

Opgave 1: Graft vs. Host disease

Denne opgave er baseret på opgave 12.3 fra DG Altman, p. 361. Data omhandler knoglemarvstransplantation af 37 leukæmipatienter, og outcome er forekomst af graft versus host disease, GvHD. Formålet med opgaven er at forsøge at prediktere, hvem der får GvHD.

Filen `gvhd.sav` indeholder følgende variable:

pnr Patientnummer

rcpage Patientens alder (recipienten)

donage Donors alder

type Leukæmitype:

1:AML (akut myeloid leukæmi)

2:ALL (akut lymfatisk leukæmi)

3:CML (kronisk myeloid leukæmi)

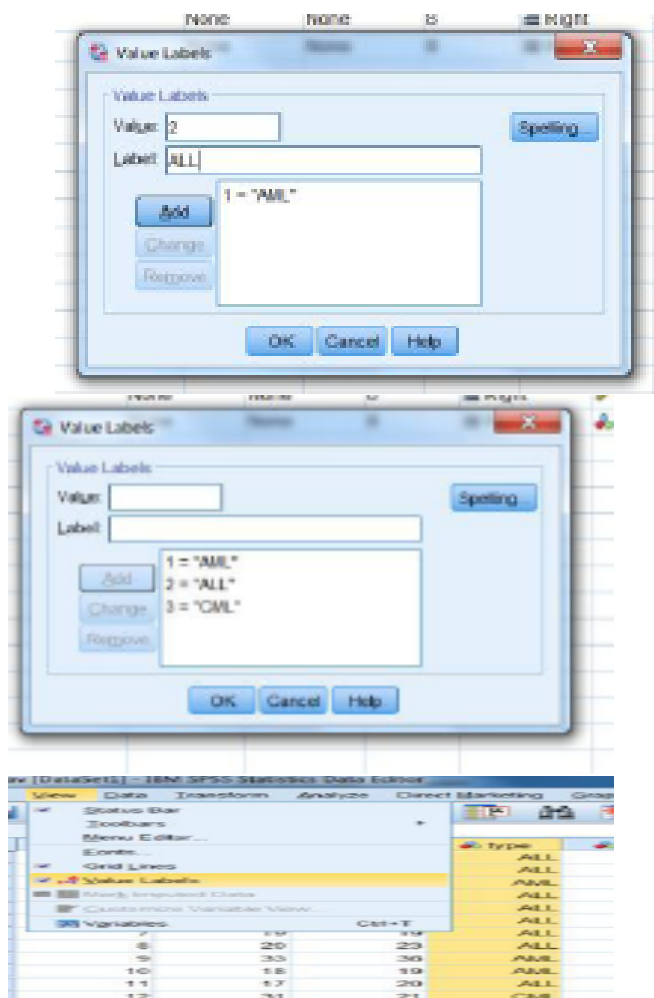
preg Indikator for hvorvidt donor har været gravid (1:ja, 0:nej)

index Index for epidermal celle-lymfocyt reaktion (kvantitativ måling)

gvhd Forekomst af GvHD (1:ja, 0:nej)

Først vil vi først konstruere nogle mere sigende navne til værdierne af variabelen **type**.

Vi går derfor ind i **Data/Define Variable Properties**, klikker **type** over i **Variables to Scan** og sætter hver værdi (1, 2 og 3) enkeltvis ind i **Changed** og den tilhørende label (**AML**, **ALL** og **CML**) i **Label**.



Inden vi går i gang med analyserne i de næste spørgsmål, vil vi lige se på en oversigt over de enkelte variable, blot for at få en slags overblik. Vi udregner derfor summary statistics, opdelt efter type. Vi går derfor først ind i Data/Split File, sætter flueben ved Compare Groups og sætte type over i Groups Based on, hvorefter vi er klar til at udregne summary statistics.

Til at udregne summary statistics benytter vi Analyze/Descriptive Statistics/Descriptives, flytter de variable, vi interesserer os for over i Variable(s) samt går ind i Statistics og sætter flueben ved de størrelser, vi ønsker udregnet (her Mean, Std. deviation, Minimum og Maximum):

Vi får så outputtet:

Descriptive Statistics						
type		N	Minimum	Maximum	Mean	Std. Deviation
1	rcpage	11	16	35	23,73	6,101
	donage	11	14	39	24,82	8,220
	preg	11	0	1	,36	,505
	Index	11	,39	10,11	3,0700	2,70549
	gvhd	11	0	1	,45	,522
	Valid N (listwise)	11				
2	rcpage	16	13	43	24,75	9,125
	donage	16	15	38	25,38	8,205
	preg	16	0	1	,25	,447
	Index	16	,27	4,52	1,5081	1,16506
	gvhd	16	0	1	,25	,447
	Valid N (listwise)	16				
3	rcpage	10	20	37	28,40	5,562
	donage	10	19	43	27,60	7,306
	preg	10	0	1	,20	,422
	Index	10	1,45	9,00	3,6680	2,41791
	gvhd	10	0	1	,80	,422
	Valid N (listwise)	10				

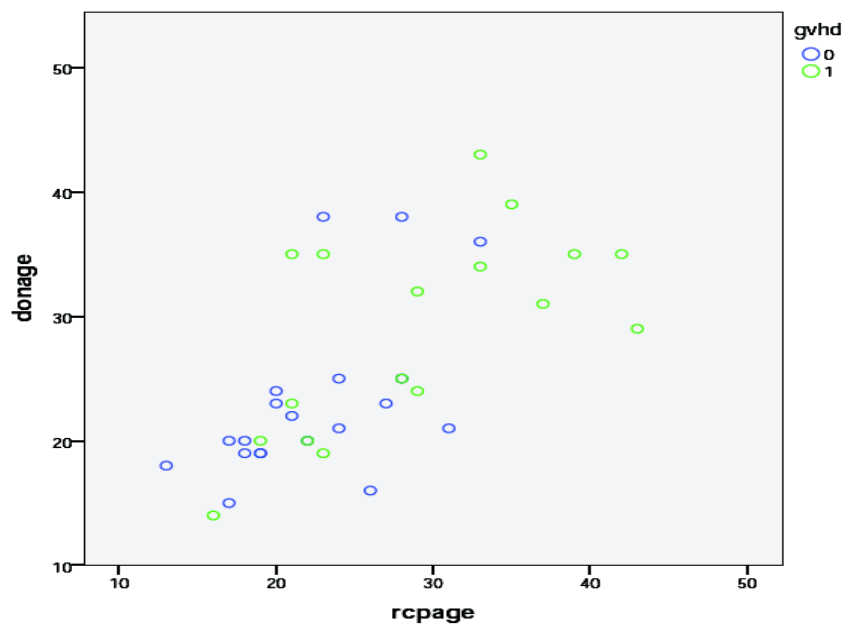
som måske er en anelse for detaljeret.

Hvis vi aflyser **Split File** (ved igen at gå ind i **Data/Split File**, men denne gang sætte flueben ved **Analyze all cases, do not create groups**) og kører analysen igen, får vi en noget mere overskuelig information:

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
rcpage	37	13	43	25,43	7,504
donage	37	14	43	25,81	7,838
preg	37	0	1	,27	,450
index	37	,27	10,11	2,5562	2,22994
gvhd	37	0	1	,46	,505
Valid N (listwise)	37				

Vi bemærker fra ovenstående, at patienterne er mellem 13 og 43 år, og at donorernes alder synes at stemme godt overens med patienternes alder. Dette er ikke noget tilfælde, idet man forsøger at matche disse, hvilket f.eks. kan ses af scatterplottet nedenfor (benyt **Graph/Chart Builder/Scatter**, vælg det opdeltte scatter plot, sæt **donage** på Y-aksen og **rcpage** på X-aksen, samt

gvhd i Set Color).



På denne figur ses i øvrigt, at højere alder synes at disponere for at få graft vs. host disease, men at det måske kan være svært at skille aldrene på de to personer ad (donor og modtager).

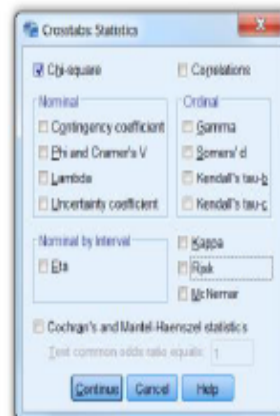
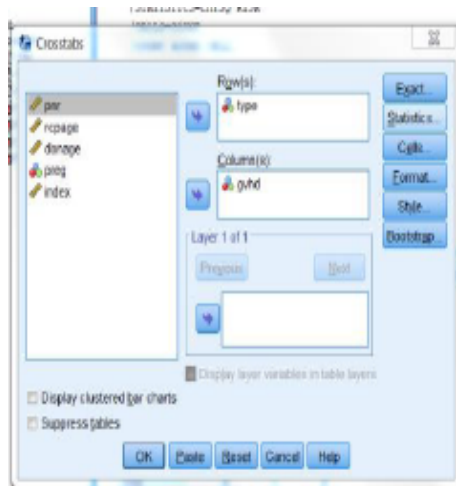
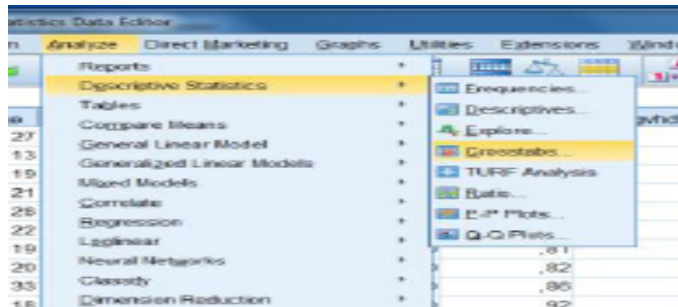
Yderligere confounding må påregnes mht. variabelen **preg**, der angiver om donoren har været gravid eller ej, da denne naturligt nok også må hænge sammen med donors alder.

Og så begynder vi på spørgsmålene:

1. *Lav først en tabel, der viser hyppigheden af GvHD for de tre sygdomstyper. Er der evidens for en forskel på disse tre hyppigheder?*

Der er her tale om en simpel 3×2 -tabel, men i modsætning til de tidligere tabeller (f.eks. astma-opgaven fra øvelsesuge 3) har vi her data på individ-niveau, så vi har ikke nogen vægt-variabel.

Vi konstruerer derfor tabellen ved at gå ind i
 Analyze/Descriptive Statistics/Crosstabs/, sætte type over i Row(s)
 og gvhd i Column(s). Og da vi også skal finde ud af, om der er evidens
 for forskelle mellem komplikationssandsynlighederne, går vi derefter ind
 i Statistics og afkrydser Chi-square:



Herved får vi:

type * gvhd Crosstabulation				
Count		gvhd		Total
		0	1	
type	AML	6	5	11
	ALL	12	4	16
	CML	2	8	10
Total		20	17	37

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7,497 ^a	2	,024
Likelihood Ratio	7,888	2	,019
Linear-by-Linear Association	2,236	1	,135
N of Valid Cases	37		

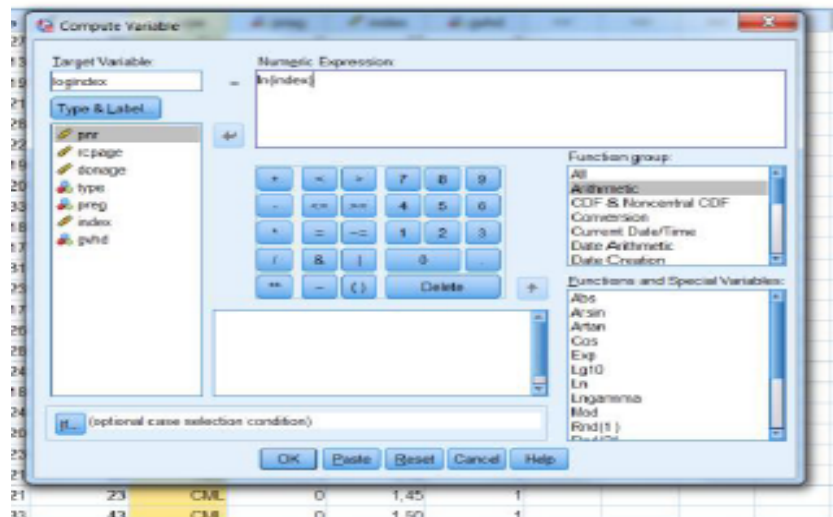
a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,59.

Vi ser af tabellen, at der totalt set observeres 17 tilfælde af graft vs. host disease blandt de 37 personer, samt at der er en tydelig overvægt af tilfælde blandt type de kroniske leukæmier (CML). Vi ser af χ^2 -testet, at der faktisk er signifikant forskel på de tre hyppigheder ($P = 0.02$), men det er *meget små antal*, og de forventede værdier (som også fremgår af tabellen) er flere steder lige omkring eller under 5.

2. Lav nu en logistisk regression med **gvhd** som outcome og med de øvrige variable som forklarende variable, idet I dog først logaritmetransformerer **index** (kald f.eks. den logaritmetransformerede variabel for **logindex**).

For at logaritmere **index**, går vi ind i Data/Transform/Compute Variable, hvor vi sætter det nye variabelnavn, **logindex**, som Target Variable og i Numeric Expression skriv-

er $\ln(\text{index})$:

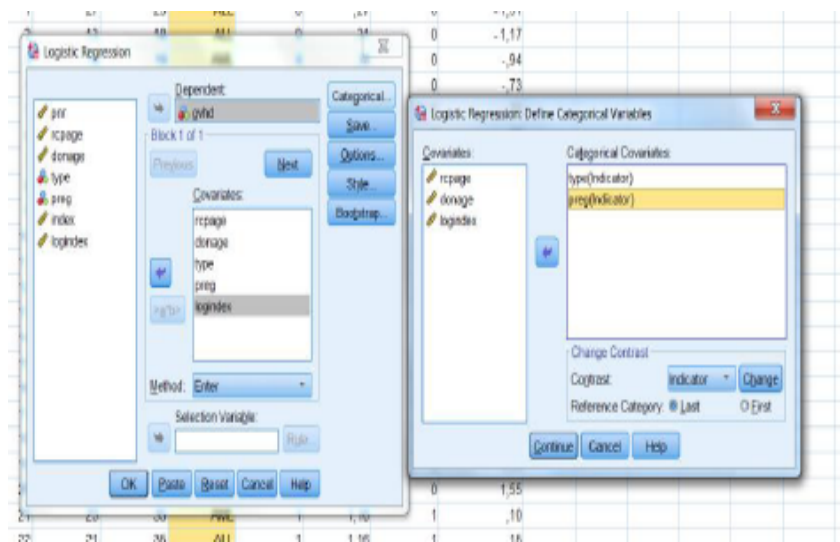
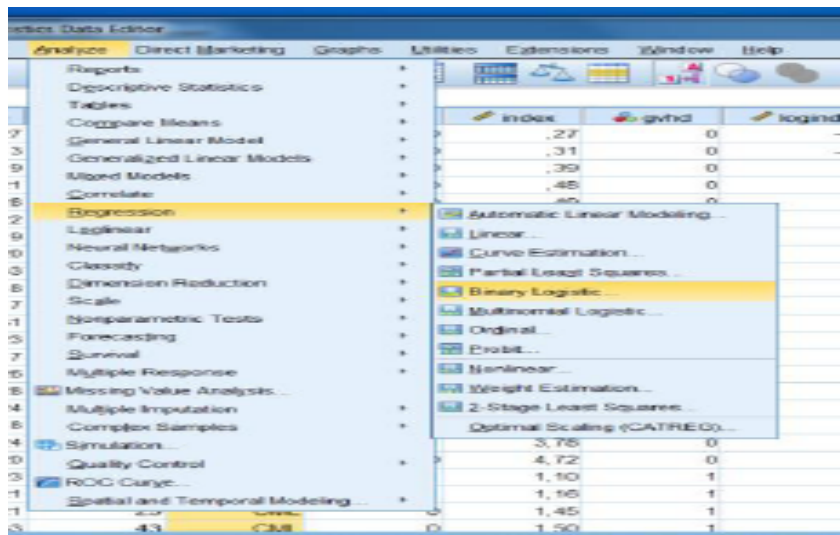


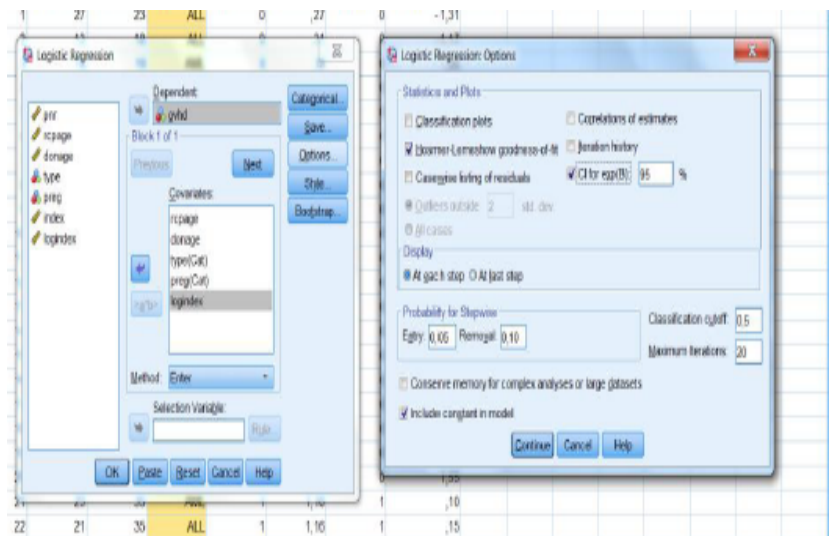
Som det ses, har vi her valgt at bruge den naturlige logaritme. Husk, at det ikke gør nogen forskel for modellen, kun for de estimater, man vælger at fokusere på. Vi kommer tilbage til dette ved fortolkningen senere.

Vi bemærker først, at der i alt er 5 potentielle kovariater, hvoraf den ene (**type**) er kategorisk med 3 niveauer, dvs. der skal estimeres 6 parametre i en model, hvor alle medtages. Dette er **klart for meget**, så vores analyse må betegnes som en **fisketur**, hvor evt. resultater skal eftervises ved en senere lejlighed!

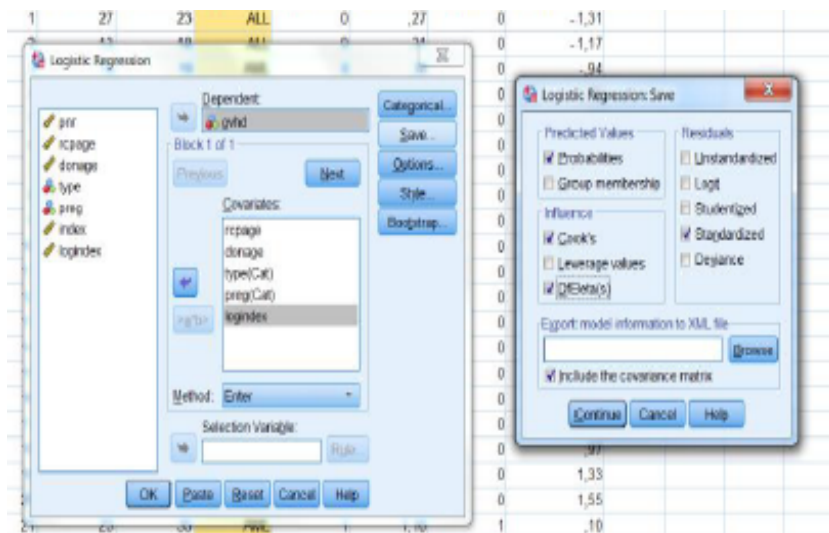
For at lave den logistiske regression benytter vi **Regression/Binary Logistic**, hvor vi sætter **gvhd** som **Dependent**, og alle 5 kovariater **rcpage**, **donage**, **type**, **preg** og **logindex** i **Covariates**. Herefter går vi ind i **Categorical Covariates** og sætter **type** og **preg** ind her, da de ikke indeholder numerisk information. Her kan man vælge, om referencen skal være første eller sidste kategori.

Herefter går vi ind i **Options** og afkrydser **Hosmer-Lemeshow Goodness-of-fit** samt **CI for $\exp(\beta)$** .





Hvis materialet havde været af en størrelsesorden, der gav mening for modelkontrol, ville man desuden gå ind i **Save** og afkrydse de ønskede størrelser, f.eks. Predicted probabilities, Cook's, Standardized residuals og DfBeta(s)



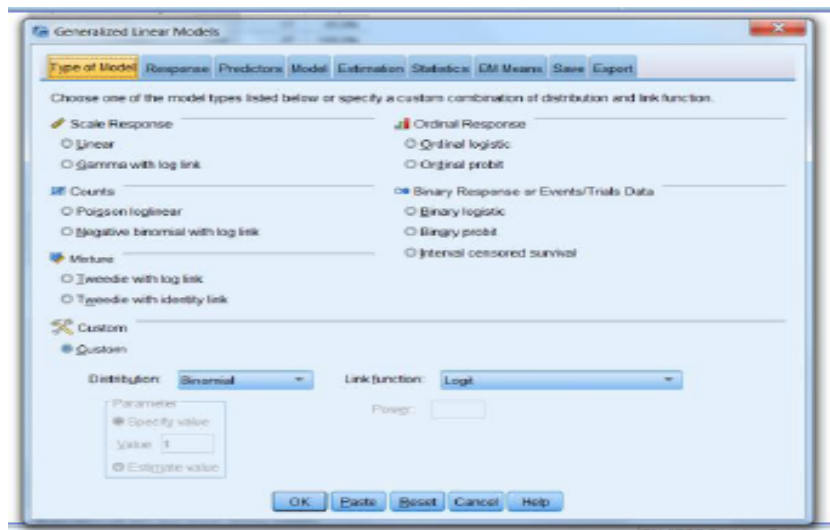
Outputtet (beskåret) fra denne *fulde* model bliver:

Variables in the Equation								
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.... Lower
Step 1 ^a	rcpage	,016	,082	,038	1	,846	1,016	,865
	donage	,107	,083	1,685	1	,194	1,113	,947
	type			2,462	2	,292		
	type(1)	-1,615	1,307	1,526	1	,217	,199	,015
	type(2)	-2,043	1,368	2,232	1	,135	,130	,009
	preg(1)	-1,662	1,181	1,981	1	,159	,190	,019
	logindex	1,819	,889	4,184	1	,041	6,164	1,079
	Constant	-2,197	3,321	,437	1	,508	,111	

Variables in the Equation		
		95% C.I. for .. Upper
Step 1 ^a	rcpage	1,193
	donage	1,309
	type	
	type(1)	2,579
	type(2)	1,891
	preg(1)	1,920
	logindex	35,217
	Constant	

a. Variable(s) entered on step 1: rcpage, donage, type, preg, logindex.

Outputtet fra denne analyse mangler det overordnede test om lighed mellem de 3 **type**-grupper, altså et test med **df=2**. Dette kan fås ved at benytte en anden procedure til den logistiske regression, nemlig **Generalized Linear Models**, hvor man skal udfylde en del faneblade (dem foroven i nedenstående box), nemlig **Response**, **Predictors**, **Model** og **Statistics** samt indgangssiden, som ses nedenfor (udfyldt med **Binary Response**, **Custom**, **Binomial** under **Distribution** og **Logit** under **Link function**):



Herved får vi (bl.a.) det ekstra output

Tests of Model Effects

Source	Wald Chi-Square	Type III	
		df	Sig.
(Intercept)	2,300	1	,129
type	2,462	2	,292
preg	1,981	1	,159
rcpage	,038	1	,846
donage	1,685	1	,194
logindex	4,184	1	,041

Dependent Variable: gvhd

Model: (Intercept), type, preg, rcpage, donage, logindex

Vi bemærker således, at vi i denne model ikke kan se forskel på de 3 sygdomsgrupper (overall F-test giver $P=0.29$).

Vi venter med at kommentere på estimerne for odds ratio til et senere spørgsmål.

3. *Fjern nu de forklarende variable successivt fra modellen, idet den mindst signifikante fjernes først (backwards elimination). Overvej gerne, om dette er en fornuftig fremgangsmåde.*

(a) *Undervejs i denne proces skal I være opmærksomme på, om estimer og P-værdier for nogle kovariater ændrer sig voldsomt ved fjernelse af andre. Se specielt på forskellen blandt leukæmigrupperne.*

Man ser, at `rcpage` (patientens alder) har den største P-værdi ($P=0.84$), så vi starter med at smide denne ud af modellen. Bemærk, at dette umiddelbart er i modstrid med vores iagttagelse fra figuren s. 3, men som nævnt er der flere variable, der hænger meget sammen og således let kan “dække” over hinanden.

og får outputtet (kraftigt beskåret til de allervigtigste dele):

Variables in the Equation								
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I....
								Lower
Step 1 ^a	donage	,114	,076	2,228	1	,136	1,120	,965
	type			2,762	2	,251		
	type(1)	-1,691	1,262	1,796	1	,180	,184	,016
	type(2)	-2,088	1,342	2,419	1	,120	,124	,009
	preg(1)	-1,722	1,145	2,260	1	,133	,179	,019
	logindex	1,841	,887	4,302	1	,038	6,301	1,107
	Constant	-1,883	2,909	,419	1	,517	,152	

Variables in the Equation		
		95% C.I. for ..
		Upper
Step 1 ^a	donage	1,300
	type	
	type(1)	2,186
	type(2)	1,721
	preg(1)	1,687
	logIndex	35,879
	Constant	

a. Variable(s) entered on step 1: donage, type, preg, logIndex.

og fra Generalized Linear Models-opsætningen endvidere:

Tests of Model Effects			
Source	Wald Chi-Square	Type III	
		df	Sig.
(Intercept)	2,575	1	,109
f _{type}	2,762	2	,251
preg	2,260	1	,133
donage	2,228	1	,136
logindex	4,302	1	,038

Dependent Variable: gvhd
Model: (Intercept), f_{type}, preg, donage, logindex

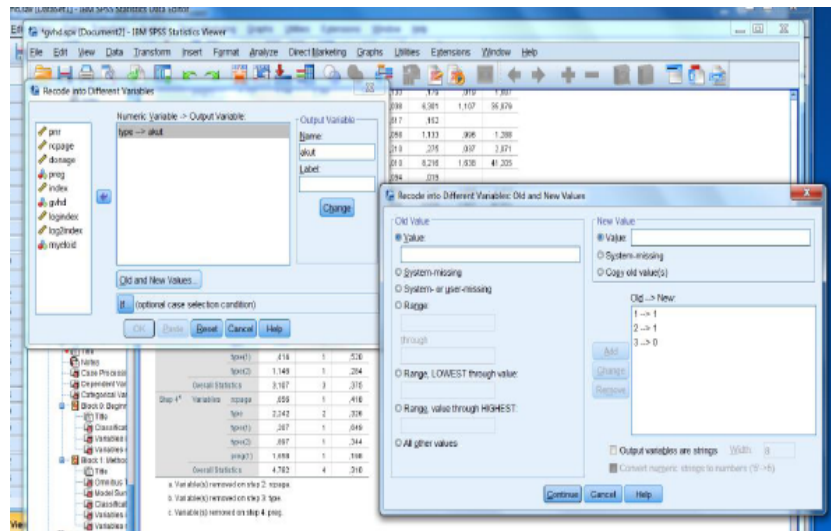
Der skete ingen særlige ændringer af de øvrige estimater, da vi smed `rcpage` ud.

Den mindst signifikante variabel blandt de tilbageværende er nu `type` ($P=0.25$), men denne har en særstatus, fordi den indeholder to sammenligninger (3 grupper, 2 frihedsgrader). Kigger vi nærmere på forskellene på de 3 grupper, ser vi, at den højeste P -værdi (**af de to viste!**) frekommer ud for `type 1:AML`, svarende til sammenligningen mellem patienter med AML og patienter med CML. Vi får dog ikke noget at vide om sammenligningen af ALL og AML, og når vi ser på estimaterne (selv om vi ikke forstår disse umiddelbart) kan vi se, at de begge ligger omkring -2, hvorimod CML som referencegruppe ligger på 0. Det må betyde, at ALL og AML faktisk ligner hinanden mest.

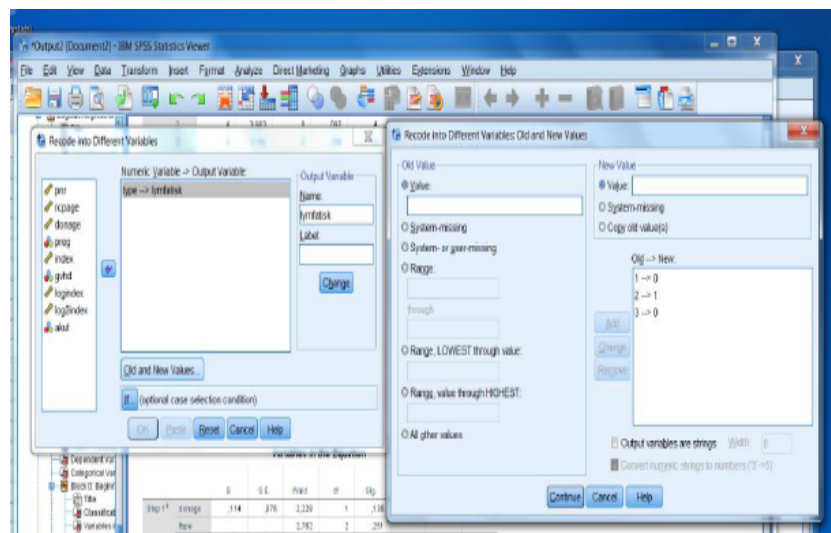
Det betyder, at den vigtigste information om typen er, hvorvidt der er tale om en akut eller kronisk form. Vi definerer derfor nye variable til erstatning for `type` nedenfor.

- (b) *I stedet for at se på leukæmitype som en kategorisk variabel med 3 niveauer kunne man have set på de to forskellige aspekter: 'kronisk versus akut' og 'lymfatisk versus myeloid'. Definer to nye 0/1-variable, der kan bruges til at vurdere hvert af de to aspekter.*

I menuen Transform/Recode into Different Variable defineres den nye variabel akut som 1 for typerne AML og ALL (1 og 2), men 0 for den kroniske form, CML:



Tilsvarende defineres lymfatisk som 0 for typerne AML og CML, men 1 for den lymfatiske ALL:



Erstat den kategoriske variabel type med de to nye variable, og foretag en ny backwards elimination.

Vi indsætter de to nye variable **akut** og **lymfatisk** i stedet for **type**, og starter forfra, men her vises kun yderst skrabet output fra de to første kørsler, idet det er de selvsamme modeller som dem, vi så på ovenfor, **bortset fra**, at vi har skiftet reference for de to kategoriske variable **preg** og **akut** til **First** (gøres under **Categorical**), således at vi ser effekten af at donor har været gravid, respektivt effekten af akut leukæmi. Dette fremgår ikke særligt tydeligt af output (men det er nok fordi jeg ikke endnu har fundet ud af, hvordan man gør det).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	donage	,107	,083	1,685	1	,194	1,113
	preg(1)	1,662	1,181	1,981	1	,159	5,272
	logindex	1,819	,889	4,184	1	,041	6,164
	akut(1)	1,615	1,307	1,526	1	,217	5,029
	lymfatisk(1)	-,428	1,254	,117	1	,733	,652
	rcpage	,016	,082	,038	1	,846	1,016
	Constant	-5,474	2,596	4,445	1	,035	,004

og når vi har fjernet **rcpage**:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.... Lower
Step 1 ^a	donage	,114	,076	2,228	1	,136	1,120	,965
	preg(1)	1,722	1,145	2,260	1	,133	5,595	,593
	logindex	1,841	,887	4,302	1	,038	6,301	1,107
	akut	-1,691	1,262	1,796	1	,180	,184	,016
	lymfatisk	-,397	1,242	,102	1	,749	,672	,059
	Constant	-3,605	2,475	2,121	1	,145	,027	

Variables in the Equation

		95% C.I.for .. Upper
Step 1 ^a	donage	1,300
	preg(1)	52,809
	logindex	35,879
	akut	2,186
	lymfatisk	7,678
	Constant	

a. Variable(s) entered on step 1: donage, preg, logindex, akut, lymfatisk.

Herfra er det naturligt at gå videre med at fjerne variabelen *lymfatisk*, hvilket svarer til at slå de to akutte former for leukæmi sammen, og kun skelne mellem akut og kronisk leukæmi.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.... Lower
Step 1 ^a	donage	,114	,076	2,228	1	,136	1,120	,965
	preg(1)	1,722	1,145	2,260	1	,133	5,595	,593
	logindex	1,841	,887	4,302	1	,038	6,301	1,107
	akut	-1,691	1,262	1,796	1	,180	,184	,016
	lymfatisk	-,397	1,242	,102	1	,749	,672	,059
	Constant	-3,605	2,475	2,121	1	,145	,027	

Variables in the Equation

		95% C.I.for .. Upper
Step 1 ^a	donage	1,300
	preg(1)	52,809
	logindex	35,879
	akut	2,186
	lymfatisk	7,678
	Constant	

a. Variable(s) entered on step 1: donage, preg, logindex, akut, lymfatisk.

Vi bemærker, at udeladelsen af **lymfatisk** fra modellen bevirker, at **akut** får en noget øget betydning, og en lavere P-værdi. Men husk, at sammenligningen nu er *en anden end før*! Før var det sammenligningen af CML og AML, men nu er det sammenligningen af CML med *både* AML og ALL. At slå disse to sammen giver samtidig større styrke i sammenligningen.

I denne model er der tre variable med *P*-værdier omkring 10-15%, **akut**, **preg** og **donage**, så det er ikke umiddelbart klart hvilken af dem, man skal fjerne først. Man skal være meget stærk i troen for at insistere på, at det skal være **donage**, der fjernes, så i stedet har vi her valgt at fitte alle tre mulige modeller hvor en af dem er fjernet:

Uden akut:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.... Lower
Step 1 ^a	donage	,125	,066	3,583	1	,058	1,133	,996
	preg(1)	1,291	1,030	1,571	1	,210	3,637	,483
	logindex	2,106	,823	6,553	1	,010	8,216	1,638
	Constant	-5,259	2,134	6,075	1	,014	,005	

Variables in the Equation

		95% C.I. for .. Upper
Step 1 ^a	donage	1,288
	preg(1)	27,394
	logindex	41,205
	Constant	

a. Variable(s) entered on step 1: donage, preg, logindex.

Uden preg:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.... Lower
Step 1 ^a	donage	,149	,069	4,597	1	,032	1,160	1,013
	logindex	2,059	,828	6,183	1	,013	7,839	1,547
	akut(1)	1,514	1,079	1,970	1	,160	4,546	,549
	Constant	-5,853	2,262	6,693	1	,010	,003	

Variables in the Equation

		95% C.I. for .. Upper
Step 1 ^a	donage	1,329
	logindex	39,734
	akut(1)	37,659
	Constant	

a. Variable(s) entered on step 1: donage, logindex, akut.

Uden donage:

Variables in the Equation								
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.... Lower
Step 1 ^a	logindex	1,488	,720	4,273	1	,039	4,427	1,080
	akut(1)	2,251	1,106	4,140	1	,042	9,493	1,086
	preg(1)	2,496	1,101	5,135	1	,023	12,128	1,401
	Constant	-2,546	,949	7,206	1	,007	,078	

Variables in the Equation			95% C.I.for .. Upper
Step 1 ^a	logindex		18,145
	akut(1)		82,965
	preg(1)		104,987
	Constant		

a. Variable(s) entered on step 1: logindex, akut, preg.

Det fremgår af ovenstående, at når en af variablene fjernes, så bliver mindst en af de andre signifikante. I den sidste model er såvel **akut** som **preg** signifikante, så det vil næppe være en god ide at udelade nogen af disse. Hvis vi får kniven for struben og **skal** vælge en af modellerne, vil vi derfor vælge denne sidste.

(c) *Overvej forskellen på resultaterne fra de to analyser i 3a og 3b.*

Nu stoppede vi jo sådan set den baglæns elimination i 3a inden vi var nået til vejs ende. Hvis vi ikke havde erstattet **type** med **akut** og **lymfatisk**, ville vi have fortsat eliminationen ved at fjerne **type** helt fra modellen, og så ville vi have fortsat helt ned til en model kun indeholdende **donage** og **logindex**, altså **en helt anden konklusion** end ovenfor. Index-variablen ville godt nok stadig være signifikant, men med en væsentlig større estimeret effekt, men **preg** er ikke med i modellen mere, og vi fik ikke opdaget, at kronisk og akut leukæmi giver forskellig risiko. Til gengæld ville vi så vurdere ældre donorer til at være mere risikable, hvilket måske kan skyldes, at ældre donorer oftere har været gravide.....

Faktisk er det ikke let at svare på, hvad der *egentlig* har betydning for forekomsten af GvHD, fordi vi kun har sølle 17 tilfælde alt i alt, og derfor absolut ikke burde bruge mere end 2 parametre på kovariater!!

Vi må (endnu en gang) konstatere, at vi har været på fisketur!!

4. *Find odds-ratioerne associeret med de binære forklarende variable i slutmodellen fra 3b, med tilhørende 95% konfidensintervaller.*

De estimerede odds-ratioer fremgår direkte af den sidste del af output fra den logistiske regression:

- Folk med kronisk leukæmi har en odds for GvHD, der er 9.49 gange højere end folk med akut leukæmi, men med et meget bredt konfidensinterval: CI=(1.09, 82.97).
- Patienter, der modtager knoglemarv fra en donor, der har været gravid, har en odds for GvHD, der er 12.13 gange højere end de, der modtager den fra en donor, der ikke har været gravid, igen med et meget bredt konfidensinterval: CI=(1.40, 104.99).

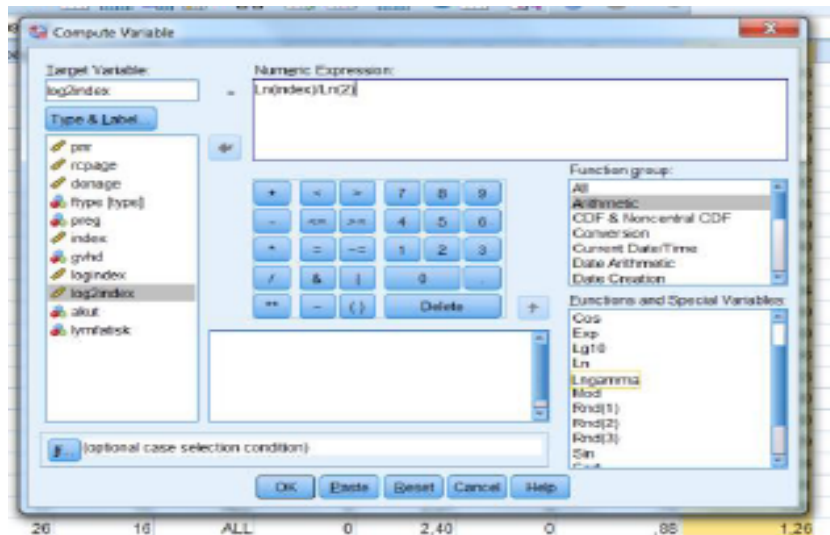
Man ser, at selv om de estimerede odds-ratioer er ret store og signifikant forskellige fra 1, er den nedre grænse for begge ret tæt på 1, så der er ikke megen evidens for substantielle effekter af disse to variable.

5. *Giv en verbal fortolkning af koefficienten til $\log(\text{index})$, gerne en, du kunne anvende, hvis du talte til en kongres.*
Hvis du på baggrund af dette hellere ville have anvendt en anden logaritmefunktion, så skift den ud nu.

Den odds ratio, der er angivet for `logindex` (altså 4.427) svarer til OR mellem to patienter der adskiller sig med 1 i værdien af `logindex`, dvs. mellem to patienter, hvis forhold mellem indexværdier er $e = 2.7183$. Det er næppe en intuitivt rimelig måde at rapportere effekten af `index` på.

Disse fortolkningsproblemer gør, at vi gerne vil skifte logaritme, f.eks. til 2-tals logaritmen, som vi så må gå ind i `Transform/Compute` og

definere. Da denne ikke findes som særskilt funktion i SPSS, må vi benytte formelen $\log_2 \text{index} = \ln(\text{index}) / \ln(2)$.



En forskel på 1 enhed i $\log_2 \text{index}$ svarer til en fordobling af indexværdien, hvilket er en noget mere mundret fortolkning.

Hvis vi fitter en model med $\log_2 \text{index}$, får vi:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	akut(1)	2,251	1,106	4,140	1	,042	9,493
	preg(1)	2,496	1,101	5,135	1	,023	12,128
	log2Index	1,031	,499	4,273	1	,039	2,804
	Constant	-2,546	,949	7,206	1	,007	,078

Variables in the Equation

		95% C.I. for EXP(B)	
		Lower	Upper
Step 1 ^a	akut(1)	1,086	82,965
	preg(1)	1,401	104,987
	log2Index	1,055	7,456
	Constant		

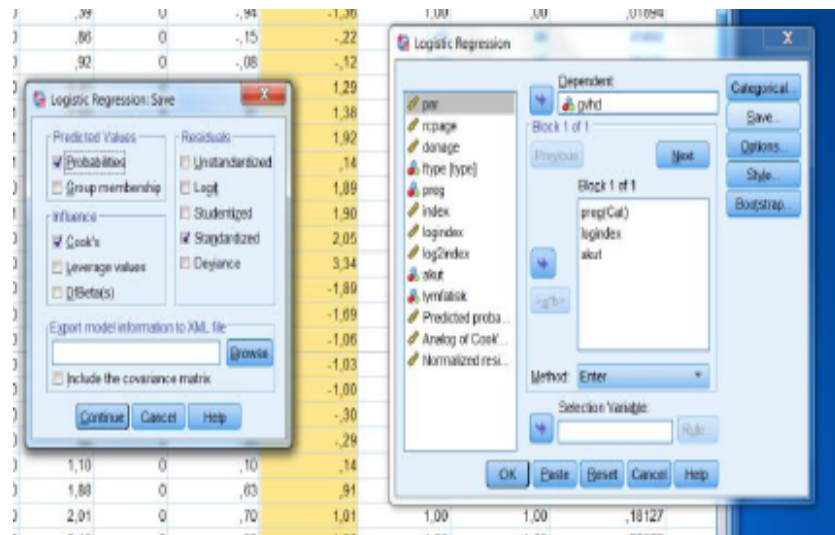
a. Variable(s) entered on step 1: akut, preg, log2Index.

Estimatet for odds-ratio associeret med $\log_2(\text{index})$, dvs. med en fordobling af index-værdien, bliver 2.80, med 95% konfidensintervallet (1.06, 7.46).

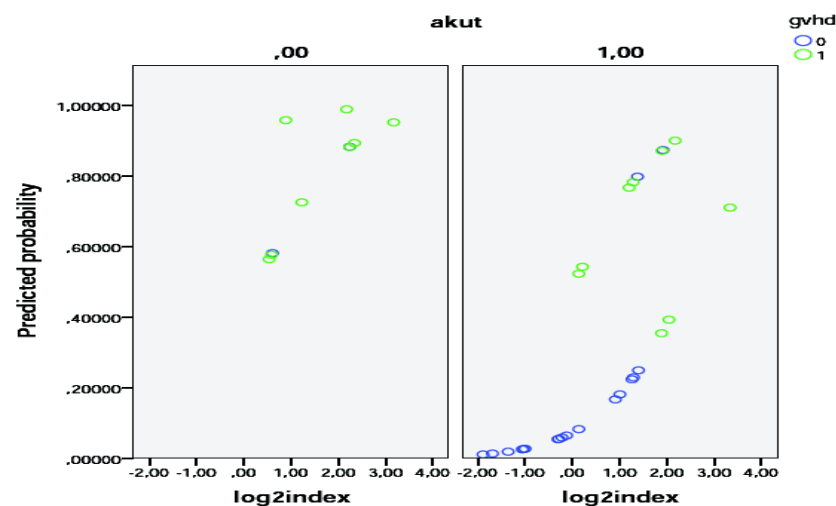
Bemærk, at fittet af de to modeller er fuldstændigt det samme, det er kun koefficienten til $\log_2\text{index}$, der er anderledes, testet og P -værdien er den samme. I princippet kunne man i stedet for at vælge 2-talslogaritmen have valgt 10-talslogaritmen. Den estimerede odds-ratio ville da svare til effekten af en 10-dobling af index-værdien. Det ville svare nogenlunde til range af index, så en reference til fordobling af værdien vil nok i dette materiale være mest klinisk relevant, mens man i andre materialer vil støde på variable som i populationen varierer meget mere, og hvor effekten af en 10-dobling ville kunne være relevant. Hvilken logaritme, man med fordel kan anvende i forbindelse med en kovariat, afhænger altså helt af den konkrete sammenhæng, og er *ikke* et statistisk spørgsmål.

Hvis man har tilstrækkeligt med datagrundlag kan man lave modelkontrol ved at at gemme relevante værdier til senere optegning. Benyt her-

til Save-knappen i opsætningen af den logistiske regressionsanalyse, og gem f.eks. Predicted probabilities, Cook's, Standardized residuals og DfBeta(s):



Herefter kunne man f.eks. optegne predikterede værdier mod log2index, opdelt efter akut:

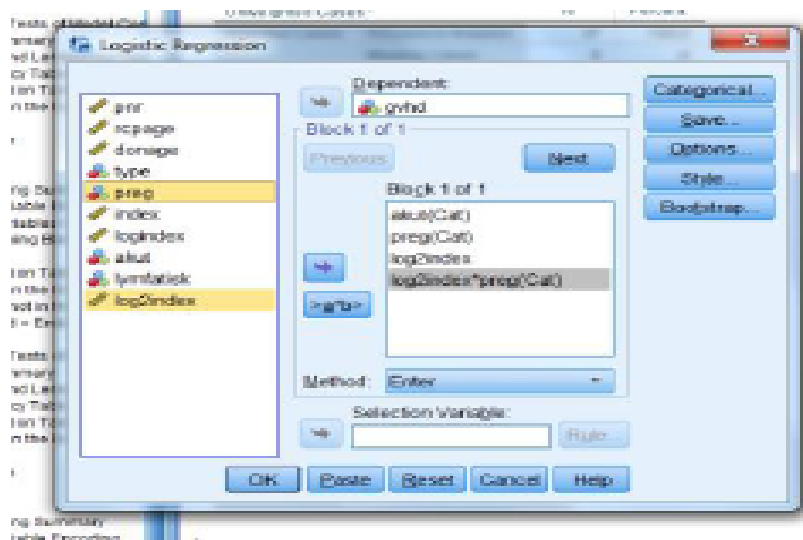


På hver af disse figurer anes to logistiske kurver, alt efter om donor har været gravid (preg=1) eller ej (preg=0).

6. Stod der mon noget i protokollen om at undersøge nogle interessante interaktioner? Prøv f.eks. interaktionen mellem `logindex` og `preg`, og giv en beskrivelse af resultatet.

Her forestiller vi os, at index-værdien kunne have forskellig prognostisk betydning, afhængig af, om donor har været gravid eller ej.

Vi tester denne interaktion *som sædvanlig*, ved at inkludere leddet `preg*log2index`:



og finder så

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	akut(1)	2,503	1,266	3,907	1	,048	12,216
	preg(1)	6,301	2,877	4,796	1	,029	544,868
	log2index	1,724	,851	4,106	1	,043	5,606
	log2index by preg(1)	-2,878	1,709	2,837	1	,092	,056
	Constant	-3,510	1,464	5,746	1	,017	,030

Variables in the Equation

		95% C.I. for EXP(B)	
		Lower	Upper
Step 1 ^a	akut(1)	1,021	146,108
	preg(1)	1,939	153139,722
	log2index	1,058	29,703
	log2index by preg(1)	,002	1,601
	Constant		

a. Variable(s) entered on step 1: akut, preg, log2index, log2index * preg .

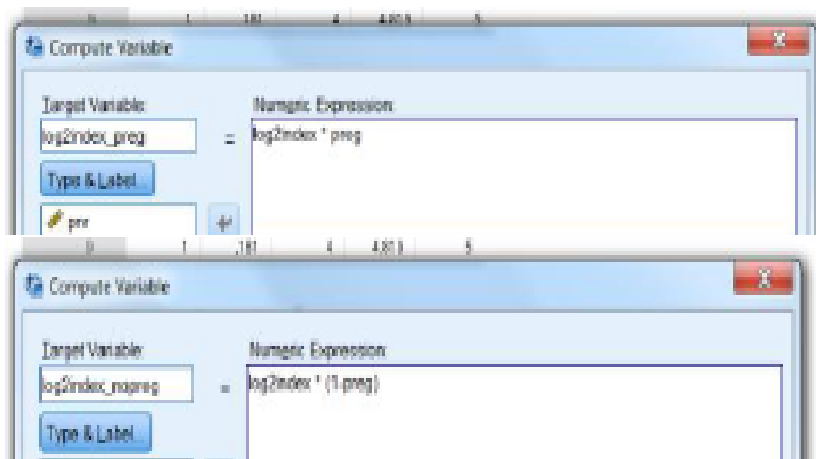
Med en P-værdi på 9% er der en svag indikation af forskel på effekten af index-variablen, afhængig af om donor har været gravid eller ej.

For at estimere de to effekter, kan man definere to versioner af **log2index**, således

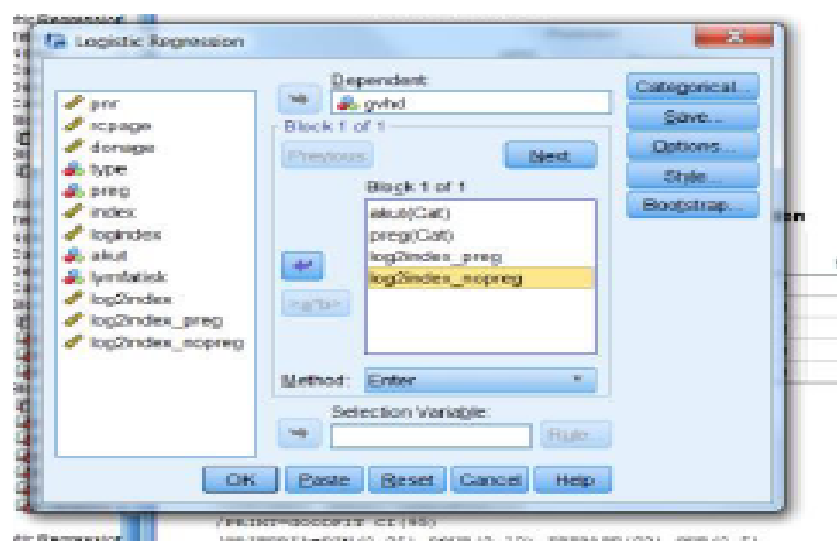
log2index_preg = log2index, hvis preg=1 og 0 ellers
 log2index_nopreg = log2index, hvis preg=0 og 0 ellers

Dette kan også skrives, som det ses i boksene nedenfor

- $\text{log2index_preg} = \text{log2index} * \text{preg}$
- $\text{log2index_nopreg} = \text{log2index} * (1 - \text{preg})$



Vi setter nu disse 2 ind som kovariater i stedet for log2index og preg*log2index:



hvorved vi får samme model som før, men nye estimater:

		Variables in the Equation					
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	akut(1)	2,503	1,266	3,907	1	,048	12,216
	preg(1)	6,301	2,877	4,796	1	,029	544,868
	log2index_preg	-1,154	1,470	,616	1	,432	,315
	log2index_nopreg	1,724	,851	4,106	1	,043	5,606
	Constant	-3,510	1,464	5,746	1	,017	,030

		Variables in the Equation	
		95% C.I. for EXP(B)	
		Lower	Upper
Step 1 ^a	akut(1)	1,021	146,108
	preg(1)	1,939	153139,722
	log2index_preg	,018	5,624
	log2index_nopreg	1,058	29,703
	Constant		

a. Variable(s) entered on step 1: akut, preg, log2index_preg, log2index_nopreg.

Vi ser, at effekten af index-variablen ser ud til at være modsat rettet for patienter med og uden gravid donor, men som nævnt er dette altså ikke signifikant og skal derfor ikke fortolkes yderligere.

Opgave 2: Mordsager

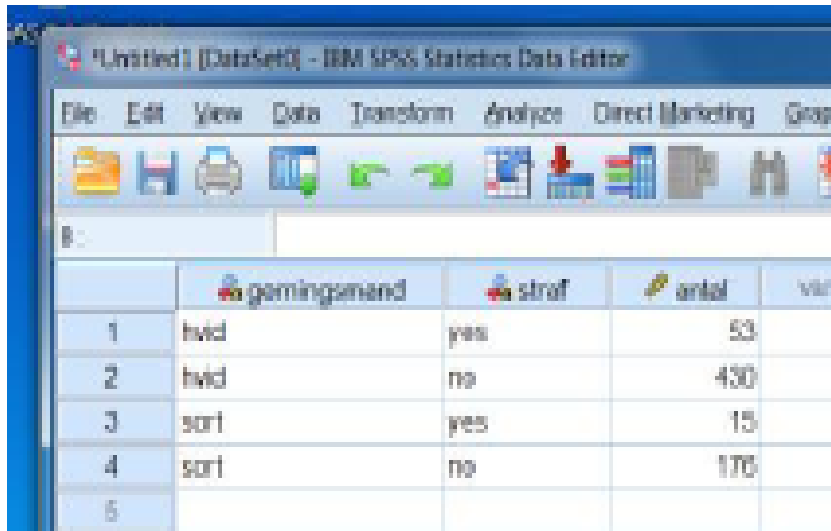
Over en årrække har man indsamlet materiale om mordsager i USA. Vi skal her se på en del af dette materiale, nemlig de sager, hvor den tiltalte blev dømt for mord.

Der var i alt 191 sorte og 483 hvide, der blev dømt for mord, og af disse fik 15 sorte og 53 hvide en dødsdom.

1. Opstil en 2×2 -tabel til sammenligning af sandsynlighederne for en dødsdom for hvide hhv. sorte, og kvantificer forskellen, med konfidensinter-

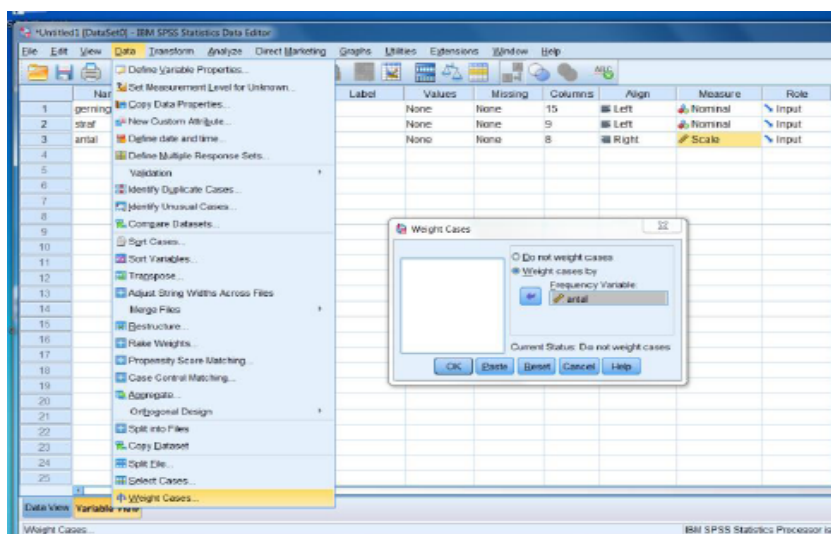
val. Ser konklusionen fornuftig ud?

Vi kan se, at der må være $191-15=176$ sorte, der *ikke* fik dødsstraf, og tilsvarende $483-53=430$ hvide. Vi indlæser derfor et datasæt med 4 linier og 3 variable: gerningsmand, straf og antal:

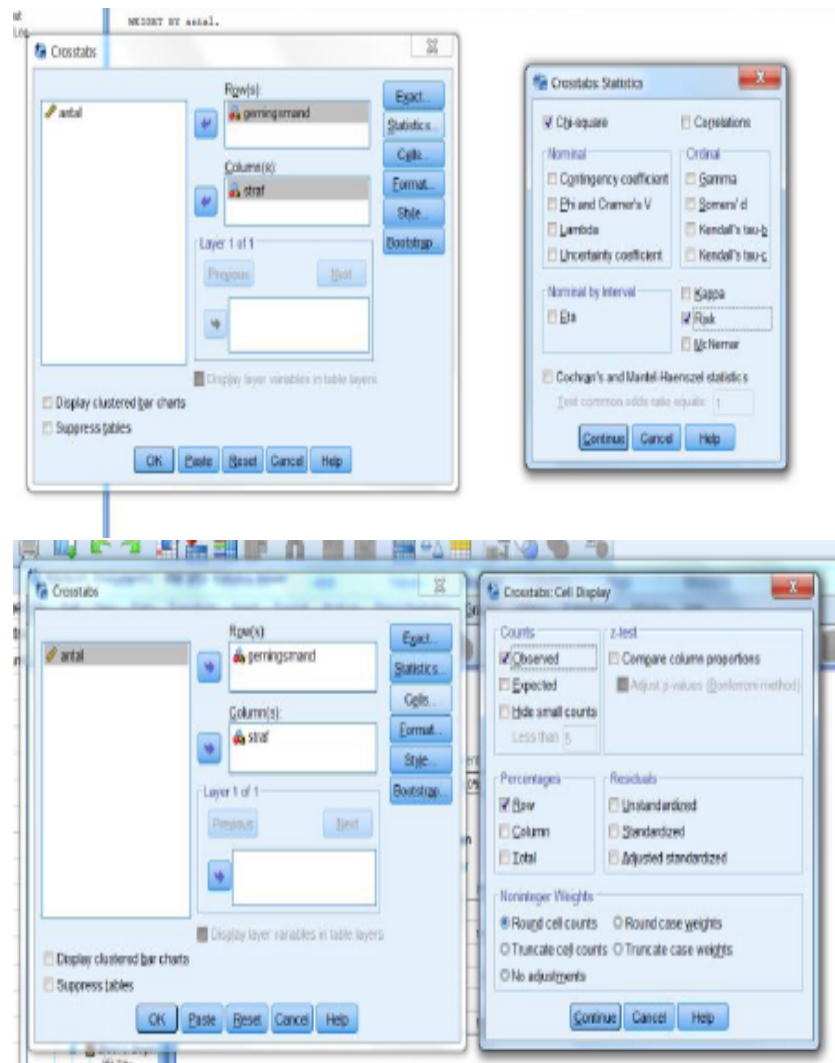


	gerningsmand	straf	antal	var
1	hvid	yes	53	
2	hvid	no	430	
3	sort	yes	15	
4	sort	no	176	
5				

Før vi kan give os til at regne på disse data, skal vi lade programmet vide, at variabelen **antal** angiver, hvor mange gange, den pågældende patienttype forekommer. Dette gøres ved at gå ind i **Data/Weight Cases**, afkrydse **Weight cases by** og sætte **antal** over i **Frequency Variable**:



Herefter kan vi opstille en 2*2-tabel og udføre χ^2 -testet for uafhængighed ved at gå ind i Analyze/Descriptive Statistics/Crosstabs/, og sætte gerningsmand over i Row(s) og straf i Column(s), samt under Statistics at afkrydse Chi-Square, Somers' d og Risk. Desuden går vi ind i Cells og afkrydser Row percentages:



Herved får vi outputtet (beskåret):

gerningsmand * straf Crosstabulation

		straf		Total
		no	yes	
gerningsmand	hvid	Count	430	53
		% within gerningsmand	89,0%	11,0%
	sort	Count	176	15
		% within gerningsmand	92,1%	7,9%
Total		Count	606	68
		% within gerningsmand	89,9%	10,1%

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)	Exact Sig. (2- sided)	Exact Sig. (1- sided)
Pearson Chi-Square	1,469 ^a	1	,226		
Continuity Correction ^b	1,145	1	,285		
Likelihood Ratio	1,536	1	,215		
Fisher's Exact Test				,258	,142
N of Valid Cases	674				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 19,27.

b. Computed only for a 2x2 table

Directional Measures

			Value	Asymptotic Standard Error ^a	Approximate T ^b
Ordinal by Ordinal	Somers' d	Symmetric	-,043	,033	-1,293
		gerningsmand Dependent	-,070	,054	-1,293
		straf Dependent	-,031	,024	-1,293

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for gerningsmand (hvid / sort)	,691	,380	1,259
For cohort straf = no	,966	,917	1,018
For cohort straf = yes	1,397	,808	2,417
N of Valid Cases	674		

Vi ser, at såvel χ^2 -testet som Fishers eksakte test (som ikke er nødvendigt her med så store antal) giver P -værdier over 0.2, og der er altså *ikke* evidens for forskelsbehandling af hvide og sorte.

Dette betyder ikke nødvendigvis, at der ikke *er* forskelsbehandling, og kigger vi på de rå hyppigheder, ser vi, at hvide har en hyppighed af dødsstraf på 11.0%, medens sorte har 7.9%. Der er således en lille **overhyppighed af dødsstraf blandt hvide gerningsmænd**, og denne kan kvantificeres på forskellig vis:

- Differens mellem hyppigheder:
3.1%point (ses under **Directional Measures** i den række, der betegnes **straf Dependent**, med approksimativ standard error på 0.024, dvs. med 95% konfidensinterval, som udregnes til ca. (-1.7%, 7.9%).
- Odds Ratio for dødsstraf blandt hvide i forhold til sorte:
 $\frac{1}{0.691} = 1.45$, med konfidensinterval $(\frac{1}{1.259}, \frac{1}{0.380}) = (0.79, 2.63)$.
- Relativ risiko for dødsstraf blandt hvide i forhold til sorte:
1.397, med CI=(0.808, 2.417).

Hvis vi opdeler data efter offerets race, får vi nedenstående to 2×2 -tabeller:

Sort offer:

Dømte	Dødsstraf?		Total
	nej	ja	
Sort	139	4	143
Hvid	16	0	16

Hvidt offer:

Dømte	Dødsstraf?		Total
	nej	ja	
Sort	37	11	48
Hvid	414	53	467

Disse data svarer til de 8 linier vist nedenfor, og vi har nu 4 kolonner, offer, gerningsmand, straf og antal:

	offer	gerningsmand	straf	antal
1	hvid	hvid	ja	53
2	hvid	hvid	nej	414
3	hvid	sort	ja	11
4	hvid	sort	nej	37
5	sort	hvid	ja	0
6	sort	hvid	nej	16
7	sort	sort	ja	4
8	sort	sort	nej	139

2. Kan offerets race tænkes at være en confounder for problemstillingen?

En confounder er pr. definition en variabel, der hænger sammen med såvel “ekspositionen” (her gerningsmandens race) som outcome (risikoen for dødsstraf).

Man kan let forestille sig, at der er en sammenhæng mellem offer og gerningsmands race, og man kan (desværre) også godt forestille sig, at der er en effekt af offerets race på resultatet af dommen. Så **ja**, offerets race kan sagtens tænkes at være en confounder i denne problematik.

Vi checker lige for øvelsens skyld. Derfor laver vi to tabeller

- En tabel, der undersøger, om der er sammenhæng mellem offerets race og gerningsmandens race (kovariaten i fokus):

offer * gerningsmand Crosstabulation

Count

		gerningsmand		Total
		hvid	sort	
offer	hvid	467	48	515
	sort	16	143	159
Total		483	191	674

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)	Exact Sig. (2- sided)	Exact Sig. (1- sided)
Pearson Chi-Square	388,807 ^a	1	,000		
Continuity Correction ^b	384,847	1	,000		
Likelihood Ratio	380,569	1	,000		
Fisher's Exact Test				,000	,000
N of Valid Cases	674				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 45,06.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for offer (hvid / sort)	86,954	47,913	157,808
For cohort gerningsmand = hvid	9,011	5,657	14,353
For cohort gerningsmand = sort	,104	,079	,136
N of Valid Cases	674		

Her ses en helt klar sammenhæng, og den relative risiko for at have en hvid gerningsmand er over 9, hvis offeret selv er hvid.

- En tabel, der undersøger, om der er sammenhæng mellem offerets race og selve dommen (outcome)

offer * straf Crosstabulation				
Count				
		straf		Total
		ja	nej	
offer	hvid	64	451	515
	sort	4	155	159
Total		68	606	674

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13,157 ^a	1	,000		
Continuity Correction ^b	12,087	1	,001		
Likelihood Ratio	16,872	1	,000		
Fisher's Exact Test				,000	,000
N of Valid Cases	674				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 16,04.

b. Computed only for a 2x2 table

Risk Estimate			
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for offer (hvid / sort)	5,499	1,970	15,349
For cohort straf = ja	4,940	1,828	13,352
For cohort straf = nej	,898	,862	,936
N of Valid Cases	674		

Her ses også en tydelig sammenhæng, idet den relative risiko for dødsstraf for hvide vs. sorte ofre er 4.94, med CI=(1.83, 13.35).

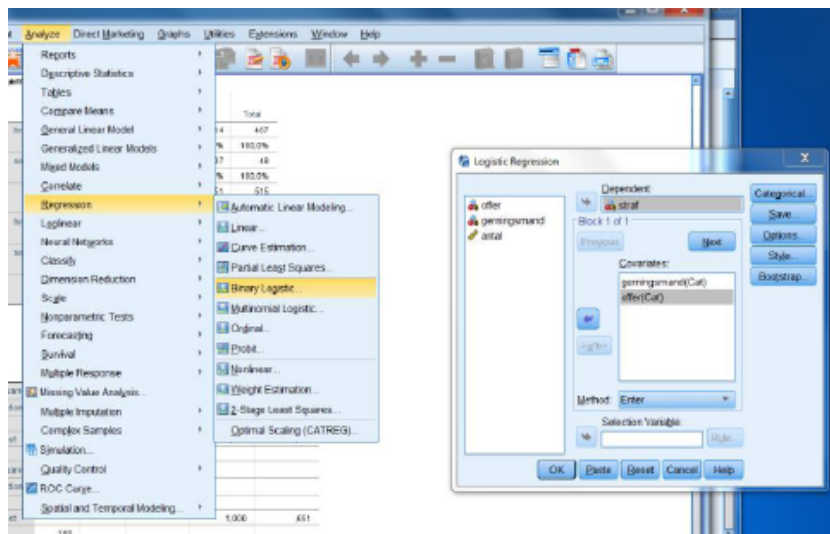
Baseret på ovenstående tabel-analyser er vi således blevet overbevist om, at offerets rolle kan agere som confounder for gerningsmandens race, når man vurderer sidstnævntes betydning for en evt. dødsdom.

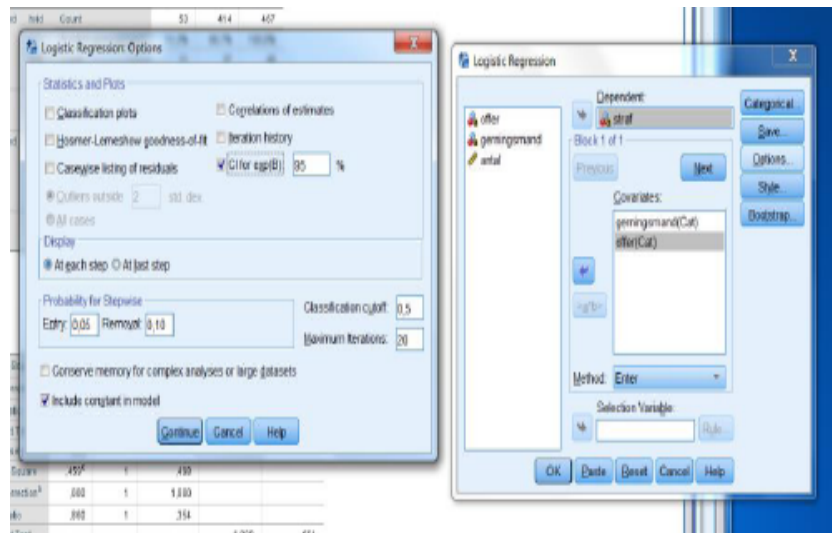
Vi ser nemlig fra den første tabel, at der i de fleste tilfælde er overensstemmelse mellem offers og gerningsmands race, faktisk i $\frac{467+143}{674}$, svarende til hele 91.51% af situationerne.

Fra den anden tabel ses, at der langt oftere gives dødsdom, når der er tale om et hvidt offer.

3. Lav en logistisk regressionsmodel for sandsynligheden for en dødsdom, med såvel offerets race som gerningsmandens race som forklarende variable.

Da vi nu har to kovariater (race for såvel offer som gerningsmand), kan vi ikke klare os med tabel-analyse, men må gå over til logistisk regression (ligesom i den forrige opgave), med **straf** som **Dependent** og såvel **offer** som **gerningsmand** som **Covariate(s)**:





Vi får derved outputtet

Dependent Variable Encoding

Original Value	Internal Value
ja	0
nej	1

Categorical Variables Codings

			Parameter coding (1)
Frequency			
offer	hvid	4	1,000
	sort	3	,000
gerningsmand	hvid	3	1,000
	sort	4	,000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	offer(1)	-2,404	,601	16,026	1	,000	,090
	gerningsmand(1)	,868	,367	5,589	1	,018	2,382
	Constant	3,596	,507	50,326	1	,000	36,456

Variables in the Equation

		95% C.I. for EXP(B)	
		Lower	Upper
Step 1 ^a	offer(1)	,028	,293
	gerningsmand(1)	1,160	4,890
	Constant		

a. Variable(s) entered on step 1: offer, gerningsmand.

Her skal man holde tungen lige i munden, for det ses af ovenstående, at vi modellerer sandsynligheden for **straf=nej** (dette ses af **Internal value=1**)! Desuden er **sort** referencen for begge de to kovariater.

Med dette i baghovedet, kan vi se, at Odds Ratio for dødsstraf for sorte i forhold til hvide (gerningsmænd) estimeres til 2.38, med CI=(1.16, 4.89), altså at sorte *oftere* får dødsstraf end hvide! Og resultatet er signifikant ($P = 0.018$).

Det er i **kraftig modsætning** til det, vi fandt i spørgsmål 1!

Endvidere ser vi, at OR for dødsstraf er mere end 10 gange højere, hvis offeret er hvidt i forhold til, hvis offeret er sort ($\frac{1}{0.090} = 11.11$).

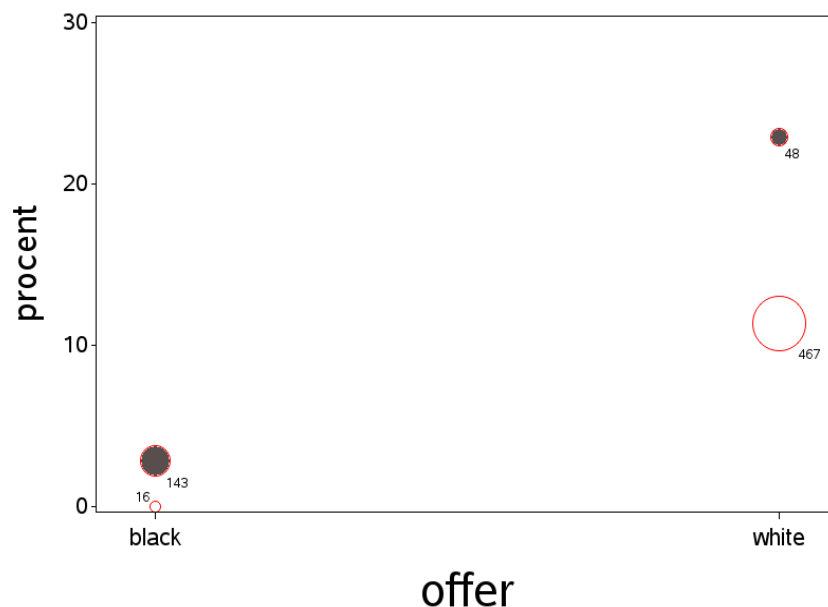
Sammenlign til resultatet i spørgsmål 1.

I spørgsmål 1 fandt vi, at Odds Ratio for dødsstraf for hvide gerningsmænd i forhold til sorte gerningsmænd var 1.45, med

CI=(0.79, 2.63), svarende til, at *hvide* oftere får dødsstraf.

Vi har altså totalt fået vendt op og ned på konklusionen, fordi vi tog hensyn til offerets race. Og vi kan pludselig se en signifikans, som vi ikke kunne se tidligere. **Hvad foregår der??**

Vi kan prøve at anskueliggøre situationen ved hjælp af en figur, hvor vi afbilder procenten af dødsstraf, som funktion af offerets race, med symboler svarende til gerningsmandens race, og hvor *størrelsen* af symbolerne (circklerne) svarer til antallet af mordsager i den pågældende kategori:



På denne tegning ses, at hvide ofre oftere giver anledning til dødsdom, samt at sorte gerningsmænd oftere bliver dødsdømt, uanset offerets race (de sorte circkler ligger højest).

Totalt set er der bare rigtigt mange af de hvide gerningsmænd, der myrder et hvidt offer, og *derfor* får de totalt set oftere dødsstraf end de sorte gerningsmænd.

4. Er der grund til at tro, at der kan være interaktion mellem offerets race og gerningsmandens race? Og er der evidens for en sådan i disse data?

Baseret på figuren ovenfor og kommentarerne til den, er der intet, der tyder på en interaktion mellem gerningsmands og offers race. Alligevel vil vi undersøge sagen.

Vi kan først prøve at analysere to 2×2 -tabeller, idet vi opdeler efter offerets race, for at vurdere effekten af gerningsmandens race i hver af de to situationer.

Vi går derfor ind i **Data/Split File**, sætter flueben ved **Compare Groups** og sætter **offer** over i **Groups Based on**. Herefter udfører vi analyserne hver for sig:

Case Processing Summary

		Valid		Cases Missing		Total	
		N	Percent	N	Percent	N	Percent
hvid	gerningsmand * straf	515	100,0%	0	0,0%	515	100,0%
sort	gerningsmand * straf	159	100,0%	0	0,0%	159	100,0%

gerningsmand * straf Crosstabulation

				straf		
offer				ja	nej	Total
hvid	gerningsmand	hvid	Count	53	414	467
			% within gerningsmand	11,3%	88,7%	100,0%
		sort	Count	11	37	48
			% within gerningsmand	22,9%	77,1%	100,0%
	Total		Count	64	451	515
			% within gerningsmand	12,4%	87,6%	100,0%
sort	gerningsmand	hvid	Count	0	16	16
			% within gerningsmand	0,0%	100,0%	100,0%
		sort	Count	4	139	143
			% within gerningsmand	2,8%	97,2%	100,0%
	Total		Count	4	155	159
			% within gerningsmand	2,5%	97,5%	100,0%

Chi-Square Tests

offer		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)
hvid	Pearson Chi-Square	5,352 ^a	1	,021	
	Continuity Correction ^b	4,342	1	,037	
	Likelihood Ratio	4,534	1	,033	
	Fisher's Exact Test				,035
	N of Valid Cases	515			
sort	Pearson Chi-Square	,459 ^c	1	,498	
	Continuity Correction ^b	,000	1	1,000	
	Likelihood Ratio	,860	1	,354	
	Fisher's Exact Test				1,000
	N of Valid Cases	159			

Chi-Square Tests

offer		Exact Sig. (1-sided)
hvid	Pearson Chi-Square	
	Continuity Correction ^b	
	Likelihood Ratio	
	Fisher's Exact Test	,024
	N of Valid Cases	
sort	Pearson Chi-Square	
	Continuity Correction ^b	
	Likelihood Ratio	
	Fisher's Exact Test	,651
	N of Valid Cases	

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,97.

b. Computed only for a 2x2 table

c. 2 cells (50,0%) have expected count less than 5. The minimum expected count is ,40.

Risk Estimate

offer		Value	95% Confidence Interval	
			Lower	Upper
hvid	Odds Ratio for gerningsmand (hvid / sort)	,431	,207	,895
	For cohort straf = ja	,495	,278	,882
	For cohort straf = nej	1,150	,982	1,346
	N of Valid Cases	515		
sort	For cohort straf = nej	1,029	1,001	1,058
	N of Valid Cases	159		

Disse tabeller viser, at hvis offeret er sort, er der ingen signifikant forskel på sorte og hvide gerningsmænd (idet det dog bemærkes, at der er *yderst få* dødsdomme i denne tabel, og slet ingen for hvide gerningsmænd). Sådan som tabellen er stillet op, forsøger programmet at udregne en OR, hvor der divideres med 0, og derfor angives ikke noget estimat. Vi må selv udregne den omvendte OR (altså OR for en dødsdom for hvide vs. sorte) til....0, selvfølgelig.

Hvis offeret derimod er hvid, ses der en signifikant forskel på hyppighederne af dødsdom for de to racer, idet sorte dobbelt så hyppigt ($P=0.035$ for det eksakte test) får en dødsdom. Relativ risiko estimeres til $\frac{1}{0.495} = 2.0$, med $CI=(\frac{1}{0.862}, \frac{1}{0.278}) = (1.1, 3.6)$, og odds ratio estimeres (ligeledes ved at invertere) til 2.32 med $CI=(1.12, 4.83)$.

På grund af de ganske få dødsdomme for sager med sort offer, kan vi ikke på denne baggrund sige, at der er interaktion mellem gerningsmands og offers race.

Vi kan også prøve at teste det direkte i en samlet logistisk regression, ved blot at tilføje interaktionsleddet **offer*gerningsmand**:

Variables in the Equation					
		B	S.E.	Wald	Sig.
Step 1 ^a	gerningsmand(1)	17,655	10048,244	,000	,999
	offer(1)	-2,335	,612	14,536	,000
	gerningsmand(1) by offer(1)	-16,812	10048,244	,000	,999
	Constant	3,548	,507	48,950	,000

Variables in the Equation			
	Exp(B)	95% C.I. for EXP(B)	
		Lower	Upper
Step 1 ^a	gerningsmand(1)	46488485,03	,000
	offer(1)	,097	,029
	gerningsmand(1) by offer(1)	,000	,000
	Constant	34,750	

Her ser vi noget **mystisk output**, og det skyldes formentlig, at der er et antal på 0, som SPSS ikke umiddelbart kan håndtere. Testet for interaktion giver i SAS en P-værdi på 0.54, således at man *ikke* kan påvise nogen interaktion mellem gerningsmands og offers racer ($P = 0.54$).

Konklusionen af vores analyser er altså:

1. Der gives langt oftere dødsdom, når offeret er hvidt i forhold til, når offeret er sort.
2. Da de fleste bliver myrdet af en af deres egen race, bevirker punkt 1, at et højt antal hvide dødsdømmes.
3. **For given race af offeret**, får flere sorte end hvide en dødsdom.