

Tentamen Onderzoeksvaardigheden 2

Onderdeel: Statistiek 2

11-4-2006, 14:00 – 17:00

NW&I

Docent: A.R.T. Donders

Het onderdeel Statistiek 2 heeft een openboek tentamen. Het tentamen bestaat uit 5 vragen. De vragen bestaan weer uit onderdelen. Voor ieder onderdeel staat het aantal punten vermeld. In totaal kun je 100 punten krijgen. Het aantal punten dat je verdient hebt wordt door 10 gedeeld om je eindcijfer te geven.

Schrijf op ieder vel dat je inlevert je naam en je studentnummer.

Lees de vraag zorgvuldig door!

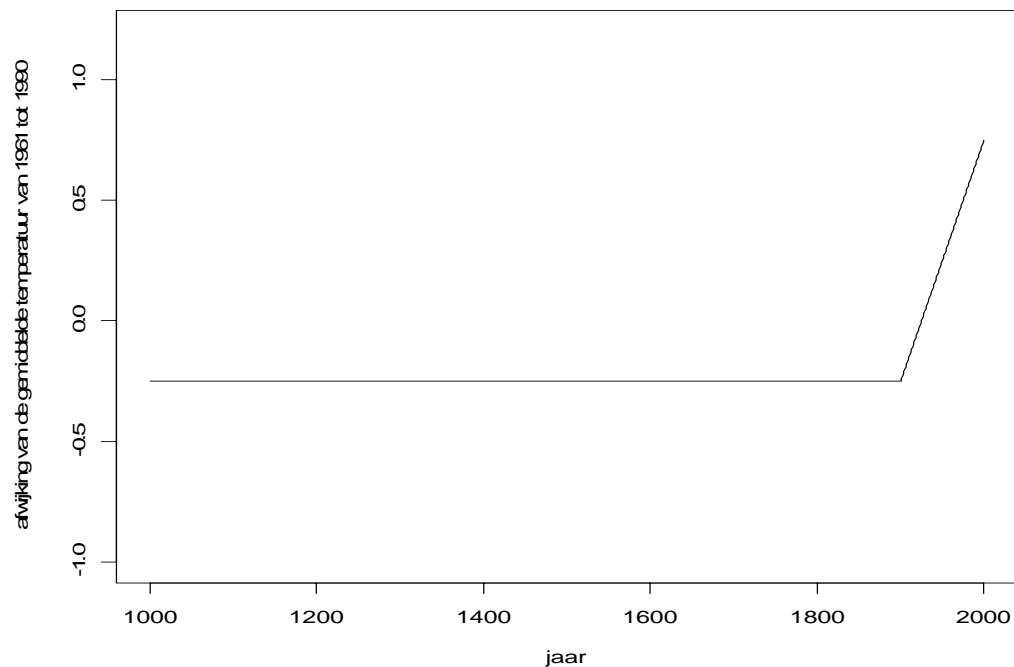
Laat goed zien hoe je aan je antwoorden komt en geef waar gevraagd een korte maar duidelijke argumentatie.

Bij de vragen staat regelmatig SPSS-output. Maak daar gebruik van!

Sommige vragen zijn makkelijker/moeilijker dan andere opgaven. Raak niet teveel tijd kwijt aan een moeilijke opgave! Mocht je niet zo snel uit een opgave komen, ga dan door met de volgende opgave en ga wanneer je alle opgaven gemaakt hebt, weer verder met deze moeilijke opgave.

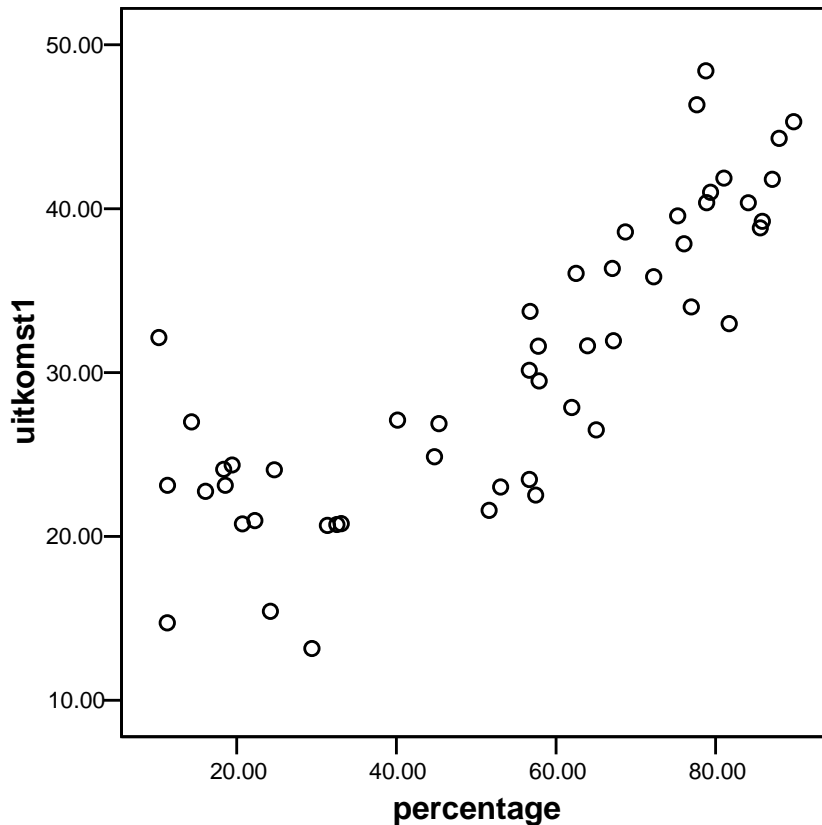
1. (30 punten)

Er is de afgelopen jaren veel te doen geweest om hockeysticks. Het ging dan niet om houten sportbenodigdheden, maar om een mogelijk bewijs voor de opwarming van de aarde als gevolg van menselijke activiteiten. Aan de hand van bepaalde gegevens die een idee geven over de temperatuur (boomringen, koraalgroei, ijsboringen en historische gegevens) had een onderzoeker (Michael Mann) geconcludeerd dat er sprake was van een veel sterkere stijging in de gemiddelde temperatuur gedurende de afgelopen 100 jaar dan ervoor; voor die tijd was er nauwelijks sprake van een stijging in de gemiddelde temperatuur terwijl na 1900 er een duidelijke stijging te zien is. Zo'n figuur bevat natuurlijk heel veel schommelingen, maar het globale plaatje ziet er als volgt uit:



Met wat fantasie kun je daar wel een (ijs)hockeystick in zien. We gaan ons hier niet met deze discussie bezighouden, maar we gaan wel kijken naar situaties waarin we verwachten dat er niet sprake is van een regressielijn maar van een lijn met een 'knik' erin. We gaan er hierbij vanuit dat we weten waar we de knik kunnen verwachten. Natuurlijk is het 'mooier' wanneer je het knikpunt zou kunnen schatten en dat is ook wel mogelijk maar het probleem wordt dan veel ingewikkelder en is niet meer met lineaire regressie op te lossen.

Stel dat er sprake is van een proces waarbij er tot een bepaalde waarde van een voorspeller nauwelijks invloed is op de uitkomst, maar er na die waarde wel degelijk een effect is. Je kan hierbij denken aan de relatie tussen de hoeveelheid cholesterol en de ernst van bloedvatwandbeschadigingen (tot een zeker niveau is cholesterol niet erg of zelf beschermend, na dat niveau is het een risico), of de hoeveelheid vervuiling en de schade aan een ecosysteem (de meeste systemen kunnen een bepaalde hoeveelheid vervuiling opvangen en verwerken, maar boven dat niveau werkt de opvang niet meer), of de hoeveelheid kenniswerkers en het dynamisch vermogen van een bedrijf (weinig kenniswerkers krijgen niets voor elkaar, maar boven een bepaalde grens is er wel een effect waarbij geldt hoe meer hoe beter). Om het wat algemener te houden geven we de voorspeller aan met PERCENTAGE en de afhankelijke variabele met UITKOMST1 en we weten dat er boven een PERCENTAGE van 40 andere effecten te verwachten zijn. Bij 50 cases zijn de variabelen PERCENTAGE en UITKOMST1 gemeten. De resultaten staan in de volgende figuur:



Hopelijk zie je aan deze figuur dat het niet echt verstandig is een eenvoudig regressiemodel te gebruiken. Zonder enige theoretische kennis zou je misschien een kwadratische term toevoegen, maar omdat we weten dat er een omslagpunt bij 40 ligt gaan we het hier anders aanpakken. De eerste benadering is om simpelweg twee regressielijnen te fitten: één lijn op alle data waarvoor geldt dat PERCENTAGE kleiner is dan (of gelijk is aan) 40 en één lijn op alle data waarvoor PERCENTAGE groter is dan 40.

Hieronder staan stukken output voor deze twee lijnen (BOVEN40 is gelijk aan 1 wanneer PERCENTAGE groter is dan 40; BOVEN40 is gelijk aan 0 wanneer PERCENTAGE kleiner of gelijk is aan 40)

Regression boven40 = .00

Model Summary^{b,c}

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.432 ^a	.186	.128	4.34933

- a. Predictors: (Constant), percentage
- b. Dependent Variable: uitkomst1
- c. boven40 = .00

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	27.309	3.295		8.288	.000
	percentage	-.264	.147	-.432	-1.790	.095

- a. Dependent Variable: uitkomst1
- b. boven40 = .00

boven40 = 1.00

Model Summary^{b,c}

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.835 ^a	.697	.688	4.14440

- a. Predictors: (Constant), percentage
- b. Dependent Variable: uitkomst1
- c. boven40 = 1.00

Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.612	3.664		.986	.332
	percentage	.449	.052	.835	8.580	.000

a. Dependent Variable: uitkomst1

b. boven40 = 1.00

- (5 punten) Stel de twee regressielijnen op (geef goed aan welke lijn bij welke situatie hoort).
- (5 punten) Is er sprake van een positief, een negatief of geen verband tussen PERCENTAGE en UITKOMST1 voor het geval dat PERCENTAGE > 40. En hoe is het verband als PERCENTAGE ≤ 40?
- (5 punten) Theoretisch zou er sprake moeten zijn van een continue functie: het effect van de variabele PERCENTAGE zou bij een kleine verandering niet sprongsgewijze groter of kleiner moeten zijn. Is daarvan bij de benadering (het fitten van twee aparte regressielijnen) sprake? Motiveer je antwoord goed. HINT: het is handig om de continuïteit te onderzoeken bij dat punt waar mogelijk problemen te verwachten zijn.

Een andere benadering is het fitten van het volgende model (wel het broken stick (letterlijk: “gebroken stok”) regressiemodel genoemd).

Er worden twee nieuwe variabelen gedefinieerd:

$$\text{percond40} = \begin{cases} 40 - \text{PERCENTAGE}, & \text{als PERCENTAGE} \leq 40 \\ 0, & \text{als PERCENTAGE} > 40 \end{cases}$$

$$\text{percbov40s} = \begin{cases} 0, & \text{als PERCENTAGE} \leq 40 \\ \text{PERCENTAGE} - 40, & \text{als PERCENTAGE} > 40 \end{cases}$$

Vervolgens worden deze twee variabelen samen met een intercept in één regressiemodel met als afhankelijke variabele UITKOMST1 gestopt. Delen van de output staan hieronder.

Regression

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	percbov40 ^a percond40	.	Enter

a. All requested variables entered.

b. Dependent Variable: uitkomst1

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.884 ^a	.782	.772	4.25597

a. Predictors: (Constant), percbov40, percond40

b. Dependent Variable: uitkomst1

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3045.799	2	1522.900	84.076	.000 ^a
	Residual	851.326	47	18.113		
	Total	3897.125	49			

a. Predictors: (Constant), percbov40, percond40

b. Dependent Variable: uitkomst1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20.364	1.472		13.835	.000
	percond40	.098	.086	.108	1.148	.257
	percbov40	.484	.048	.956	10.112	.000

a. Dependent Variable: uitkomst1

NB Ook al heb je bovenstaande misschien niet helemaal kunnen volgen, een gedeelte van de onderstaande vragen is ook zonder dat inzicht te beantwoorden.

d. (5 punten) Hoe ziet de regressievergelijking eruit (neem alle termen op)?

e. (5 punten) Wat is de op basis van dit model voorspelde waarde voor

UITKOMST1 voor een PERCENTAGE van 40?

- f. (5 punten) Stel je zou een proces hebben waarbij je 2 omslagpunten verwacht, bijvoorbeeld bij PERCENTAGES van 45 en 75. Hoe zou je dan deze methode kunnen toepassen?

2. (25 punten)

Bij de bepaling van de temperaturen voor de hockeystick werd gebruik gemaakt van verschillende type data. De vraag kan dan zijn, in welke mate stemmen de uitkomsten behaald met de verschillende methoden (aangegeven met METHODE) , bijvoorbeeld boomringen, koraal en ijskernen, met elkaar overeen.? Daartoe zou je de volgende opzet kunnen gebruiken: voor 30 (redelijk) willekeurig gekozen jaartallen (aangegeven met JAARTAL waar nodig) waarvan in voldoende mate de volgende drie gegevens na te gaan zijn bepalen we per jaar de met behulp van die methode voorspelde temperatuur (aangegeven met TEMPERATUUR). De uitkomsten zijn ingevoerd in SPSS en zijn geanalyseerd. Hieronder staan stukken van de output. (NB alle getallen zijn fictief en we gaan er zondermeer vanuit dat observaties tussen jaren onafhankelijk van elkaar zijn omdat de jaartallen ver genoeg uit elkaar liggen).

Bij de eerste analyse op de afhankelijke variabele TEMPERATUUR wordt in een variantieanalyse alleen de onafhankelijke METHODE gebruikt. Hieronder staan delen van de output

Oneway

ANOVA

temperatuur					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.235	2	.118	.497	.610
Within Groups	20.612	87	.237		
Total	20.847	89			

- a. (5 punten) Is er sprake van een significant verschil tussen de drie verschillende methoden? Motiveer je antwoord.

Vervolgens is ook de (random Block) factor jaartal in het design opgenomen. Hieronder staan delen van de output:

Univariate Analysis of Variance

Tests of Between-Subjects Effects

Dependent Variable: temperatuur

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	
Intercept	Hypothesis	.002	1	.002	.003	.958
	Error	18.923	29	.653 ^a		
methode	Hypothesis	.235	2	.118	4.042	.023
	Error	1.689	58	.029 ^b		
jaartal	Hypothesis	18.923	29	.653	22.406	.000
	Error	1.689	58	.029 ^b		

a. MS(jaartal)

b. MS(Error)

Post Hoc Tests methode

Multiple Comparisons

Dependent Variable: temperatuur

	(I) methode	(J) methode	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	.00	1.00	-.0789	.04406	.210	-.1896	.0318
		2.00	-.1237*	.04406	.025	-.2344	-.0130
	1.00	.00	.0789	.04406	.210	-.0318	.1896
		2.00	-.0449	.04406	.598	-.1556	.0658
LSD	2.00	.00	.1237*	.04406	.025	.0130	.2344
		1.00	-.0449	.04406	.598	-.0658	.1556
	.00	1.00	-.0789	.04406	.079	-.1671	.0093
		2.00	-.1237*	.04406	.007	-.2119	-.0355
	1.00	.00	.0789	.04406	.079	-.0093	.1671
		2.00	-.0449	.04406	.313	-.1331	.0433
	2.00	.00	.1237*	.04406	.007	.0355	.2119
		1.00	-.0449	.04406	.313	-.0433	.1331

Based on observed means.

*. The mean difference is significant at the .05 level.

- b. (5 punten) Is er nu sprake van een significant verschil tussen de methodes?
Motiveer je antwoord!
- c. (5 punten) Is in de huidige opzet METHODE een fixed of een random factor? En hoe zit dat met JAARTAL? Motiveer je antwoord!

- d. (5 punten) Bij de Post Hoc Tests worden de methodes met elkaar vergeleken.

Waarom zijn de betrouwbaarheidsintervallen bij Scheffé's methode breder dan de betrouwbaarheidsintervallen bij de 'gewone' t-testen (aangeven met LSD).

Op zich zullen er natuurlijk altijd wel verschillen tussen de methodes bestaan en zodra je voldoende jaren zou nemen zouden zelfs zeer betrouwbare methodes kleine maar significante verschillen laten zien. Het is eerder interessant om te weten in welke mate verschillen tussen de temperatuurmetingen veroorzaakt worden door meetfouten (error) en/of door de verschillende methodes en in welke mate de verschillen tussen de metingen veroorzaakt worden door datgene dat je wilt bepalen: verschillen tussen de jaartallen. Dit wordt wel uitgedrukt in een zogenaamde Intraclass Correlation Coefficient (ICC). Er zijn verschillende ICC's te definiëren maar voor allemaal geldt dat we eerst schattingen moeten hebben van de mate waarin verschillen tussen de metingen toe te schrijven zijn aan jaartal, de methodes en aan meetfouten (error); in SPSS worden dit wel variantiecomponenten genoemd. In onderstaande tabel staan de schattingen:

Variance Estimates

Component	Estimate
Var(jaartal)	.208
Var(methode)	.003
Var(Error)	.029

Dependent Variable: temperatuur

Method: Restricted Maximum Likelihood Estimation

- e. (5 punten) Geef een schatting van de mate waarin verschillen tussen de metingen bepaald worden door verschillen in temperatuur tussen de jaartallen. Geef een goede motivatie voor je schatter.

3. (10 punten)

Om na te gaan of een bepaalde methode een goede indicator voor de temperatuur kan zijn, worden gedurende 3 jaar per seizoen de uitkomst (aangegeven met Y_t) bepaald. We gaan in deze opgave een decompositie op deze tijdreeks uitvoeren. Er is al wat voorwerk verricht maar bepaalde getallen zijn (vanwege een onleesbaar handschrift) verloren gegaan. Vul deze getallen op het antwoordformulier in en geef een duidelijke motivatie. Het volgende is ook nog gegeven:

De op basis van G_t geschatte trendlijn is: $Y_t = 8.961 + 6.316 \times t$

t	tijd	L	Y_t	U_t	$S_t * O_t$	S_t	G_t	T_t	$C_t * O_t$	C_t	O_t
1	2001	1	22			0.801	27.466				
2	2001	2	34			2.575	13.204	21.593	0.611		
3	2001	3	10	18.500				27.909	0.803	0.854	0.940
4	2001	4	7	29.750	0.235	0.178	39.326	34.225	1.149	0.897	1.281
5	2002	1	24	41.875	0.573	0.801	29.963	40.541	0.739	0.966	0.765
6	2002	2	122	43.250	2.821	2.575	47.379	46.857	1.011	0.850	1.189
7	2002	3	19	51.125	0.372			53.173	0.801	0.887	0.903
8	2002	4	9	68.500	0.131	0.178	50.562	59.489	0.850	1.088	0.781
9	2003	1	85	79.500	1.069	0.801	106.117	65.805	1.613	1.180	1.367
10	2003	2	200		2.458	2.575	77.670	72.121	1.077	1.173	0.918
11	2003	3	29					78.437	0.829	0.945	0.877
12	2003	4	14			0.178	78.652		0.928		

4.(10 punten)

Het ligt bij de methodes om de temperatuur te bepalen voor de hand dat de observaties van opeenvolgende jaren met elkaar samenhangen. Om dit te modeleren wordt gebruik gemaakt van een ARIMA model. Dit wordt gefit op een reeks van 150 aangesloten jaren waarvoor gegevens van één methode (de boomringen) aanwezig zijn (per jaar 1 meting).

- (3 punten) Zou je gebruik kunnen maken van een ARIMA model met seizoensinvloeden (seizoensdifferencing of seizoensAR/MA-componenten)?
Waarom wel/niet?
- (2 punten) Er is sprake van een stationaire tijdreeks. Hoe kun je onderzoeken wat de orde van de AR en MA componenten moet zijn?

Het bleek dat een ARIMA(1,0,1) model het beste bij de data paste. Hieronder staan stukken van de output

Parameters:

```
AR1      _____ < value originating from estimation >
MA1      _____ < value originating from estimation >
CONSTANT _____ < value originating from estimation >
```

95.00 percent confidence intervals will be generated.

Conclusion of estimation phase.

Estimation terminated at iteration number 3 because:

Sum of squares decreased by less than .001 percent.

FINAL PARAMETERS:

```
Number of residuals 150
Standard error      .95702155
Log likelihood      -205.99024
AIC                 417.98048
SBC                 427.01239
```

Analysis of Variance:

	DF	Adj. Sum of Squares	Residual Variance
Residuals	147	136.90521	.91589024

Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	.62587826	.06616031	9.460026	.00000000
MA1	-.83093215	.04882905	-17.017169	.00000000
CONSTANT	.17384253	.37710604	.460991	.64548630

c. (5 punten) Hoe ziet het eindmodel eruit? Geef de vergelijking.

5. (25 punten)

Ten slotte nog een vraag over iets anders. Regressieanalyse (en vooral ook variantieanalyse) kan gebruikt worden om bepaalde theorieën te testen. Je kan regressieanalyse echter ook gebruiken om simpelweg voorspellingen te maken. Bij het maken van voorspellingen spelen andere dingen dan bij het testen van theorieën. We zullen een aantal van dit soort dingen in deze opgave aanbod laten komen.

Stel we willen nu eens niet theoretisch verklaren waarom sommige innovaties het wel redden en andere niet (of waarom een bepaalde plant wel in het ene gebiedje voorkomt en niet in het andere), maar we willen simpelweg voorspellen welke innovaties het goed doen (of waar we de plant kunnen vinden). Voor 250 geïntroduceerde innovaties (of 250 gebiedjes) wordt nagegaan of ze al dan niet geslaagd zijn (of er al dan niet de plant voorkomt). Dit wordt weergegeven in de variabele UITKOMST (met '0' is 'nee' en '1' is 'ja'). We hebben een set van 10 potentiële voorspellers (aangegeven met Pred1 tot en met Pred10). Via een backward selectie (op basis van de likelihoodratio) selecteren we het best passende model. Dan blijkt het bestpassende model een model te zijn met Pred1, Pred2 en Pred4. Pred1 en Pred 4 zijn discrete (categorische) variabelen. Pred1 heeft twee verschillende uitkomsten (aangegeven 0 en 1) en Pred4 heeft drie verschillende uitkomsten (aangegeven met 0, 1 en 2); Pred 2 is een continue variabele. Delen van de output van het eindmodel staan hieronder:

Logistic Regression

Categorical Variables Codings

	Frequency	Parameter coding	
		(1)	(2)
pred4 .00	97	1.000	.000
1.00	62	.000	1.000
2.00	91	.000	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step 1 Step	86.327	4	.000
Block	86.327	4	.000
Model	86.327	4	.000

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step ^a 1 pred1	1.451	.416	12.170	1	.000	4.267
pred2	.923	.162	32.354	1	.000	2.517
pred4			39.043	2	.000	
pred4(1)	-.270	.399	.456	1	.500	.764
pred4(2)	-2.671	.467	32.722	1	.000	.069
Constant	.178	.366	.238	1	.626	1.195

a. Variable(s) entered on step 1: pred1, pred2, pred4.

- (5 punten) Levert de variabele pred4 echt wel een significante bijdrage aan de voorspelling van de uitkomst? Motiveer je antwoord!
- (10 punten) Wat is de voorspelde kans voor een innovatie (gebied) met als waarde voor Pred1 van 0, een waarde voor Pred2 van 1.5 en een waarde voor Pred4 van 1.
- (5 punten) Bij het opstellen van het regressiemodel moet je een keuze maken voor de p-waarde die je gebruikt als grens voor het al dan niet verwijderen van een voorspeller. Er is in dit geval gebruikt gemaakt van een p-waarde van 0.10 in plaats van 0.05. Dit is gebruikelijk wanneer je voorspelregels wilt afleiden (er worden zelfs wel p-waardes van 0.15 of 0.20 gebruikt). Wat is het (gewenste) gevolg van deze verhoging. Motiveer je antwoord!
- (5 punten) Stel dat we nog veel meer predictoren zouden hebben gehad, b.v. geen 10 maar 100, en dat we dan ook op dit eindmodel terecht waren gekomen. Heb je dan meer vertrouwen in het eindresultaat of juist minder? Motiveer je antwoord!