Hugleiðing um námsframvinda/brottfall í Hagfræðideild HÍ

Menntavika 30. september 2011

Helgi Tómasson
Hagfræðideild HÍ

Skipulag erindis

- Bakgrunnur og saga
- Tölfræðileg líkön: Mikilvægt að taka tillit til allra þátta samtímis
- Einvíð líkön
- Margvíð líkön
- Ályktanir og framtíðarstefna

- Í okkar deild (eins og öðrum) er talað um brottfall, innritunartölur o.s.frv.
- Ákveðið að taka hlutlægt á málum og athuga hvað „gögnin segðu".
- Til að afmarka vandann var ákveðið að einbeita sér einum árgangi af nýinnrituðum BS nemum.
- Gagnavandi engu að síður mikill. Það er mikil vinna að staðla gögn.
- Byrjað á að líta á einföld líkön. Þar sem einstaklingar eru breytilegir er nauðsynlegt að nota líkön sem tillit til þess.

Gögn tala eingöngu í gegnum tölfræðilegt líkan

- Það á alltaf gera grein fyrir vali á líkani og hvernig það tengist viðfangsefninu.
- Það bera að varast að slá saman misleitum hópum og hundsa mikilvægar skýrirstærðir (þó svo þær snerti ekki viðfangsefnið beint). 2x2 töflur eru varasamar því þær samsvara regressonlíkani þar sem einungis er tekið tillit til einnar skýristærðar.
- Sýnd eru nokkur dæmi um tölfræði gildrur.
- Hér er farin sú leið, m.a. vegna rýrra gagna að beita bayesískri nálgun í mati á tölfræðilegum líkönum.
- Einvíð 0/1 líkön eru notuð og síðan margvíð 0/1 líkön.

Það má ekki slá ólíkum hópum saman

Það má ekki slá ólíkum hópum saman

	Há laun	Lág laun
Karlar	18	12
Konur	7	3

Tafla: Launadreifing í fyrirtæki A. 70% kvenna með há laun, 60% karla með há laun.

Það má ekki slá ólíkum hópum saman

	Há laun	Lág laun
Karlar	18	12
Konur	7	3

Tafla: Launadreifing í fyrirtæki A. 70% kvenna með há laun, 60% karla með há laun.

	Há laun	Lág laun
Karlar	2	8
Konur	9	21

Tafla: Launadreifing í fyrirtæki B. 30% kvenna með há laun, 20% karla með há laun.

Það má ekki slá ólíkum hópum saman

	Há laun	Lág laun
Karlar	18	12
Konur	7	3

Tafla: Launadreifing í fyrirtæki A. 70% kvenna með há laun, 60% karla með há laun.

	Há laun	Lág laun
Karlar	2	8
Konur	9	21

Tafla: Launadreifing í fyrirtæki B. 30% kvenna með há laun, 20% karla með há laun.

	Há laun	Lág laun
Karlar	20	20
Konur	16	24

Tafla: Launadreifing í fyrirtækjum A+B. 40% kvenna með há laun, 50% karla með há laun.

2x2 töflur villandi

	starf=1	starf=2
karlar	154.000	241.429
konur	126.667	200.000

Tafla: Laun eftir starfi og kyni

- Lærdómur:

2x2 töflur villandi

	starf=1	starf=2
karlar	154.000	241.429
konur	126.667	200.000

Tafla: Laun eftir starfi og kyni

	aldur=1	aldur=2
karlar	145.000	235.000
konur	122.500	190.000

Tafla: Laun eftir aldri og kyni

- Lærdómur:

2x2 töflur villandi

	starf=1	starf=2
karlar	154.000	241.429
konur	126.667	200.000

Tafla: Laun eftir starfi og kyni

	aldur=1	aldur=2
karlar	145.000	235.000
konur	122.500	190.000

Tafla: Laun eftir aldri og kyni

- Lærdómur:Þetta er hlut af kennsluefni sem ég hef notað í 20 ár um að ekki megi sleppa mikilvægum breytum

2x2 töflur villandi

	starf=1	starf=2
karlar	154.000	241.429
konur	126.667	200.000

Tafla: Laun eftir starfi og kyni

	aldur=1	aldur=2
karlar	145.000	235.000
konur	122.500	190.000

Tafla: Laun eftir aldri og kyni

- Lærdómur:Þetta er hlut af kennsluefni sem ég hef notað í 20 ár um að ekki megi sleppa mikilvægum breytum
- Gögnin eru búin til með líkani sem mismunar konum í hag, en allar 2x2 töflur gefa karla með hærri laun

Lýsing á gögnum

- Árið 2008 innrituðust 88 nýnemar í BS hagfræði. Aðrir nemar eru BA nemar og eldri BS og BA nemar, e.t.v. ca. 140 á fyrsta ári.
- BS-línan er bundin lína, en BA línan hefur miklu meira val.
- Fyrir lágu upplýsingar um úr hvaða framhaldsskóla nýnemar komu, hvaða einkunn þeir höfðu fengið og af hvaða braut (náttúrufræðibraut, o.s.frv.).
- Skoðuð var útkoma úr 5 fyrstu prófunum í BS-hagfræði.

	stl	töl	þjl	rel	hag	frh
stl	1.00	0.98	0.95	0.94	0.99	0.25
töl	0.98	1.00	0.95	0.95	0.98	0.18
þjl	0.95	0.95	1.00	0.95	0.94	0.11
rel	0.94	0.95	0.95	1.00	0.98	0.14
hag	0.99	0.98	0.94	0.98	1.00	0.21
frh	0.25	0.18	0.11	0.14	0.21	1.00

Tafla: Fylgni milli einkunna í 5 BS-hagfræðigreinum og einkunnar úr framhaldsskóla.

Hin mikla fylgni milli greina innbyrðist skýrist af því að þeir sem falla, falla í mörgum greinum. Tengsl við framhaldsskólaeinkunn virðast rýr.

- Hin há fylgni er vegna þess að þeir sem falla, falla í mörgum greinum. Af þeim (12) sem standast prófið lítur dæmið svona út.

	stl	töl	þjl	rel	hag	frh
stl	1.00	0.65	-0.14	0.43	0.35	0.15
töl	0.65	1.00	-0.25	0.18	0.16	0.26
þjl	-0.14	-0.25	1.00	0.68	0.16	0.01
rel	0.43	0.18	0.68	1.00	0.43	0.29
hag	0.35	0.16	0.16	0.43	1.00	0.49
frh	0.15	0.26	0.01	0.29	0.49	1.00

Tafla: Fylgni einkunna í fyrstu 5 greinum BS-náms í hagfræði og framhaldsskólaeinkunnar.

Einföld 0/1 líkön

- Skoðaðar tvær einvíðar útfærslur. Linear-probability (LP) líkan og probit líkan. $y_1 = 1$ þýðir viðkomandi stóðst próf 1. $\boldsymbol{X}$ er vektor af skýristærðum.

$$E(y_1) = P(y_1 = 1) = \boldsymbol{X}\beta \tag{1}$$

og probit,

$$E(y_1) = P(y_1 = 1) = \Phi(\boldsymbol{X}\beta) \tag{2}$$

Kostir þess að meta LP-líkanið, (1), er að túlkun β á þýðingu skýribreyta $\boldsymbol{X}$ er auðveldari. Galli við LP-líkanið er að við mat á því með venjulegum forritum er hugsanlegt að einstaklingi sé úthlutað neikvæðum líkum, eða líkum stærri en 1 í að fara í prófið. Það að meta líkanið þannig að $P(y_1 = 1)$ sé þvingað til að vera á bilinu [0,1] er tæknilega erfiðara. Í þessari greiningu var það hliðarskilyrði þvingað fyrir gefin gögn, en ekki er víst að þvingunin haldi fyrir öll möguleg gildi á $\boldsymbol{X}$. Kostir probit líkansins eru að $P(y_1 = 1)$ þvingast sjálfkrafa á bilið [0,1].

Tvö einvíð líkön

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.2048	1.9043	-3.78	0.0002
VÍ	0.7102	0.4193	1.69	0.0903
MR	1.5054	0.4951	3.04	0.0024
einkunn	0.8008	0.2404	3.33	0.0009

Tafla: Probit líkan fyrir stærðfræði I

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0000	0.00	1.0000
VÍ	0.0764	0.0994	0.77	0.4446
MR	0.2172	0.1270	1.71	0.0914
einkunn	0.0230	0.0083	2.77	0.0071

Tafla: LP líkan fyrir stærðfræði I

Lausleg skoðun á einvíðum líkönum gefur til kynna að hugsanlega vanti mikilvægar skýristærðir. Því gæti verið upplýsandi að meta margvítt líkan þar sem til dæmis prófþátttaka væri skýrð. Þar sem mælingar eru fáar þarf greinandi að setja viðeigandi skorður til að fá túlkanlegar útkomur. Hér var valin sú leið að meta margvítt probit líkan með bayesískum aðferðum til að skýra prófþátttöku.
Líkanið sem metið var á forminu:

$$\boldsymbol{y}_i^* = \boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N(0, \Sigma), \tag{3}$$

$$\boldsymbol{y}_i' = (y_{i1}, y_{i1}, y_{i3}, y_{i4}, y_{i5}),$$

$$y_{ij} = \begin{cases} 1 \text{ ef } y_{ij}^* > 0 \\ 0 \text{ annars} \end{cases}, \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \vdots & \ddots & & \vdots \\ \sigma_{1k} & \cdots & \cdots & \sigma_{kk} \end{bmatrix}.$$

Hér táknar $y_{ij} = 1$ að einstaklingur i hafi staðist próf j. Fylkið $\boldsymbol{X}_i$ táknar fylki af skýrbreytum fyrir einstakling i

Bayesísk mat á líkani af þessari gerð var metið fyrir 5-víða mælingu á prófárangri 88 nýnema í BS-hagfræði. Skýribreytan var einungis framhaldsskóli, (VÍ, MR annað). A priori dreifing stika var skilgreind:

$$\boldsymbol{B} \sim N(\mathbf{0}, A^{-1})$$
$$\Sigma^{-1} \sim \mathit{Wishart}(\nu, V)$$

Þar sem ákveðið hefur verið að meta fylgnifylki er eðlilegt að setja V=einingarfylki. Óvissan um Σ var ákveðin mikil, með því að velja $\nu = 3$. Fyrirfram nákvæmnin í $\boldsymbol{B}$, A var ákveðin lítil, t.d. einingarfylki margfaldað með 0.001.

Taflan sýnir metið fylgnifylki fyrir prófarangur, þegar leiðrétt hefur verið fyrir hugsanlegum mun á milli skóla. Ljóst virðist að prófgeta hvers einstaklings er að mestu einn þáttur. Í þáttagreiningu (factor analysis) hafa menn stungið upp á þeirri þumalfingursreglu að fjöldi þátta sé fjöldi eigingilda fylgnifylkisins sem eru stærri en 1. Í fylkinu í töflunni er stærsta eigingildið 4.46 og það næsta er 0.27. Einnig mætti orða það þannig að 90% eigingildasummunar séu í stærsta eigingildinu.

stl	töl	þjl	rel	hag
1.00	0.85	0.87	0.83	0.80
0.85	1.00	0.95	0.95	0.74
0.87	0.95	1.00	0.98	0.89
0.83	0.95	0.98	1.00	0.85
0.80	0.74	0.89	0.85	1.00

Tafla: Metið fylgnifylki prófgetu.

Þýðing skóla í margvíðu líkani.

Breyta	$\hat{\beta}$	s.e($\hat{\beta}$)	5%	50%	95%
(Intercept)	-0.79	0.22	-1.17	-0.78	-0.47
VÍ	0.13	0.26	-0.27	0.12	0.58
MR	0.31	0.28	-0.12	0.29	0.81

Tafla: Áhrif skóla á líkur á prófgetu.

Tæknilega er ekkert því til fyrirstöðu að fjölga breytum. Til dæmis mætti bæta við þetta líkan einkunn, og jafnvel samspili einkunnar og skóla.
Í töflunni glæru er sýnt mat á þýðingu skóla. Hér það hliðarskilyrði sett að skóli hefði sömu áhrif á prófgetu í öllum greinunum fimm. Samkvæmt þessari töflu er prófviljinn mestur há þeim sem hafa verið í MR. Nákvæmnin er eins og vænta mátti ekkert sérstök.

Sumir halda að einkunnakvarðinn sé ekki samanburðarhæfur milli skóla, hvork hvað varðar staðsetningu né skrefstærð. einkunn hafði verið bætti við. Frjálsleg túlkun töflunnar á næstu glæru væri t.d. að einstaklingur sem hvorki kemur úr VÍ né MR og hefur fengið 7 í einkunn hefði líkur $\Phi(-1.723 + 0.141 * 7) = 0.23$ á að fara í próf. Sams konar einstaklingur með 8 í einkunn hefði líkur $\Phi(-1.723 + 0.141 * 8) = 0.28$ á að fara í próf. Einnig sú (órökrétta) túlkun að einstaklingur úr VÍ með einkunn 8 hefði líkur $\Phi(-1.723 + 1.448 + 0.141 * 8 - 0.169 * 8) = 0.31$ á að fara í próf en sá sem væri með 7 í einkunn hefði líkur $\Phi(-1.723 + 1.448 + 0.141 * 7 - 0.169 * 7) = 0.32$ á að fara í próf.

Breyta	$\hat{\beta}$	s.e($\hat{\beta}$)	5%	50%	95%
(Intercept)	-1.723	0.693	-2.9400	-1.684	-0.669
VÍ	1.448	0.783	0.2626	1.398	2.797
MR	1.249	1.191	-0.6580	1.226	3.271
Einkunn	0.141	0.087	0.0095	0.137	0.291
Einkunn $\times$ VÍ	-0.169	0.107	-0.3525	-0.164	-0.005
Einkunn $\times$ MR	-0.035	0.171	-0.3110	-0.038	0.257

Tafla: Áhrif af samspili einkunnar og skóla á prófgetu.

Hver er staða hópsins 2011?

- Vorið 2011 kláruðu 3 af 88 BS-hagfræði.
- Veturinn 2011-2012 stefnir í að 4 til viðbótar klári.
- Enn eru 15 aðrir við námi í BS-hagfræði, ca. 5 eru búnir með lítið.
- 27 eru haustið 2011 við nám i HÍ í öðrum greinum.
- 17 eru hættir í HÍ.
- 22 fundust ekki.

Framhald

Á að safna frekari gögnum? Af fenginni reynslu er vitað að tæknileg vinna verður umfangsmikil. Það er ljóst að óhemju vinna liggur í því raða excel-skrám og öðru illa samræmdu efni saman. Það var mikil vinna fyrir höfund þessara lína að raða saman excel-skrám. Án efa urðu til margar gagnavillur í því ferli. Þetta er mikið gæðamál, ekki einungis fyrir Hagfræðideild heldur fyrir Háskóla Íslands í heild. Nauðsynlegt er að viðkomandi aðilar skilgreini markmið og leiðir í slíkri vinnu.

Mat á hnitmiðuðu tölfræðilíkani getur dregið mikið úr óvissu. Verðmæti viðbótargagna hlutfallslega lítið. Almennt er betra að hafa meiri gögn. Ef þau eru úr misleitari hópum kallar það á flóknari líkanasmíði.

Er það þess virði að staðla gögn, svo hægt sé að meta líkön í anda þeirra líkana sem hér hefur verið lýst? Það fer eftir því hver gagnsemi vel metins líkans er. Það er ekki augljóst að það sé besta meðferð á fé að safna gögnum. Mikilvægt er að átta sig á, hver hugsanlegur ávinningur er af aukinni nákvæmni (upplýsingum) í mati á tölfræðilegum líkönum. Mín persónulega ályktun af þessari vinnu er að einstaklingsbreytileikinn virðist það mikill, að lítil efnisleg rök séu fyrir því að meðhöndla beri framhaldsskóla mismunandi. Einkunnir eru hugsanlega þröskulsbreyta sem er illa samanburðarhæf milli skóla. Einstaklingsbreytileikinn er hinn ráðandi þáttur.