5 Niet-lineaire regressie

Als laatste van de soorten regressie zullen we in dit hoofdstuk de niet-lineaire regressie bespreken. Dit zijn modellen waarin de modelparameters(meestal aangegeven met β_i) niet lineair voorkomen in de modelvergelijking. Zoals in hoofdstuk 2 aangegeven, kan dit eenvoudig gecontroleerd worden door partiële afgeleiden naar de parameters uit te rekenen. Veel chemische verschijnselen kunnen met differentiaalvergelijkingen beschreven worden. De oplossingen van deze differentiaalvergelijkingen zijn vaak niet-lineaire functies. Dit verklaart waarom niet-lineaire regressie een zeer belangrijk hulpmiddel is in de chemie en procestechnologie.

Kernbegrippen van dit hoofdstuk:

- lineariseerbare modellen
- intrinsiek niet-lineaire modellen
- niet-lineaire regressie
- startwaarden
- invloedrijke punten
- normaliteit
- betrouwbaarheidsintervallen voor parameters
- toetsen van significantie regressiemodel
- determinatiecoëfficiënt
- residuenplots
- overfitting
- Marquardt algoritme

Er zijn drie manieren om een regressie-analyse van een niet-lineair model uit te voeren:

1. transformatie tot lineair model
2. benadering door lineair model (linearisatie)
3. niet-lineaire regressie uitvoeren

Niet-lineaire modellen kunnen soms via transformaties lineair gemaakt worden. Een voorbeeld is het exponentiële groeimodel:

$$y = \beta_0 e^{\beta_1 x}$$

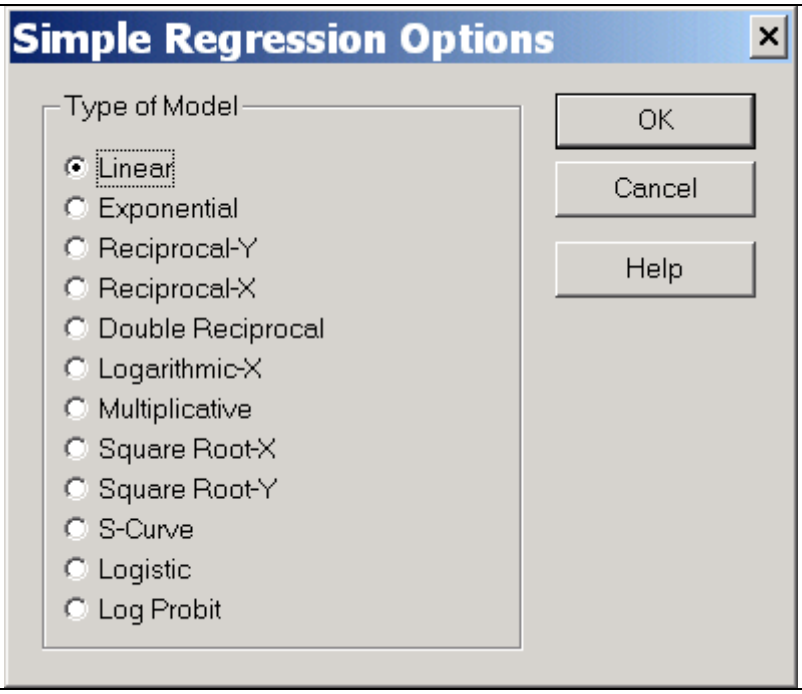
Door transformatie met behulp van ln raken we de e-macht kwijt:

$$\ln(y) = \ln(\beta_0) + \beta_1 x$$

$$y_1 = b_0 + b_1 x$$

We hebben de modelvergelijking nu omgeschreven naar een enkelvoudige lineaire regressie van $y_1=\ln(y)$ versus x. Bedenk echter dat de foutterm additief in het eindmodel moet voorkomen. Dit heeft tot gevolg dat de foutterm in het oorspronkelijke model multiplicatief moet zijn. Bij getransformeerde modellen moet men dus extra goed controleren dat de modelveronderstellingen juist zijn (via de gebruikelijke residuenplots e.d.). Niet-lineaire modellen die via een transformatie omgeschreven kunnen worden naar een lineair model heten **intrinsiek lineaire modellen**. StatGraphics kan een dergelijke transformatie uitvoeren voor een groot aantal enkelvoudige regressiemodellen via het menu **Relate, Simple Regression**. Indien men de rechtermuisknop in het venster met de regressie-

uitvoer klikt, komt een menu tevoorschijn waarin via **Analysis Options** men kiezen tussen verscheidene intrinsiek lineaire modellen. De standaardkeuze is natuurlijke het lineaire model.



Met name het multiplicatieve model $Y = aX^b$ en het S-kromme model $Y = e^{a+b/X}$.
Als er geen transformatie is tot een lineair model of de foutterm van het getransformeerde model voldoet niet aan de gebruikelijke veronderstellingen, gaat men vaak over op een 1e-orde Taylorbenadering (lineariseren). Dit is af te raden, omdat men niet kan aangeven hoe deze benadering doorwerkt op de statistische eigenschappen van de parameterschattingen.
Modellen die niet te transformeren zijn tot een lineair model heten **intrinsiek niet-lineair**. Een overzicht van veelgebruikte niet-lineaire modellen is te vinden in

Naam	Formule	Opmerking
exponentieel groeimodel	$Y = \beta_0 e^{\beta t_1}$	$Y(0) = \beta_0$
Mitscherlich model	$Y = \beta_0 \left(1 - e^{-\beta_1 (x+\delta)}\right)$	β_0 is horizontale asymptoot
inverse polynomiaal model	$Y = \frac{x+\delta}{\beta_0 + \beta_1 (x+\delta)}$	$1/\beta_1$ is horizontale asymptoot
logistiek groeimodel	$Y = \frac{\beta_0}{1 + \beta_2 e^{-\beta_1 x}}$	$Y(0) = \frac{\beta_0}{1+\beta_2}$ β_0 is horizontale asymptoot
Gompertz groeimodel	$Y = \beta_0 \quad e^{-\beta_2 e^{-\beta_1 x}}$	β_0 is horizontale asymptoot
Von Bertalanffy model	$Y = \left(\beta_0^{1-m} - \theta \quad e^{-\beta_1 x}\right)^{1/(1-m)}$	
Michaelis-Menten model	$Y = \alpha_1 \left(1 - e^{-\lambda_1 x}\right) + \alpha_2 \left(1 - e^{-\lambda_2 x}\right)$	$\alpha_1 + \alpha_2$ is horizontale asymptoot

Tabel 5.1: Overzicht van veelgebruikte niet-lineaire modellen

Zoals reeds aangegeven in het hoofdstuk 2, is het uitvoeren van niet-lineaire regressie berekeningen een stuk lastiger dan lineaire regressie berekeningen. Dit komt omdat er geen analytische oplossing mogelijk is van de vergelijkingen, die de waarden van de parameters geven waarvoor de restkwadratensom minimaal is. Door het ontbreken van deze analytische oplossing, moeten de waarden van de parameters waarvoor de restkwadratensom minimaal is, iteratief bepaald worden.

We beginnen met een voorbeeld van een iteratieve berekening, namelijk het vinden van een nulpunt voor de vergelijking: $f(x) = x^2 - 2 = 0$. Als startwaarde beginnen kiezen we $x=1$. Voor $x=1$ heeft $f(x)$ de waarde –1 en de afgeleide $f'(x) = 2x$ de

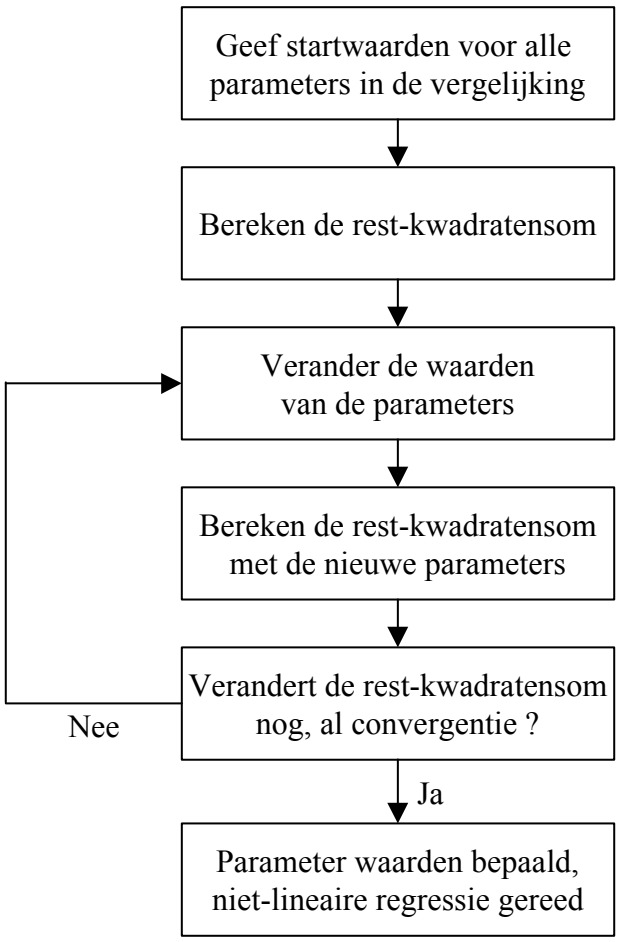
waarde 2. Daar voor $x=1$ de waarde van $f(x)$ flink van 0 verschilt', is duidelijk dat $x=1$ niet de waarde x is die we zoeken. Op basis van de gegevens die we nu hebben, zoeken we dus een betere waarde voor x. Dit doen we als volgt. Uit de definitie van differentiaalquotiënt volgt dat $f(x+\Delta x) \approx f(x) + f'(x) \cdot \Delta x$. Hierin is Δx de verandering van onze x waarde om dichter bij het gezochte nulpunt te komen. Omdat we een nulpunt zoeken, kunnen we stellen $f(x+\Delta x) = 0$. Er blijft dan over: $0 = f(x) + f'(x) \cdot \Delta x$ en hieruit kunnen we afleiden $\Delta x = -\frac{f(x)}{f'(x)}$

Uit onze gegevens kunnen we dus berekenen $\Delta x = 0{,}5$, zodat $x = 1 + 0{,}5 = 1{,}5$ onze nieuwe schatting voor het gezochte nulpunt wordt. Voor $x=1{,}5$ heeft $f(x)$ de waarde $0{,}25$ en dat is nog steeds niet voldoende nauwkeurig het nulpunt dat we zoeken. We moeten bovenstaand rekenproces dus nog een aantal keren herhalen. Dit wordt iteratie genoemd. In schema:

x	$f(x)=x^2-2$	$f'(x)=2x$	$\Delta x = -f(x)/f'(x)$	x_{nieuw}
1	-1	2	0.5	1.5
1.5	0.25	3	-0.0833	1.4167
1.4167	0.00694	2.833	-0.0025	1.4142
1.4142	6.0073E-06	2.828	-2.1239E-06	1.4142

Al na 4 iteraties zijn we het nulpunt $x=\sqrt{2}=1{,}4142$ zeer dicht genaderd. Deze berekeningsmethode is uitgevonden door Newton (1643-1727).

Voor niet-lineaire regressie, waar we niet een nulpunt maar het minimum van de restkwadratensom zoeken, ziet het iteratie proces er schematisch als volgt uit:



Het grootste probleem van niet-lineaire regressie is dat de berekeningen vaak niet volgens bovenstaand schema willen verlopen. De meest voorkomende problemen zijn:

- Niet willen convergeren.
- Extreme waarden voor de parameters, waardoor de modelvergelijkingen niet meer berekend kunnen worden.
- Het optreden van lokale minima van de restkwadratensom. Het rekenproces lijkt dan geconvergeerd, echter bij een start met andere beginwaarden voor de parameters worden er andere optimale waarden voor de parameters gevonden.

Deze problemen treden vooral op als meetgegevens en model niet zo goed bij elkaar passen. In het geval van lineaire regressie verliepen de berekeningen altijd probleemloos en werden we op deze situatie geattendeerd doordat één of meerdere parameters niet significant waren. Bij niet-lineaire regressie hebben we het niet zo gemakkelijk en moeten we ons steeds realiseren, dat moeizame niet-lineaire regressie berekeningen kunnen wijzen op een niet bij de meetgegevens passende modelvergelijking.

We bespreken nu een aantal onderdelen van het niet-lineaire regressie schema in meer detail.

5.1 Bepalen van de startwaarden voor de parameters.

Bij niet-lineaire regressie moeten we beginwaarden opgeven voor alle parameters in de modelvergelijking. Bij lineaire regressie was dit niet nodig. Het bepalen van goede beginwaarden voor de parameters is van enorm belang voor een succesvolle niet-lineaire regressieberekening. Het is een lastig probleem, dat we met wat elementair wiskundig vernuft zo goed mogelijk moeten zien op te lossen. De risico's verbonden aan de keuze van slechte beginwaarden voor de parameters zijn:

- Er zijn zeer veel iteraties nodig om de minimale rest-kwadratensom te vinden. Omdat iedere iteratie nogal wat rekenwerk met zich meebrengt, betekent dit dat de hele regressieberekening lang kan gaan duren.
- Het blijkt niet mogelijk om de parameters zo aan te passen, dat er een minimale rest-kwadratensom gevonden wordt. De niet-lineaire regressie berekening wil **niet convergeren**. Dit is in de praktijk het meest voorkomende probleem met lineaire regressie.
- Er wordt wel een minimale rest-kwadratensom gevonden, maar dit is niet de kleinst mogelijke. We zijn op een **lokaal minimum** gestuit, terwijl we op zoek zijn naar het globale minimum. Met andere, betere startwaarden voor de parameters was dit globale minimum van de rest-kwadratensom wel gevonden.

Voor het vinden van goede startwaarden voor de parameters is het zaak de modelvergelijking eens goed te bekijken. Parameters hebben soms een interpretatie in het model als asymptoot of modelwaarde bij een gegeven waarde (bijv. 0). Voorbeelden hiervan staan in Tabel 5.1. Vaak kunnen we door het toepassen van een transformatie (ln, reciproque, enz.) de niet-lineaire vergelijking omschrijven naar een lineaire vergelijking. Indien dit niet mogelijk is, De parameters in deze nieuwe lineaire vergelijking kunnen we dan met de reeds besproken lineaire regressie procedures in StatGraphics bepalen. Uit de parameters van de lineaire vergelijking kunnen vervolgens direct de startwaarden voor de parameters in de niet-lineaire vergelijking berekend worden. De lineaire regressie is dan een eerste stap voor het ruwweg bepalen van de parameter waarden waarna verfijning van de parameter waarden plaatsvindt met niet-lineaire regressie. Twee voorbeelden om dit toe te lichten.

Gegeven is de modelvergelijking:

$$y = \beta_0 e^{\beta_1 x}$$

Zoals eerder genoemd, kunnen we met behulp van logaritmen de e-macht kwijt raken:

$$\ln(y) = \ln(\beta_0) + \beta_1 x$$
$$y_1 = b_0 + b_1 x$$

We hebben de modelvergelijking nu omgeschreven naar een enkelvoudige lineaire regressie van y_1=ln(y) versus x. Deze lineaire regressie kunnen we direct uitvoeren in StatGraphics en daarmee de waarden voor b_0 en b_1 bepalen. De startwaarden voor de niet-lineaire regressie worden dan: $\beta_0 = e^{b_0}$ en $\beta_1 = b_1$. Het is in dit geval ook mogelijk StatGraphics de transformatie te laten uitvoeren door in het regressievenster met de rechtermuisknop te klikken en bij **Analysis Options** het exponentiële model te kiezen.

Als tweede voorbeeld noemen we de vergelijking van Antoine voor het beschrijven van verzadigde dampspanningen van vloeistoffen als functie van de temperatuur:

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

waarin:

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

Uit de thermodynamische theorie kan de vergelijking van Clausius-Clapeyron afgeleid worden voor het verband tussen de verzadigde dampspanningen van een zuivere vloeistof en de temperatuur:

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

Deze uit de theorie afgeleide vergelijking blijkt echter maar voor kleine temperatuur trajecten een goede beschrijving te geven. Door het toevoegen van de C parameter wordt dit temperatuur traject een stuk groter. Voor de niet-lineaire regressie is het op grond hiervan logisch om parameter C de startwaarde 0 te geven. De overblijvende vergelijking van Clausius-Clapeyron kunnen we dan weer transformeren met ln om de e-macht kwijt te raken:

$$\ln\left(P_s\right) = A - \frac{B}{T}$$
$$y = b_0 + b_1 x$$

Via een enkelvoudige lineaire regressie van y=ln(P_s) versus $x=1/T$ kunnen de waarden voor b_0 en b_1 uit de meetgegevens bepaald worden. Hiermee zijn dan de startwaarden van de niet-lineaire regressie bepaald volgens:

$$A = b_0$$
$$B = -b_1$$
$$C = 0$$

Het is in dit geval ook mogelijk StatGraphics de transformatie te laten uitvoeren door in het regressievenster met de rechtermuisknop te klikken en bij **Analysis Options** het S-kromme model te kiezen.

Verder zijn er in de modelvergelijking vaak parameters, die een bepaalde fysische betekenis hebben. Een voorbeeld is een parameter die de maximale belading van een adsorbens geeft. De startwaarde van deze parameter kan dan geschat

worden uit de meetgegevens, door in dit geval naar de beladingen bij de hoogst gemeten concentraties te kijken.

5.2 Algoritmen van de parameters om de kleinste rest-kwadratensom te vinden.

Algoritmen die de rest-kwadratensom minimaliseren als functie van de parameters zijn de essentie van de niet-lineaire regressie berekeningen. Het is van belang dat dit in zo weinig mogelijk stappen de minimale rest-kwadraten-som gevonden wordt. Het behoeft geen nadere uitleg dat dit een uiterst lastig probleem is. In de loop der jaren is een aantal methoden ontwikkeld om dit probleem aan te pakken.

Binnen StatGraphics zijn 3 van deze algoritmen beschikbaar:

- Gauss-Newton algoritme
- De "steilste helling" of gradiënt algoritme. Dit algoritme convergeert buitengewoon langzaam en het gebruik van dit algoritme wordt dan ook afgeraden.
- Marquardt algoritme, de standaard methode die StatGraphics bij niet-lineaire regressie gebruikt.

Het standaard Marquardt algoritme zal meestal voldoen. Deze methode is een slim compromis tussen de twee eerstgenoemde algoritmen. Het komt voor dat de Marquardt methode buitengewoon langzaam convergeert en letterlijk duizenden iteraties nodig heeft om te convergeren. In dat geval is het verstandig over te schakelen op de Gauss-Newton methode, deze convergeert dan meestal een stuk sneller. De Gauss-Newton methode heeft echter de neiging om over minima heen te schieten.

5.3 Convergentie

Een belangrijk punt bij de algoritmen bij niet-lineaire regressie is de bepaling van het tijdstip waarop het algoritme stop.t StatGraphics gebruikt bij zijn niet-lineaire regressie berekeningen twee criteria om vast te stellen dat de minimale rest-kwadratensom gevonden is en dat de iteratieve berekeningen beëindigd kunnen worden. Hierbij geldt dat de berekeningen gestopt worden als aan minstens één van de twee criteria voldaan wordt. StatGraphics rapporteert welk criterium dat is.

Criterium 1 is dat de rest-kwadratensom SSE niet meer verandert:

$$\frac{SSE_{i-1} - SSE_i}{SSE_i + 10^{-6}} < 10^{-5}$$

SSE_{i-1} : Rest-kwadratensom bij de $(i\text{-}1)^e$ iteratie.
SSE_i : Rest-kwadratensom bij de i^e iteratie.

Criterium 2 is dat alle parameters β in de modelvergelijking niet meer veranderen:

$$\frac{\beta_{p,i-1} - \beta_{p,i}}{\beta_{p,i} + 10^{-6}} < 10^{-4}$$

$\beta_{p,i-1}$: Waarde van de p^e parameter bij de $(i\text{-}1)^e$ iteratie.

$\beta_{p,i}$: Waarde van de p^e parameter bij de i^e iteratie.

De waarden voor de convergentie criteria 10^{-5} en 10^{-4} kunnen in StatGraphics aangepast worden als ze bijvoorbeeld tot niet voldoende nauwkeurige resultaten leiden. Om te voorkomen dat berekeningen te lang duren stopt StatGraphics ook als een van tevoren ingesteld aantal iteraties wordt bereikt. Standaard staan de niet-lineaire regressieberekeningen in StatGraphics ingesteld op maximaal 30 iteraties en maximaal 200 functie aanroepen, dat wil zeggen berekening van de modelvergelijking. In de praktijk blijkt dit vaak te weinig te zijn. Bovendien zijn deze waarden gekozen voor computers die waarschijnlijk veel langzamer zijn dan wat tegenwoordig gangbaar is. Schroom niet om dit aantal flink hoger te maken bijvoorbeeld 300 en 2000.

5.4 De vergelijking van Fritz en Schluender

Laten we het één en ander eens op een niet-lineaire vergelijking toepassen. We bekijken hiervoor de evenwichtsbeladingen van fenol op actieve kool vanuit een vloeistoffase. Een bekende actieve kool is Norit. Naast fenol is in de vloeistoffase ook de stof p-nitrofenol aanwezig. Moleculen fenol en p-nitrofenol strijden samen om de beschikbare adsorptieplaatsen op het oppervlak van de actieve kool. We spreken van competitieve adsorptie. Hoe kleiner de concentratie p-nitrofenol, hoe groter de evenwichtsbelading van fenol op de actieve kool. Actieve kool wordt gebruikt in de drinkwaterzuivering om kleine concentraties van dit soort chemische verbindingen uit het water te verwijderen.

Onderstaande tabel geeft evenwichtsbeladingen voor fenol (q1) op actieve kool bij verschillende vloeistofconcentraties fenol (C1) en p-nitrofenol (C2). Deze meetgegevens zijn afkomstig uit een artikel van Fritz en Schluender (W.Fritz, U.Schluender, "Simultaneous adsorption equilibria of organic solutes in dilute aqueous solutions on active carbon.", Chem. Eng. Science, (1974), Vol.29, 1279-1282).

Concentratie fenol C1 (mMol/l)	Concentratie p-nitrofenol C2 (mMol/l)	Belading fenol q1 (mMol/g)	Concentratie fenol C1 (mMol/l)	Concentratie p-nitrofenol C2 (mMol/l)	Belading fenol q1 (mMol/g)
0.53	0	1.9	0.21	0.1	0.14
1.37	0	2.36	1.2	0.1	0.67
1.95	0	2.53	2.29	0.1	1.01
2.91	0	2.79	2.65	0.1	1.16
3.64	0	2.97	3.35	0.1	1.36
4.4	0	3.08	4.04	0.1	1.52
4.78	0	3.02	4.66	0.1	1.63
5.4	0	3.18	5.75	0.1	1.85
6.3	0	3.33	6.34	0.1	1.97
6.88	0	3.37	7.2	0.1	2.14
0.83	0.02	1	0.84	0.5	0.17
1.14	0.02	1.26	1.42	0.5	0.31
2.19	0.02	1.76	2.34	0.5	0.46
2.87	0.02	1.99	2.66	0.5	0.55
3.1	0.02	2.05	3.41	0.5	0.67
4.14	0.02	2.25	4.1	0.5	0.79

4.85	0.02	2.44	5.24	0.5	0.92
5.66	0.02	2.61	5.6	0.5	0.98
6.21	0.02	2.65	6.29	0.5	1.07
6.79	0.02	2.75	7.39	0.5	1.23

In dit artikel stellen zij ook een vergelijking voor om de evenwichtsbelading van fenol bij verschillende concentraties p-nitrofenol te beschrijven. Deze vergelijking is opgebouwd vanuit de Freundlich isotherm, waarmee de evenwichtsbelading van 1 component op actieve kool beschreven kan worden:

$$q_1 = aC_1^{\,b}.$$

Met de volgende vergelijking wordt deze Freundlich isotherm uitgebreid om de evenwichtsbelading van fenol q_1 te corrigeren voor de aanwezigheid van p-nitrofenol C_2. Hoe groter C_2, hoe groter de correctie en hoe lager de evenwichtsbelading van fenol op de actieve kool q_1:

$$q_1 = \frac{aC_1^{\,b+X_1}}{C_1^{\,X_1} + X_2 \quad C_2^{\,X_3}}$$

waarin:

q_1 : Evenwichtsbelading fenol (mMol/g kool).
C_1 : Concentratie fenol in vloeistoffase (mMol/l).
C_2 : Concentratie p-nitrofenol in vloeistoffase (mMol/l).
a : Parameter single solute Freundlich isotherm.
b : Parameter single solute Freundlich isotherm.
X_1 : Parameter.
X_2 : Parameter.
X_3 : Parameter.

Merk op dat deze vergelijking zich reduceert tot de Freundlich isotherm voor $C_2=0$.

Gevraagd wordt nu de parameter waarden a, b, X_1, X_2 en X_3 te bepalen, waarmee deze Fritz en Schluender vergelijking de evenwichtsbelading meetgegevens in de tabel zo goed mogelijk beschrijft.

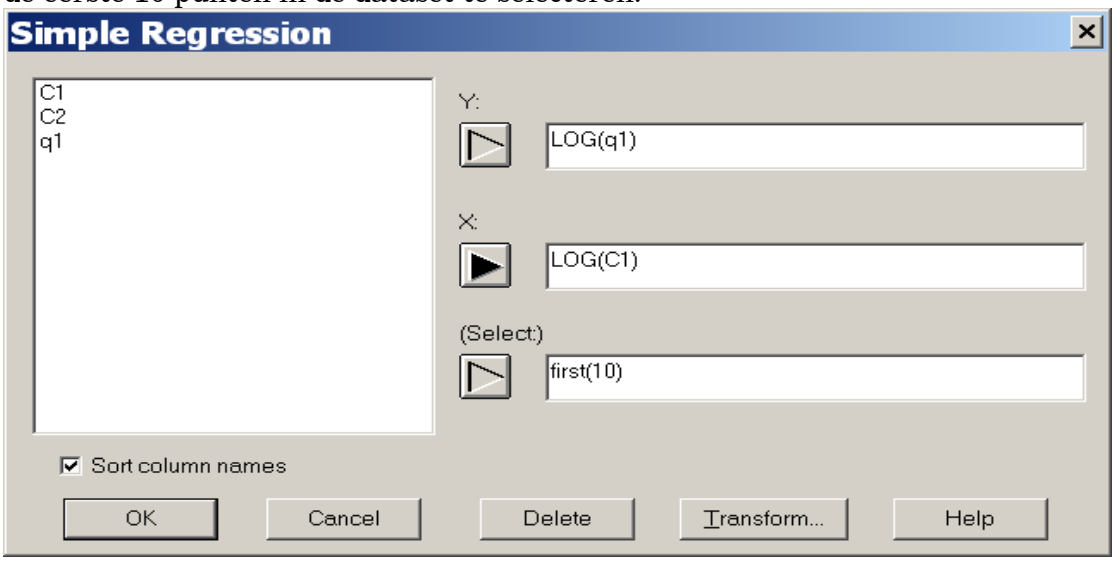
Eerst moeten we de startwaarden voor de parameters a, b, X_1, X_2 en X_3 bepaald worden. Dit is niet eenvoudig voor een dergelijke redelijk ingewikkelde vergelijking. We zullen dit in stappen doen en beginnen met de parameters a en b van de single-solute Freundlich isotherm. De eerste 10 meetwaarden in de tabel hebben namelijk C2=0 en kunnen hier mooi voor gebruikt worden.

$$q_1 = aC_1^{\,b}$$

$$\ln(q_1) \;=\; \ln(a) \;+\; b\;\ln(C_1)$$

$$y = b_0 + b_1 x$$

Via een enkelvoudige lineaire regressie van y=ln(q1) versus x=ln(C1) kunnen de waarden voor b0 en b1 uit de meetgegevens bepaald worden. Dit kan in StatGraphics via het menu **Relate, Simple Regression**. Maak bij de definitie van de variabelen gebruik van de **Transform** button en de selectie optie first (10) om de eerste 10 punten in de dataset te selecteren.



Resultaat van de lineaire regressie:
- parameter b_0 = 0,79, hieruit volgt startwaarde a = $e^{0,79}$ = 2,20.
- parameter b_1 = 0,22, hieruit volgt startwaarde b = 0,22.

Een andere mogelijkheid is om als **Analysis Option** (op te roepen via de rechtermuisknop) het multiplicatieve model te kiezen. Dit geeft natuurlijk dezelfde waarden voor a en b. Nu de startwaarden voor a en b bepaald zijn, kunnen uit de overige meetgegevens met C2≠0 de startwaarden voor X1, X2 en X3 bepaald worden. Hiervoor moet wel even met de vergelijking gewerkt worden. We beginnen met het nemen van de reciproque van q1:

$$q_1 = \frac{aC_1^{\,b+X_1}}{C_1^{\,X_1} + X_2 \;\; C_2^{\,X_3}}$$

$$\frac{1}{q_1} = \frac{C_1^{\,X_1} + X_2 \;\; C_2^{\,X_3}}{aC_1^{\,b+X_1}}$$

$$\frac{1}{q_1} = \frac{1}{aC_1^{\,b}} + \frac{X_2 \;\; C_2^{\,X_3}}{aC_1^{\,b+X_1}}$$

$$\frac{1}{q_1} - \frac{1}{aC_1^{\,b}} = \frac{X_2 \;\; C_2^{\,X_3}}{aC_1^{\,b+X_1}}$$

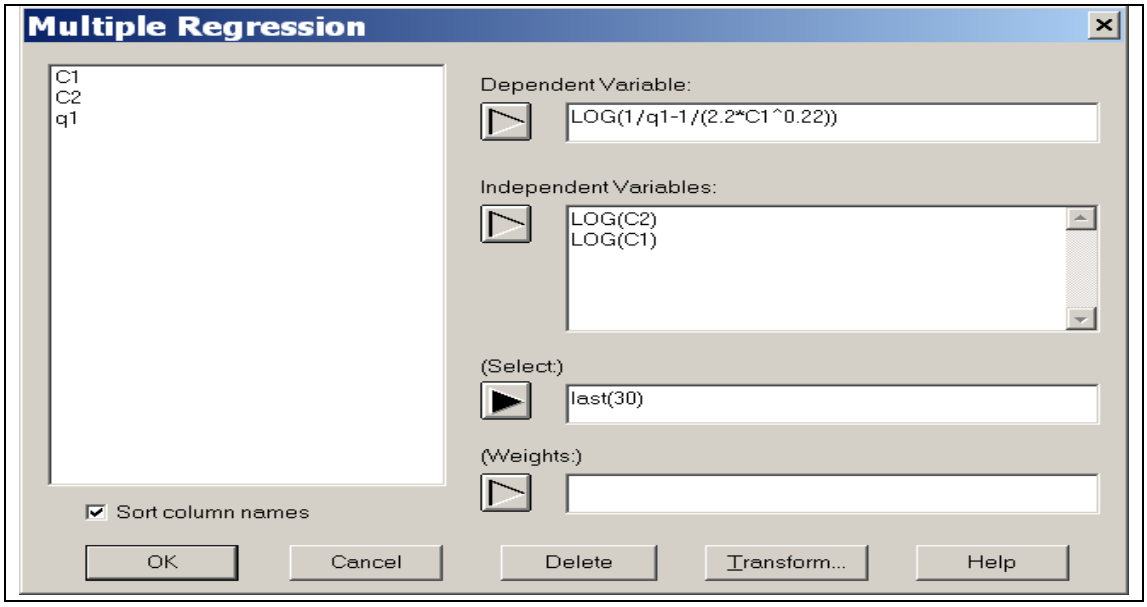
$$y = \frac{X_2 \;\; C_2^{\,X_3}}{aC_1^{\,b+X_1}}$$

$$\ln(y) = \ln(X_2) + X_3 \ln(C_2) - \ln(a) - (b + X_1)\ln(C_1)$$

$$\ln(y) = \ln\left(\frac{X_2}{a}\right) + X_3 \ln(C_2) - (b + X_1)\ln(C_1)$$

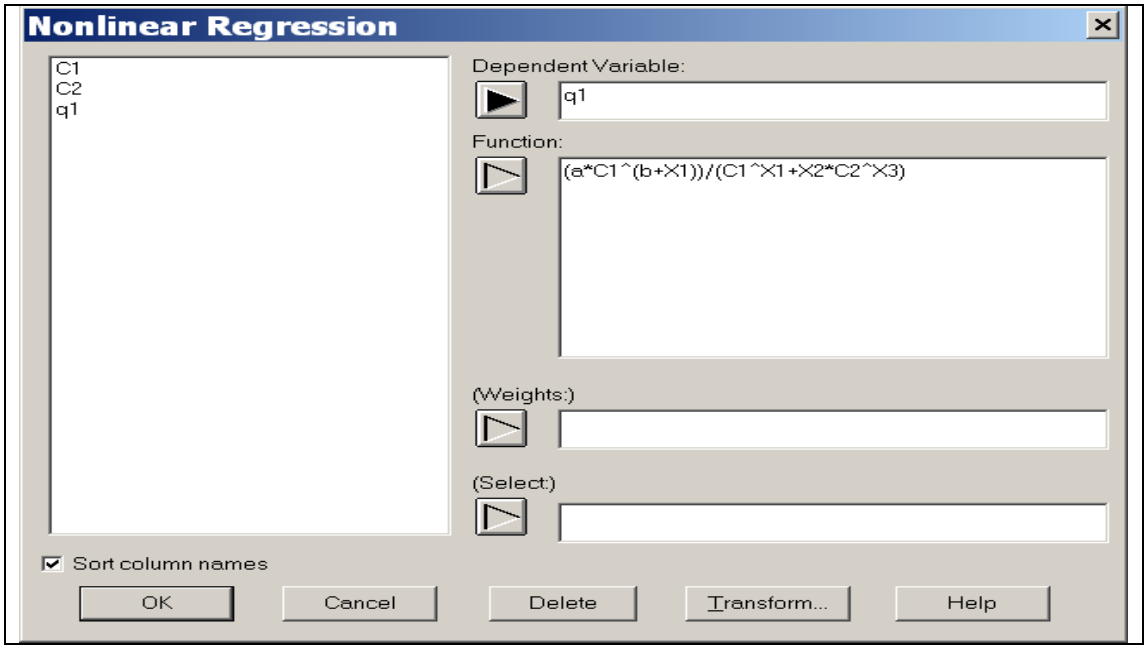
$$\ln(y) = \beta_0 + \beta_1 \ln(C_2) + \beta_2 \ln(C_1)$$

Via het menu **Relate**, **Multiple Regression** kunnen we deze resulterende vergelijking fitten aan onze meetgegevens met *C2≠0*, d.w.z. de laatste 30 meetgegevens. Merk op dat we eerste meetgegevens met *C2=0* niet kunnen gebruiken, omdat we de natuurlijke logaritme van de *C2* waarden gebruiken.

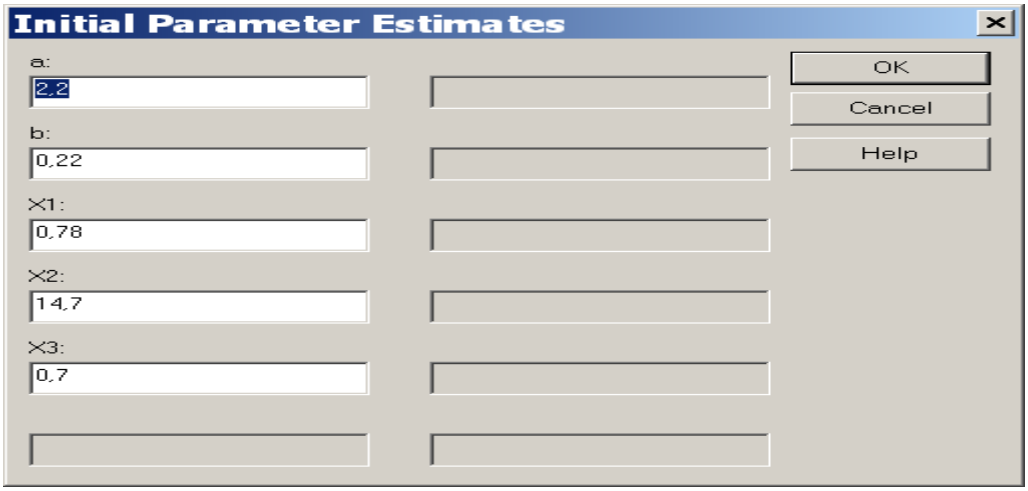


Resultaat van bovenstaande meervoudige lineaire regressie:
Parameter β_0 = 1,9, hieruit volgt door a=2,2 in te vullen startwaarde X2 = 14,7.
Parameter β_1 = 0,70, hieruit volgt startwaarde X3 = 0,70.
Parameter β_2 = -1,0, hieruit volgt startwaarde X1 = 0,78.

Hiermee zijn we voldoende voorbereid om de niet-lineaire regressie van de Fritz and Schluender vergelijking uit te voeren met StatGraphics. We kiezen in het hoofdmenu voor **Special, Advanced Regression**, **Non-linear Regression**. Evenals bij de lineaire regressie krijgen we een venster te zien waarin we de afhankelijke en onafhankelijke variabelen moeten opgeven. Echter bij niet-lineaire regressie is de onafhankelijke variabele vervangen door **Function**. Hier moeten we de vergelijking ingeven gebruik makend van de namen die we de kolommen in de datasheet met meetgegevens hebben gebruikt. Onze kolommen zijn C1, C2 en q1 genoemd. Na invullen ziet de definitie van de niet-lineaire regressieberekening er als volgt uit:



Na klikken op OK verschijnt een venster voor de startwaarden voor de parameters.



Na nogmaals klikken op OK wordt de niet-lineaire regressieberekening uitgevoerd.

```
Nonlinear Regression
--------------------
Dependent variable: q1
Independent variables:
   C1
   C2

Function to be estimated: (a*C1^(b+X1))/(C1^X1+X2*C2^X3)
Initial parameter estimates:
   a = 2,2
   b = 0,22
   X1 = 0,78
   X2 = 14,7
   X3 = 0,7

Estimation method: Marquardt
Estimation stopped due to convergence of parameter estimates.
Number of iterations: 3
Number of function calls: 19

Estimation Results
------------------------------------------------------------------------------
                                                    Asymptotic 95,0%
                                        Asymptotic  Confidence Interval
Parameter              Estimate  Standard Error     Lower         Upper
------------------------------------------------------------------------------
a                       2,19901       0,0151932    2,16817       2,22986
b                      0,220247      0,00461501   0,210878      0,229616
X1                     0,778566       0,0191097   0,739771      0,817361
X2                      14,0536        0,511334    13,0156       15,0917
X3                     0,691835      0,00792686   0,675742      0,707927
------------------------------------------------------------------------------

Analysis of Variance
-----------------------------------------------------
Source             Sum of Squares     Df  Mean Square
-----------------------------------------------------
Model                    157,296      5      31,4593
Residual                0,0259535     35   0,00074153
-----------------------------------------------------
Total                    157,322     40
Total (Corr.)             35,207     39

R-Squared = 99,9263 percent
R-Squared (adjusted for d.f.) = 99,9179 percent
Standard Error of Est. = 0,027231
Mean absolute error = 0,0193274
Durbin-Watson statistic = 1,98329
Lag 1 residual autocorrelation = -0,00577943
```

Begonnen wordt met een samenvatting van de uitgevoerde niet-lineaire regressie, welke variabelen gebruikt zijn, de gefitte vergelijking en de gebruikte startwaarden voor de parameters. Daarna volgt informatie over de berekeningen: de gebruikte methode, welk criterium aanleiding gaf tot convergentie en het aantal iteratie en functie aanroepen. Door onze zorgvuldige voorbereiding van de startwaarden van de parameters zijn maar 3 iteraties nodig om convergentie te bereiken. Dit zouden er heel wat meer kunnen zijn.
Daarna het overzicht van de gevonden optimale parameter waarden. Omdat dit niet-lineaire regressie berekeningen zijn, ziet deze output er iets anders uit vergeleken met de output van de lineaire regressie. Zo ontbreekt de duidelijke indicatie over de significantie van de parameters. Wel wordt de asymptotische standaardafwijking (asymptotic standard error) berekend. Asymptotisch wil zeggen dat de berekende waarde niet helemaal correct is en dichter bij de juiste waarde komt als het aantal meetpunten groter wordt. Ruwweg zijn 25 of meer meetpunten nodig om een redelijk juiste waarde te krijgen. Op basis van deze asymptoti-

sche standaardafwijking rekent StatGraphics dan vervolgens een asymptotisch 95% betrouwbaarheidsinterval voor de parameter waarden uit.

Bij de enkelvoudige en meervoudige lineaire regressie in de vorige hoofdstukken hebben we geleerd, dat we bijzonder goed op moeten letten dat 0 niet in het 95% betrouwbaarheidsinterval van een parameter ligt. Dit is namelijk equivalent met het toetsen of die parameter significant van 0 verschilt. De vorm van een lineair model is namelijk zodanig, dat als één van de parameters β_i gelijk aan 0 is, de bijbehorende verklarende variabele x_i uit het model wegvalt. Dit komt door de vermenigvuldiging. Een in het model gekozen verklarende variabele (bijv. temperatuur, druk, diameter, enz) valt dan weg, en dat is nogal dramatisch. De gekozen verklarende variabele blijkt geen bijdrage te kunnen leveren aan het beschrijven van y. Vandaar dat we daar geweldig goed op moeten letten om geen "onzin" verklarende variabelen in ons model te krijgen.

Bij niet-lineaire regressie ligt dit niet zo duidelijk. Kijken we bijvoorbeeld naar de parameters A en C in de Antoine vergelijking:

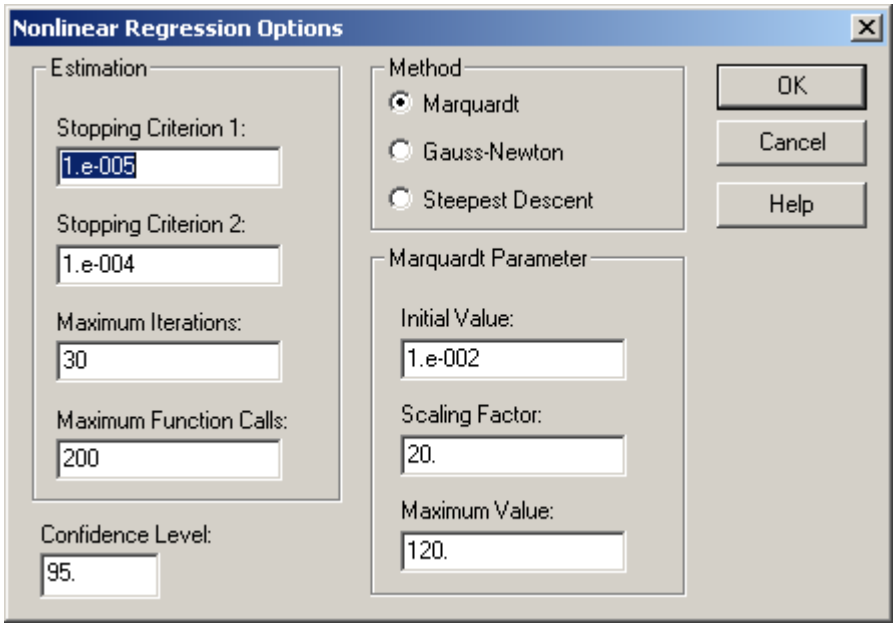
$$P_s = e^{A-\frac{B}{T+C}}$$

We zien dat waarde 0 voor de parameters A en C niet tot gevolg heeft, dat de verklarende variabele temperatuur T uit het model wegvalt. Indien parameter B de waarde 0 krijgt, dan valt de temperatuur T wel weg. Dus voor dit model is het essentieel om te controleren dat de waarde voor de parameter B significant van 0 verschilt. Indien de parameters A en C een waarde 0 hebben, dan is dit niet zo erg omdat de temperatuur in het model blijft staan. Voor $C=0$ gaat de vergelijking van Antoine over in de originele vergelijking van Clausius-Clapeyron.

Bij niet-lineaire vergelijkingen komt het veelvuldig voor dat de parameters een fysische betekenis hebben. Vaak zijn we geïnteresseerd in de waarde van de fysische grootheden. In de vergelijking van Antoine is de parameter B gelijk aan de verdampingswarmte gedeeld door de gasconstante R. In dat geval willen we de parameter zo nauwkeurig mogelijk bepalen. We bekijken dan het 95% betrouwbaarheid interval en proberen deze zo klein mogelijk te krijgen. Is dit 95% betrouwbaarheid interval te groot naar onze zin, dan kunnen we de proefopzet wijzigen, bijv. meer metingen doen of meetpunten in een bepaald gebied.
Ook komt het voor dat het asymptotisch 95% betrouwbaarheidsinterval voor een parameter waarde opvallend groot is omdat de parameter onder bepaalde condities uit de vergelijking wegvalt. Dit is bijvoorbeeld het geval bij hoge luchtvochtgehaltes voor de Cg parameter in de GAB vergelijking, een vergelijking die waterdampsorptie isothermen beschrijft. Ten aanzien van de Fritz en Schluender vergelijking geeft het 95% betrouwbaarheidsinterval van de parameters een indicatie dat er voldoende meetpunten waren om de parameters te bepalen. Daar kunnen we zondermeer tevreden mee zijn.

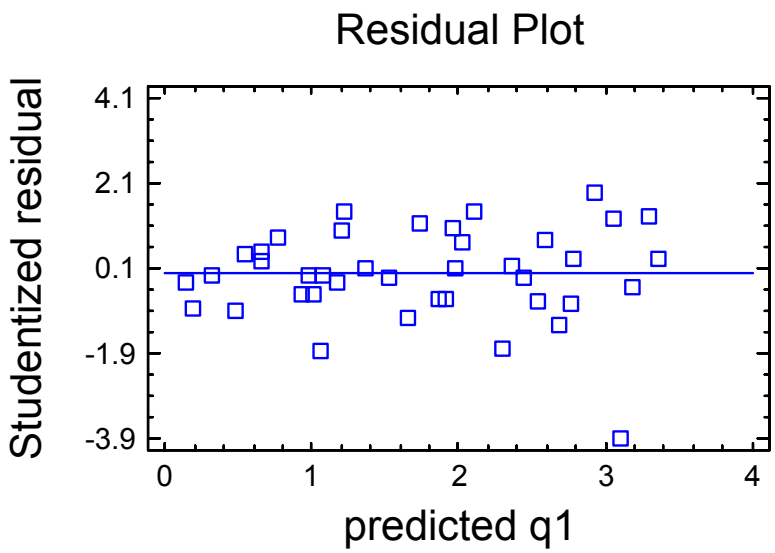
Onder de titel "Analysis of Variance" worden alle resultaten uitgeprint, die te maken hebben met de diverse kwadratensommen. Omdat het niet-lineaire regressie is ontbreken de F-toetsen.

Een aantal onderdelen van de niet-lineaire regressie berekeningen zoals de Marquardt methode en de convergentie criteria staan standaard ingesteld. Om deze te wijzigen klikken we in het analyse venster van de 'Non-linear regression' op de rechtermuisknop. Er verschijnt een menu waarin we voor Analysis Options kiezen. Vervolgens komt het volgende venster, waarin diverse onderdelen van de regressie berekeningen ingesteld kunnen worden:



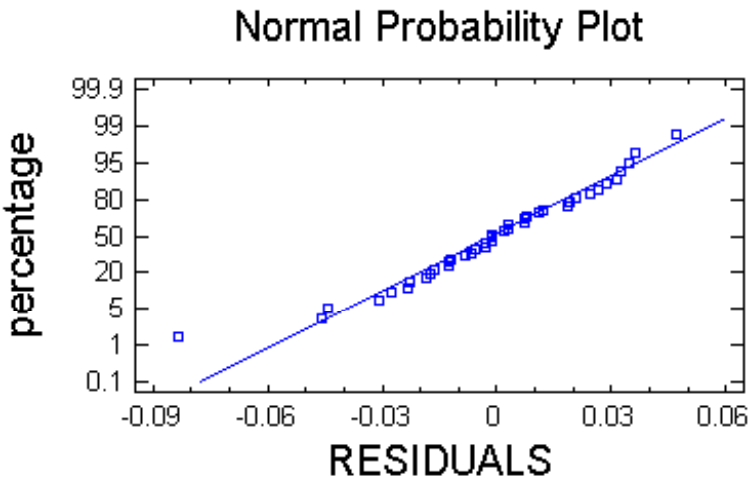
We veranderen beide Stopping criteria in de waarde 1.e-008 en klikken op OK. De niet-lineaire regressie berekeningen worden nu herhaald. In het resulterende analysevenster zien we dat de parameterwaarden nu een stukje nauwkeuriger berekend zijn. Hieruit kunnen we aflezen dat de volgende parameterwaarden in de Fritz en Schluender vergelijking onze meetgegevens beschrijft:
a = 2.199
b = 0.2202
X1 = 0.7785
X2 = 14.054
X3 = 0.6918

Hiermee zijn we nog niet klaar want overeenkomstig lineaire regressie hebben we de checklist ten aanzien van residuenplots en bijzondere meetpunten na te lopen. Aan de hand van het asymptotisch 95% betrouwbaarheidsinterval voor de parameters hebben we al gezien dat alle parameters significant zijn en thuishoren in de vergelijking. Vervolgens kijken we naar de residuenplot, die StatGraphics op identieke wijze als bij de lineaire regressie kan maken. Klik op het Graphical Options icoon en maak de keuze Residual Plots:



We zien een uitstekende, structuurloze residuenplot met 1 potentiële uitschieter bij een belading van 3 mMol/g. Na het toevoegen van de SRESIDUALS kolom aan de datasheet zien we dat deze potentiële uitschieter de meting bij C1=4,78, C2=0 en q1=3,02. We weten verder niets van deze meting, dus er is geen reden om deze meting te verwijderen en de regressie te herhalen.

De volgende stap is de controle op de normaliteit van de residuen. De residuen kunnen naar de datasheet weggeschreven worden en via de menukeuze Snap-Stats!! en vervolgens **One Sample Analysis** geanalyseerd worden.

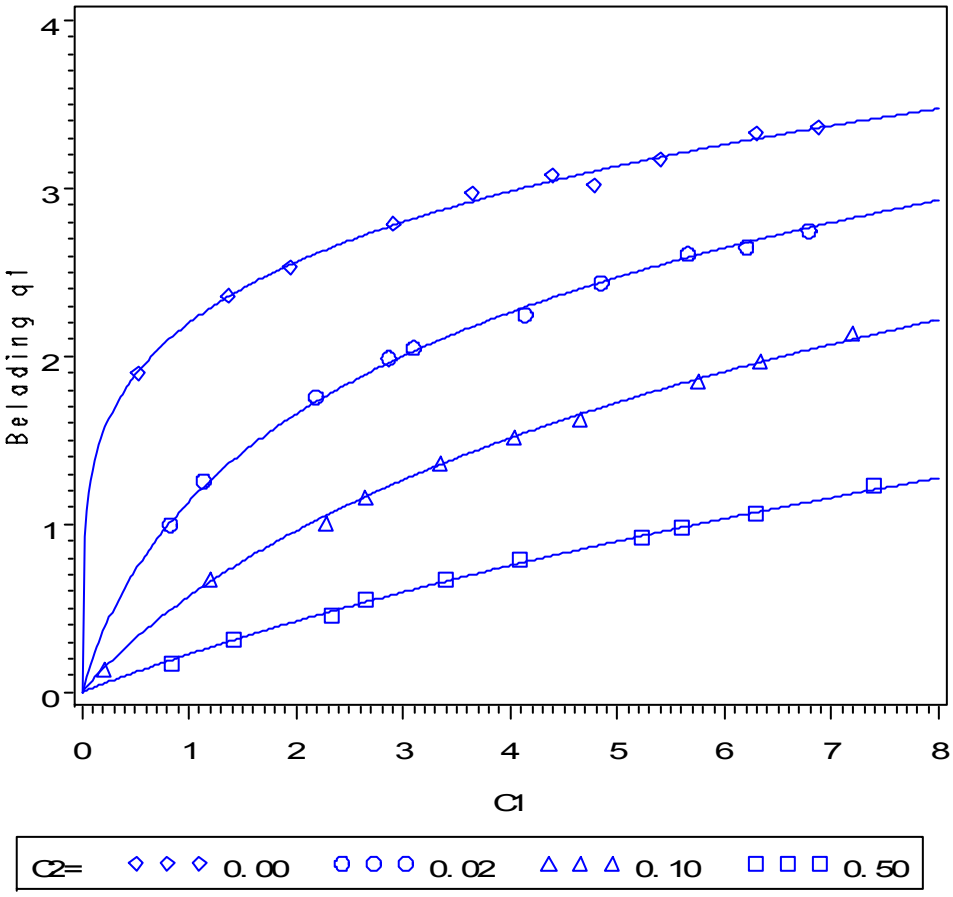


De resulterende normaliteitsplot zier er uitstekend uit met slechts 1 punt, dat er duidelijk uitspringt. Dit is natuurlijk de uitschieter, die we al eerder vastgesteld hadden. De Shapiro-Wilks P-waarde voor de toets op de normaliteit heeft een waarde van 0,3332, geen reden dus om H0: Residuen zijn normaal verdeeld te verwerpen. Verder dienen we ook de density trace te bekijken.

Als laatste laten we StatGraphics onderzoeken of er invloedrijke punten zijn. Klik in het analyse venster van de niet-lineaire regressie op het 2e icoon van links 'Tabular options' en klik de keuze 'Influential points' aan. StatGraphics vindt 2 invloedrijke punten, namelijk het 1e meetpunt met C1=0,53, C2=0 en

q1=1,9 en het 7e meetpunt met C1=4,78, C2=0 en q1=3,02. Uit het feit dat het eerste meetpunt van de single solute isotherm invloedrijk is op de parameterwaarden kunnen we concluderen, dat het verstandig is om extra single solute metingen uit te voeren bij met name nog wat lagere concentratie dan de C1=0,53. Deze extra metingen leggen de parameters a en b van de Freundlich isotherm dan beter vast. Het 7e meetpunt is precies het meetpunt dat we ook al als uitschieter aangemerkt hadden. Dat is nu niet aardig, een uitschieter die nog grote invloed op de parameterwaarden heeft ook. In het laboratorium zouden we dit punt nauwkeurig onderzoeken op vergissingen. Als die niet te vinden zijn, zouden we serieus moeten overwegen om dit punt opnieuw te meten, indien mogelijk natuurlijk.

Als afsluiting van deze niet-lineaire regressie de grafiek met de meetgegevens en de gefitte Fritz en Schluender vergelijking:



Hiermee besluiten we het voorbeeld van de vergelijking van Fritz en Schluender. Dit voorbeeld laat zien hoe uitgebreid een niet-lineaire regressie-analyse kan zijn. Het is nuttig om alle stappen na te gaan en de gebruikte technieken goed te onthouden. Deze technieken komen namelijk ook bij eenvoudiger voorbeelden goed van pas.