenaliseerde modelle om 'n bala ns te vind tussen te veel en te min gladstryking sou aansluiting kon vind by die werk wat in hierdie artikel aangebied is. Daar is egter nie in hierdie werk gepoog om aansluiting te vind by ander areas nie en dit word beskou as deel van die geleenthede wat vir verdere navorsing geskep is.

7. Gevolgtrekkings

In hierdie studie is 'n bestaande regressiemodel met minimale aannames ondersoek. In 'n poging om robuustheid en voorspellingsakkuraatheid te verbeter is die model uitgebrei om voorsiening vir die opsporing van uitskieters en moontlike oormatige passing te maak. Die uitbreidings is geïmplementeer deur gebruik te maak van gemengde heeltallige lineêre programmeringstegnieke en beperkte tweede afgeleides. Stuksgewyse lineêre regressiemodelle is ook in werking gestel ten einde die resultate van die voorgestelde uitgebreide modelle te evalueer en te vergelyk.

Die hoofbevinding van die studie was dat die implementering van die twee voorgestelde uitbreidings die voorspellingsvermoë (soos deur die gemiddelde absolute afwyking gemeet) 'n aansienlike verbetering van die oorspronklike minimale-aanname-regressiemodel tot gevolg gehad het. Hoewel die voorgestelde uitbreidings nie in al die datastelle wat ondersoek is, die laagste gemiddelde absolute afwyking bevat het nie, was die resultate altyd vergelykbaar met die resultate van die ander modelle.

Ten laaste kan vermeld word dat moderne optimimaliseringsagteware soos CPLEX kragtig genoeg blyk te wees om modelle soos dié wat in hierdie studie voorgestel word, op te los. Klein tot mediumgrootte probleme is in ‘n betreklik kort tyd opgelos.

Bibliografie

H.M. Wagner. Non-linear regression with minimal assumptions. *Journal of the American Statistical Association*, 57:572–8, 1962.

C. Hitchcock en E. Sober. Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, 55:1–34, 2004.

S.E. Ryan en L.S. Porth. *A tutorial on the piecewise regression approach applied to bedload transport data*. General Technical Report RMRS-GTR-189:1–41, 2007.

B.L. Bowerman, R.T. O'Connell en A.B. Koehler. *Forecasting, time series, and regression: an applied approach*. Thomson Brooks/Cole, Belmont, 2005.

J. Hoeting, A.E. Raftery en D. Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 22:251–270, 1996.

P.J. Rousseeuw en A.M. Leroy. *Robust regression and outlier detection*. Wiley-Interscience, Hoboken, 2003.

T.J. Hastie en R.J. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

T.J. Hastie en R.J. Tibshirani. Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82:371–386, 1987.

T.J. Hastie en R.J. Tibshirani. *Generalized additive models*. Chapman & Hall/CRC, London, 1990.

S.N. Wood. *Generalized additive models: An introduction with R*. Chapman & Hall/CRC, London, 2006.

J.M. Hattingh, H.A. Kruger en P.M. du Plessis. Linear model selection: towards a framework using a mixed integer linear programming approach. *South African Statistical Journal*, 39:197–220, 2005.

B. Efron en G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37:36–48, 1983.

Ilog, ILOG CPlex 10.1 User's manual, Ilog, France, 2006.
K.A. Brownlee. *Statistical theory and methodology in science and engineering*. Wiley, New York, 1965.

S.J. Steel en D.W. Uys. Variable selection in multiple linear regression: the influence of individual cases. *Orion*, 23:123–136, 2007.

A.C. Atkinson. [Influential observations, high leverage points, and outliers in linear regression]: Comment: aspects of diagnostic regression analysis. *Statistical Science 1:* 397–402, 1986.

S. Weisberg. *Applied linear regression*. Wiley, Hoboken, 2005.

T.P. Roux. 'n Rekenaargebaseerde stelsel om kwantifiseerbare aspekte van sosio-ekonomiese en sosio-politiese faktore van lande te ontleed, Potchefstroom: PU vir CHO, (Verhandeling-M.Com.), 1994.

[i] Kruisgeldigheidsbepaling is 'n tegniek wat gebruik word om die prestasie van 'n voorspellingsmodel te beraam. Die "laat-een-weg"-uitsnitbenadering wat deur Efron en Gong (1983) voorgestel is, is in hierdie studie gekies as tegniek vir kruisgeldigheidsbepaling.