Toelichting Assignment #2A-4: REGRESSIE MET DUMMY-VARIABELEN

De invloed van een nominale onafhankelijke variabele op een metrische afhankelijke variabele laar zich modelleren via een set dummy variabelen. Het regressiemodel is dan gelijk aan het conditionele-gemiddelden model dat we uit klassieke anova kennen. Dat conditionele gemiddelden model ziet er bijvoorbeeld als volgt uit:

```
compute incmid=incmid/1000.
means incmid by agecat7 / stat=anova linearity.
```

Report

incmid Household income in guilders

Agecat7 AGE in	Mean	N	Std. Deviation
57	3,21	315	1,538
62	2,77	386	1,637
67	2,50	374	1,617
72	2,30	369	1,336
77	2,31	339	1,207
82	2,25	295	1,393
87	2,06	174	1,150
Total	2,52	2252	1,485

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
incmid Household income in guilders * Agecat7 AGE in seven categories	Between Groups	(Combined)	265,032	6	44,172	21,108	,000
		Linearity	223,365	1	223,365	106,737	,000
		Deviation from Linea	41,668	5	8,334	3,982	,001
	Within Groups		4698,008	2245	2,093		
	Total		4963,040	2251			

Om hetzelfde in een regressiemodel onder te brengen construeren we een set ‘dummy-variabelen’::

```
recode agecat7 (57=1)(else=0) into age57.
recode agecat7 (62=1)(else=0) into age62.
recode agecat7 (67=1)(else=0) into age67.
recode agecat7 (72=1)(else=0) into age72.
recode agecat7 (77=1)(else=0) into age77.
recode agecat7 (82=1)(else=0) into age82.
recode agecat7 (87=1)(else=0) into age87.
```

Hier ontstaan zeven 0/1 variabelen. Elk van deze variabelen representeert één afzonderlijke categorie van AGECAT7. In tabelvorm ziet de codering er als volgt uit:

	Age57	Age62	Age67	Age72	Age77	Age82	Age87
57	1	0	0	0	0	0	0
62	0	1	0	0	0	0	0
67	0	0	1	0	0	0	0
72	0	0	0	1	0	0	0
77	0	0	0	0	1	0	0
82	0	0	0	0	0	1	0
87	0	0	0	0	0	0	1

Het regressiemodel ziet er nu als volgt uit:

INCMID = B0 + B1.Age57 + B2.Age62 … + B7.Age87

Dit model is echter niet geïdentificeerd, want het is logisch onmogelijk om waarden te berekenen voor alle acht coëfficiënten B0, B1, B2 .. B8. Die onmogelijkheid is snel in te zien als we ons realiseren dat de intercept B0 gedefinieerd is als de waarde van INCMIf wanneer alle variabelen Age57, Age62, .. Age87 tegelijk nul zijn. Dat punt komt niet voor.

Als we toch proberen om in spss het regressiemodel als volgt op te geven:

```
Regr /dep=INCMID  /enter=age57 to age87.
```

We bemerken dan dat in de schatting van de regressievergelijking een van de variabelen spontaan wegblijft onder de set voorspellers. We kunnen zelf bepalen welke variabele dat is, door zelf op te geven dat het de eerst of de laatste (het mag ook een andere zijn):

```
Regr /dep=INCMID  /enter=age62 to age87.
Regr /dep=INCMID  /enter=age57 to age82.
```

Hoe we het ook opgeven, de geschatte vergelijking geeft altijd dezelfde SS-verdeling en multipele correlatie. De geschatte regressiecoefficienten zien er evenwel wat anders uit.. Vergelijk de resultaten in de volgende tabel:

	Groeps-gemiddelden	Dummy-regressie #1	Dummy-regressie #2
Constante	2.52	2.064	3.210
57	3.21	1.146	0.
62	2.77	.705	-.441
67	2.50	.439	-.707
72	2.20	.233	-,913
77	2.31	.245	-.901
82	2.25	.197	-959
87	2.06	0	-1.146
SS-model	4963.04	4963.04	4963.04

Hoewel de coëfficiënten van deze drie modellen er op het eerste gezicht geheel verschillend uitzien, betekenen ze in feite precies hetzelfde. Bij dummy-regressie moeten de coëfficiënten als afwijkingen van de constante, en staat de constante voor de waarde van de weggelaten categorie.
Via dummy-variabelen kunnen we zelfs nominale waarden in regressiemodellen als onafhankelijke variabelen opnemen. Dat is handig, vooral als je naast die nominale variabele nog andere onafhankelijke variabelen hebt!