Líftölfræði 0.4.06.29

Viðskipta- og Hagfræðideild

Lesnámsskeið, 3 einingar, vorið 2005.

Kennari: Helgi Tómasson

mailto:helgito@hi.is

http://www.hi.is/~helgito/liftolfr.html

4. janúar 2005

Lýsing

Farið verður yfir nokkur áhættuhugtök, svo sem RR, OR og AR. Farið verður í grunnatriði líftímagreiningar (survival teoríu), Cox-regression ofl.

Lýsing

Farið verður yfir nokkur áhættuhugtök, svo sem RR, OR og AR. Farið verður í grunnatriði líftímagreiningar (survival teoríu), Cox-regression ofl.

Einnig verður farið í einföld líkön þar sem háða breytan er tíðni (lág tíðni), t.d. Poisson líkön.

Lýsing

Farið verður yfir nokkur áhættuhugtök, svo sem RR, OR og AR. Farið verður í grunnatriði líftímagreiningar (survival teoríu), Cox-regression ofl.

Einnig verður farið í einföld líkön þar sem háða breytan er tíðni (lág tíðni), t.d. Poisson líkön.

Líkön þar sem að háða breytan er flokkabreyta, röðuð eða óröðuð verða rifjuð upp, t.d. með logistískri aðhvarfsgreiningu (regression).

Ásamt venjulegu normal líkani mynda sum þessara líkana sem er kallaður GLM (generalize-linear-model). Slík líkön eru vinsæl vegna þess hve auðtúlkanleg þau eru.

Ásamt venjulegu normal líkani mynda sum þessara líkana sem er kallaður GLM (generalize-linear-model). Slík líkön eru vinsæl vegna þess hve auðtúlkanleg þau eru.

Tækni, forrit sem ráða vel við slík líkön verða kynnt.

Ásamt venjulegu normal líkani mynda sum þessara líkana sem er kallaður GLM (generalize-linear-model). Slík líkön eru vinsæl vegna þess hve auðtúlkanleg þau eru.

Tækni, forrit sem ráða vel við slík líkön verða kynnt.

Grundvallaratriði með einstaklingsgögn er að einstaklingar eru mismunandi. Við smíði tölfræðilegra líkana er mikilvægt að taka tillit til þessa atriðis. Það þarf að gæta varúðar ef lagðir eru saman heterogen hópar. Leið fram hjá þessu er að vinna með margar mælingar á hverjum einstakling þannig að hægt sé að leiðrétta fyrir því að einstaklingar eru mismunandi. Þetta er stundum kallað repeated-measures, longitudinal-data eða panel-data. Aðferðir til að vinna með slík gögn eru kynntar. Bæði eins og sagt er frá þeim í Woolridge og einnig verður minnst á GLMM (generalized-linear-mixed-model).

Mælingar á að túlka í samhengi við tölfræðilegt líkan. Túlkunin fer þannig fram að óþekktir hluta líkans eru metnir. Ýmis hugtök sem eru nauðsynleg við mat og prófanir líkana eru rifjaðir upp. M.a. sennileikafall (likelihood function), method-of-moments, GMM (generalized-method-of-moments), likelihood-ratio próf o.s.frv. Einnig verður minnst á least-squares, Bayes aðferðir, kernel-estimation, bootstrap o.s.frv.

Markmið

Auka skilning á þeim líkönum sem koma við sögu í túlkun heilsufarsupplýsinga.

Lesefni

- Wooldridge,J.M.: Econometric Analysis of Cross Section and Panel Data.

Lesefni

- Wooldridge,J.M.: Econometric Analysis of Cross Section and Panel Data.
- Spiegelhalter, D.J.: Incorporating Bayesian Ideas into Health-Care Evaluation, Stat. Sci, 2004

Lesefni

- Wooldridge,J.M.: Econometric Analysis of Cross Section and Panel Data.
- Spiegelhalter, D.J.: Incorporating Bayesian Ideas into Health-Care Evaluation, Stat. Sci, 2004
- Berry, D.A.: Bayesian Statistics and the Efficiency and Ethics of Clinical Trials, Stat. Sci 2004

Lesefni

- Wooldridge,J.M.: Econometric Analysis of Cross Section and Panel Data.
- Spiegelhalter, D.J.: Incorporating Bayesian Ideas into Health-Care Evaluation, Stat. Sci, 2004
- Berry, D.A.: Bayesian Statistics and the Efficiency and Ethics of Clinical Trials, Stat. Sci 2004

- Deb, P. and Holmes, A.M.: Estimates and Costs of Behavioural Health Care: A Comparison of Standard and Finite Mixture Models: Health Ecnomics, 2000

- Deb, P. and Holmes, A.M.: Estimates and Costs of Behavioural Health Care: A Comparison of Standard and Finite Mixture Models: Health Ecnomics, 2000
- Grootendorst, P.V.: A Comparison of Alternative Models of Prescription Drug Utilization, Health Ecnomics, 1995.

Efni sem tekið var í fyrra sem verður sleppt eða breytt

- Diagnostic literatur: Í fyrra var dreift ljósriti úr Belsley, Kuh og Welsch

Efni sem tekið var í fyrra sem verður sleppt eða breytt

- Diagnostic literatur: Í fyrra var dreift ljósriti úr Belsley, Kuh og Welsch
- Í fyrra hélt ég sér fyrirlestur um eigin grein.
- Í fyrra var kynnt R-forritið.

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR
- Um survival teóríu, kafi 20 í bók. Hér á að þekkja helstu hugtök hazard-fall,

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR
- Um survival teóríu, kafi 20 í bók. Hér á að þekkja helstu hugtök hazard-fall,

  censoring, ýmsar tengundir af sampling.

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR
- Um survival teóríu, kafi 20 í bók. Hér á að þekkja helstu hugtök hazard-fall,

  censoring, ýmsar tengundir af sampling.

  Einnig á að koma yfir sig eiginleikum einfaldra parametríska líkana.

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR
- Um survival teóríu, kafi 20 í bók. Hér á að þekkja helstu hugtök hazard-fall,

  censoring, ýmsar tengundir af sampling.

  Einnig á að koma yfir sig eiginleikum einfaldra parametrískra líkana.
- Um binary regression, kafli 15 í bók.

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR
- Um survival teóríu, kafi 20 í bók. Hér á að þekkja helstu hugtök hazard-fall,

  censoring, ýmsar tengundir af sampling.

  Einnig á að koma yfir sig eiginleikum einfaldra parametrískra líkana.
- Um binary regression, kafli 15 í bók.

  Þekkja logit og probit líkön.

Námsyfirferð

- Ýmis áhættuhugtök, OR, RR, AR
- Um survival teóríu, kafi 20 í bók. Hér á að þekkja helstu hugtök hazard-fall,

  censoring, ýmsar tengundir af sampling.

  Einnig á að koma yfir sig eiginleikum einfaldra parametrískra líkana.
- Um binary regression, kafli 15 í bók.

  Þekkja logit og probit líkön.

- Um counting gögn kafli 19 í bók.

- Um counting gögn kafli 19 í bók.

  Poisson-regression, overdispersion, negative-binomial líkan.

- Um counting gögn kafli 19 í bók.

  Poisson-regression, overdispersion, negative-binomial líkan.

- Treatment greining, kafli 18 í bók.

- Um counting gögn kafli 19 í bók.

  Poisson-regression, overdispersion, negative-binomial líkan.

- Treatment greining, kafli 18 í bók.

  Þekkja á hugtök eins og ATE, selection-process og átta sig á áhrifum þess að vera með non-random sample.

- Um counting gögn kafli 19 í bók.

  Poisson-regression, overdispersion, negative-binomial líkan.

- Treatment greining, kafli 18 í bók.

  Þekkja á hugtök eins og ATE, selection-process og átta sig á áhrifum þess að vera með non-random sample.

- Kaflar 10 og 11 í bók um panel data.

- Um counting gögn kafli 19 í bók.

  Poisson-regression, overdispersion, negative-binomial líkan.

- Treatment greining, kafli 18 í bók.

  Þekkja á hugtök eins og ATE, selection-process og átta sig á áhrifum þess að vera með non-random sample.

- Kaflar 10 og 11 í bók um panel data.

  Átta sig á hvers vegna panel-gögnum er safnað.

- Um counting gögn kafli 19 í bók.

  Poisson-regression, overdispersion, negative-binomial líkan.

- Treatment greining, kafli 18 í bók.

  Þekkja á hugtök eins og ATE, selection-process og átta sig á áhrifum þess að vera með non-random sample.

- Kaflar 10 og 11 í bók um panel data.

  Átta sig á hvers vegna panel-gögnum er safnað.

Skilja hvers vegna orðanotkunin fixed effect og random effect í ekonometríu er ekki alveg heppileg. Skilja af hverju stundum vilja menn nota random-effects aðferðir og hvers vegna stundum fixed-effects aðferðir.

Skilja hvers vegna orðanotkunin fixed effect og random effect í ekonometríu er ekki alveg heppileg. Skilja af hverju stundum vilja menn nota random-effects aðferðir og hvers vegna stundum fixed-effects aðferðir.

Þekkja á hugtök eins og dulinn einstaklingsbreytileika, mixture models, GLMM.

Skilja hvers vegna orðanotkunin fixed effect og random effect í ekonometríu er ekki alveg heppileg. Skilja af hverju stundum vilja menn nota random-effects aðferðir og hvers vegna stundum fixed-effects aðferðir.

Þekkja á hugtök eins og dulinn einstaklingsbreytileika, mixture models, GLMM.

- Æskilegt að lesa fyrstu blaðsíður í köflum 16 og 17.

Skilja hvers vegna orðanotkunin fixed effect og random effect í ekonometríu er ekki alveg heppileg. Skilja af hverju stundum vilja menn nota random-effects aðferðir og hvers vegna stundum fixed-effects aðferðir.

Þekkja á hugtök eins og dulinn einstaklingsbreytileika, mixture models, GLMM.

- Æskilegt að lesa fyrstu blaðsíður í köflum 16 og 17.

  Þekkja þarf censoring og truncation hugtökin.

Af hverju repeated-measures (panel)

Grundvallar atriði þegar verið er að vinna með mælingar sem eru ekki óháðar að byggja inn í líkanið hvernig þær eru háðar. Það er t.d. vel hugsanlegt að breytileiki milli einstaklinga yfirgnæfi algerlega.

- Meðal við dvergvexti
- Hugsanlegt að öll fyrirtæki í landinu mismuni konum í hag í launum en samt virðist summan mismuna körlum í hag.
- Ein tegund lausna byggir á að meta GLMM líkan.

- Nemendur eiga að ráða við að taka gögn inn í forrit, fá útkomu og túlka (GLM og GLMM).

- Nemendur eiga að ráða við að taka gögn inn í forrit, fá útkomu og túlka (GLM og GLMM).

  Einnig þarf að hafa hugmyndir um hvernig skuli framkvæma einhvers konar diagnostics á útkomum. (Ekki alveg ákveðin áætlun hér)

- Nemendur eiga að ráða við að taka gögn inn í forrit, fá útkomu og túlka (GLM og GLMM).

  Einnig þarf að hafa hugmyndir um hvernig skuli framkvæma einhvers konar diagnostics á útkomum. (Ekki alveg ákveðin áætlun hér)

- Almennt þarf að kunna skil á grundvallar atriðum við líkanasmíði.

- Nemendur eiga að ráða við að taka gögn inn í forrit, fá útkomu og túlka (GLM og GLMM).

  Einnig þarf að hafa hugmyndir um hvernig skuli framkvæma einhvers konar diagnostics á útkomum. (Ekki alveg ákveðin áætlun hér)

- Almennt þarf að kunna skil á grundvallar atriðum við líkanasmíði.

  Þekkja hugtökin, likelihood-fall,

- Nemendur eiga að ráða við að taka gögn inn í forrit, fá útkomu og túlka (GLM og GLMM).

  Einnig þarf að hafa hugmyndir um hvernig skuli framkvæma einhvers konar diagnostics á útkomum. (Ekki alveg ákveðin áætlun hér)

- Almennt þarf að kunna skil á grundvallar atriðum við líkanasmíði.

  Þekkja hugtökin, likelihood-fall,

  information, standard-error of estimate o.s.frv. Þessi atriði eru tekin fyrir í ýmsum bókum.

- Nemendur eiga að ráða við að taka gögn inn í forrit, fá útkomu og túlka (GLM og GLMM).

  Einnig þarf að hafa hugmyndir um hvernig skuli framkvæma einhvers konar diagnostics á útkomum. (Ekki alveg ákveðin áætlun hér)

- Almennt þarf að kunna skil á grundvallar atriðum við líkanasmíði.

  Þekkja hugtökin, likelihood-fall,

  information, standard-error of estimate o.s.frv. Þessi atriði eru tekin fyrir í ýmsum bókum.

  Í Wooldridge er þetta rifjað upp í köflum 2-10 og 12-14. Það þarf ekki að hafa öll smáatriði úr þeim köflum á hraðbergi. Kunnátta úr öðrum skyldum

námsskeiðum ætti að duga.

Hvað eiga nemendur að gera?

- Lesa efnið

Hvað eiga nemendur að gera?

- Lesa efnið
- Skila ritgerðum, endursögnum á Berry og Spiegelhalter

Hvað eiga nemendur að gera?

- Lesa efnið
- Skila ritgerðum, endursögnum á Berry og Spiegelhalter
- Helmingur heldur fyrirlestur um Deb and Holmes og helmingur heldur fyrirlestur um Grootendorst. Með fyrirlestur er átt við glærsusýningu og munnlegan fyrirlestur ca. 30 mínútur þar sem gert er grein fyrir efni greinanna

Hvað eiga nemendur að gera?

- Lesa efnið
- Skila ritgerðum, endursögnum á Berry og Spiegelhalter
- Helmingur heldur fyrirlestur um Deb and Holmes og helmingur heldur fyrirlestur um Grootendorst. Með fyrirlestur er átt við glærusýningu og munnlegan fyrirlestur ca. 30 mínútur þar sem gert er grein fyrir efni greinanna
- Sérhver þátttakandi velur grein, t.d. úr Health Economics og gerir ritgerð/endursögn og heldur fyrirlestur/glærusýningu

Hvað eiga nemendur að gera?

- Lesa efnið
- Skila ritgerðum, endursögnum á Berry og Spiegelhalter
- Helmingur heldur fyrirlestur um Deb and Holmes og helmingur heldur fyrirlestur um Grootendorst. Með fyrirlestur er átt við glærusýningu og munnlegan fyrirlestur ca. 30 mínútur þar sem gert er grein fyrir efni greinanna
- Sérhver þátttakandi velur grein, t.d. úr Health Economics og gerir ritgerð/endursögn og heldur fyrirlestur/glærusýningu

- Reikna létt dæmi úr Wooldridge. Reyna að fá sömu útkomur og í sýnidæmum í bók.
- Mæta í próf og reikna nokkur dæmi og skýra nokkur hugtök með eigin orðum.

Nokkur áhættuhugtök

Hér verður skoðuð hending Y sem getur tekið gildin 0 eða 1 og við ætlum að bera saman hættuna á að

$\boldsymbol{Y} = 1$ milli tveggja hópa.

$$p_1 = P(\boldsymbol{Y} = 1|\text{Hópur 1}) =$$

líkur á að einstaklingur úr hóp 1 fái $\boldsymbol{Y} = 1$

$$p_2 = P(\boldsymbol{Y} = 1|\text{Hópur 2}) =$$

líkur á að einstaklingur úr hóp 2 fái $\boldsymbol{Y} = 1$

$$\text{RR} = \frac{p_1}{p_2}$$

$$\text{OR} = \frac{\frac{p_1}{(1-p_1)}}{\frac{p_2}{(1-p_2`)}}$$

$$AR = p_1 - p_2$$

RR=relative-risk, OR=Odds-ratio,
AR=Attributable-risk. Athugið að

- Ef p_1 og p_2 eru lítil þá er RR $\approx$ OR.

- Ef OR$<$1 þá er OR $\leq$ RR $\leq$ 1
- Ef OR$>$1 þá er OR $\geq$ RR $\geq$ 1
- Ef OR$=$1 þá RR$=$1

Nokkur grunnhugtök í líftímagreiningu

Látum $\boldsymbol{T}$ tákna hendingu sem mælir líftíma. Dreififall hendingarinnar er:

$$F(t) = P(\boldsymbol{T} < t) \qquad \text{og þéttifallið er}$$

$$f(t) = F'(t)$$

Þá er skilgreind „hazard¨-fall líftímans með:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

$\lambda(t)$ mælir þá hættu sem einstaklingur á aldri t er í. $S(t) = 1 - F(t)$ er kallað „survivor"-fall (T). Athugið að:

$$F(t) = 1 - e^{\int_0^t \lambda(s)ds}$$

Sjá nánar í kafla 20 í Woolridge.

Spurningar

1. Menn spila rússneska rúllettu. Hægt er að velja gula byssu með 1 skoti og 5 púðurskotum eða rauða byssu með tveim skotum og 4 púðurskotum. Finnið
   - RR rauðrar byssu miðað við gula?
   - OR rauðrar byssu miðað við gula?
   - Ef að 2/3 spilar með gulri byssu og 1/3 með rauðri, hve mikill hluti mannfallsins er vegna þess að byssurnar eru mismunandi?
2. Hve mikið myndi lungnakrabbameinstilfellum fækka ef reykingar leggðust af?

3. Ef rannsóknir sýnir að munur á þyngd tveggja bæja er mjög marktækur, hvernig á þá að bregðast við?

4. Hermið líftímadreifingar sem byggja á a) expoential dreifingu, b) Weibull dreifingu c) Pareto dreifingu. Detta ykkur fleiri dreifingar í hug?

5. Á að mótefnamæla alla Íslendinga vegna AIDS?

Survival teoría, ch 20. í W

Duration data, transition data, single spell=survival data.

Látum $\boldsymbol{T}$ tákna hendingu sem mælir líftíma. Dreififall hendingarinnar er:

$$F(t) = P(\boldsymbol{T} < t) \qquad \text{og þéttifallið er}$$

$$f(t) = F'(t)$$

Þá er skilgreind „hazard¨-fall líftímans með:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

$\lambda(t)$ mælir þá hættu sem einstaklingur á aldri t er í.

$S(t) = 1 - F(t)$ er kallað „survivor"-fall (T). Athugið að:

$$F(t) = 1 - e^{\int_0^t \lambda(s)ds}$$

Nokkur einföld parametrísk líkön

- $\boldsymbol{T}$ er exponential, $E(\boldsymbol{T}) = \lambda$.
- $\boldsymbol{T}^\alpha$ er exponential, γ, þ.e. $\boldsymbol{T}$ er Weibull með parametra α, γ.

$$f(t) = \alpha\gamma t^{\alpha-1} exp(-\gamma t^\alpha)$$

- Hazardfall fyrir exponential:

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda exp(-\lambda t)}{1 - (1 - exp(-\lambda t))} = \lambda$$

- Fyrir Weibull dreifingu er hazardfallið:

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{\gamma \alpha t^{\alpha - 1} exp(-\gamma t^{\alpha})}{1 - (1 - exp(-\gamma t^{\alpha}))} = \gamma \alpha t^{\alpha - 1}$$

- Einnig má skilgreina líkanið með því að skilgreina hazardfallið beint, t.d. log-logisic:

$$\lambda(t) = \frac{\alpha t^{\alpha-1}}{1+\gamma t^{\alpha}}$$

o.s.frv. Það er hægt að velja hvaða pósítívt fall sem er. Athugið að ef $\int_0^\infty \lambda(s)ds < \infty$ þá er maður að segja að það séu jákvæðar líkur á eilífu lífi. Það sem stendur í kennslubókum er oft þar vegna þess að það er stærðfræðilega þægilegt.

Að taka inn skýribreytur

- Skilgreini regression-jöfnu, læt líftímann (eða fall af honum) vera háða stærð, beiti einhvers konar línulegum/ólínulegum aðferðum, allt eftir því hvaða líkindadreifingu ég nota mér. Þ.e. beiti einhvers konar GLM (generalized-linear-model) aðferðum.
- Set skýribreytur inn í hazardfallið, reikna likelihood-fall og beiti einhvers konar maximum-likelihood aðferðum með því að hámarka likelihood fallið númeríska.

- Proportional hazard, þ.e. geng út frá því að það megi þátta megi hazardfallið

$$\lambda(\boldsymbol{x}, t) = \kappa(\boldsymbol{x})\lambda_0(t)$$

  þ.e. að það sé einhver grunn áhætta og svo sé henni hliðrað með einhverju falli af skýribreytunum.

- Flokka þarf skýribreytur eftir því hvort þær eru fastar fyrir hvern einstakling eða hvort þær þróast í tíma. Ef þær þróast í tíma þarf strangt til tekið að hafa allar upplýsingar um feril hverrar breytu í tíma.

- Líkön má smíða fyrir samfelldan tíma eða discrete.

Í hagnýtum tilfellum má nálga sanna hazardfallið smooth falli eða tröppufalli.

Sampling ferli

- Fyrir fólk, nánast útilokað að fá random úrtak.
- Ýmis konar selection kerfi ráða því hvaða gögn við sitjum uppi með. Við mat á líkani þarf að taka tillit til þess.
- Get oftast ekki fylgt einstakling til enda, þ.e. **right censoring**.
- Byrjunin getur verið með ýmsum hætti, t.d. óþekkt **left censoring** (óvanalegt).
- Gögn geta t.d. samanstaðið af einungis heilbrigðum einstaklingum í byrjun og síðan er beðið, **flow sampling**.

- **Stock sampling** gengur út á að velja úrtak úr þeim sem eru þegar í einhverju ástandi á gefnum tímapunkti. Þá er hafa glatast upplýsingar um þá sem dóu snemma út úr eldri kynslóðum. (**left truncated**)
- **Unobserved heterogeneity**. Hugsanlegt er að einstaklingar séu í mismikilli hættu þá að ekki sé mæld nein ákveðin skýristærð, t.d. erfða og umhverfisþættir.

Önnur atriði

- Competing risks
- Cox-regression

Smá tækniskammtur

Hvernig á að meta parametra þegar right-censoring er til staðar? Reikna log-likelihood:

$$l_i = d_i log(f_i) + (1 - d_i) log(1 - F_i)$$

$d_i = 1$ ef einstaklingur númer i er mældur, þ.e. ekki censored. hámarka síðan $\sum_{i=1}^{n} l_i$ með tilliti til óþekktra parametra. Lesið sjálf um piecewise constant hazard.

Binary data, logistic regression

Oft auðvelt í framkvæmd og túlkun. Hentar vel til við rannsóknir þegar þarf að meta RR ýmissa þátta vegna sjaldgæfra sjúkdóma. Breytan $Y = 1$ ef viðkomandi er með tiltekinn sjúkdóm $Y = 0$ annars. Til reiðu eru skýribreytur $\boldsymbol{x}$. Gerum ráð fyrir að líkur á að $Y = 1$ séu á forminu:

$$p = P(Y = 1|\boldsymbol{x}) = \frac{exp(\boldsymbol{x\beta})}{1 + exp(\boldsymbol{x\beta})}$$

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

Binary data, logistic regression

$p/(1-p) = Oddsratio$. Einn af kostum þessa forms eru túlkunarmöguleikarnir, þ.e. að fyrir tiltekna hnit, j, í $\boldsymbol{x}$ vektornum, þá er OR per einingu í $x_j = exp(\beta_j)$

Nokkrar léttar æfingar

1. Skrifið niður likelihood-fall fyrir parameterinn λ í poisson líkani miðað við random úrtak.
2. Hvernig breytist likelihood-fallið ef $\lambda = \boldsymbol{X}\boldsymbol{\beta}$?
3. Metið LPM(linear-probabiltiy-model), logistic-líkan og probit líkan fyrir MROZ.RAW gögnin.

4. Takið burt tvær breytur og framkvæmið LR-próf á kenningunni að þær skipti ekki máli.

5. Metið Poisson-regression líkan fyrir FERTIL2.RAW gögnin.

6. Takið burt tvær breytur og framkvæmið LR-próf á kenningunni að þær skipti ekki máli.

7. Af hverju eru quasi-Poisson aðferðir notaðar?

Grein um áhrif geðlyfja, frásögn

Til að átta sig vel á þessu þarf að skilja grundvallaratriði tímaraðagreiningar, sbr. þætti í tölfræði fyrir jól og tíðni regression, sbr. kafla 19. Lýst er líkani fyrir tíðna sjálfsvíga.

- Er Poisson dreifing viðeigandi?
- Er tímastrúktur?
- Residual diagnostics, eru horfur á því að mikilvægar breytur vanti?
- Hvernig má bæta líkanið?
- Hvernig gagna ætti að afla frekar?

Dæmi w20-1.

	Value	Std. Error	z	p
(Intercept)	4.09939	0.347535	11.796	4.11e-32
workprg	-0.06257	0.120037	-0.521	6.02e-01
priors	-0.13725	0.021459	-6.396	1.59e-10
tserved	-0.01933	0.002978	-6.491	8.51e-11
felon	0.44399	0.145087	3.060	2.21e-03
alcohol	-0.63491	0.144217	-4.402	1.07e-05
drugs	-0.29816	0.132736	-2.246	2.47e-02
black	-0.54272	0.117443	-4.621	3.82e-06
married	0.34068	0.139843	2.436	1.48e-02
educ	0.02292	0.025397	0.902	3.67e-01
age	0.00391	0.000606	6.450	1.12e-10
Log(scale)	0.59359	0.034412	17.249	1.13e-66

GLM: Hlutar úr köflum 15, 19,20.

- GLM=generalized-linear-model, fall af væntalegu gildi er línulegt í x-breytum.

$$g(E(Y_i|\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

g er kölluð link-function.

- Y getur verið ýmiskonarbreyta við höfum haft

Y	hugsanlegar dreifingar
biðtíma breyta	exponential, Weibull, gamma, log-normal, o.s.frv
0/1 (binary)	Bernoulli
count	Poisson, quasi-poisson, negatív-binomial o.s.frv.
samfelld	normal, t, o.s.frv.

- Höfum metið svona líkön og prófað kenningar um parametra.

- Fyrir survival-líkön er censoring áberandi atriði
- Fyrir counting-líkön er Poisson algeng viðmiðun, í praxís er overdispersion oft fyrir hendi
- Í logistískri regression er exp(parameter) odds-ratio per breytingu um eina einingu í x-breytu.
- Vantar að skipuleggja diagnostics, skoða leyfaliði (residuals),

Um Diagnostics

Upprifjun ??

- Hvað getur farið úrskeiðis í tölfræðilegri líkanagerð?
- Forsendur í normal línulegu líkani
- Misdreifni (heteroskedacity), þ.e. varíans hópa ekki jafn
- Mælingar ekki óháðar
- Mæliskekkjur í breytum, t.d. mistök i flokkun
- Breytur vantar

Hverjar eru afleiðingarnar?

- Tegundir afleiðinga: i) brestur í „consistency'', ii) skert nýtni á upplýsingum
- Misdreifni og sjálffylgni (autocorrelation) leiða (venjulega) til skertrar nýtni. Þ.e. t-gildi parametra gefa villandi mynd af nákvæmni. Slíkt má komast yfir með því að vera með stórt úrtak.
- Mæliskekkjur og að breytur vanti þýðir (venjulega) inconsistent parameter möt. Slíkt er ekki hægt laga með því að stækka úrtak
- Ranglega skilgreind dreifing, t.d. ranglega notuð normaldreifing þýðir venjulega að t-gildi parametra gefa villandi mynd af nákvæmni. T.d. ýkjukennda marktækni

Hugleiðingar um val á dreifingu fyrir tíðnigögn

- Hvaða dreifing af hverju ekki normal-ANCOVA?
- Normal ANCOVA þarf ekki að vera svo slæmt
- Poisson dreifing oft eðlileg fyrir tíðni
- Uppástunga: Nota poisson-regression og met parametra með maximum-likelihood
- Upprifjun: Munið eftir eiginleikum maximum-likelihood
- Ef sett er upp Poisson líkan fást consistent parametermöt (ef allar mikilvægar breytur eru settar í líkan á réttu formi)
- Til að réttar ályktanir (t-gildi) séu dregnar er nauðsynlegt að bæði væntalegt gildi og varíans sé rétt skilgreind. Þ.e. allar breytur þurfa að vera á réttu formi og varíans þarf að vera jafn væntanlegu gildi. Annars fást ýkjukennd t-gildi (yfirleitt).
- Hægt að leiðrétta fyrir því að varíans sé ekki jafn meðaltali og bæta nýtni í Poisson regression þó að sanna dreifingin sé ekki Poisson, t.d. með eins konar quasi/pseudo-Poisson

Um treatment effect, kafli 18, nokkur hugtök

- ATE = average treatment effect og ATE_1 = average treatment effect for treated.
- Mikilvægar breytur eru (Y_0, Y_1, W) þar sem W ákvarðar meðferð, og Y_1 útkomu með einni meðferð $W = 1$ og Y_0 útkomu með $W = 0$, Vandinn er að fyrir hvern einstakling er einungis mælt annað hvort Y_0 eða Y_1.
- Auðvelt að meta treatment effect ef hægt er að randomísera. Í praxís er einhvers konar val-prócess í gangi sem ákvarðar hvor meðferðin er notuð.

Meira um treatment mat úr kafla 18.

- Vil meta $ATE = E(y_1 - y_0)$ eða $ATE_1 = E(y_1 - y_0|w = 1)$
- Breyturnar eru (y_0, y_1, w) en á hverjum einstakling mælum við annað hvort y_0 eða y_1.
- Ef hægt væri að randomísera væri hægt að finna consistent mat á treatment, en vandinn er að gera verður ráð fyrir self-selection.
- Ein lausn: Ef hægt er að finna $\boldsymbol{x}$ þannig að gefið $\boldsymbol{x}$ þá eru vektorinn (y_0, y_1) óháður w.
- Til vara að $E(y_1 - y_0|\boldsymbol{x}, w) = E(y_1 - y_0|\boldsymbol{x})$
- Ath. þá er $ATE_1(\boldsymbol{x}) = E(y_1 - y_0|\boldsymbol{x}, w = 1) = E(y_1 - y_0|\boldsymbol{x}) = ATE(\boldsymbol{x})$, það er miklu auðveldara að eiga við 1. moment heldur en skilyrðið óháður.
- Skilgreini: $r(\boldsymbol{x} = ATE(\boldsymbol{x})$
- Metum ATE með regression

$$E(y|\boldsymbol{x}, w) = E(y_0|\boldsymbol{x}) + w(E(y_0|\boldsymbol{x}) - E(y_0|\boldsymbol{x}))$$

- Náum í úrtak af $(y, w, \boldsymbol{x})$ reiknum

$$r_1(\boldsymbol{x}) = E(y|\boldsymbol{x}, w = 1)$$
$$r_0(\boldsymbol{x}) = E(y|\boldsymbol{x}, w = 0)$$

- $\boldsymbol{x}$-gildi einstaklings i er $\boldsymbol{x}_i$.
- reiknum $\hat{r}_1(\boldsymbol{x}_i)$ og $\hat{r}_0(\boldsymbol{x}_i)$ og fáum

$$\hat{ATE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{r}_1(\boldsymbol{x}_i) - r_0(\boldsymbol{x}_i))$$

$$\hat{ATE}_1 = \frac{1}{\sum_{i=1}^{N} w_i}\sum_{i=1}^{N} w_i(\hat{r}_1(\boldsymbol{x}_i) - r_0(\boldsymbol{x}_i))$$

- Skoðið proposition 18.1. Þar er búin til control-function sem er hugsuð sem leiðrétting fyrir selection processnum.
- Hér er stundum gengið út frá þeirri forsendu að einstaklingar með sömu $\boldsymbol{x}_i$-gildi bregðist við með jafnstóru fráviki frá meðaltali. Mætti breyta með því t,d, að taka inn í líkan interaction parametra við treatment.
- Takið eftir hugmyndinni að leiðrétta með propensity score. Þ.e. regressa y_i á $1, w, \hat{p}(\boldsymbol{x}_i)$, þar sem $\hat{p}(\boldsymbol{x}_i)$ eru líkur á að einstaklingur með $\boldsymbol{x}$-gildi ($\boldsymbol{x}_i$ fái meðferð. Proposition 18.4 gefur skilyrði fyrir consistent mati.
- Einnig hugsanlegt að regressa á $1, w, \hat{p}, w\hat{p}(1-\hat{p})$.
- Matching aðferðir, þ.e. einstaklingar með lík $\boldsymbol{x}$-gildi matchaðir.
- Lauslega kafla 18.4 um instrumentalaðferðir.

Loglínulegt dæmi

- Farið í tjónagögn frá Lloyds tryggingafélaginu. Áhætta per mánaðarnotkun metin. Áhættuþættir kvarðaðir.

Um diagnostics, að hverju er verið að leita?

- Outliers, vitlaus mæling eða skrýtinn einstaklingur?
- High-leverage, óvenjulega blanda x-gilda
- Áhrifamikil mæling, að taka hana burt hefur mikil áhrifa á mat
- Multi-collinearity
- Hverju líkist dreifing afgangsliða

Tæknitól við diagnostics

- Gröf
- hat-fylki
- dfbeta o.s.frv.
- vif

Tölvuæfing

1. Búið til gögn úr normal-dreifingu og gerið pp-plot og qq-plot á móti normaldreifingu. Gerið töflum með descriptive statistics.

2. Búið til gögn úr t-dreifingu með 3 frígráður og gerið pp-plot og qq-plot á móti normaldreifingu. Gerið töflu með descriptive statistics.

3. Blað og blýantsdæmi

Um panel data, stiklað í kafla 10 og 11

- Panel data=longitudinal data=repeated measures
- Endurteknar mælingar á sama einstakling
- Einfalt dæmi

Fyrir	Eftir
120	121
130	131
140	142
150	151
160	161

Augljóst að hér skiptir máli hvort þetta er sama fólkið eða tvö óháð úrtök. Gott að geta borið einstakling saman við fyrri gildi af sjálfum sér.

- Vil álykta um áhrifa x-breytu en leiðrétta fyrir einstaklings áhrifum c.
- Geng út frá líkani

$$y_t = \beta_0 + \boldsymbol{x}_t'\boldsymbol{\beta} + c + u_t$$

- Safna T mælingum, $t = 1, \ldots, T$, á N einstaklingum, hef þá **balanced panel**. (Oft ekki

staðan í praxís).

- Mælilíkanið er þá

$$y_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + u_{it}$$

- Hvernig á að túlka c_i? Fixed effect versus random effect. Athugið að econometría notar skrýtna flokkun hér, þ.e. að fixed effect eru ekki fixed heldur random á ákveðinn hátt.

- Nánari skýring, random-effect er í microeconometríu látið þýða $Cov(x_{it}, c_i) = 0)$ og e.t.v. einnig $E(c_i|\boldsymbol{x}_{it}) = E(c_i)$. Fixed effect er látið þýða að $Cov(x_{it}, c_i) \neq 0$.

- Þetta er afar ólánleg orðanotkun (sem gerir að flestir skilja alls ekki hvað er á seyði). Höfundurinn er meðvitaður um þetta reynir að forðast þessa notkun og kalla þetta frekar ómældan einstaklingsþátt eða eitthvað þess háttar.

- Höfundur heldur sig við að kalla matsaðferðir „fixed-effects'' eða „random-effect'' og afsakar sig með að þessar nafngiftir séu svo inngrónar að það þýði ekki að andæfa þeim.
- Exogeneity of observed effect er atriði, leyst með instrumentum.

Random effects aðferð

- Forsendur

$$E(u_{it}|\boldsymbol{x}_i, c_i) = 0 \quad \boldsymbol{x}_i = (x_{i1}, \ldots, x_{iT})'$$
$$E(u_{it}|c_i) = 0$$
$$E(u_{it}^2) = \sigma_u^2$$
$$V(c_i|\boldsymbol{x}_i) = \sigma_c^2$$

- Get skrifað líkan á forminu:

$$\boldsymbol{y}_i = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{v}_i$$

þar sem $\boldsymbol{v}_i$ er samsettur stókastískurliður, vegna skella og vegna breytileika einstaklinga.

$$V(\boldsymbol{v}_i) = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & \cdots & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

- Venjuleg least-squares aðferðafræði gefur

$$\hat{\boldsymbol{\beta}}_{RE} = (\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{y}_i)$$

- Að ýmsu að hyggja, misdreifni, sjálffylgni o.s.frv. Einnig getur verið áhugavert að álykta um σ_c.

- Fixed effect estimation, framkvæmd þegar grunur leikur á tengslum x-breyta og einstaklingsþáttarins. Forsenda:

$$E(u_{it}|\boldsymbol{x}_i, c_i) = 0$$

- Reynum að meta líkan á formi þar sem c_i hefur verið fjarlægt. T.d. með því að draga frá meðaltal.
- Skiljið muninn á **within** og **between** (berið saman 10.45 og 10.50).
- First difference aðferð. Önnur leið til að ná burt c_i. x-breytur verða að breytast í tíma til að þetta sé gagnlegt.
- Ef aðeins eru fyrir hendi tvö tímabil er þetta venjulegur paraður samanburður.

Nokkur atriði um kafla 11

- Byrjar á því að slaka á kröfunni um slökun á kröfunni um exogenitet. Sequentially conditioned moments er svipað og predetermined hugtakið í klassískri económetríu.
- Random slopes er áhugavert. Þið ættuð að skilja hugtakið þó að þið kafið ekki í formúlurnar í kaflanum.