

OPLOSSINGEN BIJ HOOFDSTUK 10

1. Volgende regressievergelijking werd opgesteld na onderzoek:

$$Y_i = 6 + 2.5 * X_{i1} + 3 * X_{i2} + \varepsilon_i$$

Bereken de voorspelde score op variabele Y voor een case die een score 4 haalt op variabele X1 en score 6 op variabele X2.

Deze opdracht lossen we eenvoudig op door in de vergelijking X1 en X2 te vervangen door de geobserveerde waarden van deze variabelen:

$$Y_i = 6 + 2.5 * 4 + 3 * 6 + \varepsilon_i$$

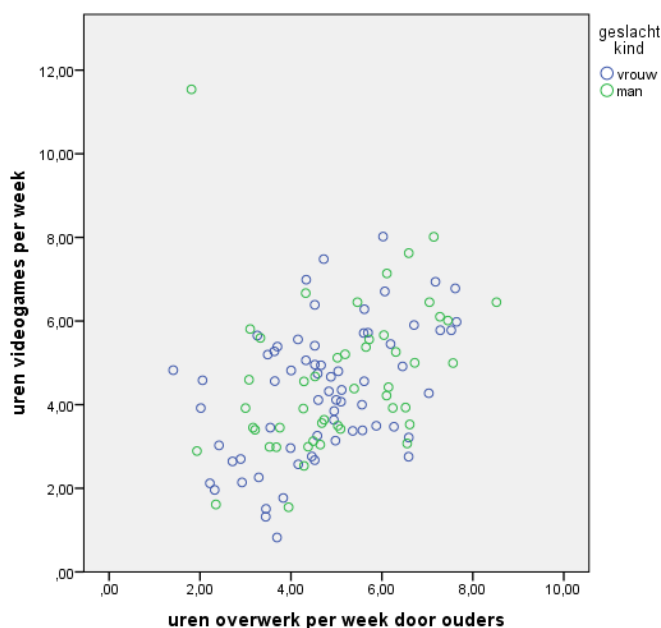
$$Y_i = 6 + 10 + 18 + \varepsilon_i$$

$$Y_i = 34 + \varepsilon_i$$

De voorspelde score op variabele Y is dus 34, plus of min een onbekende fout die inherent is aan de regressiemethode.

2. We proberen te voorspellen hoeveel uren kinderen per week spenderen aan videogames aan de hand van het aantal uren dat hun ouders overwerken. Ga na of dit een zinvolle predictie is. Stel in dat geval ook het regressiemodel op. Je vindt de data in het bestand *opdr_overuren-video.sav*.

In het bestand vinden we vier variabelen, waarvan we eigenlijk enkel de variabelen overwerk en video nodig hebben. Om te beginnen maken we een scatterplot om onze data te verkennen: Graphs > Chart Builder, Scatter/Dot slepen naar het werkveld en de beide variabelen naar de twee assen slepen. Het maakt niet uit welke variabele je waar zet, al is de conventie min of meer dat je de predictor op de horizontale as zet en de criteriumvariabele op de verticale as. Eventueel, maar niet noodzakelijk kan je ook via Groups/point ID Grouping variable aanvinken en dan Geslacht naar het veld Set Color slepen. Je krijgt dan dit:

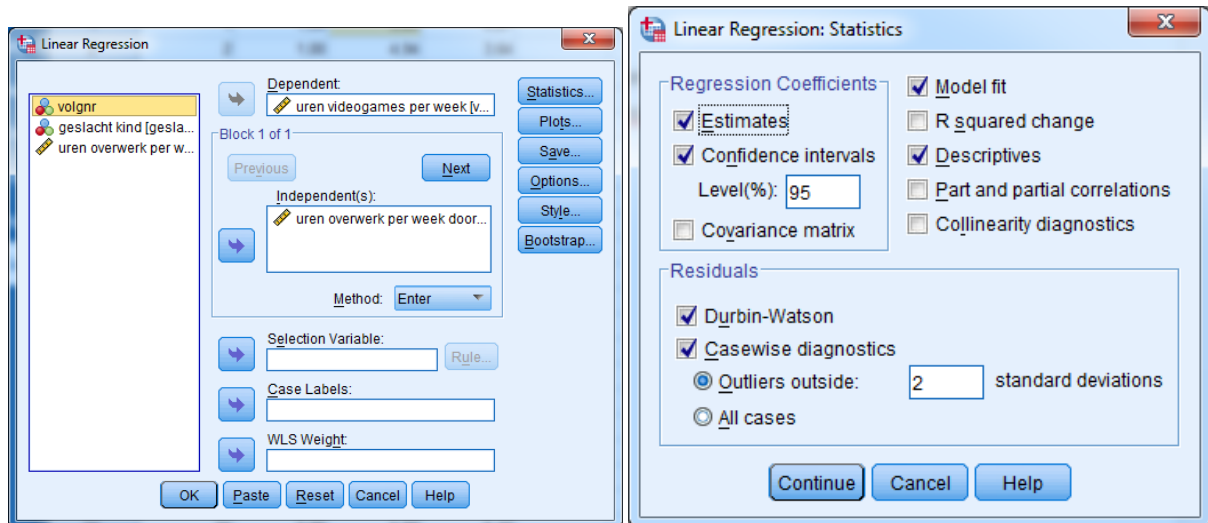


De puntenwolk is vrij breed, al lijkt er wel enige richting in te zitten. We zien ook linksboven een eenzame case, die we verderop in het oog zullen houden in het kader van storende outliers.

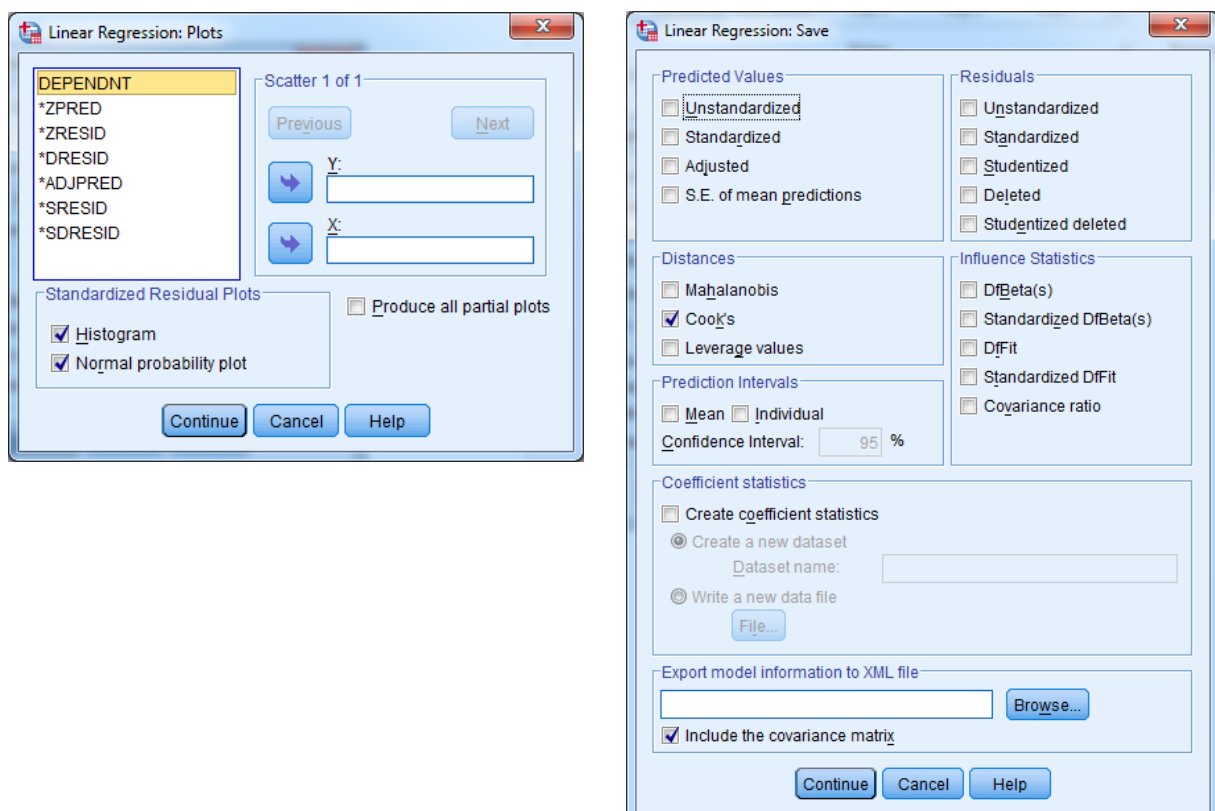
Vervolgens starten we de regressieanalyse:

Analyze > Regression > Linear

Het algemene venster en het venster Statistics:



De vensters Plots en Save:



Klik dan op ok en aanschouw de output:

In de tabel Correlations zien we dat er een correlatie van .375 bestaat tussen beide variabelen. Dat geeft ons alvast het idee dat beide variabelen wel degelijk iets met elkaar te maken hebben. Diezelfde correlatie vinden we ook terug in de tabel Model Summary, waar we ook de R^2 van .14 terugvinden. We zien hierin bovendien dat de Durbin-Watson statistiek 2.22 bedraagt, wat ons niet hoeft te verontrusten in verband met autocorrelatie tussen de residuen.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,375 ^a	,140	,133	1,56433	2,221

a. Predictors: (Constant), uren overwerk per week door ouders

b. Dependent Variable: uren videogames per week

De ANOVA tabel geeft aan dat ons model significant is: $p = .000$ of $p < .001$

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45,133	1	45,133	18,443	,000 ^b
	Residual	276,525	113	2,447		
	Total	321,658	114			

a. Dependent Variable: uren videogames per week

b. Predictors: (Constant), uren overwerk per week door ouders

In de Coefficients tabel stellen we vast dat de predictor overwerk significant is, en dat de B-waarde, die we nodig hebben om de regressievergelijking op te stellen, gelijk is aan .41.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,453	,489		5,017	,000	1,484	3,421
	uren overwerk per week door ouders	,413	,096	,375	4,295	,000	,223	,604

a. Dependent Variable: uren videogames per week

Om de voorwaarden na te gaan, kijken we nog even naar de tabel Casewise Diagnostics. Daarin zien we dat case 70 een erg hoge Std. Residual krijgt. Als we in de data view gaan kijken, stellen we vast dat dit de case is die we ook in de scatterplot al hadden gedetecteerd als een behoorlijk afwijkende case. We gaan ervan uit dat er bij de meting van deze case iets is foutgelopen, of dat deze case gewoon niet representatief is voor de populatie te we bestuderen. We zullen daarom de case verwijderen en de analyse opnieuw uitvoeren.

In de nieuwe analyse zien we dat de correlatie is toegenomen en dat het model uiteraard nog steeds significant is:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,498 ^a	,248	,241	1,34903	2,083

- a. Predictors: (Constant), uren overwerk per week door ouders
b. Dependent Variable: uren videogames per week

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	67,220	1	67,220	36,936	,000 ^b
	Residual	203,826	112	1,820		
	Total	271,046	113			

- a. Dependent Variable: uren videogames per week
b. Predictors: (Constant), uren overwerk per week door ouders

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1,891	,431		4,389	,000	1,037	2,745
	uren overwerk per week door ouders	,514	,085	,498	6,078	,000	,346	,681

- a. Dependent Variable: uren videogames per week

We zien verder geen eigenaardigheden in de tabellen Casewise Diagnostics (alle Std. Residuals in absolute waarde kleiner dan 3) en Residuals Statistics (Cook's distance niet groter dan 1) meer.

Ook de opgevraagde plots vertonen geen eigenaardige kronkels, dus we kunnen rapporteren:

Een enkelvoudige regressieanalyse werd uitgevoerd met aantal uren videogames als criterium en het aantal uren overwerk als predictor. Dit model bleek significant, met $R^2 = .25$, $F = 36.94$, $p < .001$. Tabel 1 geeft een overzicht van de coëfficiënten. Deze analyse wijst uit dat er een positief verband bestaat tussen de bestudeerde variabelen. De verklaarde variantie is beperkt, wat suggereert dat er andere factoren zijn die het aantal uren videogames bepalen.

Tabel 1: Resultaten enkelvoudige regressie met fuifsatisfactie als criterium en aantal alcoholische consumpties als predictor.

	<i>B</i>	<i>SE B</i>	<i>β</i>	<i>t</i>
constante	1.89	.43		4.39***
aantal overuren	.51	.09	.50	6.08***

*** $p < .001$

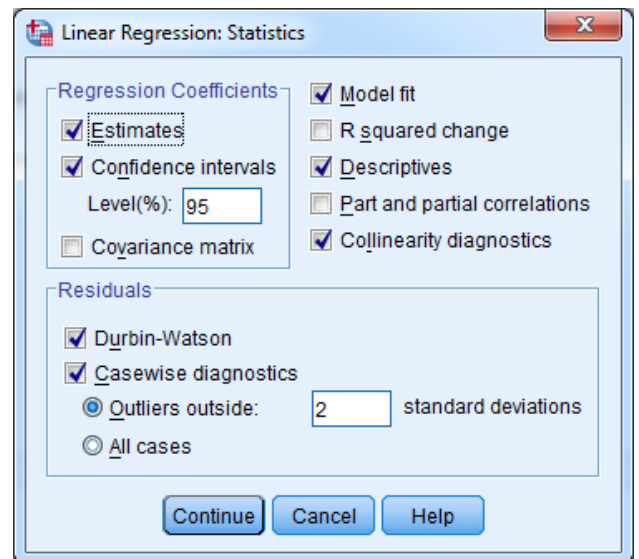
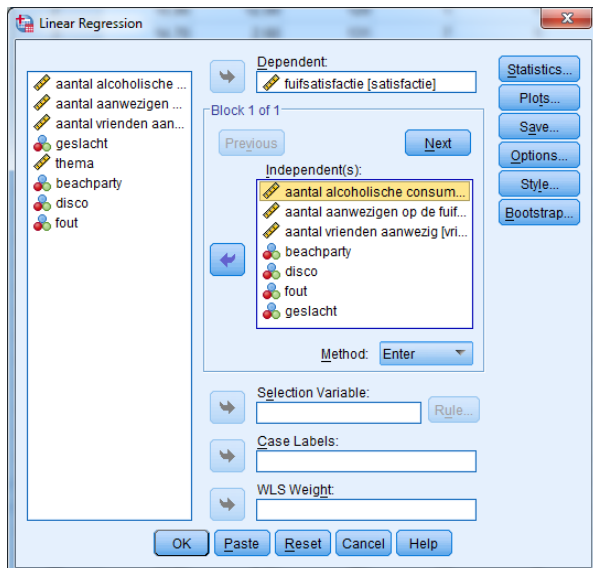
Het regressiemodel stellen we eenvoudig op door de gevonden coëfficiënten in de vergelijking te plaatsen:

$$\text{uren videogames}_i = 1.89 + .51 * \text{aantal overuren}_i + \varepsilon_i$$

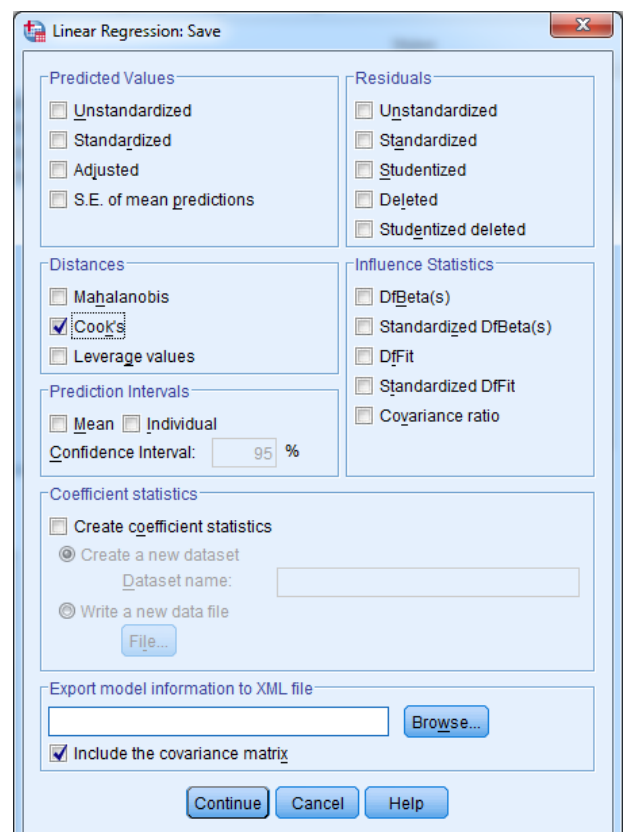
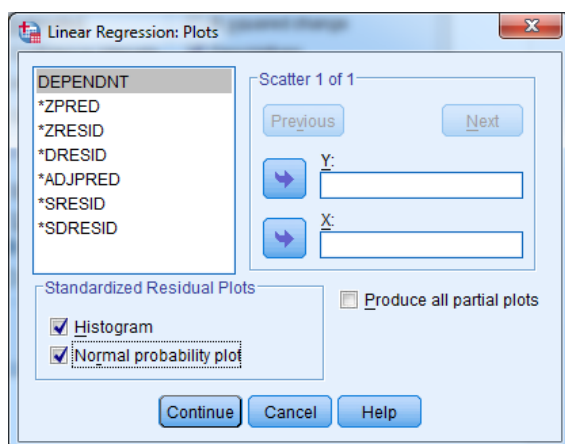
- Voer de regressieanalyse uit dit hoofdstuk opnieuw uit, maar betrek er ook de twee extra predictoren uit paragraaf 10.3 bij. Je kan deze vinden in het bestand *opdr_fuifevaluatie_regr.sav*. Volg de instructies om dummy variabelen toe te voegen aan de analyse en bereken de regressieanalyse. Ga na of er voldaan is aan alle voorwaarden en bekijk welke predictoren kunnen worden behouden. Stel ook het definitieve regressiemodel op.

In deze opdracht wordt gevraagd om de nominale variabelen geslacht en thema meen in de analyse te brengen. Voor geslacht is dat geen probleem, aangezien deze variabele slechts twee niveaus heeft (man – vrouw). De variabele thema heeft daarentegen vier niveaus (beach, disco, fout, geen), wat betekent dat we deze variabele zullen moeten opsplitsen in verschillende dummy variabelen. Aangezien er vier niveaus zijn, zullen we drie dummy variabelen nodig hebben. In dit geval lijkt het logisch om het niveau “geen thema” als basislijn te hanteren en de andere drie niveaus als dummy te creëren. In paragraaf 10.3 van het boek staat netjes uitgelegd hoe je via Recode deze dummy’s kan aanmaken.

Als de dummy variabelen zijn gemaakt, dan kunnen we aan de slag met de eigenlijke analyse: kies Analyze > Regression > Linear om te starten. Vul de vensters Linear Regression en Statistics aan zoals in de figuren:



Achter de knopjes Plots en Save duiden we deze opties aan:



De Model Summary en de ANOVA tabel geven aan dat het model 52% van de variantie kan verklaren en dat dit model significant is. De Durbin Watson waarde ligt tussen 1 en 3, wat suggereert dat er geen probleem is met autocorrelatie van de residuen.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,720 ^a	,518	,453	1,18797	2,234

- a. Predictors: (Constant), geslacht, aantal alcoholische consumpties, aantal aanwezigen op de fuif, fout, disco, aantal vrienden aanwezig, beachparty
- b. Dependent Variable: fuifsatisfactie

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	78,797	7	11,257	7,976	,000 ^b
	Residual	73,386	52	1,411		
	Total	152,183	59			

- a. Dependent Variable: fuifsatisfactie
- b. Predictors: (Constant), geslacht, aantal alcoholische consumpties, aantal aanwezigen op de fuif, fout, disco, aantal vrienden aanwezig, beachparty

De tabel Coefficients toont ons welke predictoren een significante bijdrage leveren aan het model. Een eerste vaststelling is dat het aantal aanwezigen geen significante predictor blijkt. Dat is niet verrassend – we hebben dat immers in de het voorbeeld in paragraaf 10.2 ook gezien. Van de nieuwe dummy variabelen blijkt enkel de dummy Beachparty significant. De interpretatie hiervan is evenwel een belangrijk aandachtspunt: we hebben

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	10,156	,927		10,955	,000	8,295	12,016		
	aantal alcoholische consumpties	-,176	,053	-,349	-3,338	,002	-,282	-,070	,847	1,181
	aantal aanwezigen op de fuif	,005	,003	,151	1,494	,141	-,002	,012	,908	1,101
	aantal vrienden aanwezig beachparty	,294	,053	,624	5,514	,000	,187	,401	,724	1,381
	disco	,941	,454	,266	2,074	,043	,031	1,852	,562	1,779
	fout	,501	,448	,130	1,120	,268	-,397	1,399	,692	1,445
	geslacht	-,153	,465	-,039	-,330	,743	-1,087	,780	,679	1,472
		,136	,331	,041	,410	,683	-,528	,799	,911	1,098

- a. Dependent Variable: fuifsatisfactie

bij het aanmaken van de dummy's geopteerd om de categorie "geen thema" als basis te gebruiken. De huidige significantie betekent dan ook dat het thema beachparty zorgt voor meer satisfactie in vergelijking met het ontbreken van een thema. Verder leiden we af dat de thema's disco en fout geen satisfactie verhogen in vergelijking met het ontbreken van een thema.

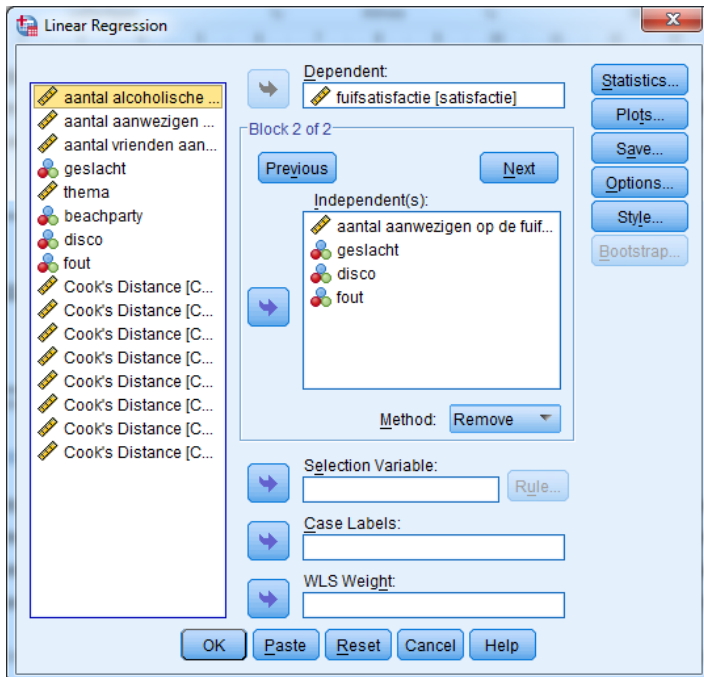
Wat betreft de Colinearity Statistics in de Coefficients tabel zien we geen grote problemen: de waarden van de Tolerance zijn allemaal groter dan .20, en de waarden van de VIF zijn niet groter dan 5.

In de tabel Collinearity Diagnostics stellen we vast dat de predictor aantal aanwezigen een hoge lading krijgt op meer dan één dimensie. Dat geldt ook voor de predictoren disco en fout. In combinatie met de t-toetsen

kunnen we besluiten dat deze predictoren niet erg nuttig zijn voor onze analyse, en dat we deze beter kunnen verwijderen uit het model.

We controleren ten slotte nog de tabellen Casewise Diagnostics en Residuals Statistics op tekenen van eventuele outliers, maar we zien daar geen verontrustende zaken.

De volgende stap is dan het verwijderen van de minder interessante predictoren en de analyse opnieuw uitvoeren om de verschillen te bekijken. We hoeven eigenlijk simpelweg opnieuw het dialoogvenster Linear Regression te openen en een block toe te voegen. Alle andere opties blijven immers behouden uit de vorige analyse. Klik dus op het knopje Next, sleep de te verwijderen predictoren naar rechts en kies onderaan voor Remove:



Op deze manier krijgen we straks twee modellen die we kunnen vergelijken. Je kan nu ook nog achter het knopje Statistics de optie R Squared Change aanduiden.

In de tabel Variables Entered/Removed kan je nakijken of beide modellen correct zijn ingevoerd.

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	fout, aantal aanwezigen op de fuif, aantal alcoholische consumpties, geslacht, disco, aantal vrienden aanwezig, beachparty ^b		Enter
2	^b	fout, disco, aantal aanwezigen op de fuif, geslacht ^c	Remove

a. Dependent Variable: fuijsatisfactie

b. All requested variables entered.

c. All requested variables removed.

In de tabellen Model Summary en ANOVA kunnen we beide modellen vergelijken. We stellen vast dat het gereduceerde model (2) iets minder variantie verklaart, maar de reductie blijft beperkt. De ANOVA tabel geeft aan dat ook het tweede model nog steeds significant is. We zien dat de F-waarde van model 2 een pak groter is dan die van model 1. Dat heeft te maken met het kleinere aantal predictoren in model 2, waardoor de df van model 2 kleiner is.

Model Summary^c

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,720 ^a	,518	,453	1,18797	,518	7,976	7	52	,000	2,227
2	,694 ^b	,482	,454	1,18630	-,036	,961	4	52	,437	

a. Predictors: (Constant), fout, aantal aanwezigen op de fuif, aantal alcoholische consumpties, geslacht, disco, aantal vrienden aanwezig, beachparty

b. Predictors: (Constant), aantal alcoholische consumpties, aantal vrienden aanwezig, beachparty

c. Dependent Variable: fuifsatisfactie

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	78,797	7	11,257	7,976	,000 ^b
	Residual	73,386	52	1,411		
	Total	152,183	59			
2	Regression	73,374	3	24,458	17,379	,000 ^c
	Residual	78,809	56	1,407		
	Total	152,183	59			

a. Dependent Variable: fuifsatisfactie

b. Predictors: (Constant), fout, aantal aanwezigen op de fuif, aantal alcoholische consumpties, geslacht, disco, aantal vrienden aanwezig, beachparty

c. Predictors: (Constant), aantal alcoholische consumpties, aantal vrienden aanwezig, beachparty

De tabel Coefficients toont dat de drie resterende predictoren in model 2 significant zijn, waarbij het aantal aanwezige vrienden blijkbaar de grootste impact heeft op de satisfactie. De collinearity Statistics zien er ok uit.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	10,156	,927		10,955	,000	8,295	12,016		
	aantal alcoholische consumpties	-,176	,053	-,349	-3,338	,002	-,282	-,070	,847	1,181
	aantal aanwezigen op de fuif	,005	,003	,151	1,494	,141	-,002	,012	,908	1,101
	aantal vrienden aanwezig	,294	,053	,624	5,514	,000	,187	,401	,724	1,381
	geslacht	,136	,331	,041	,410	,683	-,528	,799	,911	1,098
	beachparty	,941	,454	,266	2,074	,043	,031	1,852	,562	1,779
	disco	,501	,448	,130	1,120	,268	-,397	1,399	,692	1,445
	fout	-,153	,465	-,039	-,330	,743	-1,087	,780	,679	1,472
2	(Constant)	11,173	,602		18,555	,000	9,967	12,380		
	aantal alcoholische consumpties	-,187	,051	-,370	-3,645	,001	-,289	-,084	,897	1,115
	aantal vrienden aanwezig	,293	,049	,622	6,034	,000	,196	,391	,871	1,148
	beachparty	,844	,381	,239	2,217	,031	,081	1,606	,798	1,254

a. Dependent Variable: fuifsatisfactie

Ook de tabel Collinearity Diagnostics toont een normaal beeld voor model 2 (elke predictor laadt hoog op slechts 1 dimensie).

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions								
				(Constant)	aantal alcoholische consumpties	aantal aanwezigen op de fuif	aantal vrienden aanwezig	geslacht	beachparty	disco	fout	
1	1	5,200	1,000	,00	,01	,00	,00	,00	,00	,00	,00	,01
	2	1,038	2,239	,00	,01	,00	,00	,00	,00	,23	,10	,10
	3	1,007	2,272	,00	,00	,00	,00	,00	,00	,00	,29	,24
	4	,325	3,999	,00	,27	,03	,02	,02	,02	,08	,26	,28
	5	,216	4,907	,00	,64	,00	,00	,00	,00	,51	,20	,35
	6	,125	6,445	,00	,03	,26	,03	,49	,00	,00	,00	,00
	7	,068	8,748	,00	,00	,57	,58	,09	,04	,00	,00	,01
	8	,021	15,683	,99	,05	,12	,37	,39	,13	,16	,01	,01
2	1	3,062	1,000	,01	,03		,01		,03			
	2	,667	2,144	,01	,00		,02		,65			
	3	,233	3,628	,02	,91		,05		,19			
	4	,038	8,945	,97	,06		,92		,13			

a. Dependent Variable: fuifsatisfactie

Verder vinden we geen eigenaardigheden bij het checken van de voorwaarden ivm outliers en de verdeling van de residuen. We kunnen dus rapporteren:

Een meervoudige regressieanalyse werd uitgevoerd met de fuifsatisfactie als criterium en het aantal alcoholische consumpties, het aantal aanwezigen, het aantal vrienden, geslacht en fuifthema als predictoren (model 1). De variabele fuifthema werd dummygecodeerd in de variabelen beachparty, disco en foute fuif. Dit model bleek significant, met $R^2 = .52$, $F = 7.98$, $p < .001$. Zoals aangegeven in Tabel 1 waren het aantal aanwezigen, geslacht, disco en foute fuif geen significante predictoren. Deze predictoren werden daarom niet opgenomen in model 2, dat ook significant bleek met $R^2 = .48$, $F = 17.38$, $p < .001$. Van de resterende predictoren blijkt het aantal aanwezige vrienden het meeste invloed uit te oefenen op de fuifsatisfactie.

Tabel 1: Resultaten enkelvoudige regressie met fuifsatisfactie als criterium en aantal alcoholische consumpties als predictor.

	<i>B</i>	<i>SE B</i>	<i>β</i>	<i>t</i>
model 1				
constante	10.16	.93		10.96***
aantal consumpties	-.18	.05	-.35	-3.34**
aantal aanwezigen	.01	.00	.15	1.49
aantal vrienden	.29	.05	.62	5.51***
geslacht	.14	.33	.04	.41
beachparty	.94	.45	.27	2.07*
disco	.50	.45	.13	1.12
foute fuif	-.15	.47	-.04	-.33
model 2				
constante	11.17	.60		18.56***
aantal consumpties	-.19	.05	-.37	-3.65***
aantal vrienden	.29	.05	.62	6.03***
beachparty	.84	.38	.24	2.22*

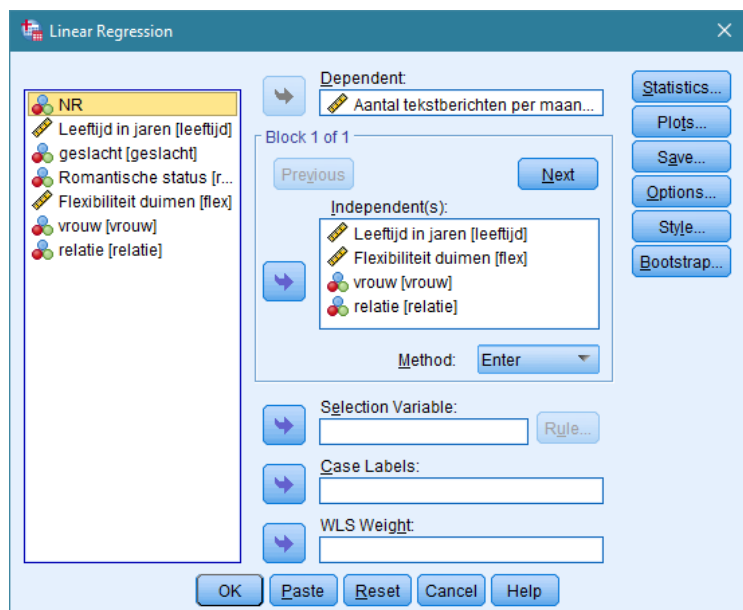
* $p < .05$, ** $p < .01$, *** $p < .001$

- Een telecommarketeer is geïnteresseerd in het aantal tekstberichten die verstuurd worden door de klanten van het bedrijf. Hij vraagt zich af of dit aantal wordt beïnvloed door predictoren als leeftijd, geslacht, romantische status en flexibiliteit van de duimen. Hij verzamelde data in een willekeurige steekproef, die je kunt vinden in het bestand *opdr_teksberichten.sav*. Ga na welke variabelen een invloed uitoefenen op het aantal verstuurde tekstberichten en stel hiervoor het regressiemodel op.

In deze oefening vragen we ons af in welke mate het aantal verstuurde tekstberichten kan voorspeld worden door de predictoren leeftijd, geslacht, relatie en duimflexibiliteit. Leeftijd en flexibiliteit zijn intervalvariabelen, terwijl geslacht en relatie duidelijk nominaal zijn. Dat betekent dat we voor deze twee laatste variabelen dummy variabelen zullen moeten maken. Dit zal dan ook de eerste stap zijn. Gebruik Recode into different variables om zowel voor geslacht als relatie 1 dummy variabele te maken.

	NR	leeftijd	geslacht	rstatus	flex	tekstberichten	vrouw	relatie
1	1	50	1	1	10,18	321	1	0
2	2	48	2	1	8,79	85	0	0
3	3	37	1	2	6,35	258	1	1
4	4	54	2	1	6,55	240	0	0
5	5	57	1	1	11,93	251	1	0
6	6	45	1	1	8,84	266	1	0
7	7	32	1	2	7,81	344	1	1
8	8	41	1	2	10,73	105	1	1
9	9	59	2	1	8,73	169	0	0
10	10	55	2	2	10,86	149	0	1
11	11	29	1	1	11,98	275	1	0
12	12	38	1	1	10,86	210	1	0
13	13	58	1	2	11,01	181	1	1
14	14	18	1	1	9,16	222	1	0
15	15	33	2	2	10,55	161	0	1
16	16	40	2	1	10,69			
17	17	50	1	1				
18	18	25	1					

Vervolgens voeren we de regressieanalyse uit met alle opties zoals ook getoond in paragraaf 10.2.7. We gebruiken de enter methode omdat we geen specifieke verwachtingen hebben omtrent de relatieve bijdragen van de verschillende predictoren. Denk eraan om de dummy variabelen in de analyse te brengen en niet de oorspronkelijke variabelen.



In de output kijken we eerst naar de tabel

Correlations. Daarin zien we een aantal redelijke correlaties, waarvan een aantal significant. De correlatie tussen tekstberichten en relatie is weliswaar erg klein en niet significant. Voorlopig doen we daar niets mee, maar we houden die informatie in het achterhoofd voor het verdere verloop.

Correlations

		Aantal tekstberichten per maand	Leeftijd in jaren	Flexibiliteit duimen	vrouw	relatie
Pearson Correlation	Aantal tekstberichten per maand	1,000	-,326	-,148	-,182	,066
	Leeftijd in jaren	-,326	1,000	,044	,084	,010
	Flexibiliteit duimen	-,148	,044	1,000	,037	-,096
	vrouw	-,182	,084	,037	1,000	-,069
	relatie	,066	,010	-,096	-,069	1,000
Sig. (1-tailed)	Aantal tekstberichten per maand	.	,000	,005	,001	,124
	Leeftijd in jaren	,000	.	,223	,072	,433
	Flexibiliteit duimen	,005	,223	.	,257	,048
	vrouw	,001	,072	,257	.	,113
	relatie	,124	,433	,048	,113	.
N	Aantal tekstberichten per maand	305	305	305	305	305
	Leeftijd in jaren	305	305	305	305	305
	Flexibiliteit duimen	305	305	305	305	305
	vrouw	305	305	305	305	305
	relatie	305	305	305	305	305

Volgende interessante gegevens zijn die in de tabellen Model Summary en ANOVA. Daarin zien we dat de R²-waarde afgerond gelijk is aan .15, wat niet geweldig hoog is. De Durbin-Watson waarde is 2.1, wat geheel normaal is en niet tot bezorgdheid moet leiden. De ANOVA tabel toont dat het model significant is.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,386 ^a	,149	,138	82,832	2,103

a. Predictors: (Constant), relatie, Leeftijd in jaren, Flexibiliteit duimen, vrouw

b. Dependent Variable: Aantal tekstberichten per maand

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	360421,014	4	90105,253	13,133	,000 ^b
	Residual	2058339,852	300	6861,133		
	Total	2418760,866	304			

a. Dependent Variable: Aantal tekstberichten per maand

b. Predictors: (Constant), relatie, Leeftijd in jaren, Flexibiliteit duimen, vrouw

Daarna bekijken we in de tabel Coefficients eerst de Standardized Coefficients en de significanties. Daaruit blijkt dat alle predictoren behalve Relatie significant zijn. Leeftijd blijkt de sterkste predictor te zijn. In dezelfde tabel zien we dat er voorlopig geen collineariteitsproblemen zijn, want de waarden voor Tolerance en VIF zijn normaal.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	383,730	29,116		13,180	,000	326,433	441,027		
	Leeftijd in jaren	-2,193	,380	-,308	-5,765	,000	-2,942	-1,445	,991	1,009
	Flexibiliteit duimen	-5,830	2,501	-,125	-2,331	,020	-10,751	-,909	,988	1,012
	vrouw	-26,427	9,584	-,148	-2,757	,006	-45,288	-7,566	,987	1,013
	relatie	8,394	9,580	,047	,876	,382	-10,458	27,245	,986	1,014

a. Dependent Variable: Aantal tekstberichten per maand

In de tabel Collinearity Diagnostics zien we wel een probleem opduiken: op dimensies 2 en 3 zien we grote gewichten voor zowel Vrouw als Relatie en dat is geen goede zaak. Op basis van wat we al vaststelden in verband met de correlaties en de Standardized Coefficients kunnen we besluiten dat de predictor Relatie beter geschrapt kan worden uit de analyse.

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Leeftijd in jaren	Flexibiliteit duimen	vrouw	relatie
1	1	3,951	1,000	,00	,01	,00	,02	,02
	2	,579	2,613	,00	,00	,00	,44	,48
	3	,377	3,237	,01	,03	,01	,53	,47
	4	,077	7,170	,03	,89	,12	,00	,00
	5	,016	15,633	,96	,08	,87	,00	,03

a. Dependent Variable: Aantal tekstberichten per maand

De overige tabellen en grafieken waarin we outliers en residuen controleren zien er allemaal normaal uit, dus we kunnen gewoon de analyse opnieuw uitvoeren, maar dan zonder Relatie als predictor.

De correlatietabel bevat uiteraard dezelfde correlaties als voorheen, maar dan zonder de variabele Relatie. In de volgende tabellen zien we dat het model nog steeds significant is, en dat de R^2 nauwelijks gedaald is. Ook de Durbin-Watson waarde zit in het veilige bereik.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,383 ^a	,147	,138	82,800	2,094

a. Predictors: (Constant), vrouw, Flexibiliteit duimen, Leeftijd in jaren

b. Dependent Variable: Aantal tekstberichten per maand

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	355153,397	3	118384,466	17,268	,000 ^b
	Residual	2063607,468	301	6855,839		
	Total	2418760,866	304			

a. Dependent Variable: Aantal tekstberichten per maand

b. Predictors: (Constant), vrouw, Flexibiliteit duimen, Leeftijd in jaren

De tabel Coefficients toont opnieuw drie significante predictoren, met leeftijd als sterkste predictor. We vragen ons nu ook af op welke manier de predictoren een voorspellende waarde hebben. Aan de hand van het teken van de Standardized Coefficients stellen we vast dat er telkens een negatief verband is. Dat wil zeggen: hoe

lager de waarde van de predictor, hoe hoger de waarde van de AV (aantal tekstberichten). Met andere woorden:

- Hoe lager de leeftijd, hoe hoger het aantal tekstberichten;
- Hoe lager de waarde voor duimflexibiliteit, hoe hoger het aantal tekstberichten;
- Hoe lager de waarde voor "vrouw", hoe hoger het aantal tekstberichten. Om te weten wat dit betekent voor de predictor Geslacht dien je te kijken naar hoe je zelf de dummy-codering hebt uitgevoerd: heb je de waarde '1' gebruikt voor vrouwen of mannen? In ons geval hebben we waarde 1 toegekend aan vrouwen. Dat wil zeggen dat '0' eigenlijk overeenstemt met mannen. De negatieve Standardized Coefficient geeft dus aan dat een '0' samenhangt met meer tekstberichten dan een '1'. Dus: mannen sturen meer tekstberichten dan vrouwen.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	389,682	28,301		13,769	,000	333,989	445,376		
	Leeftijd in jaren	-2,187	,380	-,308	-5,752	,000	-2,935	-1,439	,991	1,009
	Flexibiliteit duimen	-6,036	2,489	-,129	-2,425	,016	-10,933	-1,139	,997	1,003
	vrouw	-26,994	9,559	-,151	-2,824	,005	-45,805	-8,184	,992	1,008

a. Dependent Variable: Aantal tekstberichten per maand

In de tabel Collinearity Diagnostics zien we dat elke predictor nu slechts op 1 dimensie hoog scoort.

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Leeftijd in jaren	Flexibiliteit duimen	vrouw
1	1	3,451	1,000	,00	,01	,00	,03
	2	,455	2,754	,00	,01	,00	,97
	3	,077	6,696	,03	,90	,11	,00
	4	,017	14,369	,96	,08	,88	,00

a. Dependent Variable: Aantal tekstberichten per maand

De tabellen Casewise Diagnostics en Residual Statistics zien er normaal uit: geen Std. Residuals groter dan 3 en Cook's Distance blijft beperkt tot max .047.

We rapporteren dus:

Een meervoudige regressieanalyse werd uitgevoerd met Aantal Tekstberichten als criterium en Leeftijd, Geslacht, Romantische status en Duimflexibiliteit als predictoren (model 1). Model 1 bleek significant met $R^2 = .015$, $F = 13.13$, $p < .001$. De predictor Romantische status bleek echter geen significante predictor en bleek bovendien een ongewenst effect te hebben op de multicollineariteit. Deze predictor werd weggelaten uit Model 2. Ook dit model was significant met $R^2 = .015$, $F = 17.27$, $p < .001$. Leeftijd blijkt de sterkste predictor te zijn van de drie resterende predictoren. We concluderen dat het aantal verstuurde berichten per maand groter is bij jongere mensen; dat mannen meer berichten sturen dan vrouwen; en dat een geringere duimflexibiliteit eigenaardig genoeg leidt tot meer verstuurde berichten per maand.

Tabel 1: Resultaten meervoudige regressie met Aantal Tekstberichten als criterium en Leeftijd, Geslacht, Romantische status en Duimflexibiliteit als predictoren.

	<i>B</i>	<i>SE B</i>	<i>β</i>	<i>t</i>
model 1				
constante	383.73	29.12		13.18***
leeftijd	-2.19	.38	-.31	-5.77***
geslacht	-26.43	9.58	-.15	-2.76**
duimflexibiliteit	-5.83	2.50	-.13	-2.33*
romantische status	8.39	9.58	.047	.88
model 2				
constante	389.68	28.30		13.77***
leeftijd	-2.19	.38	-.31	-5.75***
geslacht	-26.99	9.56	-.15	-2.82**
duimflexibiliteit	-6.04	2.49	-.13	-2.43*

* $p < .05$, ** $p < .01$, *** $p < .001$