Punktar úr fyrirlestrum í hagrannsóknum I haustið 2012

Helgi Tómasson
Hagfræðideild HÍ
helgito@hi.is

Forkröfur

- Econometría = hagnýt tölfræðigrein með áherslu á hvernig ber að álykta út frá hagmælinum.
- Forkröfum = grunnþekking á líkindafræði og tölfræði. T.d. námskeiðið líkeindareikningur og tölfræði. Einnig er mjög gagnlegt að vera vel að sér í línulegri algebru.
- Nokkur hugtök úr líkindafræði (probability). Random variable=slembibreyta, dreifing, væntanlegt gildi, einvíðar/margvíðar slembibreytur, skilyrtar líkur, óháðar slembibreytur o.s.frv.
- Nokkur atriði úr tölfræði (ályktunarfræði, inference). Mat og metill (estimate og estimator), prófanir (testing), núllkenning, H_0, valkostur H_1, villa I, villa II, power. Hvað þýðir marktækur (significant)?

Hvað þýðir regression?

- Tvenns konar svar. a) Líkindafræðilegt, eins konar skilyrt væntanlegt gildi $E(Y|X=x)$ og $V(Y|X=x)$, og b) rúmfræðilegt=minnsta fjarlægð á milli Y og X.
- Í þessu námskeiði er lögð áhersla á línulega regression, þ.e. að $E(Y|X=x)=\beta_1+\beta_2 x$. Þetta er línulegt því að þetta fall er línulegt í β_1 og β_2.
- Með línulegri regression er verið að meta línuleg sambönd. Fylgnistuðlar er náskyldir regression-stuðlum. Ef slembibreytur eru óháðar þá er fylgni þeirra núll. Hugsanlegt er að fylgni sé 0 þó að slembibreytur

séu háðar, sbr. dæmi 6 og 7 á dæmablaði 2. Ef slembibreytur eru normaldreifðar og fylgni þeirra 0 þá eru þær óháðar.

Setning FWL

- Samanburður á meðaltölum (ANOVA) sem margir hafa lært í fyrri námskeiðum (t.d. lík. og töl.) er regression með 0/1 skýristærðum. Marktækni þýðir ca. sama og t-gildi stærra en 2 (mjög varasamt hugtak).
- Ef mikilvægum breytum er sleppt fást bjagaðar hugmyndir um þýðingu skýristærða, sbr. dæmablað 3.
- Á dæmablaði 3 var sýnidæmi um kyn og laun. Kynbreytan er að hluta vinnutímabreyta því einstaklingar af öðru kyninu unnu meira en einstaklingar af hinu kyninu.
- Það má leiðrétta fyrir þessum með því að bæta vinnutíma og fleiri mikilvægum breytum í líkanið og fá mat á þýðingu kynbreytunnar. Í þessu dæmi voru gögnin búin þannig til að þegar rétt líkan var metið fékkst 100% skýrigeta og konur fengu ákveðið umfram karla þó svo annað virðist í einfaldari líkönum.
- Í stað þess að bæta breytum í líkanið má margfalda í gegnum um jöfnuna með fylki sem hreinsar vinnutíma og fleiri atriði út úr kynbreytunni og þannig fæst sama gildi og áður.
- Bjagaða myndin kemur fram vegna þess að breytan sem sleppt er tengist áhugaverðu breytunni. Ef hægt er að brjóta þau tengsl má fá óbjagað mat.
- Með fylkja-algebru er hægt að hafa góða stjórn á útreikningum.

Setning GM

- $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{u}$

  1. Form líkans ($E(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x})$, rétt allar mikilvægar breytur með.
  2. $V(\boldsymbol{u}) = \sigma^2 I$.
  3. $\boldsymbol{X}$ non-random.
  4. $\boldsymbol{X}$-fylkið full-rank.

  Þá gildir að OLS-metill er BLUE. Sönnun auðveld með fylkjaalgebru.

Hvað þýða brestir í forsendum?

- Brestur í 1 þýðir nánast alltaf að mat bjagast. Undantekning er þegar hægt er að „orthogonalisera", t.d. í tilraunavísindum með því að negla niður x-gildin. T.d. að hafa vinnutímadreifingu nákvæmlega eins í öllum hópum. Ef breytu vantar þarf að bæta henni við eða margfalda í gegnum kerfið með breytu sem er orthogonal á hana, sbr. FWL.

- Brestur í 2 hefur í för með sér minnkandi nýtni. Þ.e. mat verður ekki eins nákvæmt og það gæti verið. Rætt er um tvenns konar bresti í forsendu 2.

  (a) $V(\boldsymbol{u})$ hornalínufylki en ekki sama talan á allri hornalínunni. Þetta er kallað heteroskedacity (misdreifni).

  (b) $V(\boldsymbol{u})$ ekki hornalínufylki. Algengt form hér er að stökin sem eru nálægt hornalínunni séu stærri en hin. Þegar mælingar eru í tímaröð er talað um auto-correlation (sjálffylgni). Annað form er ef $V(\boldsymbol{u})$ er samsett úr blokkum, t.d. sami einstaklingur mældur oft. Hér eru til ýmsar leiðréttingaraðferðir. Það þarf að skilja hvernig GLS metillinn verður til.

- Brestir í 3, þ.e. $\boldsymbol{X}$ er random stærð. Þetta er algeng staða í hagrannsóknum og öðrum „observational"tölfræðigreinum. X-gildunum er ekki stjórnað af vísindamanninum heldur rignir þeim yfir hann. Hér eru ekki til einföld svör fyrir endanlegan mælingafjölda og því nauðsynlegt að treysta á „asymptotic"niðurstöður. Líkindafræðilegir eiginleikar $\boldsymbol{X}$ og tengslin við $\boldsymbol{u}$ skipta máli hér. „Unconditionalęiginleikar OLS-metilsins verða flóknir.

- Brestur í 4 þýðir að andhverfa $\boldsymbol{X'X}$ er ekki til og því ekki hægt að aðgreina mikilvægi einstakra X-breyta. Ef forsenda 4 er nálægt því að bresta þýðir það að erfitt er að aðgreina þýðingu X-breyta.

Ýmsar mikilvægar líkindadreifingar

Það þarf að þekkta til og vita hvernig ýmsar líkindadreifingar tengjast. Svo sem margvíð normaldreifing, χ^2, t- og F-dreifingar. Það þarf að vita hvernig dreifing línulegar falla og kvaðratískra forma af margvíðu normaldreifingunni er. Það þarf að átta sig á því hvernig idempotent fylki koma hér við sögu.

Prófstærðir fyrir ýmsar kenningar.

Ef forsendur GM gilda og $\boldsymbol{u}$ er normaldreifður þá er dreifing OLS-metilsins nákvæmlega þekkt. Hér er nauðsynlegt að geta reiknað prófstærðir fyrir línuleg hliðarskilyrði á $\boldsymbol{\beta}$. Einnig er nauðsynlegt að átta sig á því hvernig eigi að álykta um σ.

Asymptotic teóría

- Fyrir mörg vandamál eru bara til niðurstöður fyrir „large-sample". Nokkur hugtök eru nauðsynleg (varist frasann *jafnar þetta sig ekki út*)
- Nauðsynlegt er að þekkja skil á nokkrum „convergence-hugtökum".
  (a) a.s.=almost-surely
  (b) m.s.=mean-square
  (c) plim=convergence-in-probability
  (d) convergence-in-distribution
- CLT=central-limit-theorem (margar útgáfur til): Segir hvenær dreifing summu stefnir í dreifingu á normal-dreifingu.

Likelihood-teóría

- Nokkrar aðferðir við tölfræðilegt mat
  (a) Rúmfræðilegar, LS=least-squares, LAD=least-absolute-deviation
  (b) MM=method-of-moments. Ýmsar útfærslur GMM=generalized-method-of-moment tekið í hagrannsóknum II.
  (c) Maximum-likelihood
  (d) Bayes aðferðir

Likelihood-fall=þéttifall mælivektors sem fall af óþekktum parameter.

$$L(\boldsymbol{\theta}|y_1, \ldots, y_n) = f(y_1, \ldots, y_n; \boldsymbol{\theta})$$

- Oft þægilegra að vinna með $l(\theta|y_1,\ldots,y_n) = \log(L(\boldsymbol{\theta}|y_1,\ldots,y_n))$. Maximum-likelhood (ML) mat á $\boldsymbol{\theta}$ er það gildi á $\boldsymbol{\theta}$ sem hámarkar L eða l.

- ML-aðferðin er stundum þýdd sem aðferð mesta sennileika (til að hafa orð sem hljómar ólíkt orðinu „líkur). Orðið „probability"hljómar öðruvísi en likelihood.

- ML-matið, $\hat{\boldsymbol{\theta}}$, er alls ekki líklegasta gildi á $\boldsymbol{\theta}$, heldur það gildi sem hefur mælingarnar $(y_1,\ldots,y_n)$ sem líklegasta gildi.

- ML-aðferð er í einhverjum skilningi „best". Ef ákveðin regularity skilyrði gilda þá:

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} N(\boldsymbol{\theta}, I_{\boldsymbol{\theta}}^{-1}),$$

$$I_{\boldsymbol{\theta}} = -E\left(\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = \text{Fisher Information fylki.}$$

- Cramer-Rao ójafnan segir að varíans óbjagaðs metils hljóti að vera meiri en $I_{\boldsymbol{\theta}}^{-1}$. Nákvæmnin í matinu (öryggismörk) ræðst af upplýsingafylkinu.

- Gott að átt sig á delta-method, sem gefur nálgun á dreifingu $g(\hat{\boldsymbol{\theta}}_{ML})$.

- LR=likelihood-ratio próf gengur út á að reikna hámarksgildi sennileikafalls gefið H_0 rétt og bera saman við hámarksgildi sennileikafallsins gefir H_1 rétt. Ath. $H_0 \subset H_1$. $-2\log(LR) \xrightarrow{d} \chi^2_k$, þar sem k sem fjöldi hliðarskilyrða sem H_0 setur á H_1.

- Önnur asymptótísk próf í sama anda eru Wald-próf og LM próf.

Um instrumental aðferðir

- Í haglíkönum er oft eðlilegt að gera ráð fyrir ómælanlegum bretum.

- Það gilda sömu lögmál þegar þeim er sleppt og öðrum mikilvægum breytum.

- Instrumental breytur geta stundum verið gagnlegar í slíkum tilfellum. Dæmi um gagnsemi slíkra breyta er í simultaneous kerfislíkönum og þegar x-breyturnar eru mældar með mæliskekkju.

- Það þarf a.m.k. eina instrumentalbreytu fyrir hverja X-breytu. Ef X-breytur eru „exogen", eða mældar án mæliskekkju getur breytan sjálf verið instrúmentið.

- Hugmyndin er að í líkaninu:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

Hef fylki af instrumentbreytum, $\boldsymbol{Z}$. Þær þurfa að vera fleiri en X-breyturnar (order condition), vera óháðar (ókorreleraðar) ómældu breytunum (mæliskekkjunni) og nægilega háðar sönnu X-breytunum (rank condition) og nægilega óháðar u-breytunum. Nægilegt hér þýðir

$$\text{plim}(\frac{\boldsymbol{Z'X}}{n}) = Q_{ZX} < \infty,$$
$$\text{plim}(\frac{\boldsymbol{Z'u}}{n}) = \boldsymbol{0},$$
$$\text{plim}(\frac{\boldsymbol{Z'Z}}{n}) = Q_{ZZ} < \infty.$$

- Spái $\boldsymbol{X}$ með $\hat{\boldsymbol{X}}$=regression á $\boldsymbol{Z}$ og met $\boldsymbol{Y} = \hat{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{v}$. Kalla niðurstöðuna b_{IV}. Sá metill er consistent mat á β.

- Athugið að σ^2 verður að meta með:

$$s^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}_{IV})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}_{IV})}{n}, \quad \text{ekki } \hat{\boldsymbol{X}}.$$

- Forritið sem bjó til skónúmeradæmið býður upp á að prófa eiginleika instrumentaðferðarinnar. Hér er forritið:

```
#  Hef n mælingar
#
n=1000
#
#  Fylgni intstrumental breytu og áhugaverðrar breytu er rho
#
rho=0.9
#
#  Mæliskekkjan er +/- delta
#
delta=10

#
#  Choleski-þátta fylgnifylki
#

L=t(chol(matrix(c(1,rho,rho,1),ncol=2)))

#
#  Endurtek nsim sinnum
#
nsim=1000

#
#  Núllistilli útkomuvektora
#

biv=matrix(rep(0,2*nsim),ncol=2)

btrue=biv

sigmax=rep(0,nsim)

#
# Loop
```

```
#

for (ii in 1:nsim){
#
# Bý til n stykki af tvívíðum normal vectorum
#
eps1=matrix(rnorm(2*n),ncol=2)%*%t(L)
#
#
#

haed=180+7*eps1[,1]
skonumer=43+1.6*eps1[,2]
#
#
#
maeliskekkja=-delta+2*delta*rbinom(100,1,0.5)

maeldhaed=haed+maeliskekkja

eps2=rnorm(n)

thyngd=-100+haed+10*eps2

#
#  Bý til instrument
#
z=cbind(rep(1,n),skonumer)
#
#  Mælt x
#
x=cbind(rep(1,n),maeldhaed)
#
#  Reikna instrumental-vektor
#
bivx=solve(t(z)%*%x)%*%t(z)%*%thyngd
#
#  Met líkan án mæliskekkju
#
btruex=lm(thyngd~haed)
```

```
#
# safna beta-hat án mæliskekkju
#
btrue[ii,]=btruex$coefficients
#
# safna sigma^2 reiknað með instrumental beta
#
sigmax[ii]=sum((thyngd-x%*%bivx)^2)/n
#
#  safna instrumental betum
#


biv[ii,]=bivx


}


apply(btrue,2,quantile,c(0.05,0.10,0.25,0.5,0.90,0.95))


apply(biv,2,quantile,c(0.05,0.10,0.25,0.5,0.90,0.95))


qzx=1/n*t(z)%*%x
qzz=1/n*t(z)%*%z
#
# Skoða fræðilega varíans biv.
#
mean(sigmax)/n*solve(qzx)%*%qzz%*%(solve(t(qzx)))
#
# Á að bera saman við
#
var(biv)
```

Um tímaraðir

- Mikið af haggögnum er á formi tímaraða.
- Nauðsynlegt að hafa sérstök tímaraðalíkön.
- Stundum er tímaröð sundurliðuð í trend, cycle, season og irregular.
- Það má segja að við séum með venjulega líkanið:

  $$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad E(\boldsymbol{u}) = \boldsymbol{0}, \quad V(\boldsymbol{u}) = \boldsymbol{\Omega},$$

  en að nú sé aðaláherslan á $\boldsymbol{\Omega}$.
- Þurfum að skilgreina nokkra núllpunkta. Í samfelldum tíma má hugsa sér minnislausan ferli sem er samfelldur, Wiener-ferli og feril sem fer allra sinna ferða í eins skrefs hoppum, Poisson-ferli. Í strjálum tíma (discrete-time), höfum við white-noise og random-walk.
- Stöðugleikahugtök nauðsynleg. Stationary (weak og strong) og ergodic.
- Auto-covariance, sjálffylgni (ACF) og partial sjálffylgni (PACF).
- AR, MA, ARMA.
- Non-stationary ARIMA.
- Spurious regression. Hver er villan? Athugið dæmi Yule, hermun á óháðum RW á dæmablaði og dæmi 14.6 í ETM.
- Unit-root próf, Dickey-Fuller, ADF.
- Co-integration og ECM. Aðferð Granger og Engle við að meta cointegrated kerfi.
- Exogeneity hugtök, pre-determined/strictly exogenous og weak/strong/super. Granger-causality.