

Hoofdstuk 7: Regressie

Inhoud

- 7.0 Formuleblad regressie**
- 7.1 Inleiding**
- 7.2 Schatten en modelcontrole**
- 7.3 Meervoudige lineaire regressie**
- 7.4 Betrouwbaarheidsintervallen en voorspellingsinterval**
- 7.5 Toetsingstheorie**
- 7.6 Transformaties, niet-lineaire verbanden**
- 7.7 ANOVA (1 factor), toetsingstheorie**
- 7.8 Opgaven**
- 7.9 Uitwerkingen van opgaven**

7.0 Formuleblad regressie

Enkelvoudige lineaire regressie

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{S_{xx}} \quad \text{met } S_{xx} = \sum_i (x_i - \bar{x})^2, \quad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$S^2 = \frac{\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n - 2}$$

(stochastisch) betrouwbaarheidsinterval voor $\beta_0 + \beta_1 x$:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - cS \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}, \hat{\beta}_0 + \hat{\beta}_1 x + cS \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right) \quad \text{met } P(T_{n-2} \leq c) = 1 - \frac{1}{2} \alpha$$

(stochastisch) voorspellingsinterval voor Y bij de waarde x :

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - cS^*, \hat{\beta}_0 + \hat{\beta}_1 x + cS^* \right)$$

$$\text{met } P(T_{n-2} \leq c) = 1 - \frac{1}{2} \alpha \quad \text{en} \quad S^* = S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

Meervoudige lineaire regressie

$$SSE = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik})^2$$

$$S^2 = \frac{SSE}{n - (k + 1)}$$

$$S_{YY} = SSR + SSE, \quad S_{YY} = \sum_i (Y_i - \bar{Y})^2$$

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)} \times \frac{SSE}{S_{YY}}, \quad R^2 = 1 - \frac{SSE}{S_{YY}}$$

(stochastisch) betrouwbaarheidsinterval voor β_i : $\left(\hat{\beta}_i - cS_{\hat{\beta}_i}, \hat{\beta}_i + cS_{\hat{\beta}_i} \right)$

met $P(T_{n-(k+1)} \leq c) = 1 - \frac{1}{2} \alpha$

7.1 Inleiding

In regressie-analyse gaat het erom de relatie tussen variabelen weer te geven. In dit hoofdstuk zullen de volgende voorbeelden een centrale rol spelen.

Voorbeeld 7.1.1 Voor 61 grote steden in Engeland en Wales zijn m.b.t. de periode 1958-1964 de waarden van de volgende variabelen x en y verzameld.

x : de (gemiddelde) calcium-concentratie in het drinkwater ('parts per million')

y : het (gemiddelde) jaarlijkse sterftecijfer voor mannen (aantal per 100 000)

x	y	x	y	x	y	x	y
105	1247	10	1637	68	1369	91	1569
44	1702	6	1696	39	1428	60	1527
17	1668	84	1359	50	1257	138	1096
59	1309	101	1236	122	1318	53	1627
5	1466	73	1392	75	1587	16	1591
133	1259	13	1711	21	1260	122	1486
14	1800	12	1755	71	1713	37	1402
27	1427	14	1444	44	1723	81	1485
18	1609	78	1307	13	1557	15	1772
6	1724	49	1591	94	1379	71	1378
10	1558	96	1254	57	1640	8	1828
107	1175	8	1987	8	1742	21	1519
15	1807	20	1491	71	1709	26	1704
5	1486	14	1495	9	1574	14	1581
78	1299	39	1555	20	1625	13	1625
90	1456						

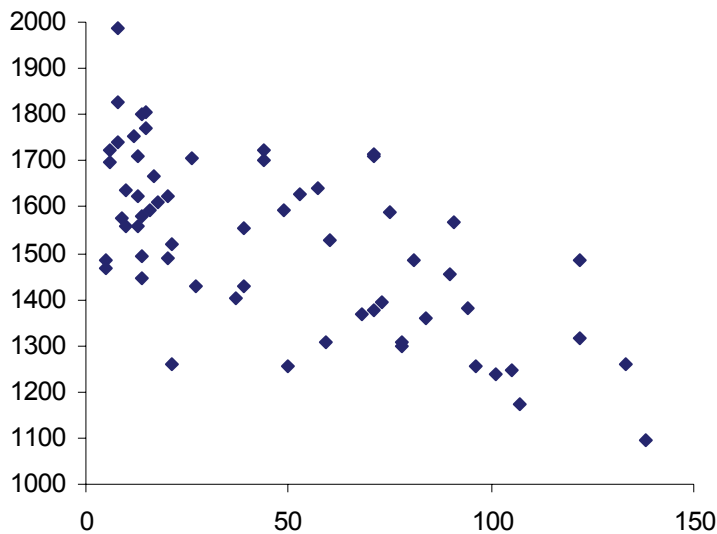
(Herkomst data: Prof. M.J. Gardner, Medical Research Council Environmental Epidemiology Research Unit, Southampton.) Het is interessant te bestuderen hoe de variabele y afhangt van de variabele x . □

Voorbeeld 7.1.2 We bestuderen hoe de weerstand van rubber tegen afschuren beïnvloed wordt door de hardheid (x_1) en de treksterkte (x_2) van rubber. Dertig monsters rubber zijn getest op hardheid (in graden Shore, hoe hoger het getal des te harder is het rubber) en treksterkte (gemeten in kg/cm^2). Daarna is elk monster rubber een tijd lang afgeschuurd (alle monsters op dezelfde wijze) en is het afgeschuurde gewicht aan rubber per uur (y) bepaald (eenheid: g/uur). Het is van belang te analyseren hoe y van de twee variabelen x_1 en x_2 afhangt. De waarden van x_1 , x_2 en y staan in de volgende tabel. □

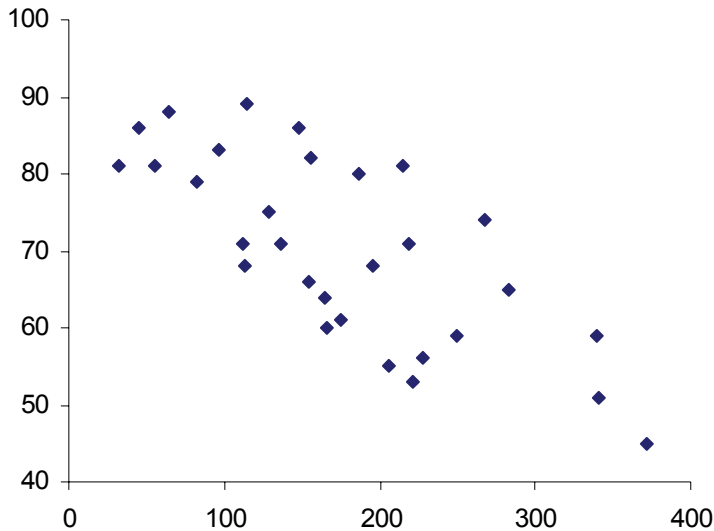
y	hardheid	treksterke	y	hardheid	treksterke	y	hardheid	treksterke
372	45	162	164	64	210	219	71	151
206	55	233	113	68	210	186	80	165
175	61	232	82	79	196	155	82	151
154	66	231	32	81	180	114	89	128
136	71	231	228	56	200	341	51	161
112	71	237	196	68	173	340	59	146
55	81	224	128	75	188	283	65	148
45	86	219	97	83	161	267	74	144
221	53	203	64	88	119	215	81	134
166	60	189	249	59	161	148	86	127

Bij regressie gaat het erom een variabele y zo goed mogelijk te beschrijven of te voorspellen op grond van een of meer verklarende variabelen. In voorbeeld 7.1.1 gaat het erom het sterftcijfer y zo goed mogelijk te beschrijven met de variabele x , de hardheid van drinkwater. In voorbeeld 7.1.2 willen we bestuderen hoe het afgeschuurd gewicht y afhangt van de hardheid en treksterkte van rubber. Laten we ons eerst beperken tot 1 verklarende variabele x . Ten aanzien van voorbeeld 7.1.2 beperken we ons derhalve voorlopig tot de relatie tussen y en x_1 . In figuur 7.1 hebben we de waarden van x en y geplot. De puntenwolk laat een dalende trend zien. In figuur 7.2 hebben we de waarden van x_1 (hardheid) en y (afgeschuurd gewicht) van voorbeeld 7.1.2 geplot. Ook nu is er sprake van een (dalende) trend. Net als in figuur 7.1 is er geen kromming in de puntenwolk te zien. De punten liggen met enige spreiding langs een rechte lijn in beide figuren. Dit verspreid liggen langs een rechte lijn willen we vertalen in een statistisch model.

Figuur 7.1: sterftcijfer (y) versus calcium-concentratie (x)



Figuur 7.2: afgeschuurd gewicht versus hardheid



Vaak is het ontwikkelen van een goede beschrijving, een goed model, een kwestie van ‘trial and error’. Eén van de eenvoudigste modellen is als volgt. Als we een variabele y zo goed mogelijk willen verklaren op grond van een variabele x terwijl we n onafhankelijke waarnemingen $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ hebben, dan beschouwen we y_1, y_2, \dots, y_n als uitkomsten van onafhankelijke stochastische variabelen Y_1, Y_2, \dots, Y_n die verdeeld zijn volgens

$$(7.1.1) \quad Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

waar β_0, β_1 en σ^2 onbekende parameters zijn. Een gelijkwaardige manier van modelleren is als volgt:

$$(7.1.2) \quad Y_i = \beta_0 + \beta_1 x_i + U_i$$

waar U_1, U_2, \dots, U_n onafhankelijke stochastische variabelen zijn verdeeld volgens

$$(7.1.3) \quad U_i \sim N(0, \sigma^2).$$

Wellicht is de tweede manier van modelleren het duidelijkst: we gaan uit van een lineair verband dat gemaskeerd is door “verstoringen” U_i . Het model (7.1.2) resulteert in punten (x_i, y_i) die langs een rechte lijn liggen. Omdat in figuren 7.1 en 7.2 de punten ook langs rechte lijnen liggen, zullen we bovenstaand regressiemodel toepassen op de beide

voorbeelden van deze sectie. Het model dat we hanteren heet het model van de (enkelvoudige) lineaire regressie.

Let wel dat de waarden x_i behandeld worden als vaste getallen terwijl de waarden y_i niet opgevat worden als vaste getallen doch als uitkomsten van stochastische variabelen. Deze handelwijze in regressie is vaak het gevolg van de probleemstelling. In voorbeeld 7.1.2 zijn we geïnteresseerd in de weerstand van rubber tegen afschuren. De hardheid en treksterkte van het rubber vatten we op als gegevens, gegeven getallen, hoewel hardheid en treksterkte ook gemeten moeten worden. Hoe de weerstand van rubber van deze gegevens afhangt, is de vraag van voorbeeld 7.1.2. In bijvoorbeeld landbouwkundige experimenten is het vaak zo dat de waarden x_1, x_2, \dots, x_n al van tevoren vastgesteld zijn. Dit klopt dan precies met het regressiemodel waar x_1, x_2, \dots, x_n staan voor vaste gegeven getallen. Soms kan de variabele x een sterk stochastisch karakter hebben. Het opvatten van x_1, x_2, \dots, x_n als vaste gegeven getallen is dan louter een kunstgreep om de analyse eenvoudig te houden.

7.2 Schatten

We laten in deze sectie zien hoe de parameters β_0, β_1 en σ^2 geschat kunnen worden. We zullen de resultaten direct toepassen op de data van sectie 7.1.

Eén van de manieren om de parameters β_0 en β_1 te schatten is de methode van de kleinste kwadraten. De kleinste-kwadratenschattingen $\hat{\beta}_0$ en $\hat{\beta}_1$ zijn die waarden van β_0 en β_1 waarvoor

$$(7.2.1) \quad \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

als functie van β_0 en β_1 minimaal is. Merk op dat (7.2.1) de som is van de kwadraten van de “verticale” afstanden van de punten $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ tot de lijn $y = \beta_0 + \beta_1 x$. Op grond van het soort functie is het duidelijk dat (7.2.1) een minimum heeft. Om formules voor de kleinste-kwadratenschattingen $\hat{\beta}_0$ en $\hat{\beta}_1$ op te sporen nemen we de afgeleides naar β_0 en β_1 , en stellen we die afgeleides gelijk aan nul. Dit levert de volgende twee vergelijkingen op voor $\hat{\beta}_0$ en $\hat{\beta}_1$:

$$(7.2.2) \quad \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$(7.2.3) \quad \sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Vergelijking (7.2.2) kan herschreven worden als

$$(7.2.4) \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

met $\bar{y} = \sum_i x_i / n$ en $\bar{x} = \sum_i y_i / n$. Blijft over een formule voor $\hat{\beta}_1$ te vinden. Met behulp van (7.2.4) herschrijven we (7.2.3):

$$(7.2.5) \quad \sum_i x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = \sum_i x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_i x_i (x_i - \bar{x}) = 0,$$

zodat we krijgen:

$$(7.2.6) \quad \hat{\beta}_1 = \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})}.$$

Het is gebruikelijk de formule voor $\hat{\beta}_1$ nog een beetje te herschrijven. Vanwege $\sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = \sum_i x_i - \sum_i x_i = 0$ en $\sum_i (y_i - \bar{y}) = 0$ gelden de volgende gelijkheden

$$\sum_i x_i (y_i - \bar{y}) = \sum_i x_i (y_i - \bar{y}) - \bar{x} \sum_i (y_i - \bar{y}) = \sum_i (x_i - \bar{x})(y_i - \bar{y}),$$

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x})y_i,$$

$$\sum_i x_i (x_i - \bar{x}) = \sum_i x_i (x_i - \bar{x}) - \bar{x} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x})(x_i - \bar{x}) = \sum_i (x_i - \bar{x})^2.$$

Hieruit volgt

$$(7.2.7) \quad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})y_i}{\sum_i (x_i - \bar{x})^2}.$$

We zullen nu de meest aannemelijke schattingen (maximum likelihood schattingen) van alle drie de parameters bepalen. Omdat de kansdichtheid van de $N(\beta_0 + \beta_1 x_i, \sigma^2)$ -verdeling, de verdeling van Y_i , gelijk is aan

$$(7.2.8) \quad \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right),$$

met $\exp(z) = e^z$, is de aannemelijkheidsfunctie $L(\beta_0, \beta_1, \sigma^2)$ als volgt te schrijven:

$$(7.2.9) \quad L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

Let wel dat β_0 en β_1 alleen via de som van kwadraten (7.2.1) in de aannemelijkheidsfunctie voorkomen en dat (7.2.9) als functie van β_0 en β_1 maximaal is als (7.2.1) minimaal is. Voor willekeurige waarden voor β_0 , β_1 en σ^2 geldt dus

$$(7.2.10) \quad L(\beta_0, \beta_1, \sigma^2) \leq L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2).$$

Als we nu nog $L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)$ maximaliseren als functie van σ^2 , dan zijn we klaar met de meest aannemelijke schattingen. Zonder bewijs vermelden we dat $\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 / n$ de meest aannemelijke schatting is van σ^2 . Deze schatting zullen we echter niet gebruiken. In plaats van deze formule gebruiken we een aanpassing.

In de enkelvoudige lineaire regressie worden de volgende schatters gebruikt:

$$(7.2.11) \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

$$(7.2.12) \quad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2},$$

$$(7.2.13) \quad S^2 = \frac{\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}.$$

Deze schatters zijn zuiver, dit zullen we alleen voor $\hat{\beta}_0$ en $\hat{\beta}_1$ aantonen in deze sectie. Zonder bewijs vermelden we dat deze schatters ook de beste schatters zijn, in de zin van minimale verwachte kwadratische fout. Voor de zuiverheid van de schatters $\hat{\beta}_0$ en $\hat{\beta}_1$ gaan we $E(\hat{\beta}_0) = \beta_0$ en $E(\hat{\beta}_1) = \beta_1$ aantonen:

$$(7.2.14) \quad E(\hat{\beta}_1) = E\left(\frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2}\right) = \frac{E\left(\sum_i (x_i - \bar{x})Y_i\right)}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})E(Y_i)}{\sum_i (x_i - \bar{x})^2}$$

Uit (7.1.1) of (7.1.2-3) volgt $E(Y_i) = \beta_0 + \beta_1 x_i$. Met behulp van $\sum_i (x_i - \bar{x}) = 0$ en $\sum_i x_i(x_i - \bar{x}) = \sum_i (x_i - \bar{x})^2$, zie eerder deze sectie, vinden we:

$$(7.2.15) \quad E(\hat{\beta}_1) = \frac{\beta_0 \sum_i (x_i - \bar{x}) + \beta_1 \sum_i x_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \beta_1 \frac{\sum_i x_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \beta_1.$$

$$(7.2.16) \quad \begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\bar{Y}) - \bar{x}E(\hat{\beta}_1) = E(\bar{Y}) - \bar{x}\beta_1 \\ &= E\left(\frac{\sum_i Y_i}{n}\right) - \bar{x}\beta_1 = \frac{\sum_i E(Y_i)}{n} - \bar{x}\beta_1 = \frac{\sum_i (\beta_0 + \beta_1 x_i)}{n} - \bar{x}\beta_1 \\ &= \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0 \end{aligned}$$

Voor voorbeeld 7.1.1 hebben we SPSS laten rekenen, een deel van de output is als volgt:

Coefficients	Unstandardized Coefficients	
	B	Std. Error
(Constant)	1676.356	29.298
HARDHEID	-3.226	.485

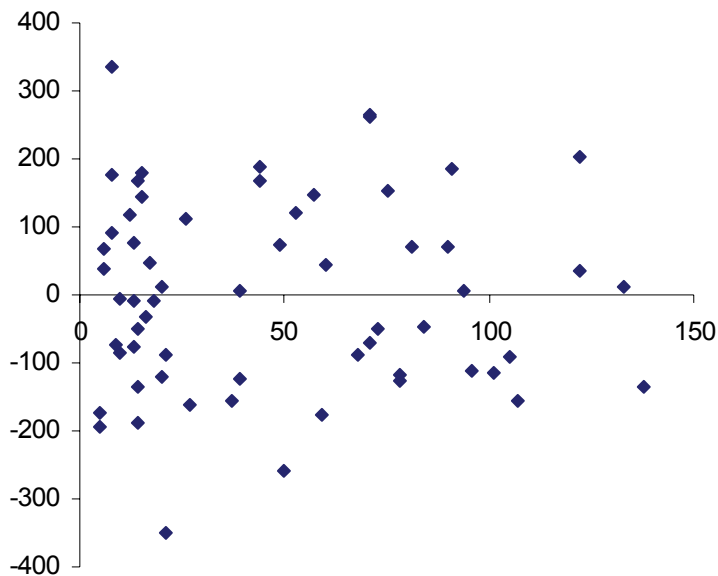
a Dependent Variable: STERFTE

De schatting voor β_1 is -3.226 en 1676.356 is de schatting voor β_0 . De schatting voor σ^2 komt in later te tonen output aan de orde. De betekenis van ‘standard error’ (Std. Error) leggen we ook later uit. Voor voorbeeld 7.1.2 (zonder treksterkte) is de output als volgt.

Coefficients	Unstandardized Coefficients	
	B	Std. Error
(Constant)	550.415	65.787
HARDHEID	-5.337	.923

a Dependent Variable: Y

residu versus hardheid (vb 7.1.1)



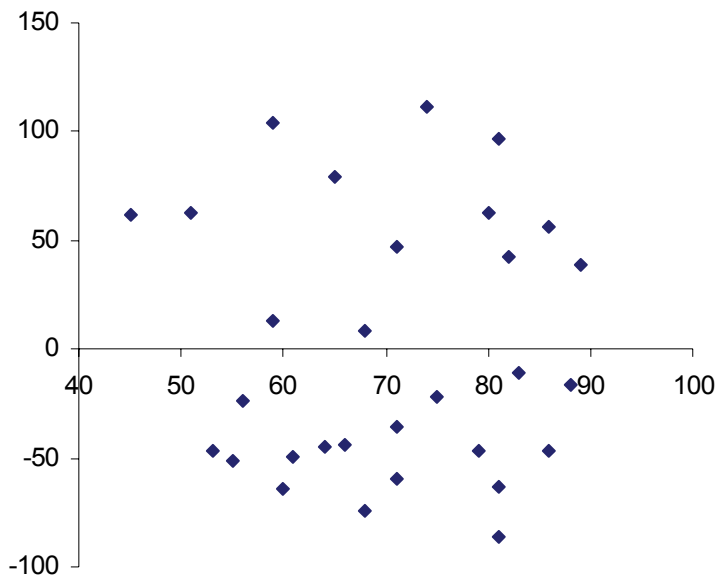
Uitgaande van het regressiemodel beschreven door (7.1.2) en (7.1.3), of gelijkwaardig (7.1.1), zijn de residuen R_i als volgt gedefinieerd:

$$(7.2.17) \quad R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

met $\hat{\beta}_0$ en $\hat{\beta}_1$ de kleinste-kwadratenschatters van β_0 en β_1 . Het is zinvol de waarden van de R_i uit te zetten tegen de waarden x_i , de waarden van de verklarende variabele. Met de plot van het residu en de x -variabele heb je de mogelijkheid de aannames van het regressiemodel te controleren. Ten eerste kun je kijken of er een patroon te zien is in de puntenwolk van de punten (x_i, R_i) . Met R_i schat je $Y_i - \beta_0 - \beta_1 x_i = U_i$. Omdat de ruistermen U_i geacht worden onderling onafhankelijk te zijn dien je een chaotische puntenwolk te zien. Elk soort patroon duidt erop dat een of meer van de modelaannames niet klopt. In de residuenplots van voorbeelden 7.1.1 en 7.1.2 menen we geen patroon te zien, dus het model is gezien deze modelcontrole OK.

Naast het speuren naar een patroon in de residuen, kan men ook kijken naar de verdeling van de residuen. Aangezien de storingen U_i volgens het model $N(0, \sigma^2)$ -verdeeld zijn, moeten de residuen R_i bij benadering $N(0, \sigma^2)$ -verdeeld zijn daar je U_i met R_i schat. Controleren op normaliteit kan uitgevoerd met statistische technieken die in collegejaar 2004/2005 in het vak Verkeer aan de orde komen. Vaak echter laat men dit achterwege en beperkt men zich tot het bekijken van residuenplots zoals getoond in deze sectie.

residu versus hardheid (vb 7.1.2)



Uitgaande van normaliteit is het interessant te kijken of er residuen zijn die sterk afwijken. De gestandaardiseerde residuen R_i/S , met S de schatter van σ , is bij benadering $N(0,1)$ -verdeeld. (In sommige boeken verbetert men de formule R_i/S maar dit laten wij achterwege.) Men kan nagaan dat hoge waarden zoals 3 en 4 voor $|R_i/S|$ zeer zelden mogen voorkomen vanwege de $N(0,1)$ -verdeling. Als dergelijke waarden toch voorkomen, is dit een reden om het model te heroverwegen ofwel de herkomst van de metingen te onderzoeken. Dit soort onderzoek stellen we uit omdat we nog geen uitkomst van S getoond hebben.

7.3 Meervoudige Lineaire Regressie

Meervoudige lineaire regressie is een uitbreiding van enkelvoudige lineaire regressie, waarbij het mogelijk is meer verklarende variabelen op te nemen. Enkelvoudige lineaire regressie wordt wel kortweg aangeduid met lineaire regressie. In plaats van meervoudige lineaire regressie spreekt men wel over multipele lineaire regressie of kortweg multipele regressie.

Veel van de theorie van de multipele regressie is een directe generalisatie van de theorie van de lineaire regressie. Hoewel de formules conceptueel veelal niet moeilijker zijn dan bij lineaire regressie, zijn de formules technisch wel ingewikkelder en moet meestal een beroep op de computer gedaan worden om de berekeningen uit te voeren. In dit hoofdstuk laten we output van SPSS zien, we zijn daar in de vorige sectie al mee begonnen.

Net als in de vorige secties beschouwen we een variabele y , de **afhankelijke variabele**, die we zo goed mogelijk willen verklaren/voorspellen. In de lineaire regressie is er 1 verklarende variabele x waarmee y verklaard/voorspeld wordt. Nu we multiple regressie behandelen hebben we in het algemeen k **verklarende variabelen** x_1, x_2, \dots, x_k (het aantal k hangt van de situatie af, k is een bekend geheel getal). Verklarende variabelen worden ook wel onafhankelijke variabelen genoemd. Laten we uitgaan van n waarnemingen. We hebben dan waarden y_1, y_2, \dots, y_n van de afhankelijke variabele en bij elke y_i de bijbehorende waarden $x_{i1}, x_{i2}, \dots, x_{ik}$ van de verklarende variabelen. In de multipele regressie veronderstellen we het volgende. De waarden y_1, y_2, \dots, y_n vatten we op als uitkomsten van stochastische variabelen Y_1, Y_2, \dots, Y_n waarvoor geldt:

$$(7.3.1) \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + U_i,$$

met onafhankelijke stochastische variabelen U_1, U_2, \dots, U_n die $N(0, \sigma^2)$ -verdeeld zijn, en waar $\beta_0, \beta_1, \dots, \beta_k$ en σ^2 onbekende parameters zijn.

De parameters $\beta_0, \beta_1, \dots, \beta_k$ worden geschat door de som van de kwadratische afwijkingen

$$(7.3.2) \quad \sum_i (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik})^2$$

te minimaliseren als functie van $\beta_0, \beta_1, \dots, \beta_k$. We noteren deze kleinste-kwadraten-schatters met $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$. De resulterende minimale kwadraatsom noteren we met SSE , dus

$$(7.3.3) \quad SSE = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik})^2.$$

Als schatter van σ^2 (de variantie van de “storingen” U_i) nemen we

$$(7.3.4) \quad S^2 = \frac{SSE}{n - (k + 1)}.$$

Dit is de onmiddellijke generalisatie van (7.2.13). Het getal $n - (k + 1)$ is het aantal **vrijheidsgraden**. We kunnen dit onthouden door de vuistregel dat voor elke geschatte parameter β_i één vrijheidsgraad verloren gaat. We schatten hier $\beta_0, \beta_1, \dots, \beta_k$ ($k + 1$ parameters) en komen uit op $n - (k + 1)$ vrijheidsgraden.

We noemen

$$(7.3.5) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

de **regressievergelijking**, die we dus nooit en te nimmer te weten komen, omdat $\beta_0, \beta_1, \dots, \beta_k$ onbekend zijn en blijven. Als $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ de kleinste-kwadraten-schattingen zijn van $\beta_0, \beta_1, \dots, \beta_k$ (we maken weer geen onderscheid in notatie tussen schatters en schattingen) dan heet

$$(7.3.6) \quad y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

de aangepaste vergelijking.

We gaan nu terug naar voorbeeld 7.1.2 en proberen meervoudige lineaire regressie toe te passen. Het afgeschuurde gewicht rubber per uur is de afhankelijke variabele. De verklarende variabelen zijn hardheid (= x_1) en treksterkte (= x_2). Met SPSS kun je o.a. de volgende output krijgen.

ANOVA

	Sum of Squares	df	Mean Square	F
Regression	189061.623	2	94530.811	70.997
Residual	35949.744	27	1331.472	
Total	225011.367	29		

a Predictors: (Constant), TREKSTER, HARDHEID

Coefficients

	Unstandardized Coefficients		T
	B	Std. Error	
(Constant)	885.161	61.752	14.334
HARDHEID	-6.571	.583	-11.267
TREKSTER	-1.374	.194	-7.073

a Dependent Variable: Y

HARDHEID en TREKSTER zijn de namen die we aan SPSS hebben doorgegeven als namen van de verklarende variabelen. (Voor de tweede naam: we konden maar 8 letters gebruiken.) De schattingen $\hat{\beta}_i$ kunnen we halen uit de tweede kolom van de tweede tabel. De aangepaste vergelijking is voor voorbeeld 7.1.2 derhalve

$$(7.3.7) \quad y = 885.161 - 6.571x_1 - 1.374x_2 .$$

In een vergelijking als (7.3.7) kun/mag je variabelen x_1, x_2 en y vervangen door hun namen. In het algemene regressiemodel met k verklarende variabelen is het residu van de waarneming i gedefinieerd als

$$(7.3.8) \quad R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik} .$$

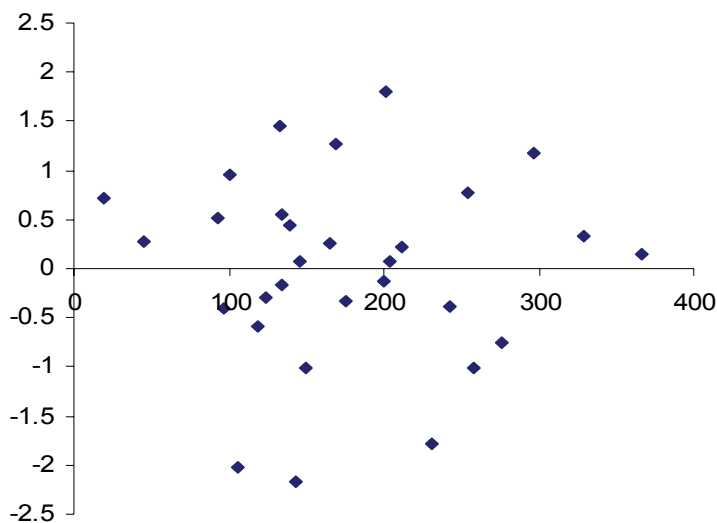
De kwadraatsom SSE is letterlijk de som van de kwadraten van de residuen. De kwadraatsom SSE wordt daarom ook wel Residual Sum of Squares genoemd en afgekort met RSS . Voor SSE moeten we naar de eerste tabel van de output kijken, kijken naar rij ‘residual’ en kolom ‘Sum of Squares’: 35949.744 . Het juiste aantal vrijheidsgraden is vermeld in dezelfde rij: $n - (k + 1) = 30 - 3 = 27$. Omdat we σ^2 schatten met $S^2 = SSE / (n - (k + 1))$ is de schatting van σ^2 gelijk aan $35949.744 / 27 = 1331.472$, te vinden als Residual Mean Squares.

De uitkomsten van

$$(7.3.9) \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}$$

noemt men wel de aangepaste waarden, ofwel ‘fitted values’ ofwel ‘predicted values’. De predicted values staan voor het verklaarde deel van de waarden y . Net als bij enkelvoudige lineaire regressie schat je met de residuen R_i de storingen U_i . De residuen representeren het onverklaarde deel van de data, van de waarden van y . Als je de residuen plot dien je een chaotische puntenwolk te zien, omdat de storingen (ruistermen) U_i geacht worden onderling onafhankelijk te zijn. Elk soort patroon duidt erop dat een of meer van de modelaannames niet klopt. Je kunt de waarden van de residuen uitzetten tegen de waarden van elk van de verklarende variabelen. Een plot van de residuen (verticaal) en de predicted values (horizontaal) is ook zeer gebruikelijk om te bekijken voor een modelcontrole. Voor voorbeeld 7.1.2 komen we zo op het volgende plotje. We zien een chaotisch beeld zoals we dat horen te zien.

vb 7.1.2: residu versus predicted value



In de meervoudige lineaire regressie geldt de volgende gelijkheid die we niet zullen bewijzen:

$$(7.3.10) \quad S_{YY} = SSR + SSE,$$

waar $S_{YY} = \sum_i (Y_i - \bar{Y})^2$, met $\bar{Y} = \sum_i Y_i / n$, de totale spreiding (in de afhankelijke variabele) vertegenwoordigt en de kwadraatsom SSR als volgt gedefinieerd is:

$$(7.3.11) \quad SSR = \sum_i (\hat{Y}_i - \bar{Y})^2.$$

De kwadraatsom SSR kunnen we ook in de output vinden: het is de Sum of Squares due to Regression, zoek bij kolom Sum of Squares en rij Regression, het is 189061.623 voor voorbeeld 7.1.2. In vergelijking (7.3.10) wordt de totale spreiding S_{YY} opgedeeld in twee stukken: SSE staat voor het deel van de spreiding dat niet verklaard door de verklarende variabelen en SSR staat voor het deel van de spreiding dat we kennelijk wel verklaren met de verklarende variabelen. Men noemt

$$(7.3.12) \quad R^2 = \frac{SSR}{S_{YY}} = 1 - \frac{SSE}{S_{YY}}$$

de **fractie verklaarde variantie**. De grootheid R^2 wordt ook de gekwadrateerde multiële correlatie coëfficiënt genoemd, en ook wel de multiële coëfficiënt of determination. Uit (7.3.10) volgt onmiddellijk $0 \leq R^2 \leq 1$. Verder duidt een kleine waarde van R^2 op weinig invloed van de verklarende variabelen x_1, x_2, \dots, x_k en een waarde vlakbij 1 op grote invloed. De interpretatie van R^2 is echter niet zo eenvoudig. Hoe meer verklarende variabelen men opneemt, des te beter volgen de aangepaste waarden de waargenomen y , des te kleiner SSE en des te dichter R^2 bij 1. De kwadraatsom SSE wordt echter (vrijwel) **altijd** kleiner ook al voeg je onzinnige variabelen toe als verklarende variabelen. Toch is het natuurlijk geen goede zaak alsmaar meer verklarende variabelen te introduceren. Er komt dan steeds meer “ruis” binnen. Er zijn verschillende methoden om een “straf” te leggen op het opnemen van veel verklarende variabelen. Een vaak toegepaste methode is de berekening van een zogenaamde ‘adjusted’ R^2 gegeven door

$$(7.3.13) \quad R_a^2 = 1 - \left(\frac{n-1}{n-(k+1)} \right) \frac{SSE}{S_{YY}} = 1 - \frac{S^2}{S_{YY}/(n-1)}.$$

Merk op dat $S_{YY}/(n-1)$ de steekproefvariantie van Y is. Als we meer verklarende variabelen in ons model opnemen, wordt k groter en SSE kleiner. Door het kleiner worden van SSE komt R^2 dichter bij 1. Daarentegen zorgt een grotere k ervoor dat $(n-1)/(n-(k+1))$ groter wordt. Het kleiner worden van SSE wordt hierdoor

gecorrigeerd in R_a^2 . Wordt SSE veel kleiner, dan komen we met R_a^2 toch dichterbij 1 en geeft R_a^2 terecht een verbetering aan.

De R_a^2 voor voorbeeld 7.1.2 en het model met 2 verklarende variabelen is gelijk aan

$$(7.3.14) \quad R_a^2 = 1 - \frac{29}{27} \times \frac{35949.744}{225011.367} = 1 - \frac{29}{27} \times 0.1598 = 1 - 0.1716 = 82.8\% .$$

Als we alleen hardheid als verklarende variabele opnemen, dan blijkt (deze berekening tonen we hier niet) R_a^2 uit te komen op 52.8%. Kennelijk verklaren we met het toevoegen van een tweede verklarende variabele in voorbeeld 7.1.2 veel meer van de spreiding.

7.4 Betrouwbaarheidsintervallen en voorspellingsinterval

In hoofdstuk 5 van het dictaat zijn voor allerlei parameters betrouwbaarheidsintervallen gegeven. Uitgaande van het model van de meervoudige lineaire regressie, zie (7.3.1), ziet het betrouwbaarheidsinterval voor β_i met betrouwbaarheid $\gamma = 1 - \alpha$ er als volgt uit:

$$(7.4.1) \quad \left(\hat{\beta}_i - cS_{\hat{\beta}_i}, \hat{\beta}_i + cS_{\hat{\beta}_i} \right),$$

waarbij $S_{\hat{\beta}_i}$ de geschatte standaardafwijking van $\hat{\beta}_i$ is, en c gegeven wordt door

$$(7.4.2) \quad P(T_{n-(k+1)} \leq c) = 1 - \frac{1}{2} \alpha .$$

De geschatte standaardafwijking van $\hat{\beta}_i$ wordt ook wel standaardfout of **standard error (s.e.)** genoemd. In de output van een statistiekpakket als SPSS kan het wemelen van standard errors. Voor elke gebruikte schatter kan een standaardafwijking afgeleid worden, als deze geschat wordt spreekt men van standard error van de betrokken schatter/schatting.

In het kader van regressie zullen we standard errors uit de computer output halen, in het algemeen. We zullen in het algemeen ons dus niet inlaten met formules voor standard errors. Alleen in het geval van enkelvoudige lineaire regressie en het betrouwbaarheidsinterval voor β_1 bepalen we de betrokken standard error, de s.e. van $\hat{\beta}_1$. De formule van de schatter $\hat{\beta}_1$ is (zie 7.2.12):

$$(7.4.3) \quad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})Y_i}{S_{xx}} \quad \text{met } S_{xx} = \sum_i (x_i - \bar{x})^2$$

De variantie hiervan is als volgt:

$$\begin{aligned}
\text{var}(\hat{\beta}_1) &= \text{var}\left(\frac{\sum_i (x_i - \bar{x})Y_i}{S_{xx}}\right) = \frac{\text{var}\left(\sum_i (x_i - \bar{x})Y_i\right)}{S_{xx}^2} \\
(7.4.4) \quad &= \frac{\sum_i (x_i - \bar{x})^2 \text{var}(Y_i)}{S_{xx}^2} \quad (Y_1, Y_2, \dots, Y_n \text{ zijn o.o.}) \\
&= \frac{\sum_i (x_i - \bar{x})^2 \sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

De standaardafwijking van $\hat{\beta}_1$ is $\sigma/\sqrt{S_{xx}}$ en dus is de standard error van $\hat{\beta}_1$ gelijk aan (de uitkomst van) $S/\sqrt{S_{xx}}$, met S bepaald door (7.2.13). De standard error van $\hat{\beta}_1$ in geval van enkelvoudige lineaire regressie kan in dit vak op twee manieren gevonden worden: via computer output of via formule $S/\sqrt{S_{xx}}$. Het betrouwbaarheidsinterval voor β_1 is gebaseerd op het feit dat

$$(7.4.5) \quad \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

geldt en dat bovendien bewezen kan worden (zullen we niet doen):

$$(7.4.6) \quad \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}.$$

Laten we het voorgaande toepassen op de twee voorbeelden. Van voorbeeld 7.1.2 is output in sectie 7.3. Laten we het 95%-betrouwbaarheidsinterval bepalen voor β_2 , de regressiecoëfficiënt van treksterkte. De schatting van β_2 is -1.374, de bijbehorende standard error is 0.194 en met behulp van de t_{27} -verdeling vinden we de waarde van c : 2.05. Het (numeriek) 95%-betrouwbaarheidsinterval voor β_2 is derhalve

$$(7.4.7) \quad (-1.374 - 2.05 \times 0.194, -1.374 + 2.05 \times 0.194) = (-1.77, -0.98).$$

Output van voorbeeld van 7.1.1 staat in sectie 7.2. We bepalen het 95%-betrouwbaarheidsinterval voor β_1 , de regressiecoëfficiënt van de (enige) verklarende variabele hardheid. We lezen af: uitkomsten van $\hat{\beta}_1$ en $S_{\hat{\beta}_1}$ zijn resp. -3.226 en 0.485. Het aantal vrijheidsgraden is $n - (k + 1) = 61 - 2 = 59$, uit de t_{59} -verdeling volgt $c = 2.00$ (interpolatie). Het 95%-betrouwbaarheidsinterval voor β_1 is daarom:

$$(7.4.8) \quad (-3.226 - 2.00 \times 0.485, -3.226 + 2.00 \times 0.485) = (-4.20, -2.26).$$

In de rest van deze sectie beperken we ons tot enkelvoudige lineaire regressie en introduceren nog een betrouwbaarheidsinterval en, tenslotte, een voorspellingsinterval.

Enkelvoudige Lineaire Regressie: betrouwbaarheidsinterval voor $E(Y) = \beta_0 + \beta_1 x$

Veelal zijn we erin geïnteresseerd bij een nieuwe waarde van x de bijbehorende waarde van Y te voorspellen. Allereerst merken we op dat het voorspellen van een waarde van Y behorende bij een x die niet in het door x_1, x_2, \dots, x_n bepaalde interval behoort een hachelijke zaak is. Immers, we hebben voor zo'n x geen enkele indicatie of het model van de enkelvoudige lineaire regressie wel geldt.

In de rest van deze sectie gaan we er vanuit dat we wel een waarde x hebben waarvoor het model van de enkelvoudige lineaire regressie geldt: de bijbehorende (nog niet gerealiseerde) waarneming Y is dan $N(\beta_0 + \beta_1 x, \sigma^2)$ -verdeeld. Allereerst zullen we het betrouwbaarheidsinterval construeren voor de verwachting van Y , $E(Y) = \beta_0 + \beta_1 x$, gebaseerd op de (gerealiseerde) waarnemingen Y_1, Y_2, \dots, Y_n en de bijbehorende waarden x_1, x_2, \dots, x_n van de verklarende variabele. Voor het betrouwbaarheidsinterval voor $E(Y) = \beta_0 + \beta_1 x$ moeten we de volgende stochastische variabele bestuderen:

$$(7.4.9) \quad \hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1(x - \bar{x}),$$

waar we (7.2.11) gebruikt hebben. Om rekenpartijen te beperken gaan we uit van de volgende gelijkheid die we niet zullen bewijzen:

$$(7.4.10) \quad \text{cov}(\bar{Y}, \hat{\beta}_1) = 0.$$

Volgens het model zijn Y_1, Y_2, \dots, Y_n onderling onafhankelijk en is Y_i verdeeld, volgens $N(\beta_0 + \beta_1 x_i, \sigma^2)$. Hieruit volgt dat $\hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1(x - \bar{x})$ ook normaal verdeeld is, want $\hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1(x - \bar{x})$ is uiteindelijk een lineaire functie van Y_1, Y_2, \dots, Y_n . Verwachting en variantie van \bar{Y} zijn als volgt:

$$(7.4.11) \quad E(\bar{Y}) = E\left(\frac{\sum Y_i}{n}\right) = \frac{\sum E(Y_i)}{n} = \frac{\sum (\beta_0 + \beta_1 x_i)}{n} = \beta_0 + \beta_1 \bar{x},$$

$$(7.4.12) \quad \text{var}(\bar{Y}) = \text{var}\left(\frac{\sum_i Y_i}{n}\right) = \frac{\sum_i \text{var}(Y_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Bij (7.4.12) is de onafhankelijkheid van de Y_i gebruikt. De verwachting van $\hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1(x - \bar{x})$ is gelijk aan

$$(7.4.13) \quad E(\bar{Y}) + E(\hat{\beta}_1)(x - \bar{x}) = \beta_0 + \beta_1 \bar{x} + \beta_1(x - \bar{x}) = \beta_0 + \beta_1 x,$$

en de variantie gelijk aan:

$$(7.4.14) \quad \text{var}(\bar{Y}) + \text{var}(\hat{\beta}_1(x - \bar{x})) + 2 \text{cov}(\bar{Y}, \hat{\beta}_1(x - \bar{x})).$$

Op grond van de definitie van covariantie zal duidelijk zijn dat $\text{cov}(\bar{Y}, \hat{\beta}_1(x - \bar{x}))$ gelijk is aan $(x - \bar{x}) \times \text{cov}(\bar{Y}, \hat{\beta}_1)$ en vanwege (7.4.10) nul is. De variantie van $\hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1(x - \bar{x})$ is daarom:

$$(7.4.15) \quad \text{var}(\bar{Y}) + \text{var}(\hat{\beta}_1(x - \bar{x})) = \frac{\sigma^2}{n} + (x - \bar{x})^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right).$$

Kortom: $\hat{\beta}_0 + \hat{\beta}_1 x \sim N\left(\beta_0 + \beta_1 x, \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)\right)$. Standaardiseren levert:

$$(7.4.16) \quad \frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x)}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim N(0,1).$$

Zonder bewijs vermelden we:

$$(7.4.17) \quad \frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x)}{S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

met S bepaald door (7.2.13). Het betrouwbaarheidsinterval voor $E(Y) = \beta_0 + \beta_1 x$ met betrouwbaarheid $\gamma = 1 - \alpha$ is gebaseerd op (7.4.17) en ziet er zo uit:

$$(7.4.18) \quad \left(\hat{\beta}_0 + \hat{\beta}_1 x - cS \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}, \hat{\beta}_0 + \hat{\beta}_1 x + cS \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right),$$

met c gegeven door $P(T_{n-2} \leq c) = 1 - \frac{1}{2}\alpha$.

Laten we dit betrouwbaarheidsinterval toepassen op voorbeeld 7.1.1. Een deel van de output verkregen met SPSS is als volgt.

Coefficients	Unstandardized Coefficients	
	B	Std. Error
(Constant)	1676.356	29.298
HARDHEID	-3.226	.485

a Dependent Variable: STERFTE

ANOVA				
	Sum of Squares	Df	Mean Square	F
Regression	906185.333	1	906185.333	44.296
Residual	1206988.339	59	20457.429	
Total	2113173.672	60		

a Predictors: (Constant), HARDHEID
b Dependent Variable: STERFTE

(Het eerste deel van deze output hebben we al in sectie 7.2 getoond.) Laten we het 95%-betrouwbaarheidsinterval bepalen voor $E(Y) = \beta_0 + \beta_1 x$ met $x = 80$, oftewel het gemiddelde sterftcijfer voor mannen voor steden met een hardheid van 80 (calcium-concentratie in drinkwater, parts per million). We hebben niet genoeg aan de getoonde output. Het is echter genoeg ook het steekproefgemiddelde en de steekproefvariantie van de verklarende variabele x te weten:

	Steekproefgemiddelde	Steekproefvariantie
x	47.18	1451.15

Let wel: S_{xx} is $n - 1$ maal de steekproefvariantie van x , $60 \times 1451.15 = 87069$ voor onze data set. We vullen het betrouwbaarheidsinterval stukje bij beetje in:

$$\hat{\beta}_0 + \hat{\beta}_1 x = 1676.356 - 3.226 \times 80 = 1418.28, \text{ uitkomst van } S \text{ is } \sqrt{20457.429} = 143.03,$$

$$c = 2.00 \text{ (} t_{59} \text{-verdeling), } \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} = \sqrt{\frac{1}{61} + \frac{(80 - 47.18)^2}{87069}} = 0.170, \text{ deze resultaten}$$

invullen levert het 95%-betrouwbaarheidsinterval voor $E(Y) = \beta_0 + \beta_1 x$ voor $x = 80$:

$$(7.4.19) \quad (1418.28 - 2.00 \times 143.03 \times 0.170, 1418.28 + 2.00 \times 143.03 \times 0.170) \\ = (1369.6, 1466.9) .$$

Enkelvoudige lineaire regressie: voorspellingsinterval

Als we in het enkelvoudige lineaire regressiemodel bij een gegeven waarde x de waarde van Y willen voorspellen, dan hebben we te maken met **twee bronnen van variatie**:

(1) Y zelf is een stochastische variabele (normaal verdeeld met verwachting $\beta_0 + \beta_1 x$ en variantie σ^2),

(2) we kennen de verwachting $\beta_0 + \beta_1 x$ van Y niet en moeten deze schatten, dit brengt een schattingsfout mee.

Voor de constructie van het voorspellingsinterval voor Y moet de “voorspellingsfout”

$Y - \hat{\beta}_0 - \hat{\beta}_1 x$ bestudeerd worden, deze is als volgt verdeeld:

$$(7.4.20) \quad Y - \hat{\beta}_0 - \hat{\beta}_1 x \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)\right).$$

Voor (7.4.20) hebben we gebruik gemaakt van de onafhankelijkheid van Y en Y_1, Y_2, \dots, Y_n : aangezien $\hat{\beta}_0 + \hat{\beta}_1 x$ een functie is van Y_1, Y_2, \dots, Y_n dan zijn Y en $\hat{\beta}_0 + \hat{\beta}_1 x$ ook (onderling) onafhankelijk en is de variantie van het verschil de som van de varianties. Als we $Y - \hat{\beta}_0 - \hat{\beta}_1 x$ standaardiseren komen we op

$$(7.4.21) \quad \frac{Y - \hat{\beta}_0 - \hat{\beta}_1 x}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim N(0,1).$$

Het voorspellingsinterval voor Y bij de waarde x met betrouwbaarheid $\gamma = 1 - \alpha$ is gebaseerd op

$$(7.4.22) \quad \frac{Y - \hat{\beta}_0 - \hat{\beta}_1 x}{S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

(dit bewijzen we niet) en ziet er als volgt uit:

$$(7.4.23) \quad (\hat{\beta}_0 + \hat{\beta}_1 x - cS^*, \hat{\beta}_0 + \hat{\beta}_1 x + cS^*),$$

met c gegeven door $P(T_{n-2} \leq c) = 1 - \frac{1}{2}\alpha$ en

$$(7.4.24) \quad S^* = S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}.$$

Laten we dit interval ook toepassen in de context van voorbeeld 7.1.1. Stel we beschouwen een andere (grote) stad uit Engeland of Wales, we kennen de waarde van x , dit blijkt 80 te zijn, en we willen de waarde van Y voorspellen (of schatten zoals anderen zeggen) met behulp van een interval omdat we de waarde van Y (nog) niet kennen. We willen het nu dus niet hebben over een gemiddeld sterftcijfer voor mannen, nee we willen het nu hebben over het sterftcijfer voor mannen van deze ene stad. Voor het model betekent dit dat we het nu willen hebben over een waarde van Y en niet over een verwachting $E(Y) = \beta_0 + \beta_1 x$. De uitkomst van S^* is gelijk aan

$$(7.4.25) \quad 143.03 \sqrt{1 + \frac{1}{61} + \frac{(80 - 47.18)^2}{87069}} = 143.03 \times 1.014 = 145.07.$$

Het 95%-voorspellingsinterval voor Y is nu:

$$(7.4.26) \quad (1418.28 - 2.00 \times 145.07, 1418.28 + 2.00 \times 145.07) = (1128.1, 1708.4).$$

Dit interval is een stuk breder dan het corresponderende betrouwbaarheidsinterval voor $\beta_0 + \beta_1 x$.

7.5 Toetsingstheorie

In het algemeen probeert men (kans)modellen eenvoudig te houden. Uitgaande van een model van de meervoudige lineaire regressie met k verklarende variabelen is het een belangrijke vraag of we niet te ingewikkeld bezig zijn met de introductie van de k verklarende variabelen. We zouden ons kunnen afvragen of we de afhankelijke variabele ook goed kunnen beschrijven/voorspellen met 1 of meer verklarende variabelen minder. Voor elk van de verklarende variabelen kunnen we de nulhypothese toetsen of de bijbehorende regressiecoëfficiënt nul is.

Uitgaande van het model (7.3.1) beschouwen we nu het toetsen van

$$(7.5.1) \quad H_0 : \beta_i = 0 \text{ tegen } H_1 : \beta_i \neq 0,$$

voor een gegeven geheel getal i ($1 \leq i \leq k$). De toetsingsgrootte is als volgt:

$$(7.5.2) \quad T = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}.$$

Onder H_0 is T $t_{n-(k+1)}$ -verdeeld, vergelijk dit met (7.4.6). Als α de gekozen onbetrouwbaarheid (sdrempel) is verwerpen we H_0 als $T \leq -c$ of $T \geq c$, met c gegeven door $P(T_{n-(k+1)} \leq c) = 1 - \frac{1}{2}\alpha$.

Laten we teruggaan naar de context van voorbeeld 7.1.2 en voor elk van de twee verklarende variabelen (hardheid= x_1 en treksterkte= x_2) toetsen of ze wel in het regressiemodel thuishoren. We toetsen eerst

$$(7.5.3) \quad H_0 : \beta_1 = 0 \text{ tegen } H_1 : \beta_1 \neq 0.$$

De toetsingsgrootheid is $T = \hat{\beta}_1 / S_{\hat{\beta}_1}$ en we verwerpen H_0 als $T \leq -2.05$ of $T \geq 2.05$, werkend met een t_{27} -verdeling en 5% als onbetrouwbaarheidsdrempel kiezend. De uitkomst van T is hier $-6.571/0.583 = -11.27$. De uitkomst ligt in het kritiek gebied, dus verwerpen we H_0 . Dit betekent dat we “ $\beta_1 \neq 0$ ” bewezen achten: de verklarende variabele hardheid heeft betekenis voor het beschrijven/voorspellen van de afhankelijke variabele. Laten we nu ook

$$(7.5.4) \quad H_0 : \beta_2 = 0 \text{ tegen } H_1 : \beta_2 \neq 0$$

toetsen. De toetsingsgrootheid is nu $T = \hat{\beta}_2 / S_{\hat{\beta}_2}$ en we verwerpen H_0 als $T \leq -2.05$ of $T \geq 2.05$, opnieuw onbetrouwbaarheidsdrempel 5% kiezend. De uitkomst van T is $-1.374/0.194 = -7.1$: ook nu verwerpen we H_0 . Dit keer achten we “ $\beta_2 \neq 0$ ” bewezen: ook de verklarende variabele treksterkte is van belang voor het voorspellen van de afhankelijke variabele.

Eerder hadden we aan de hand van waarden voor R_a^2 al een sterke indicatie dat het zinvol is alle twee verklarende te gebruiken in plaats van alleen de eerste.

Alleen in het geval van enkelvoudige lineaire regressie en het toetsen van $H_0 : \beta_1 = 0$ tegen $H_1 : \beta_1 \neq 0$ hebben we een expliciete formule voor T : $T = \frac{\hat{\beta}_1}{S / \sqrt{S_{xx}}}$, met $\hat{\beta}_1$ geven door (7.2.12), S bepaald door (7.2.13) en $S_{xx} = \sum_i (x_i - \bar{x})^2$.

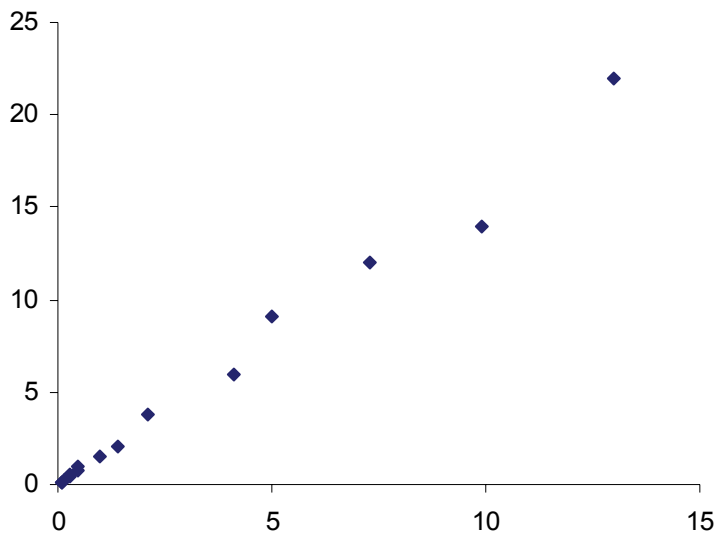
7.6 Transformaties, niet-lineaire verbanden

We bestuderen de volgende data set.

Voorbeeld 7.6.1 Van een bepaalde bacteriesoort kan de dichtheid in grond op twee manieren gemeten worden: (1) via de telplaat-methode, (2) via de immuno-fluorescentie-methode (IF). De laatstgenoemde methode werkt sneller en is goedkoper. De IF-methode telt systematisch te weinig bacteriën, maar is wellicht wel als voorspeller te gebruiken voor de uitkomsten die men met de telplaat-methode zou verkregen hebben. Om dit te onderzoeken is voor 15 grondmonsters beide methoden gebruikt. De waarnemingen ($E6$ betekent $\times 10^6$) staan in de volgende tabel. □

Telplaat	IF	telplaat	IF	telplaat	IF
5.9E6	4.1E6	1.5E6	9.8E5	1.1E5	7.1E4
4.1E5	2.6E5	7.8E5	4.5E5	1.4E7	9.9E6
2.1E6	1.4E6	5.1E5	2.6E5	1.2E7	7.3E6
1.4E5	9.8E4	9.1E6	5.0E6	2.1E5	1.3E5
3.8E6	2.1E6	9.4E5	4.9E5	2.2E7	1.3E7

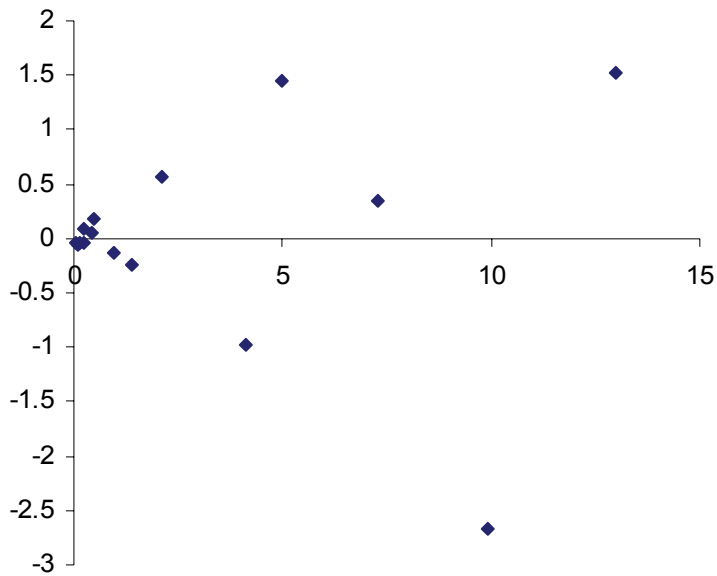
figuur 7.6.1: telplaatmeting versus IF-meting



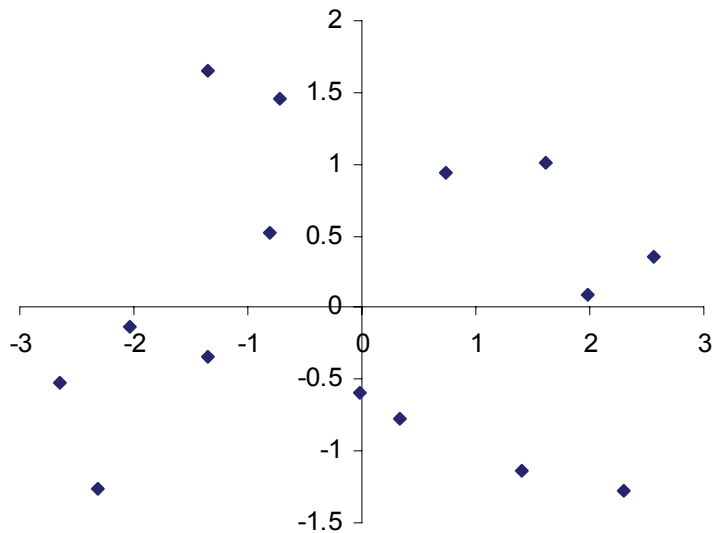
We willen enkelvoudige lineaire regressie toepassen, met de telplaat-meting als afhankelijke variabele y en de IF-meting als verklarende variabele x . In figuur 7.6.1 hebben we de waarden van de twee metingen tegen elkaar uitgezet. In figuur 7.6.2 zijn de residuen uitgezet tegen de waarden van de verklarende variabele, de IF-meting. We hebben daarbij $E6=10^6$ als nieuwe eenheid van x en y gekozen.

Wat opvalt in figuur 7.6.2 is dat de spreiding in de residuen niet constant is. Bij kleine waarden van de verklarende variabele is de spreiding van de residuen veel kleiner dan bij de grote waarden van de verklarende variabele. Daar de residuen schattingen van de uitkomsten van de storingen U_i zijn, moet de conclusie zijn dat de variantie van de U_i niet constant is. Een van de modelaannames is dat de variantie van de storingen wel constant is: $\text{var}(U_i) = \sigma^2$. In dit geval moeten we concluderen dat het model van de enkelvoudige lineaire regressie niet klopt. Hoe nu verder?

figuur 7.6.2: residu versus IF-meting



figuur 7.6.3: residu versus ln(IF-meting)



Wat men in zo'n situatie kan proberen is nagaan of het model van de enkelvoudige lineaire regressie wel klopt na transformatie van y en/of x . Veel gebruikte transformaties zijn $z \rightarrow \ln(z)$ en $z \rightarrow z^p$ (met p nog te kiezen, voor $p = \frac{1}{2}$ krijgen we transformatie $z \rightarrow \sqrt{z}$). Vaak blijkt $z \rightarrow \ln(z)$ een variantie stabiliserende transformatie in die zin dat de spreiding constant wordt na toepassing van deze transformatie. Dit blijkt

hier ook te werken. Als we enkelvoudige lineaire regressie toepassen met als afhankelijke variabele de (natuurlijke) logaritme van de telplaatmeting en als verklarende variabele de logaritme van de IF-meting dan is in het plotje van het plotje “residu versus x ”, zie figuur 7.6.3, niets meer te zien van de ongelijke spreiding in de residuen.

Als we een residuenplot bekijken ter controle van de modelaannames dan staat figuur 7.6.2 voor een van de patronen waarop men moet letten. Een ander patroon waar men alert op moet zijn is een “banaanvormige” puntenwolk, als regel duidt dit op een niet-lineair verband. Men kan proberen met het toevoegen van een kwadratische term het model te verbeteren. In geval van enkelvoudige lineaire regressie kan men proberen

$$(7.6.1) \quad Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + U_i$$

als modelvergelijking in plaats van (7.1.2). Men spreekt wel over kwadratische regressie. Een model als (7.6.1) kan gemakkelijk ingepast worden in meervoudige lineaire regressie. Model (7.6.1) kan opgevat worden als meervoudige lineaire regressie met twee verklarende variabelen. De eerst verklarende variabele heeft als waarden de getallen x_i en de tweede heeft als waarden de getallen x_i^2 (dus $x_{i1} = x_i$ en $x_{i2} = x_i^2$).

Alternatief voor modellen als (7.6.1) is ook weer het transformeren van de afhankelijke variabele en/of de verklarende variabelen. Men kan proberen een niet-lineair verband te transformeren naar een lineair verband. Het vinden van een geschikte transformatie is vaak een zaak van ‘trial and error’ en het lukt ook niet altijd. Soms ook heeft men wel een idee hoe het niet-lineaire verband eruit ziet en kan men deze wetenschap benutten om een lineair verband te krijgen zodat (meervoudige lineaire) regressie toepasbaar wordt. Als men bijvoorbeeld het volume hout van bomen wil voorspellen op grond van de verklarende variabelen diameter en hoogte, dan zou y zo ongeveer gelijk kunnen zijn aan $\text{constante} \times \text{diameter}^2 \times \text{hoogte}$: een niet-lineair verband dus. Echter $\ln(\text{volume})$ zou dan wel lineair afhangen van $\ln(\text{diameter})$ en $\ln(\text{hoogte})$, dit zouden we dan met meervoudige lineaire regressie kunnen onderzoeken.

7.7 ANOVA (1 factor), toetsingstheorie

In de voorgaande secties hebben we verschillende keren output van SPSS gepresenteerd. Een deel van de output wordt voorafgegaan door de naam ANOVA. ANOVA staat voor **A**nalysis of **v**ariance, variantie-analyse in het Nederlands. We willen het met name hebben over de grootheid F in dit deel van de output. In het ANOVA-deel staan de SSR en SSE getabelleerd. Uitgaande van het model met k verklarende variabelen staat SSR voor het deel van de spreiding dat we verklaren met deze k verklarende variabelen en SSE is het deel van de spreiding dat onverklaard blijft. De gelijkheid $S_{yy} = SSR + SSE$ geldt. De grootheid F is de verhouding van SSR en SSE gecorrigeerd naar de bijbehorende aantallen vrijheidsgraden,

$$(7.7.1) \quad F = \frac{SSR / k}{SSE / (n - (k + 1))}$$

en is de toetsingsgrootheid voor het toetsen van

$$(7.7.2) \quad H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

tegen

$$(7.7.3) \quad H_1 : \beta_i \neq 0 \text{ voor een zekere waarde } i .$$

Onder H_0 heeft F een $F_{n-(k+1)}^k$ -verdeling. We verwerpen H_0 als $F \geq c$, met c gegeven door $P(F_{n-(k+1)}^k > c) = 1 - \alpha$ als α de gekozen onbetrouwbaarheidsdrempel is. Teller en noemer van F zijn in output te vinden in de kolom Mean Square. Met deze toets vergelijk je eigenlijk het model met alle k verklarende variabelen met het model zonder (alle) verklarende variabelen. Meestal is dit niet zinvol: vaak gaat het erom een enkele verklarende variabele toe te voegen of te schrappen, toetsingstheorie van sectie 7.5 is dan toepasbaar.

In variantie-analyse met 1 factor, ofwel het k -steekproeven-probleem, is het bovenstaande toetsingsprobleem wel goed toepasbaar. In het k -steekproeven-probleem hebben we waarnemingen/metingen Y_1, Y_2, \dots, Y_n , deze zijn onderling onafhankelijk en als Y_i tot de steekproef j behoort dan is Y_i verdeeld volgens $N(\mu_j, \sigma^2)$. De waarnemingen/metingen zijn dus normaal verdeeld met een gemeenschappelijke variantie σ^2 en een verwachting die (mogelijkerwijs) van de steekproef afhangt. Om te onderzoeken of de steekproeven werkelijk wel verschillen is het interessant

$$(7.7.4) \quad H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

te toetsen tegen

$$(7.7.5) \quad H_1 : \mu_i \neq \mu_j \text{ voor zekere } i \text{ en } j .$$

Dit toetsingsprobleem kunnen we zodanig herschrijven dat we de relevante toetsingsgrootheid in de ANOVA-output kunnen vinden. We gaan het volgende voorbeeld gebruiken.

Voorbeeld 7.7.1 Als onderdeel van een groter onderzoek wordt de brandbaarheid van vier stoffen gemeten door een onderzoeksbureau. Voor elk van de stoffen wordt vijf keer de brandtijd van een japon gemeten (gemaakt van de betreffende stof) nadat de japon vlam gevat heeft via een label op de boord. De data zijn als volgt:

stof 1	stof 2	stof 3	stof 4
17.8	11.2	11.8	14.9
16.2	11.4	11.0	10.8
17.5	15.8	10.0	12.8
17.4	10.0	9.2	10.7
15.0	10.4	9.2	10.7

We willen weten of de vier stoffen werkelijk wel verschillen met betrekking tot de brandbaarheid. \square

Laten we uitgaan van het model van normale verdelingen met gelijke variantie voor alle metingen en een (mogelijkerwijs) verschillende verwachting voor elke stof. Dus voor de vijf metingen van stof j gaan we uit van een $N(\mu_j, \sigma^2)$ -verdeling ($j = 1, 2, 3, 4$). De vier verwachtingen $\mu_1, \mu_2, \mu_3, \mu_4$ en de variantie σ^2 zijn onbekend. We gaan

$$(7.7.6) \quad H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

toetsen tegen de alternatieve hypothese dat H_0 niet waar is ($H_1 : \mu_j \neq \mu_k$ voor zekere j en k). Laat Y_1, Y_2, \dots, Y_{20} de 20 metingen van voorbeeld 7.7.1 zijn. We gaan de metingen herschrijven als

$$(7.7.7) \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + U_i$$

met geschikt gekozen waarden x_{ij} ($i = 1, 2, \dots, 20; j = 1, 2, 3$) en onderling onafhankelijke storingsen U_1, U_2, \dots, U_{20} die $N(0, \sigma^2)$ -verdeeld zijn, zodat we theorie van de meervoudige lineaire regressie kunnen toepassen. Gelijkwaardig met (7.7.7) en de bijbehorende aannames is te zeggen dat Y_1, Y_2, \dots, Y_{20} onderling onafhankelijk zijn en verdeeld volgens:

$$(7.7.8) \quad Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \sigma^2).$$

We moeten de waarden x_{ij} zo kiezen dat geldt:

$$(7.7.9) \quad \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} = \mu_j \text{ als } Y_i \text{ een meting van stof } j \text{ is.}$$

Dit kan als volgt:

$$(7.7.10) \quad \begin{aligned} x_{i1} &= 1 \text{ als } Y_i \text{ een meting van stof 1 is, anders } x_{i1} = 0, \\ x_{i2} &= 1 \text{ als } Y_i \text{ een meting van stof 2 is, anders } x_{i2} = 0, \\ x_{i3} &= 1 \text{ als } Y_i \text{ een meting van stof 3 is, anders } x_{i3} = 0. \end{aligned}$$

Zo hebben we 3 verklarende variabelen gedefinieerd, deze duiden we verder aan met x_1, x_2 en x_3 . Variabelen als x_1, x_2 en x_3 worden wel indicatorvariabelen genoemd of dummy variabelen. We combineren (7.7.9) en (7.7.10), we vinden:

$$(7.7.11) \quad \begin{array}{ll} \mu_1 = \beta_0 + \beta_1 & \beta_0 = \mu_4 \\ \mu_2 = \beta_0 + \beta_2 & \text{en} \quad \beta_1 = \mu_1 - \mu_4 \\ \mu_3 = \beta_0 + \beta_3 & \beta_2 = \mu_2 - \mu_4 \\ \mu_4 = \beta_0 & \beta_3 = \mu_3 - \mu_4 \end{array}$$

We hebben de vier verwachtingen μ_1, μ_2, μ_3 en μ_4 herschreven in termen van regressiecoëfficiënten $\beta_0, \beta_1, \beta_2$ en β_3 . De nulhypothese (7.7.6) is gelijkwaardig met

$$(7.7.12) \quad H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$$

We kunnen dus de toetsingsgrootheid F gegeven door (7.7.1) toepassen. Voor de regressie-analyse hebben we de volgende “data matrix” ingevoerd.

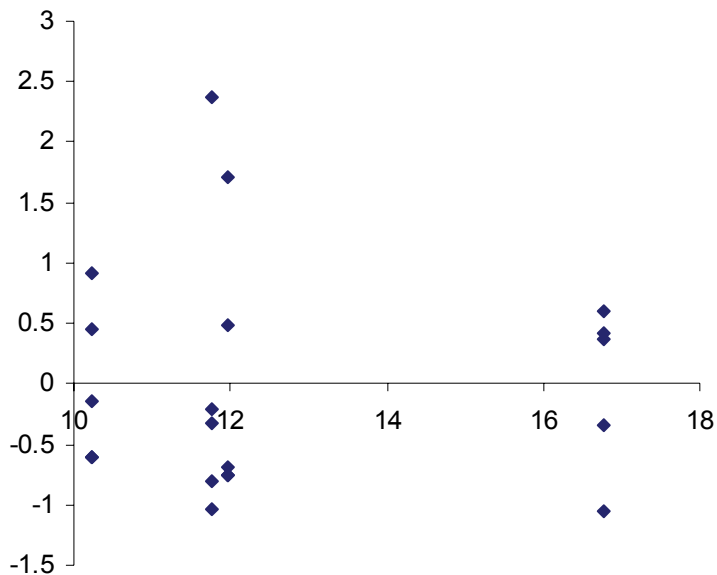
x_1	x_2	x_3	y
1	0	0	17.8
1	0	0	16.2
1	0	0	17.5
1	0	0	17.4
1	0	0	15.0
0	1	0	11.2
0	1	0	11.4
0	1	0	15.8
0	1	0	10.0
0	1	0	10.4
0	0	1	11.8
0	0	1	11.0
0	0	1	10.0
0	0	1	9.2
0	0	1	9.2
0	0	0	14.9
0	0	0	10.8
0	0	0	12.8
0	0	0	10.7
0	0	0	10.7

Het ANOVA-deel van de regressie-output is als volgt:

	SS	df	MS	F
Regression	120.498	3	40.166	13.892
Residual	46.260	16	2.891	
Total	166.758	19		

Met SS hebben we Sums of Squares afgekort, MS staat voor Mean Square. Bij het toetsen van de nulhypothese (7.7.12) tegen de alternatieve hypothese $H_1 : \beta_i \neq 0$ voor zekere i , verwerpen we H_0 als $F \geq 3.22$ (F_{16}^3 -verdeling, interpolatie). Aangezien 13.892 de uitkomst is van de toetsingsgrootheid F verwerpen we H_0 en concluderen dat (tenminste twee van de) vier stoffen verschillen in brandtijd. Voor de volledigheid hebben we het gestandaardiseerde residu geplott tegen de predicted value, het vertoont redelijk een chaotisch beeld.

residu versus fitted value (vb 7.7.1)_



We hebben in deze sectie in feite laten zien hoe je kwalitatieve variabelen kan representeren met dummy variabelen als verklarende variabelen. Uiteraard kan dit soort verklarende variabelen gecombineerd worden met “continue” verklarende variabelen in een regressiemodel.

We hebben hier maar een glimp laten zien van ANOVA, variantie-analyse. We hebben ons beperkt tot “1 factor”, de vier steekproeven verschilden in de factor “stof”. We verwijzen naar de literatuur voor hoe je moet werken met 2 of meer factoren en voor de eigen notatie van de variantie-analyse (deze notatie hebben we hier genegeerd).

7.8 Opgaven

Opgave 1

We beschikken over de volgende gegevens over het jaarlijkse inkomen (x) en de uitgaven (y) voor voeding van 16 gezinnen. De bedragen zijn uitgedrukt in duizenden guldens.

Gezin	inkomen (x)	uitgaven (y)	gezin	inkomen (x)	uitgaven (y)
1	55	25	9	57	28
2	58	30	10	47	24
3	56	24	11	58	27
4	61	31	12	50	23
5	58	27	13	54	26
6	61	29	14	64	28
7	56	24	15	59	26
8	50	21	16	52	23

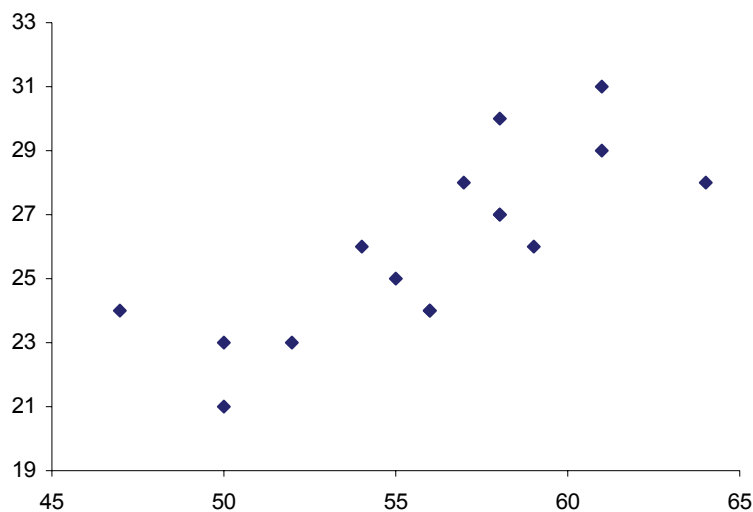
ANOVA

	Sum of Squares	df	Mean Square	F
Regression	71.616	1	71.616	22.590
Residual	44.384	14	3.170	
Total	116.000	15		

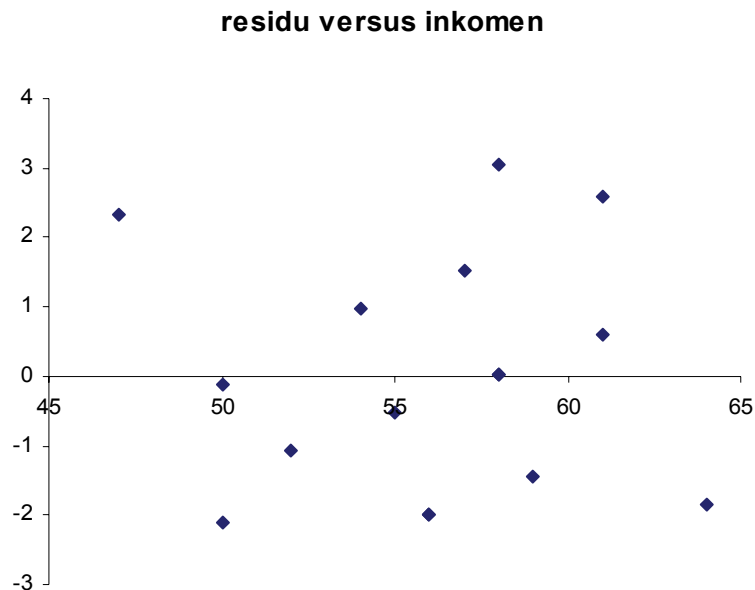
Coefficients

	B	Std. Error	t
(Constant)	-.916	5.681	-.161
INKOMEN	.481	.101	4.753

uitgaven versus inkomen

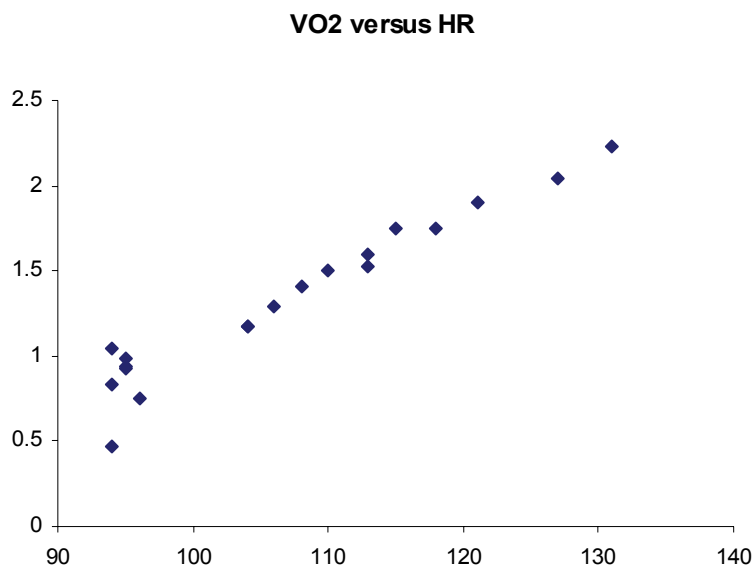


- Formuleer het model van de enkelvoudige lineaire regressie, in de context van deze opgave.
- Geef de schattingen van de parameters β_0, β_1 en σ^2 .
- Bepaal de schatting van de toename van de uitgaven voor voeding corresponderende met een vermeerdering van het inkomen met 100 gulden.
- De gulden is inmiddels vervangen door de euro. We nemen nu als nieuwe eenheid 1000 euro voor zowel inkomen als uitgaven, in plaats van 1000 gulden. Geef bij deze nieuwe eenheid opnieuw schattingen voor de parameters β_0, β_1 en σ^2 . (Gebruik: 1 euro is 2.20371 gulden.)
- Beschrijf wat we bij enkelvoudige lineaire regressie verstaan onder residuen.
- Beoordeel de plot van het residu (verticaal) uitgezet tegen de verklarende variabele (inkomen), zie volgend figuur.

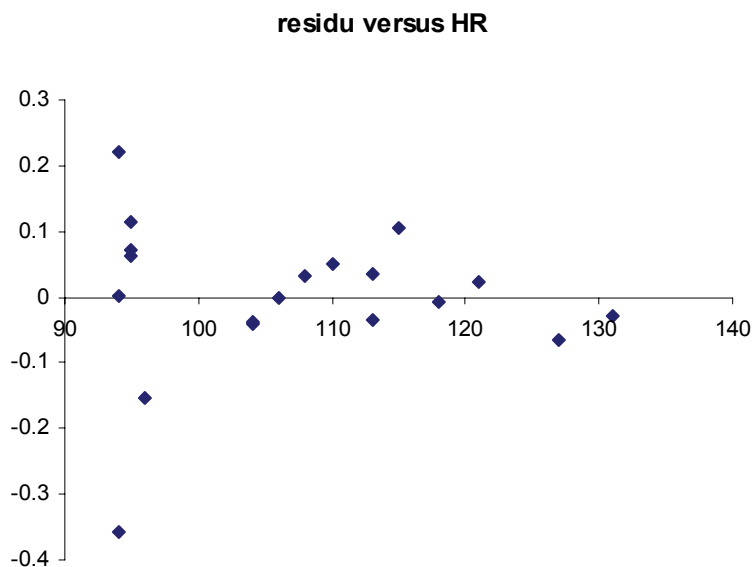


Opgave 2

Het menselijk lichaam verbruikt bij inspanning meer zuurstof dan in ruststand. Om de spieren van meer zuurstof te voorzien, moet het hart sneller kloppen. De hartslag is gemakkelijk te meten, maar het meten van de hoeveelheid opgenomen zuurstof vereist ingewikkelde apparatuur. Als de zuurstofopname (VO₂) nauwkeurig kan worden voorspeld uit de hartslag (HR, 'heart rate'), kunnen bij onderzoek de voorspelde waarden de feitelijk gemeten waarden vervangen. Helaas zijn niet alle menselijke lichamen gelijk, daarom is er niet één enkele voorspellingsvergelijking die voor alle mensen geldig is. Onderzoekers kunnen echter voor één persoon zowel HR als VO₂ meten bij variërende inspanningsniveaus en proberen de samenhang tussen HR en VO₂ te modelleren om uiteindelijk tot voorspellingen te komen voor VO₂ bij gegeven waarden van HR. Hier volgt de plot van de gegevens voor één persoon. (Data verstrekt door Paul Wadsmith, uit experimenten verricht in het laboratorium van Don Corrigan, Purdue University.)



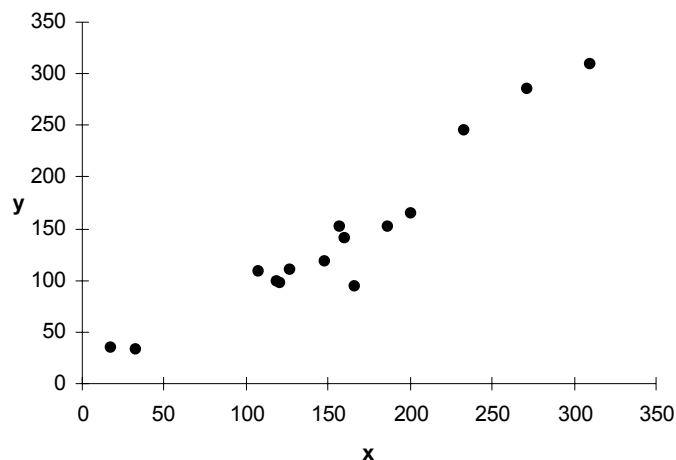
De plot van de residuen (van enkelvoudige lineaire regressie) en de waarden van HR zijn als volgt.



Beoordeel aan de hand van deze plots of we enkelvoudige lineaire regressie kunnen toepassen.

Opgave 3

Gedurende een jaar werd in 15 Duitse Gaststätten de pilsomzet gemeten (in hl). Op grond van de herfst + winter omzet (x) wil men de voorjaar + zomer omzet (y) voorspellen. Een plot van de data is als volgt.



Output van enkelvoudige lineaire regressie:

ANOVA				
	Sum of Squares	df	Mean Square	F
Regression	84134.94	1	84134.94	142.59
Residual	7670.40	13	590.03	
Total	91805.34	14		

Coefficients			
	B	Std. Error	T
(Constant)	-11.489	---	---
INKOMEN	0.98737	0.08269	11.94

Verdere gegevens:

steekproefgemiddelde van x : 156.88

steekproefstandaardafwijking van x : 78.51

- Geef het kansmodel van de klassieke enkelvoudige lineaire regressie voor deze situatie.
- Bereken de aangepaste vergelijking.
- Bereken de fractie verklaarde variantie en geef commentaar op het resultaat.
- Bereken de voorspelling van de voorjaar + zomer omzet bij een herfst + winter omzet van 140 hl.
- Bereken het 95%-voorspellingsinterval voor de voorjaar + zomer omzet bij een herfst + winter omzet van 140 hl.
- Bereken het 95%-betrouwbaarheidsinterval voor de verwachte voorjaar + zomer omzet bij een herfst + winter omzet van 140 hl.
- Bij een zekere Gaststätte is de herfst + winter omzet 140 hl. Om er vrij zeker van te zijn ("95%-betrouwbaarheid") voldoende bier in voorjaar + zomer te hebben, wordt voor deze periode 140.62 hl besteld. Zal deze hoeveelheid inderdaad met een kans van 95% voldoende zijn? Argumenteer uw antwoord.

Opgave 4

We willen het gewichtsverlies Y van een zekere chemische stof in verband brengen met de tijd x_1 (in uren) dat de stof blootgesteld is aan de lucht en met de relatieve vochtigheid x_2 gedurende die tijd. We beschikken over gegevens van het gewichtsverlies in 12 verschillende situaties. Als model wordt gehanteerd:

$$(1) \quad Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + U_i \quad (i = 1, 2, \dots, 12).$$

De parameters β_0, β_1 en β_2 zijn onbekend en de storingstermen U_1, U_2, \dots, U_{12} zijn onderling onafhankelijk en $N(0, \sigma^2)$ -verdeeld (σ^2 is ook onbekend). De computer-output is als volgt:

multiple R	0.9791
R square	0.9586
Std.error	0.3866

Analysis of variance				
	df	SS	MS	F
regression	2	31.1242	15.5621	104.1329
residual	9	1.3450	0.1494	

Variables in the equation			
variable	B	std. error B	T
constant	0.6667		
X1	1.3167	0.0998	13.1911
X2	-8.0000	1.3668	-5.8532

- Bereken de voorspelling van het gewichtsverlies als de stof 6.5 uur blootgesteld is aan de lucht bij een relatieve vochtigheid van 0.35 .
- Geef de schatting voor σ^2 .
- Bereken het 95%-betrouwbaarheidsinterval voor β_2 .
- Iemand beweert dat de gegevens erop wijzen dat het beter is ook een term $\beta_3 x_1 x_2$ in de regressievergelijking op te nemen, omdat dan de fractie verklaarde variantie groter wordt. Leg uit waarom dit geen goed argument is. Met welke grootte zou je beter de beide modellen kunnen vergelijken? Bereken de waarde van deze grootte voor het model gegeven door (1).

Opgave 5

Uit een onderzoek naar twee methoden voor het meten van de bloedstroming in de hondenmaag komen de volgende gegevens.

bolletjes (x)	ader (y)	bolletjes (x)	ader (y)
4.0	3.3	15.9	16.4
4.7	8.3	17.4	15.4
6.3	4.5	18.1	17.6
8.2	9.3	20.2	21.0
12.0	10.7	23.9	21.7

‘Bolletjes’ is een experimentele methode waarvan de onderzoekers hopen dat hij ‘ader’, de moeilijke standaardmethode, goed zal voorspellen. Een vooronderzoek van de gegevens brengt geen redenen aan het licht om te twijfelen aan de geldigheid van het model van enkelvoudige lineaire regressie, met ‘ader’ als afhankelijke variabele y en ‘bolletjes’ als verklarende variabele x .

ANOVA

	Sum of Squares	df	Mean Square	F
Regression	360.570	1	360.570	116.849
Residual	24.686	8	3.086	
Total	385.256	9		

a Predictors: (Constant), BOLLETJE

b Dependent Variable: ADER

Coefficients

	Unstandardized Coefficients		t
	B	Std. Error	
(Constant)	1.031	1.224	.843
BOLLETJE	.902	.083	10.810

a Dependent Variable: ADER

Steekproefgemiddelde van ‘bolletjes’ is 13.07, de steekproefvariantie van ‘bolletjes’ is 49.245.

- Formuleer voor deze data het model van de enkelvoudige lineaire regressie.
- Bereken de aangepaste vergelijking.
- Veronderstel dat we voor een hond een ‘bolletjes’-waarde waarnemen die gelijk is aan 15.0. Geef een interval dat de meting van ‘ader’, de moeilijke standaardmethode, omvat met kans 95%.
- Bereken het residu voor de waarneming $(x, y) = (4.7, 8.3)$.

Opgave 6

Van een (aselecte) steekproef van 50 maatschappelijke werkers hebben we de volgende gegevens: het salaris (eenheid: dollar) en het aantal jaren ervaring. We passen het model toe van (enkelvoudige) lineaire regressie met als afhankelijke variabele y de (natuurlijke) logaritme van het salaris en als verklarende variabele het aantal jaren ervaring. Een plot van de waarden van x (horizontaal) en y (verticaal) is te zien op de volgende bladzijde.

Met Excel is de volgende computer output verkregen.

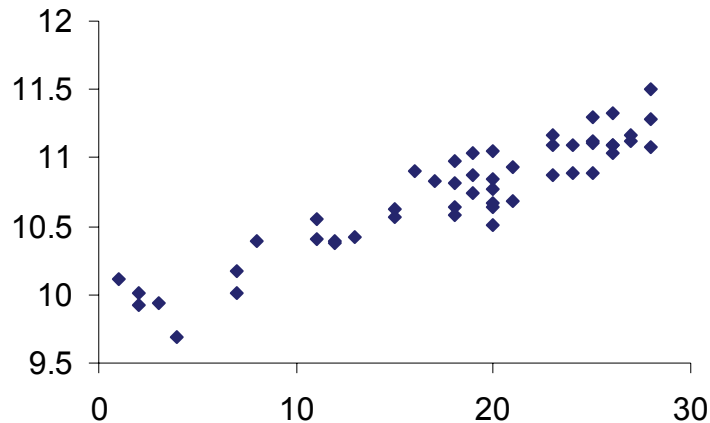
Analysis of variance

	df	SS	MS	F
Regression	1	7.212	7.212	303.6
Residual	48	1.140	0.024	
Total	49	8.352		

Variables in the equation

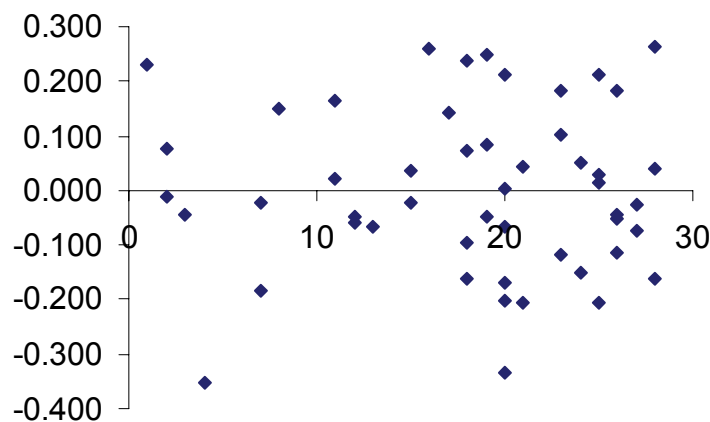
	B	SE B	T
constant	9.841	0.056	174.6
x	0.050	0.003	17.43

y versus x



- Formuleer het model van de enkelvoudige lineaire regressie, toegespitst op de data van deze opgave.
- Bereken het 99%-betrouwbaarheidsinterval voor β_1 .
- Geef de (gebruikelijke) schatting voor σ^2 .
- Bereken de waarde van R^2 . Geef commentaar op het resultaat.
- Om te beoordelen of het model van de (enkelvoudige) lineaire regressie goed past bij de data, kunnen de residuen (verticaal) uitgezet worden tegen de waarden van x . Hieronder staat de plot voor de data van deze opgave. Wat is je oordeel ten aanzien van deze residuenplot?

residu versus x



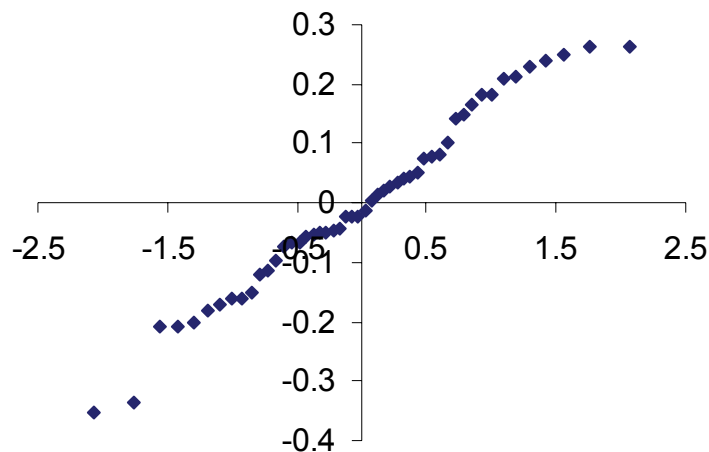
Opgave 7 (Dit is een opgave van Statistische Technieken, dus geen tentamenstof.)

De 50 waarden van de residuen van het model van de vorige opgave zijn als volgt (de waarden zijn al gerangschikt van klein naar groot).

-0.354	-0.335	-0.207	-0.207	-0.202	-0.183	-0.170	-0.162
-0.160	-0.151	-0.120	-0.112	-0.096	-0.074	-0.066	-0.066
-0.058	-0.053	-0.050	-0.049	-0.046	-0.044	-0.025	-0.023
-0.022	-0.012	0.005	0.015	0.021	0.029	0.034	0.041
0.043	0.052	0.074	0.078	0.083	0.103	0.141	0.150
0.164	0.182	0.183	0.211	0.211	0.229	0.239	0.249
0.261	0.264						

- Maak van deze 50 waarden een boxplot (inclusief eventuele uitschieters).
- Zijn de residuen (bij benadering) normaal verdeeld? Beantwoord deze vraag door de volgende normale Q-Q plot te beoordelen. Betrek ook de boxplot in je antwoord.

normale Q-Q plot



Opgave 8

Een projectontwikkelaar is geïnteresseerd in een model waarmee de verkoopprijs geschat kan worden van kavels aan het strand van de kust van Oregon. Voor 20 recent verkochte kavels werden de volgende gegevens verkregen:

- y : verkoopprijs van de kavel (in duizenden dollars)
- x_1 : oppervlakte van de kavel (in honderden vierkante meters)
- x_2 : hoogte van de kavel
- x_3 : glooiing van de kavel

De projectontwikkelaar paste multiple lineaire regressie toe en een deel van de computeruitdraai is als volgt.

Analysis of Variance

	DF	Sum of Squares	Mean Square	F Ratio
Regression	?	21.409	?	?
Residual	?	5.903	?	

Variable	Coefficient	Std. Error	F-value
(Constant)	-2.491		
Oppervlakte	0.099	0.058	2.935
Hoogte	0.029	0.006	23.327
Glooiing	0.086	0.031	7.705

- Vul de waarden in op de plaatsen van de vraagtekens.
- Geef de aangepaste vergelijking.
- Onderzoek met behulp van geschikte statistische toetsen voor elk van de 3 verklarende variabelen of deze statistisch significant is in aanwezigheid van de overige 2. Neem in alle 3 gevallen als onbetrouwbaarheidsdrempel 5%.
- Veronderstel dat van tevoren (dat wil zeggen voor dat de gegevens bekeken zijn) het idee bestaat dat hoe meer glooiing des te aantrekkelijker de kavel is. Onderzoek met behulp van een geschikte statistische toets of de verkoopprijzen stijgen bij een grotere helling in aanwezigheid van de andere variabelen. Neem als onbetrouwbaarheidsdrempel 5%. Waarom moet een dergelijke te toetsen stelling **van tevoren** geformuleerd worden en niet op grond van de data?
- Bereken het 90%-betrouwbaarheidsinterval voor de regressieparameter die de verkoopprijs relateert aan de oppervlakte in aanwezigheid van hoogte en glooiing. Geef een interpretatie van het betrouwbaarheidsinterval bij gegeven hoogte en glooiing.

Opgave 9

Een manager van een elektronicafabriek voerde een studie uit naar het verband tussen de omvang van de productie en het aantal defecte artikelen bij de productie. De omvang van de productie wordt weergegeven in honderden eenheden en wordt genoteerd met x . Het aantal defecte artikelen noemen we Y . De gegevens hebben betrekking op 12 aselect gekozen productiedagen. Toepassing van kwadratische regressie leverde o.a. de volgende output op:

Variable	Coef	Std.error	t-ratio	p
Constant	7.296	5.066	1.44	0.184
x	2.153	1.854	1.16	0.275
x^2	0.1043	0.1484	0.70	0.500

$s = 2.817$ $R - sq = 93\%$ $R - sq(adj) = 91.4\%$

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	949.23	474.62	59.80	0.000
Error	9	71.43	7.94		
	11	1020.67			

- Geef de aangepaste vergelijking
- Voorspel het aantal defecte artikelen als er 500 eenheden worden geproduceerd.
- Uit de t -waarden blijkt dat de coëfficiënten van x en x^2 niet statistisch significant van 0 verschillen bij de gebruikelijke onbetrouwbaarheidsdrempels. Betekent dit dat de termen met x en x^2 beide uit de regressievergelijking weggelaten kunnen worden? Motiveer je antwoord.
- Toepassing van enkelvoudige regressie van Y op x levert een waarde van $R-sq(adj)$ op van 91.9%. Men kiest op grond daarvan voor enkelvoudige lineaire regressie en neemt de in onderdeel a gevonden vergelijking met weglating van de x^2 -term. Geef commentaar op deze handelwijze.

Opgave 10

Men heeft de opwarmtijden (in seconden) van drie typen apparaten gemeten. Voor elk type heeft men 15 metingen, de data zijn als volgt:

	type A					type B					type C				
data	19	23	26	18	20	20	20	32	27	40	16	26	15	18	19
	20	18	35	27	31	24	22	18	24	25	17	19	18	14	18
	25	22	23	27	29	29	31	24	25	32	19	21	17	19	18
steekproef-gemiddelde	24.200					26.200					18.267				
steekproef-variantie	25.171					33.457					7.638				

Om uit te maken of er een systematisch verschil in opwarmtijd bestaat tussen de 3 typen apparaten hebben we meervoudige lineaire regressie toegepast met de opwarmtijd als afhankelijke variabele en met 2 verklarende variabelen x_1 en x_2 die als volgt gedefinieerd zijn:

$x_1 = 1$ als apparaat van type A is, anders $x_1 = 0$;

$x_2 = 1$ als apparaat van type B is, anders $x_2 = 0$.

Deze verklarende variabelen zullen we met “type A” en “type B” aanduiden.

Een deel van de computer output is als volgt.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	510.711	2	255.356	11.560	.000
Residual	927.733	42	22.089		
Total	1438.444	44			

a Predictors: (Constant), TYPE_B, TYPE_A

b Dependent Variable: TIJD

Coefficients

	Unstandardized Coefficients		t	Sig.
	B	Std. Error		
(Constant)	18.267	1.214	15.053	.000
TYPE_A	5.933	1.716	3.457	.001
TYPE_B	7.933	1.716	4.623	.000

a Dependent Variable: TIJD

- Formuleer het model van de meervoudige lineaire regressie, toegespitst op de data van deze opgave.
- Geef de aangepaste vergelijking.
- Bepaal de schatting van σ^2 .
- Is er een systematisch verschil tussen de 3 typen apparaten? Voer een toets uit om deze vraag te beantwoorden. Neem onbetrouwbaarheidsdrempel 1%.

Opgave 11

Een besteldienst baseert de prijs voor het vervoer op het gewicht van het te vervoeren pakket en de afstand. Voor een juiste prijsstelling wordt een onderzoek uitgevoerd naar het verband tussen de kosten y (in dollars), het gewicht x_1 (in ponden) en de te overbruggen afstand x_2 (in mijlen). Voor 20 aselekt gekozen pakketten zijn de gegevens geregistreerd. Als regressievergelijking neemt men

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 .$$

De computeruitdraai geeft o.a.

ANOVA

	DF	SS	MS	F
Regression	3	445.45	148.48	358
Residual	16	6.63	0.41	

Variable	Coefficient	Std. error
Intercept	-0.1405	0.6481
x_1	0.0191	0.1582
x_2	0.0077	0.0039
$x_1 \times x_2$	0.0078	0.0009

- Geef de aangepaste vergelijking.
- Bereken de schatting van de verwachte toename van de kosten per extra mijl voor een pakket van 2 pond.
- Ga door middel van een geschikte toets na of de interactieterm $\beta_3 x_1 x_2$ een relevante bijdrage geeft bij het voorspellen van y . Neem als onbetrouwbaarheidsdrempel 5%.
- Bereken de waarde van R_a^2 en geef commentaar op het resultaat.

7.9 Uitwerkingen van opgaven

Opgave 7.1

7.1a

Model: We hebben zestien paren getallen (x_i, y_i) ($i = 1, 2, \dots, 16$). De waarden y_i vatten we op als uitkomsten van onderling onafhankelijke stochastische variabelen

$$Y_i = \beta_0 + \beta_1 x_i + U_i$$

waar de U_i onderling onafhankelijke storingen zijn, verdeeld volgens $N(0, \sigma^2)$. Y_i staat hier voor de uitgaven voor voeding van gezin i en x_i voor het inkomen van gezin i .

7.1b

Uit de output halen we de volgende schattingen:

$$\hat{\beta}_0 = -0.916 \text{ en } \hat{\beta}_1 = 0.481 .$$

Schatting van σ^2 : 'residual mean square'=3.170

7.1c

Eenheid is duizend gulden. De parameter β_1 geeft aan hoeveel de uitgaven (gemiddeld) stijgen als het inkomen toeneemt met 1 eenheid. Als inkomen stijgt met 100 gulden, dan is dit een stijging van 0.1 en nemen de uitgaven (gemiddeld) toe met $0.1 \times \beta_1$ maal duizend gulden. Schatting hiervan is $0.1 \times 0.481 \times 1000 = 48.1$ gulden.

7.1d

Omzetten naar euro's houdt de volgende vervangingen in:

$$x_i \rightarrow x_i / 2.20371 \text{ en } y_i \rightarrow y_i / 2.20371 .$$

We moeten onderzoeken hoe de uitkomsten van de schatters veranderen. We beginnen met

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} .$$

Het is, denk ik, redelijk snel in te zien dat de uitkomst van $\hat{\beta}_1$ niet verandert. De uitkomst van $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ verandert wel, de nieuwe uitkomst hiervan wordt $-0.916 / 2.20371 = -0.416$, want \bar{x} en \bar{Y} worden beide een factor 2.20371 kleiner. Bekijken we tenslotte de schatter van σ^2 :

$$S^2 = \frac{\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n - 2} .$$

Elke term $Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ wordt een factor 2.20371 kleiner en derhalve wordt de nieuwe uitkomst van S^2 : $\frac{3.170}{(2.20371)^2} = 0.6528$.

7.1e

Onder de residuen verstaan we (de uitkomsten van): $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

Met de residuen schatten we (de uitkomsten van) de storingen U_i .

7.1f

De residuenplot laat een chaotische puntenwolk zien zonder enig patroon. Dit is precies datgene wat we willen zien: dit spoort met het model van de (enkelvoudige) lineaire regressie. Let wel: de residuen zijn schattingen van storingen die onafhankelijk geacht worden te zijn en gelijk verdeeld. Dit houdt in dat geen enkele relatie tussen de residuen zichtbaar moet zijn.

Opgave 7.2

De residuenplot moet beoordeeld op dezelfde manier als bij opgave 7.1f. Nu is er op zijn minst twijfel over de correctheid van het model. De spreiding in de residuen lijkt links groter dan rechts. In het model van de (enkelvoudige) lineaire regressie nemen we aan dat de variantie (maat voor spreiding) van de storingen U_i constant is, altijd gelijk is aan σ^2 (deze parameter is onbekend). De constante variantie van de storingen lijkt niet te kloppen.

Opgave 7.3

7.3a

Als y_i de voorjaar+zomer omzet van Gaststätte i is, dan vatten we y_i op als uitkomst van een stochastische variabele $Y_i = \beta_0 + \beta_1 x_i + U_i$, met x_i de bijhorende herfst+winter omzet en met onderling onafhankelijke “storingen” U_i die $N(0, \sigma^2)$ -verdeeld zijn.

De parameters β_0, β_1 en σ^2 zijn onbekend.

7.3b

Van een berekening is niet echt sprake. De schattingen $\hat{\beta}_0$ en $\hat{\beta}_1$ kunnen we uit de output halen: ze zijn respectievelijk -11.489 en 0.98737 . Daarmee wordt de aangepaste vergelijking: $y = -11.489 + 0.98737x$.

7.3c

De fractie verklaarde variantie $R^2: \frac{84134.94}{91805.34} = 0.916 = 91.6\%$

Omdat de fractie verklaarde variantie dichtbij 1 is, voldoet het regressiemodel goed in die zin dat we de variabele y kennelijk goed kunnen voorspellen.

7.3d

Voorspelling: $\hat{y} = -11.489 + 0.98737 \times 140 = 126.7$

7.3e

95%-voorspellingsinterval: $(\hat{\beta}_0 + \hat{\beta}_1 x - cS^*, \hat{\beta}_0 + \hat{\beta}_1 x + cS^*)$

Verdere uitwerking:

$$\hat{\beta}_0 + \hat{\beta}_1 x = -11.489 + 0.98737 \times 140 = 126.7$$

$c = 2.16$ (t_{13} -verdeling)

$$\text{Nu het moeilijkste: } S^* = S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

De residual Mean Square geeft uitkomst van $S^2: 590.03$.

Als we $S_{xx} = \sum_i (x_i - \bar{x})^2$ delen door $n-1$ dan krijgen we de steekproefvariantie van de waarden x_i , omgekeerd is S_{xx} $n-1$ maal de steekproefvariantie van x :

$$S_{xx} = 14 \times (78.51)^2 = 86293.5$$

$$\text{Uitkomst van } S^* : \sqrt{590.03} \times \sqrt{1 + \frac{1}{15} + \frac{(140 - 156.88)^2}{86293.5}} = 24.291 \times 1.0344 = 25.13$$

$$95\text{-voorspellingsinterval: } (126.7 - 2.16 \times 25.13, 126.7 + 2.16 \times 25.13) = (72.4, 181.0)$$

7.3f

Gevraagd 95%-BI voor verwachting voor Y als $x = 140$, dus 95%-BI voor $\beta_0 + \beta_1 x$, met $x = 140$:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x - cS \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}, \hat{\beta}_0 + \hat{\beta}_1 x + cS \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right)$$

met $c = 2.13$ (t_{13} -verdeling).

$$\text{Uitkomst } S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} : 24.291 \times \sqrt{\frac{1}{15} + \frac{(140 - 156.88)^2}{86293.5}} = 24.291 \times 0.26452 = 6.425$$

$$95\text{-BI voor } \beta_0 + \beta_1 x : (126.7 - 2.16 \times 6.425, 126.7 + 2.16 \times 6.425) = (112.8, 140.6)$$

7.3g

De waarde 140.62 zit aan de bovenkant van het BI. We moeten echter de waarde niet vergelijken met het BI. Het gaat om een toekomstige waarde van y en het **voorspellingsinterval** geeft hiervoor waarden. De waarde 140.62 zit behoorlijk in het midden van het voorspellingsinterval, we kunnen er dus niet (95%) zeker van zijn dat 140.62 hl bier genoeg is. Het 95%-voorspellingsinterval geeft waarden tot ca 180.

Opgave 7.4

7.4a

Aangepaste vergelijking: $y = 0.6667 + 1.3167x_1 - 8.0000x_2$, invullen van $x_1 = 6.5$ en $x_2 = 0.35$ geeft voorspelling $\hat{y} = 6.425$.

7.4b

De schatting voor σ^2 is de residual MS: 0.1494

7.4c

$$\text{BI voor } \beta_2 : \left(\hat{\beta}_2 - cS_{\hat{\beta}_2}, \hat{\beta}_2 + cS_{\hat{\beta}_2} \right)$$

met $c = 2.26$ (t_9 -verdeling),

De schatting $\hat{\beta}_2$ en bijbehorende standaardfout (standard error) $S_{\hat{\beta}_2}$ zijn in de output te vinden, invullen levert BI:

$$\begin{aligned} & (-8.0000 - 2.26 * 1.3668, -8.0000 + 2.26 * 1.3668) \\ & = (-8.0000 - 3.0890, -8.0000 + 3.0890) = (-11.09, -4.91) \end{aligned}$$

7.4d

Omdat de SSE altijd daalt bij het toevoegen van verklarende variabelen, wordt

$R^2 = 1 - \frac{SSE}{S_{YY}}$ **altijd** groter. De fractie verklaarde variantie kan ook dalen bij toevoeging

van verklarende variabelen, als we een correctie doorvoeren naar k , het aantal verklarende variabelen. Zo komen we op

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)} * \frac{SSE}{S_{YY}},$$

deze grootheid kunnen we beter gebruiken om beide modellen te vergelijken. Voor model (1) krijgen we:

$$R_a^2 = 1 - \frac{11}{9} * \frac{1.3450}{31.1242 + 1.3450} = 1 - \frac{11}{9} * 0.04142 = 0.949.$$

Opgave 7.5

7.5a

Laat de $(x_1, y_1), \dots, (x_{10}, y_{10})$ de data vertegenwoordigen met y_i ($i = 1, \dots, 10$) een 'ader'-meting en x_i de bijbehorende 'bolletjes'-meting. De y_i 's vatten we als uitkomsten van stochastische variabelen Y_i waarvoor geldt:

$$Y_i = \beta_0 + \beta_1 x_i + U_i,$$

met U_1, \dots, U_{10} onderling onafhankelijk en verdeeld volgens een $N(0, \sigma^2)$ -verdeling;

β_0, β_1 en σ^2 zijn onbekende parameters.

7.5b

Aangepaste vergelijking: $y = \hat{\beta}_0 + \hat{\beta}_1 x$, dus $y = 1.031 + 0.902x$.

7.5c

Gevraagd: 95%-voorspellingsinterval voor Y bij de waarde $x = 15.0$:

$$(\hat{\beta}_0 + \hat{\beta}_1 x - cS^*, \hat{\beta}_0 + \hat{\beta}_1 x + cS^*)$$

met:

$$\hat{\beta}_0 + \hat{\beta}_1 x = 1.031 + 0.902 * 15.0 = 14.56$$

$$c = 2.31 \text{ (uit } t_8 \text{-verdeling)}$$

$$S^* = S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

Uitkomst van S^2 : residual mean square 3.086

$$S_{xx}: n-1 \text{ maal steekproefvariantie van } x: 9 \times 49.245 = 443.21$$

$$\text{Uitkomst van } S^* \text{ zo: } \sqrt{3.086} \times \sqrt{1 + \frac{1}{10} + \frac{(15.0 - 13.07)^2}{443.21}} = 1.757 \times 1.053 = 1.851$$

Invullen levert voorspellingsinterval:

$$(14.56 - 2.31 \times 1.851, 14.56 + 2.31 \times 1.851) = (14.56 - 4.28, 14.56 + 4.28) = (10.28, 18.84)$$

7.5d

$$\text{Residu is } y - \hat{\beta}_0 - \hat{\beta}_1 x = 8.3 - 1.031 - 0.902 * 4.7 = 3.03$$

Opgave 7.6

7.6a

Laat (x_i, y_i) ($i = 1, 2, \dots, 50$) de waarden van (x, y) zijn, x_i is het aantal jaren ervaring van maatschappelijk werker i en y_i is de bijbehorende logaritme van het salaris. We beschouwen de y_i als uitkomsten van stochastische variabelen $Y_i = \beta_0 + \beta_1 x_i + U_i$ met storingen U_i die onderling onafhankelijk zijn en $N(0, \sigma^2)$ -verdeeld (β_0, β_1 en σ^2 zijn onbekend).

7.6b

$$99\% \text{-BI voor } \beta_1: (\hat{\beta}_1 - cS_{\hat{\beta}_1}, \hat{\beta}_1 + cS_{\hat{\beta}_1})$$

met $S_{\hat{\beta}_1}$ de standaardfout (standard error) van $\hat{\beta}_1$.

We passen in feite theorie van multiple regressie toe.

Uitkomsten (uit output): 0.050 voor $\hat{\beta}_1$ en 0.003 voor $S_{\hat{\beta}_1}$, $c = 2.68$ (t_{48} -verdeling)

$$\Rightarrow 99\% \text{-BI voor } \beta_1: (0.050 - 2.68 * 0.003, 0.050 + 2.68 * 0.003) = (0.042, 0.058)$$

7.6c

schatting voor σ^2 : residual MS danwel MS(error), dus 0.024

7.6d

$$R^2 = 1 - \frac{1.140}{8.352} = 86\%$$

Dus 86% van de spreiding is verklaard. Dit is wel een goede waarde, echter vaak streven we naar een hogere waarde.

7.6e

Bij de residuenplot speuren we naar patronen. Zodra we een patroon zien is het "mis", want dit geeft aan dat het model verbeterd kan worden. Patronen zijn bijvoorbeeld een gekromde puntenwolk en een kegelachtige puntenwolk. Bij de getoonde residuenplot is geen patroon te zien, we zien alleen maar chaos en dit is OK.

Opgave 7.7

7.7a

$$\text{steekproefmediaan: } \frac{-0.022 - 0.012}{2} = -0.017 \quad (\text{gem. van } 25^e \text{ en } 26^e)$$

laagste steekproefkwartiel: -0.096 (13^e)

hoogste steekproefkwartiel: 0.103 (13^e van groot naar klein)

$$1.5 * \text{steekproefkwartielafstand: } 1.5 * (0.103 + 0.096) = 0.2985$$

Uitschieters?

Dit zijn waarden kleiner dan $-0.096 - 0.2985 = -0.3945$ en waarden groter dan $0.103 + 0.2985 = 0.4015$: dus geen uitschieters.

Verder op te nemen in boxplot: kleinste waarde (-0.354) en grootste waarde (0.264).

Boxplot (zie sectie 1.6.2 van hoofdstuk 2 van Statistische Technieken) in goede verhoudingen tekenen.

7.7b

De punten van de normale Q-Q plot liggen redelijk langs een rechte lijn. De afwijkingen ten opzichte van een rechte lijn zijn te verklaren door toevalsfluctuatie. De getoonde normale Q-Q plot duidt daarom op een normale verdeling.

Deze conclusie wordt niet tegengesproken door de boxplot: er is niet sprake van (flinke) scheefheid en er zijn geen uitschieters.

Opgave 7.8

7.8a

Analysis of Variance

	DF	Sum of Squares	Mean Square	F Ratio
Regression	3	21.409	7.136	19.3
Residual	16	5.903	0.369	

7.8b

$$y = -2.491 + 0.099x_1 + 0.029x_2 + 0.086x_3$$

Eventueel x_1, x_2, x_3 en y vervangen door de namen van de variabelen.

7.8c

We gaan eerst onderzoeken of de glooiing van de kavel van invloed is op de verkoopprijs als we ook al het oppervlak en de hoogte van de kavel gebruiken als verklarende variabelen.

Laat Y_i de verkoopprijs van kavel i zijn, x_{1i} de bijbehorende oppervlakte, x_{2i} de bijbehorende hoogte en x_{3i} de bijbehorende glooiing.

De acht stappen van de toets zijn als volgt:

1. Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + U_i$.

De storingen U_i zijn onderling onafhankelijk en $N(0, \sigma^2)$ -verdeeld.

(De parameters $\beta_0, \beta_1, \beta_2, \beta_3$ en σ^2 zijn onbekend.)

2. $H_0 : \beta_3 = 0$ tegen $H_1 : \beta_3 \neq 0$

3. Toetsingsgrootte: $T = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}}$

4. Onder $H_0 : T \sim t_{16}$

5. Uitkomst van $T : \frac{0.086}{0.031} = 2.77$

6. We verwerpen H_0 als $T \leq -c$ of $T \geq c$.

Onbetrouwbaarheidsdrempel 5%, t -tabel $\Rightarrow c = 2.12$

7. Uitkomst 2.77 ligt in kritiek gebied $\Rightarrow H_0$ verwerpen.

8. We achten bewezen, bij onbetrouwbaarheidsdrempel 5%, dat de glooiing iets bijdraagt tot het voorspellen van de verkoopprijs, in toevoeging op het oppervlak en de hoogte van de kavel.

Voor het toetsen of de hoogte van de kavel iets bijdraagt aan het voorspellen van de verkoopprijs in toevoeging tot de andere twee verklarende variabelen kunnen we het bovenstaande schema op de voor de liggende wijze aanpassen.

Stappen 1, 4 en 6: blijven hetzelfde.

Stappen 2 en 3: $3 \rightarrow 2$ (vier keer)

Stappen 5, 7 en 8: uitkomst van toetsingsgrootte wordt nu 4.8, dus H_0 ook nu verwerpen. Ook de hoogte draagt bij tot het voorspellen van de verkoopprijs in toevoeging tot de bijdragen van de twee andere verklarende variabelen.

Voor het toetsen of het oppervlak van de kavel iets bijdraagt aan het voorspellen van de verkoopprijs in toevoeging tot de andere twee verklarende variabelen kunnen we het bovenstaande schema op de voor de liggende wijze aanpassen.

Stappen 1, 4 en 6: blijven hetzelfde.

Stappen 2 en 3: $3 \rightarrow 1$ (vier keer)

Stappen 5, 7 en 8: uitkomst van toetsingsgrootte wordt nu 1.71, dus nu wordt H_0 niet verworpen. Kennelijk verschillen de oppervlaktes van de 20 kavels niet op zodanige wijze dat het de verkoopprijs aantoonbaar beïnvloedt.

7.8d.

In eerste instantie lijkt het erop dat naar een toets gevraagd wordt, die we al bij onderdeel b hebben uitgevoerd. Er is echter één verschil: er wordt nu naar een eenzijdige toets gevraagd. We moeten nu $H_0 : \beta_3 = 0$ toetsen tegen $H_1 : \beta_3 > 0$ omdat een negatieve waarde voor β_3 kennelijk niet kan.

We moeten ervoor waken dat de vaststelling van H_0 en H_1 niet van de data gaat afhangen, daarmee zou de inhoud van H_0 en/of H_1 ook stochastisch worden, terwijl in de theorie alleen voorzien is in vaste hypothesen (van tevoren bekend en niet veranderend). De tweezijdige toets veranderend in de eenzijdige toets, krijgen we de volgende acht stappen:

1. Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + U_i$.

De storingen U_i zijn onderling onafhankelijk en $N(0, \sigma^2)$ -verdeeld.

(De parameters $\beta_0, \beta_1, \beta_2, \beta_3$ en σ^2 zijn onbekend.)

2. $H_0 : \beta_3 = 0$ tegen $H_1 : \beta_3 > 0$

3. Toetsingsgrootte: $T = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}}$

4. Onder $H_0 : T \sim t_{16}$

5. Uitkomst van $T : \frac{0.086}{0.031} = 2.77$

6. We verwerpen H_0 als $T \geq c$.

Onbetrouwbaarheidsdrempel 5%, t -tabel $\Rightarrow c = 1.75$

7. Uitkomst 2.77 ligt in kritiek gebied $\Rightarrow H_0$ verwerpen.

8. We achten bewezen, bij onbetrouwbaarheidsdrempel 5%, dat een grotere glooiing bijdraagt tot een hogere verkoopprijs, als het oppervlak en de hoogte van de kavel ook verklarende variabelen in het model zijn.

7.8e

Gevraagd: 90%-BI voor β_1 : $(\hat{\beta}_1 - cS_{\hat{\beta}_1}, \hat{\beta}_1 + cS_{\hat{\beta}_1})$

De interpretatie is als bij elk 90%-BI: als we keer op keer dit soort intervallen gaan uitrekenen, dan zal in 90% van de gevallen de waarde van de echte parameter (nu β_1) in het interval vallen. (We weten echter niet wanneer wel en wanneer niet.)

Over de betekenis van β_1 het volgende:

De parameter β_1 geeft niet aan hoe y (verkoopprijs) afhangt van x_1 (oppervlak), maar geeft aan hoe de verkoopprijs verandert ten gevolge van een verandering in het oppervlak in de situatie dat er al rekening is gehouden met de hoogte (x_2) en glooiing (x_3) van de kavel.

Berekening van BI:

Uit output vinden we 0.099 en 0.058 voor resp. $\hat{\beta}_1$ en $S_{\hat{\beta}_1}$,

verder: $c = 1.75$ (90%-BI, t_{16}). Invullen levert: $(-0.003, 0.201)$.

Opgave 7.9

7.9a

aangepaste vergelijking

$$y = 7.296 + 2.153x + 0.1043x^2$$

7.9b

voorspelling voor $x = 5$

Invullen in aangepaste vergelijking: $\hat{y} = 7.296 + 2.153 \times 5 + 0.1043 \times 25 = 20.7$

7.9c

De nulhypothese $H_0 : \beta_1 = 0$ en $H_0 : \beta_2 = 0$ kunnen kennelijk beide geaccepteerd worden. Voor het schrappen van beide termen (zowel de term met x als de term met x^2) zou de nulhypothese $H_0 : \beta_1 = \beta_2 = 0$ niet verworpen moeten worden. Voor het wel of niet verwerpen van de laatste nulhypothese moeten we een andere toetsingsgrootheid gebruiken, namelijk de F van de gepresenteerde analysis-of-variance-tabel. (Daaruit blijkt dat we $H_0 : \beta_1 = \beta_2 = 0$ moeten verwerpen, dus tenminste 1 van de twee termen hoort wel thuis in het model.)

We moeten beide termen dus niet schrappen. Eerst voorlopig maar 1 term schrappen.

7.9d

Dit is hartstikke fout. Met het schrappen van de term met x^2 krijgen de overblijvende parameters/termen een andere betekenis. De overblijvende parameters moeten opnieuw geschat worden.

Opgave 7.10

7.10a

We hebben dus 45 apparaten, met voor elk een waarde voor x_1 en x_2 (zoals in de tekst aangegeven) en een waarde voor de opwarmtijd die we met y aangeven.

Voor de 15 opwarmtijden van apparaat type A zijn de bijbehorende waarden van x_1 en x_2 als volgt: $x_1 = 1$ en $x_2 = 0$.

Voor de 15 opwarmtijden van apparaat type B hebben we telkens $x_1 = 0$ en $x_2 = 1$ en voor de 15 opwarmtijden van apparaat type C geldt telkens $x_1 = 0$ en $x_2 = 0$. Laat y_i de opwarmtijd zijn van apparaat i ($i = 1, 2, \dots, 45$), met x_{i1} en x_{i2} de bijbehorende waarden van x_1 en x_2 . De waarden y_i vatten we op als uitkomsten van stochastische variabelen $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + U_i$. De storingsen U_i zijn stochastische variabelen die onderling onafhankelijk zijn en $N(0, \sigma^2)$ -verdeeld. (De parameters $\beta_0, \beta_1, \beta_2$ en σ^2 zijn onbekend.)

7.10b

Aangepaste vergelijking: $y = 18.267 + 5.933x_1 + 7.933x_2$

Hiermee kunnen we dus (punt)voorspellingen maken voor de opwarmtijd y . Voor type A wordt de voorspelling $\hat{y} = 18.267 + 5.933 \times 1 + 7.933 \times 0 = 18.267 + 5.933 = 24.200$, voor type B krijgen we $\hat{y} = 18.267 + 7.933 = 26.200$ en voor type C $\hat{y} = 18.267$.

7.10c

De residual mean square is de schatting voor σ^2 : 22.089

7.10d

De 8 stappen zijn als volgt.

1. Model: zie antwoord bij 7.10a.

2. $H_0 : \beta_1 = \beta_2 = 0$ tegen $H_1 : \beta_1 \neq 0$ en/of $\beta_2 \neq 0$

3. Toetsingsgrootheid: $F = \frac{SSR/k}{SSE/(n-(k+1))} = \frac{SSR/2}{SSE/42}$

4. Onder $H_0 : F \sim F_{42}^2$

5. Uitkomst van $F : \frac{510.711/2}{927.733/42} = \frac{255.356}{22.089} = 11.56$ (uitkomst staat al in output)

6. We verwerpen H_0 als $F \geq c$

F_{42}^2 -verdeling, onbetrouwbaarheidsdrempel 1% $\Rightarrow c = 5.16$

7. Uitkomst 11.56 ligt in kritiek gebied $\Rightarrow H_0$ verwerpen.

8. Bij onbetrouwbaarheidsdrempel 1% achten we bewezen dat de drie typen apparaten systematisch verschillen met betrekking tot de opwarmtijd.

Opgave 7.11

Opmerking: de term $\beta_3 x_1 x_2$ in de regressievergelijking $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ duidt men wel aan met “interactie”. Met de toevoeging van de interactieterm blijft de relatie tussen y en x_1 lineair als we de waarde van x_2 vasthouden. (Omgekeerd blijft de relatie tussen y en x_2 ook lineair als we de waarde van x_1 vasthouden.) Het is wel zo dat de relatie tussen y en x_1 bij vaste x_2 afhangt van de waarde van x_2 , aan deze beïnvloeding dankt de term $\beta_3 x_1 x_2$ zijn naam, interactie. (Omgekeerd hangt de relatie tussen y en x_2 bij vaste x_1 ook af van de waarde van x_1 .)

We moeten de term $\beta_3 x_1 x_2$ opvatten als een term $\beta_3 x_3$. We introduceren een derde verklarende variabele x_3 . Voor elke waarneming/pakket krijgen we een waarde voor x_3 door het product te nemen van de waarden van x_1 en x_2 .

7.11a

$$y = -0.1405 + 0.0191x_1 + 0.0077x_2 + 0.0078x_1x_2$$

7.11b

Gevraagd wordt de (gemiddelde) toename in y te schatten voor een extra mijl in geval van $x_1 = 2$. Op grond van de aangepaste vergelijking schatten we dit met:

$$0.0077 + 0.0078 \times 2 = 0.0233 \text{ (dollar).}$$

7.11c

We gaan toetsen of de interactieterm thuishoort in het model. Laat Y_i de kosten zijn van het vervoeren van pakket i , x_{1i} het bijbehorende gewicht en x_{2i} de afstand die voor pakket i overbrugd moet worden. De acht stappen van de toets zijn als volgt.

1. Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + U_i$.

De storingen U_i zijn onderling onafhankelijk en $N(0, \sigma^2)$ -verdeeld.

(De parameters $\beta_0, \beta_1, \beta_2, \beta_3$ en σ^2 zijn onbekend.)

2. $H_0 : \beta_3 = 0$ tegen $H_1 : \beta_3 \neq 0$

3. Toetsingsgrootte: $T = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}}$

4. Onder $H_0 : T \sim t_{16}$

5. Uitkomst van $T : \frac{0.0078}{0.0009} = 8.7$

6. We verwerpen H_0 als $T \leq -c$ of $T \geq c$.

Onbetrouwbaarheidsdrempel 5%, t -tabel $\Rightarrow c = 2.12$

7. Uitkomst 8.7 ligt in kritiek gebied $\Rightarrow H_0$ verwerpen.

8. We achten bewezen, bij onbetrouwbaarheidsdrempel 5%, dat de interactieterm thuishoort in het model.

7.11d

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)} \times \frac{SSE}{S_{YY}} = 1 - \frac{19}{16} \times \frac{6.63}{6.63 + 445.45} = 1 - \frac{19}{16} \times 0.01467 = 0.0983 = 98.3\%$$

Dit is een goede waarde, want is bijna 100%.