Nýtísku aðhvarfsgreining (GLIM)

Helgi Tómasson

28. mars 1998

- Dagur 1: Hvað er tölfræði?

  Líkindafræði, ályktunarfræði, líkanasmíði

- Dagur 2: Upprifjun, línuleg líkön

  (ANOVA, aðhvarfsgreining), One-way, two-way o.s.frv.

- Dagur 3: Praktísk dæmi með tölvu,

  útvíkkun á venjulegu línulegu líkani.

- Dagur 4: GLIM-líkön, t.d. logisic og log-linear líkön

- Dagur 5: Praktísk dæmi, notkun á R forritinu.

- Dagur 6: Praktísk dæmi, yfirlit

- Bækur: Allar bækur með regression í titlinum

McCullagh, P., and Nelder, F.R.S.: Generalized Linear Models, Chapman Hall (1989 eða 1983)

Venables, W.N. and Ripley B.D.: Modern Applied Statistics with S-plus, Springer , 1996

Hvað er tölfræði?

- Líkindafræði (Probabilty) + Ályktunarfræði (Inference)
- Líkanasmíði (modelling): Einfölduð mynd af raunveruleika
- Ályktun = líkan + gögn
- Upprifjun Líkindafræði, nokkur hugtök
- Tilraun, atburður, líkur
- Reiknireglur um líkur, óháðir atburði, skilyrtar líkur, Regla Bayes
- Random breyta = Hending (e. Random variable)

- Eiginleikar hendinga, samfelldar/sundurslitnar
- Dreifing: Dreififall(cdf) , þéttifall (density, pfd), líkandafall (pmf=probability mass function)
- Lýsitölur fyrir dreifingar, væntanleg gildi (expected value), dreifni (variance), miðgildi (median) o.s.frv.
- Margvíðar hendingar, samand hendinga, fylgni.
- Reiknireglur fyrir væntanlegt gildi og varíans(dreifni)
- Ýmsar dreifingar, hvers vegna hafa þær nafn, verkefni.

- Ályktunarfræði, Mat (punktmat, bilmat) Kenningaprófanir
- Mat, Punktmat, Metill(estimator), hlutdrægni(bias), samkvæmni (consistency), nýtni (efficiency)
- Mat, Bilmat, öryggismörk (confidence limits).
- Kenningaprófanir, Núllkenning, valkostur, villa I, villa II, marktækni(significance), styrkur (power) Verkefni
- Tveir skólar ályktunarfræði, Hefðbundinn(tíðni-skóli), Bayesian skóli.

- Aðhvarf(regression) og fylgni(correlation)
- Ýmsar tegundir af breytum.
- Aðhvarf meðs einni skýristærð,

  aðhvarf með mörgum skýristærðum,

  tengsl við ANOVA,

  tengsl við samanburð meðaltala Verkefni
- Tegundir parametra (í. stika), main effects, interactions, intercept, slope. Túlkanir.
- Nokkur atriði um matsaðferðir.
- GLIM fjölskyldan

  Venjuleg ANOVA, aðhvarfslíkön,

- Tilraun (experiment), aðgerð sem gefur útkomu

  Ùtkoma óviss, atburður = mengi af útkomum

- Mengja-stærðfræði hentug,

  A $\cap B$ = atburðurinn A og B

  A $\cup$ B = atburðurinn A eða B (geta verið báðir)

  Ω = mengi af öllum mögulegum útkomum

- Líkur

  P vörpun sem úthlutar atburði tölu á milli 0 og 1.

  P(A $\cup$ B) = P(A) + P(B) - P(A $\cap$ B)

- Óháðir (independent) atburðir

  $P(A \cap B) = P(A)\ P(B)$

- skilyrtar líkur (conditional probability)

  $A \mid B$ = atburðurinn A gefið atburðurinn B

  $P(A \mid B) = P(A \cap B)/P(B)$

- regla Bayes

  $P(A \mid B) = P(B \mid A)\ P(A)/P(B)$

  Dæmi: Á að mótefnamæla population?

  Segjum 250.000 manna population,

  prevalence = 500

  sensitivity = specificty =0.95

  reiknið $P(S{=}+ \mid T{=}+)$

  svar: $P(S{=}+ \mid T{=}+)$*250.000

  = mörg þúsund

- Dreififall(Cdf) hendingar

  F(x) = P(X $\leq$ x) (ath. stórt og lítið x)

- þéttifall(density, pdf) líkindamassafall (pmf)

  f(x) = F$'$(x) ef X er samfelld

  f(x) = P(X=x) ef X er sundurslitin

- Lýsitölur fyrir hendingar

  E(X) = expected value =væntanlegt gildi

  E(X) = $\int$ x f(x) dx ef X er samfelld

  E(X) = $\sum$ x f(x) ef X er sundurslitin

  V(X) = variance X = dreifni X

  V(X) = ${\sigma_X}^2$ = E(X-E(X))^2

  σ_X = staðalfrávik X

  x$_q$ = q-kvantíll $\iff$ F(x$_q$)=q

- Margvíðar hendingar (multivariate random variables)

  F(x,y) = P(X $\leq$ x $\cap$ Y$\leq$ y)

  f(x,y) = F-diffrað fyrir samfelldar hendingar

$f(x,y) = P(X{=}x \cap Y{=}y)$ fyrir sundurslitnar

- X og Y óháðar $\iff$ $F(x,y)=F(x)F(y)$
- Y|X hefur dreififall $F_{Y|X}(y|x)$
- Reiknireglur væntanlegt gildi og varíans
- $E(aX) = a\,E(X)$
- $V(aX) = a^2V(X)$

- Fylkja-algebra (matrix-algebra mjög þægileg

  Ef X er vektor af hendingum

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

$$V(\mathbf{X}) = \begin{bmatrix} V(X_1) & \ldots & COV(X_1, X_n) \\ \vdots & \vdots & \vdots \\ COV(X_1, X_n) & \vdots & V(X_N) \end{bmatrix}$$

$$COV(X_i, X_j) = E(X_i - E(X_i))(X_j - E(X_j))$$

$$CORR(X_i, X_j) = \frac{COV(X_i, X_j)}{\sqrt{V(X_i)V(X_j)}}$$

- Nokkrar tegundur sundurslitinna (discrete) dreifinga

  Bernoulli (p)

  Binomial, B(n,p)

  Poisson, λ

  Geometric (p)

  Negatív binomial (r,p)

  Uniform (a,b)

- Nokkrar tegundir samfelldra hendinga

  Uniform (a,b)

  Exponential λ

  Gamma (r, λ)

  Normal N(μ, σ^2)

  χ^2(k)

  t(k) cauchy=t(1)

  F(n,m)

Verkefni Búið til 30 mælingar úr a) exponential dreifingu með meðaltal 180, b) N(180,7^2) og c) cauchy með miðgildi 180, og 5% meira en 190. Reiknið úrtaksmeðaltal, úrtaksstaðalfrávik og staðalfrávik úrtaksmeðaltalsins. Endurtakið fyrir úrtaksstærð 300 og 3000.

Verkefni Búið til 30 mælingar úr normaldreifingu með meðaltal 0 og staðalfrávik 1. Reiknið 95Endurtakið 50-1000 sinnum. Notið tölvu og viðeigandi forrit. T.d. EXCEL eða einhvern annan töflureikni.

Verkefni Búið til 30 mælingar úr normaldreifingu með meðaltal 0 og staðalfrávik 1. Prófið kenninguna að meðaltalið sé 0 og gangið út frá að vitað sé að staðalfrávik sé 1. Er kenningunni hafnað miðað við α=0.05? Enrdurtakið 50-1000 sinnum.

Verkefni Búið til 30 mælingar úr normaldreifingu með meðaltal 0.2 og staðalfrávik 1. Prófið kenninguna að meðaltalið sé 0 og gangið út frá að vitað sé að staðalfrávik sé 1. Er kenningunni hafnað miðað við α=0.05? Enrdurtakið 50-1000 sinnum.

Verkefni Búið til 30 mælingar úr normaldreifingu með meðaltal 0.4 og staðalfrávik 1. Prófið kenninguna að meðaltalið sé 0 og gangið út frá að vitað sé að staðalfrávik sé 1. Er kenningunni hafnað miðað við α=0.05? Enrdurtakið 50-1000 sinnum.

Nokkur atriði um logistic regression

Það sem meta á eru líkur á ákveðnum atburði, t.d. bilun sjúkdómi eða þess háttar. Köllum þá breytu Y. p=P(Y=1), 1-p=P(Y=0).

Gerum ráð fyrir að

p= $e^{x\beta}$ / (1 + $e^{x\beta}$)

sem er jafngilt og

log(p/(1-p)) = $x\beta$

Likelihood-fall fyrir eina mælingu er:

$L(\beta \mid y,x) = p^{y} \, (1\text{-}p)^{1-y}$

$\log(L) = y \log(p) + (1\text{-}y) \log(1 - p)$

Fyrir n óháðar mælingar er log-likelihood fallið því

$\Sigma_{i=1}^{n} \left(y_i \log(p_i) + (1\text{-}y_i) \log(1\text{-}p_i)\right)$

$p_i = e^{x_i\beta} / (1 + e^{x_i\beta})$

Nokkur atriði um survival-greiningu.

Helstu hugtök

S(t) survival fall, S(t) = 1 - F(t) þar sem hendining T er endingartími þess sem rannsaka skal.

h(t) hazard fall, h(t) = f(t)/S(t)

H(t) cumulative hazard fall, H(t) = $\int_0^t$ h(s) ds

Dæmi, Líftími exponential dreifður, f(t) = $\lambda e^{-\lambda t}$

þá er S(t) = 1 - F(t) = 1 - (1 - $e^{-\lambda t}$) = $e^{-\lambda t}$

og

h(t) = f(t)/S(t) = $\lambda e^{-\lambda t}$ /$e^{-\lambda t}$ = λ

H(t) = λt

Censoring, það liggur í hlutarins eðli að við getum ekki séð heildar feril hvers og eins. Það munu verða margir sem við vitum ekki dánardægur á. Til eru ýmsar tegundir af censoring.

Aðferðir til að meta survival feril. E.t.v. best þekkt Kaplan-Meier og Flemming-Harrington. Einni hugsanlegt að negla niður fyrirfram ákveðið form eins og t.d. Weibull-dreifingu.

Oft er S(t) sjálft ekki aðaláhugamálið heldur samanburður ýmissa hópa og áhrif skýristærða.

Get gefið mér fjölskyldu af líftímadreifingum, weibull, log-normal, log-logistic o.s.frv. og síðan framkvæmd einhvers konar aðhvarfsgreiningu. (eins konar GLM)

Cox-regression, gengur út á að hægt sé að þátta hazardfallið h(t)á á eftirfarandi hátt

h(t,x) = hazard fall fyrir einstakling með eiginleika x

= h(t) exp(xβ)

Þ.e. að ein grunn-hazard lína gildi fyrir alla og síðan sé munurinn eingöngu fólginn í þættinum exp(xβ).

Nokkur atriði um tölfræðilega líkanagerð

Skilgreina markmið

Skilgreina áhugaverð tengsl, átta sig á truflandi þáttum

Við experimental aðstæður, kynna sér experimental design til að fá hámarksupplýsingar út úr mælingunum.

Setja fram hugsanlegt líkan, þ.e. átta sig á hugsanlegri dreifingu mælinga, hugsanlegu formi líkans.

Meta Líkan, t.d. með maximum-likelihood aðferð.

Framkvæma ítarlega diagnostics, leifaskoðun (residual analysis). Àtta sig á að hvernig líkan misstígur sig. Átta sig á outliers og influential observations. Athugið að til eru margs konar leifar, í venjulegum normal líkönum, t.d. least-squares residuals og recursive residuals. Ì GLIM og survival eru margskonar leifar.

Ef líkan stendst skoðun túlka niðurstöður. Varist over-fitting, þ.e. að finna líkan sem hentar mældum gögnum afar vel en gæti misstígið sig illa ef því væri beitt á annað gagnasett. Almennt gildir að velja skuli líkönin sem minnst (principle of parsimony)