3 Enkelvoudige lineaire regressie

In dit hoofdstuk zullen we een begin maken met de introductie van statistische technieken die het mogelijk maken uit meetgegevens zinvolle verbanden tussen fysische en chemische grootheden te halen. Deze technieken worden vaak aangeduid met de namen curve fitting of regressie-analyse.

Kernbegrippen van dit hoofdstuk:

- kwadratensom
- variabelen
  - afhankelijk
  - onafhankelijk
- determinatiecoëfficiënt
- residu
- toets van D
- betrouwbaarheidsinterval
- significantie toets
- invloedrijk punt
- lack-of-fit

3.1 Methode van de kleinste kwadratensom.

Gegeven is een algemene (chemisch/fysische) vergelijking met 1 of meer parameters. Hiervan zijn meetgegevens beschikbaar. In dit hoofdstuk zullen de uitgangspunten besproken worden om optimale waarden voor de parameters te vinden. Onder optimaal wordt hier verstaan dat de gekozen vergelijking de meetgegevens zo goed mogelijk beschrijft.

Tijd (sec)	Gemeten afstand (m)
1	36.754
2	71.845
3	60.479
4	101.149
5	103.150
6	111.148
7	142.170
8	157.334
9	161.843
10	206.030

Om deze uitgangspunten duidelijk te maken wordt uitgegaan van een voorbeeld. In de tabel worden 10 meetgegevens vermeld. Dit zijn tijden t in seconden en bijbehorende metingen van de afgelegde weg in meters.

Op basis van deze meetgegevens wordt gevraagd naar de snelheid over het tijdsinterval dat er meetgegevens beschikbaar zijn. We nemen hierbij aan dat de snelheid gedurende de metingen vrijwel constant is.

Om de snelheid uit deze meetgegevens te kunnen afleiden, kiezen we de volgende vergelijking om de meetgegevens te beschrijven:

$s = s_0 + v\,t$

waarin:

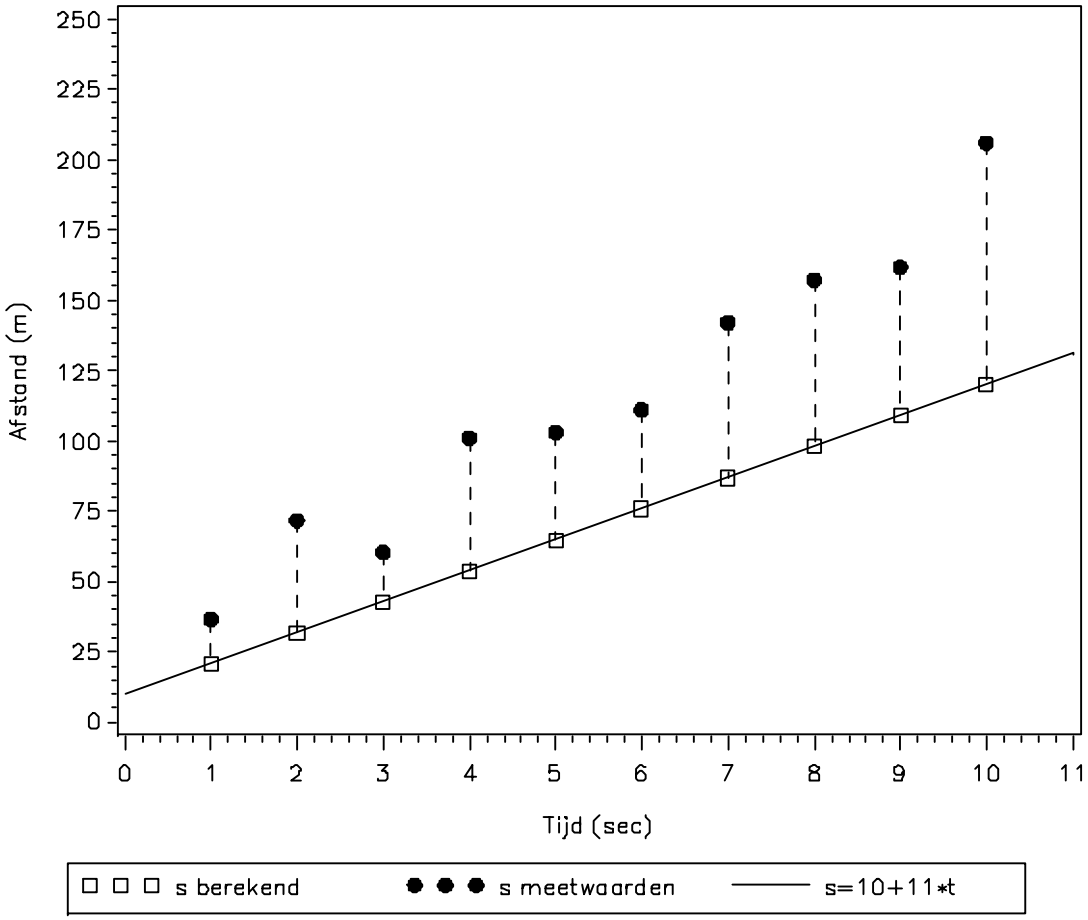
s : de afgelegde weg op tijdstip t (m)

s_0 : de afgelegde weg op tijdstip t=0 (m)
v : snelheid (m/sec)
t : tijd (sec)

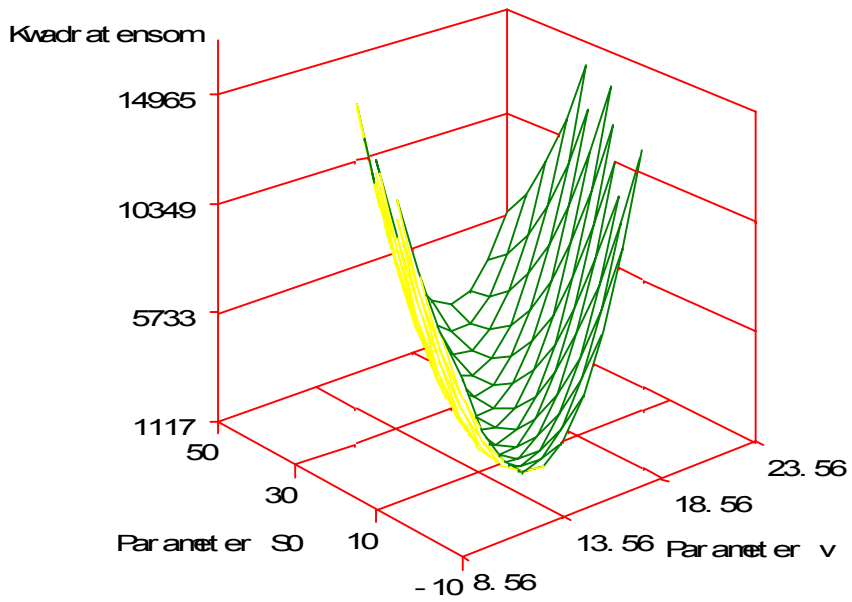
In de gekozen vergelijking zijn s_0 en v parameters. Van deze parameters willen we nu de waarden zodanig bepalen, dat de resulterende vergelijking de meetgegevens in de tabel zo goed mogelijk beschrijft. Precies beschrijven zal niet mogelijk zijn doordat enerzijds onze metingen meetfouten zullen bevatten en anderzijds de snelheid van het object niet precies constant zal zijn. Hiervoor gaan we als volgt te werk.
Kies een waarde voor de parameters s_0 en v bijvoorbeeld s_0 =10 en v=11. Nu de vergelijking volledig vastligt, kan voor iedere gemeten tijd t de afgelegde weg s berekend worden volgens de vergelijking. Voor ieder meetpunt worden de gemeten en berekende afstanden van elkaar afgetrokken en het resulterende verschil gekwadrateerd. Van alle meetpunten worden de op deze wijze berekende kwadraten bij elkaar opgeteld. Het resultaat is één getal dat de kwadratensom genoemd wordt. De volgende tabel geeft de resultaten van deze berekeningen:

Tijd (sec)	Gemeten afstand	Berekende afstand	Gemeten – Berekende afstand	Kwadraat	
1	36.754	21	15.754	248.19	
2	71.845	32	39.845	1587.62	
3	60.479	43	17.479	305.52	
4	101.149	54	47.149	2223.03	
5	103.150	65	38.15	1455.42	
6	111.148	76	35.148	1235.38	
7	142.170	87	55.17	3043.73	
8	157.334	98	59.334	3520.52	
9	161.843	109	52.843	2792.38	
10	206.030	120	86.03	7401.16	+
			Kwadratensom	23812.96	

In onderstaande figuur wordt de berekening van de kwadratensom nog eens getoond. De zwarte rondjes markeren de meetgegevens, de vierkanten de bijbehorende berekende waarden voor de vergelijking $s=10+11*t$. Deze vergelijking wordt zelf weergegeven door de getrokken lijn. De stippellijnen markeren het verschil tussen de gemeten en berekende afstand s. De lengte van alle stippellijnen gekwadrateerd en bij elkaar opgeteld levert de kwadratensom.

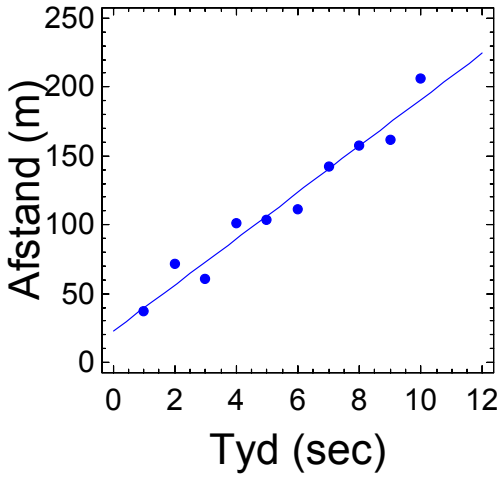


Door het kiezen van andere waarden voor de parameters s_0 en v wordt een andere kwadratensom gevonden. In onderstaande figuur wordt in drie dimensies het verband tussen de kwadratensom en combinaties van s_0 en v weergegeven.



Onder de optimale waarden voor de parameters s_0 en v worden nu die waarden verstaan, waarbij de berekende kwadratensom minimaal is. Deze aanpak wordt aangeduid als de methode van de kleinste kwadraten.

Met behulp van de methode welke beschreven zal worden in het vervolg van dit hoofdstuk kunnen de optimale waarden s_0 =22.89665 en v=16.78 bepaald worden. De kwadratensom voor deze parameters is 1084.039 en dit is tevens de kleinste kwadratensom die gevonden kan worden. De gevraagde gemiddelde snelheid v over het tijdsinterval dat er meetgegevens beschikbaar zijn, is dus 16.78 m/sec. De figuur hiernaast toont hiervoor de meetgegevens en de bijbehorende optimale vergelijking



Het hier uitgewerkte probleem, is natuurlijk vrij eenvoudig en kan in feite op een simpele zakrekenmachine tot een goede oplossing gebracht worden. Het gaat echter niet om de oplossing van dit specifieke probleem maar om de algemene methode van aanpak. Kijk bijvoorbeeld naar de vergelijking van Antoine:

$$P_s = e^{A-\frac{B}{T+C}}$$

waarin:

P_s : verzadigde dampspanning (Pa)
T : absolute temperatuur (K)
A,B,C : parameters

Het vinden van waarden voor de parameters A, B en C om een set gemeten verzadigde dampspanningen bij verschillende temperaturen te kunnen beschrijven, gaat op exact dezelfde wijze als boven beschreven. Neem waarden aan voor de parameters A, B en C. Bereken vervolgens de kwadratensom. Neem nieuwe waarden voor de A, B en C en herhaal dit net zolang tot weer de kleinst mogelijke waarde van de kwadratensom is gevonden. Het vinden van deze optimale waarden voor A, B en C is nu echter een wat lastiger probleem. In het hoofdstuk over 'Niet-lineaire regressie' zal besproken worden hoe dit probleem opgelost kan worden.

Er worden nu een aantal algemene begrippen geïntroduceerd. Deze hebben te maken met de vergelijking, die de meetgegevens zou moeten beschrijven en die de weergave is van het model dat de scheikundige heeft geformuleerd. Deze modelvergelijking kan als volgt geformuleerd worden:

$$Y = F(x_j, \beta_k)_{\substack{j=1,\ldots,m \\ k=1,\ldots,p}} + \varepsilon$$

waarin:

Y : De grootheid die beschreven wordt, bijvoorbeeld afstand, verzadigde dampspanning, enz. Y wordt de **afhankelijke variabele** (Engels: dependent variable) genoemd.

x : De bij de meting ingestelde eigenschappen of waarden, bijvoorbeeld tijd, absolute temperatuur. Men spreekt hier van af **onafhankelijke variabelen** of **instelvariabelen** (Engels: independent variables). Hiervan kunnen er meerdere zijn, zoals een vergelijking die de dampspanning beschrijft bij gegeven absolute temperatuur en druk. Om de verschillende onafhankelijke variabelen van elkaar te onderscheiden, wordt de index j gebruikt. Er zijn dus m verschillende onafhankelijke variabelen.

β : De parameters zoals de snelheid v in het eerste voorbeeld en A, B, C in de vergelijking van Antoine. In totaal bevat de vergelijking p parameters.

F : Een willekeurig (algebraïsch) functievoorschrift.

ε : Foutterm.

Deze algemene vergelijking beschrijft de meetgegevens. Deze meetgegevens zijn opgebouwd uit individuele metingen. Iedere individuele meting bestaat uit m ingestelde waarden voor $x_{i,j}$ en een bijbehorende meetwaarde y_i. Er zijn n metingen uitgevoerd, dus i loopt van 1 t/m n. Bij de i^e meting worden m waarden voor x ingesteld, dus x krijgt 2 indices i en j waarbij i loopt van 1 t/m n en j van 1 t/m m. Merk op dat het toegestaan is dat bij verschillende metingen dezelfde instelwaarden worden gebruikt. We zullen later in dit hoofdstuk aangeven waarom het nuttig kan zijn metingen bij bepaalde instelwaarden te herhalen.

Het aanpassen van de modelvergelijking aan de meetgegevens bestaat uit het vinden van die waarden voor de parameters $\beta_{k=1,p}$ dat

$$SSQ = \sum_{i=1}^{n} \left[y_i - F(x_j, \beta_k) \right]^2$$

minimaal is. Deze kwadratensom SSQ noemen we de doelfunctie. Dit rekenproces wordt **regressie** genoemd.

3.2 Soorten regressie

Met betrekking tot het vinden van de optimale waarden voor de parameters β in modelvergelijkingen met behulp van regressie worden een drietal mogelijkheden onderscheiden. Als eerste wordt er een onderscheid gemaakt tussen lineaire en niet-lineaire regressie. Er is sprake van lineaire regressie als voor alle parameters in de modelvergelijking geldt, dat de partiële afgeleiden van Y naar de parameter β_k geen parameters bevat. Bevat deze partiële afgeleide wel 1 of meerdere parameters uit de modelvergelijking, dan is er sprake van niet-lineaire regressie.

Voorbeelden:

Modelvergelijking: $y = \beta_1 x + \beta_2 x^2$

Partiële afgeleide naar β_1: $\frac{\partial y}{\partial \beta_1} = x$ bevat geen β_1 en/of β_2.

Partiële afgeleide naar β_2: $\frac{\partial y}{\partial \beta_2} = x^2$ bevat geen β_1 en/of β_2.

Dus het vinden van de optimale waarden voor parameters β_1 en β_2 in deze modelvergelijking kan met lineaire regressie.

Een ander voorbeeld:

Modelvergelijking: $y = \beta_1 C^{\beta_2}$

Partiële afgeleide naar β_1: $\frac{\partial y}{\partial \beta_1} = C^{\beta_2}$ bevat parameter β_2.

Partiële afgeleide naar β_2: $\frac{\partial y}{\partial \beta_2} = \beta_1 C^{\beta_2} \ln(C)$ bevat zowel β_1 als β_2.

Omdat de afgeleide van y naar β_1 de parameter β_2 nog bevat, moeten de parameters in deze modelvergelijking aangepast worden met niet-lineaire regressie.

Het onderscheid tussen lineaire en niet-lineaire regressie is van belang voor de rekenmethode om de optimale waarden voor de parameters aan te passen. In het geval van lineaire regressie is er een analytische oplossing, d.w.z. er is een formule waarmee de oplossing direct uitgerekend kan worden. Anders ligt dit voor de niet-lineaire regressie. Hierbij is geen analytische oplossing voorhanden. De waarden van de parameters moeten met een iteratief rekenproces gevonden worden. Dit iteratieve rekenproces heeft een startwaarde nodig voor alle parameters. Als de keuze van de startwaarden ongelukkig is, kunnen pseudo-optimale oplossingen gevonden worden, zgn. lokale minima van de kwadratensom. Niet-lineaire regressie is een stuk lastiger uit te voeren dan lineaire regressie.

Nu er onderscheid gemaakt is tussen lineaire en niet-lineaire regressie, wordt er met betrekking tot lineaire regressie nog onderscheid gemaakt tussen enkelvoudige en meervoudige lineaire regressie.

Enkelvoudige lineaire regressie: De relatie bevat slechts 1 instelvariabele x.
$Y = \beta_0 + \beta_1 x + \varepsilon$

Meervoudige lineaire regressie: De relatie bevat meer dan 1 instelvariabele.
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon$

Samenvattend worden de volgende 3 soorten regressie onderscheiden:
Enkelvoudige lineaire regressie, meestal kortweg aangeduid als lineaire regressie.
Meervoudige lineaire regressie.
Niet-lineaire regressie.

Er zijn nog andere, zeer specifieke soorten regressie zoals ridge regressie. Deze soorten regressie vallen buiten het bestek van dit college. In het vervolg van dit hoofdstuk zullen we ons beperken tot enkelvoudige lineaire regressie. Meervoudige lineaire en niet-lineaire regressie zullen in volgende hoofdstukken besproken worden.

3.2 Enkelvoudige lineaire regressie.

Bij enkelvoudige lineaire regressie heeft de vergelijking, die aangepast wordt aan de set meetgegevens, de volgende algemene vorm:
$Y = \beta_0 + \beta_1 x$.

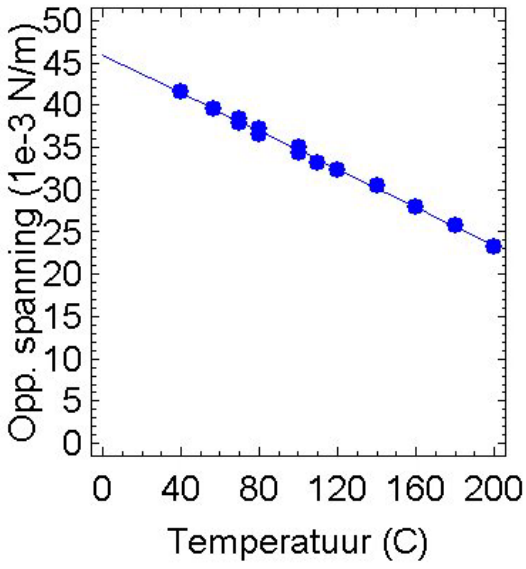
Deze vergelijking is een rechte lijn met een richtingscoëfficiënt β_1 en afgesneden stuk bij de oorsprong β_0. De instelvariabele x mag ook een willekeurige functie zijn, zolang β_0 of β_1 er maar niet in voorkomen, bijv. $\sqrt{x}$, x^2 , $\ln(x)$, enz.
Beschikbaar zijn meetgegevens van de oppervlaktespanning van nitrobenzeen over het temperatuur traject 40 tot 200 ºC (R.C. Reid, "Properties of gases and liquids"). Deze meetgegevens worden gegeven in onderstaande tabel. In de bijbehorende figuur zijn deze meetgegevens tevens in een grafiek weergegeven.

Voor gereduceerde temperaturen T_r tussen 0.3 en 0.7 geldt over het algemeen, dat de oppervlaktespanning van vloeistoffen als functie van temperatuur een rechte lijn is:
$\gamma = \beta_0 + \beta_1 T$
waarin:

γ	:	Oppervlaktespanning (N/m)
T	:	Temperatuur (ºC).
β_0	:	Parameter
β_1	:	Parameter

De grafiek voor de oppervlaktespanningen in de tabel laat zien, dat dit ook voor nitrobenzeen geldt. Gevraagd wordt nu de optimale waarden voor de parameter constanten β_0 en β_1 te bepalen, waarmee de algemene vergelijking de gegevens in de tabel optimaal beschrijft.

Temp (ºC)	Oppervlakte spanning (1e-3 N/m)
40	41.6
57	39.5
69.7	38.3
70	37.8
80	37.2
80	36.6
100	35.0
100	34.4
110	33.2
120	32.3
140	30.4
160	27.9
180	25.7
200	23.3

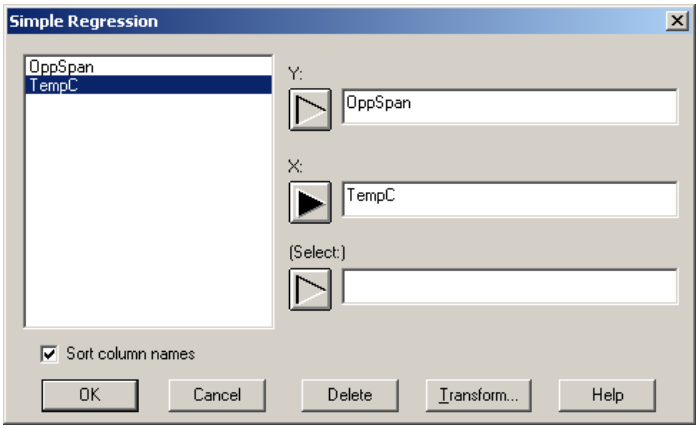


Om deze vraag te beantwoorden maken we gebruik van StatGraphics. We beginnen met het invoeren van de beschikbare meetgegevens. De tekst boven de kolommen is gewijzigd door de kolom met de muis te selecteren door bovenaan de kolom te klikken, vervolgens de rechtermuisknop te gebruiken en dan **Modify Column** te kiezen. De toetsencombinatie Shift-F5 doet hetzelfde.
Nu de meetgegevens zijn ingetypt, kunnen we de regressieberekeningen laten

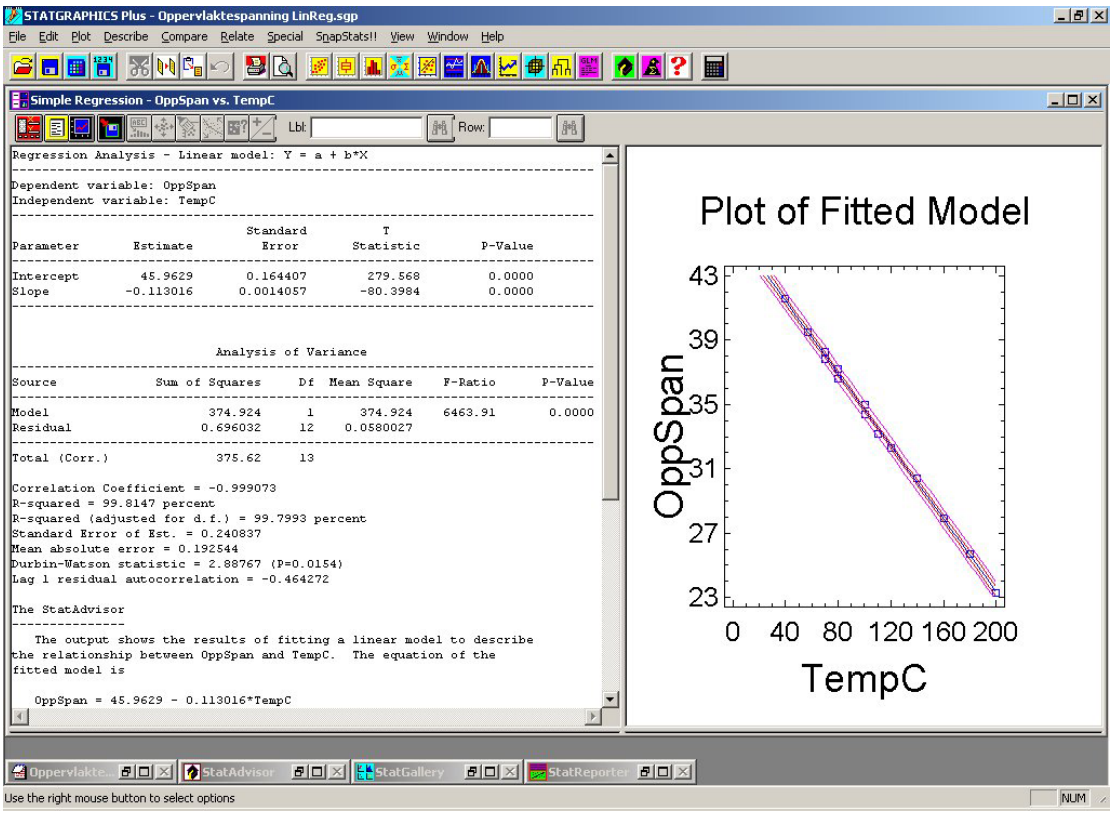
Oppervlakte spanning LinReg.sf3

	TempC	OppSpan	Col_3	Col_4	Col_5	Col_6	Col_7
1	40	41.6					
2	57	39.5					
3	69.7	38.3					
4	70	37.8					
5	80	37.2					
6	80	36.6					
7	100	35					
8	100	34.4					
9	110	33.2					
10	120	32.3					
11	140	30.4					
12	160	27.9					
13	180	25.7					
14	200	23.3					
15							
16							
17							
18							
19							
20							

uitvoeren om de optimale waarden voor β_0 en β_1 te berekenen. In het hoofdmenu van StatGraphics kiezen we voor **Relate**, **Simple Regression**… We krijgen een venster te zien, waarin we moeten aangeven welke kolommen de gegevens voor de afhankelijke variable (Y) en onafhankelijke variable (X) bevat. Klik links op OppSpan en vervolgens op de button met driehoek onder Y (of dubbelklik op OppSpan). Evenzo om TempC aan X toe te kennen. Resultaat:



Klik op de OK-knop en de regressieberekeningen worden uitgevoerd met als resultaat:



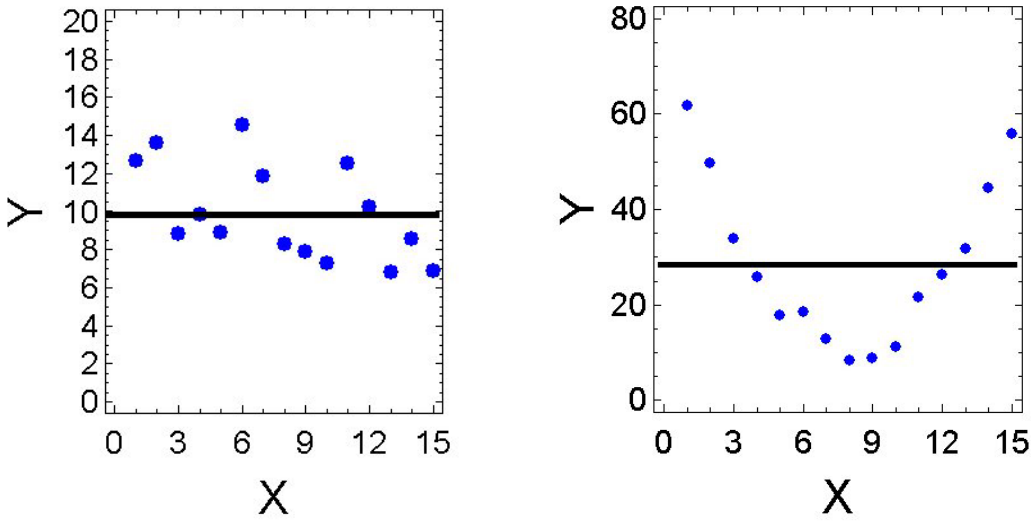
Links zien we de berekenings- en analyseresultaten. Rechts een grafiek met de meetpunten, de regressielijn, het 95% betrouwbaarheidsinterval voor de gefitte regressielijn en het 95% voorspellingsinterval. We zullen de verschillende onderdelen van deze output nu 1 voor 1 gaan bespreken.

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: OppSpan
Independent variable: TempC
------------------------------------------------------------------------------
                                  Standard          T
Parameter       Estimate            Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept         45.9629        0.164407         279.568         0.0000
Slope           -0.113016       0.0014057        -80.3984         0.0000
------------------------------------------------------------------------------
```

Onder de titel " Regression Analysis - Linear model: Y = a + b*X" worden alle resultaten uitgeprint, die te maken hebben met de gevonden parameters in de opgegeven lineaire vergelijking. De waarden van de optimaal aangepaste parameters staan onder de kop " Estimate". Het bovenste getal (Intercept) is de gevonden waarde voor de intercept β_0. Het getal daaronder (Slope) is de gevonden waarde voor β_1.
De gemeten oppervlaktespanningen voor nitrobenzeen in de tabel worden dus beschreven door:
γ (N/m) = 45.9629 . 10^{-3} - 0.113016 . 10^{-3} . T (ºC)

Bij lineaire regressie berekeningen moeten we ons altijd afvragen of de regressie significant is. Dit doen we door te toetsen of parameter β_1 (helling) een waarde heeft, die significant van 0 verschilt. We toetsen dus de hypotheses:
H0: $\beta_1 = 0$, er is geen lineair verband.
H1: $\beta_1 \neq 0$, er is wel een lineair verband.

We hopen dat we de nulhypothese H0 mogen verwerpen. Wordt H0 **NIET** verworpen, dan hebben we de situatie dat β_1 gelijkgesteld moet worden aan 0. Daardoor valt x uit de lineaire vergelijking weg en blijft de vergelijking $y = \beta_0 = \bar{y}$ over. Kennis van de waarde van x levert geen extra kennis op over de bijbehorende waarde voor y. Als deze situatie optreedt, dan is de gekozen vergelijking niet in staat de meetpunten correct te beschrijven. Er moet een nieuwe vergelijking geformuleerd worden en de regressieberekening moet herhaald worden. Grafisch kan deze situatie er als volgt uit zien:



StatGraphics levert ons de gegevens om bovenstaande hypotheses te kunnen toetsen. Van iedere parameter wordt de standaard deviatie (se) gegeven onder de kop "Standard Error". Aangetoond kan worden dat:

$$T = \frac{\beta_1}{se(\beta_1)}$$

een toetsingsgrootheid is voor bovenstaande hypotheses en een Student t-verdeling heeft met, in het geval van enkelvoudige lineaire regressie , n-2 vrijheidsgraden Hierbij is n het totaal aantal meetpunten, in ons geval voor het oppervlaktespanningsvoorbeeld geldt n=14.

Op basis van de berekende waarde voor de toetsingsgrootheid T (T=-80.3984 in dit geval) en het gegeven dat deze toetsingsgrootheid T een Student t_{n-2} verdeling heeft (t_{12} in ons voorbeeld), rekent StatGraphics een maat voor de waarschijnlijkheid van de nulhypothese H0 uit. Deze waarschijnlijkheid heet P-value en wordt door StatGraphics in de gelijknamige kolom gerapporteerd. Als deze mate van waarschijnlijkheid van H0 kleiner is dan een gekozen grenswaarde α, kunnen we de nulhypothese H0 verwerpen. Deze gekozen grenswaarde α is de fout van de

eerste soort en meestal wordt hiervoor de waarde α=0,05 genomen. In ons voorbeeld wordt voor de P-value een waarde 0.0000 berekend. Deze waarde is kleiner dan α=0.05. We kunnen de nulhypothese H0: $\beta_1 = 0$ verwerpen en concluderen dat er een lineair verband is tussen oppervlaktespanning en temperatuur.

Onder de titel "Analysis of Variance" worden alle resultaten uitgeprint, die te maken hebben met de diverse kwadratensommen voor de opgegeven lineaire vergelijking.

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	374.924	1	374.924	6463.91	0.0000
Residual	0.696032	12	0.0580027		
Total (Corr.)	375.62	13			

We hebben gezien dat de optimale waarden voor parameters β_0 en β_1 bepaald worden door de rest-kwadratensom SSE te minimaliseren. Deze rest-kwadratensom SSE wordt berekend met de formule:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

waarin:

n : Aantal meetpunten.

y_i : De y-waarde van het meetpunt (x_i,y_i).

$\hat{y}_i$: Met de aangepaste vergelijking berekende y waarde voor het meetpunt (x_i,y_i)

Als we geen regressiemodel zouden hebben, dan zou $\bar{y}$ voor iedere x_i de beste schatting zijn. Vandaar dat we een totale kwadratensom definiëren ten opzichte van $\bar{y}$ volgens:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Aangetoond kan worden dat kwadratensommen bij elkaar opgeteld mogen worden volgens:

SST = SSR + SSE

Deze vergelijking geeft aan dat van de oorspronkelijke totale kwadratensom SST na het uitvoeren van de regressie een rest-kwadratensom SSE overblijft. Door deze regressie is de oorspronkelijke kwadratensom SST met SSR verminderd, we zeggen dat de regressie een kwadratensom SSR van de totale kwadratensom SST verklaard heeft.

In de gegeven “Analysis of Variance” output van StatGraphics kunnen we onder de kop “Sum of Squares” en achter “Total (Corr.)” aflezen dat voor de oppervlaktespanning versus temperatuur regressie de totale kwadratensom SST=375.62, achter “Residual” dat de rest-kwadratensom SSE=0.696032 is en tot slot achter “Model” dat de kwadratensom verklaard door regressie SSR=374.924 is.

Het belang van deze kwadratensommen is, dat ze een andere mogelijkheid bieden om de significantie van het regressiemodel te toetsen. Met deze kwadratensommen kunnen we namelijk de volgende hypotheses toetsen:
H0: Geen van de parameters in het model is significant
H1: Minstens 1 van de parameters in het model is significant.

Omdat we bij enkelvoudige lineaire regressie maar 1 parameter hebben, namelijk de helling β_1 (het afgesneden stuk β_0 telt hierbij niet mee als parameter), is het toetsen van deze hypotheses identiek aan het eerdere toetsen van de hypotheses H0: $\beta_1 = 0$, H1: $\beta_1 \neq 0$.
Als toetsingsgrootheid gebruiken we:

$$F = \frac{MSR}{MSE}$$

waarin MSR de gemiddelde kwadratensom verklaard door de regressie is. MSE is de gemiddelde rest-kwadratensom. Deze gemiddelde kwadratensommen worden berekend door de kwadratensommen te delen door hun respectievelijke vrijheidsgraden. De totale kwadratensom SST heeft n vrijheidsgraden, waarvan er 1 verloren gaat door het berekenen van $\bar{y}$, blijven er dus *n-1* over. Het aantal vrijheidsgraden van de kwadratensom verklaard door regressie SSR is gelijk aan het aantal parameters in het regressiemodel, waarbij β_0 niet meetelt. Bij enkelvoudige lineaire regressie is dat dus 1. Het aantal vrijheidsgraden voor de rest-kwadratensom SSE is nu het verschil van het aantal vrijheidsgraden van de totale kwadratensom SST en de kwadratensom verklaard door regressie SSR, wat voor enkelvoudige lineaire regressie neerkomt op *n-2* vrijheidsgraden. Onder de kop “Analysis of Variance” in de StatGraphics output worden al deze berekeningen voor ons uitgevoerd en gerapporteerd. Het aantal vrijheidsgraden onder de kop Df (Degrees of freedom), de gemiddelde kwadratensommen MSR en MSE onder de kop “Mean Square” en tot slot de waarde van de toetsingsgrootheid F.
Deze toetsingsgrootheid F heeft een F-verdeling met in de teller het aantal vrijheidsgraden van SSR namelijk 1 en in de noemer het aantal vrijheidsgraden van SSE namelijk *n*-2. Op identieke wijze als besproken bij het toetsen van $\beta_1 = 0$ wordt er een P-value berekend als maat voor de waarschijnlijkheid van H0. Deze P-value heeft een waarde van $0{,}0000 < \alpha = 0{,}05$ dus we kunnen H0 verwerpen en concluderen dat minstens 1 van de parameters in het model significant is. Omdat er maar 1 parameter β_1 is, is deze significant. Dit hadden we eerder ook al aangetoond. Bij enkelvoudige lineaire regressie geven de twee besproken methoden om de significantie van de regressie te toetsen dus een identiek resultaat, voor hun toetsingsgrootheden blijkt dan ook te gelden $F=T^2$.

Van de resterende output bespreken we nog de volgende onderdelen:

```
Correlation Coefficient = -0.999073
R-squared = 99.8147 percent
Standard Error of Est. = 0.240837
```

De correlatiecoëfficiënt R is een maat voor de sterkte van het lineaire verband tussen de onafhankelijke variabele Y en de instelvariabele X. De waarde van R ligt tussen –1 en +1 waarbij waarden in de buurt van –1 en +1 een grote lineaire associatie aangeven. Een waarde voor R rond 0 geeft aan dat er geen lineair ver-

band aanwezig is tussen Y en X. In ons geval is er met een R= –0.999073 dus een zeer sterk negatief lineair verband tussen oppervlaktespanning en temperatuur. Negatief betekent dat de oppervlaktespanning afneemt als de temperatuur toeneemt. Dit klopt met de eerder gegeven grafiek van oppervlaktespanning versus temperatuur.

Het kwadraat van de steekproefcorrelatiecoëfficiënt R^2 wordt de determinatiecoefficiënt genoemd. De coefficiënt R^2 kan berekend worden volgens de formule:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Omdat de restkwadratensom SSE altijd tussen 0 en de totale kwadratensom SST ligt, volgt hieruit dat R^2 een waarde tussen 0 en 1 zal hebben. De betekenis van R^2 is dat de waarde van R^2 aangeeft welk gedeelte van de totale variantie in de afhankelijke variabele Y verklaard wordt door het regressiemodel. Vandaar dat StatGraphics de waarde van R^2 presenteert in de vorm van een percentage. Voor de meetgegevens van oppervlaktespanning en temperatuur wordt dus 99.8147 % van de variantie in oppervlaktespanning verklaart door het gefitte lineaire regressiemodel.

Het lineaire regressiemodel heeft de volgende algemene vorm:

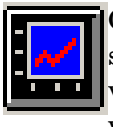
$$Y = \beta_0 + \beta_1 x + \varepsilon$$

waarin aangenomen wordt dat de residuen $\varepsilon \sim NID(0, \sigma^2)$ verdeeld zijn. Aangetoond kan worden dat de gemiddelde restkwadratensom MSE een schatting is voor de variantie σ^2 van de residuen. Deze schatting kan dus direct afgelezen worden uit de “Analysis of Variance” gegevens in de StatGraphics output. Voor de oppervlaktespanning meetgegevens volgt hieruit een waarde voor de schatting van σ^2 gelijk aan 0.0580027. De wortel hieruit is een schatting voor de standaarddeviatie σ van de residuen. De waarde daarvan wordt door StatGraphics berekend en gerapporteerd als Standard Error of Est. = 0.240837.

Reeds eerder is uitgelegd dat het uitvoeren van deze lineaire regressie volgens de methode van de kleinste kwadratensom onder andere is gebaseerd op het uitgangspunt, dat de gekozen vergelijking de meetgegevens zo goed mogelijk beschrijft. De residuen zijn dan een gevolg van onwillekeurige en onbeheersbare invloeden (experimentele ruis). De residuen zijn bij benadering normaal verdeeld met verwachtingswaarde 0 en constante variantie σ^2. In de praktijk is dit een belangrijke controle of de vergelijking gebruikt mag worden om de meetgegevens te beschrijven. Om deze controle uit te voeren kijken we naar de grafiek van de studentized residuals tegen $\hat{y}_i$.

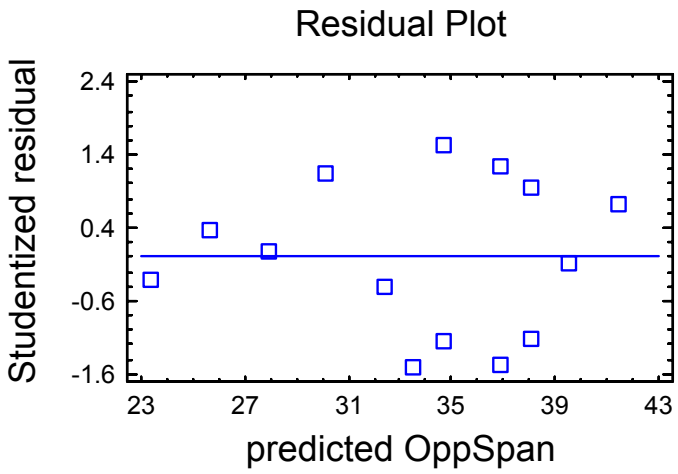
Studentized residuals (sr) zijn residuen, die gestandaardiseerd zijn door ze te delen door hun standaardafwijking σ. In formule vorm:

$sr_i = (y_i - \hat{y}_i) / \sigma = e_i / \sigma$

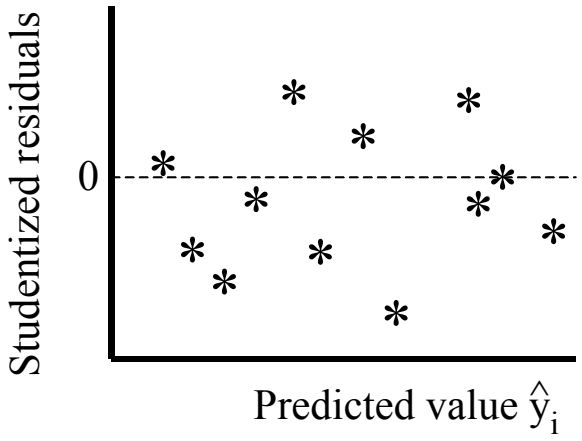


Om deze grafiek op te roepen klikken we in het Simple Regression venster van StatGraphics op het icoon dat links staat afgebeeld. Dit icoon bevindt zich op derde plaats van links. Als de muispijl boven dit icoon staat wordt de hint tekst ‘Graphical options’ getoond. Na aanklikken van dit

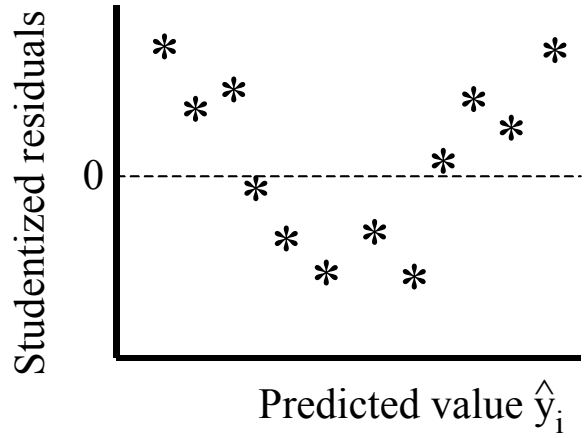
icoon verschijnt een menu met 5 grafieken, die aangevinkt kunnen worden. We vinken de grafiek 'Residuals versus Predicted' aan. De volgende grafiek verschijnt:



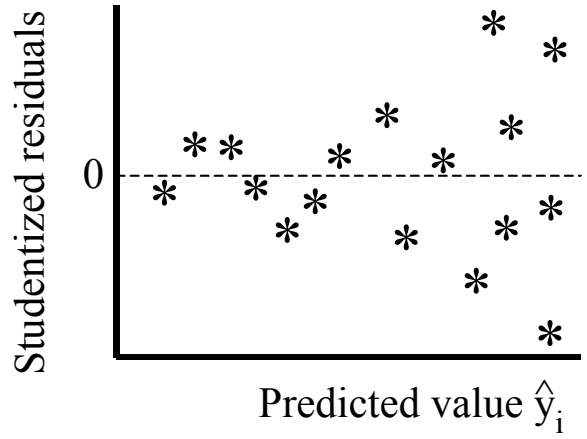
De grafiek van studentized residuals tegen predicted values $\hat{y}_i$ moet een structuurloos, random patroon rondom de getrokken referentie lijn $sr_i=0$ te zien geven. Daarnaast mogen de afwijkingen ten opzichte van de referentie lijn $sr_i=0$ niet afhankelijk zijn van $\hat{y}_i$. Bijvoorbeeld groter worden als $\hat{y}_i$ groter wordt. We geven nu een drietal veel voorkomende patronen in de grafieken van studentized residuals tegen predicted values $\hat{y}_i$.



Deze residualplot is een voorbeeld van een correcte modelaanpassing, die voldoet aan de aannamen. We zien het gewenste structuurloze, schot-hagelpatroon van de residuen rondom de referentielijn $sr_i=0$. Men moet niet te lang naar dit soort grafieken kijken: als er op het eerste gezicht niet duidelijk een verband te zien is, dan is de modelaanpassing wat dit onderdeel betreft in orde. Als men lang kijkt, ontdekt men altijd wel een patroon.



Bovenstaande residualplot vertoont een veel voorkomend patroon. Hieruit moet de conclusie getrokken worden dat de gekozen vergelijking niet in staat is de meetgegevens correct te beschrijven. Hier is een kwadratisch verband beschreven met een vergelijking voor een rechte lijn. Als men een dergelijk patroon tegenkomt, is er een probleem met de gekozen vergelijking. Het beste en meest voor de hand liggende is het formuleren van een nieuw model met een nieuwe vergelijking en het herhalen van de regressieberekeningen. Dit echter niet altijd mogelijk. In dat geval zou men de vergelijking kunnen accepteren, echter met een zeer duidelijke vermelding van de geconstateerde tekortkomingen. Nog een mogelijkheid is dat een aantal meetpunten niet aan het model voldoen en daardoor aanleiding geven tot dit patroon. Dit komt bijvoorbeeld voor bij opstartverschijnselen. Zoals bij het opwarmen van een bol, pas bij voldoende grote tijd wordt het verband tussen de logaritme van de temperatuur en de tijd een rechte lijn. Worden de korte opwarmtijden meegenomen in het aanpassen van deze rechte lijn, dan resulteert deze residualplot. In dat geval laat men de meetgegevens met korte opwarmtijden weg en herhaalt men de regressieberekeningen. Of men breidt de modelvergelijking uit met de hogere orde termen en voert een meervoudige lineaire regressie uit. Een patroon als in deze residualplot vereist dus een zorgvuldige evaluatie van de modelvergelijking en de meetgegevens.

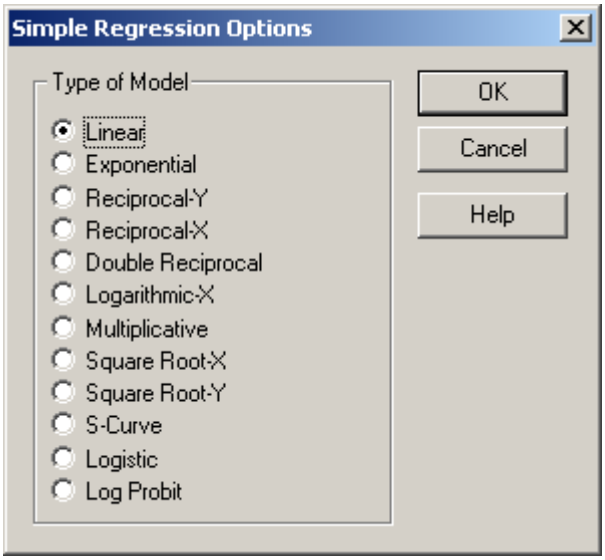


De derde residualplot is een voorbeeld van een situatie waarbij de residuen toenemen bij toenemende $\hat{y}_i$. Hier wordt dus niet voldaan aan de eis dat de spreiding in de residuen constant moet zijn. Oplossingen voor dit probleem zijn:

- Het transformeren van y bijvoorbeeld 1/y of ln(y) en dan pas de regressie uitvoeren. Door keuze van de juiste transformatie van y lukt het meestal de

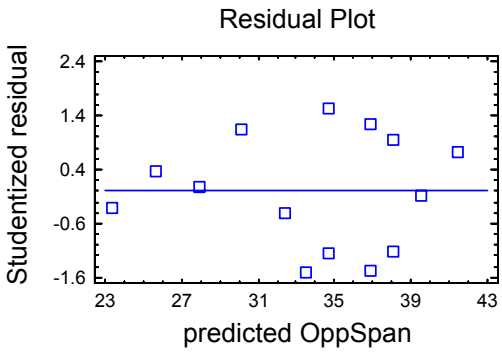
spreiding in de residuen nu wel constant te maken. Transformaties van y kunnen eenvoudig uitgevoerd worden in StatGraphics via:

- De button Transform bij het definiëren van de afhankelijke en onafhankelijke variabelen voor de lineaire regressie.
- Voer een gewone lineaire regressie uit en klik in het venster met de analyse resultaten met de rechtermuisknop. Er verschijnt een menu waar uit 12 verschillende transformaties gekozen kan worden.

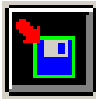


Het uitvoeren van een **gewogen regressie**. Bij het berekenen van de kwadratensom worden dan niet alle meetpunten gelijkwaardig meegenomen maar eerst vermenigvuldigd met een weegfactor w_i. Hoe groter de weegfactor, hoe zwaarder de punten meetellen in de berekening van de kwadratensom. Deze weegfactor w_i kan dan omgekeerd evenredig genomen worden met de waarde van de afhankelijke variabele y. Hoe groter y, hoe kleiner de weegfactor en hoe minder het punt meetelt bij het berekenen van de kwadratensom. Nauwkeurige meetpunten tellen nu zwaarder. Deze gewogen regressieberekeningen kunnen weer op vrijwel identieke wijze met StatGraphics uitgevoerd worden via de nog te bespreken menukeuze 'Multiple regression'.

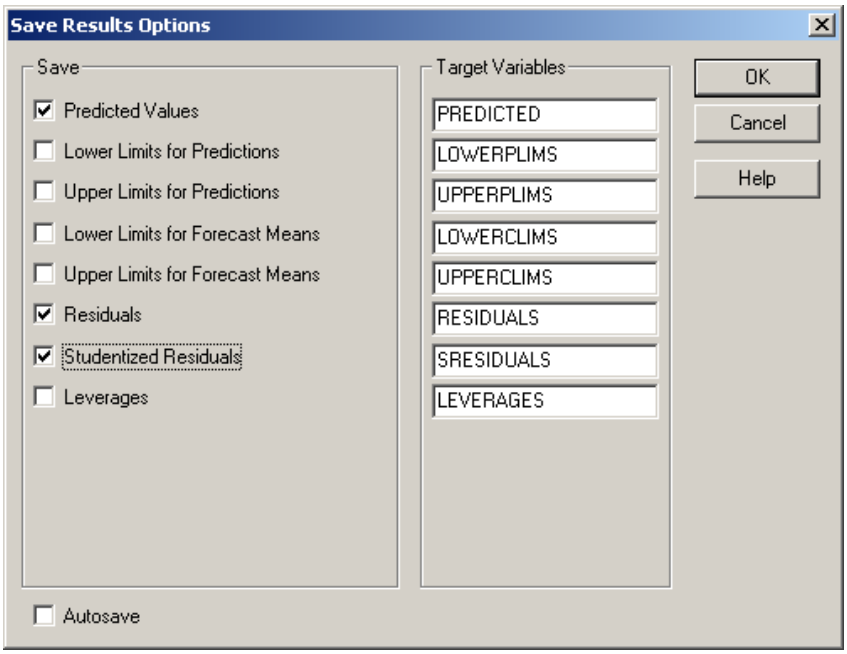
Kijken we nog even naar de grafiek van studentized residuals tegen predicted values $\hat{y}_i$ voor het beschrijven van de oppervlaktespanning van nitrobenzeen, dan kan geconcludeerd dat deze residualplot een bevredigend structuurloos beeld geeft en dat de gekozen lineaire vergelijking de meetgegevens goed kan beschrij-



ven.

Naast deze residualplot is het belangrijk om te controleren of de residuen normaal verdeeld zijn. Om deze controle met StatGraphics uit te voeren, moeten we de residuen eerst toevoegen aan het spreadsheet met meetgegevens, de datafile. Dit bereiken we door in het venster met de analyse resultaten van de lineaire regressie te klikken op het icoon 

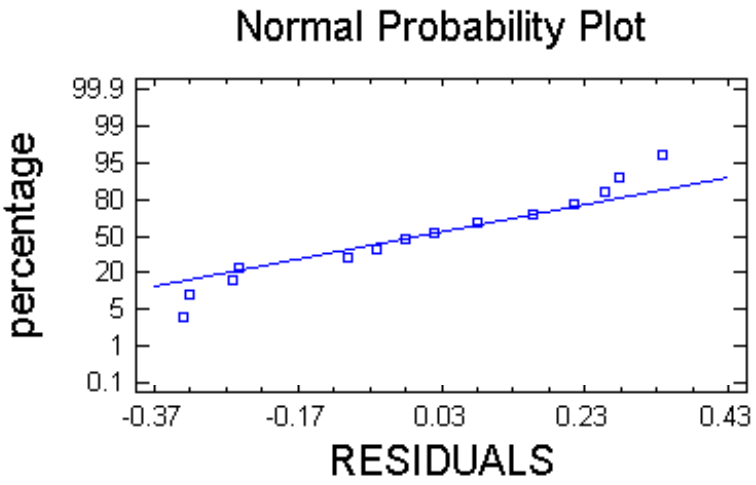
Dit is het vierde icoon van links en geeft als hint 'Save results'. Er verschijnt het volgende venster:



We vinken Predicted Values, Residuals en Studentized Residuals aan en klikken op OK. Aan het spreadsheet met meetdata worden nu de kolommen PREDICTED, RESIDUALS en SRESIDUALS toegevoegd. Kolom PREDICTED bevat de berekende waarden voor $\hat{y}_i$.

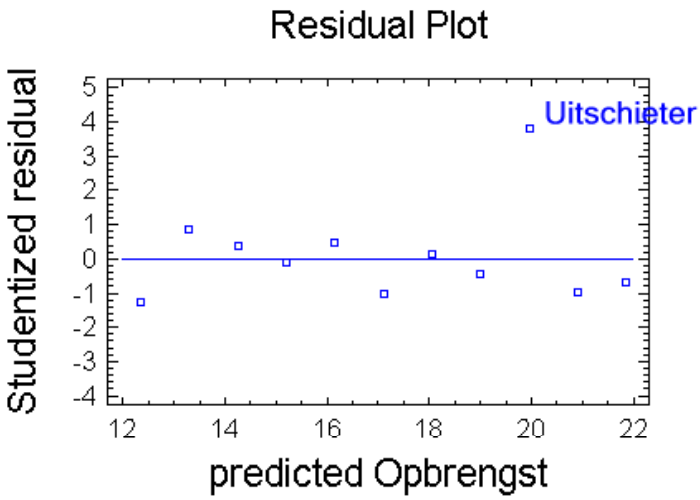
Om te onderzoeken of deze residuen normaal verdeeld zijn kiezen we in het StatGraphics hoofdmenu voor SnapStats!! en vervolgens voor One Sample Analysis. Er verschijnt een venster met een ruime hoeveelheid statistische informatie over de residuen. Hiervan zijn twee onderdelen voor ons interessant.

Als eerste kijken we naar de normal probability plot, zoals behandeld in hoofd-



stuk 1.
Naast deze grafische controle is er een formele statistische toets van Shapiro-Wilks op de normaliteit (zie hoofdstuk 1). Op basis van de berekende waarde voor deze toetsingsgrootheid W, berekent StatGraphics weer de P-value uit. De P-value is een maat voor de waarschijnlijkheid van de nulhypothese H0 uit. Onder Diagnostics zien we staan Shapiro-Wilks P-value = 0.3669. Als grenswaarde hanteren we 1% of 0,01. De berekende P-value van 0.3669 is niet kleiner dan de grenswaarde 0,01 en er is dus geen reden om de nul-hypothese te verwerpen. Er is dus geen reden om aan te nemen dat onze residuen niet normaal verdeeld zouden zijn.

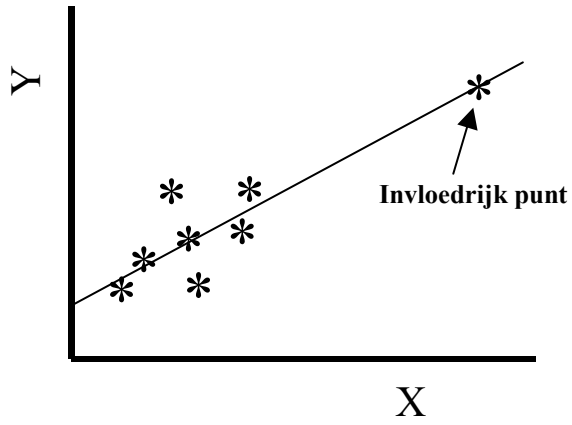
In de praktijk komt het heel vaak voor dat de residuen niet normaal verdeeld zijn, omdat er 1 of meerdere foute meetgegevens aanwezig zijn. Oorzaken hiervan kunnen storingen in de apparatuur zijn (lekkage, spanningsuitval, vervuilde kolom), schrijffouten (64 i.p.v. 46) en vele andere mogelijkheden. Het is belangrijk te onderzoeken of er potentiële uitbijters aanwezig zijn in de meetgegevens, zodat ze verwijderd en bij voorkeur opnieuw gemeten kunnen worden. Over het algemeen dient men zeer terughoudend te zijn om meetgegevens als uitbijter aan te merken en te verwijderen uit de set meetgegevens. Meetgegevens verwijderen zou men uitsluitend moeten doen, als er ook een duidelijke oorzaak aangegeven kan worden waarom de meetgegevens niet goed zijn.
Controle op de aanwezigheid van potentiële uitbijters in de meetgegevens is eenvoudig binnen StatGraphics. We gebruiken hiervoor de residualplot, waarin de studentized residuals zijn uitgezet tegen de berekende y-waarden. Potentiële uitbijters vallen in de residualplot meestal direct op, doordat ze er in de grafiek ook echt uitschieten. Als grenswaarde wordt 2.5 gehanteerd, meetpunten met een studentized residual groter 2.5 of kleiner –2.5 worden als potentiële uitbijter aangemerkt.



Als er potentiële uitbijters aanwezig zijn, is het in de praktijk handig om ze in relatie met alle meetgegevens te bekijken. Daarvoor hebben we reeds een kolom aan de spreadsheet met meetgegevens toegevoegd, waarin de waarden van de studentized residual per meting zijn gekomen, SRESIDUALS. Waarden in deze kolom SRESIDUALS >2.5 of <-2.5 wijzen de potentiële uitbijters aan.
Ook is het mogelijk om StatGraphics een onderzoek te laten uitvoeren naar uitbijters genaamd Unusual Residuals. Klik in het lineaire regressie venster op het tweede icoon van links met de hint 'Tabular options'.

Vink in het venster dat verschijnt de keuze 'Unusual Residuals' aan. Er wordt een nieuw venster toegevoegd binnen het lineaire regressie venster met een overzicht van de alle meetpunten met een absolute studentized residual waarde groter 2.
Controleren we op deze wijze of er uitbijters voorkomen in de meetgegevens voor de oppervlaktespanning van nitrobenzeen, dan concluderen we dat deze niet aanwezig zijn.

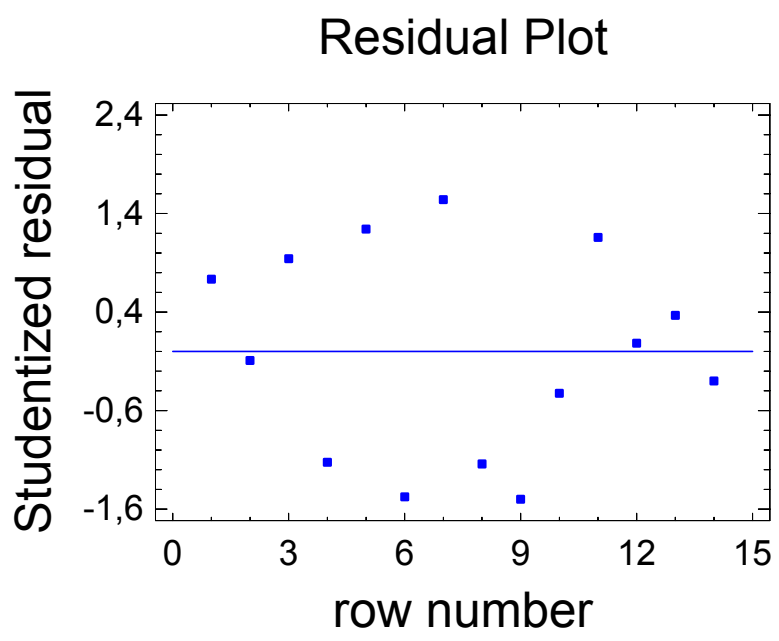
Naast deze meetpunten die uitbijters kunnen zijn en bij voorkeur gecorrigeerd moeten worden, is er een ander belangrijk type meetpunten, de zogenaamde **invloedrijke punten**. Dit zijn meetpunten die een onevenredig grote invloed hebben op de parameter waarden β_0 en/of β_1. In onderstaande grafiek een voorbeeld van een dergelijk invloedrijk punt:



De ligging van het invloedrijke punt rechtsboven (ook wel **hefboompunt** genoemd) bepaalt vrijwel volledig de helling β_1 van de regressielijn. We moeten er dus zeker van zijn dat dit meetpunt goed gemeten is. Bij voorkeur zullen we er naar streven om extra metingen te doen rondom invloedrijke punten en op deze manier de regressielijn beter vast te leggen. In tegenstelling tot uitbijters, die op een verstoring van het meetsysteem duiden en daarom verwijderd of gecorrigeerd moeten worden, zijn invloedrijke punten geen verstoring en mogen NOOIT verwijderd worden. Invloedrijke meetpunten dienen extra gecontroleerd te worden en indien mogelijk uitgebreid te worden met extra metingen. Een web site die dit laat zien is te vinden op http://www.stat.sc.edu/~west/javahtml/Regression.html .
Het onderzoek naar invloedrijke punten (Influential points) kan StatGraphics voor ons uitvoeren. Net als bij 'Unusual Residuals' klikken we in het lineaire regressie venster op het tweede icoon van links met de hint 'Tabular options' en vinken de keuze 'Influential Points'. Er wordt een nieuw venster toegevoegd binnen het lineaire regressie venster met een overzicht van alle invloedrijke meetpunten. StatGraphics rekent hierbij voor ieder meetpunt een waarde voor de zgn. leverage uit. De waarde van de grootheid leverage is een maat voor de invloed die dit meetpunt heeft op de parameterwaarden β_0 en β_1. Is de leverage van een meetpunt 3 keer groter dan de gemiddelde leverage van de andere meetpunten, dan wordt dit meetpunt als invloedrijk gekarakteriseerd en in het overzicht getoond. Voeren we deze analyse uit voor de oppervlaktespanning meetdata van nitrobenzeen, dan zien we dat er geen invloedrijke punten zijn in deze set meetgegevens.

Tenslotte dienen we te onderzoeken of de waarnemingen onderling onafhankelijk zijn. Ook dit doen we via de residuen. Enerzijds door de residuen tegen de tijd te bestuderen via een grafiek, anderzijds via een formele toets (de toets van Durbin-Watson). Deze laatste toets wordt automatisch uitgevoerd in StatGraphics bij een regressie-analyse. De toets van Durbin-Watson geeft aan dat de data gecorreleerd zou kunnen zijn. Echter een residual plot geeft geen duidelijk patroon te zien. Wel is het vreemd dat de residuen in het begin afwisselend positief en negatief zijn. Hoewel dit nader onderzoek verdient, is er in dit geval geen reden om aan de uitkomst van het regressie-model te twijfelen.

```
Standard Error of Est. = 0,240837
Mean absolute error = 0,192544
Durbin-Watson statistic = 2,88767 (P=0,0154)
Lag 1 residual autocorrelation = -0,464272
```



Hiermee is het beschrijven van de oppervlaktespanning van nitrobenzeen als functie van de temperatuur via enkelvoudige lineaire regressie gereed. De waarden van de parameters β_0 en β_1 zijn bepaald, beide parameters zijn significant en de residuen voldoen aan de aannamen dat ze normaal verdeeld zijn met een constante spreiding. Er is vastgesteld dat de set meetgegevens geen uitbijters en geen invloedrijke punten bevat.

3.3 Lack-of-fit toets

In het voorgaande voorbeeld werden geen van de metingen herhaald. Indien dit wel het geval is, kunnen modelafwijkingen ook via een zogenaamde lack-of-fit toets onderzocht worden. Het idee achter deze toets is twee verschillende schat-

tingen voor de variantie van de meetfout te vergelijken: de gewone en één gebaseerd op de varianties binnen de herhaalde waarnemingen. Het is goed om het verschil te zien tussen significantie toetsen en lack-of-fit toetsen:

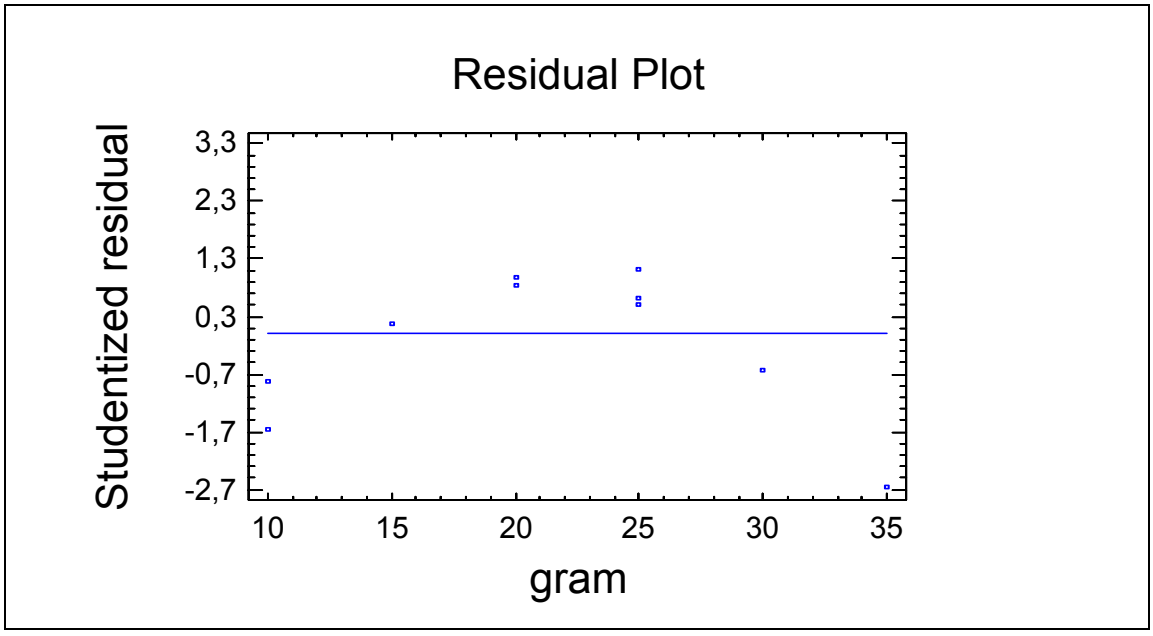
- significantie toetsen: toetsen of de onafhankelijke variabele de verschillende uitkomsten van de afhankelijke variabele geheel of gedeeltelijk verklaart. De nulhypothese is dat er geen verband is tussen beide variabelen
- lack-of-fit toetsen: toetsen of het model de uitkomsten van de metingen voldoende verklaart. De nulhypothese is hier dat het gegeven model de metingen adequaat beschrijft; de alternatieve hypothese is dat er een ander model is dat de werkelijkheid beter beschrijft.

Als voorbeeld nemen we meetgegevens van de groeisnelheid van ratten die een bepaald voedingssupplement krijgen. Uit eerdere onderzoekingen bleek dat deze groeisnelheid een lineaire functie is van het aantal grammen voedselsupplement. Als we in StatGraphics het menu **Relate**, **Simple Regression ...** kiezen en dan bij de Tabular Options de optie **Lack-of-Fit tests** aanvinken, dan krijgen we de volgende uitvoer te zien:

Analysis of Variance with Lack-of-Fit

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	24,5017	1	24,5017	0,29	0,6076
Residual	686,398	8	85,7998		
Lack-of-Fit	659,398	4	164,85	24,42	0,0045
Pure Error	27,0	4	6,75		
Total (Corr.)	710,9	9			

Aangezien de P-waarde van de lack-of-fit toets onder de 0,05 ligt, volgt hieruit dat het lineaire model bij deze waarnemingen niet goed past. De toets geeft echter geen inzicht in de reden hiervoor. Daarom is het aan te raden naast deze toets ook altijd naar de residuen te kijken. In dit geval is duidelijk te zien wat de reden is voor het falen van het lineaire model:



Stappenplan: enkelvoudige lineaire regressie

1. Voer de lineaire regressieberekeningen uit met StatGraphics.
2. Controleer of de parameters significant zijn. Mochten één of meerdere parameters niet significant zijn, formuleer dan een nieuw model en start opnieuw met het uitvoeren van de regressie. Indien er meerdere waarnemingen zijn met dezelfde waarde van de instelvariabele, voer dan een lack-of-fit toets uit.
3. Bestudeer de residualplot waarin voor ieder meetpunt de studentized residual uitgezet wordt tegen de berekende waarde $\hat{y}_i$. Onderzoek of deze grafiek een onwillekeurig, random patroon te zien geeft, waarin bovendien de grootte van de residuen constant moet zijn. Vertoont deze grafiek een duidelijk patroon, dan voldoet het gekozen model niet. Formuleer een ander model en herhaal de regressieberekeningen of geef duidelijk de beperkingen van het gekozen model aan.
4. Onderzoek de residuen op de aanwezigheid van uitbijters. Potentiële uitbijters zijn te herkennen:
   - In de residualplot waar ze er echt uit moeten schieten en y-waarden groter 2.5 en kleiner –2.5 hebben.
   - In de SRESIDUALS kolom in de datasheet aan absolute waarden groter 2.5.

Komt men tot de conclusie dat er uitbijters aanwezig zijn in de meetgegevens, corrigeer deze uitbijters dan indien mogelijk en voer de regressie opnieuw uit. Wees terughoudend in het zomaar weggooien van meetpunten.
5. Controleer de normaliteit van de residuen. Bekijk de normaliteitsplot van de residuen en toets de normaliteit formeel met de Shapiro-Wilks toetsingsgrootheid.
6. Controleer of de waarnemingen onderling onafhankelijk zijn via een residual plot tegen het waarnemingsnummer en de toets van Durbin-Watson.
7. Onderzoek of in de set meetgegevens invloedrijke punten voorkomen. Als dit het geval is, verzamel dan extra meetgegevens in de buurt van deze invloedrijke punten en voer de regressie opnieuw uit.

Als een modelvergelijking voldoet aan alle bovenstaande eisen, dan is dit nog steeds geen bewijs dat de modelvergelijking juist is. Er zijn vele nonsens vergelijkingen (bijv. verband tussen aantal ooievaars en aantal geboorten) die perfect voldoen aan alle bovenvermelde eisen, maar desondanks niet juist zijn. Echter voor de procestechnoloog geldt dat als de modelvergelijking gebaseerd is op correcte fysisch/chemische grondslagen en het aanpassen van deze modelvergelijking aan de meetgegevens voldoet aan bovenvermelde eisen, met redelijke zekerheid mag worden aangenomen dat de modelvergelijking goed is.