Statistiek: Werkcollege 5 / hoofdstuk 6

# Vraag 1

* 1. Voorwaarden lineaire regressie:
		1. Alleen y-waardes bevatten fouten. Bij een dergelijke dataset met experimenteel gevonden waarden is de keuze bijna arbitrair. Misschien kun je beredeneren dat de sterftecijfers iets betrouwbaardere data zijn en je deze daarom prefereert op de x-as.
		2. Die fouten zijn normaal verdeeld.

Hier kun je moeilijk een uitspraak over doen, maar er is geen reden aan te nemen dat dit niet zo is.

* + 1. Homoscedasticiteit: overal dezelfe variantie.

Het is nog maar de vraag of hieraan voldaan is. De metingen “waaieren uit”, dwz de metingen lijken steeds verdere afwijkingen van een denkbeeldige lijn erdoorheen te vertonen. Daarbij is het logisch dat mensen die weinig roken er minder naast zitten (bv. een paar sigaretten ernaast) als mensen die veel roken (bv. een paar pakjes ernaast).

* 1. Het gaat hier om de grootte van de groepen. Als je alle mensen afzonderlijk zou plotten zou je zien dat de lijn wel door het gemiddelde van de data zou gaan. Hier kan één punt echter meer of minder mensen dan een ander punt voorstellen, en is *de facto* dus sprake van een gewogen gemiddelde.
	2. Hiervoor kijken we naar de waarde van b, want als deze 0 is, dan is er geen helling, en dus geen verband.

Het betrouwbaarheidsinterval van b is b ± t(n-2,α) ∙ sb, in dit geval dus b ± t(25-2,0.05) ∙ sb = 1.09 ± 2.086[[1]](#footnote-1)∙ 0.22. De waarde 0 ligt dus niet in dit interval, en dus is er een significant verband. (Moet je eigenlijk niet ook controleren voor de waarde b = ∞?)

* 1. Dat is wel een erg gevaarlijke conclusie. Misschien ligt er wel een andere oorzaak aan ten grondslag die op dezelfde wijze wordt gedeeld als roken (of zelfs in grotere mate), of misschien zorgt “het kankergen” er ook wel voor dat mensen graag roken . Er is in elk geval een verband, maar of dit causaal is zal onderzocht moeten worden.

# Vraag 2

* + - 1. $b=\frac{SS\_{xy}}{SS\_{xx}}=\frac{\sum\_{}^{}\left(x\_{i}-\overbar{x}\right)(y\_{i}-\overbar{y})}{\sum\_{}^{}(x\_{i}-\overbar{x})^{2}}$.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **%iso-octaan** | *xi-xgem* | (xi-xgem)^2 |  | **piekoppervlak** | *yi-ygem* |  | ((xi-xgem)\*(yi-ygem)) |
|  | 0,532 | -0,577 | 0,332929 |  | 1,09 | -1,412 |  | 0,814724 |  |
|  | 0,803 | -0,306 | 0,093636 |  | 1,78 | -0,722 |  | 0,220932 |  |
|  | 1,08 | -0,029 | 0,000841 |  | 2,6 | 0,098 |  | -0,00284 |  |
|  | 1,38 | 0,271 | 0,073441 |  | 3,03 | 0,528 |  | 0,143088 |  |
|  | 1,75 | 0,641 | 0,410881 | + | 4,01 | 1,508 |  | 0,966628 | + |
|  | Gem.: 1,109 |  | 0,911728 |  | Gem.: 2,502 |  |  | 2,14253 |  |

|  |  |  |
| --- | --- | --- |
| b | = | 2,34996622 |
|  |  |  |
| a=y-bx | = | -0,10411254 |

|  |  |
| --- | --- |
| (yi-yber)^2 | /3 |
| 0,00314379 | 0,001048 |
| 8,4701E-06 | 2,82E-06 |
| 0,0276055 | 0,009202 |
| 0,01184633 | 0,003949 |
| 2,7944E-06 | 9,31E-07 |
|  | 0,014202 |
|  |  |
| sqrt | 0,119173 |

Het blijkt dus dat b = 2.35 en a = -0.104. Dus de detectielimiet is:

 $y\_{dl}=y\_{blanco}+3s\_{blanco} $

 Met $s\_{res}=\sqrt{\sum\_{}^{}\frac{(y\_{i}-\hat{y}\_{i})^{2}}{n-2}}=$

 Dus 0.119 is sres.

 ydl = -0.104 + 3 ∙ 0.119 =

 0.253.

1. Als je een waarde meet gelijk aan ($y\_{dl}=y\_{blanco}+3s\_{blanco}$ =) 0.253, wat is dan de bijbehorende x-waarde, uitgedrukt als 95% betrouwbaarheidsinterval?

De x-waarde ($\hat{x}$) is dus te bepalen uit de vergelijking van de lijn, en is

$$y=2.35∙x-0.104\rightarrow $$

$$\hat{x}=\frac{y+0.104}{2.35}=\frac{0.253+0.104}{2.35}=0.152$$

$$s\_{x0}=\frac{s\_{res}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(y\_{0}-\overbar{y})^{2}}{b^{2}\sum\_{}^{}(x\_{i}-\overbar{x})^{2}}}=\frac{0.1192}{2.350}\sqrt{\frac{1}{1}+\frac{1}{5}+\frac{(0.253-2.502)^{2}}{2.350^{2}∙0.912}}=0.0753$$

Dus het 95% betrouwbaarheidsinterval is $\hat{x}$ ± t(n-2,α) ∙ sx0 = $\hat{x}$ ± t(3,α=0.05) ∙ sx0 = 0.152 ± 3.182 ∙ 0.0753 = 0.152 ± 0.240.

# Vraag 3

1. De correlatiecoefficient wordt berekend via formule 6.1. Dit doe ik weer met Excel:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | x | y | x-xgem | y-ygem | (x-xgem)(y-ygem) | (x-xgem)^2 | (y-ygem)^2 |
|  | 0,01 | 127,6 | -0,86308 | 18,7846154 | -16,21256805 | 0,744901775 | 352,861775 |
|  | 0,48 | 124 | -0,39308 | 15,1846154 | -5,968721893 | 0,154509467 | 230,572544 |
|  | 0,71 | 110,8 | -0,16308 | 1,98461538 | -0,32364497 | 0,026594083 | 3,93869822 |
|  | 0,95 | 103,9 | 0,076923 | -4,91538462 | -0,378106509 | 0,00591716 | 24,1610059 |
|  | 1,19 | 101,5 | 0,316923 | -7,31538462 | -2,318414201 | 0,100440237 | 53,5148521 |
|  | 0,01 | 130,1 | -0,86308 | 21,2846154 | -18,37026036 | 0,744901775 | 453,034852 |
|  | 0,48 | 122 | -0,39308 | 13,1846154 | -5,182568047 | 0,154509467 | 173,834083 |
|  | 1,44 | 92,3 | 0,566923 | -16,5153846 | -9,362952663 | 0,321401775 | 272,757929 |
|  | 0,71 | 113,1 | -0,16308 | 4,28461538 | -0,698721893 | 0,026594083 | 18,357929 |
|  | 1,96 | 83,7 | 1,086923 | -25,1153846 | -27,29849112 | 1,181401775 | 630,782544 |
|  | 0,01 | 128 | -0,86308 | 19,1846154 | -16,55779882 | 0,744901775 | 368,049467 |
|  | 1,44 | 91,4 | 0,566923 | -17,4153846 | -9,873183432 | 0,321401775 | 303,295621 |
|  | 1,96 | 86,2 | 1,086923 | -22,6153846 | -24,58118343 | 1,181401775 | 511,455621 |
| gemiddelde: | 0,873076923 | 108,8153846 |  |  |   |   |   |
| som: |  |  |  |  | -137,1266154 | 5,708876923 | 3396,61692 |
|  |  |  |  |  |  | maal elkaar = |   |
|  |  |  |  |  |  | 19390,86797 |   |
|  |  |  |  |  | sqrt | 139,2510968 |  |
|  |  |  |  |  |  |  |  |
|  | Dus  |  |  |  |  |  |  |
|  | r(x,y) =  | 134391,6695 | / | 292678,625 | = | -0,984743521 |  |
|  | r^2= |  |  |  |  | 0,969719802 |  |
|  |  |  |  |  |  |  |  |

De r(x,y) is dus -0,985.

1. Dit is exact hetzelfde wat ik bij 2 ook al deed (alleen dan met deze gegevens), plus het invullen van de formules 6.9 en 6.10. Deels zijn deze waarden ook door Excel bepaald (zie boven).
2. Dit is waarschijnlijk wel het geval, want de punten in de figuur hierboven lijken niet meer of minder af te wijken van de trendlijn bij een hogere of lagere x. Echter zou je dit beter precies kunnen doen door telkens de formule in te vullen en dat van de waarde van een punt af te trekken en dit te plotten. Dan zou je ook in de tijd kunnen kijken of er sprake is van homoscdasticiteit.

*Aanvulling van de nakijker: ‘Als je dat plotje zou maken voor de homoscedasticiteit, zou je kunnen zien dat de punten op een sinusoide liggen... Maar dat ziet niet iedereen en op zich vind ik je antwoord wel acceptabel.’*

De punten zijn niet goed *gedisigned*, want als je ze aan de uiterste x waardes legt dan krijg je een smaller betrouwbaarheidsinterval. Beter was het dus geweest om de meeste punten rond 0 en 2 te verzamelen, en enkele in het midden te zetten (ter controle van het lineaire verband).

# Vraag 4

(Tentamenopgave blz 152, 2a).

1. De regressie is significant als 0 niet in het betrouwbaarheidsinterval van de helling (b) ligt. Het 777 betrouwbaarheidsinterval van b is b ± t(n-2,α) ∙ sb, in dit geval dus b ± t(10-2,0.05) ∙ sb = -0.002 ± 2.306∙ 0.002, dus hier is geen significante regressie. Het 888 betrouwbaarheidsinterval van b is b ± t(n-2,α) ∙ sb, in dit geval dus b ± t(10-2,0.05) ∙ sb = 0.019 ± 2.306∙ 0.008, dus hier is net wel significante regressie. Dit laatste is natuurlijk wel een twijfelgeval, want de 95% grens is arbitrair.
1. 23 vrijheidsgraden stond er niet in, tcal ligt tussen 2.060 en 2.086. [↑](#footnote-ref-1)