Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Bokmål

Faglig kontakt under eksamen: Professor Jarle Tufto
Telefon: 99705519

Statistisk modellering for biologer og bioteknologer, ST2304

24. mai, 2013
Kl. 9–13
Sensur: 14. juni, 2013

Tillatte hjelpemidler: Et håndskrevet gult A4 ark, kalkulator, "Tabeller og formler i statistikk" (Tapir forlag), K. Rottmann: Matematisk formelsamling.

Hjelpesider for noen R funksjoner det kan hende du får bruk for følger på side 8.

Oppgave 1 Vi ønsker å undersøke om variansene σ_X^2 og σ_Y^2 i to normalfordelte populasjoner er forskjellige fra hverandre og trekker utvalg av størrelse 10 fra den en populasjonen og 20 fra den andre. Det kan da vises at testobservatoren $F = S_X^2/S_Y^2$, hvor S_X^2 og S_Y^2 er de to utvalgsvariansene, er F-fordelt med 9 og 19 frihetsgrader under nullhypotesen $H_0 : \sigma_X^2 = \sigma_Y^2$.

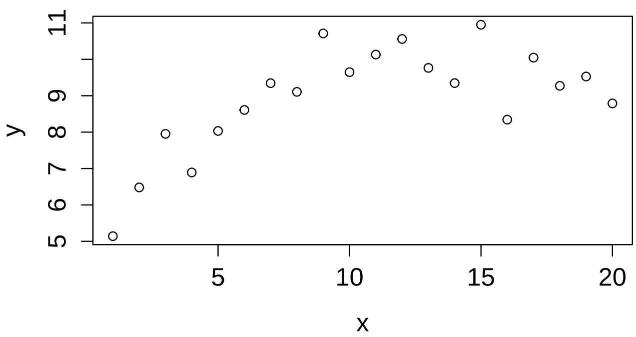
a) Skriv et R uttrykk som beregner de kritiske verdiene for testen hvis vi velger et signifikansnivå på 0.05.

b) Anta at estimatet av de to variansene blir henholdsvis 13.5 og 5.2. Skriv et R-uttrykk som beregner testens signifikanssannsynlighet.

c) Er testobservatoren diskret eller kontinuerlig fordelt? Hva er sannsynligheten for at testobservatoren tar en verdi eksakt lik 1?

d) Skriv et R-uttrykk som simulerer 1000 realisasjoner av testobservatoren under antakelsen om at H_0 er sann og som plotter et histogram av disse realisasjonene.

Oppgave 2 Vi ønsker å finne optimal temperatur for vekst av kveiteyngel og måler vekst y (gram/uke) ved 19 ulike vanntemperaturer x (°C) under ellers like betingelser. De observerte dataene er vist under.



Anta at vi modellerer sammenhengen mellom responsvariabelen y (vekst) og temperatur x ved hjelp av multippel regresjon hvor vi inkluderer temperatur x og kvadratet av temperatur, x^2, som forklaringsvariable som følger.

```
> x2 <- x^2
> modell <- lm(y ~ x + x2)
> summary(modell)

Call:
lm(formula = y ~ x + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55021 -0.31092  0.03203  0.38368  1.04305

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.862606   0.497222   9.780 2.14e-08 ***
x            0.816051   0.109049   7.483 8.95e-07 ***
x2          -0.031351   0.005044  -6.215 9.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.6683 on 17 degrees of freedom
Multiple R-squared: 0.8165,Adjusted R-squared: 0.7949
F-statistic: 37.82 on 2 and 17 DF,  p-value: 5.504e-07
```

a) Skriv opp modellen vi har tilpasset i matematisk (algebraisk) notasjon og hva som er forutsetningen for modellen. Hvilke forklaringsvariabler har en statistisk signifikant effekt på responsvariabelen?

b) Vis at optimal veksttemperatur x_0 er funksjonen

$$x_0 = f(b_1, b_2) = -\frac{b_1}{2b_2}, \tag{1}$$

hvor b_1 er regresjonskoeffisienten for temperatur x og b_2 er regresjonskoeffisienten for kvadratet av temperatur x^2. Hint: Sett den deriverte av vekst y m.h.p. temperatur x lik 0 og løs for x.

Hva blir estimatet $\hat{x}_0$ av optimal veksttemperatur? Virker dette rimelig ut i fra de observerte dataene?

c) Beregn standardfeilen til $\hat{x}_0$. Du vil blant annet trenge et estimat av kovariansen mellom $\hat{b}_1$ og $\hat{b}_2$ som kan leses ut av følgende R-utskrift. Se hjelpsiden for `vcov` for mer informasjon.

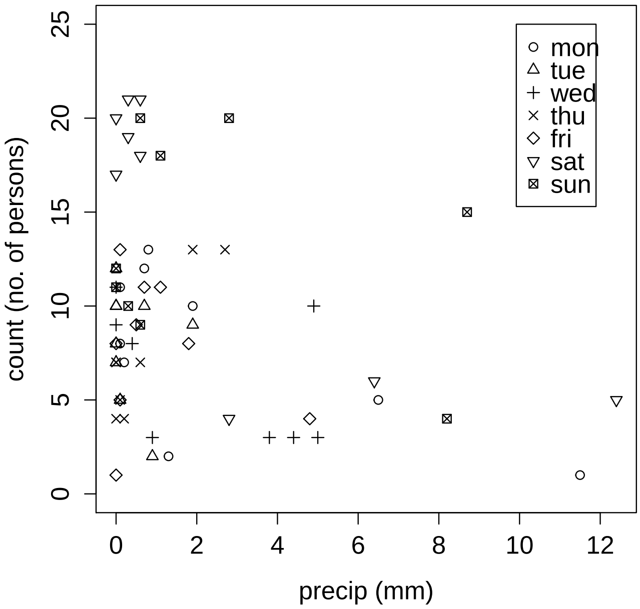
```
> vcov(modell)
             (Intercept)         x        x2
(Intercept)   0.247229 -0.048192  0.001959
x            -0.048192  0.011891 -0.000534
x2            0.001959 -0.000534  0.000025
```

Kommenter igjen om svaret virker rimelig ut i fra figuren over.

Oppgave 3 Som del av en kartlegging av rekreasjonell bruk av Bymarka i Trondheim telles antall turgåere, syklister og joggere som passerer et gitt punkt på vei inn i marka mellom kl 0900 og 2000 hver dag i løpet av juni og juli i 2012. Vi legger inn dataene som følgende dataramme i R.

```
> bymarka
   weekday weekend precip count
1      fri      no    0.1    13
2      sat     yes    0.3    21
3      sun     yes    2.8    20
4      mon      no    0.2     7
5      tue      no    0.7    10
6      wed      no    0.0    11
7      thu      no    1.9    13
8      fri      no    0.5     9
9      sat     yes    0.0    20
10     sun     yes    0.6    20
11     mon      no    0.1    11
12     tue      no    0.0    10
13     wed      no    0.4     8
14     thu      no    2.7    13
15     fri      no    1.8     8
16     sat     yes    0.6    21
17     sun     yes    8.7    15
18     mon      no    0.8    13
19     tue      no    0.0    10
20     wed      no    0.0     9
21     thu      no    0.0     7
.
.
.
57     fri      no    0.0     1
58     sat     yes   12.4     5
59     sun     yes    8.2     4
60     mon      no    6.5     5
61     tue      no    0.0    12
```

Variablene `weekday` og `weekend` er faktorer som representer henholdsvis ukedag og hverdag/helg. Variabelen `precip` er antall millimeter nedbør i hvert døgn og `count` antall personer som passerte hver av dagene. Dataene er grafisk framstilt i følgende figur.



Vi tilpassar først følgende modell (modell A).

```
> fitA <- glm(count ~ weekend + precip,fam=poisson)
> summary(fitA)

Call:
glm(formula = count ~ weekend + precip, family = poisson(link="log")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2315  -1.2276  -0.0180   0.9036   2.3345

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.12138    0.05893  35.999  < 2e-16 ***
weekendyes   0.69069    0.08493   8.133 4.20e-16 ***
precip      -0.08896    0.01830  -4.860 1.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 188.07  on 60  degrees of freedom
Residual deviance: 108.11  on 58  degrees of freedom
AIC: 352.32

Number of Fisher Scoring iterations: 5

> drop1(fitA,test="Chisq")
Single term deletions

Model:
count ~ weekend + precip
        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       108.11 352.32
weekend  1   171.21 413.41 63.093 1.972e-15 ***
precip   1   136.50 378.71 28.390 9.919e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) Forklar hvorfor Poisson-antakelsen kan være et rimelig utgangspunkt. Hvor mange forventes å passere på helgedager i forhold til på hverdager? Er forskjellen statistisk signifikant? Hva blir forventet antall passerende i en helg på en dag med 5mm nedbør?

For å teste om det er en ytterligere forskjell mellom forventet antall turgåere på forskjellige ukedager utover helge-effekten tilpasser vi følgende alternative modell (modell B).

```
> fitB <- glm(count ~ weekday + precip,fam=poisson)
> summary(fitB)

Call:
glm(formula = count ~ weekday + precip, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2586  -1.1416  -0.0993   0.9164   2.4785

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.22000    0.12422  17.871  < 2e-16 ***
```

```
weekdaytue  -0.09275    0.16958  -0.547 0.584407
weekdaywed  -0.19018    0.18585  -1.023 0.306167
weekdaythu  -0.14767    0.17758  -0.832 0.405633
weekdayfri  -0.08671    0.17030  -0.509 0.610648
weekdaysat   0.63720    0.14876   4.283 1.84e-05 ***
weekdaysun   0.54393    0.15132   3.595 0.000325 ***
precip      -0.08874    0.01857  -4.777 1.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 188.07  on 60  degrees of freedom
Residual deviance: 106.33  on 53  degrees of freedom
AIC: 360.54

Number of Fisher Scoring iterations: 5

> drop1(fitB,test="Chisq")
Single term deletions

Model:
count ~ weekday + precip
        Df Deviance    AIC    LRT  Pr(>Chi)
<none>       106.33 360.54
weekday  6   171.21 413.41 64.877 4.572e-12 ***
precip   1   133.60 385.81 27.269 1.770e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

b) Forklar hvorfor modell A er nøstet i modell B. Hva er antall estimerte parametere under henholdsvis modell A og B? Utfør en test av om det er en innbyrdes forskjell mellom henholdsvis ulike hverdager og ulike helgedager i forventet antall passerende utover helg/hverdag-effekten. Hint: Dette kan ikke leses direkte ut av utskriften over men må gjøres «for hånd». Er modell A eller B å foretrekke ut i fra de observerte dataene? Forutsett at det ikke er overdispersjon i dataene.

c) Test om det er overdispersjon under den valgte modellen. Diskuter mulige mekanismer som kan tenkes å generere overdispersjon i dette tilfelle.

```
FDist                 package:stats                  R Documentation

The F Distribution

Description:

     Density, distribution function, quantile function and random
     generation for the F distribution with 'df1' and 'df2' degrees of
     freedom (and optional non-centrality parameter 'ncp').

Usage:

     df(x, df1, df2, ncp, log = FALSE)
     pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
     qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
     rf(n, df1, df2, ncp)

Arguments:

    x, q: vector of quantiles.

       p: vector of probabilities.

       n: number of observations. If 'length(n) > 1', the length is
          taken to be the number required.

df1, df2: degrees of freedom.  'Inf' is allowed.

     ncp: non-centrality parameter. If omitted the central F is
          assumed.

log, log.p: logical; if TRUE, probabilities p are given as log(p).

lower.tail: logical; if TRUE (default), probabilities are P[X <= x],
          otherwise, P[X > x].

Details:

     The F distribution with 'df1 =' n1 and 'df2 =' n2 degrees of
     freedom has density

     f(x) = Gamma((n1 + n2)/2) / (Gamma(n1/2) Gamma(n2/2))
         (n1/n2)^(n1/2) x^(n1/2 - 1)
         (1 + (n1/n2) x)^-(n1 + n2)/2

     for x > 0.

     It is the distribution of the ratio of the mean squares of n1 and
     n2 independent standard normals, and hence of the ratio of two
     independent chi-squared variates each divided by its degrees of
     freedom.  Since the ratio of a normal and the root mean-square of
     m independent normals has a Student's t_m distribution, the square
     of a t_m variate has a F distribution on 1 and m degrees of
     freedom.

     The non-central F distribution is again the ratio of mean squares
     of independent normals of unit variance, but those in the
     numerator are allowed to have non-zero means and 'ncp' is the sum
     of squares of the means.  See Chisquare for further details on
     non-central distributions.

Value:

     'df' gives the density, 'pf' gives the distribution function 'qf'
     gives the quantile function, and 'rf' generates random deviates.

     Invalid arguments will result in return value 'NaN', with a
     warning.

Note:

     Supplying 'ncp = 0' uses the algorithm for the non-central
     distribution, which is not the same algorithm used if 'ncp' is
     omitted.  This is to give consistent behaviour in extreme cases
     with values of 'ncp' very near zero.

     The code for non-zero 'ncp' is principally intended to be used for
     moderate values of 'ncp': it will not be highly accurate,
     especially in the tails, for large values.

Source:

     For the central case of 'df', computed _via_ a binomial
     probability, code contributed by Catherine Loader (see 'dbinom');
     for the non-central case computed _via_ 'dbeta', code contributed
     by Peter Ruckdeschel.

     For 'pf', _via_ 'pbeta' (or for large 'df2', _via_ 'pchisq').

     For 'qf', _via_ 'qchisq' for large 'df2', else _via_ 'qbeta'.

References:

     Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) _The New S
     Language_.  Wadsworth & Brooks/Cole.

     Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) _Continuous
     Univariate Distributions_, volume 2, chapters 27 and 30.  Wiley,
     New York.

See Also:

     Distributions for other standard distributions, including 'dchisq'
     for chi-squared and 'dt' for Student's t distributions.

Examples:

     ## the density of the square of a t_m is 2*dt(x, m)/(2*x)
     # check this is the same as the density of F_{1,m}
     x <- seq(0.001, 5, len=100)
     all.equal(df(x^2, 1, 5), dt(x, 5)/x)

     ## Identity:  qf(2*p - 1, 1, df)) == qt(p, df)^2)  for  p >= 1/2
     p <- seq(1/2, .99, length=50); df <- 10
     rel.err <- function(x,y) ifelse(x==y,0, abs(x-y)/mean(abs(c(x,y))))
     quantile(rel.err(qf(2*p - 1, df1=1, df2=df), qt(p, df)^2), .90)# ~= 7e-9

----------------------------------------------------------------------------
vcov                  package:stats                  R Documentation

Calculate Variance-Covariance Matrix for a Fitted Model Object

Description:

     Returns the variance-covariance matrix of the main parameters of a
     fitted model object.

Usage:

     vcov(object, ...)

Arguments:

  object: a fitted model object, typically.  Sometimes also a
          'summary()' object of such a fitted model.

     ...: additional arguments for method functions.  For the 'glm'
          method this can be used to pass a 'dispersion' parameter.

Details:

     This is a generic function.  Functions with names beginning in
     'vcov.' will be methods for this function.  Classes with methods
     for this function include: 'lm', 'mlm', 'glm', 'nls',
     'summary.lm', 'summary.glm', 'negbin', 'polr', 'rlm' (in package
     'MASS'), 'multinom' (in package 'nnet') 'gls', 'lme' (in package
     'nlme'), 'coxph' and 'survreg' (in package 'survival').

     ('vcov()' methods for summary objects allow more efficient and
     still encapsulated access when both 'summary(mod)' and 'vcov(mod)'
     are needed.)

Value:

     A matrix of the estimated covariances between the parameter
     estimates in the linear or non-linear predictor of the model.
     This should have row and column names corresponding to the
     parameter names given by the 'coef' method.
```