UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 301 — Statistiske metoder og anvendelser.
Eksamensdag:	Fredag 2. juni 2000.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 7 sider.	
Vedlegg:	Ingen.
Tillatte hjelpemidler:	Alle trykte og skrevne, kalkulator.

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

Oppgave 1.

Denne oppgaven omhandler boligmarkedet i Oslo. Fra Aftenpostens boligannonser er det registrert for $n = 100$ leiligheter

1) Prisantydning/Verditakst (i 1000 kr)
2) Leilighetens areal (Boa) i kvadratmeter
3) Antall rom i leiligheten
4) Månedlig husleie (i kroner)
5) Om det er balkong tilknyttet leiligheten (Ja = 1, Nei= 0)
6) Om det er garasje tilknyttet leiligheten (Ja = 1, Nei = 0)
7) x-koordinat (i km, retning øst)
8) y-koordinat (i km, retning nord)

x- og y-koordinatene angir altså hvor i Oslo leilighetene ligger. Koordinatsystemets origo er plassert på Bygdøy, mens det nordøstlige hjørnet for leilighetene som er registrert ligger på Grefsen.

Dataene skal analyseres med lineær regresjon der responsvariabel er prisantydning/verditakst (unøyaktig referert til som pris).

a) Som en innledende analyse skal du se på regresjonene mot antall kvadratmeter og antall rom og mot begge kovariater.
Under finner du utskrift av regresjonsanalysene samt av deskriptiv

(Fortsettes side 2.)

statistikk og korrelasjoner for variablene som inngår.
Gi en beskrivelse og intuitiv forklaring på fenomenet du observerer.

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Pris	100	1207.4	1055.0	1154.8	525.3	52.5
Kvm	100	60.89	58.00	59.52	24.85	2.49
Rom	100	2.3500	2.0000	2.3111	0.9143	0.0914

Regression Analysis: Pris versus Kvm

The regression equation is
Pris = 135 + 17.6 Kvm

Predictor	Coef	SE Coef	T	P
Constant	134.54	77.53	1.74	0.086
Kvm	17.620	1.180	14.94	0.000

S = 291.7 R-Sq = 69.5% R-Sq(adj) = 69.2%

Regression Analysis: Pris versus Rom

The regression equation is
Pris = 254 + 406 Rom

Predictor	Coef	SE Coef	T	P
Constant	253.6	103.5	2.45	0.016
Rom	405.87	41.08	9.88	0.000

S = 373.7 R-Sq = 49.9% R-Sq(adj) = 49.4%

Regression Analysis: Pris versus Kvm; Rom

The regression equation is
Pris = 133 + 17.5 Kvm + 3.2 Rom

Predictor	Coef	SE Coef	T	P
Constant	133.07	82.64	1.61	0.111
Kvm	17.519	2.221	7.89	0.000
Rom	3.23	60.37	0.05	0.957

S = 293.2 R-Sq = 69.5% R-Sq(adj) = 68.9%

Correlations: Kvm; Rom

Pearson correlation of Kvm and Rom = 0.846

b) Resultater fra en regresjonsanalyse med alle kovariater inkludert er oppgitt under.

Regression Analysis: Pris versus Kvm; Rom; Garasje; Balkong; Leie; x; y

The regression equation is
Pris = 543 + 20.2 Kvm + 8.0 Rom + 89.6 Garasje + 116 Balkong - 0.150 Leie
- 106 x - 7.2 y

Predictor	Coef	SE Coef	T	P
Constant	542.59	80.62	6.73	0.000
Kvm	20.189	1.376	14.67	0.000
Rom	7.99	38.05	0.21	0.834
Garasje	89.56	74.39	1.20	0.232
Balkong	115.83	40.70	2.85	0.005
Leie	-0.15042	0.01920	-7.84	0.000
x	-106.32	13.58	-7.83	0.000
y	-7.25	17.18	-0.42	0.674

S = 175.2 R-Sq = 89.7% R-Sq(adj) = 88.9%

Forklar hva hver enkelt av de estimerte regresjonsparameterne forteller om hvordan prisantydning avhenger av kovariatene. I dette punktet skal du ikke se på signifikans av parametrene.

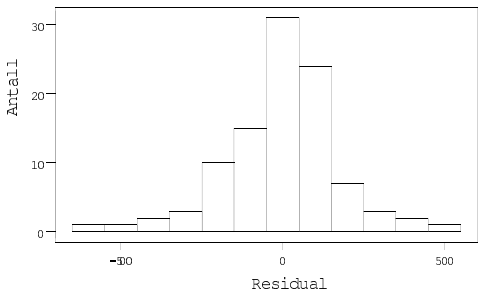
(Fortsettes side 3.)

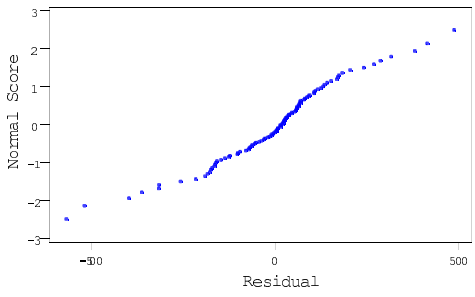
c) Angi hvilke kovariater som har en signifikant effekt. Diskuter kort fordeler og ulemper ved å fjerne insignifikante kovariater fra modellen. Sammenlign med analysen under der noen kovariater er fjernet fra modellen.

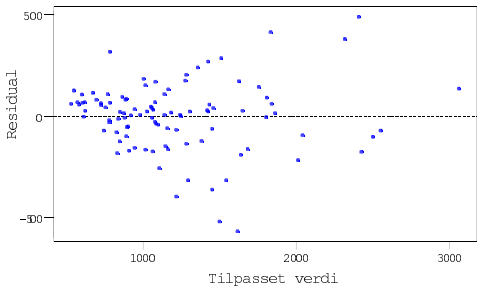
Regression Analysis: Pris versus Kvm; Balkong; Leie; x

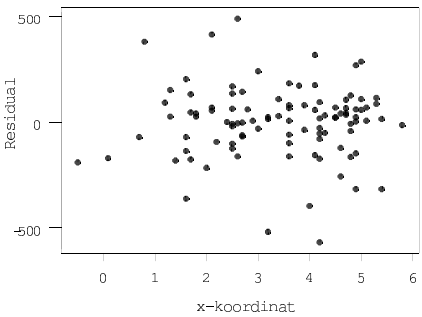
```
The regression equation is
Pris = 528 + 20.3 Kvm + 131 Balkong - 0.139 Leie - 108 x

Predictor        Coef     SE Coef          T        P
Constant       527.99       67.22       7.85    0.000
Kvm           20.3425      0.7971      25.52    0.000
Balkong        130.58       38.25       3.41    0.001
Leie        -0.13941     0.01672      -8.34    0.000
x             -108.38       13.23      -8.19    0.000

S = 173.9       R-Sq = 89.5%     R-Sq(adj) = 89.0%
```

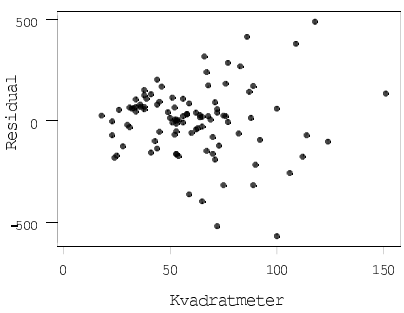
d) Gi et kort sammendrag av hvordan residualplott kan benyttes til å sjekke om modellantagelsene holder.
Anvend metodene på plottene under. Disse er laget utfra den reduserte modellen.

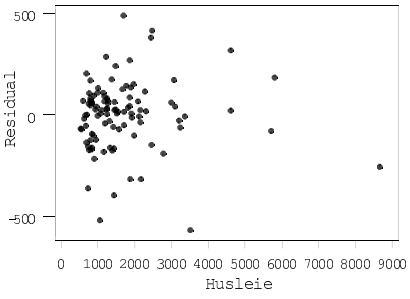












(Fortsettes side 4.)

e) For den fulle modellen finner man en kryssvalidert R^2 på 0.873. For den reduserte modellen ble kryssvalidert R^2 lik 0.882.
Sammenlign med R^2 fra utskriftene. Forklar hva som ligger i disse begrepene og diskuter forskjellen mellom dem. Hvordan kan man bruke kryssvalidert R^2 til å sammenligne full og redusert modell?

f) Sammenlign estimater og signifikans for variablene balkong og garasje fra den fulle modellen. Kommenter.

Anta at parameterestimatene er lik de faktiske verdiene. Man planlegger å samle inn mer data. Anta at kovariatfordeling og sammenheng med prisantydning forblir den samme.

Hvor stort må datamaterialet være for at studiens styrke mht nullhypotesen "Garasje har ingen innvirkning på prisantydning" skal bli minst 50%? Hvor stort må utvalget være for at styrken skal bli 80%? Benytt i begge tilfeller ensidig test med nivå 2.5%.

Oppgave 2.

I denne oppgaven skal du se nærmere på dataene om giftdose og dødelighet av biller fra kapittel 7 i læreboka.

a) Under er angitt resultatene av den logistiske regresjonsmodellen

$$p_1(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$

som ekvivalent kan skrives

$$\text{logit}(p_1(x)) = \log\left(\frac{p_1(x)}{1-p_1(x)}\right) = a + bx$$

(Dette skal du ikke vise.) Her er x giftdose og $p_1(x)$ sannsynligheten for at en bille skal dø ved denne dosen.

```
Link Function:  Logit

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -60.717      5.181   -11.72 0.000
Dose          34.270      2.912    11.77 0.000 7.65E+14 2.54E+12 2.30E+17

Log-Likelihood = -186.235
Test that all slopes are zero: G = 272.970; DF = 1; P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                  10.027     6  0.124
Deviance                 11.232     6  0.081
Hosmer-Lemeshow          10.027     6  0.124
```

Beskriv kort to tester for å undersøke om giftdosen påvirker dødeligheten. Hva blir konklusjonen?

(Fortsettes side 5.)

b) Et alternativ til en logistisk regresjonsmodell er en komplementær log-log (eller gombit) modell der sannsynligheten for å dø ved dose x er gitt ved

$$p_2(x) = 1 - \exp(-\exp(a + bx))$$

Ekvivalent kan dette skrives

$$\log(-\log(1 - p_2(x)) = a + bx$$

(Dette skal du ikke vise.)

Resultatene av en komplementær log-log modell er angitt under. Tester kan gjøres helt analogt med den logistiske modellen.
Sammenlign resultatene med hverandre og med de observerte andelene.
Hvilken modell synes du passer best og hvorfor?

```
Link Function:  Gompit

Logistic Regression Table

Predictor       Coef    SE Coef        Z     P
Constant     -39.572      3.240   -12.21 0.000
Dose          22.041      1.799    12.25 0.000

Log-Likelihood = -182.343
Test that all slopes are zero: G = 280.756; DF = 1; P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                   3.295     6  0.771
Deviance                  3.446     6  0.751
Hosmer-Lemeshow           3.295     6  0.771
```

c) Den logistiske regresjonsmodellen utvides med et kvadrat ledd slik at

$$\text{logit}(p_3(x)) = a + bx + cx^2$$

der $p_3(x)$ er sannsynligheten for død ved dose x under denne modellen.
Resultatene er angitt under.

```
Link Function:  Logit

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant       431.1       180.7     2.39 0.017
Dose          -520.6       204.5    -2.55 0.011     0.00     0.00     0.00
Dose2          156.41       57.86     2.70 0.007        * 4.73E+18        *

Log-Likelihood = -182.217
Test that all slopes are zero: G = 281.008; DF = 2; P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                   3.004     5  0.699
Deviance                  3.195     5  0.670
Hosmer-Lemeshow           3.004     6  0.808
```

Sammenlign med modellene i punkt a) og punkt b). Hvilken modell foretrekker du?
Angi en ulempe med kvadratleddsmodellen.

(Fortsettes side 6.)

d) Anta at $x = 0$ eller $x = 1$ samt at $p_1(1) = 0.02$ og $p_1(0) = 0.01$. Finn e^b og sammenlign med relativ risk $p_1(1)/p_1(0)$.
Gjør tilsvarende for $p_2(x)$. Kommenter.
Foreslå en generell tilnærming for e^b under den komplementære log-log modellen når $p_2(x)$ er små. Bevis kreves ikke.

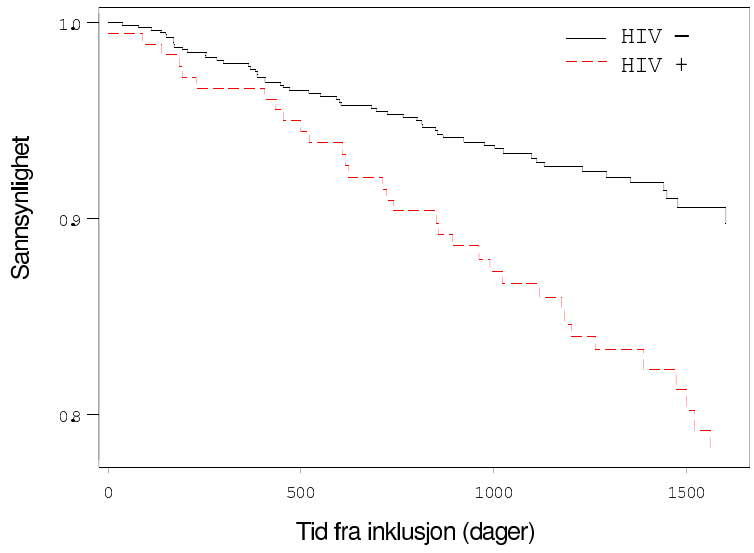
e) Diskuter gyldigheten av tilnærmelsen når $p_2(x)$ ikke er små. Sammenlign med tilsvarende tilnærmelse for $p_1(x)$. Det kan være nyttig å beregne e^b f.eks. når $p_j(0) = 0.4$ og $p_j(1) = 0.6$ for $j = 1, 2$.

Oppgave 3.

Sprøytenarkomane inkluderes i en levetids studie når de avlegger en HIV-test. For i alt 1004 personer er det registrert

1) Tid fra inklusjon i studien til død eller til avslutning av studien
2) Indikator for dødsfall før studiens avslutning
3) Alder ved inklusjon
4) Kjønn (Mann = 1, Kvinne = 0)
5) Om personen var HIV-positiv ved inklusjon (Ja = 1, Nei = 0)

a) Plottene under viser Kaplan-Meier estimatoren blant HIV-positive og HIV-negative.



(Fortsettes side 7.)

Beskriv hvordan Kaplan-Meier estimatoren beregnes. Forklar hvorfor man med denne type data må benytte en slik estimator for å anslå sannsynlighetene for overlevelse. Beskriv ut i fra plottet forskjellen i dødelighet i de to gruppene.

b) Under er det angitt resultatene av en log-rank test mellom HIV-positive og HIV-negative.

```
           N Observed Expected (O-E)^2/E (O-E)^2/V
hivpos=0 823       55     69.1      2.87      13.4
hivpos=1 181       33     18.9     10.48      13.4

 Chisq= 13.4  on 1 degrees of freedom, p= 0.000245
```

Forklar hva utskriften forteller og sammenhold med punkt a).

c) Anta at dødeligheten kan beskrives ved en proporsjonal intensitetsmodell

$$\lambda_i(t) = \lambda_0(t)e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}}$$

der $\lambda_i(t)$ er intensiteten for individ nr. i, $\lambda_0(t)$ er en underliggende intensitet og kovariatene er $x_{i1} =$ alder, $x_{i2} =$ kjønn og $x_{i3} =$ indikator for å være HIV-positiv. Resultatene av denne modellen er angitt under.

```
           coef exp(coef) se(coef)       z       p
 alder  0.06295     1.065   0.0209  3.0056 0.00270
 kjonn -0.00893     0.991   0.2274 -0.0393 0.97000
hivpos  0.79483     2.214   0.2213  3.5914 0.00033

Likelihood ratio test=20.6  on 3 df, p=0.000126  n= 1004
```

Fortolk e^{β_j} og finn 95% konfidensintervall for disse. Sammenhold resultatet for HIV-positivitet med de foregående punktene.

d) Anta at studiedesignet var slik at vi kun kjente til om levetiden var større enn eller mindre enn en tid τ. Man kan vise at dersom den proporsjonale intensitetsmodellen i punkt c) holder blir sannsynligheten for et dødsfall gitt ved

$$1 - \exp(-\Lambda_0(\tau)e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}})$$

der $\Lambda_0(z) = \int_0^\tau \lambda_0(s)ds$. (Dette skal du ikke vise.)
Foreslå en metode for å estimere β_1, β_2 og β_3 fra et slikt studiedesign.

SLUTT