Tentamen Statistische Methoden
MST–STM
1 april 2009, 9.00–12.00 uur

Toelichting. Een antwoord alleen is *niet* voldoende: er dient een motivatie, toelichting of berekening aanwezig te zijn. Gebruik, tenzij anders voorgeschreven, voor toetsen een α van 5% en voor betrouwbaarheidsintervallen een betrouwbaarheid van 95%. Rapporteer bij toetsen altijd een p-waarde *of* een kritieke waarde.
Beoordeling vindt plaats op basis van uw schriftelijke uitwerking. Uw SPSS "output file" of uw Matlab "diary" file kunnen door ons zonodig geraadpleegd worden. U moet dan wel uw computernummer op de tentamenuitwerking vermelden en de betrokken bestanden schonen.
De laatste vraag is drie van de tien punten waard. Alle overige deelvragen hebben hetzelfde gewicht.

---

1. (a) Waarom heeft een *sample statistic* een *sampling distribution* maar geldt dat niet voor een *population statistic*? Leg kort en duidelijk uit.

   (b) Geef van de volgende modellen aan of ze lineair of niet-lineair zijn in de parameters.

$$\mathrm{E}\,Y = \beta_1 + \beta_2 e^{2/x}$$
$$\mathrm{E}\,Y = \beta_1 \beta_2 x$$
$$\mathrm{E}\,Y = \ln(\beta_1 x) - \ln \beta_1 + \beta_2$$

2. Gegeven is het bestand `apr09twee.dat`.

   (a) Karakteriseer de dataset grafisch door middel van een schets. Geef ook minstens twee kentallen voor locatie (*measures of location*) en twee voor spreiding (*measures of spread*).

   (b) Neem aan dat de getallen realisaties zijn van een stochast met een normale verdeling, $N(\mu, \sigma^2)$. Toets $H_0 : \mu = 3$ tegen $H_1 : \mu \neq 3$.

   (c) Wat is het onderscheidingsvermogen (de *power*) van deze toets indien de werkelijke waarde van μ 4 is?

   (d) Wat denkt u aan de hand van de in (a) verkregen grafiek over de veronderstelling van normaliteit van de gegevens?

3. Gegeven zijn de data in het bestand `apr09drie.xls` en het model

$$\mathrm{E}\,Y = \beta_1 \exp(3x) + \beta_2 \cos(14x)$$

   (a) Fit het model aan de data en geef een betrouwbaarheidsinterval voor de variantie van de meetfout. Leg uit hoe je hier een schatting van de variantie kunt geven zonder een model te fitten (niet uitvoeren).

   (b) Geef aan de hand van de residuen uw oordeel over de modeladequaatheid, dat wil zeggen, een oordeel over de vraag of de data werkelijk past bij het gegeven model.

   (c) Beschrijf kort de Lack-of-Fit toets en leg uit waarom hiermee een uitspraak over de modeladequaatheid gedaan kan worden.

(d) Voer de toets uit op bovenstaand model en op de fit van een simpele kwadratische kromme $\mathrm{E}\,Y = \beta_3 + \beta_4 x^2$ op deze data. Rapporteer bij elke toets één relevant getal en je uitspraak over de bijbehorende nulhypothese.

(e) Ga na dat in het oorspronkelijke model de nulhypothese $H_0 : \beta_1 = \beta_2$ getoetst kan worden met behulp van het model

$$\mathrm{E}\,Y = \beta_5[\exp(3x) + \cos(14x)] + \beta_6 \cos(14x).$$

Leg uit en voer de toets uit.

4. Het percentage sludge (altijd > 0 en $< 100\%$) in het afvalwater van een continu productieproces is afhankelijk van een drietal factoren. Op het hele uur worden telkens twee monsters genomen, waarvoor deze factoren en het sludge-percentage worden gemeten. De resultaten staan in het bestand `sludge.xls` met de kolommen X_1, X_2, X_3, en Y (sludge); de eerste kolom correspondeert met de meettijdstippen. Men wil het sludge percentage terugbrengen en onderzoekt de mogelijkheid dit te bereiken door te sturen met X_1, X_2 en X_3. Hiertoe wenst men een statistische analyse, waarvoor u verantwoordelijk bent.

   Geef uw analyse en rapporteer uw conclusies voor de plantmanager.

Antwoorden

1a *Een sample statistic (Nederlands: steekproefgrootheid) is een grootheid die je aan de hand van een steekproef (sample) berekent. Bij het nemen van een nieuwe steekproef wordt de uitkomst (als regel) anders. De* sampling distribution *is de hierbij horende kansverdeling die beschrijft* hoe *de grootheid varieert. Een* population statistic *is een kenmerk van de populatie als geheel, en derhalve een vast getal.*

Concreet: de fractie op 1 april 2009 om 9:00 uur in Nederland aanwezige personen die 42 jaar of ouder zijn is een population statistic*; de fractie van zulk soort mensen in een steekproef van 1000 personen is een* sample statistic *waarvan de steekproefverdeling bij benadering een normale verdeling heeft. Zie het boek, rond formule 3.14.*

1b *We moeten hier nagaan of de betreffende functie uit te schrijven is in de lineaire vorm*

$$\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \ldots .$$

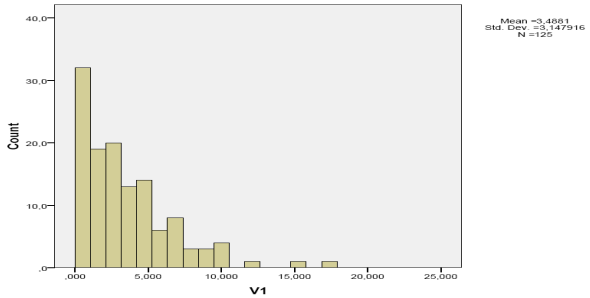
De coëfficiënten mogen allerlei functies zijn, maar niet afhangen van andere parameters.

$\beta_1 + \beta_2 e^{2/x}$ *Lineair*[1]*.* $x_{i1} = 1$ *en* x_{i2} *is een functie van* x*, die toevallig exponentieel is.*

$\beta_1 \beta_2 x$ *Niet lineair. Dit valt niet te ontbinden in bovengenoemde som. Met het invoeren van een nieuwe parameter,* $\beta_3 = \beta_1 \beta_2$*, is dit wel zo, maar het is niet lineair in de gegeven parameters.*

$\ln(\beta_1 x) - \ln \beta_1 + \beta_2$ *Lineair (noot 1). Hiervoor moeten we het uitschrijven:* $\ln(\beta_1) + \ln(x) - \ln \beta_1 + \beta_2 = \ln(x) + \beta_2$*, wat lineair is met* $x_{i1} = 0$ *en* $x_{i2} = 1$*.*

2a *Hier kan men een schets geven van het histogram (20 cellen is genoeg, 12 cellen aan de krappe kant), dan wel de boxplot schetsen. Merk op:* $n = 125$*. Kentallen voor ligging: gemiddelde = 3.49, mediaan = 2.50. Kentallen voor variabiliteit: range = 17.28, kwartielafstand of IQR = 3.90, standaardafwijking = 3.15, variantie = 9.91.*



[1]Sommige kandidaten gebruikten een transformatie: $\ln(\beta_1 + \beta_2 e^{2/x})$, maar dit valt geheel niet te ontbinden. Ernstiger, de aangehaalde uitkomst $\ln(\beta_1) + \ln(\beta_2)2/x$ is fout. Een ander voorbeeld $\exp(\ln(\beta_1 x) - \ln \beta_1 + \beta_2)$ is niet gelijk aan $\beta_1 x - \beta_1 + \exp(\beta_2)$. Het toont een structureel probleem met wiskunde.

2b *In verband met vraag c bepalen we eerste het kritieke gebied C dat wordt bepaald door $P(\bar{x} \in C|H_0) \leq \alpha$. C is dan van de vorm $\bar{x} \geq R$ plus $\bar{x} \leq L$ met $R = 3 + s/\sqrt{n} * t_{0.025}(124)$ en L analoog met minteken. We vinden $R = 3.556$, wegens $\bar{x} = 3.49$; de nulhypothese wordt niet verworpen.*

Als alternatief kan de p-waarde worden bepaald: $t = (3.49 - 3)/0.281 = 1.744$ (bij 124 vrijheidsgraden), dus $p = 0.084$ (tweezijdig). In SPSS wordt dit gevonden met de One-sample-t-test (0.085*).*

2c *Het onderscheidingsvermogen voor $\mu = 4$ wordt gegeven door $P(\bar{x} \in C|\mu = 4)$, met C zoals daarnet bepaald. Enig rekenwerk leert ons dat $\mathrm{P}\big(\bar{X} \geq 3.556 \,|\, \mu = 4\big)$ ongeveer 0.9426, de kans op de linkerstaart is nihil.*

Via `tcdf` *in Matlab: de kritieke t-waarde is 1.9397; $ES = 1/3.15 = 0.3175$, en de power (rechterstaart)* `1-tcdf(tc-ES*sqrt(125),124)`*, hetgeen 0.9405 oplevert.*

Het onderscheidingsvermogen is dus ongeveer 0.94. Dit is ook logisch, want $\mu = 4$ is vrij ver weg van $\mu = 3$ (vanwege het grote aantal waarnemingen). Veel gemaakte fout: doen alsof de toets eenzijdig is (waarbij dus $t_c = 1.6572$); de power komt dan te hoog uit (0.97).

2d *De scheefheid van de dataset is evident, zowel bij het histogram als bij de boxplot. Andere observaties: a) als de waarnemingen normaal verdeeld zouden zijn, zouden er — rekening houdend met de standaarddeviatie — punten < 0 aanwezig zijn; b) er is een duidelijk aanwijsbare asymmetrie; c) de gevonden verdeling lijkt sterk op een exponentiële verdeling met $\mu = 3.5$.*

3a *Na fitten in SPSS of Matlab (*zonder intercept!*) vinden we $\hat{\sigma^2} = 188$ en 78 vrijheidsgraden, waaruit volgen de methode van paragraaf 3.4 het 95% betrouwbaarheidsinterval volgt: $\sigma^2 \in (141, 265)$.*

Omdat er replicaties in de metingen zijn, kan de variantie van de meetfout ook onafhankelijk bepaald worden, namelijk door telkens de steekproefvariantie te bepalen van de metingen bij dezelfde x-waarde en deze te combineren. We bepalen de zogenaamde pure error sum of squares *(SSPE) en* pure error mean square *(MSPE) die een zuivere schatting voor σ^2 is.*

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	459424,902	2	229712,451	1219,862	,000[a]
	Residual	14688,198	78	188,310		
	Total	474113,100[b]	80			

a. Predictors: cos14x, exp3x

b. This total sum of squares is not corrected for the constant because the constant is zero for regression through the origin.

c. Dependent Variable: y

d. Linear Regression through the Origin

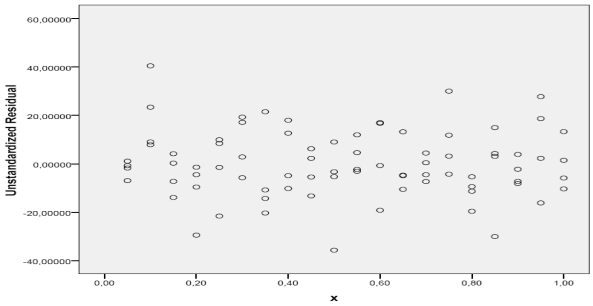
Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	exp3x	8,495	,179	,972	47,537	,000
	cos14x	5,443	2,274	,049	2,394	,019

a. Dependent Variable: y

b. Linear Regression through the Origin

3b *De plot van de residuen versus x ziet er erg mooi uit. Bij $x = 0.1$ lijkt er iets aan de hand, alsof er bias in zit. Als ik kon zou ik deze set controleren.*



3c *Zie het boek voor de beschrijving van de toetsen. Een korte samenvatting:*

De Lack-of-Fit toets is van toepassing bij herhaalde metingen en een model dat aan de metingen gefit is. Men toetst of de geschatte variantie die verkregen wordt dankzij (uitsluitend) de herhaalde metingen gelijk is aan de variantie van de residuen.

Onder de hypothese dat "de data volgens dit model tot stand zijn gekomen" is de variantie van de 'pure error' gelijk aan de variantie geassocieerd met de residuen zonder 'pure error'. Omdat beide slechts geschat kunnen worden, is dit geformaliseerd in een F-toets, waarbij gekeken wordt naar de verhouding MSLF/MSPE tussen de twee geschatte varianties. Is dit verschil te groot, dan wordt er verworpen vanwege de vermoedelijk aanwezige systematische component (lack of fit) in de residuen.

3d *Voor de Lack-of-Fit toets was het Excelsheet, dat bij de som gegeven was, reeds voorbereid om de Pure Error te bepalen. Bij elk van de 18 instelpunten was het gemiddelde en de residuen al gegeven. Men had dus de ANOVA tabel voor de ze toets geheel in Excel kunnen uitvoeren, mits men begreep hoe de berekeningen zijn opgebouwd.*

We vinden dan een F-toetsingsgrootheid van 1.4370*, met een p-waarde van* 0.1480 *(De resultaten met SPSS zijn hieronder). Geen problemen hier dus met de fit.*

Dependent Variable:y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	459424,902 [a]	2	229712,451	1219,862	,000
exp3x	425528,944	1	425528,944	2259,723	,000
cos14x	1079,125	1	1079,125	5,731	,019
Error	14688,198	78	188,310		
Total	474113,100	80			

a. R Squared = .969 (Adjusted R Squared = .968)

Lack of Fit Tests

Dependent Variable:y

Source	Sum of Squares	df	Mean Square	F	Sig.
Lack of Fit	4424,597	18	245,811	1,437	,148
Pure Error	10263,601	60	171,060		

Voor het kwadratische model vinden we bij de Lack-of-Fit toets een p-waarde van 0.00024*: er zijn stevige aanwijzingen dat het kwadratische model niet adequaat is. Een blik op de residuen kan dat bevestigen.*

Dependent Variable:y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	180111,408 [a]	1	180111,408	684,868	,000
Intercept	510,196	1	510,196	1,940	,168
x2	180111,408	1	180111,408	684,868	,000
Error	20512,980	78	262,987		
Total	474113,100	80			
Corrected Total	200624,388	79			

a. R Squared = .898 (Adjusted R Squared = .896)

Lack of Fit Tests

Dependent Variable:y

Source	Sum of Squares	df	Mean Square	F	Sig.
Lack of Fit	10249,379	18	569,410	3,329	,000
Pure Error	10263,601	60	171,060		

3e *Hiertoe moet je zelf een nieuwe predictor-kolom maken door de* $\exp(x)$*-kolom en de* $\cos(x)$*-kolom op te tellen (in zowel Matlab als SPSS een eenvoudige stap) en daarna het model fitten en de hypothese* $H_0 : \beta_6 = 0$ *toetsen. Want als* $\beta_6 = 0$ *dan stelt dit model precies het oorspronkelijke voor met* $\beta_1 = \beta_2$*.*

We fitten en vinden $\hat{\beta}_6 = -3.051$ *met een standard error van* 2.32*, hetgeen oplevert* $t = -1.315$ *en* $p = 0.192$*. De hypothese* $H_0 : \beta_1 = \beta_2$ *wordt dus niet verworpen.*

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	459424,902	2	229712,451	1219,862	,000 [a]
	Residual	14688,198	78	188,310		
	Total	474113,100 [b]	80			

a. Predictors: som, cos14x

b. This total sum of squares is not corrected for the constant because the constant is zero for regression through the origin.

c. Dependent Variable: y

d. Linear Regression through the Origin

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	cos14x	-3,051	2,320	-,027	-1,315	,192
	som	8,495	,179	,992	47,537	,000

a. Dependent Variable: y

b. Linear Regression through the Origin

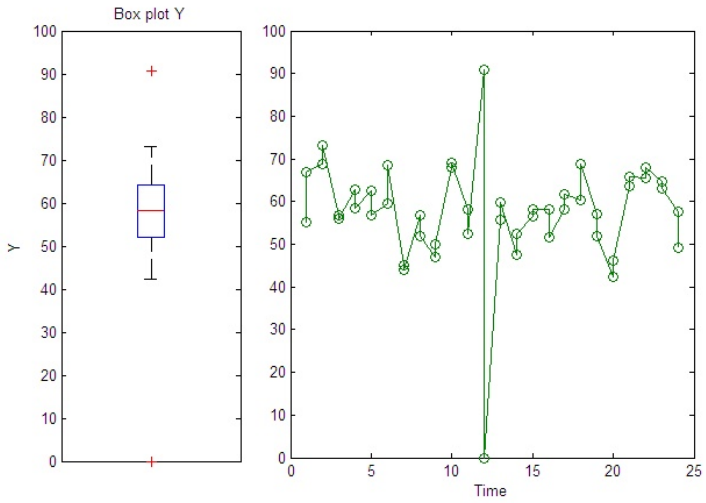
4 *In het volgende wordt een zeer uitgebreid antwoord gegeven waarin alle mogelijke statistische aspecten aan bod komen. Typisch wordt 50% hiervan verwacht en de nadruk ligt op de kwaliteit van de redenering achter de antwoorden. Wat hieronder niet in zit, zijn de antwoorden wanneer de kandidaat andere wegen heeft gevolgd. Een voorbeeld: men kan drie uitbijters detecteren, maar men kan en mag doorgaan met de data na het verwijderen van één, twee, drie of zelfs geen uitbijters.*

Wanneer de 48 datapunten worden ingelezen vindt men een tabel met een kolom voor de tijd en de rest zijn de genoemde variabelen X_i *en* Y*. Er is hier geen enkele reden – zoals uit de uitwerking zal blijken - om bij elke punt in de tijd met 2 waarnemingen er een gemiddelde bij te bepalen*[2]*.*

Eerste verkenning data: visueel inspecteren van de metingen *We bekijken het signaal als functie van de tijd en zien dat er op tijd* $t = 12$ *er*

[2]Bij grote praktijk datasets is vaak verstandig om inderdaad naar gemiddeldes te gaan, maar dan moet men eerst zorgvuldig zien dat er geen potentiële uitbijters zijn. Ten tweede, bij model regressie moet men dan ook vaak rekening houden met de standaardfout die hoort bij dat gemiddelde en die kan dan verschillen. Een onderwerp op zichzelf.

mogelijk twee uitbijters zitten. Dit wordt bevestigd in de box plot van Y.



Eén van de twee uitbijters is 0 *en volgens het gegeven in de vraag kan dat fysisch niet. Deze verwijderen we*[3]*. Bij de andere is er een indicatie, maar dat kan het gevolg zijn van een toevallige combinatie van de instelvariabelen X_i. Er is geen fysische reden om deze uitbijter te verwijderen.*

Een tweede verkenning data: correlaties *De correlatiecoëfficiënt van Y met elk van de drie instelvariabelen is respectievelijk 0.236, 0.217 en 0.631. Er is dus de sterkste afhankelijkheid van Y met X_3 en de minste met X_2.*

De correlatiecoëfficiënten van de instelvariabelen is indicatief voor multicollineariteit. Ze zijn $c_{12} = -0.068, c_{13} = -0.401$ en $c_{23} = 0.158$. Dit zijn betrekkelijk normale waardes die geen indicatie geven van een potentiële moeilijkheid.

Eerste verkenning model: eenvoudige regressie *We gaan hiermee nu eenvoudige regressie met intercept uitvoeren op de beschikbare variabelen. Resultaat is een model, dat met $P = 9 \times 10^{-11}$ duidelijk bij de waarnemingen past. De fouten variantie is 26.8.* [4] *Verder is $R^2 = 0.682$ waaraan we zien dat dit model een 68.2% van de variatie in Y verklaart. En inspecteren wederom de residuen. Het grootste residu is nu weer bij $t = 12$ met een waarde van +22, de volgende is bij $t = 2$ en heeft een residu van +11. Een histogram en een boxplot van de residuen ondersteunen deze identificatie. In principe laten we niets weg,*[5] *maar in dit geval zou geargumenteerd kunnen worden dat rond dit middaguur al een echte uitbijter was gevonden en dat we daarom andere metingen op dat moment mogen wantrouwen. Hier laten we dit de doorslag geven en laten dit punt uit de verzameling. We gaan door met 46 datapunten.*

Regressie met direct de instelvariabelen *Eenvoudige regressie met het mo-*

[3]Een kandidaat verving dit punt door de andere waarneming op hetzelfde tijdstip. Dit is gewoon fout.

[4]In de praktijk heb je vaak variantie gegevens uit historische data, waarmee de waarde van deze variantie getoetst kan worden.

[5]Als een kandidaat dit had gedaan, dan was dat even goed of zelfs beter geweest. De volgende uitwerking zal anders zijn. Zie opmerking aan begin van de vraag.

del

$$\mathrm{E}\,Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{1}$$

geeft nu het volgende resultaat. De toets of het model bij de waarnemingen past heeft $P = 7 \times 10^{-12}$*. De foutenvariantie is 16.8.* $R^2 = 0.726$ *wat betekent dat 72.6% van de variatie in* Y *door het model wordt verklaard.* $R^2_{adj} = 0.706$*.*

De kwadratensom: $SSE = 706.64$ *en de parameters met 'standard error' en betrouwbaarheidsgebied zijn:*

$$\begin{array}{ll} \beta_0 = 38.5 \pm 2.1 & \in (34.2, 42.8) \\ \beta_1 = 0.164 \pm 0.023 & \in (0.117, 0.212) \\ \beta_2 = 0.26 \pm 0.22 & \in (-0.17, 0.70) \\ \beta_3 = 0.214 \pm 0.022 & \in (0.170, 0.260) \end{array}$$

Het is duidelijk van deze gegevens dat de nulhypothese $\beta_2 = 0$ *niet zal worden verworpen, omdat 0 omsloten wordt door zijn betrouwbaarheidsgebied. We kunnen de hypothese ook afzonderlijk toetsen voor de vier parameters:*

H0: $\beta_0 = 0$	*t=18.03*	$P = 2.2 \times 10^{-21}$	
H0: $\beta_1 = 0$	*t=7.02*	$P = 1.4 \times 10^{-8}$	
H0: $\beta_2 = 0$	*t=1.21*	$P = 0.234$	*niet verworpen*
H0: $\beta_3 = 0$	*t=9.66*	$P = 3.1 \times 10^{-12}$	

We gaan dan door met de instelvariabelen X_1 *en* X_3*.*

$$\mathrm{E}\,Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 \tag{2}$$

De toets of het model bij de waarnemingen past heeft $P = 2 \times 10^{-12}$*. De foutenvariantie is 17.0.* $R^2 = 0.716$ *wat betekent dat 71.6% van de variatie in* Y *door het model wordt verklaard.* $R^2_{adj} = 0.703$*.*

De kwadratensom: $SSE = 731.13$ *en de parameters met 'standard error' en betrouwbaarheidsgebied zijn:*

$$\begin{array}{ll} \beta_0 = 39.5 \pm 2.0 & \in (35.5, 43.4) \\ \beta_1 = 0.164 \pm 0.024 & \in (0.117, 0.212) \\ \beta_3 = 0.218 \pm 0.022 & \in (0.174, 0.262) \end{array}$$

Dit is dan het beste model met de gegeven instelvariabelen.

Regressie met uitgebreide modellen *Om de alternatieven te bekijken is de eerste optie die waarbij de tweede order en de kruistermen met* X_i *worden toegevoegd.*

$$\begin{aligned} \mathrm{E}\,Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1 X_1 + \beta_{22} X_2 X_2 \\ + \beta_{33} X_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 \end{aligned} \tag{3}$$

met de 'stepwise' procedure in het boek wordt dit gereduceerd tot

$$\mathrm{E}\,Y = \beta_0 + \beta_3 X_3 + \beta_{11} X_1 X_1 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 \tag{4}$$

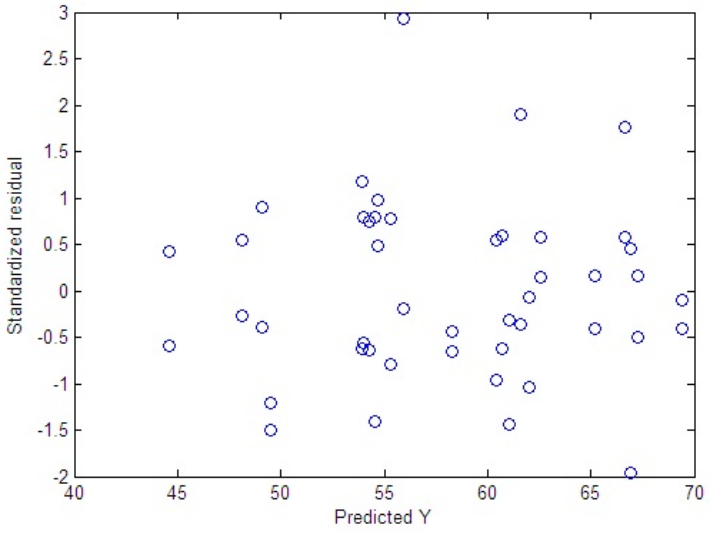
De toets of het model bij de waarnemingen past heeft $P = 5 \times 10^{-13}$*. De foutenvariantie is 17.0.* $R^2 = 0.781$ *wat betekent dat 78.1% van de variatie in* Y *door het model wordt verklaard.* $R^2_{adj} = 0.759$*, wat laat zien dat ondanks het grotere aantal parameters dan in vergelijking (2) dit een iets betere beschrijving geeft. Dit is dan het beste model met de gegeven instelvariabelen.*

Rapportage over model en meetfout *De kwadratensom:* $SSE = 564.67$ *en de parameters met 'standard error' en betrouwbaarheidsgebied zijn:*

$$\begin{array}{ll} \beta_0 = 39.7 \pm 1.9 & \in (36.0, 43.5) \\ \beta_3 = 0.259 \pm 0.041 & \in (0.176, 0.342) \\ \beta_{11} = 0.00215 \pm 0.00033 & \in (0.00149, 0.00280) \\ \beta_{23} = -0.00160 \pm 0.00072 & \in (-0.00306, -0.00014) \\ \beta_{33} = 0.0099 \pm 0.0042 & \in (0.0013, 0.0185) \end{array}$$

De geschatte meetfout is $\hat{\sigma} = \sqrt{SSE/(n-p) = 564.67/(46-5)} = 3.7$.

Evaluatie van het model *Een histogram zal laten zien dat de waarnemingen normaal verdeeld zijn. Een plot van de residuen als functie van de berekende* Y *laat zien dat er geen structuur is.*



Ten derde, men heeft hier herhaalde waarnemingen, waarmee er een Lack-of-fit test gedaan kan worden[6]*:*

	Sum Squares	*df*	*Mean Squares*	*F*	*P*
SSLF	*239.44*	*18*	*13.30*	*0.941*	*0.453*
SSPE	*325.23*	*23*	*14.14*		
SSE	*564.67*	*41*			

Hieruit blijkt dat er geen reden bestaat dit model te verwerpen.

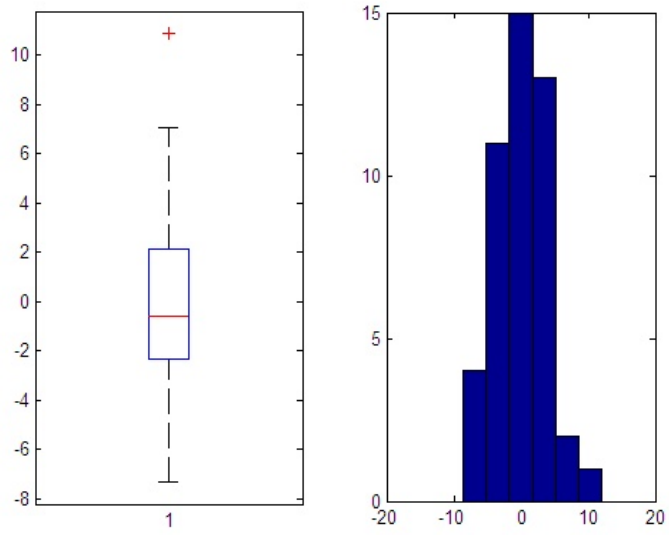
Regressie met alternatieve modellen *Een optie is het toevoegen van de tijd als factor. Wanneer we deze toevoegen aan het model van vergelijkingen (1) en (4) dan wordt in beide gevallen de nulhypothese* $\beta_{tijd} = 0$ *niet verworpen met respectievelijk* $P = 0.74$ *en* $P = 0.65$.

[6] In een eerdere vraag was ook om de Lack-of-fit toets gevraagd. Daar zat de hint, hoe op een gemakkelijke manier SS_{PE} uit het spreadsheet kon worden gekregen. In feite was het in MatLab nog eenvoudiger omdat er per instelpunt twee observaties waren waardoor het korte statement `sum(1*var(reshape(Y,2,23)))` de gevraagde sum van kwadraten geeft, 325.23.

Het is ook mogelijk om te kijken naar termen die bijvoorbeeld afhangen van $\ln(X_i), \exp(X_i), X_i^3, \sin(X_i), \tanh(X_i), \ldots$.

Karakterisering van de residuen *Aangezien de residuen de enige stochastische variabele vormt waarvan het interessant is het te karakteriseren, hier enige observaties.*

Ten eerste, een histogram laat zien dat het er - op het oog - normaal verdeeld uitziet. In het volgende figuur is de optimale bin-breedte genomen op basis van de formule uit het boek, $3.49s/n^{1/3} = 3.45$.



Ten tweede, een aantal kentallen, waarbij het lokatie kental niet interessant is, omdat per definitie 0 *moet zijn. Het zijn: gemiddelde is* 4.1×10^{-14}*, median is -0.56 en het getrimde gemiddelde is -0.08. Spreidingskentallen zijn standaard deviatie, 3.54, interkwartiel afstand, 4.46, en de range, 18.2. De vorm van de verdeling kan gekarakteriseerd worden met de skewness, 0.56, en de kurtosis, 0.95. Beide geven aan dat de verdeling redelijk normaal is*[7].

Al het bovenstaande gaat rondom de statistische analyse van de data. De laatste - en voor het vak de minst interessante - vraag is wat wij de plant manager zouden aanraden.

- *Uit de correlatie en eenvoudige regressie blijkt dat de instelvariabele* X_3 *de belangrijkste factor is, en dat je zou aanraden om in ieder geval deze instelvariabele zo klein mogelijk te maken.*
- *Uit de betere vergelijking (4) blijkt dat de relatie echter iets gecompliceerder ligt.*
- *Je kunt verder laten zien, dat als alle instelvariabelen op* 0 *geplaatst waren, de minimum slurry percentage tussen de 36 en 43.5% zal liggen.*

[7] In feite hebben we bij de stof niet gesproken over wat redelijk is. Er bestaat een zogenaamde bootstrap methode, die aangeeft dat de skewness ligt op het interval (-0.7,1.5) en de kurtosis op (-4,2).

- *De variabiliteit, is zodanig dat als je bijvoorbeeld 1% zou willen halen, elke meting* $(1/3.7)^2 \approx 14$ *moet uitvoeren in plaats van de huidige twee keer.*