UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 301 — Statistiske metoder og anvendelser.
Eksamensdag:	Fredag, 2. juni, 1995.
Tid for eksamen:	09.00 – 14.00.
Oppgavesettet er på 9 sider.	
Vedlegg:	Tabeller over normal, t-, χ^2- og F-fordelingen.
Tillatte hjelpemidler:	Alle trykte og skrevne samt regneutstyr.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

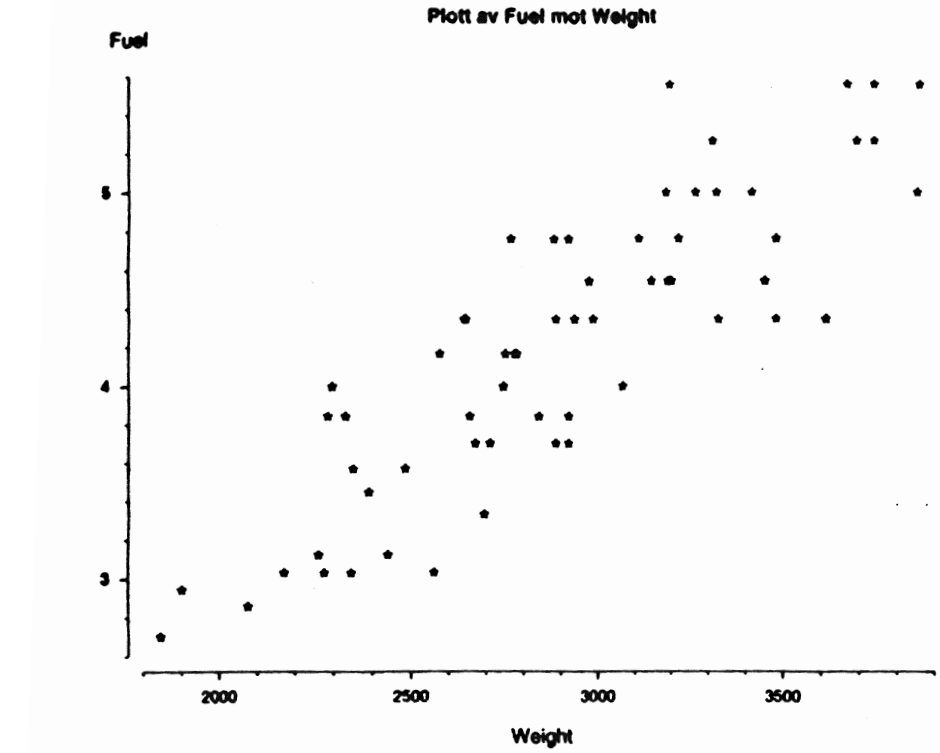
Vi vil i denne oppgaven studere sammenhengen mellom bensinforbruk (*Fuel*) for ulike biler og bilenes vekt (*Weight*), motorens slagvolum (*Disp*) og biltype (*Type*).

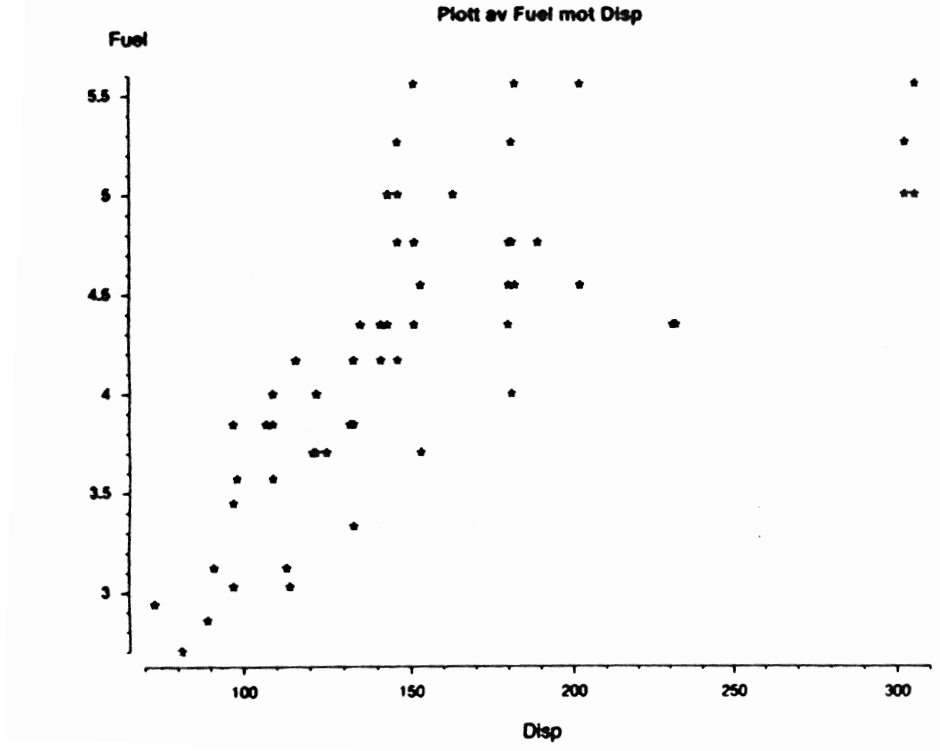
Fuel, *Weight* og *Disp* er kontinuerlige variable, mens *Type* antar 5 verdier, "Small", "Sporty", "Compact", "Medium", "Large" og "Van".

a) Nedenfor er gitt plott av *Fuel* mot *Weight*, *Fuel* mot *Disp* og en tabell over gjennomsnittsverdier for *Fuel*, *Weight* og *Disp* for de ulike biltyper. En tabell over korrelasjoner mellom de kontinuerlige variablene er også gitt.

   Kommenter de ulike plott og tabeller og diskuter dem i forhold til regresjonsanalyse av dataene.

(Fortsettes side 2.)





(Fortsettes side 3.)

Type	Weight	Disp	Fuel
Small	2257.692	97.308	3.273
Sporty	2798.889	164.111	3.958
Compact	2821.000	140.400	4.168
Medium	3195.769	175.846	4.601
Large	3676.667	279.333	4.968
Van	3517.143	164.429	5.313

Gjennomsnitt av *Weight, Disp* og *Fuel* for hver biltype.

b) En lineær regresjonsanalyse med *Weight*, *Disp* og *Type* som forklaringsvariable og *Fuel* som respons gir følgende resultat:

```
Dependent variable:       Fuel
Independent variables:    Weight, Disp., Type[1 2 3 4 5]
Observations:  60         Parameters:  8

Parameter     Estimate       SE          t-Ratio     P-Value
intercept     3.9181       0.73029       5.3652      0.0000
 Weight       4.164e-05    0.0002585     0.1611      0.8726
 Disp.        0.0075941    0.0017411     4.3616      0.0001
 Type[1]     -1.4777       0.29052      -5.0865      0.0000
 Type[2]     -1.3234       0.24361      -5.4324      0.0000
 Type[3]     -0.93417      0.20861      -4.4780      0.0000
 Type[4]     -0.78519      0.17727      -4.4294      0.0000
 Type[5]     -1.2247       0.27626      -4.4333      0.0000

Residual SD   0.31392     Residual Variance   0.098549
Multiple R    0.92122     Multiple R-squared  0.84864
```

Hvorfor er det her blitt 5 parametre for *Type*?

Diskuter de to siste kolonnene i tabellen.

c) Vi vil se på en litt utvidet modell der også $(Disp)^2$ er tatt med som forklaringsvariabel. Det ga følgende resultat:

```
Dependent variable:       Fuel
Independent variables:    Weight,Disp,Disp^2,Type[1 2 3 4 5]
Observations:  60         Parameters:  9

Parameter     Estimate       SE          t-Ratio     P-Value
intercept     3.8872       0.88235       4.4056      0.0001
 Weight       3.611e-05    0.0002750     0.1313      0.8960
 Disp         0.0081002    0.0081252     0.9969      0.3235
 Disp^2      -1.195e-06    1.873e-05    -0.0638      0.9494
 Type[1]     -1.4721       0.30628      -4.8065      0.0000
 Type[2]     -1.3209       0.24904      -5.3038      0.0000
 Type[3]     -0.93474      0.21083      -4.4336      0.0000
 Type[4]     -0.78802      0.18438      -4.2738      0.0001
 Type[5]     -1.2202       0.28765      -4.2422      0.0001

Residual SD   0.31697     Residual Variance   0.10047
Multiple R    0.92122     Multiple R-squared  0.84865
```

Hvorfor får vi en større R^2-verdi i dette tilfellet?

Diskuter tolkningen av R^2.

Ser det ut som om det er fornuftig å ta med $(Disp)^2$?

(Fortsettes side 4.)

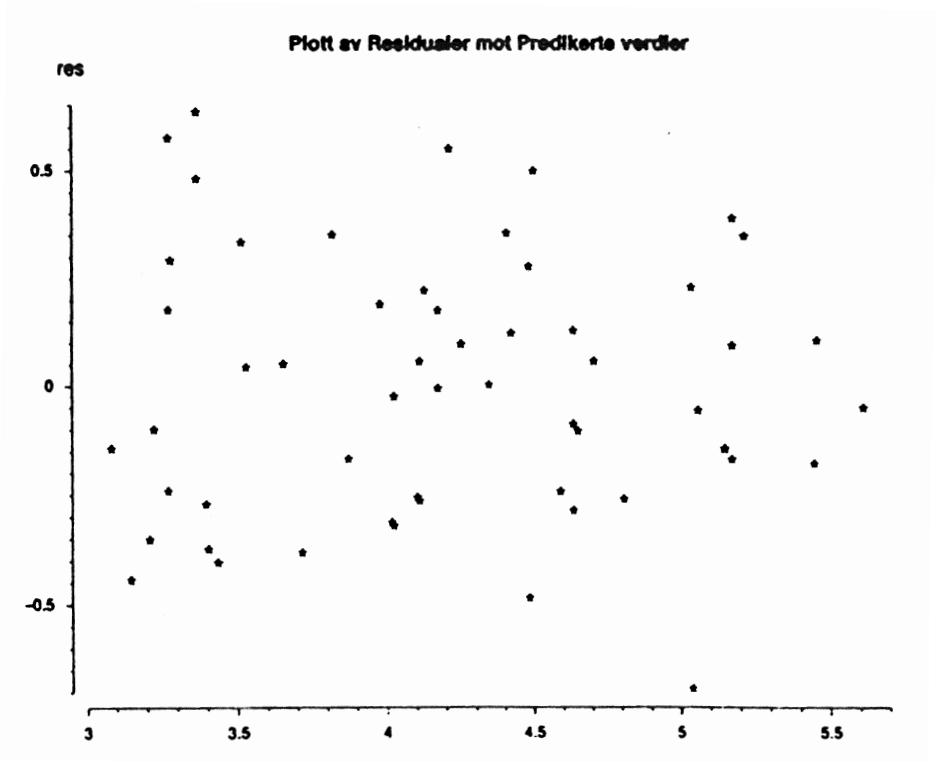
d) I tabellen nedenfor er et kryss-validert estimat for R^2 gitt for ulike modeller:

Forklaringsvariable								R^2_{cross}
W	*D*	D^2	T_1	T_2	T_3	T_4	T_5	0.7873
W	*D*		T_1	T_2	T_3	T_4	T_5	0.7940
	D		T_1	T_2	T_3	T_4	T_5	0.8097
			T_1	T_2	T_3	T_4	T_5	0.6481

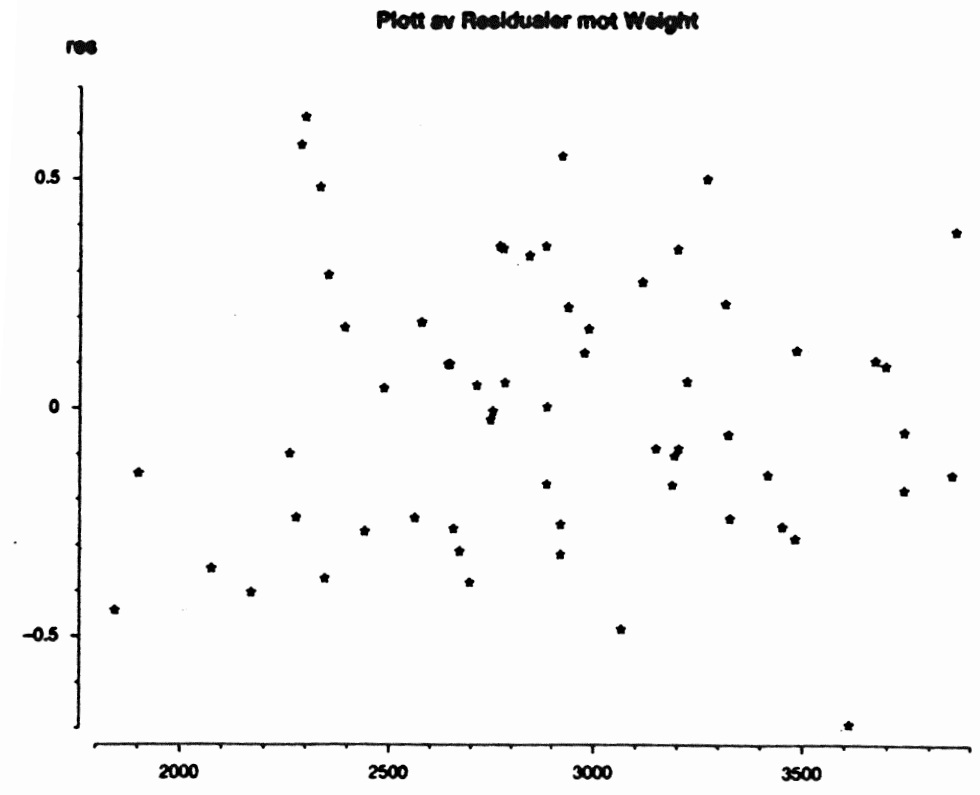
Kryss-validert estimat for R^2 for ulike modeller.
Her står W for *Weight*, D for *Disp* og T_j for *Type* [j].

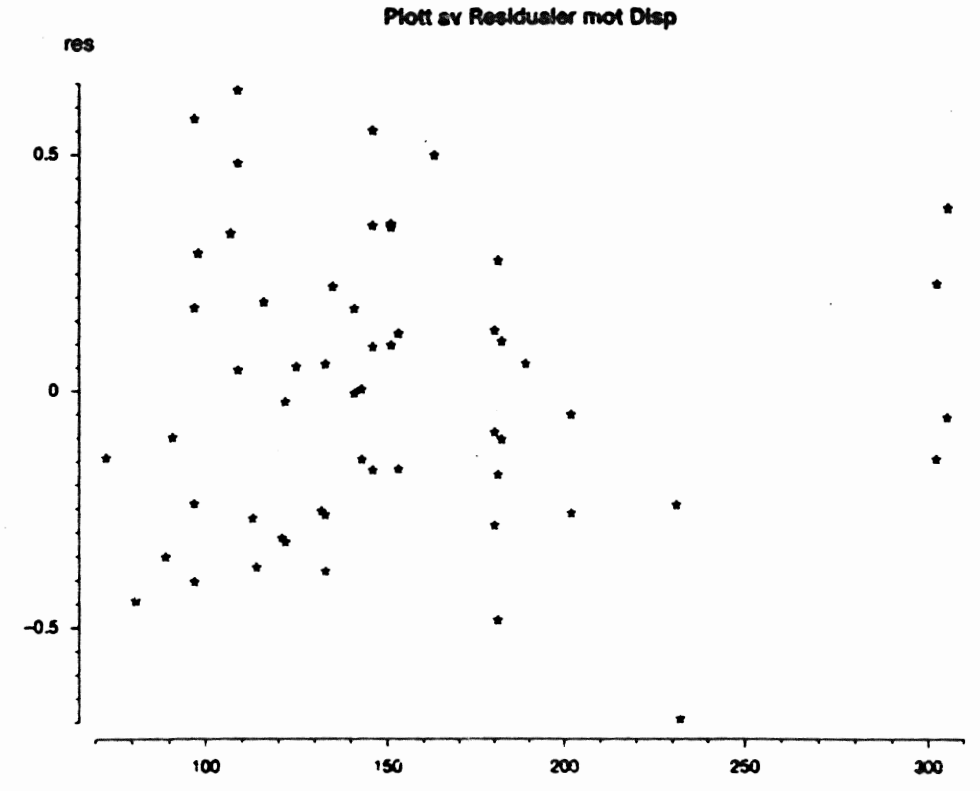
Forklar hva som menes med et kryss-validert estimat for R^2 og hvorfor dette er et mer fornuftig mål enn den vanlige R^2.
Hvilken modell ville du valgt?

e) Plottene nedenfor viser ulike residualplott.
Forklar hva hver type plott kan brukes til og gi en vurdering av plottene i forhold til modell.

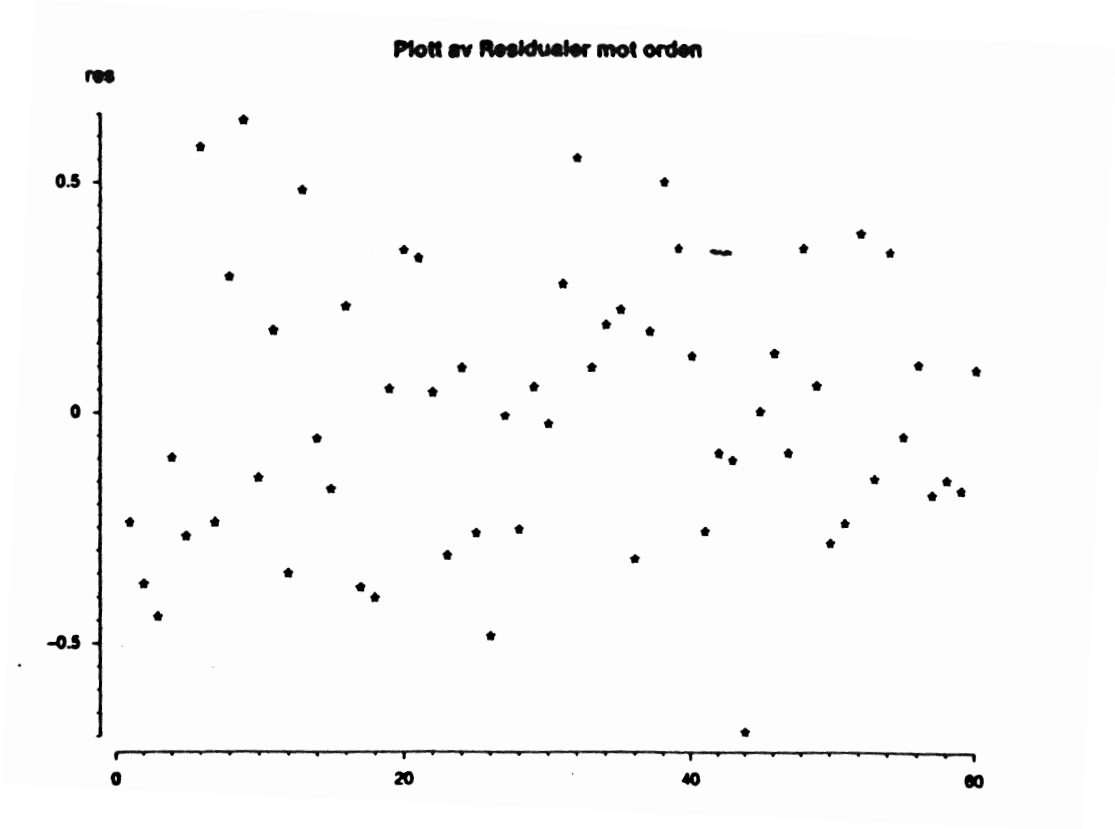


(Fortsettes side 5.)





(Fortsettes side 6.)



f) Anta vi ønsker å predikere bensinforbruket for en bil med vekt 2745 kg, motorslagvolum 125 og av type "Compact".
Hva blir punktestimatet?
Hvilke problemer får du her når et usikkerhetsanslag skal bestemmes?

Oppgave 2.

Dataene i tabellen nedenfor angir antall melanoma (kreft) tilfeller i perioden 1969–1971 blandt hvite menn for ulike aldersgrupper og to ulike regioner.

	Melanoma tilfeller			
Aldersgruppe	Nord		Syd	
< 35	61	(2880262)	64	(1074246)
$35 - 44$	76	(564535)	75	(220407)
$45 - 54$	98	(592983)	68	(198119)
$55 - 64$	104	(450740)	63	(134084)
$65 - 74$	63	(270908)	45	(70708)
≥ 75	80	(161850)	27	(34233)

Antall nye Melanoma tilfeller blandt hvite menn fra 1969 til 1971 fordelt over aldersgrupper.
Totalt antall personer i hver aldersgruppe er gitt i parantes.

En er interessert i å undersøke forskjeller mellom regioner.

(Fortsettes side 7.)

a) Forklar hvorfor Poisson regresjon er rimelig å benytte for analyse av slike data.

b) Nedenfor er gitt en utskrift av 3 ulike modeller:

Modell 1:
Region og *alder* som kontinuerlige forklaringsvariable, dvs.

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

der $x_1 = \begin{cases} 1 & \text{hvis nordlig region} \\ 0 & \text{ellers} \end{cases}$

og $x_2 =$ alder.

(Vi setter her alderen til å være 30, 40, 50, 60, 70 og 80 for de ulike aldersgruppene.)

Parameter	Estimate	SE	W	P-Value
beta0	−10.630	0.081830	−129.9038	0.0000
beta1	0.84021	0.071030	11.8291	0.0000
beta2	0.51646	0.019345	26.6971	0.0000

Log-likelihood = 374.433

Model 2:
Region, *alder* og $(alder)^2$, dvs.

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

der x_1 og x_2 er som ovenfor.

Parameter	Estimate	SE	W	P-Value
beta0	−11.549	0.14796	−78.0557	0.0000
beta1	0.82616	0.070971	11.6408	0.0000
beta2	1.2708	0.096719	13.1394	0.0000
beta3	−0.11340	0.014269	−7.9472	0.0000

Log-likelihood = 407.5835

Modell 3:
Region som kontinuerlig variabel og *alder* som kategorisk variabel, dvs.

$$\eta = \beta_0 + \beta_1 x_1 + \beta_4 z_1 + \beta_5 z_2 + \beta_6 z_3 + \beta_7 z_4 + \beta_5 z_5$$

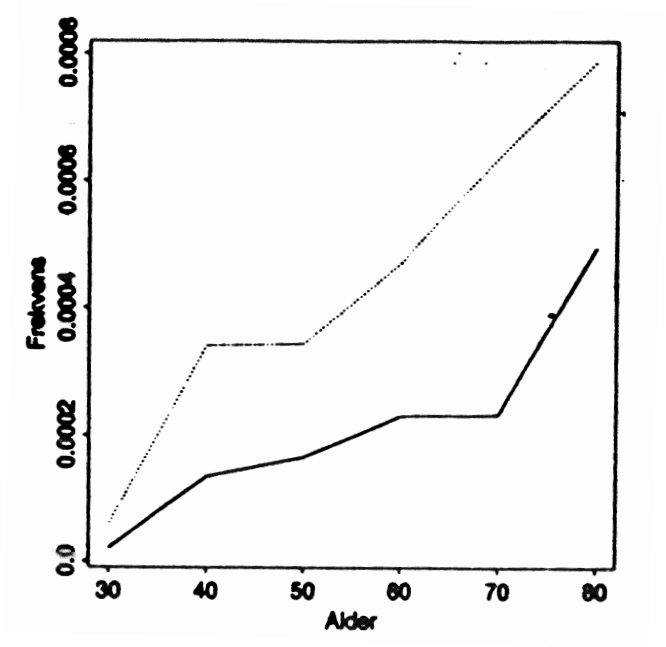
der $z_i = \begin{cases} 1 & \text{hvis data tilhører aldersgruppe } i \\ 0 & \text{ellers} \end{cases}$.

(Fortsettes side 8.)

Parameter	Estimate	SE	W	P-Value
beta0	−7.7136	0.099380	−77.6173	0.0000
beta1	0.81948	0.071028	11.5375	0.0000
beta4	−2.94447	0.13205	−22.3002	0.0000
beta5	−1.1473	0.12679	−9.0490	0.0000
beta6	−1.0316	0.12421	−8.3054	0.0000
beta7	−0.70288	0.12396	−5.6704	0.0000
beta8	−0.57896	0.13644	−4.2432	0.0000

Log-likelihood = 407.5835

Nedenfor er også gitt et plott av alder (x_2) mot frekvens for de ulike regionene (heltrukne for den nordlige region, stiplet for den sydlige). Diskuter de ulike modeller, fordeler, ulemper.
Hvilken modell vil du foretrekke?



c) En ønsker å teste om det er forskjell mellom de to regionene. Nedenfor er gitt utskrift av Poisson regresjon for de 3 modellene i b), men der *region* ikke er tatt med som forklaringsvariabel.

Modell 4: $\eta = \beta_0 + \beta_2 x_2$

Parameter	Estimate	SE	W	P-value
beta0	−10.295	0.073419	−140.2251	0.0000
beta2	0.49713	0.019190	25.9048	0.0000

Log-likelihood = 309.306

Modell 5: $\eta = \beta_0 + \beta_2 x_2 + \beta_3 x_2^2$

Parameter	Estimate	SE	W	P-value
beta0	−11.246	0.14408	−78.0496	0.0000
beta2	1.2728	0.096627	13.1724	0.0000
beta3	−0.11656	0.014257	−8.1757	0.0000

Log-likelihood = 344.419

(Fortsettes side 9.)

Modell 6: $\eta = \beta_1 + \beta_4 z_1 + \beta_5 z_2 + \beta_6 z_3 + \beta_7 z_4 + \beta_8 z_5$

Parameter	Estimate	SE	W	P-Value
beta0	−7.75135	0.096647	−77.7199	0.0000
beta4	−2.8486	0.13170	−21.6288	0.0000
beta5	−1.0426	0.12637	−8.2508	0.0000
beta6	−0.95573	0.12398	−7.7090	0.0000
beta7	−0.64761	0.12383	−5.2298	0.0000
beta8	−0.54585	0.13640	−4.0018	0.0001

Log-likelihood = 382.692

Sett opp to ulike tester for hypotesen om at det ikke er ulikheter mellom regionene. Prøv ut med alle de 3 variasjoner av modellen.
Diskuter hvorfor resultatene blir så like.

Oppgave 3.

Dataene i denne oppgaven skriver seg fra en studie av virkningen av 6-mercaptopurin på varigheten av steroid-indusert remisjon i akutt leukemi. Observasjonstider i uker til tilbakefall er angitt i tabellen under. Observasjonstidene er gitt for to grupper, en behandlingsruppe og en kontrollgruppe. Observasjonene markert med $*$ har ikke fått tilbakefall ved tidspunktet innsamling av data avsluttes.

Behandlings-gruppe (1)	**6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*.**
Kontrol-gruppe (2)	**1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.**

Tabel 3.3. Remissionstider i uger fra en undersøgelse af 6-MP. Censurerede observationer er markeret med *.

En ønsker å undersøke om dataene gir grunnlag for å påstå at remisjonstidene er høyere i behandlingsgruppen enn i kontrollgruppen.
Diskuter metoder for å analysere slike data og hvordan en test på at det ikke er forskjell mellom gruppene kan utføres.
Du behøver ikke utføre de konkrete beregninger, men bør ha med hvilke beregninger som må utføres.

SLUTT