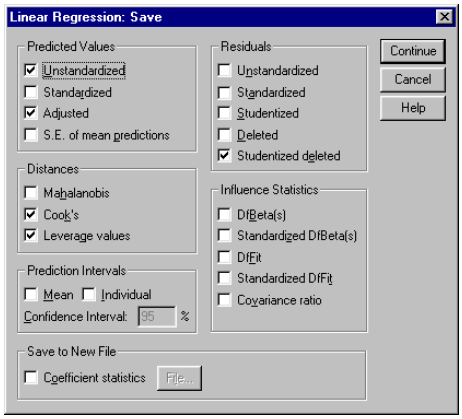
Greiningarstuðlar í leif

Greiningarstuðlar í leif (*diagnostics*) gefa upplýsingar um það hvernig aðfallslíkanið passar við þau einstöku mæligildi sem það byggir á. Í meginatriðum eru tvenns konar vandamál sem finna má með slíkum greiningarstuðlum. Annars vegar geta verið einstök mæligildi sem líkanið nær ekki að lýsa á fullnægjandi hátt. Þetta eru frávillingar (*outliers*) í leif, þ.e. mæligildi sem sýna mikil frávik frá spágildi líkansins. Hins vegar geta einstök mæligildi haft óvenjulega mikil áhrif á líkanið, þ.e. hallastuðlana. Þetta verður til þess að líkanið lagar sig að þessum áhrifamiklu (*influential*) mæligildum og lýsir öðrum mæligildum ver sem því nemur. Sama mæligildið getur verið bæði áhrifamikið og frávillingur í leif þótt það fari ekki alltaf saman.



Greiningarstuðla fást í SPSS með því að fara í Linear Regression valmyndina og ýta á Save. Þá birtist valmyndin hér til hliðar þar sem má velja ýmsa greiningarstuðla. Yfirleitt er fullnægjandi að biðja um Cook's distance, Leverage, spágildi líkansins og staðlaða leif (við veljum að fá Studentized deleted).[1] Við ýtum síðan á Continue, breytum hugsanlega einhverju á öðrum valmyndum en ýtum að lokum á OK á aðalvalmyndinni. Þá tekur SPSS við, reiknar líkanið og vistar umbeðna greiningarstuðla í gagnaskrána sem unnið er með.

```
REGRESSION
  …
/SAVE PRED ADJPRED LEVER COOK SDRESID.
```

Taflan hér við hliðar sýnir greiningarstuðlanna. Fyrsti dálkurinn er kunnuglegur enda sýnir hann grunnskóla í Reykjavík. Líkanið sem unnið er með byggist á því að spá fyrir um meðaleinkunn einstakra skóla á grundvelli menntunarstigs skólahverfis og fjölda óheimilla fjarvista.[2] Næsti dálkur sýnir staðlaða leif (Studentized) fyrir hvern skóla. Þar er oft miðað við að mæligildi sem eru innan markanna ±2 sé vel lýst af líkaninu, vafi leiki á um mæligildi fyrir utan mörkin ±2 og að mæligildi sem eru ±3 eða hærri í staðlaðri leif sé illa lýst af líkaninu. Taka verður þessum leiðsagnarreglum með fyrirvara. Búast má við því að 5% mæligilda séu fyrir utan ±2 í normaldreifðri leif og um 0,1% fyrir utan ±3. Því er að meðaltali eitt mæligildi fyrir utan ±3 í þúsund manna úrtaki ef leifin er normaldreifð. Þótt þetta virðist mikið frávik er ástæðulaust að elta ólar við slíkt mæligildi í stóru úrtaki; frávikið endurspeglar eðlilegan breytileika fremur en einhvern grundvallareiginleika líkansins eða mæligildisins sjálfs.

Skóli	Leif	Vogarafl	Cook
Hagaskóli	0,20	**0,21**	0,00
Hlíðaskóli	0,87	0,15	0,05
Hvassaleitis	–0,11	0,20	0,00
Ölduselssk	1,93	0,05	0,06
Álftamýrask	0,15	0,12	0,00
Austurbæjar	0,24	**0,24**	0,01
Háteigsskóli	-0,01	0,08	0,00
Laugalækjar	0,22	**0,26**	0,01
Foldaskóli	1,59	0,11	0,09
Réttarholts	0,31	0,12	0,00
Vogaskóli	-0,01	0,16	0,00
Árbæjarskóli	-1,78	0,11	0,11
Seljaskóli	0,22	0,07	0,00
Hamraskóli	0,98	0,12	0,04
Langholtsskóli	-0,69	0,06	0,01
Hólabrekkuskóli	0,09	0,19	0,00
Rimaskóli	-0,47	0,10	0,01
Húsaskóli	**-3,21**	0,05	0,13
Breiðholtsskóli	-0,13	**0,26**	0,00
Fellaskóli	-0,95	**0,33**	0,15

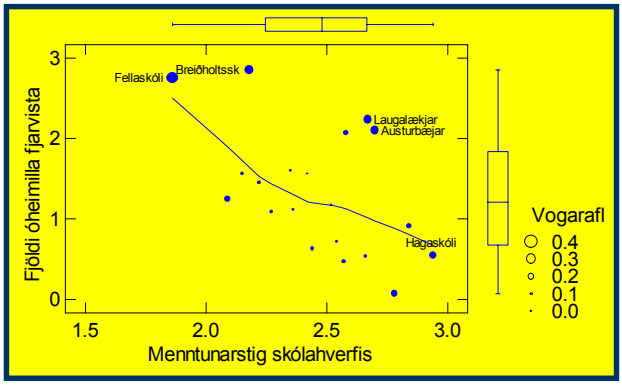
Í töflunni sjáum við aðeins eitt mæligildi sem er athugunarvert vegna mikillar staðlaðrar leifar. Þetta er Húsaskóli sem er með meðaltal sem er rúmum þremur staðalfrávikum lægra en spágildið. Þetta er mjög óvenjulegt í þetta litlu úrtaki og bendir til þess að Húsaskóli sé annaðhvort mjög sérstakur skóli eða að líkanið sé á einhvern hátt ófullnægjandi. Ljóst er að líkanið lýsir Húsaskóla illa og rétt að líta á hann sem frávilling (*outlier*) frá líkaninu.

Þriðji dálkur töflunnar sýnir vogarafl (*leverage*) mæligildanna. Vogarafl er mælikvarði á það hversu óvenjulegt mæligildið er með tilliti til frumbreytanna. Mæligildi er með óvenjuleg gildi á frumbreytum ef það er jaðargildi (*extreme value*) á einhverri frumbreytanna eða með mjög ólíklega samsetningu gilda á frumbreytum. Dæmi um það síðarnefnda væri t.d. ef einhver skólanna væri í skólahverfi með háu menntunarstigi en jafnframt mörgum óheimiluðum fjarvistum (hér er gert ráð fyrir neikvæðri fylgni milli þessara frumbreyta). Almennt séð er vogtalan mælikvarði á það hvort mæligildi sé frávillingur þegar horft er eingöngu til frumbreyta.

---

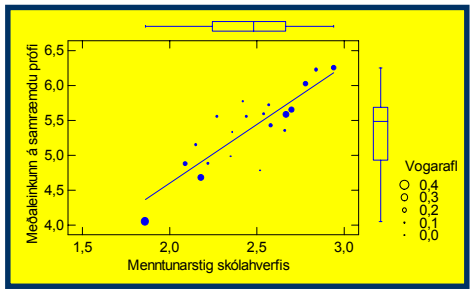
[1] Á valmyndinni hefur verið hakað við óstöðluð (unstandardized) og leiðrétt (adjusted) spágildi. Óstaðlaða talan er þetta hvunndagsspágildi sem við getum handreiknað út frá fasta og hallatölum líkansins. Leiðrétta spágildið er hins vegar reiknað með því að fjarlægja viðkomandi færslu úr líkaninu, reikna líkanið aftur og nota spágildið sem fæst þannig. Yfirleitt skiptir litlu hvor talan sé notuð.

[2] Ítarlega er fjallað um þetta líkan á öðru leiðbeiningarblaði.



Meðaltal vogarafls er p/N þar sem p er samanlagður fjöldi hallastuðla og fasta í líkaninu og N er fjöldi einstaklinga (staka; færslna). Erfitt er að meta hvenær veita skal mæligildi athygli vegna vogarafls, en oft er miðað við að vogarafl á bilinu 0,2–0,5 sé athugunarvert og vogarafl yfir 0,5 sé orðið hættulega mikið (SPSS, 1999, bls. 376). Ýmis önnur viðmið hafa verið nefnd, t.d. að vogarafl sé ekki meira en tvöfalt meðaltalið (þ.e. ekki hærra en 0,3 í okkar tilfelli; Fox, 1991, bls. 33; Fox, 1997, bls. 280). Almennt séð er þó rétt að veita þeim mæligildum athygli sem skera sig úr, t.d. á kassariti, fyrir hátt vogarafl.

Við höfum tvær frumbreytur og 20 skóla svo meðaltal vogarafls ætti að vera 3/20= 0,15 og tvöföld sú tala er 0,3. Samkvæmt töflunni eru fimm skólar með vogarafl yfir 0,2. Fellaskóli sker sig úr með vogtölu yfir 0,3 en hann er með lægsta menntunarstigið og hlutfallslega flestar fjarvistir. Vogarafl Breiðholtsskóla er töluvert lægri vegna þess að skólinn sker sig aðeins úr vegna fjarvista. Laugarlækjar- og Austurbæjarskóli skera sig hvorki úr vegna fjarvista né vegna menntunarstigs skólahverfis. Hins vegar er staðsetning þeirra óvenjuleg þegar breyturnar eru skoðaðar saman, því þeir eru með óvenjulega miklar fjarvistir þegar til þess er tekið hve hátt menntunarstig skólahverfisins er. Að síðustu er Hagaskóli með vogarafl rétt yfir 0,2 vegna þess að hann er bæði með hæsta menntunarstigið og mjög fáar óheimilar fjarvistir.



Vogarafl segir til um hversu fast mæligildið getur togað í aðfallslínuna og haft þannig áhrif á líkanið. Myndin hér til hliðar sýnir samband einnar frumbreytu við fylgibreytu. Vogarafl er greinilega mest til endanna á línunni, þ.e. hjá þeim mæligildum sem eru lengst frá meðaltali frumbreytunnar. Þetta má hugsa þannig að mæligildin togi í línuna og reyni að breyta halla hennar. Eðlilega hafa þau mæligildi sem eru til endanna meiri möguleika á því að breyta halla línunnar en þau sem eru nær miðju frumbreytunnar. Þetta má hugsa þannig að línan leiki á ás sem liggur við meðaltal frumbreytunnar og mæligildin leitist við að láta línuna snúast um þennan ás. Eðlilega verður átakið meira ef við erum við enda línunnar heldur en ef við erum rétt við ásinn; vogaraflið er meira eftir því sem við fjarlægjum snúningsás línunnar. Ef frumbreytur eru margar, hefur mæligildi því meira vogarafl því óvenjulegra sem það er með tilliti til frumbreyta; þ.e. því fjarlægara sem það er heildarmiðju gagnasafnsins.

Síðasti dálkurinn sýnir fjarlægð Cooks (*Cook's distance*). Fjarlægð Cooks er mælikvarði á það hversu áhrifamikið mæligildið er. Mæligildi með hátt Cooksgildi hefur mikil áhrif á hallastuðla líkansins. Áhrifin má sjá með því að fjarlægja mæligildið úr líkaninu og bera líkanið saman við upprunalega líkanið. Ýmis viðmið eru til um stærð á fjarlægð Cooks, ýmis föst viðmið eða breytilegt eftir fjölda frumbreyta og fjölda færslna (þátttakenda, staka). Stundum er miðað við að fjarlægð Cooks fari ekki yfir 2,0 (SPSS, 1999, bls. 375), fari ekki yfir $\frac{4}{n-k-1}$ (Fox, 1991, bls. 34; Fox, 1997, bls. 281) en einnig má bera tölugildin saman við F-dreifingu ($\sim F(p, n-p)$) og miða þannig við einhverjar tiltekin líkindi, t.d. 5%.[3]

Í okkar gögnum eru engin mæligildi með hátt Cooksgildi. Við gætum miðað við $\frac{4}{n-k-1}$, þ.e. 0,24, eða það gildi sem gæfi 5% líkindi ($F_{0,95}(3,17)$), þ.e. 3,2. Ekkert mæligildi er nálægt þessum mörkum og því engin áhrifamikil mæligildi í gögnunum. Fellaskóli er með hæsta Cooksgildið en þó aðeins 0,15 sem er langt frá báðum viðmiðunum.

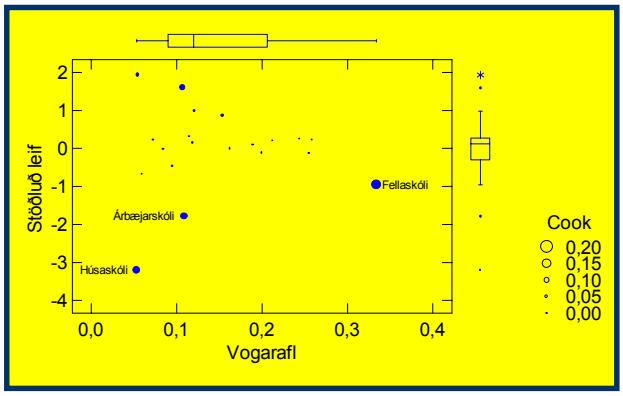
Við höfum fundið a.m.k. tvo skóla sem eru athugunarverðir. Húsaskóli er frávillingur í leif og illa lýst af líkaninu. Fellaskóli er hugsanlegur frávillingur ef litið er eingöngu til frumbreyta. Enginn skólanna virðist þó skekkja líkanið svo nokkru nemi. Líkanið virðist því ágætlega lýsandi og rétt svo langt sem það nær fyrir 19 af skólunum 20.

Skoðum nú samhengið milli þessara þriggja greiningarstuðla. Húsaskóli er með mikil frávik í leif en lítið vogarafl. Þótt hann sé vel innan marka þá er hann með annað hæsta Cooksgildið í úrtakinu. Fellaskóli er með mikið vogarafl en tiltölulega lítið frávik í leif. Hann er með hæsta

[3] n: fjöldi færslna; k: fjöldi frumbreyta; p: fjöldi hallastuðla að fastanum meðtöldum; $p = k + 1$.

Cooksgildið. Austurbæjar-, Laugarlækjar- og Breiðholtskóli eru allir með tiltölulega hátt vogarafl, en nánast engin frávik í leif og því næst 0,0 í Cooksgildi.

Samhengið milli greiningarstuðlanna þriggja sést á myndinni hér til hliðar. Skólar með lítið vogarafl geta verið tiltölulega áhrifamiklir ef þeir hafa mikil frávik í leif og að sama skapi geta skólar haft mikil áhrif á líkanið ef þeir hafa mikið vogarafl jafnvel þótt frávik í leif sé aðeins miðlungi mikið. Áhrifamestu mæligildin eru þau sem eru með mikið vogarafl samfara miklum frávikum í leif. Þetta fer hvergi saman hjá skólunum 20 og því finnum við engin áhrifamikil mæligildi í gagnasafninu. Við getum dregið þetta saman með eftirfarandi gervijöfnu: Áhrif á líkan= Vogarafl × Frávik í leif. Vogaraflið segir til um hversu gott tak mæligildi hefur á líkanið, frávikið í leif gefur til kynna hversu sterkt mæligildið togar í hallastuðla líkansins og fjarlægð Cooks gefur til kynna heildaráhrif mæligildisins á líkanið.



Berum í lokin saman líkönin eftir að einstakir skólar hafa verið felldir brott. Við viljum sjá upprunalega líkanið og hafa það til samanburðar. Til álita kemur að fella brott Húsaskóla, sem hafði mikil frávik í leif en lágt Cooksgildi. Við eigum ekki von á miklum breytingum á líkaninu en forspárhæfni þess ætti að batna.[4] Einnig væri áhugavert að fella brott Fellaskóla því hann hafði hæsta Cooksgildið. Við eigum von á lítils háttar breytingum á líkaninu en forspárhæfnin ætti ekki að batna að neinu ráði þar sem Fellaskóli var ekki frávillingur í leif líkansins.

Breyta	Hallatala	SE	p
Allir skólar			
Fasti	2,06	0,76	0,01
MenntStig	1,44	0,27	0,01
Fjarvist	–0,17	0,10	0,10
R^2= 0,76 SE_{Est}= 0,28			
Allir nema Húsaskóli			
Fasti	2,05	0,61	0,01
MenntStig	1,46	0,22	0,01
Fjarvist	–0,18	0,08	0,04
R^2= 0,85 SE_{Est}= 0,23			
Allir nema Fellaskóli			
Fasti	2,31	0,80	0,01
MenntStig	1,33	0,30	0,01
Fjarvist	–0,15	0,10	0,17
R^2= 0,67 SE_{Est}= 0,28			

Ef við skoðum töfluna hér til hliðar sjáum við að staðalfrávik spágildis lækkar umtalsvert (um 20%) við það að fella Húsaskóla brott. Skýrð dreifing hækkar sömuleiðis um tæp 10 prósentustig. Staðalvillur hallatalna lækka einnig sem felur í sér nákvæmara mat á líkaninu. Þetta er því greinilega betra líkan, þ.e. ef við getum skilgreint Húsaskóla út úr þýðinu. Hugsanlega er þetta þó rangt líkan; ef hægt er að finna ástæðu þess að Húsaskóli er svona mikið frávik frá líkaninu getur það orðið til að nýtt líkan uppgötvist sem lýsir öllum skólunum 20.

Það að fella brott Fellaskóla breytir hallastuðlum lítillega en hefur lítil áhrif á staðalvillur. Forspárhæfni líkansins breytist ekkert, samanber óbreytta staðalvillu spágildis. Skýrð dreifing minnkar hins vegar um svo mikið sem 9 prósentustig. Það skýrist af því að Fellaskóli var langlægstur á báðum frumbreytum. Við brottfall hans minnkar verulega dreifing beggja frumbreyta sem verður til þess að minna það hlutfall af dreifingu fylgibreytunnar sem skýrist af dreifingu frumbreytanna. Hér erum við því að sjá hin velþekktu áhrif þess að minnka dreifisvið frumbreyta. Þetta sýnir einnig hversu villandi mælikvarði R^2 getur verið; hér minnkar skýrð dreifing vegna minnkaðrar dreifingar frumbreyta án þess að forspárhæfni líkansins hafi í reynd breyst.

Þessi litla athugun okkar á greiningarstuðlum hefur gefið okkur aukna innsýn inn í eðli líkansins og þeirra fyrirbæra (skólanna) sem það lýsir. Þetta er hinn almenni tilgangur tölfræðilegrar úrvinnslu að lýsa tengslum fyrirbæra og fá vitneskju um gæði og eiginleika þeirrar lýsingar sem við setjum fram með tölfræðilegum aðferðum. Þekking á greiningarstuðlum og færni í túlkun þeirra er mikilvæg ef öðlast á leikni í tölfræðilegri úrvinnslu.

[4] Auðvitað er villandi að orða þetta svona. Forspárhæfnin skánar ekki hót við að fjarlægja fráviksgildið heldur breytast niðurstöður líkansins og væntanlega ofmeta forspárgildið miðað við að fráviksgildið (Húsaskóli) sé enn hluti af þýðinu.

Heimildir

Fox, J. (1991). *Regression diagnostics: An introduction. Exploratory data analysis. Sage university paper series on quantitative applications in the social sciences, 07-79.* Thousand Oaks, CA and London: Sage.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA og London: Sage.

SPSS (1999). *Systat. Statistics I*. Chicaco: Höfundur.