Voorspellen met Modellen
Antwoorden op de niet gesterde opgaven

Opmerking: Dit document bevat de antwoorden op (de meeste van) de niet gesterde opgaven. De antwoorden op de gesterde opgaven staan in het boek. Sommige antwoorden zijn verkregen door de resultaten van tussenstappen af te ronden (zoals weergegeven in de tekst), zodat het antwoord zonder tussentijdse afrondingen hiervan soms iets kan afwijken.

Opgave 2.1

(i) $P(Y > 0) = 0,50$; $P(Y > 2) = 0,023$; $P(|Y| > 2) = 0,045$; en $P(|Y| > 4) = 0,000$ (dit betekent dat de kans kleiner is dan 0,0005, maar de kans is natuurlijk wel iets groter dan nul).

(ii) $\mathrm{N}(-1;1) : P(Y > 0) = 0,159$; $\mathrm{N}(1;1) : P(Y > 0) = 0,841$; en $\mathrm{N}(0;0.25) : P(Y > 0) = 0,500$.

Opgave 2.3

(i) De uitkomst van Y is 1 met kans p en 0 met kans $1-p$. Het gemiddelde is daarom $p \times 1 + (1-p) \times 0 = p$. De afwijking van Y ten opzichte van het gemiddelde is $1-p$ met kans p en $0-p = -p$ met kans $1-p$. De variantie is daarom $p(1-p)^2 + (1-p)(-p)^2 = p(1-p)$, en de standaardafwijking is de wortel hieruit, dus $\sqrt{p(1-p)}$.

(ii) Laat $f(p) = \sqrt{p(1-p)}$, dan is de afgeleide $f'(p) = (1-2p)/(2\sqrt{p(1-p)})$, waaruit de in de opgave genoemde resultaten volgen.

(iii) De standaardafwijking is een maatstaf voor de onzekerheid in de uitkomst van Y. De onzekerheid is het grootst voor $p = 1/2$ omdat de uitkomsten 1 en 0 dan even waarschijnlijk zijn, terwijl er geen onzekerheid is als $p = 1$ of als $p = 0$.

Opgave 2.4

(i) Y = aantal leerlingen dat een prijs wint. Het verwachte aantal is $np = 1000 \times 0,01 = 10$.

(ii) $P(Y = 0) = (0,99)^{100} = 0,000$ (dus kleiner dan 0,0005).

Opgave 3.6

(i) $a = -0,523$ en $b = 0,173$.

(ii) Voor $X = 3,0$ is de kans op een voldoende $p = -0,523 + 0,173 \times 3,0 = -0,004$, als $X = 6,5$ dan is $p = 0,602$, en als $X = 9,0$ dan is $p = 1,034$.

(iii) De slaagkansen liggen niet altijd tussen de 0 en 1, zodat het model niet goed te interpreteren is.

Opgave 3.8

(i) De gegevens zijn samengevat in de volgende tabel.

Inkomensklasse	1	2	3	4	5	6
Aantal buitenland	0	2	11	14	3	1
Aantal niet buitenland	2	6	6	0	0	0

Opgave 3.10

Geef de waarden van (X, Y) uitgedrukt in miljoenen euro's aan met (x_i, y_i), en uitgedrukt in euro's met (x_i^*, y_i^*). Dan geldt dat $x_i^* = mx_i$ en $y_i^* = my_i$, waarin $m = 10^6$.

(i) De parameters in de regressielijn $y = a + bx$ (in miljoenen euro's) worden gegeven door $a = \overline{y} - b\overline{x}$ en $b = s_{xy}/s_x^2$, en die van de regressielijn $y^* = a^* + b^*x^*$ (in euro's) door $b^* = s_{x^*y^*}/s_{x^*}^2$ en $a^* = \overline{y^*} - b^*\overline{x^*}$. Uit de definities van steekproefvariantie en steekproefcovariantie in §3.2 volgt dat $\overline{x^*} = m\overline{x}$, $\overline{y^*} = m\overline{y}$, $s_{x^*y^*} = m^2 s_{xy}$ en $s_{x^*}^2 = m^2 s_x^2$. Dit geeft $b^* = s_{x^*y^*}/s_{x^*}^2 = (m^2 s_{xy})/(m^2 s_x^2) = s_{xy}/s_x^2 = b$ en $a^* = \overline{y^*} - b^*\overline{x^*} = m\overline{y} - bm\overline{x} = ma = 10^6 a$. De regressielijn in

euro's is dus $y^* = a^* + b^* x^* = 10^6 a + bx^*$. Laat verder $e_i^* = y_i^* - a^* - b^* x_i^*$ de fouten zijn in het model uitgedrukt in euro's, dan volgt dat $e_i^* = m(y_i - a - bx_i) = me_i$ en dus $e_i^{*2} = m^2 e_i^2$ en $\text{SKF}^* = m^2\text{SKF}$. Hieruit volgt dat $s^{*2} = \text{SKF}^*/n = m^2\text{SKF}/n = m^2 s^2 = 10^{12} s^2$.

(ii) Geef de bijbehorende regressielijn aan met $y = \tilde{a} + \tilde{b}x^*$. Dan is $\tilde{b} = s_{x^*y}/s_{x^*}^2 = (ms_{xy})/(m^2 s_x^2) = b/m = 10^{-6} b$ en $\tilde{a} = \overline{y} - \tilde{b}\overline{x^*} = \overline{y} - (b/m)m\overline{x} = \overline{y} - b\overline{x} = a$. De regressielijn is nu dus $y = a + 10^{-6} bx^*$. De fouten zijn nu $\tilde{e}_i = y_i - \tilde{a} - \tilde{b}x_i^* = y_i - a - (b/m)mx_i = y_i - a - bx_i = e_i$, zodat voor de foutenvariantie geldt dat $\tilde{s}^2 = \widetilde{\text{SKF}}/n = \text{SKF}/n = s^2$.

(iii) Als X en Y beide in euro's worden gemeten, dan is volgens (i) $s_{y^*}^2 = m^2 s_y^2$ en $s^{*2} = m^2 s^2$, zodat $R^{*2} = 1 - (s^{*2}/s_{y^*}^2) = 1 - s^2/s_y^2 = R^2$. Als X in euro's wordt gemeten en Y in miljoenen euro's, dan blijft s_y^2 onveranderd en volgens (ii) geldt dit ook voor s^2, zodat R^2 ook niet verandert.

Opgave 4.3

(i) Het 99% voorspelinterval voor de logaritme van de verkopen is gedefinieerd als $[a + bX - 2{,}6s,\ a + bX + 2{,}6s]$, waarbij $s = 0{,}0385$ en $2{,}6s = 0{,}1001$. In 2002 is $a + bX = 3{,}487$ en in 2003 is $a + bX = 3{,}552$, zie het antwoord van Opgave 4.1(i). Voor de logaritme van de verkopen geeft dit voor 2002 het interval $[3{,}386,\ 3{,}588]$ en voor 2003 het interval $[3{,}451, 3{,}653]$. Het interval voor de verkopen krijgen we door de exponent te nemen, en het interval voor 2002 is $[29{,}55,\ 36{,}16]$ en voor 2003 is het $[31{,}53,\ 38{,}59]$.

(ii) Een 99% voorspelinterval houdt in dat de kans op een uitkomst binnen het interval gelijk is aan 99%. Bij een 95% voorspelinterval is deze kans 95%. Omdat er meer uitkomsten vallen in het 99% interval moet dat interval breder zijn dan het 95% interval.

Opgave 4.5

(i) Het voorspelinterval is van de vorm $[a + bX - cs,\ a + bX + cs]$ waarin, volgens het antwoord op Opgave 4.1(i), $a + bX = 3{,}487$ en $s = 0{,}0385$. De logaritme van de werkelijke verkopen heeft in 2002 de waarde $\ln(28{,}0) = 3{,}332$, en deze ligt in het interval als $3{,}332 \geq 3{,}487 - 0{,}0385c$, dus als $c \geq 4{,}021$. Voor de standaardnormale verdeling is een uitkomst in het interval $[-4{,}021,\ 4{,}021]$ gelijk aan 0,99994. De minimaal vereiste betrouwbaarheid is 99,994%, dus praktisch 100%.

(ii) Het interval moet enorm worden opgeblazen voordat de werkelijk waargenomen waarde erin valt. Met andere woorden, de werkelijke verkopen wijken zeer duidelijk af van wat redelijkerwijze voorspeld kon worden op basis van het verleden.

Opgave 5.2

(i) Regressie van het cijfer voor statistiek op het geslacht levert $a = 5{,}92$ en $b = -0{,}33$.

(ii) Met $n = 30$, $s_x^2 = 0{,}210$, $s_y^2 = 0{,}490$, $s^2 = 0{,}467$ en $s = 0{,}684$ volgt dat $R^2 = 1 - s^2/s_y^2 = 1 - 0{,}467/0{,}490 = 0{,}047$ en $t = b \times \sqrt{(n-2)s_x^2/s^2} = -0{,}33 \times \sqrt{(30-2) \times 0{,}210/0{,}467} = -1{,}171$.

(iii) De t-waarde valt binnen het interval $[-2,\ 2]$, zodat b niet significant is. We concluderen dat er geen significant verschil is tussen de resultaten van mannen en vrouwen.

(iv) Regressie van het verwachte cijfer voor statistiek op het geslacht levert $a = 6{,}824$, $b = -0{,}402$, $s = 0{,}898$, $s^2 = 0{,}806$, $s_y^2 = 0{,}840$ en $s_x^2 = 0{,}210$. Door gebruik te maken van dezelfde formules als in onderdeel (ii) volgt dat $R^2 = 1 - 0{,}806/0{,}840 = 0{,}040$ en $t = -0{,}402 \times \sqrt{(30-2) \times 0{,}210/0{,}806} = -1{,}086$. Er is dus geen significant verschil tussen de verwachtingen van mannen en vrouwen.

Opgave 5.3

(i) Regressie van het statistiek cijfer op het wiskunde cijfer levert $a = 1,424$ en $b = 0,637$.

(ii) Er geldt dat $s = 0,274$, $s_y^2 = 0,490$ en $s_x^2 = 1,023$, zodat $R^2 = 1 - 0,274^2/0,490 = 0,847$ en $t = 0,637 \times \sqrt{(30-2) \times 1,023/0,274^2} = 12,44$.

(iii) De t-waarde valt buiten het interval $[-2,\ 2]$, zodat b significant afwijkt van 0. Er is dus sprake van een significant verband tussen het wiskunde cijfer op de middelbare school en het statistiek cijfer op de universiteit. De waarde van R^2 ligt redelijk dicht bij 1, zodat de twee cijfers vrij nauw met elkaar samenhangen.

Opgave 5.4

De formule voor het 95% voorspelinterval is $[a + bX - 2s,\ a + bX + 2s]$, waarbij (volgens het antwoord op Opgave 5.3) voor $X = 7$ volgt dat $a + bX = 1,424 + 0,637 \times 7 = 5,883$ en $s = 0,274$. Het interval is dus gelijk aan $[5,883 - 2 \times 0,274,\ 5,883 + 2 \times 0,274] = [5.335,\ 6.431]$. Het behaalde cijfer (een 6,2) ligt in dit interval.

Opgave 5.5

De tabel vat enige resultaten samen (de eerste kolom verwijst naar het onderdeel van de opgave).

(i)	X	5	6	7	8	9
(i)	p	0,533	0,520	0,700	0,764	0,883
(i)	$q = ln(p/(1-p))$	0,132	0,080	0,847	1,175	2,021
(ii)	$a + bX$	-0,125	0,362	0,849	1,336	1,823
(iv)	$p = e^{a+bX}/(1+e^{a+bX})$	0,469	0,590	0,700	0,792	0,861

(i) Zie de derde regel van de tabel.

(ii) Regressie geeft $a = -2,560$ en $b = 0,487$, met model $q = -2,560 + 0,487X$.

(iii) Omdat $s = 0,205$, $s^2 = 0,042$, $s_y^2 = 0,517$ en $s_x^2 = 2,000$ volgt dat $R^2 = 0,92$ en $t = 5,81$.

(iv) Uit $q_i = \ln(p_i/(1-p_i))$ volgen de kansen $p_i = e^{q_i}/(1+e^{q_i})$ volgens het model in (ii). Deze kansen staan in de onderste regel van de tabel.

(v) Voor $X = 5$ is de modelkans van 46,9% wat lager dan de kans van 53,3% in de steekproef, terwijl voor $X = 6$ de modelkans van 59,0% wat groter is dan de kans van 52,0% in de steekproef. Volgens het model is de kans op een voldoende groter naarmate het wiskundecijfer hoger is, en dat is ook logisch. In de steekproef is dit weliswaar niet het geval voor de wiskundecijfers $X = 5$ en $X = 6$, maar dat is zeer vermoedelijk te wijten aan het toeval. Helemaal zeker is dit niet, want het zou bijvoorbeeld kunnen dat studenten met $X = 5$ zich harder gaan inspannen dan die met $X = 6$ omdat ze op school al gewaarschuwd zijn. Dit soort mogelijke psychologische effecten is niet verwerkt in het model.

Opgave 5.6

De leerling had een 6 voor zijn eindexamen wiskunde en de bijbehorende kans op een voldoende is volgens het model in Opgave 5.5(iv) gelijk aan 59%. De kans op een voldoende is dus groter dan die op een onvoldoende. Hij behaalde een 6,2, dus inderdaad een voldoende.

Opgave 5.7

De percentages van studenten in de groepen $X = 1, \ldots, 5$ zijn respectievelijk als volgt: 4,75%, 23,73%, 37,03%, 24,58% en 9,92%. Het model uit Opgave 5.5 voorspelt per groep de volgende slagingspercentages: 46,9%, 59,0%, 70,0%, 79,2% en 86,1%. Het slagingspercentage van de totale groep is dan het gewogen gemiddelde: $0,0475 \times 0,469 + 0,2373 \times 0,590 + 0,3703 \times 0,700 + 0,2458 \times 0,792 + 0,0992 \times 0,861 \approx 70\%$. Deze schatting ligt iets hoger dan het werkelijke percentage voldoendes voor statistiek, 67,3%. Merk op dat het model geschat is op basis van een steekproef van slechts $n = 30$ studenten, en dat de voorspellingen worden gemaakt voor een groep van 948 studenten.