Een docent heeft gegevens verzameld van 60 studenten bij zijn vak. Onder andere heeft hij beschikking over informatie wat betreft het aantal uren dat deze studenten gestudeerd hebben voor dit vak en het cijfer wat zij gehaald hebben. Met behulp van onderstaande output wil deze docent achterhalen in hoeverre deze twee variabelen aan elkaar gerelateerd zijn. Vraag 1-4 gaan over de output.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,642[a]	,412	,401	1,79745

a. Predictors: (Constant), zelfstudie

b. Dependent Variable: cijfer

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	131,088	1	131,088	40,574	,000[a]
	Residual	187,389	58	3,231		
	Total	318,477	59			

a. Predictors: (Constant), zelfstudie

b. Dependent Variable: cijfer

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,547	,592				1,363	3,732
	zelfstudie	,069	,011	,642			,047	,090

a. Dependent Variable: cijfer

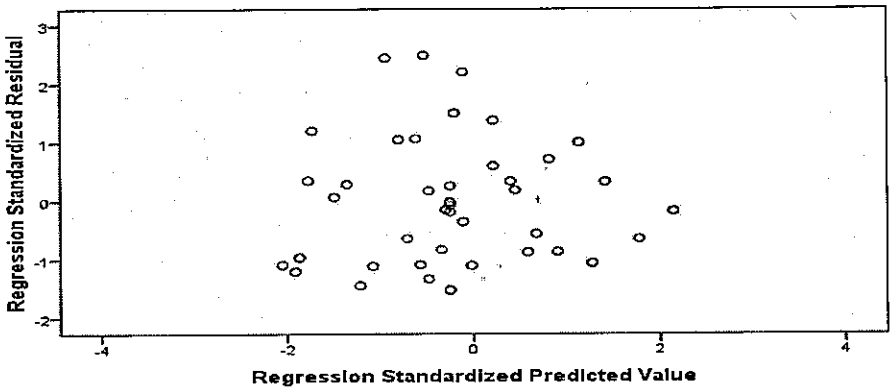
1. Welke informatie vertelt je concreet hoe goed het model bij deze data past?
   a. **De proportie verklaarde variantie, 0.412**
   b. De helling $B_1 = 0.069$
   c. De F-waarde 40,574
   d. De constante $B_0 = 2,547$

2. De helling van de regressievergelijking is 0.069 in deze steekproef. Als je een significantieniveau van 5% hanteert, wat mag je dan concluderen?
   a. In de populatie zal de voorspelling van het cijfer stijgen met 0.069 punt als het aantal uren zelfstudie stijgt met één uur
   b. **Er lijkt een significant positieve relatie te zijn tussen zelfstudie en cijfer**
   c. Gezien het betrouwbaarheidsinterval voor de helling kan er bij het gegeven significantieniveau geen uitspraak gedaan worden over een eventuele relatie tussen zelfstudie en cijfer
   d. De relatie tussen zelfstudie en cijfer is dusdanig zwak dat deze niet significant zal zijn

3. Een bepaalde student heeft 80 uur gestudeerd en een 8.5 gehaald op het tentamen. Wat zal de waarde van het residu zijn voor deze student?
   a. 8.067
   b. 5.457
   c. **0.433**
   d. 3.043

4. Wat is hier de betekenis van het intercept B_0?
   a. Het intercept B_0 geeft aan hoeveel uur je gestudeerd hebt als je een 1 hebt gehaald op je tentamen
   b. Het intercept B_0 geeft aan wat de voorspelde waarde op cijfer is als je geen tijd hebt gestopt in zelfstudie
   c. Het intercept B_0 drukt de relatie tussen zelfstudie en cijfer uit
   d. Het intercept B_0 heeft nooit een zinvolle betekenis

5. Bij een regressieanalyse horen assumpties die gecontroleerd dienen te worden. Hieronder staat een residuplot. Welke assumptie lijkt het meest geschonden te zijn volgens dit residuplot?
   a. Onafhankelijke waarnemingen
   b. Normaliteit van het residu
   c. Homoscedasticiteit van het residu
   d. Lineaire relatie tussen de onafhankelijke variabele en de afhankelijke variabele



6. De docent heeft naast het aantal uren zelfstudie ook bijgehouden hoeveel college uren elke student gevolgd heeft. Hieronder staan de correlaties tussen college uren, zelfstudie en het behaalde cijfer. Hoeveel procent verklaarde variantie zal het aantal gevolgde college uren toevoegen als we deze IV betrekken bij de regressie van cijfer?
   a. 0.131
   b. 0.222
   c. 0.361
   d. 0.439

Correlations

		cijfer	zelfstudie	college
Pearson Correlation	cijfer	1,000	,642	,661
	zelfstudie	,642	1,000	,565
	college	,661	,565	1,000
Sig. (1-tailed)	cijfer	.	,000	,000
	zelfstudie	,000	.	,000
	college	,000	,000	.
N	cijfer	60	60	60
	zelfstudie	60	60	60
	college	60	60	60

Ongeacht eventuele problemen (die je dus mag negeren) voert de docent een multipele regressieanalyse uit over de data. Hieronder staat een gedeelte van de output. Vraag 7-8 gaan over deze output

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	172,897	2	86,449	33,848	,000[b]
	Residual	145,579	57	2,554		
	Total	318,477	59			

a. Dependent Variable: cijfer

b. Predictors: (Constant), college, zelfstudie

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-1,075	1,038		-1,035	,305			
	zelfstudie	,042	,012		3,625	,001	,642	,433	,325
	college	,175	,043		4,046	,000	,661	,472	,362

a. Dependent Variable: cijfer

7. Welke uitspraak is het best verdedigbaar?
   a. College uren heeft een behoorlijk sterkere relatie met het cijfer dan zelfstudie gezien het grote verschil tussen beide regressiegewichten (0.175 versus 0.042)
   b. College uren heeft een net iets sterkere relatie met cijfer dan zelfstudie gezien het kleine verschil in gestandaardiseerde regressiegewichten (0.439 versus 0.393)
   c. College uren en zelfstudie hebben een even sterke relatie met cijfer aangezien beiden significant zijn
   d. Aangezien het intercept negatief is kan er geen waarde gehecht worden aan dit model

8. Wat is de nulhypothese die getoetst wordt met de ANOVA F-procedure bij deze output?
   a. De partiële regressiegewichten voor zelfstudie en college uren zijn gelijk aan nul in de populatie
   b. Het intercept en de partiële regressiegewichten voor zelfstudie en college uren zijn gelijk aan nul in de populatie
   c. Zelfstudie en college uren hebben een gemiddelde in de populatie van nul
   d. Zelfstudie, college uren en cijfer hebben een gemiddelde in de populatie van nul

9. Wat kun je onderzoeken met een QQ-plot (ook wel "normal quantile plot")?
   a. De assumptie van gelijkheid van varianties
   b. De assumptie van normaliteit
   c. De mate van samenhang tussen de afhankelijke en de onafhankelijke variabele
   d. Multicollineariteit

10. Een betrouwbaarheidsinterval wordt breder als:
    a. de margin of error (foutenmarge) kleiner wordt
    b. de standaarddeviatie in de steekproef kleiner wordt
    c. de steekproefgrootte (*n*) kleiner wordt
    d. het percentage betrouwbaarheid kleiner wordt

Een groep onderzoekers wil weten in hoeverre de tijd die iemand besteedt aan het scheiden van afval (deze variabele wordt "uren" genoemd) afhangt van de waarde die iemand toekent aan 1) de lokale effecten van niet scheiden ("lokaal"), 2) de globale effecten van niet scheiden ("globaal"), 3) de sociale druk van andere leden van de huishouding ("huisgenoten") en 4) de sociale druk van de buurt ("buurtgenoten"). De waarde die iemand toekent aan deze vier factoren worden gemeten op een schaal van 0 tot 15. De vragen 11 tot en met 15 gaan over onderstaande output.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,713[a]	,509	,492	4,55610

a. Predictors: (Constant), lokaal, globaal, huisgenoten, buurtgenoten

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2492,162	4	623,040	30,014	,000[a]
	Residual	2407,930	116	20,758		
	Total	4900,091	120			

a. Predictors: (Constant), lokaal, globaal, huisgenoten, buurtgenoten

b. Dependent Variable: uren

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5,780	3,518		1,643	,103	-1,188	12,747
	globaal	-,021	,414	-,004	-,052	,959	-,841	,798
	huisgenoten	,747	,191	,282	3,907	,000	,369	1,126
	buurtgenoten	2,028	,232	,653	8,729	,000	1,568	2,488
	lokaal	-1,398	,236	-,428	-5,931	,000	-1,865	-,931

a. Dependent Variable: uren

11. Het betrouwbaarheidsinterval voor de waarde van globale effecten loopt grofweg afgerond van -0.8 tot 0.8. Wat betekent dit betrouwbaarheidsinterval?
    a. **Er is 95% kans dat het interval de parameter voor het partiële regressiegewicht horend bij de waarde van globale effecten omvat**
    b. Er is 95% kans dat het interval de statistic voor het partiële regressiegewicht horend bij de waarde van globale effecten omvat
    c. De waarde van globale effecten speelt geen rol bij het aantal uren
    d. De waarde van globale effecten verhoogt bij 95% van de deelnemers het aantal uren met minstens -0.8 en maximaal 0.8

12. Welke van de volgende conclusies is juist?
    a. Het aantal aan scheiden bestede uren in de steekproef lijkt voor ongeveer 49% te voorspellen op basis van de verschillende IV in deze steekproef.
    b. **Het aantal aan scheiden bestede uren in de steekproef lijkt voor ongeveer 51% te voorspellen op basis van de verschillende IV in deze steekproef**
    c. Het aantal aan scheiden bestede uren in de steekproef lijkt voor ongeveer 71% te voorspellen op basis van de verschillende IV in deze steekproef
    d. Het aantal aan scheiden bestede uren lijkt in de steekproef te voorspellen op basis van de verschillende IV in deze steekproef, maar hoe veel verklaard is in de steekproef is op basis van deze gegevens niet te zeggen

13. De standaardfouten in de derde kolom van het blokje *Coefficients* zijn behoorlijk verschillend van elkaar. Welke van de volgende uitspraken is juist?
    a. Dit duidt op een schending van de assumptie van gelijke varianties
    b. Aangezien de onafhankelijke variabelen gemeten zijn op verschillende schalen zegt dit niet zo veel over de assumptie van gelijke varianties
    c. **Dit heeft niks met de assumptie van gelijke varianties te maken aangezien die op de residuen en niet op de afzonderlijke variabelen slaat**
    d. De assumptie van gelijke varianties is waarschijnlijk alleen voor *globaal* niet geschonden

14. Anna heeft 40 uur besteed aan het scheiden van afval. De scores die zij voor dit vak toekent aan de verschillende onafhankelijke variabelen zijn de volgende: huisgenoten en buurtgenoten 10, en lokaal en globaal 0. Wat is voor haar het residu voor het aantal uren?
    a. **6.47**
    b. 12.25
    c. 27.75
    d. 33.53

15. Bert heeft de waardes voor de vier onafhankelijke variabelen als volgt ingevuld: globaal 11, lokaal 13, buurtgenoten 3, en huisgenoten 5. Op basis hiervan kan een voorspelling gemaakt worden van het aantal aan scheiden van afval bestede uren. Bij het inleveren van zijn briefje met getallen voor de verschillende opbrengsten realiseert Bert zich dat hij de waarde voor buurtgenoten verkeerd heeft ingevuld: hij had eigenlijk een 7 willen invullen. Wat betekent deze aanpassing voor de voorspelling op basis van het regressiemodel van het aantal uren voor Bertt?
    a. Dit wordt na aanpassing 2.028 uren hoger
    b. Dit wordt na aanpassing 2.028 uren lager
    c. **Dit wordt na aanpassing 8.112 uren hoger**
    d. Dit wordt na aanpassing 8.112 uren lager

Onderstaande output heeft betrekking op een multipele regressie analyse. Gebruik waar nodig deze output voor de vragen 8-15

Descriptive Statistics

	Mean	Std. Deviation	N
y	7,9506	1,27836	162
x1	7,2160	1,17806	162
x2	76,2037	12,57037	162
x3	7,8160	1,07183	162

Correlations

	y	x1	x2	x3
y	1,000	,496	,578	,551
x1	,496	1,000	,454	,451
x2	,578	,454	1,000	,552
x3	,551	,451	,552	1,000

Model Summary

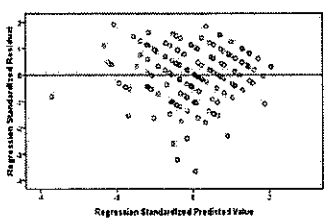
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,496[a]	,246	,241	1,11359	,246		1	160	,000
2	,634[b]	,402	,395	,99440	,157		1	159	,000
3	,670[c]	,449	,439	,95775	,047		1	158	,000

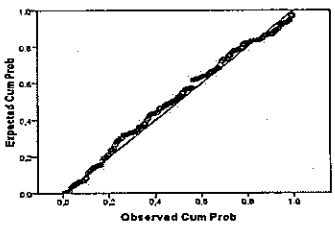
ANOVA[d]

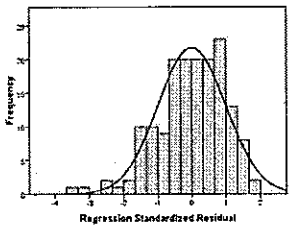
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	64,693	1	64,693	52,169	,000[a]
	Residual	198,412	160	1,240		
	Total	263,105	161			
2	Regression	105,880	2	52,940	53,538	,000[b]
	Residual	157,225	159	,989		
	Total	263,105	161			
3	Regression	118,173	3	39,391	42,943	,000[c]
	Residual	144,932	158	,917		
	Total	263,105	161			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,068	,545		7,469	,000		
	x1	,538	,074	,496	7,223	,000	1,000	1,000
2	(Constant)	2,206	,566		3,900	,000		
	x1	,319	,075	,294	4,275	,000	,794	1,260
	x2	,045	,007	,444	6,454	,000	,794	1,260
3	(Constant)	1,141	,617		1,847	,067		
	x1	,245	,076	,226	3,265	,001	,736	1,359
	x2	,033	,007	,326	4,433	,000	,643	1,556
	x3	,321	,088	,269	3,661	,000	,645	1,551







16. Bovenstaande output heeft betrekking op een stapsgewijze regressieanalyse. Wat voor methode lijkt hier te zijn gekozen?
    a. Een backward regressieanalyse
    b. Een forward regressieanalyse
    c. Een hiërarchische regressieanalyse

17. Wat zal dan de waarde van de F-toets zijn waarbij getoetst wordt in hoeverre de toevoeging van X2 aan de enkelvoudige regressievergelijking van Y op X1 leidt tot een significante verbetering van het model..?
    a. 6.454
    b. 41.65
    c. 53.54

18. Wat zal de waarde van de F-toets zijn waarbij getoetst wordt in hoeverre de toevoeging van X3 aan de enkelvoudige regressievergelijking van Y op X2 leidt tot een significante verbetering van het model?
    (pas op, lastige vraag... eigenlijk zijn het meerdere vragen ineen. Ik heb de volgorde veranderd ten opzichte van de output, dus je moet alles zelf uit gaan rekenen in plaats van op de output te vertrouwen. Hint: maak gebruik van de correlatietabel om te bepalen wat de proportie verklaarde variantie van model 1 is en wat de toegevoegde proportie verklaarde variantie is in model 2 als je X3 aan dit enkelvoudige model toevoegt)
    a. Ongeveer 13.5
    b. Ongeveer 21
    a. Ongeveer 42.5

19. Welke assumptie lijkt gezien de output het meest geschonden?
    a. homoscedasticiteit
    b. normaliteit
    c. lineariteit

20. Welke predictor levert in het uiteindelijke model de sterkste bijdrage aan de voorspelling van de afhankelijke variabele
    a. X3, aangezien het geschatte regressiegewicht van X3 het verst van nul afligt
    b. X2, aangezien het geschatte gestandaardiseerde regressiegewicht van X2 het verst van nul afligt
    c. X2, aangezien de correlatie van X2 met Y het sterkst is

21. In model 1 heeft X1 een geschat gestandaardiseerd regressiegewicht van 0.496. Het geschatte regressiegewicht van X1 in model 3 is gedaald naar 0.226. Waar heeft dit mee te maken?
    a. De lineaire samenhang tussen de verschillende IV
    b. Modelfluctuatie
    c. Zowel a, als b

22. Hoeveel procent van de variantie van X1 kan verklaard worden door X2 en X3?
    a. 0.226
    b. 0.245
    c. 0.264

23. In model 3 is de totale proportie verklaarde variantie van het model 0.449. Hoe kan dit uitgeschreven worden?
    a. $R^2_{y\hat{y}} = r^2_{yx1} + r^2_{yx2} + r^2_{yx3}$
    b. $R^2_{y\hat{y}} = r^2_{yx1} + r^2_{yx2.x1} + r^2_{yx3.x1x2}$
    c. $R^2_{y\hat{y}} = r^2_{yx1} + r^2_{y(x2.x1)} + r^2_{y(x3.x1x2)}$