

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.txt` 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)

Data indlæses direkte fra hjemmesiden, og samtidig konstruerer vi nogle dummy variable for hver af de tre leukæmi typer, en mere sigende variabel for leukæmityper, samt to forskellige logaritmetransformationer af **index**, som vi senere får brug for.

Bemærk, at variablen angivet ovenfor som **type** her indlæses som **grp** og bagefter laves om til en mere informativ version (med karakter værdier), som så kaldes **type**.

Endelig laver vi en `proc means;`, i håb om at få et overblik af en art:

```

FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/gvhd.txt";

data gvhd;
infile navn firstobs=2;
input pnr rcpage donage grp preg index gvhd;

* Dannelselse af mere beskrivende variabel for type ;
if grp=1 then type="aml";
if grp=2 then type="all";
if grp=3 then type="cml";

* Dannelselse af indikatorvariable for alle tre grupper ;
aml = (grp = 1) ;
all = (grp = 2) ;
cml = (grp = 3) ;

* Logaritmetransformation med naturlig og 2-talslogaritme ;
logindex = log(index) ;
log2index = log2(index) ;
run ;

proc means data = gvhd;
run;

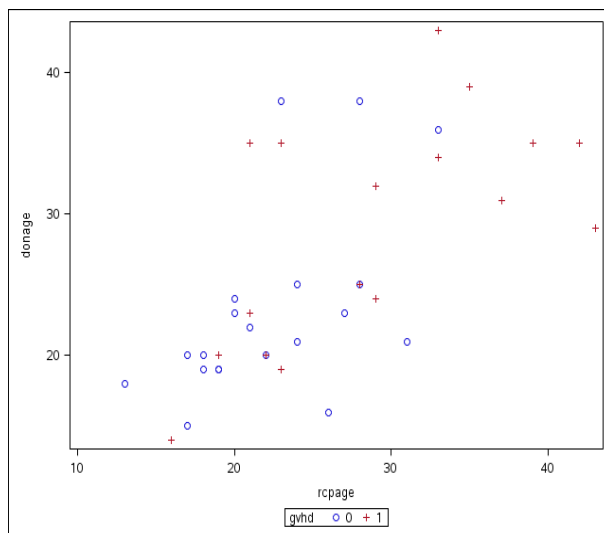
```

Herved får vi outputtet

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
pnr	37	19.0000000	10.8243553	1.0000000	37.0000000
rcpage	37	25.4324324	7.5038529	13.0000000	43.0000000
donage	37	25.8108108	7.8380760	14.0000000	43.0000000
grp	37	1.9729730	0.7632710	1.0000000	3.0000000
preg	37	0.2702703	0.4502252	0	1.0000000
index	37	2.5562162	2.2299382	0.2700000	10.1100000
gvhd	37	0.4594595	0.5052279	0	1.0000000
aml	37	0.2972973	0.4633732	0	1.0000000
all	37	0.4324324	0.5022472	0	1.0000000
cml	37	0.2702703	0.4502252	0	1.0000000
logindex	37	0.5722071	0.9177197	-1.3093333	2.3135250
log2index	37	0.8255203	1.3239896	-1.8889687	3.3377111

Vi bemærker fra `proc means`, 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 et simpelt scatterplot, hvor der også er tilføjet farver alt efter om patienten *har* fået graft vs host disease (rød) eller *ikke har* (blå).



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, så vi benytter koden:

```
proc freq data = gvhd;
tables type*gvhd / nocol nopercnt expected chisq;
run;
```

og får outputtet

The FREQ Procedure
Table of type by gvhd

type	gvhd		
Frequency			
Expected			
Row Pct	0	1	Total
-----+-----+-----+			
all	12	4	16
	8.6486	7.3514	
	75.00	25.00	
-----+-----+-----+			
aml	6	5	11
	5.9459	5.0541	
	54.55	45.45	
-----+-----+-----+			
cml	2	8	10
	5.4054	4.5946	
	20.00	80.00	
-----+-----+-----+			
Total	20	17	37

Statistics for Table of type by gvhd

Statistic	DF	Value	Prob

Chi-Square	2	7.4970	0.0236
Likelihood Ratio Chi-Square	2	7.8884	0.0194

Sample Size = 37

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 `lindex`).

Vi trodser forslaget om variablenavn og ser i første omgang på den logaritmetransformerede variabel `logindex=log(index)`, hvor vi altså

har brugt den naturlige logaritme. Husk, at det ikke gør nogen forskel for modellen, kun for de estimer, man vælger at fokusere på.

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!

Vi benytter `proc logistic`, da dette er det letteste, når man skal lave traditionel logistisk regression, og vi afstår fra modelkontrol (derfor udkommenteret af programmet nedenfor), da materialet er for lille til, at det vil give mening:

```
proc logistic data=gvhhd;
class type preg / param=GLM;
model gvhhd(event="1") = type logindex preg donage rcpage /
                        /*lackfit*/ link=logit;
*output out=pred lower=lower pred=phat upper=upper reschi=reschi;
run;
```

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

The LOGISTIC Procedure

Model Information

Data Set	WORK.GVHD
Response Variable	gvhd
Number of Response Levels	2
Model	binary logit

Number of Observations Read	37
Number of Observations Used	37

Response Profile

Ordered Value	gvhd	Total Frequency
1	0	20
2	1	17

Probability modeled is gvhhd=1.

Type 3 Analysis of Effects

Wald

Effect	DF	Chi-Square	Pr > ChiSq
type	2	2.4617	0.2920
logindex	1	4.1836	0.0408
preg	1	1.9812	0.1593
donage	1	1.6847	0.1943
rcpage	1	0.0379	0.8457

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.1966	3.3211	0.4375	0.5083
type all	1	-2.0434	1.3677	2.2320	0.1352
type aml	1	-1.6152	1.3074	1.5261	0.2167
type cml	0	0	.	.	.
logindex	1	1.8187	0.8892	4.1836	0.0408
preg 0	1	-1.6624	1.1810	1.9812	0.1593
preg 1	0	0	.	.	.
donage	1	0.1072	0.0826	1.6847	0.1943
rcpage	1	0.0160	0.0820	0.0379	0.8457

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
type all vs cml	0.130	0.009 1.891
type aml vs cml	0.199	0.015 2.579
logindex	6.164	1.079 35.215
preg 0 vs 1	0.190	0.019 1.920
donage	1.113	0.947 1.309
rcpage	1.016	0.865 1.193

Vi bemærker, 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.

```
proc logistic data=gvhhd;
class type preg / param=GLM;
model gvhd(event="1") = type logindex preg donage;
run;
```

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

Type 3 Analysis of Effects				
Effect	DF	Wald Chi-Square	Pr > ChiSq	
type	2	2.7621	0.2513	
logindex	1	4.3020	0.0381	
preg	1	2.2601	0.1327	
donage	1	2.2277	0.1356	

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-1.8830	2.9094	0.4189	0.5175	
type all	1	-2.0880	1.3424	2.4194	0.1198	
type aml	1	-1.6913	1.2620	1.7959	0.1802	
type cml	0	0	.	.	.	
logindex	1	1.8407	0.8875	4.3020	0.0381	
preg 0	1	-1.7219	1.1453	2.2601	0.1327	
preg 1	0	0	.	.	.	
donage	1	0.1135	0.0761	2.2277	0.1356	

Der skete ingen særlige ændringer af de øvrige estimer, da vi smed denne variabel 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 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, så vi ved faktisk ikke om forskellen mellem AML og CML er den mindst signifikante forskel mellem diagnostiske grupper. For at belyse forskellen mellem AML og ALL skal en af disse gøres til reference-niveau. Vi vælger her ALL. Det kan lettest gøres ved at tilføje (ref="all") i class statementet:

```
proc logistic data=gvhhd;
class type(ref="all") preg / param=GLM;
model gvhhd(event="1") = type logindex preg donage;
run;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.9710	2.8662	1.9195	0.1659
type	aml	1	0.3968	1.2424	0.1020	0.7495
type	cml	1	2.0880	1.3424	2.4194	0.1198
type	all	0	0	.	.	.
logindex		1	1.8407	0.8875	4.3020	0.0381
preg	0	1	-1.7219	1.1453	2.2601	0.1327
preg	1	0	0	.	.	.
donage		1	0.1135	0.0761	2.2277	0.1356

Bemærk, at tests og parameterestimer (bortset fra de, der vedrører diagnosen) er uændrede. Det er nemlig den samme model, blot parametriseret på en anden måde. Koefficienten til `type cml` svarer nu til log-odds-ratio mellem patienter med CML og patienter med ALL, og den er derfor præcis den modsatte af koefficienten til `type all` fra den forrige model. Tilsvarende er koefficienten til `type aml` log-odds-ratio mellem patienter med AML og ALL, og den er derfor differensen mellem koefficienterne til `type aml` og `type all` (= 0.3968) fra den forrige model.

Det er nu klart, at den mindst signifikante forskel er mellem AML og ALL patienter, dvs. at den eneste betydende prognostiske forskel på leukæmigrupperne mht. GvHD ser ud til at være, hvorvidt der er tale om en akut eller kronisk leukæmi.

Vi kunne også have opnået dette resultat ved simpelthen at tilføje en `estimate`-sætning til vores `logistic`-kode:

```
proc logistic data=gvhhd;
class type preg / param=GLM;
model gvhhd(event="1") = type logindex preg donage;
estimate "aml vs all" type -1 1 0 / exp;
run;
```

som ville give en ekstra linie i output

Estimate					
Label	Estimate	Standard Error	z Value	Pr > z	Exponentiated
aml vs. all	0.3968	1.2424	0.32	0.7495	1.4870

der (selvfølgelig) stemmer ganske med resultatet, som vi fik ovenfor.

- (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 SAS kan man f.eks. benytte variabelen

`cml = (type=3);`

for sammenligningen 'kronisk versus akut', idet der er den konvention, at værdien af et logisk udsagn er 1, hvis det er sandt, og 0, hvis det er falsk, så variabelen ovenfor bliver 1 for CML-patienter og 0 for de øvrige.

Tilsvarende kan man benytte

`all = (type=2);`

for sammenligningen 'lymfatisk versus myeloid'.

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

Vi indsætter de to nye variable `cml` og `aml` (defineret i forbindelse med indlæsningen på s. 2) i stedet for `type`, og starter forfra, men her vises kun yderst skrabt output fra de to første kørsler, idet det er de selvsamme modeller som dem, vi så på ovenfor:

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-2.6249	3.6979	0.5039	0.4778	
cml 0	1	-1.6152	1.3074	1.5261	0.2167	
cml 1	0	0	.	.	.	
all 0	1	0.4282	1.2539	0.1166	0.7327	
all 1	0	0	.	.	.	
logindex	1	1.8187	0.8892	4.1836	0.0408	
preg 0	1	-1.6624	1.1810	1.9812	0.1593	
preg 1	0	0	.	.	.	

donage	1	0.1072	0.0826	1.6847	0.1943
rcpage	1	0.0160	0.0820	0.0379	0.8457

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2798	3.2637	0.4879	0.4848
cml	0 1	-1.6913	1.2620	1.7959	0.1802
cml	1 0	0	.	.	.
all	0 1	0.3968	1.2424	0.1020	0.7495
all	1 0	0	.	.	.
logindex	1	1.8407	0.8875	4.3020	0.0381
preg	0 1	-1.7219	1.1453	2.2601	0.1327
preg	1 0	0	.	.	.
donage	1	0.1135	0.0761	2.2277	0.1356

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

```
proc logistic data=gvh;
class cml preg / param=GLM;
model gvhd(event="1") = cml logindex preg donage;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8073	2.8767	0.3947	0.5298
cml	0 1	-1.8702	1.1421	2.6813	0.1015
cml	1 0	0	.	.	.
logindex	1	1.8972	0.8672	4.7865	0.0287
preg	0 1	-1.7346	1.1485	2.2810	0.1310
preg	1 0	0	.	.	.
donage	1	0.1092	0.0735	2.2060	0.1375

Odds Ratio Estimates				
Effect		Point Estimate	95% Wald Confidence Limits	
cml	0 vs 1	0.154	0.016	1.445
logindex		6.667	1.218	36.482
preg	0 vs 1	0.176	0.019	1.676
donage		1.115	0.966	1.288

Vi bemærker, at udeladelsen af `all` fra modellen bevirker, at `cml` får en noget øget betydning, og en lavere P-værdi. Men husk, at sammenligningen nu er *en anden en 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%, `cml`, `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:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.9674	2.3687	2.8055	0.0939
logindex	1	2.1060	0.8227	6.5526	0.0105
preg 0	1	-1.2912	1.0302	1.5711	0.2100
preg 1	0	0	.	.	.
donage	1	0.1245	0.0658	3.5829	0.0584

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
logindex	8.216	1.638 41.205
preg 0 vs 1	0.275	0.037 2.071
donage	1.133	0.996 1.288

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.3388	2.2517	3.7128	0.0540
cml 0	1	-1.5142	1.0788	1.9703	0.1604
cml 1	0	0	.	.	.
logindex	1	2.0592	0.8281	6.1833	0.0129
donage	1	0.1488	0.0694	4.5972	0.0320

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
cml 0 vs 1	0.220	0.027 1.822
logindex	7.839	1.547 39.734

donage	1.160	1.013	1.329
--------	-------	-------	-------

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.1996	1.4466	2.3122	0.1284
cml	0	-2.2506	1.1061	4.1401	0.0419
cml	1	0	.	.	.
logindex	1	1.4877	0.7197	4.2725	0.0387
preg	0	-2.4955	1.1012	5.1354	0.0234
preg	1	0	.	.	.

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits
cml	0 vs 1	0.105	0.012 0.921
logindex		4.427	1.080 18.145
preg	0 vs 1	0.082	0.010 0.714

Det fremgår af ovenstående, at når en af variablene fjernes, så bliver mindst en af de andre signifikante. I den sidste model er såvel cml 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.

(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 **cml** og **all**, ville vi have fortsat eliminationen ved at fjerne **type** helt fra modellen, og så ville vi have fortsat således:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.7349	2.6882	3.1024	0.0782
logindex	1	2.0446	0.8237	6.1613	0.0131
preg	0	-1.2232	1.0514	1.3535	0.2447
preg	1	0	.	.	.
donage	1	0.0997	0.0750	1.7658	0.1839
rcpage	1	0.0559	0.0859	0.4234	0.5152

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.9674	2.3687	2.8055	0.0939
logindex	1	2.1060	0.8227	6.5526	0.0105
preg	0	-1.2912	1.0302	1.5711	0.2100
preg	1	0	.	.	.
donage	1	0.1245	0.0658	3.5829	0.0584

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4540	2.0816	6.8647	0.0088
logindex	1	2.1777	0.7899	7.6004	0.0058
donage	1	0.1459	0.0647	5.0950	0.0240

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
logindex	8.826	1.877	41.511
donage	1.157	1.019	1.313

Som det ses, ville vi altså ende med 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 vil vi nu 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 `logistic`, men umiddelbart vender to af dem “den gale vej”, idet de er mindre end 1. Når vi inverterer dem, finder vi konklusionerne

- Folk med kronisk leukæmi har en odds for GvHD, der er $\frac{1}{0.105} = 9.49$ gange højere end folk med akut leukæmi, CI=(1.09,82.97).
- Patienter, der modtager knoglemarv fra en donor, der har været gravid, har en odds for GvHD, der er $\frac{1}{0.082} = 12.13$ gange højere end de, der modtager den fra en donor, der ikke har været gravid, 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 definerede allerførst i indlæsningsdelen, og kaldte `log2index`.

En forskel på 1 enhed i `log2index` svarer til en fordobling af indexværdien, hvilket er en noget mere mundret fortolkning.

Hvis vi fitter en model med `log2index`, får vi:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq

Intercept		1	2.1996	1.4466	2.3122	0.1284
cml	0	1	-2.2506	1.1061	4.1401	0.0419
cml	1	0	0	.	.	.
log2index		1	1.0312	0.4989	4.2725	0.0387
preg	0	1	-2.4955	1.1012	5.1354	0.0234
preg	1	0	0	.	.	.

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
cml	0 vs 1	0.105	0.012	0.921
log2index		2.804	1.055	7.456
preg	0 vs 1	0.082	0.010	0.714

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.

6. *Stod der mon noget i protokollen om at undersøge nogle interessante interaktioner? Prøv f.eks. interaktionen mellem lindex 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 $\text{preg} \times \log_2\text{index}$, og samtidig tilføjer vi et par `estimate`-sætninger for at få estimater for hver af de to effekter af $\log_2\text{index}$:

```

proc logistic data=gvh;
class cml preg / param=GLM;
model gvhd(event="1") = cml log2index preg preg*log2index;
estimate "log2index for preg=0" log2index 1 preg*log2index 1 0 / exp;
estimate "log2index for preg=1" log2index 1 preg*log2index 0 1 / exp;
run;

```

og finder output (kraftigt beskåret):

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
cml	1	3.9072	0.0481
log2index	1	0.1139	0.7358
preg	1	4.7964	0.0285
log2index*preg	1	2.8374	0.0921

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, er der i ovenstående kode tilføjet nogle `estimate`-sætninger, som giver outputtet:

Label	Estimate	Standard Error	z Value	Pr > z	Exponentiated
log2index for preg=0	1.7238	0.8507	2.03	0.0427	5.6060
log2index for preg=1	-1.1542	1.4700	-0.79	0.4324	0.3153

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

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 konfidensinterval. 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 laver en 2×2 -tabel:

```
data a1;
input gerningsmand$ death$ antal;
datalines;
white yes 53
white no 430
black yes 15
black no 176
;
run;

proc freq data=a1;
    tables gerningsmand*death / nopercnt nocol chisq riskdiff relrisk;
weight antal;
run;
```

Herved får vi outputtet (beskåret):

```
The FREQ Procedure

Table of gerningsmand by death

gerningsmand      death
```

Frequency			
Row Pct	no	yes	Total
-----+-----+-----+			
black	176	15	191
	92.15	7.85	
-----+-----+-----+			
white	430	53	483
	89.03	10.97	
-----+-----+-----+			
Total	606	68	674

Statistics for Table of gerningsmand by death

Statistic	DF	Value	Prob

Chi-Square	1	1.4685	0.2256

Fisher's Exact Test

Two-sided Pr <= P		0.2578	

Statistics for Table of gerningsmand by death

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.0785	0.0195	0.0404	0.1167	0.0446	0.1262
Row 2	0.1097	0.0142	0.0819	0.1376	0.0833	0.1411
Total	0.1009	0.0116	0.0782	0.1236	0.0792	0.1261
Difference	-0.0312	0.0241	-0.0784	0.0161		

Difference is (Row 1 - Row 2)

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.4462	0.7941	2.6337
Cohort (Col1 Risk)	1.0350	0.9827	1.0902
Cohort (Col2 Risk)	0.7157	0.4137	1.2382

Sample Size = 674

Vi ser, at såvel χ^2 -testet som Fishers eksakte test 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 10.97%, medens sorte har 7.85%. Der er således en lil-

le **overhyppighed af dødsstraf blandt hvide gerningsmænd**, og denne kan kvantificeres på forskellig vis:

- Differens mellem hyppigheder:
3.12%, med 95% konfidensinterval på (-1.61%, 7.84%).
- Odds Ratio for dødsstraf blandt hvide i forhold til sorte:
1.45 (0.79, 2.63).
- Relativ risiko for dødsstraf blandt hvide i forhold til sorte:
 $\frac{1}{0.7157} \approx 1.40$, med $CI = (\frac{1}{1.2382}, \frac{1}{0.4137}) = (0.81, 2.42)$.

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 har faktisk (i et senere spørgsmål) fået nogle flere oplysninger:

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

Vi udnytter nu disse oplysninger til at undersøge, hvorvidt offerets race virkelig ser ud til at være en confounder.

Først indlæser vi det udbyggede datasæt, nu med 8 linier, fordi vi opdeler efter 3 binære variable (offerets race, gerningsmandens race og udfaldet af dommen). Derefter laver vi to tabeller

- En tabel, der undersøger, om der er sammenhæng mellem offerets race og gerningsmandens race (kovariaten i fokus)
Bemærk brugen af `norow` her i stedet for det sædvanlige `nopercent`
- En tabel, der undersøger, om der er sammenhæng mellem offerets race og selve dommen (outcome)

```
data a1;
input offer$ gerningsmand$ death$ antal;
datalines;
white white yes 53
white white no 414
white black yes 11
white black no 37
black white yes 0
black white no 16
black black yes 4
black black no 139
;
run;

proc freq data=a1;
    tables offer*gerningsmand / norow nocol chisq riskdiff relrisk;
weight antal;
run;

proc freq data=a1;
    tables gerningsmand*death / nopercent nocol chisq riskdiff relrisk;
weight antal;
run;
```

Herved får vi to tabel-analyser med tilhørende test for uafhængighed:

The FREQ Procedure

Table of offer by gerningsmand

offer	gerningsmand		
	black	white	Total
black	143	16	159

	21.22 2.37 23.59
white	48 467 515
	7.12 69.29 76.41
Total	191 483 674
	28.34 71.66 100.00

Statistics for Table of offer by gerningsmand

Statistic	DF	Value	Prob
Chi-Square	1	388.8069	<.0001

Fisher's Exact Test

Two-sided Pr <= P	<.0001
-------------------	--------

The FREQ Procedure

Table of offer by death

offer	death
Frequency	
Row Pct no yes Total	
black 155 4 159	
97.48 2.52	
white 451 64 515	
87.57 12.43	
Total 606 68 674	

Statistics for Table of offer by death

Statistic	DF	Value	Prob
Chi-Square	1	13.1570	0.0003

Fisher's Exact Test

Two-sided Pr <= P	0.0001
-------------------	--------

Baseret på ovenstående tabel-analyser er vi 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 (faktisk $21.22\% + 69.29\% = 91.51\%$, som det fremgår af procenterne i tabellen)

er overensstemmelse mellem offers og gerningsmands race.

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:

```
proc genmod descending data = a1;
weight antal;
class offer gerningsmand;
model death = offer gerningsmand /
    dist=binomial link=logit;
estimate "offer b vs w" offer 1 -1 / exp;
estimate "offer w vs b" offer -1 1 / exp;
estimate "gerningsmand b vs w" gerningsmand 1 -1 / exp;
run;
```

The GENMOD Procedure

Model Information

Data Set	WORK.A1
Distribution	Binomial
Link Function	Logit
Dependent Variable	death
Scale Weight Variable	antal
Number of Observations Read	8
Number of Observations Used	7
Sum of Weights	674
Number of Events	3
Number of Trials	7
Missing Values	1

Class Level Information

Class	Levels	Values
offer	2	black white

gerningsmand 2 black white

Response Profile

Ordered Value	death	Total Frequency	Total Weight
1	yes	3	68
2	no	4	606

PROC GENMOD is modeling the probability that death='yes'.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square
Intercept	1	-2.0595	0.1458	-2.3453 -1.7736	199.40
offer black	1	-2.4044	0.6006	-3.5816 -1.2273	16.03
offer white	0	0.0000	0.0000	0.0000 0.0000	.
gerningsmand black	1	0.8678	0.3671	0.1483 1.5872	5.59
gerningsmand white	0	0.0000	0.0000	0.0000 0.0000	.
Scale	0	1.0000	0.0000	1.0000 1.0000	

Parameter	Pr > ChiSq
Intercept	<.0001
offer black	<.0001
offer white	.
gerningsmand black	0.0181
gerningsmand white	.
Scale	

NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits	L'Beta Estimate	Standard Error
offer b vs w	0.0828	0.0271 0.2267	-2.4044	0.6006
Exp(offer b vs w)			0.0903	0.0542
offer w vs b	0.9172	0.7733 0.9729	2.4044	0.6006
Exp(offer w vs b)			11.0723	6.6502
gerningsmand b vs w	0.7043	0.5370 0.8302	0.8678	0.3671
Exp(gerningsmand b vs w)			2.3817	0.8742

Label	Alpha	L'Beta Confidence Limits	Chi-Square	Pr > ChiSq
offer b vs w	0.05	-3.5816 -1.2273	16.03	<.0001
Exp(offer b vs w)	0.05	0.0278 0.2931		
offer w vs b	0.05	1.2273 3.5816	16.03	<.0001
Exp(offer w vs b)	0.05	3.4119 35.9320		
gerningsmand b vs w	0.05	0.1483 1.5872	5.59	0.0181
Exp(gerningsmand b vs w)	0.05	1.1599 4.8903		

Af den nederste **estimate**-sætning ser vi, 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!

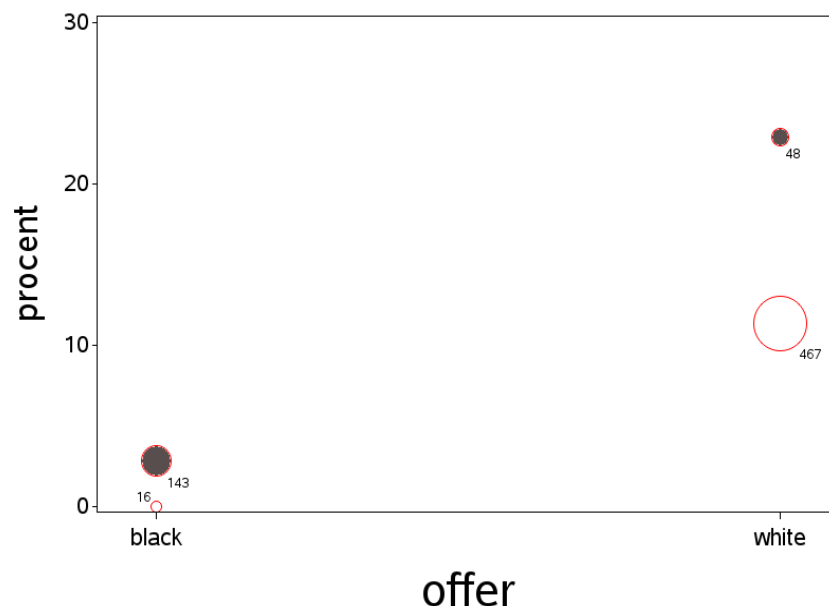
Endvidere ser vi (af den midterste **estimate**-sætning), at OR for dødsstraf er mere end 10 gange højere, hvis offeret er hvidt i forhold til, hvis offeret er sort.

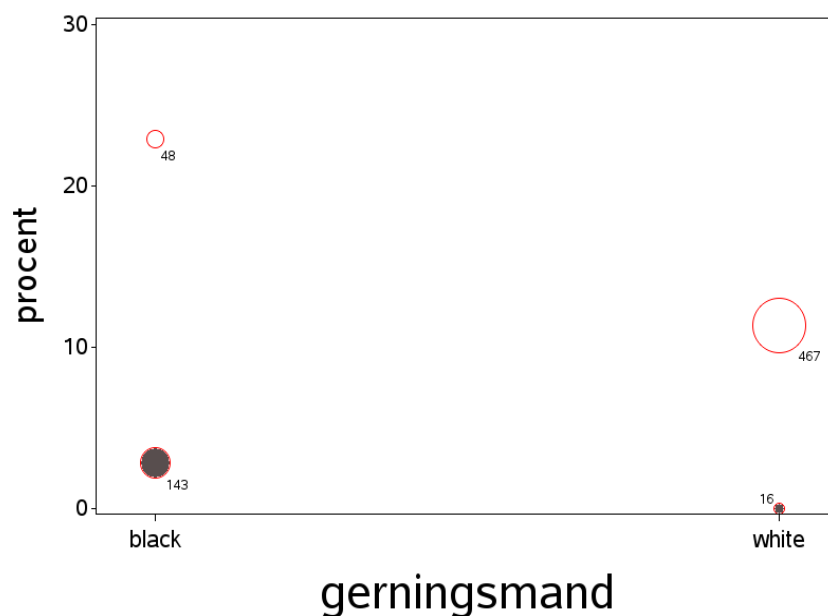
Sammenlign til resultatet i spørgsmål 1.

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 et par figurer, hvor vi afbilder procenten af dødsstraf, som funktion af hver af de to kovariater, med symboler svarende til den anden kovariat, og hvor *størrelsen* af symbolerne (circklerne) svarer til antallet af mordsager i den pågældende kategori:





På den øverste af disse tegninger ses, at hvide ofre oftere giver anledning til dødsdom, samt at sorte gerningsmænd oftere bliver dødsdømt, uanset offerets race (de sorte cirkler ligger højest).

Det samme ses i princippet på den anden figur, så det er nok mest et spørgsmål om smag og behag, hvilken man bedst kan lide.

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 figurerne ovenfor og kommentarerne til dem, 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 rolle, for at vurdere effekten af gerningsmandens race i hver af de to situationer.

```
proc sort data=a1;
```

```

        by offer;
run;

proc freq data=a1; by offer;
    tables gerningsmand*death / nopercnt nocol chisq riskdiff relrisk;
weight antal;
run;

```

offer=black

The FREQ Procedure

Table of gerningsmand by death

gerningsmand		death		Total
Frequency				
Row Pct	no	yes		
-----+-----+-----+				
black	139	4		143
	97.20	2.80		
-----+-----+-----+				
white	16	0		16
	100.00	0.00		
-----+-----+-----+				
Total	155	4		159

Statistics for Table of gerningsmand by death

Statistic	DF	Value	Prob

Chi-Square	1	0.4591	0.4980

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Two-sided Pr <= P 1.0000

Statistics for Table of gerningsmand by death

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.0280	0.0138	0.0009	0.0550	0.0077	0.0701
Row 2	0.0000	0.0000	0.0000	0.0000	0.0000	0.2059
Total	0.0252	0.0124	0.0008	0.0495	0.0069	0.0632
Difference	0.0280	0.0138	0.0009	0.0550		

Difference is (Row 1 - Row 2)

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Cohort (Col1 Risk)	0.9720	0.9454	0.9994

One or more risk estimates not computed --- zero cell.

Sample Size = 159

offer=white

The FREQ Procedure

Table of gerningsmand by death

gerningsmand		death	
Frequency			
Row Pct	no	yes	Total
black	37	11	48
	77.08	22.92	
white	414	53	467
	88.65	11.35	
Total	451	64	515

Statistics for Table of gerningsmand by death

Statistic	DF	Value	Prob
Chi-Square	1	5.3518	0.0207
Likelihood Ratio Chi-Square	1	4.5341	0.0332
Continuity Adj. Chi-Square	1	4.3416	0.0372

Fisher's Exact Test

Two-sided Pr <= P 0.0350

Statistics for Table of gerningsmand by death

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.2292	0.0607	0.1103	0.3481	0.1203	0.3731
Row 2	0.1135	0.0147	0.0847	0.1423	0.0862	0.1458
Total	0.1243	0.0145	0.0958	0.1528	0.0970	0.1559
Difference	0.1157	0.0624	-0.0067	0.2380		

Difference is (Row 1 - Row 2)

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.4306	0.2072	0.8947
Cohort (Col1 Risk)	0.8695	0.7427	1.0180
Cohort (Col2 Risk)	2.0193	1.1335	3.5973

Sample Size = 515

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 får en dødsdom. Relativ risiko estimeres til 2.0 (1.1, 3.6), og odds ratio estimeres (ved at invertere) til 2.32 (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`. Samtidig vises det, hvordan man kan tilføje `estimate`-sætninger til estimation af Odds Ratio svarende til opdelte tabeller:

```
proc genmod descending plots=all data = a1;
weight antal;
class offer gerningsmand;
model death = offer gerningsmand offer*gerningsmand /
    dist=binomial link=logit type3;
estimate "offer w: gerningsmand b vs w"
    gerningsmand 1 -1 offer*gerningsmand 0 0 1 -1 / exp;
```

```

estimate "offer b: gerningsmand b vs w"
      gerningsmand 1 -1 offer*gerningsmand 1 -1 0 0 / exp;
run;

```

Class Level Information

Class	Levels	Values
offer	2	black white
gerningsmand	2	black white

Response Profile

Ordered Value	death	Total Frequency	Total Weight
1	yes	3	68
2	no	4	606

PROC GENMOD is modeling the probability that death='yes'.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error
Intercept		1	-2.0556	0.1459
offer	black	1	-22.3097	0.6125
offer	white	0	0.0000	0.0000
gerningsmand	black	1	0.8426	0.3731
gerningsmand	white	0	0.0000	0.0000
offer*gerningsmand	black black	0	19.9746	0.0000
offer*gerningsmand	black white	0	0.0000	0.0000
offer*gerningsmand	white black	0	0.0000	0.0000

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		Wald 95% Confidence Limits	Wald Chi-Square
Intercept		-2.3415 -1.7696	198.53 <.0001
offer	black	-23.5102 -21.1093	1326.81 <.0001
offer	white	0.0000 0.0000	. .
gerningsmand	black	0.1112 1.5739	5.10 0.0239
gerningsmand	white	0.0000 0.0000	. .
offer*gerningsmand	black black	19.9746 19.9746	. 0.5377
offer*gerningsmand	black white	0.0000 0.0000	. .
offer*gerningsmand	white black	0.0000 0.0000	. .
offer*gerningsmand	white white	0.0000 0.0000	. .
Scale		1.0000 0.0000	. .

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
offer	1	14.59	0.0001
gerningsmand	1	1.76	0.1843
offer*gerningsmand	1	0.38	0.5377

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits	L'Beta Estimate
offer w: gerningsmand b vs w	0.6990	0.5278 0.8283	0.8426
Exp(offer w: gerningsmand b vs w)			2.3223
offer b: gerningsmand b vs w	1.0000	1.0000 1.0000	20.8171
Exp(offer b: gerningsmand b vs w)			1.0984E9

Label	Standard Error	Alpha	L'Beta Confidence Limits
offer w: gerningsmand b vs w	0.3731	0.05	0.1112 1.5739
Exp(offer w: gerningsmand b vs w)	0.8665	0.05	1.1177 4.8252
offer b: gerningsmand b vs w	0.3731	0.05	20.0858 21.5484
Exp(offer b: gerningsmand b vs w)	4.0985E8	0.05	5.2865E8 2.2823E9

Label	Chi- Square	Pr > ChiSq
offer w: gerningsmand b vs w	5.10	0.0239
Exp(offer w: gerningsmand b vs w)		
offer b: gerningsmand b vs w	3112.7	<.0001
Exp(offer b: gerningsmand b vs w)		

Vi ser, at der ganske rigtigt *ikke* kan påvises nogen interaktion mellem gerningsmands og offers racer ($P = 0.54$).

For hvide ofre fås (ligesom for de opdelte analyser ovenfor) en odds ratio for dødsdom for sorte vs. hvide på 2.32 (1.12, 4.83).

For sorte ofre angives den tilsvarende odds ratio til 1.098×10^9 , hvilket er SAS's måde at indikere, at der nok er divideret med 0.

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 sort end hvide en dødsdom.