Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen:
Jo Eidsvik 90 12 74 72

Bokmål

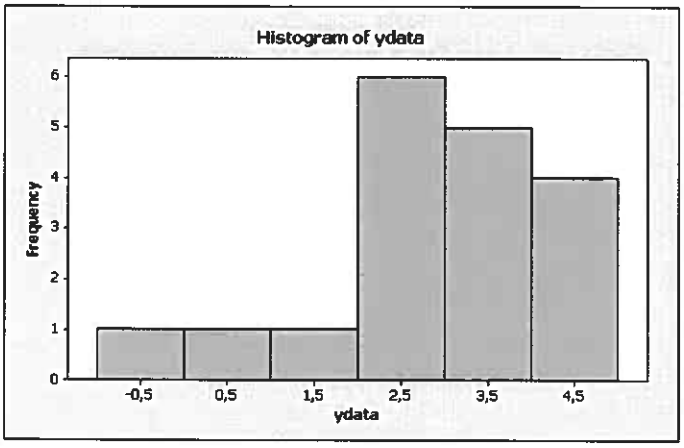
EKSAMEN I FAG ST2202 ANVENDT STATISTIKK

Tirsdag 11. desember 2007
Tid: 09:00–14:00

Tillatte hjelpemidler:
Alle trykte og håndskrevne hjelpemidler, alle kalkulatorer.
Sensur: 21.des 2007

Oppgave 1

Figur 1 viser et histogram av antatt uavhengige dataverdier $y_1, \ldots, y_{18}$, som har den reelle tallinjen som utfallsrom.



Figur 1: Histogram av 18 dataverdier.

Histogrammet er lagd ved bruk av 6 bins (intervaller). Disse er $(-\infty, 0)$, $(0, 1)$, $(1, 2)$, $(2, 3)$, $(3, 4)$, og $(4, \infty)$. Gjennomsnittet for datasettet er $\hat{\mu} = \frac{1}{18}\sum_{i=1}^{18} y_i = 2.97$, estimert standardavvik er $s = \sqrt{\frac{1}{17}\sum_{i=1}^{18}(y_i - \hat{\mu})^2} = 1.37$.

a) Anta at data er normalfordelt $N(\mu, \sigma^2)$. Vi vil teste hypotesen H_0: $\mu = 2$, mot H_1: $\mu \neq 2$. Finn testobservator og cirka p-verdi?

Hva skiller en ikke-parametrisk hypotesetest fra testen over? Bruk fortegnstesten til å undersøke om medianen er ulik 2, dvs test H_0: $\tilde{\mu} = 2$, mot H_1: $\tilde{\mu} \neq 2$. Finn testobservator og cirka p-verdi. Sammenlign med resultatet over.

Vis at en approksimativ fortegnstest, på 5 prosent signifikansnivå, i dette tilfelle er å forkaste H_0 dersom $|\frac{x-9}{\sqrt{4,5}}| > 1.96$, der x er antall data mindre enn 2.

b) Er det grunnlag for å påstå at data ikke er normalfordelte? Formuler dette som en hypotesetest.

Sett opp og beregn en testobservator basert på intervalloppdelingen i histogrammet, og estimert normalfordeling med sin forventning og varians. Spesifiser frihetsgradene til testobservatoren.

Utfør testen på signifikansnivå $0,05$ og tolk resultatet.

Oppgave 2

Anta at noen vurderer å sette opp en vindmølle et bestemt sted. Man registrerer vindstyrke fra værvarslet på dette stedet gitte dager, og logaritmen til dette varslet benevnes $x_{i,1}$ for $i = 1, \ldots, 20$ dager. I tillegg måler man gjennommsnittlig avlest vindstyrke på stedet de samme dagene. Logaritmen til denne benevnes y_i, for $i = 1, \ldots, 20$ dager. Basert på dette datamaterialet ønsker man å tilpasse en modell for vind basert på varsel. Man vurderer modellen

$$y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \text{uavhengige}, i = 1, \ldots, 20. \tag{1}$$

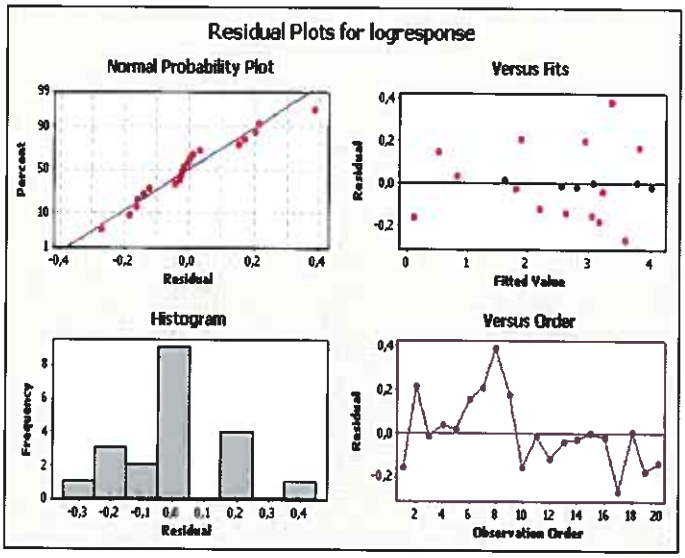
Regresjonsanalysen er gjengitt i Tabellen under.

Tabell 1: Tabell for regresjonskoeffisienter med faktor varsel.

Predictor	Coef	SE Coef	T	P
Constant	-0,21692	0,09993	-2,17	0,044
logvarsel	1,24067	0,04187	29,63	0,000

I tillegg oppgis at $SSE = \sum_{i=1}^{20}(y_i - \hat{y}_i)^2 = 0,48$, $SST = \sum_{i=1}^{20}(y_i - \bar{y})^2 = 24,0$, og at kovariansen mellom $\hat{\beta}_0$ og $\hat{\beta}_1$ er $-0,0037$.

Et plot av residualene er gitt i Figur 2.



Figur 2: Residualplott.

a) Regn ut s som er et estimat for σ. Regn videre ut R^2.

Hva plottes i hvert av residualplottene i Figur 2, og hva kan hvert plott brukes til? Gir modellen en god tilpasning til data? Hvorfor kan det her virke naturlig å bruke logaritmen til varslet og målt vindstyrke?

b) Lag et 95 prosent konfidensintervall for β_0.

Test hypotesen $H_0 : \beta_1 = 1$, mot alternativet $H_1 : \beta_1 \neq 1$. Bruk signifikansnivå $0,05$.

Hva er spesielt med varselverdien $x_1 = \beta_0/(1-\beta_1)$? Finn en estimator for denne, kalt $\hat{x}_1$, og regn ut estimatet i vårt tilfelle. Finn også approksimativ varians til $\hat{x}_1$.

Av de 20 dagene er 10 på sommeren og 10 på vinteren. En modell som tar hensyn til dette er:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i, \qquad \epsilon_i \sim N(0,\sigma^2), \text{uavhengige}, i = 1,\ldots,20. \tag{2}$$

der $x_{i,2} \in \{0,1\}$ er en indikatorvariabel for vinter, og $x_{i,3} = x_{i,1}x_{i,2}$.

Tabell 2: Tabell for regresjonskoeffisienter i modell med faktorer for varsel og sesong.

Predictor	Coef	SE Coef	T	P
Constant	-0,2957	0,1795	-1,65	0,119
logvarsel	1,23652	0,06758	18,30	0,000
sesong	0,0183	0,1949	0,09	0,926
sesonglogvarsel	0,10140	0,07743	1,31	0,209

Det oppgis i tillegg at $SSE = 0,2$, $SSR = 23,8$, og $SST = 24$ for denne modellen.

c) Lag basert på utskriften en enkel skisse av regresjonslinjen for y mot logaritmen til varslet vindstyrke, en linje for sommer og en for vinter.

Bruk en Fisher test til å undersøke om modellen med 4 parametre bør foretrekkes framfor modellen fra punkt a) og b) med 2 parametre. Bruk signifikansnivå $0,05$.

Anta, uansett konklusjon i c), at sesong ikke har noen effekt. Enda en alternativ modell er at logaritme til varsel og logaritme til målt vindstyrke egentlig begge gir den sanne log-vindstyrke z_i hver dag i, men at vi tillater et konstant avvik β_0 for varsel. Denne modellen formuleres med responser y_i og $x_{i,1}$ gitt ved

$$\begin{aligned} y_i &= z_i + \epsilon_i, \qquad \epsilon_i \sim N(0,\sigma^2), i = 1,\ldots,20, \\ x_{i,1} &= z_i + \beta_0 + \delta_i, \qquad \delta_i \sim N(0,\tau^2), i = 1,\ldots,20, \end{aligned} \tag{3}$$

der alle støyleddene antas uavhengige.

d) Sett modellen i (3) på matrise / vektor form $\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{v}$, der $\boldsymbol{y} = (y_1,\ldots,y_{20},x_{1,1},\ldots,x_{20,1})'$, og med vektor av ukjente parametre $\boldsymbol{\beta} = (z_1,\ldots,z_{20},\beta_0)'$. Spesifiser $\boldsymbol{X}$ og egenskapene til $\boldsymbol{v}$.

Den vanlige minste kvadraters estimatoren tar ikke hensyn til ulike varianser som vi har her. La $\boldsymbol{W}$ være en diagonal matrise med $1/\sigma^2$ i de første 20 diagonalelementene, og $1/\tau^2$ i de 20 siste.

Vis at minste kvadraters estimator for $\boldsymbol{\beta}$ nå kan skrives som

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}. \tag{4}$$

Hva er varians/kovariansen til $\hat{\boldsymbol{\beta}}$?

Hvordan forenkles utrykket dersom $\sigma = \tau$?

Oppgave 3

I søk i databaser brukes ofte nøkkelord. Utover dette forsøker man å søke gjennom kun en delmengde av dokumenter ved å bruke informasjon søkeren ikke 'ser'. Dette er typisk lengde på setninger i abstract, eller ordfordeling i abstract, osv., som kan klassifiseres utfra tidligere dokumenter innen et delområde. Anta her at vi har en samling av abstract fra en statistikkjournal som publiserer både anvendte og teoretiske artikler. I alt 71 artikler blir gjennomgått. Man teller antall setninger og en indikator om abstract har en formel eller ikke. I tillegg klassifiseres hver artikkel som 'anvendt' (ikke-suksess) eller 'teoretisk' (suksess). Datasettet er gjengitt i Tabellen under med kovariater $x_{i,1}$ for antall setninger, $x_{i,2}$ for formel eller ikke, y_i som antall suksesser, og n_i som antall abstract for hver kombinasjon.

Tabell 3: Datamateriale for klassifisering av abstract.

Setninger, x_1	Formel, x_2	Suksesser, y	antall, n
3	0	4	6
4	0	5	7
5	0	4	9
6	0	3	7
7	0	1	6
8	0	1	5
9	0	0	4
3	1	6	6
4	1	5	7
5	1	7	9
6	1	3	4
7	1	2	3

En logistisk regresjon gir følgende estimater:

Tabell 4: Estimering ved logistisk regresjon.

Predictor	Coef	SE Coef	Z	P
Constant	2,64	1,05	*	0,012
formel	1,38	*	2,36	0,018
setninger	*	0,18	-2,93	*

Log likelihood for denne modellen er $-39,4$.

a) Skriv opp de ulike elementene som inngår i en logistisk regresjonsmodell.

Fyll inn elementene merket med $*$ i Tabellen.

Gi utfra Tabellen en begrunnelse for at alle kovariatene skal være med i modellen over.

Hva er predikert sannsynlighet for suksess hvis setninger= 6 og formel= 1?

b) Oddsen for kovariat $\boldsymbol{x} = (1, x_1, x_2)$ er definert ved $O(\boldsymbol{x}) = p(\boldsymbol{x})/(1 - p(\boldsymbol{x}))$, der $p(\boldsymbol{x})$ er modellert sannsynlighet for suksess under kovariat $\boldsymbol{x}$.

Anta at antall setninger= 4. Estimer hvor mye oddsen relativt øker når kovariat 'formel' går fra lav til høy? (Dvs hvor mye den øker med en multiplikativ faktor.)

Finn approksimativ varians til denne relative økningen i odds, og lag et approksimativt 95 prosent konfidensintervall for endringen i odds.

c) Anta at en annen person har brukt et annet datamateriale og funnet at oddsen i tilfellet fra b) øker med $3,0$. Tilhørende varians er lik $1,0$. Anta at både løsning fra b) og denne personen sin estimator er forventningsrette.

Bruk estimat fra b), uansett hva det ble, og denne personen sitt estimat til å lage et forbedret kombinert estimat for oddsen.

Hva er variansen til den nye kombinerte estimatoren?

d) Dersom vi fjerner kovariat 'formel' blir log-likelihood $-42,3$.

Bruk forskjellen i log-likelihood (deviansen) til å undersøke om modellen med 'formel' er å foretrekke framfor modellen uten 'formel'. Formuler dette som en hypostesetest, finn en testobservator og gi approksimativ fordeling under hypostesen.

Regn ut p-verdi. Sammenlign med Punkt a).