4 Meervoudige lineaire regressie

In het vorige hoofdstuk is enkelvoudige lineaire regressie besproken. Hierbij was er slechts één onafhankelijke variabele. In de praktijk zijn er echter gevallen waarin één onafhankelijke variabele niet voldoende is. We hebben dit bijvoorbeeld gezien bij de behandeling van lack-of-fit toetsen. In dit hoofdstuk gaan we daarom de situatie bekijken waarbij meerdere onafhankelijke variabelen toegelaten worden. Men spreekt hier van meervoudige lineaire regressie. Bij meervoudige lineaire regressie heeft de vergelijking, die aangepast wordt aan de set meetgegevens, de volgende algemene vorm:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ ..... \ + \beta_n x_n .$$

De instelvariabele x_i mag ook een willekeurige functie van x_i zijn, zolang parameters β_i er maar niet in voorkomen, bijv. $\sqrt{x_i}$, $x_i^{\,2}$, $\ln(x_i)$, enzovoort. Zo als in het hoofdstuk over enkelvoudige lineaire regressie is uitgelegd, gaat het dus om het lineair voorkomen van de parameters β. We zullen zien dat de basisbegrippen en procedures uit het hoofdstuk over enkelvoudige lineaire regressie ook bij meervoudige lineaire regressie van toepassing zijn. In die zin is dit hoofdstuk gedeeltelijk een herhaling. Het belangrijkste verschil met enkelvoudige lineaire regressie is dat er door het aanwezig zijn van meerdere onafhankelijke variabelen verschijnselen optreden die bij enkelvoudige lineaire regressie niet voorkomen. Dit geldt met name voor het begrip multicollineariteit.

Kernbegrippen van dit hoofdstuk:

- meervoudige lineaire regressie
- polynoomregressie
- invloedrijke punten
- normaliteit
- betrouwbaarheidsintervallen voor parameters
- toetsen van significantie regressiemodel
- determinatiecoëfficiënt
- residuenplots
- interactie
- multicollineariteit
- prestatiekentallen van regressiemodellen
  - MSE
  - C_p
  - R^2_{adj}

We gaan als eerste voorbeeld meervoudige lineaire regressie toepassen om de soortelijke warmte Cp van waterdamp bij constante druk van waterdamp tussen 280 en 400 K te beschrijven. Deze Cp moet niet verward worden met de statistische grootheid C_p die verder in dit hoofdstuk aan de orde komt. De meetgegevens zijn gehaald uit Perry (R.H. Perry et al., "Perry's Chemical Engineers' Handbook", Mc Graw Hill, 6e editie, blz. 3-239).

Temperatuur (K)	Soortelijke warmte (J/kg.ºC)
280	1858
285	1861
290	1864
295	1868
300	1872
305	1877
310	1882
315	1888
320	1895
325	1903
330	1911
335	1920
340	1930
345	1941
350	1954
355	1968
360	1983
365	1999
370	2017
375	2036
380	2057
385	2080
390	2104
400	2158

Tabel 4.1: Soortelijke warmte van waterdamp.

In thermodynamica boeken (bijv. K. Denbigh, "The principles of chemical equilibrium") wordt aangegeven, dat de soortelijke warmte Cp gewoonlijk beschreven kan worden met een tweedegraads polynoom in de temperatuur:

$$C_p = \beta_0 + \beta_1 T + \beta_2 T^2$$

waarin:

Cp : Soortelijke warmte (J/kg ºC)

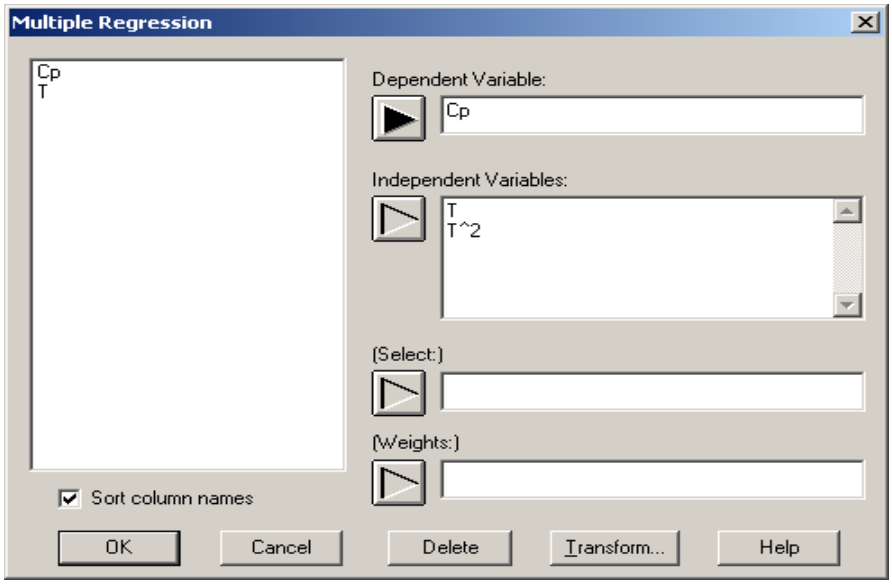
T : Absolute temperatuur (K)

β_0, β_1, β_2 : Parameters.

Volgens Denbigh is het gebruik van een tweedegraads polynoom gewoonlijk voldoende nauwkeurig. De vraag is nu de parameters β_i, d.w.z. de coëfficiënten van het tweedegraads polynoom te bepalen, die de soortelijke warmte van waterdamp als functie van de temperatuur beschrijft.

We beginnen met het inlezen van de meetgegevens in een StatGraphics datasheet, waarbij de kolom met de temperatuur meetwaarden de naam T krijgt en de soortelijke waarden kolom de naam Cp. Bestudering van het voorgestelde verband tussen temperatuur en soortelijke warmte laat zien, dat deze lineair is in de parameters en dat er tevens meerdere verklarende variabelen in voorkomen (T, T^2). Deze vergelijking valt dus in de categorie "Meervoudige lineaire regressie".

Vanwege het feit dat het hier om een polynoom in één variabele gaat, spreken we hier van **polynoomregressie**. Om in dat geval de optimale waarden voor de parameters te vinden, gebruiken we in StatGraphics de menukeuze **Relate** en vervolgens **Multiple Regression**. Gezien het feit dat we hier met polynoomregressie te maken hebben, hadden we ook de menukeuze **Relate**, **Polynomial Regression** kunnen doen. Het enige verschil is dat de invoer iets makkelijker maakt. Het voordeel van **Multiple Regression** is dat het voor alle meervoudige lineaire regressiemodellen werkt. Zoals we reeds eerder zagen bij het uitvoeren van enkelvoudige lineaire regressie in StatGraphics verschijnt er een venster waarin de afhankelijke en onafhankelijke variabele(n) opgegeven moeten worden. Het verschil is dat we bij meervoudige lineaire regressie meerdere onafhankelijke variabelen hebben. Voor ons soortelijke-warmteprobleem moet het venster als volgt ingevuld worden:



Na een klik op de OK button, voert StatGraphics de regressieberekeningen uit met als resultaat:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Cp
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                3590,36        76,3041        47,0533         0,0000
T                      -12,1386       0,454369       -26,7153         0,0000
T^2                   0,0213415    0,000670762        31,8169         0,0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                  169252,0      2      84626,2    6227,13       0,0000
Residual                285,388     21      13,5899
-----------------------------------------------------------------------------
Total (Corr.)          169538,0     23

R-squared = 99,8317 percent
R-squared (adjusted for d.f.) = 99,8156 percent
Standard Error of Est. = 3,68645
Mean absolute error = 2,94042
Durbin-Watson statistic = 0,310971 (P=0,0000)
Lag 1 residual autocorrelation = 0,640511
```

Onder de titel " Multiple Regression Analysis " worden eerst de resultaten afgedrukt, die te maken hebben met de gevonden parameters in de opgegeven meervoudige lineaire vergelijking. De getalswaarden van de optimaal aangepaste parameters staan onder de kop " Estimate". Het bovenste getal (CONSTANT) is de gevonden waarde voor de asafsnede β_0. Het getallen daaronder zijn de gevonden parameterwaarden, die voor de gegeven onafhankelijke variabele staan. De meetgegevens voor de Cp van waterdamp in de tabel worden dus beschreven door:
$Cp = 3590.36 - 12.1386\ T + 0.021342\ T^2$,

waarbij T de absolute temperatuur is.

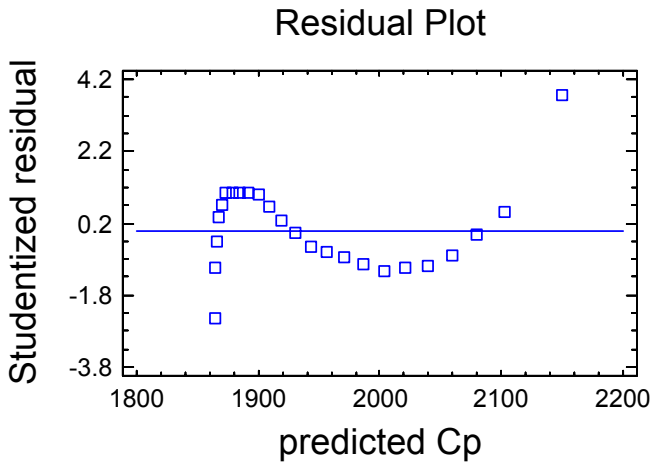
De betekenis van de berekende gegevens onder Analysis of Variance is identiek aan enkelvoudige lineaire regressie en daar besproken. We zien dat maar liefst 99.8% van de variantie in Cp verklaard wordt door deze kwadratische vergelijking. Ook zien we dat het model als geheel significant is, d.w.z. er is een significant verband tussen soortelijke warmte en de onafhankelijke variabelen als groep leveren. Bij meervoudige lineaire regressie is het erg belangrijk om ook de afzonderlijke onafhankelijke variabelen te toetsen op hun significantie, dat wil zeggen of ze voldoende van de waarde 0 verschillen. We doen dit door voor iedere parameter de volgende hypotheses te toetsen:
H0: $\beta_i = 0$, i^e parameter is NIET significant.
H1: $\beta_i \neq 0$, i^e parameter is significant.
We hopen dat we de nul hypothese H0 mogen verwerpen. Wordt H0 **NIET** verworpen, dan hebben we de situatie dat β_i gelijkgesteld moet worden aan 0. Daardoor valt de bijbehorende onafhankelijke variabele uit de lineaire functie weg. We moeten een nieuwe vergelijking formuleren, waarin deze onafhankelijke variabele niet meer voorkomt en de regressieberekeningen herhalen. Vaak zijn de onafhankelijke variabelen fysische grootheden zoals temperatuur, druk, luchtsnelheid, enz. Als een dergelijke fysische grootheid niet significant is en wegvalt uit de vergelijking, heeft dat ook consequenties voor het beeld dat we hebben van het proces dat we met de vergelijking proberen te beschrijven. We dachten bijvoorbeeld dat de temperatuur een rol speelt. Als deze dan wegvalt krijgen we een heel ander beeld van ons proces. Dit werkt sterk inzicht verhogend en is de belangrijkste reden om deze regressieberekeningen zorgvuldig te overdenken.
Voor het toetsen van significantie van regressieparameters wordt de toetsingsgrootheid

$$T = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

gebruikt. StatGraphics berekent en rapporteert de standaardafwijking van alle parameters en rekent op grond hiervan voor iedere parameter de P-value uit. P-value is een maat voor de waarschijnlijkheid van H0. Als deze mate van waarschijnlijkheid van H0 kleiner is dan een gekozen grenswaarde α=0,05, kunnen we de nulhypothese H0 verwerpen. We zien dat voor alle 3 de parameters P-values 0,0000 berekend worden. Deze waarden zijn veel kleiner dan de grenswaarde 0,05 waarmee we de nulhypothese kunnen verwerpen en kunnen concluderen dat alle 3 de parameters significant zijn.

Evenals bij enkelvoudige lineaire regressie controleren we de modelaannamen via o.a. residualplots en onderzoeken we of er uitschieters en/of invloedrijke pun-

ten aanwezig zijn. De residualplot als functie van de voorspelde waarden ziet er als volgt uit:



We zien dat de residuen nog een duidelijk patroon vertonen. Op grond hiervan kunnen we concluderen dat een kwadratisch verband nog niet voldoet en verbetert kan worden. De meest voor de hand liggende verbetering is het toevoegen van een polynoom term T^3. Controleer zelf dat de residualplot dan een structuurloos, rondom patroon vertoont. Verder blijkt de asafsnede β_0 niet significant te zijn. Dit is niet zo'n probleem, het geeft alleen maar aan dat de gevonden functie vrijwel door de oorsprong gaat. Controleer ook dat er 2 potentiële uitschieters en 3 invloedrijke meetpunten zijn. Tenslotte merken we op dat normaliteit in orde is in het derde-ordemodel. Aangezien niet duidelijk wat de tijdsvolgorde van de metingen is, heeft het geen zin om onafhankelijkheid te onderzoeken.

We bekijken nu een andere dataset, waarin de opbrengst (%) van een chemische reactie als functie van de temperatuur T (ºC) en druk (bar) gegeven wordt. Gevraagd wordt een vergelijking te bepalen, die de opbrengst als functie van temperatuur en druk beschrijft. Vervolgens kan deze vergelijking dan gebruikt worden om af te schatten bij welke temperatuur en druk de maximale opbrengst verkregen wordt.

T (ºC)	P (bar)	Yield (%)
50	1	68
50	1	65
50	5.5	81
50	5.5	82
50	10	57
50	10	65
125	1	52
125	1	53
125	5.5	82
125	5.5	85
125	5.5	80
125	5.5	83
125	10	76
125	10	76

200	1	28
200	1	27
200	5.5	70
200	5.5	68
200	10	80
200	10	84

Tabel 4.2: Opbrengst van een chemische reactie

Om de opbrengst Yield als functie van T en P te beschrijven, beginnen we maar eens met het meest voor de hand liggende:

$$\text{Yield} = \beta_0 + \beta_1 T + \beta_2 P$$

Bij dit soort vergelijkingen is het gebruikelijk om ook de term β_3 T*P mee te nemen:

$$\text{Yield} = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 T * P$$

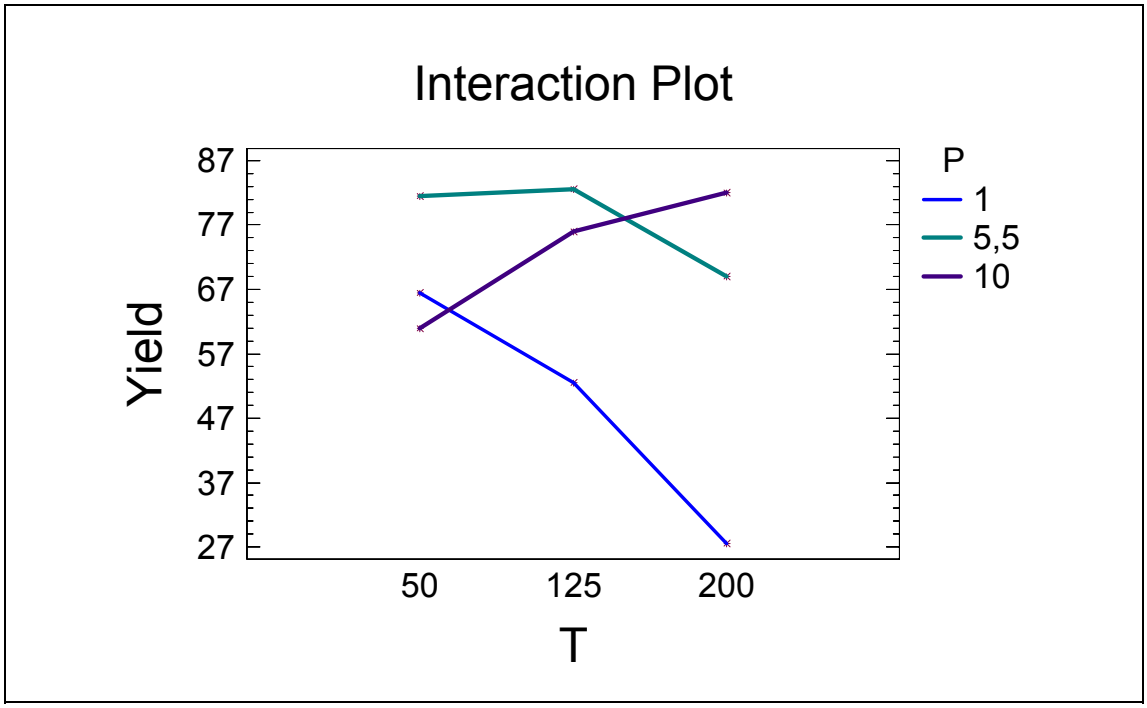
De term T*P wordt de **interactie** tussen T en P genoemd. Om te begrijpen wat de interactie tussen T en P betekent, kijken we naar de helling van Yield als functie van de temperatuur T. Zonder interactie is deze gelijk aan:

$$\frac{\partial \text{Yield}}{\partial \text{T}} = \beta_1$$

en dus constant. Met de interactie term T*P is deze gelijk aan:

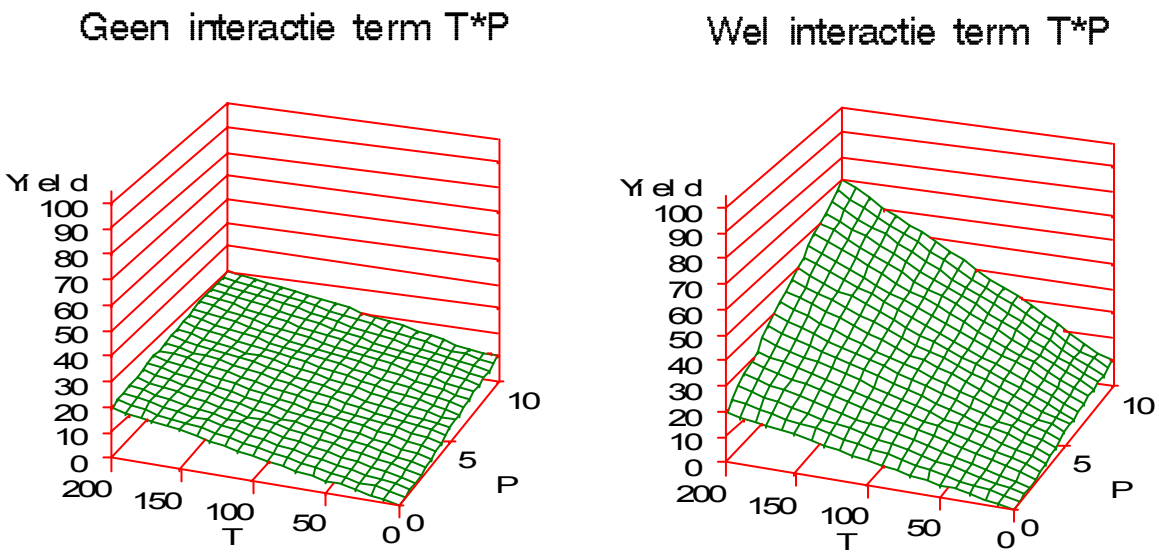
$$\frac{\partial \text{Yield}}{\partial \text{T}} = \beta_1 + \beta_3 \text{P}$$

en niet langer constant maar afhankelijk van P. De interactie term T*P zorgt er dus voor dat de helling van Yield versus T kan veranderen, als P verandert. Ditzelfde geldt ook voor de helling van de Yield versus de druk P, die door de interactie term T*P kan veranderen als T verandert.



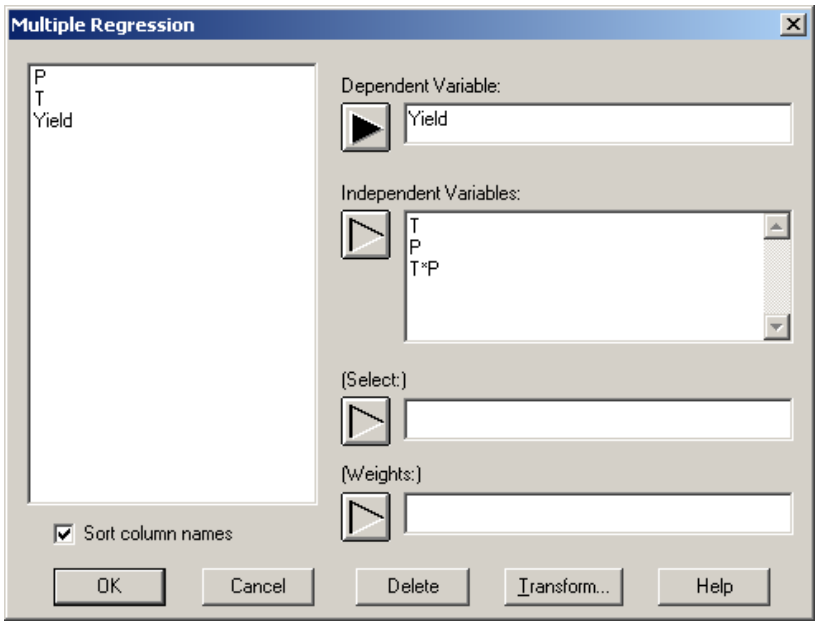
Tabel 4.3

Driedimensionaal ziet het verschil tussen wel en geen interactie term T*P er als volgt uit:



Zonder interactie term hebben we een plat vlak. Door het toevoegen van de interactieterm kan een gebogen oppervlak ontstaan, dat vaak beter in staat is de meetgegevens te beschrijven.

De volgende stap is het uitvoeren van de meervoudige lineaire regressie in StatGraphics. Eerst voeren we de meetgegevens in een StatGraphics datasheet en noemen de kolommen respectievelijk T, P en Yield. Daarna kiezen we voor de menukeuze **Relate** en vervolgens **Multiple Regression**. De afhankelijke – en onafhankelijke variabelen definiëren we als volgt:



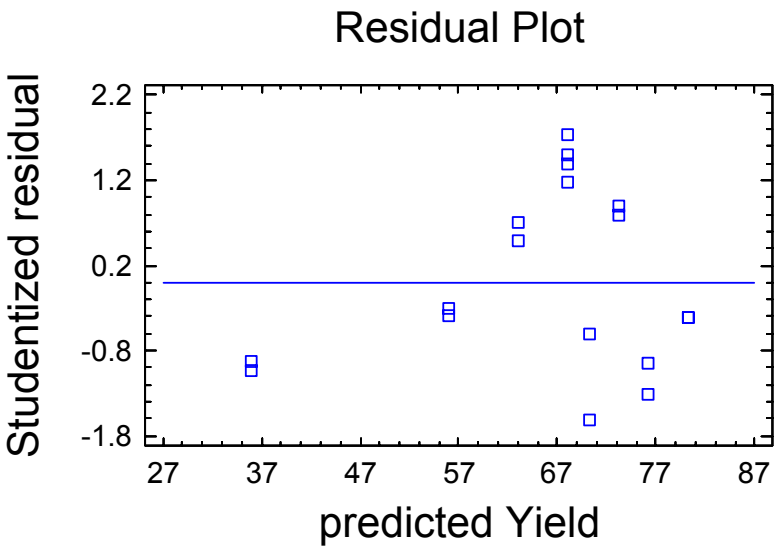
Na een klik op de OK button worden de regressieberekeningen uitgevoerd met het volgende resultaat:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Yield
-----------------------------------------------------------------------------
                                        Standard          T
Parameter               Estimate          Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 92.3593        10.1698        9.08172         0.0000
T                       -0.312222      0.0732812       -4.26061         0.0006
P                        -2.87037        1.54214       -1.86129         0.0812
T*P                     0.0444444      0.0110791        4.01157         0.0010
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    3862.17      3      1287.39      11.51       0.0003
Residual                 1789.63     16      111.852
-----------------------------------------------------------------------------
Total (Corr.)             5651.8     19

R-squared = 68.3352 percent
R-squared (adjusted for d.f.) = 62.398 percent
Standard Error of Est. = 10.576
Mean absolute error = 8.62
```

We zien dat van onze vergelijking

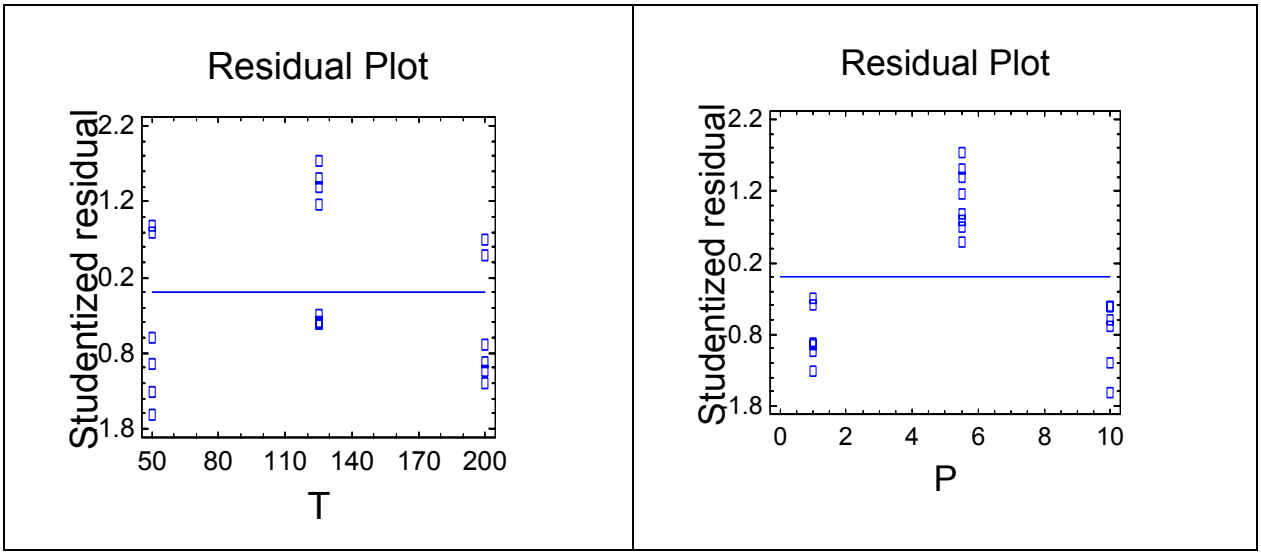
$$\text{Yield} = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 T * P$$

de asafsnede β_0 en de parameters β_1 en β_3 significant zijn. De parameter β_2 is dat niet maar een P-value van 0,08 ligt zo dicht bij de grenswaarde 0,05 dat we kunnen concluderen dat de druk P wel enige invloed heeft. Laten we eens naar de residualplot kijken en zien wat die ons leert:



We zien dat de residualplot niet het gewenste random patroon van de residuen vertoont. Er is een duidelijk verband, eerst beneden nul, dan boven nul en tenslotte weer beneden nul. We missen het willekeurige patroon rondom de horizontale lijn $e_i=0$. Het is aan te raden om bij meervoudige regressie ook de kijken naar

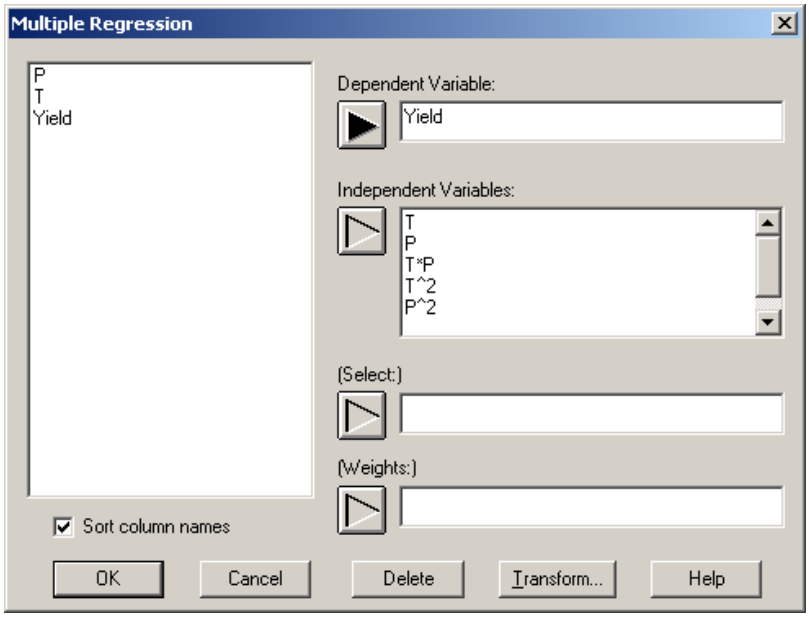
de plot van de residuen tegen de onafhankelijke instelvariabelen. In dit geval worden dat plots tegen temperatuur T en druk P. Deze plots zien er als volgt uit:



Beide residualplots vertonen niet het gewenste random patroon en uit beide residualplots is duidelijk af te leiden, dat er in het model een kwadratische term voor zowel T als P ontbreekt. We breiden de vergelijking dus uit met T^2 en P^2. De totale regressiemodel ziet er dan als volgt uit:

$$\text{Yield} = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 T * P + \beta_4 T^2 + \beta_5 P^2$$

Het fitten van deze vergelijking met StatGraphics gaat door de variabelen als volgt in te stellen.



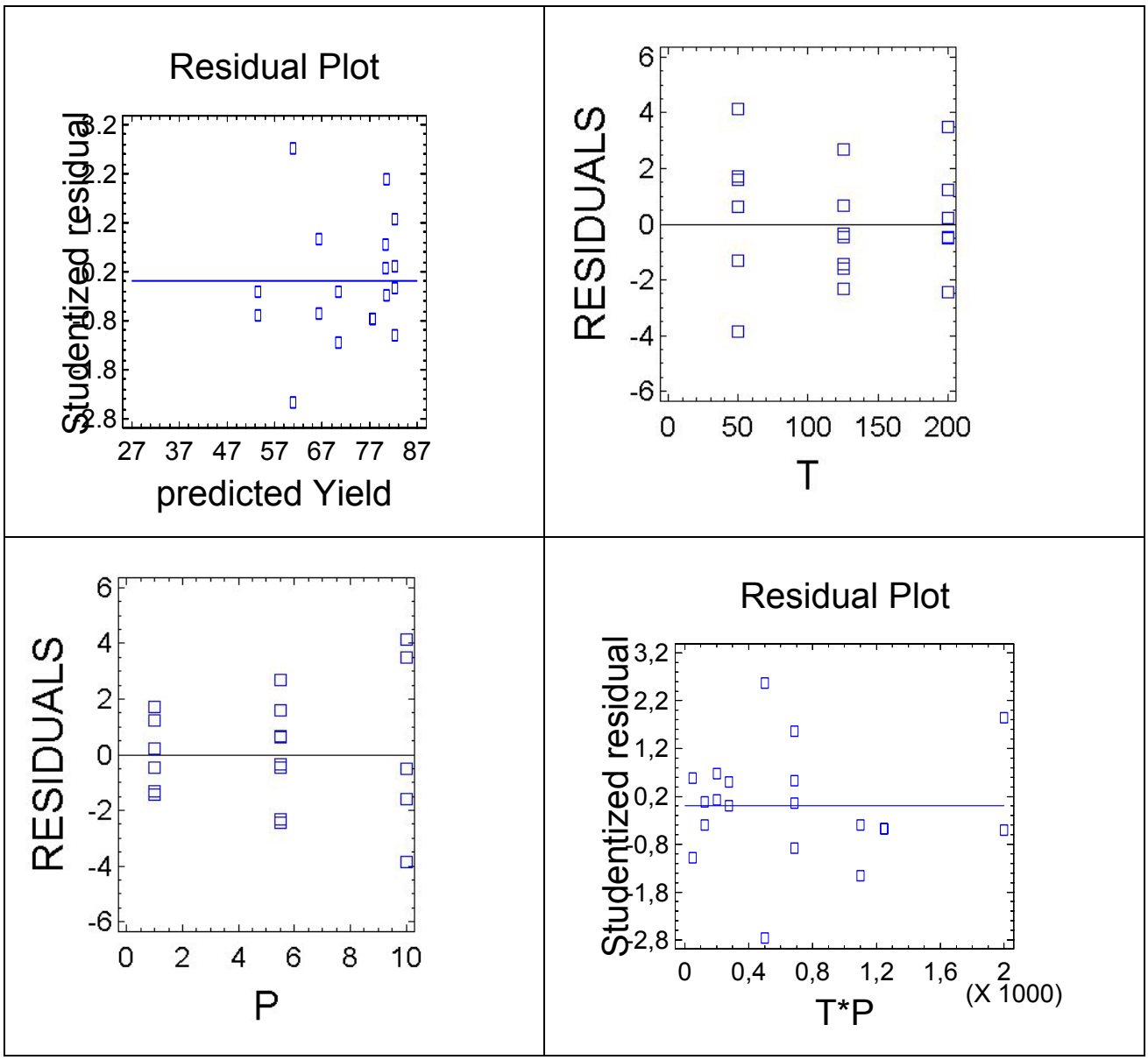
We zien een enorme verbetering van de beschrijving van de meetgegevens door het toevoegen van T^2 en P^2.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Yield
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 64,2072       3,26458       19,6678         0,0000
T                      -0,0471429     0,0504541      -0,934371        0,3659
P                         6,3448      0,675556       9,39196         0,0000
T*P                     0,0444444    0,00243463       18,2551         0,0000
T^2                   -0,00106032   0,000191261      -5,54383         0,0001
P^2                    -0,837743      0,053128      -15,7684         0,0000
-----------------------------------------------------------------------------

                          Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    5576,18      5      1115,24     206,47       0,0000
Residual                  75,619     14      5,40136
-----------------------------------------------------------------------------
Total (Corr.)            5651,8     19

R-squared = 98,662 percent
R-squared (adjusted for d.f.) = 98,1842 percent
Standard Error of Est. = 2,32408
Mean absolute error = 1,51905
Durbin-Watson statistic = 2,71456 (P=0,0021)
Lag 1 residual autocorrelation = -0,427245
```

Uit de meervoudige regressie-analyse van StatGraphics volgt, dat de onafhankelijke variabele T in dit model niet meer significant is. Deze term verwijderen we uit onze regressiemodel. Hierna herhalen we de meervoudige regressieberekening. Dit leidt tot het volgende resultaat:

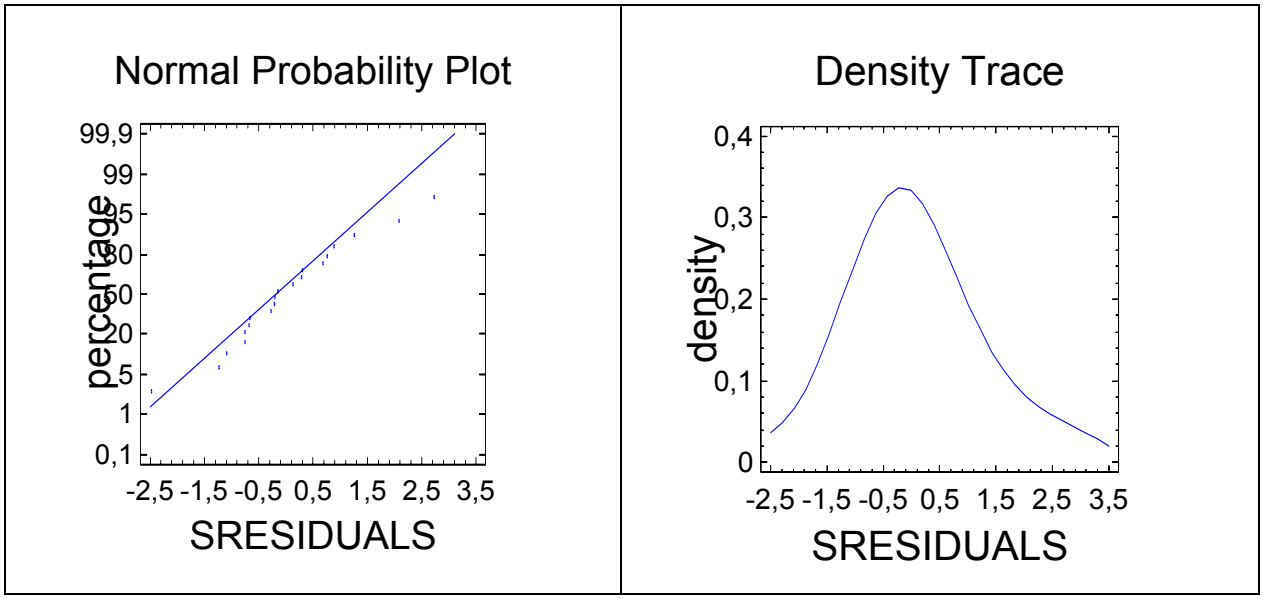
```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Yield
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 61.6742       1.81109       34.0537         0.0000
P                        6.33402      0.672593       9.41731         0.0000
T*P                    0.0438407    0.00233737       18.7565         0.0000
T^2                  -0.00122968  0.0000607863      -20.2295         0.0000
P^2                    -0.829902     0.0522386      -15.8868         0.0000
-----------------------------------------------------------------------------

                          Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    5571.47      4      1392.87     260.07       0.0000
Residual                 80.3347     15      5.35565
-----------------------------------------------------------------------------
Total (Corr.)             5651.8     19

R-squared = 98.5786 percent
R-squared (adjusted for d.f.) = 98.1996 percent
Standard Error of Est. = 2.31423
Mean absolute error = 1.63523
```

Alle parameters zijn nu significant. De belangrijkste residualplots zien er als volgt uit:



Dit is een uitstekend willekeurig patroon. Tevens zien we meteen dat er geen uitschieters zijn, behalve bij een voorspelde opbrengst van 61%.

De normal probability plots en de density trace van de residuen zien er uitstekend uit. De Shapiro-Wilks P-value voor de toets op de normaliteit heeft een waarde van 0.9183, geen reden dus om H0: Residuen zijn normaal verdeeld te verwerpen.



Onderzoek naar invloedrijke punten levert 3 invloedrijke meetpunten op, het 5e, 6e en 20e meetpunt. De gefitte parameters hebben voor onze doelstelling voldoende kleine standaard deviaties, dus er is geen reden om hier verder actie te ondernemen.

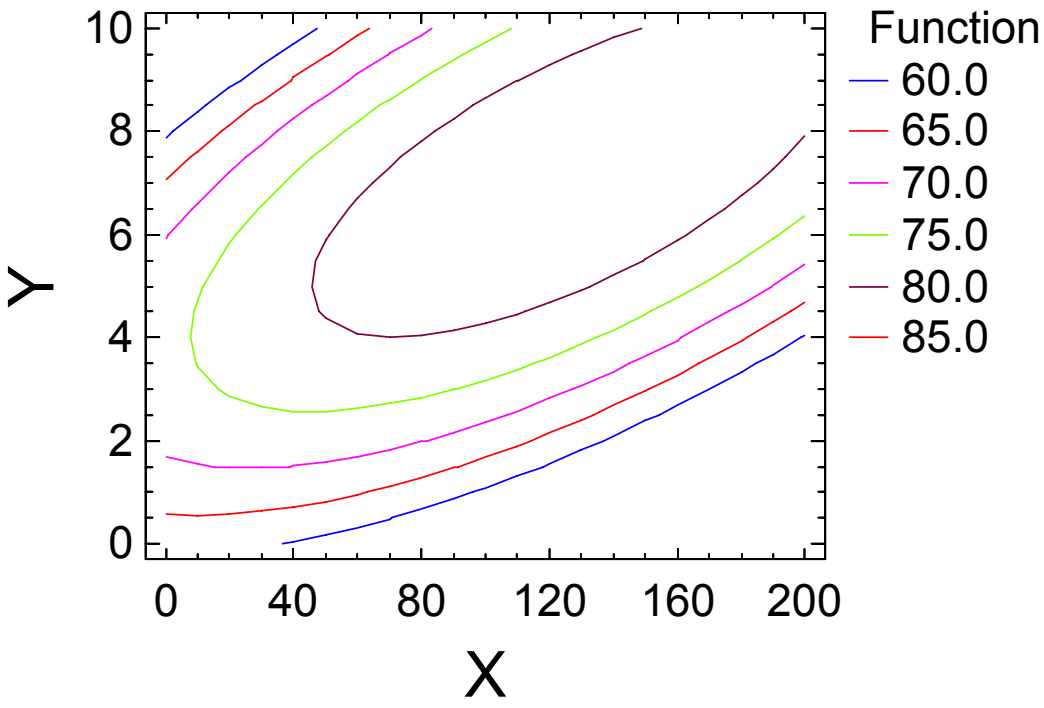
We kunnen nu dus de gevonden regressiemodel aanvaarden. De vergelijking waarmee de meetgegevens van Yield (%) als functie van T (ºC) en P (bar) het beste beschreven worden luidt nu:

$$\text{Yield} = \beta_0 + \beta_1 P + \beta_2 T * P + \beta_3 T^2 + \beta_4 P^2$$

waarin:

β_0	=	61,6742
β_1	=	6,33402
β_2	=	0,0438407
β_3	=	-0,00122968
β_4	=	-0,829902

In StatGraphics kan via de menu keuze **Plot** en **Response Surfaces** een hoogtelijnen kaart gemaakt van dit soort regressiemodellen (een andere mogelijkheid staat beschreven in paragraaf 3.6 van het StatGraphics dictaat). Voor deze vergelijking ziet die hoogtelijnenkaart of contourplot er als volgt uit:



In deze grafiek is X de temperatuur T, Y de druk P en Function de waarde voor de opbrengst Yield. Uit deze grafiek kunnen we aflezen dat het maximum in de buurt van temperatuur T=130 ºC en druk P=7 bar ligt. We kunnen dit maximum ook analytisch berekenen:

$$\frac{\partial \text{Yield}}{\partial T} = \beta_2 P + 2\beta_3 T = 0$$

$$\frac{\partial \text{Yield}}{\partial P} = \beta_1 + \beta_2 T + 2\beta_4 P = 0$$

Oplossing van dit stelsel van twee vergelijkingen met twee onbekenden geeft:
T_{max} = 128.6 ºC
P_{max} = 7.2 bar

StatGraphics is ons behulpzaam bij het berekenen van Yield bij T=128.6 en P=7.2 met de gefitte regressiemodel. Daarvoor moeten we de gewenste waarden voor T en P onderaan de datasheet met meetgegevens toevoegen en daarbij geen waarden invullen voor de kolom Yield:

Dictaat T P Yield meetset.sf3

	T	P	Yield	PREDICTED	RESIDUALS	SRESIDUALS
9	125	5.5	82	82.3335	-0.333472	-0.149407
10	125	5.5	85	82.3335	2.66653	1.25957
11	125	5.5	80	82.3335	-2.33347	-1.08791
12	125	5.5	83	82.3335	0.666528	0.299344
13	125	10	76	77.6113	-1.6113	-0.75616
14	125	10	76	77.6113	-1.6113	-0.75616
15	200	1	28	26.7593	1.24074	0.678924
16	200	1	27	26.7593	0.240742	0.129693
17	200	5.5	70	70.4443	-0.444338	-0.212228
18	200	5.5	68	70.4443	-2.44434	-1.22665
19	200	10	80	80.5184	-0.518399	-0.271892
20	200	10	84	80.5184	3.4816	2.08489
21	128.6	7.2				
22						
23						
24						
25						
26						
27						
28						

Vervolgens gaan we in StatGraphics naar het Multiple Regression Analysis venster en klikken op het tweede gele icoon van links “Tabular Options” en vinken de keuze Reports aan:

Regression Results for Yield

Row	Fitted Value	Stnd. Error for Forecast	Lower 95.0% CL for Forecast	Upper 95.0% CL for Forecast
21	84.5136	2.4396	79.3137	89.7135

Er wordt een maximale opbrengst van 84,5 % met de gefitte vergelijking uitgerekend. Verder berekent StatGraphics het 95% voorspellingsinterval. Dit 95% voorspellingsinterval geeft aan, dat als we een nieuwe meting gaan uitvoeren bij T=128,6 ºC en P=7,2 bar, de gevonden opbrengst Yield met 95% zekerheid zal liggen tussen 79,3 en 89,7.

In het bovenstaande voorbeeld was het doel van de regressie-analyse het doen van voorspellingen. Het is dan in feite niet belangrijk welke variabelen het regressiemodel bevat, zolang er maar nauwkeurige voorspellingen gedaan kunnen worden. Een niet-significante parameter is dan eigenlijk geen probleem. Indien men echter een regressie-analyse uitvoert om een verklarend model te maken, dan is het storend als het model onafhankelijke variabelen bevat die geen significante bijdrage leveren tot het te onderzoeken verschijnsel. Omgekeerd is het ook storend als onafhankelijke variabelen wel een bijdrage leveren aan de verklaring van het te onderzoeken verschijnsel, maar dat ze in de regressie-analyse ten onrechte als niet significant gekenmerkt worden. Dit laatste kan gebeuren als de onafhankelijke variabelen bijna aan een lineair verband voldoen. Men spreekt dan van **multicollineariteit**. De kleinste-kwadratenmethode die ten grondslag ligt aan de regressie-analyses kan in zulke gevallen verkeerde antwoorden ople-

veren of parameterschattingen met grote varianties. Multicollineariteit is niet eenduidig vast te stellen. Kenmerken die duiden op multicollineariteit zijn:
1. verkeerde tekens van parameters in het regressiemodel
2. veel niet-significante parameters, terwijl het regressiemodel als geheel wel significant is.
3. hoge waarden (dichtbij 1 of –1) in de correlatiematrix (in StatGraphics te vinden onder Tabular Options.

Terugkerend naar het voorbeeld van de opbrengstmetingen, zien we dat $T*P$ en T^2 mogelijk gecorreleerd zijn. We zullen aan het eind van het hoofdstuk zien dat weglaten van één van deze variabelen tot een duidelijke verslechtering van het model leidt.

Correlation matrix for coefficient estimates

	CONSTANT	P	T*P	T^2
CONSTANT	1,0000	-0,8113	0,5442	-0,6658
P	-0,8113	1,0000	-0,4720	0,4011
T*P	0,5442	-0,4720	1,0000	-0,8173
T^2	-0,6658	0,4011	-0,8173	1,0000
P^2	0,5564	-0,8735	0,0440	-0,0539

	P^2
CONSTANT	0,5564
P	-0,8735
T*P	0,0440
T^2	-0,0539
P^2	1,0000

Er zijn geavanceerde regressietechnieken zoals ridge regression die dit probleem ondervangen. Dit valt echter buiten de doelstelling van dit college. Een eenvoudige praktische aanpak is experimenteren met het weglaten van één of meerdere onafhankelijke variabelen.

Bij meervoudige lineaire regressie worden vaak uitgebreide regressiemodellen gefit met veel parameters β. In het uitgewerkte voorbeeld van Yield als functie van T en P werd een regressiemodel met 5 parameters gefit:

$$\text{Yield} = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 T * P + \beta_4 T^2 + \beta_5 P^2$$

Hier bleek de temperatuurterm T met parameter β_1 niet significant te zijn en moest de vergelijking aangepast en opnieuw gefit worden. Stel nu eens dat niet alleen temperatuur T en druk P maar ook de partieel spanning zuurstof O een rol had gespeeld. De volledige regressiemodel zou er dan als volgt hebben uitgezien:

$$\text{Yield} = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 O + \beta_4 T * P + \beta_5 T * O + \beta_6 P * O + \beta_7 T^2 + \beta_8 P^2 + \beta_9 O^2$$

Meestal blijken een aantal parameters hierin niet significant te zijn. Welke parameters dat zijn, is niet op voorhand te zeggen. Het is dus een heel werk om dit soort uitgebreide vergelijkingen terug te brengen tot die termen, die significant zijn.

We presenteren hier drie methoden om dit probleem aan te pakken:
1. alle modellen automatisch doorrekenen
2. voorwaartse regressie
3. achterwaartse regressie.

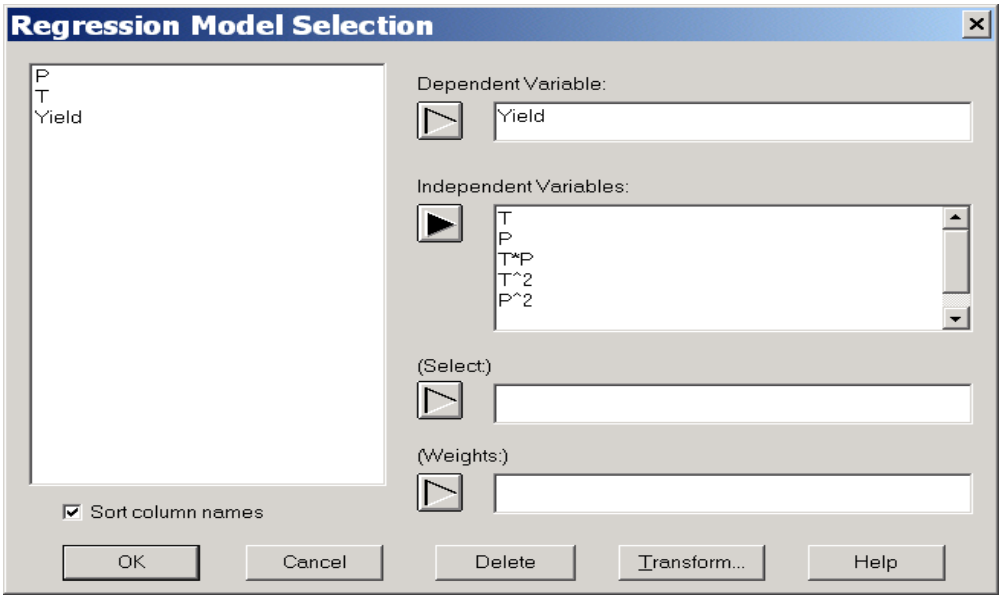
Alvorens we deze methoden bespreken, is het noodzakelijk na te denken wat een goed model inhoudt. Als we dit niet weten, kunnen we geen modellen met elkaar vergelijken. Het is niet praktisch modellen te vergelijken op alle mogelijke controles (residuenplots, normaliteittoetsen e.d.). In plaats daarvan is het nuttig om een vergelijking tussen modellen te baseren op een beperkt aantal getallen (1 tot 3) en daarna het beste model grondig te controleren op de gebruikelijke manier. De volgende getallen worden hierbij vaak gebruikt:

1. aangepaste determinatiecoëfficiënt R^2_{adj}
2. MSE (Mean Square Error)
3. Mallows C_p

De gewone determinatiecoëfficiënt R^2 geeft de hoeveelheid verklaarde variantie aan, d.w.z. hoeveel van de variantie in de metingen door de onafhankelijke variabelen verklaard wordt. Deze waarde moet natuurlijk hoog zijn (liefst boven de 90%). Er zijn echter twee kanttekeningen. Allereerst is een hoge R^2 geen garantie voor een goed model, zoals we in het eerste voorbeeld van dit hoofdstuk hebben gezien. Anderzijds neemt de R^2 altijd toe als we een onafhankelijke variabele toevoegen. Dit gaat in tegen het algemene natuurwetenschappelijke principe om zo weinig mogelijk verklarende variabelen te gebruiken. De **aangepaste determinatiecoëfficiënt** R^2_{adj} komt hieraan tegemoet door een correctiefactor te gebruiken voor het aantal onafhankelijke variabelen. De aangepaste determinatiecoëfficiënt R^2_{adj} moet zo hoog mogelijk zijn.

De **MSE** is een schatter voor de variantie van de meetfout en dus het kwadraat van de standaardafwijking van de meetfout, die in de regressie output als standard error staat aangegeven. Hoe kleiner de MSE, hoe preciezer het model past. De MSE neemt automatisch het aantal onafhankelijke variabelen met zich mee. Het is dus mogelijk dat door het toevoegen van extra onafhankelijke variabele de MSE stijgt. De MSE moet zo laag mogelijk zijn.
Een variant op de MSE is **Mallows C_p** (niet te verwarren met de soortelijke warmte Cp). Deze grootheid neemt niet alleen de variantie van de meetfout mee, maar houdt ook rekening met eventuele systematische afwijkingen tussen meetwaarden en modelvoorspellingen. Net als de MSE dient de C_p zo klein mogelijk te zijn.

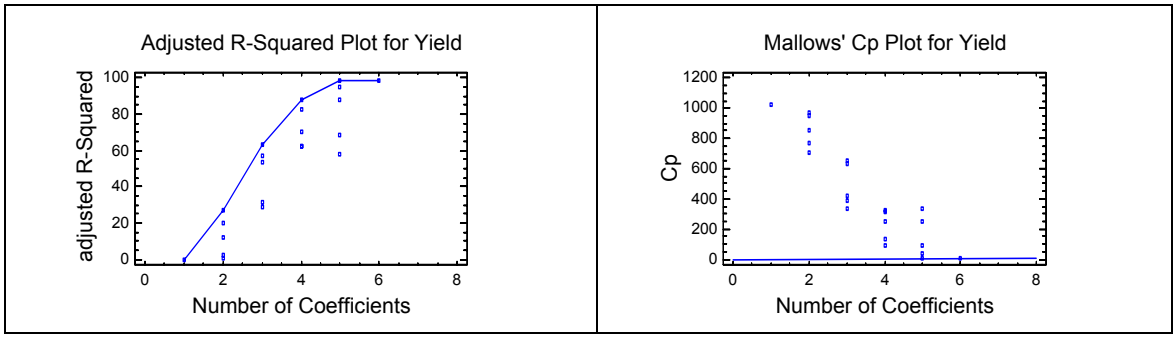
StatGraphics biedt via het menu **Special, Advanced Regression, Regression Model Selection** de mogelijkheid alle mogelijke modellen met gegeven onafhankelijke variabelen door te rekenen. Merk op dat we alle variabelen (ook de interacties) zelf moeten invoeren.



Bij Tabular Options kunnen we StatGraphics het model met de beste R^2_{adj} of beste C_p laten berekenen. De uitvoer van beide opties is een lijst van modellen waarbij alle drie de kentallen worden uitgerekend. Het verschil zit in het sorteren. Door met de rechtermuisknop in het bovenste venster (Analysis Summary) te klikken kunnen we het maximale en minimale aantal onafhankelijke variabelen instellen. Het maximale aantal zal meestal het aantal onafhankelijke variabelen zelf zijn; het minimum kan nuttig zijn om te verhinderen dat StatGraphics modellen doorrekent die vanwege het te kleine aantal variabelen op voorhand niet interessant zijn. Daarnaast kunnen we bij de grafische opties grafieken laten tekenen van de kentallen afzonderlijk als functie van het aantal variabelen.

```
Regression Model Selection

Dependent variable: Yield
Independent variables:
   A=T
   B=P
   C=T*P
   D=T^2
   E=P^2
```

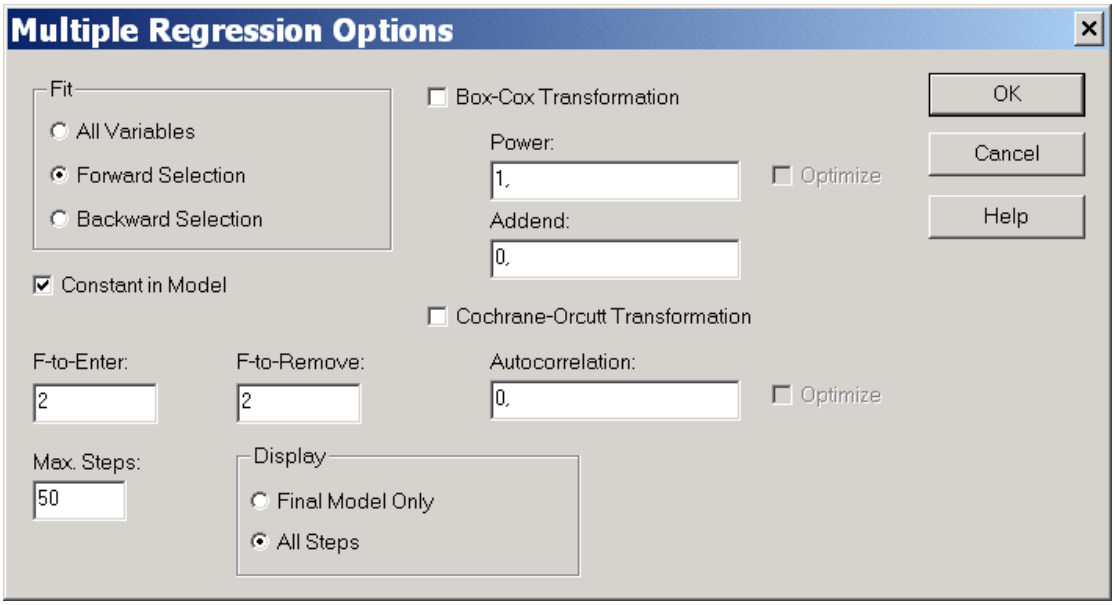
```
Models with Largest Adjusted R-Squared

Model Results
-------------------------------------------------------------------
                        Adjusted                  Included
MSE         R-Squared   R-Squared   Cp            Variables
-------------------------------------------------------------------
5,35565     98,5786     98,1996     4,87305       BCDE
5,40136     98,662      98,1842     6,0           ABCDE
16,1083     95,7248     94,5848     34,7341       ABCE
34,7065     90,1747     88,3325     90,8081       CDE
36,8045     90,232      87,6272     92,209        ACDE
50,9081     85,5881     82,8859     138,801       ACE
```



Hieruit kunnen we concluderen dat onze keuze van een volledig kwadratisch model in T en P inderdaad een goede keuze was. Bedenk echter dat het model dat als beste tevoorschijn komt uit deze procedure, altijd nog apart gecontroleerd moet worden op normaliteit, residuen etc. . Bij onze bespreking van multicollineariteit hadden we gezien dat T^*P en T^2 mogelijk gecorreleerd zijn. Als we echter één van deze variabelen weglaten, dan verslechtert het model duidelijk. Modellen zonder beide variabelen komen niet in de top van de lijst voor en hebben daarom bijvoorbeeld een veel grotere MSE. Het is dus niet aan te raden één van deze variabelen weg te laten.

Met een computer is bovenstaande procedure snel uit te voeren. Het kan echter voorkomen dat er een groot aantal variabelen is, waardoor het aantal regressiemodellen te groot wordt om allemaal door te rekenen. In dit soort gevallen kan men voorwaartse en achterwaartse regressie toepassen. Bij **voorwaartse regressie** begint met een regressiemodel zonder variabelen en voegt dan bij elke stap die onafhankelijke variabele toe die de sterkste verbetering van het model geeft. De procedure stopt als toevoegen van een variabele geen significante verbetering van het regressiemodel levert. Bij **achterwaartse regressie** begint men daarentegen met een volledig regressiemodel en laat men bij elke step die variabele weg die de minste invloed heeft. De procedure stopt als er geen variabele meer weggelaten kan worden zonder het model significant slechter te maken. Een belangrijk probleem met beide procedures is dat er geen rekening gehouden wordt afhankelijkheden tussen de onafhankelijke variabelen. Zo hebben we in dit hoofdstuk gezien dat toevoegen van een onafhankelijke variabele een al eerder in het model opgenomen variabele van significant in niet significant kan doen veranderen en omgekeerd. Dit probleem wordt gedeeltelijk ondervangen door de zogenaamde **stapsgewijze regressie**, die beide procedures combineert en bij elke stap controleert of alle onafhankelijke variabelen in het regressiemodel nog steeds significant zijn. Helaas biedt StatGraphics deze mogelijkheid niet. Merk op dat voorwaartse en achterwaartse regressie niet hetzelfde model hoeft op te leveren en dat er ook geen garantie is dat het gevonden model adequaat is. Het is daarom aan te raden zowel voorwaartse als achterwaartse regressie uit te voeren en de uitkomsten te vergelijken.
Om achterwaartse of voorwaartse regressie uit te voeren in StatGraphics beginnen we met het uitvoeren van een normale Multiple Regression. Bij het opgeven van de onafhankelijke variabelen definiëren we alle termen, die een rol kunnen spelen in de beschrijving van de afhankelijke variabele. Vervolgens klikken we op de OK button om dit volledige model te fitten. Hierbij zijn meestal 1 of meerdere onafhankelijke variabelen niet significant. In het Multiple Regression Analysis venster klikken we nu met de rechtermuisknop en kiezen in het menu dat verschijnt voor de keuze Analysis Options. In het venster dat verschijnt kiezen we voor Forward Selection. Vervolgens vullen we bij F-to-Enter en F-to-Remove de

waarden 2 in. Deze waarde geeft aan wanneer de procedure moet stoppen. Strikt genomen is het niet juist een vaste waarde te gebruiken. De juiste F-waarde hangt namelijk af van het aantal variabelen en dat varieert natuurlijk tijdens de voorwaartse of achterwaartse regressie. Het is daarom verstandig om te experimenteren met de waarden voor F-to-Enter en F-to-Remove en te kijken naar de regressiemodellen die als eindmodel tevoorschijn komen.



Vervolgens klikken we op de OK button en de voorwaartse regressie start. Door de optie All Steps aan te vinken laat StatGraphics de stappen zien die genomen worden:

```
Stepwise regression
-------------------
Method: forward selection
F-to-enter: 2,0
F-to-remove: 2,0

     Step 0:
     ---------
     0 variables in the model.  19 d.f. for error.
     R-squared =   0,00%     Adjusted R-squared =   0,00%     MSE = 297,463

     Step 1:
     ---------
     Adding variable P with F-to-enter = 8,08713
     1 variables in the model.  18 d.f. for error.
     R-squared =  31,00%     Adjusted R-squared =  27,17%     MSE = 216,651

     Step 2:
     ---------
     Adding variable P^2 with F-to-enter = 11,1902
     2 variables in the model.  17 d.f. for error.
     R-squared =  58,39%     Adjusted R-squared =  53,49%     MSE = 138,336

     Step 3:
     ---------
     Adding variable T^2 with F-to-enter = 3,15388
     3 variables in the model.  16 d.f. for error.
     R-squared =  65,24%     Adjusted R-squared =  58,72%     MSE = 122,78

     Step 4:
     ---------
     Adding variable T*P with F-to-enter = 351,805
     4 variables in the model.  15 d.f. for error.
     R-squared =  98,58%     Adjusted R-squared =  98,20%     MSE = 5,35565

     Final model selected.
```

In dit geval levert de voorwaartse regressie hetzelfde model op dat we al eerder hadden gevonden. Als we echter de standaardwaarde 4 van StatGraphics hadden aangehouden, dan hadden we een kwadratisch model in P gevonden, wat duidelijk geen goed model kan zijn.
Er is ook de mogelijkheid van Backward Selection. Dit levert het volgende resultaat op:

```
Stepwise regression
-------------------
Method: backward selection
F-to-enter: 2,0
F-to-remove: 2,0

     Step 0:
     ---------
     5 variables in the model.  14 d.f. for error.
     R-squared =  98,66%     Adjusted R-squared =  98,18%     MSE = 5,40136

     Step 1:
     ---------
     Removing variable T with F-to-remove = 0,87305
     4 variables in the model.  15 d.f. for error.
     R-squared =  98,58%     Adjusted R-squared =  98,20%     MSE = 5,35565

     Final model selected.
```

Ook hier vinden het volledige model met alleen T verwijderd. Als we de standaardwaarde 4 van StatGraphics hadden aangehouden voor F-to-Enter en F-to-Remove, dan hadden we hetzelfde model gevonden. Zoals boven opgemerkt, zou

voorwaartse regressie dan niet hetzelfde model als achterwaartse regressie hebben opgeleverd.

We besluiten dit hoofdstuk met een uitbreiding van het stappenplan dat we voor enkelvoudige lineaire regressie hebben gegeven.

Stappenplan voor meervoudige regressie

1. Voer de lineaire regressieberekeningen uit met StatGraphics.
2. Controleer of het model significant is. Mochten het model niet significant zijn, formuleer dan een nieuw model en start opnieuw met het uitvoeren van de regressie.
3. Controleer of alle parameters significant zijn. Indien er meerdere waarnemingen zijn met dezelfde waarde van de instelvariabele, voer dan een lack-of-fit toets uit.
4. Onderzoek of er sprake is van multicollineariteit door te kijken naar de correlatiematrix en door de tekens van de parameters te controleren.
5. 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. Voer desgewenst een modelselectie uit door alle mogelijke regressiemodellen met een gegeven verzameling onafhankelijke variabelen door te rekenen. Indien dit niet mogelijk is vanwege een te groot aantal onafhankelijke variabelen, voer dan voorwaartse en achterwaartse regressie uit. Controleer de eindmodellen door terug te gaan naar stap 2.
6. 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 dan –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.
7. Controleer de normaliteit van de residuen. Bekijk de normaliteitsplot van de residuen en toets de normaliteit formeel met de Shapiro-Wilks toetsingsgrootheid.
8. Controleer of de waarnemingen onderling onafhankelijk zijn via een residual plot tegen het waarnemingsnummer en de toets van Durbin-Watson.
9. 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.