Multivariate statistiek regressie les 3

§13.4 Meervoudige regressie, dummyvariabelen

Determinatiecoëfficiënt, correlatiematrix, meervoudige correlatie, multicollineariteit

Meervoudige regressie, twee onafhankelijke variabelen

Bij meervoudige regressie zijn er naast één afhankelijke variabele twee of meer onafhankelijke variabelen.

We gaan eerst uit van twee onafhankelijke variabelen X_1 en X_2, en een afhankelijke variabele Y.

Voorbeeld shampoofabrikant

Tabel 13.12 Gegevens shampoofabrikant

Maand	Omzet per maand	Prijs per stuk	Advertentie-uitgaven
1	18.000 stuks	€ 4,00	€ 10.000
2	20.000	€ 3,75	€ 10.000
3	24.000	€ 3,25	€ 12.000
4	17.000	€ 3,75	€ 8.000
5	22.000	€ 3,00	€ 12.000
6	21.000	€ 3,50	€ 14.000
7	25.000	€ 2,75	€ 20.000
8	23.000	€ 3,00	€ 18.000

Afhankelijke variabele: Omzet per maand

Onafhankelijke variabelen: 1 Prijs per stuk
2 Advertentie-uitgaven

Model en regressievergelijking

Het lineaire regressiemodel met twee onafhankelijke variabelen:

$$\underline{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underline{\varepsilon}$$

De te schatten regressievergelijking is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

De punten (X_1, X_2, Y) die hieraan voldoen liggen op een vlak.

Notatieverschillen t.o.v. boek van Buijs:

- Voor de constante term: β_0 in plaats van α
- Voor de geschatte Y: $\hat{Y}$ in plaats van Y^c

Kleinstekwadratenmethode

De voorspellingsfout is: $e_t = Y_t - \widehat{Y}_t$

De te minimaliseren kwadratensom is:

$$\sum {e_t}^2 = \sum \left(Y_t - \widehat{Y}_t\right)^2 = \sum (Y_t - b_0 - b_1 X_1 - b_2 X_2)^2$$

Deze is een functie van b_0, b_1 en b_2 .

We zoeken dus de combinatie van waarden van b_0, b_1 en b_2 waarvoor een minimale waarde wordt bereikt van:

$$\sum (Y_t - b_0 - b_1 X_1 - b_2 X_2)^2$$

Hoe minimaliseren?

Minimaliseren van deze kwadratensom kan op verschillende manieren, zoals:

- Analytisch met behulp van de theorie uit calculus (functie van meerdere variabelen)
- Analytisch met behulp van de theorie van lineaire algebra (matrixrekening), en dan berekenen bijv. m.b.v. R
- Numeriek m.b.v. de Oplosser in Excel
- Numeriek m.b.v. statistische software

Meervoudige regressie, willekeurig aantal onafhankelijke variabelen

Het lineaire regressiemodel met k onafhankelijke variabelen:

$$\underline{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \underline{\varepsilon}$$

De te schatten regressievergelijking voor de (voorspelling van de) onafhankelijke variabele bij k onafhankelijke variabelen:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots b_k X_k$$

De aanpak met de kleinstekwadratenmethode is vergelijkbaar met die bij 1 of 2 onafhankelijke variabelen.

Variantie en covariantie

Variantie van X (in een steekproef)

$$\mathrm{v}ar(x) = {s_x}^2 = \frac{\sum\left(X_i-\overline{X}\right)^2}{n-1}$$

Covariantie tussen X en Y (in een steekproef):

$$cov(x, y) = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

Covariantie en correlatie

Er zijn twee maatstaven voor onderlinge samenhang tussen twee variabelen in een steekproef.

Covariantie tussen X en Y:

$$cov(x,y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Correlatiecoëfficiënt tussen X en Y:

$$r(x,y) = \frac{cov(x,y)}{s_x \cdot s_y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

waarin s_x resp. s_y de standaarddeviatie van x en y.

Correlatiematrix

De correlatiematrix is de matrix met de correlaties tussen elk tweetal variabelen. Bijv. bij variabelen Y, X_1, X_2, X_3:

	Y	X_1	X_2	X_3
Y	1	r_{yx_1}	r_{yx_2}	r_{yx_3}
X_1		1	$r_{x_1x_2}$	$r_{x_1x_3}$
X_2			1	$r_{x_2x_3}$
X_3				1

Op de diagonaal staan de correlaties van elke variabele met zichzelf; die zijn 1.

Linksonder staat hetzelfde als rechtsboven, want $r_{ab} = r_{ba}$

Multicolineariteit

Multicolineariteit is een te sterke samenhang tussen onafhankelijke variabelen onderling.

Algemene regels:

- De correlatie tussen onafhankelijke variabelen onderling mag (in absolute waarde) niet groter zijn dan 0,7;
- De correlatie tussen onafhankelijke variabelen onderling mag niet sterker zijn dan de correlatie van één van deze met de afhankelijke variabele.

Voorbeeld

Dataset met gegevens over een bepaald product, over een aantal maanden, met per maand:

Afhankelijke variabele:

- Y: de afzet (aantal verkopen) van een bepaald product

Onafhankelijke variabelen:

- X_1: de prijs van het product
- X_2: reclame-uitgaven voor het product
- X_3: de prijs van hetzelfde product bij een grote concurrent

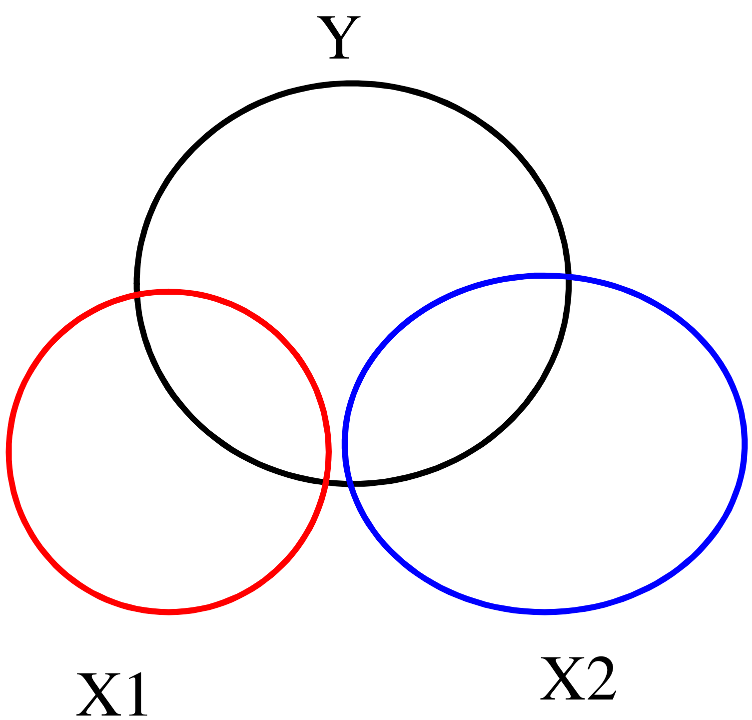
Voorbeeld van multicolineariteit

De correlatiematrix is

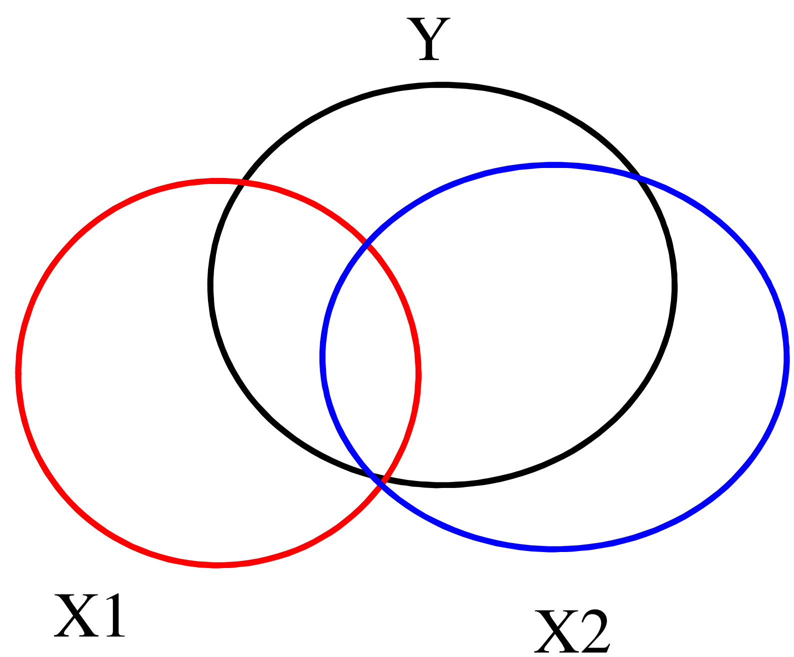
	Afzet (Y)	**Prijs** (X_1)	**Reclame** (X_2)	**Prijs conc** (X_3)
Afzet (Y)	1	$-0{,}79$	0,81	0,51
Prijs (X_1)		1	$-0{,}58$	0,83
Reclame(X_2)			1	$-0{,}46$
Prijs conc (X_3)				1

Hier is de correlatie tussen X_1 en X_3 te groot, want die is groter dan 0,7.
Er is dus multicolineariteit.
Mogelijke oplossing: X_1 of X_3 uit het model verwijderen.
Dat is nogal rigoreus; enkele dia's verderop meer mogelijkheden.

Correlaties in een schema

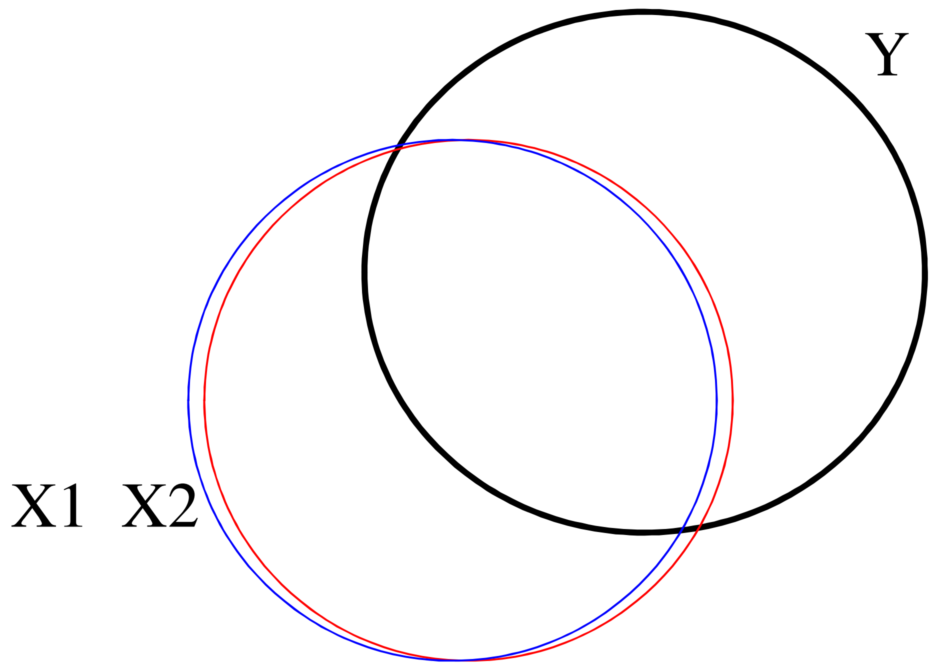


Geen correlatie tussen X1 en X2. Elke variabele verklaart een deel van de variantie van Y.



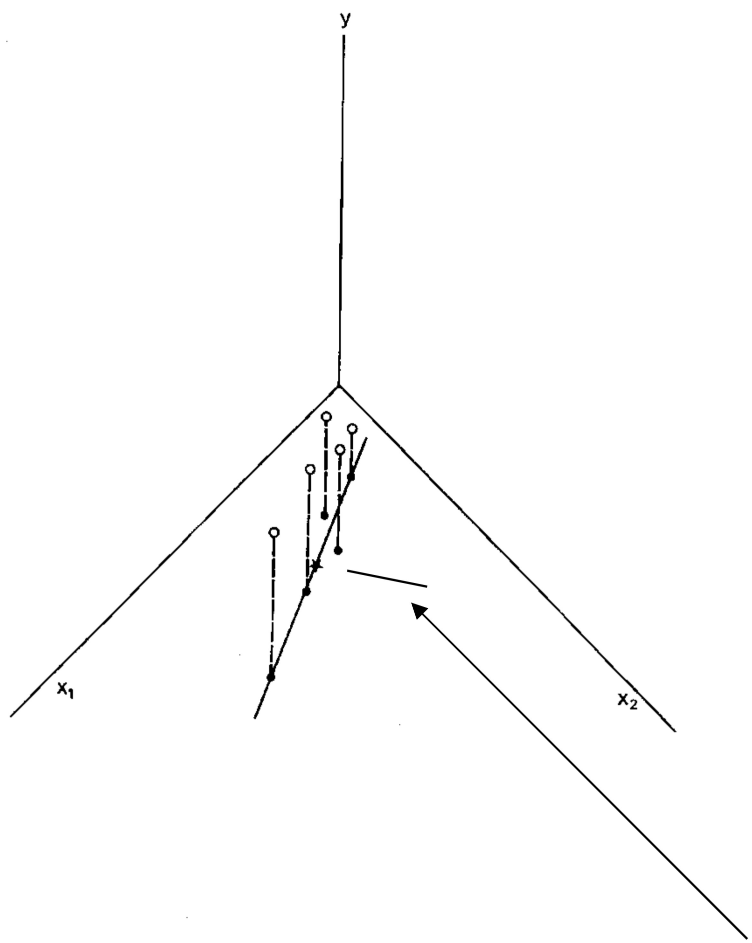
Wel correlatie tussen X1 en X2. Een deel van de variantie van Y wordt door beide verklaard (multicolineariteit).

Bijna volledige multicolineariteit



Het is hier zinloos om zowel X1 als X2 in het model op te nemen.

Het gevolg van multicolineariteit



We schatten de regressievergelijking:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

De punten die aan deze vergelijking voldoen liggen op een vlak V.

Stel: X_1 en X_2 hebben een sterke correlatie, dus de datapunten liggen bijna op een lijn l.
Vlak V gaat (bij benadering) door l.
De onnauwkeurigheid in de ligging van V is groot: een kleine verandering van de data geeft een grote schommeling rond lijn l.

Oplossingen voor multicolineariteit

1. Een onafhankelijke variabele eruit (nadeel: de informatie van die variabele laat je dan geheel weg).

2. Kritisch kijken naar de definitie van variabelen, bijvoorbeeld:

A) we verklaren omzet (Y) uit promotiebudget en aantal distributiepunten, uit data op maandbasis.

promotiebudget en aantal distributiepunten correleren sterk met elkaar; beter: variabele *'promotiebudget PER distributiepunt'*

B) we verklaren omzet in een regio (Y) uit aantal inwoners in regio, en promotiebudget in regio.

beter: variabele *'omzet PER hoofd vd bevolking in een regio'*

Oplossingen voor multicolineariteit

3. ‘Eerste’ verschillen nemen; d.w.z. in plaats van X_t nemen we $X_t - X_{t-1}$ als onafhankelijke variabele; voorbeeld:

   Afhankelijke variabele:

   - Y: de afzet (aantal verkopen) van een product in een maand

   Onafhankelijke variabelen:

   - X_1: de prijs van het product
   - X_2: reclame-uitgaven voor het product
   - X_3: de prijs van hetzelfde product bij een grote concurrent

   Stel: X_1 en X_3 hebben een sterke correlatie, dan kan je voor X_3 de ‘eerste verschillen’ nemen. Dus we gebruiken dan niet de prijs van de concurrent, maar zijn prijsVERSCHIL t.o.v. de vorige maand.

Variantie-inflatiefactor (VIF)

Bereken voor elke onafhankelijke variabele de determinatiecoëfficiënt R_j^2 met alle andere onafhankelijke variabelen

Bij drie onafhankelijke variabelen dus:

$$X_1 = a_1 + b_1 X_2 + c_1 X_3$$
$$X_2 = a_2 + b_2 X_1 + c_2 X_3$$
$$X_3 = a_3 + b_3 X_1 + c_3 X_2$$

levert R_1^2, R_2^2 en R_3^2; bereken de VIF met : $VIF_j = \dfrac{1}{1 - R_j^2} \qquad j = 1,2,3,\ldots n$

$VIF_j=1$: geen correlatie
$VIF_j>5$: te sterke correlatie (ofwel $R_j^2 > 0{,}8$)
Betekenis: VIF_j is een factor in de onzekerheid van de schatting van de regressieparameter β_j

Determinatiecoëfficiënt

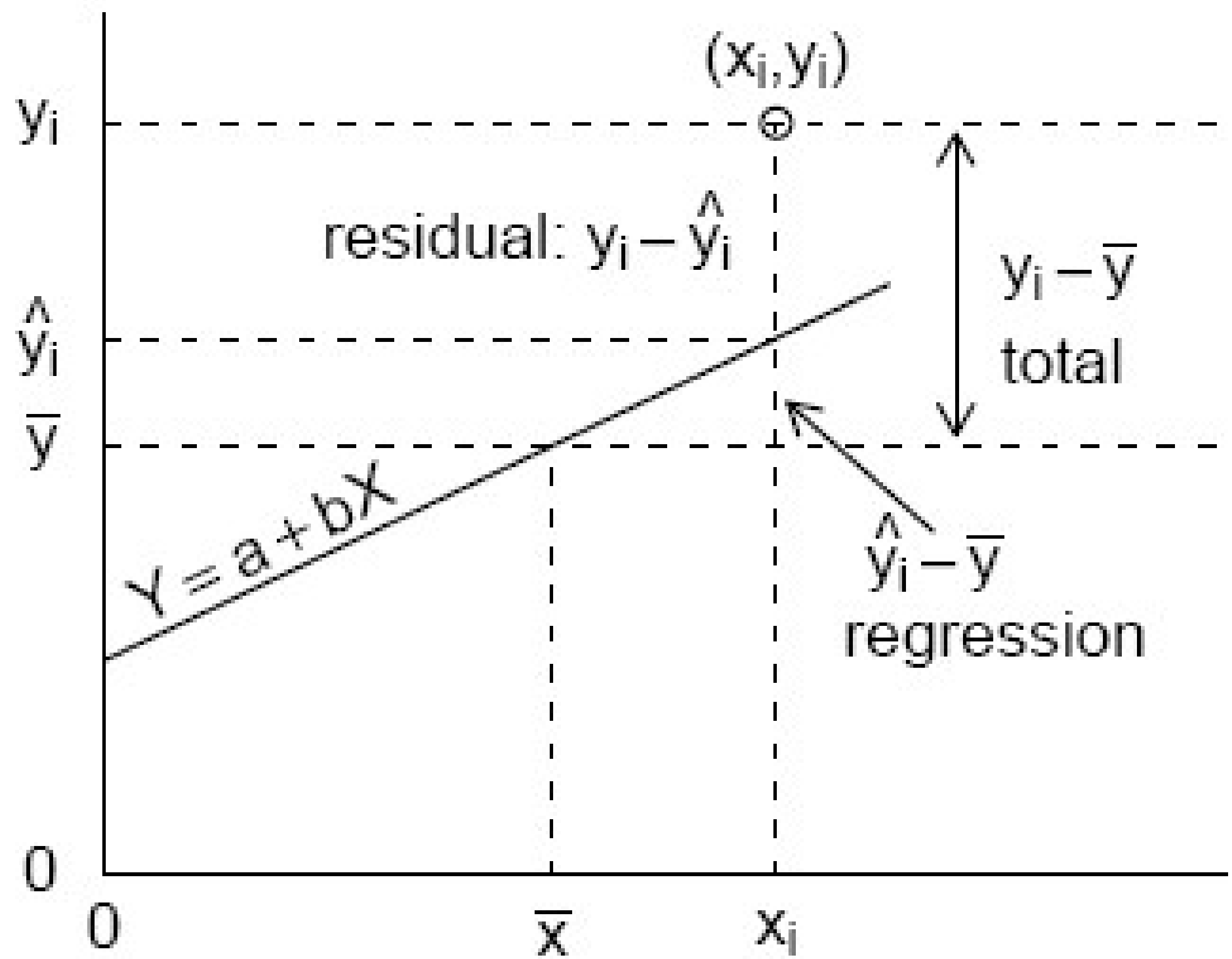
Illustratie ‘Multivariate statistiek, Regressie, Les 3’ (Excel)

De drie soorten verschillen

Waargenomen uitkomst t.o.v. gemiddelde	$Y_i - \bar{Y}$
Voorspelling t.o.v. gemiddelde	$\widehat{Y}_i - \bar{Y}$
Waargenomen uitkomst t.o.v. voorspelling	$Y_i - \widehat{Y}_i$

Er geldt: $Y_i - \bar{Y} = \left(\widehat{Y}_i - \bar{Y}\right) + \left(Y_i - \widehat{Y}_i\right)$

De drie soorten verschillen



Kwadraatsommen

Waargenomen uitkomst t.o.v. gemiddelde	$Y_i - \bar{Y}$
Voorspelling t.o.v. gemiddelde	$\widehat{Y}_i - \bar{Y}$
Waargenomen uitkomst t.o.v. voorspelling	$Y_i - \widehat{Y}_i$

Er geldt: $Y_i - \bar{Y} = \left(\widehat{Y}_i - \bar{Y}\right) + \left(Y_i - \widehat{Y}_i\right)$

Aangetoond kan worden dat, als de a en b in $Y = a + bX$ zijn bepaald met de kleinstekwadratenmethode, er geldt:

$$\sum (Y_i - \bar{Y})^2 = \sum \left(\widehat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \widehat{Y}_i\right)^2$$

Zie bijvoorbeeld Buijs, Statistiek om mee verder te werken, blz.138-139

SST=SSR+SSE

$$\sum (Y_i - \bar{Y})^2 = \sum \left(\widehat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \widehat{Y}_i\right)^2$$

kunnen we interpreteren als:

SST = SSR + SSE

waarin:

SST: kwadratensom totaal;

SSR: kwadratensom door regressie verklaard;

SSE: kwadratensom niet door regressie verklaard.

SSE is hetzelfde als $\sum {e_t}^2$, die m.b.v. de kleinstekwadratenmethode geminimaliseerd is

Verklaarde en onverklaarde variantie

$$\frac{\sum(Y_i - \bar{Y})^2}{n-1} = \frac{\sum\left(\widehat{Y}_i - \bar{Y}\right)^2}{n-1} + \frac{\sum\left(Y_i - \widehat{Y}_i\right)^2}{n-1}$$

Totale variantie = Verklaarde variantie + Onverklaarde variantie

Hoe beter de regressielijn, hoe groter de verklaarde variantie (en dus hoe kleiner de onverklaarde variantie).

De verhouding van verklaarde variantie t.o.v. de totale variantie is een maat voor de kwaliteit van de regressie; daarom de definitie van de determinatiecoëfficiënt:

Determinatiecoëfficiënt

Determinatiecoëfficiënt R^2 is:

$$\frac{verklaarde\ variantie}{totale\ variantie}$$

$$= 1 - \frac{onverklaarde\ variantie}{totale\ variantie}$$

$$= 1 - \frac{\sum (Y_i - \widehat{Y_i})^2}{\sum (Y_i - \bar{Y})^2}$$

Interpretatie van de determinatiecoëfficiënt

Voor de determinatiecoëfficiënt R^2 geldt:

$$0 \leq R^2 \leq 1$$

Hoe dichter R^2 bij 1, hoe beter de verklaring van de Y-waarden.

Bij enkelvoudige lineaire regressie is de determinatiecoëfficiënt het kwadraat van de correlatiecoëfficiënt r, dus $R^2 = r^2$.

Dummyvariabelen

Een dummyvariabele neemt de waarden 0 en 1 aan.

Te gebruiken bij ‘dichotome’ **onafhankelijke** variabelen: nominale variabelen met twee mogelijke waarden.

Bijvoorbeeld man/vrouw:	x=1: man;	x=0: vrouw
Of: roker/niet_roker:	x=1: roker;	x=0: niet-roker

Met dummyvariabelen kun je ‘gewoon’ rekenen:

Bijvoorbeeld $Y = 2.000 + 1.200X + 4.500D$, met:

Y=Inkomen, X=Leeftijd, D=Geslacht (man=1, vrouw=0).

Nominale variabelen met meer dan 2 niveaus

Bijvoorbeeld vooropleiding = havo/mbo/vwo

(aanname: precies 1 van de 3 geldt).

Coderen met twee dummy variabelen x1 en x2:

x1=0: havo nee x1=1: havo ja

x2=0: mbo nee x2=1: mbo ja

Voor vooropleiding vwo coderen we: x1=0 en x2=0

(en x1=1 en x2=1 kan niet voorkomen)

In het algemeen: bij een nominale variabele met n mogelijke waarden, zijn er $n-1$ dummyvariabelen nodig.

Logistische regressie

Als de **afhankelijke** variabele een 0/1 variabele is, dan gebruik je geen dummy-variabele, maar logistische regressie, bijvoorbeeld:

- wel (1) of niet (0) doorbranden van een onderdeel van een machine, afhankelijk van de tijd dat dit onderdeel verhit wordt.
- slagen (1) of zakken (0) voor een tentamen, afhankelijk van de bestede studietijd door de student.

- *Logistische regressie wordt later behandeld*

Niet-lineaire regressie; voorbeeld

De formule voor Y is niet-lineair, bijvoorbeeld :

$$\underline{Y} = \alpha \cdot \beta^X + \underline{\varepsilon}$$

$$\hat{Y} = \alpha \cdot \beta^X$$

Deze formule is **lineariseerbaar**:

$\ln Y = ln(\alpha \cdot \beta^X) = \ln\alpha + ln(\beta^X) = \ln\alpha + X \cdot ln(\beta)$

- De waarden van Y transformeren tot $\ln Y$
- Je bepaalt de lineaire regressievergelijking met als onafhankelijke variabele X en afhankelijke variabele $\ln Y$; dus de coëfficiënten ln (α) en ln (β)
- Daarmee bepaal je α en β.