Marghliða aðfallsgreining

Fyrirlestur í Tölfræði III (SÁL308G)

31.08.15

Marghliða aðfallsgreining

Í marghliða aðfallsgreiningu höfum við tvær eða fleiri frumbreytur. Við gerum ráð fyrir að frumbreyturnar hafi línuleg tengsl við fylgibreytuna og villan sé normaldreifð með einu sameiginlegu staðalfráviki.

Takið eftir að engar sérstakar kröfur eru til breytanna sjálfra.

Þótt líkanið sé svipað og í einfaldri aðfallsgreiningu verður túlkunin mun vandasamari við það að frumbreytum fjölgar.

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$
$$\varepsilon \sim N(0, \sigma)$$

$$\mu_y = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

Eins og áður segir β til um hvað fylgibreytan muni breytast mikið þegar frumbreytan hækkar eða lækkar um eina einingu.

Það sem er öðru vísi er að núna gefur β_p upp breytinguna sem verður þegar x_p breytist en allar aðrar frumbreytur haldast óbreyttar.

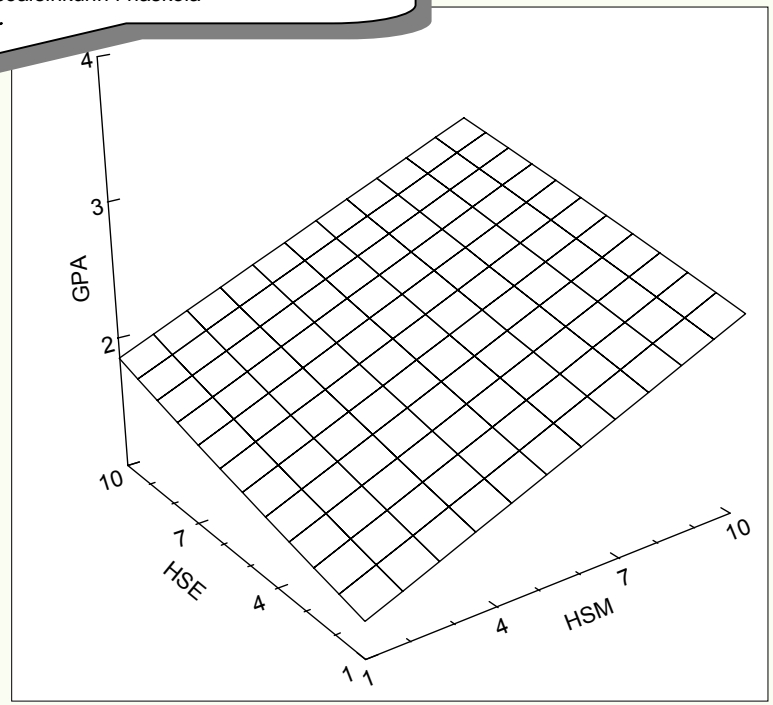
Engin aðfallslína

Gögnin koma úr athugun á tengslum einkunna í ensku, stærðfræði og náttúrufræði (*science*) við meðaleinkunn í háskóla

Myndin sýnir dæmi um tengsl tveggja frumbreyta við fylgibreytu. Taktu eftir því að tvær frumbreytur mynda flöt en ekki línu eins og þegar aðeins er ein frumbreyta.

Fastinn segir til um gildi fylgibreytu þegar allar frumbreytur eru 0.

Ef enskueinkunn (HSE) helst óbreytt, hækkar GPA um 0,18 fyrir hverja einingu sem HSM hækkar. Ef HSM er óbreytt, hækkar GPA um 0,06 fyrir hverja einingu sem HSE hækkar.



$$\hat{\mu}_y = 0{,}62 + 0{,}06 \cdot HSE + 0{,}18 \cdot HSM$$

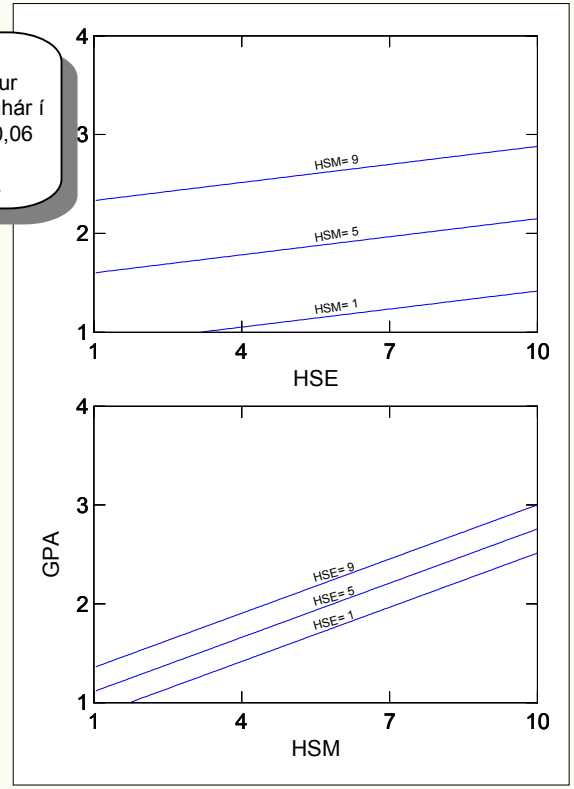
Skilyrtar hallatölur

Efri myndin sýnir tengsl HSE og GPA fyrir þrjú gildi á HSM. Við erum með beina línu með hallatöluna 0,06 sama hvert gildi HSM er.

Neðri myndin sýnir að fyrir hvert gildi HSE fáum við beina línu sem lýsir tengslum HSM og GPA, hallatalan er 0,18.

Þannig lýsir hallatala í marghliða aðfallsgreiningu áhrifum frumbreytu þegar öðrum frumbreytum er haldið föstum.

Nemandi sem er einum heilum hærri í ensku heldur en annar nemandi en jafnhár í stærðfræði er að jafnaði 0,06 hærri í GPA—sama hver einkunnin í stærðfræði er.



Túlkun hallatalna

Hallatölur í aðfallsgreiningu gefa til kynna áhrif frumbreytu þegar öðrum frumbreytum er haldið föstum í einhverjum ákveðnum gildum.

Þetta samsvarar því að við metum skilyrt eða sérhæf áhrif frumbreyta.

Stundum er talað um að leiðrétt sé fyrir áhrif annarra frumbreyta og hallatalan sýni því þau áhrif sem frumbreytan deilir ekki með öðrum frumbreytum líkansins.

Yfirleitt eru frumbreytur tengdar innbyrðis. Það væri því óvenjulegt ef ein frumbreyta breyttist án þess að aðrar breyttust um leið.

Ef við berum saman tvo nemendur sem eru misgóðir í íslensku, má búast við samsvarandi mun í stærðfræði og ensku.

Aðfallsgreining er aðferð til að greina niður áhrif frumbreyta og meta sérhæf áhrif hverrar fyrir sig.

Mat á líkaninu

Við þekkjum yfirleitt ekki líkanið heldur verðum að meta það í úrtakinu. Matið er óvisst og því metum við óvissuna með því að reikna staðalvillu hallatalnanna.

Við getum reiknað spágildi fyrir tiltekin gildi frumbreytanna og einnig metið leifina fyrir hvern og einn þátttakanda.

Leifin er mat okkar á villunni og því er mikilvægt að skoða hana vel.

Við fáum einnig R^2 sem er mat á forspárhæfni líkansins.

$$y_i = b_o + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + e_i$$

$$\hat{y}_i = b_o + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

$$e_i = y_i - \hat{y}_i$$

Marktektarpróf á líkanið

Við fáum allsherjar *F*-próf fyrir líkanið í heild sinni. Það prófar þá núlltilgátu að allar hallatölur séu 0,0 í þýði. Yfirleitt gerum við þá kröfu að geta hafnað þessari allsherjartilgátu áður en einstakir hallastuðlar eru prófaðir.

Við fáum einnig staðalvillu og *t*-próf fyrir hvern hallastuðul fyrir sig. Þessi marktektarpróf prófa þá núlltilgátu að *sérhæf* áhrif viðkomandi frumbreytu séu engin í þýði.

$$H_0 : \beta_1 = \beta_2 = \cdots \beta_p = 0$$

Marktektarpróf í aðfallsgreiningu eru fremur vandræðaleg. Ef *F*-prófið er marktækt, ályktum við að hallastuðlarnir, einn eða fleiri, séu ekki núll í þýði.

Vandinn er sá að stundum er *F*-prófið marktækt en enginn af hallastuðlunum. Við getum einnig fengið ómarktækt *F*-próf en einn eða fleiri marktæka hallastuðla. Það getur því verið áberandi ósamræmi í niðurstöðum.

Almenna reglan er að túlka ekki hallastuðla nema *F*-prófið sé marktækt.

Athuga lýsandi tölfræði

Einföld lýsandi tölfræði er fyrsta skref allrar úrvinnslu. Hér sjáum við hvort gögnin hafi verið lesin rétt inn, hvort einhver sérkennilegheit eru til staðar eða önnur frávik.

Við sjáum að það er ekkert brottfall. Hæstu og lægstu gildi eru öll eðlileg. Meðaltölin liggja dálítið hátt fyrir einkunnir í framhaldsskóla en það er væntanlega til marks um háar einkunnir í bandarískum skólum.

Staðalfrávik virka eðlileg þótt oft sé erfitt að meta hvað teljist eðlilegt.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
OBS	224	1	224	112,50	64,807
GPA	224	,12	4,00	2,6352	,77939
HSM	224	2	10	8,32	1,639
HSS	224	3	10	8,09	1,700
HSE	224	3	10	8,09	1,508
Valid N (listwise)	224				

Aldrei gleyma þessu mikilvæga skrefi. Hér koma yfirleitt alvarlegustu mistökin í ljós.

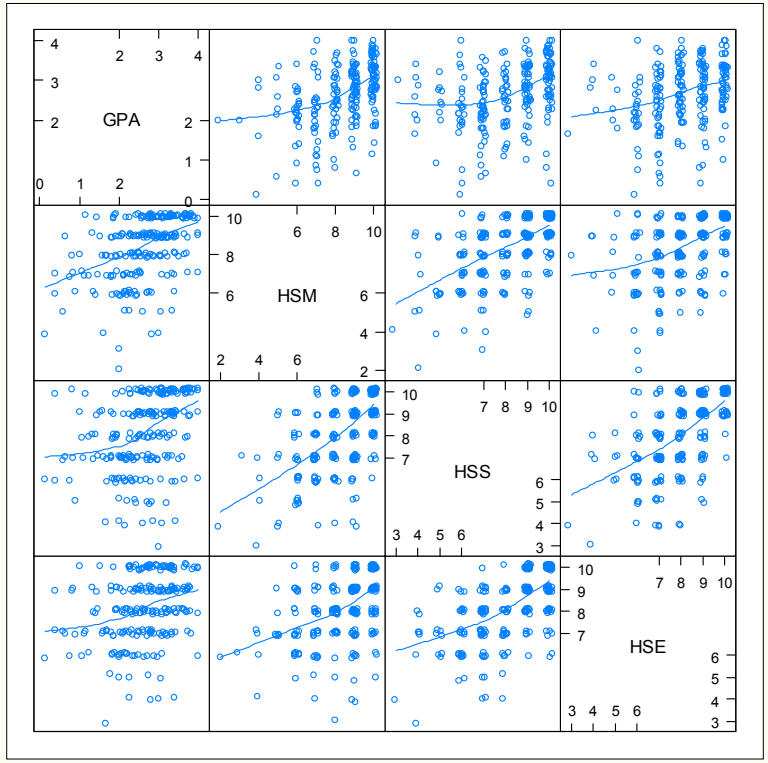
Það er tímafrekt að endurtaka úrvinnslu og einnig alvarlegt ef ekkert er að marka niðurstöður vegna annmarka í gögnunum. Þetta er því einföld aðferð til að sjá algengar misfellur í gögnum.

Forathuganir á tengslum breyta

Aðfallsgreining gerir ekki formlega kröfu um að frumbreytur tengist fylgibreytu eða innbyrðis með beinni línu. Líkanið er hins vegar línulegt og þarf að passa við gögnin.

Ef tengslin sýna áberandi frávik frá beinum línum, er ástæða til að hafa áhyggjur.

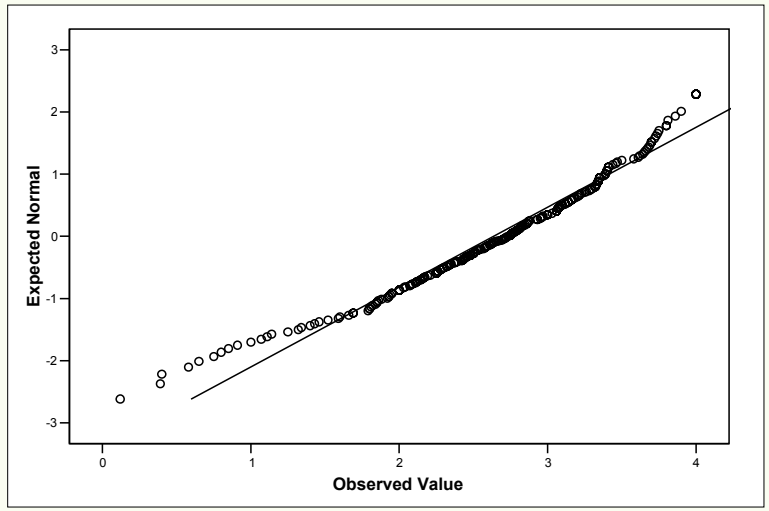
Myndin sýnir fylki fylgnirita með tregum línum. Frávik frá beinni línu eru ekki stórvægileg þótt mögulega séu einhver frávik til staðar í þýði.



Forathugun á fylgibreytu

Aðfallsgreining gerir engar formlegar kröfur til fylgibreytunnar. Það er samt mikilvægt að skoða hana vel til að átta sig á eiginleikum hennar. Það er forsenda þess að við getum túlkað niðurstöðuna með nægjanlegu öryggi.

Í þessu tilviki virðist breytan með lítil frávik frá normaldreifingu nema að neðri hali dreifingarinnar er mjög langur og með tilhneigingu til fráviksgilda. Við gætum því þurft að gæta okkar á frávillingum. Einnig eru væg rjáfurhrif.

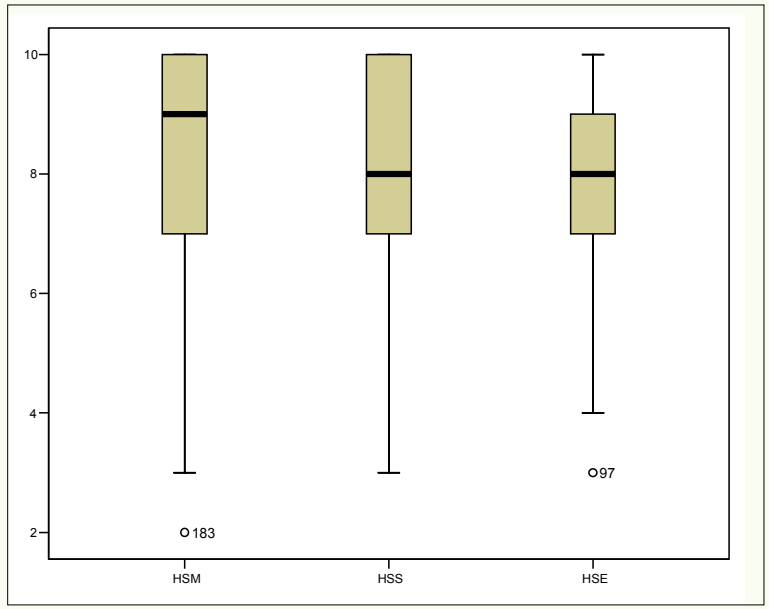


Forathugun á frumbreytum

Kassaritin eru í samræmi við þessi háu meðaltöl sem við höfum séð.

Það eru greinileg rjáfurhrif á öllum frumbreytum og væg tilhneiging til fráviksgilda.

Frumbreyturnar þurfa ekki að vera normaldreifðar. Skekkju og rjáfur í dreifingum þeirra þarf þó að hafa í huga. Það er hugsanlegt að það verði einhver fráviksgildi í úrvinnslunni vegna þess en þó alls ekki víst að svo verði.



Forspárhæfni

Við ætlum að lýsa tengslum einkunna í ensku, stærðfræði og náttúrufræði (*science*) við meðaleinkunn í háskóla.

Taflan segir okkur að forspárhæfnin sé ekkert sérstök. Við getum skýrt um fimmtung af dreifingu einkunna í háskóla. Við sjáum einnig að leifin hefur staðalfrávikið 0,7 sem er dálítið mikið þar sem ein eining skilur að stafaeinkunnir (þ.e. A, B, C og D).

R^2 er að jafnaði ofmetið í úrtaki en leiðrétt R^2 reynir að leiðrétta það.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,452[a]	,205	,194	,69984

a. Predictors: (Constant), HSE, HSM, HSS

b. Dependent Variable: GPA

Það þarf aðgát við túlkun R^2. Það þarf ekki að vera undarlegt að forspárhæfni sé ekki meiri en þetta. Hér er farið á milli skólastiga og því breytingar á áherslum, kröfum og viðfangsefnum.

Með fleiri eða betri forspárbreytum má hugsanlega fá betri forspá en við verðum þó að gera ráð fyrir því að það séu einhver efri mörk í forspánni. Við gætum verið að nálgast þau.

Dreifigreiningartaflan

Dreifigreiningartöfluna túlkum við eins og fyrr. Frígráður líkansins eru þrjár eins og fjöldi frumbreyta.

Ef núlltilgátan er rétt, meta MS_M og MS_E bæði villuna. Í okkar tilviki er MS_M næstum 19 sinnum stærri og því ótrúverðugt að núlltilgátan sé rétt. Svona mikill munur myndi finnast í minna en 0,1% tilvika.

Við ályktum því að hallatölurnar geti ekki allar verið 0,0 í þýði.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,712	3	9,237	18,861	,000[a]
	Residual	107,750	220	,490		
	Total	135,463	223			

a. Predictors: (Constant), HSE, HSM, HSS

b. Dependent Variable: GPA

Við ályktum að einhver tengsl séu milli frumbreyta og fylgibreytu, þ.e. einkunnir í framhaldsskóla tengist meðaleinkunn fyrstu þriggja missera í tölvufræði.

Við vitum hins vegar ekki hver tengslin eru, til þess þurfum við að skoða hallatölurnar.

Munum líka að niðurstöðurnar eiga við tiltekin bandarískan háskóla og þurfa því ekki að lýsa aðstæðum almennt.

Túlkun hallatalna

Allar hallatölur eru lágar, hugsanlega sökum þess að frumbreyturnar deila áhrifum á milli sín.

Áhrif einkunna í náttúrufræði og ensku eru ekki marktæk. Þau eru einnig lítil, jafnvel þótt litið sé til efri marka öryggisbila.

Stærðfræði hefur mun meiri áhrif og með marktekt, en áhrifin eru samt fremur lítil.

Kannski eru áhrifin bara svona lítil en einnig gætu frumbreytur hafa deilt þeim á milli sín.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	,590	,294		2,005	,046	,010	1,170
	HSM	,169	,035	,354	4,749	,000	,099	,239
	HSS	,034	,038	,075	,914	,362	-,040	,108
	HSE	,045	,039	,087	1,166	,245	-,031	,121

a. Dependent Variable: GPA

Við túlkun hallastuðla lítum við á marktekt, stærð hallastuðlana og öryggisbilin.

Ómarktækur hallastuðull gæti haft veruleg áhrif í þýði ef öryggisbilið gefur það sem sennilegan möguleika.

Í stóru úrtaki getur hallastuðull verið marktækur en öryggisbilið samt sýnt að sennilega eru áhrifin ekki umtalsverð í þýði.

Skilyrtir og óskilyrtir hallastuðlar

Hér ber ég saman hallastuðla í marghliða aðfallsgreiningu við þá sem fást ef hver frumbreyta er sett inn í sína einföldu aðfallsgreiningu.

Stærðfræðieinkunnir halda sínu en náttúrufræði og enska hafa mun minni áhrif þegar hinum frumbreytunum er haldið föstum.

Nemandi sem er hærri á HSS (eða HSE) er að jafnaði einnig hærri á HSM og fær hærra GPA. Ef HSS hækkar *án þess að HSM eða HSE hækki*, eru áhrifin afar lítil á GPA.

	Skilyrt	Óskilyrt
HSM	0,17	0,21
HSS	0,03	0,15
HSE	0,05	0,15

Tveir nemendur með ólíka enskueinkunn (HSE) eru að jafnaði einnig með ólíka einkunn í stærðfræði. Þetta skilar sér að jafnaði sem ólík einkunn í tölvunarfræði í háskóla.

Ef við hins vegar berum saman nema með ólíka enskueinkunn en *sömu einkunn í stærðfræði*, er munurinn á GPA afar lítill.

Sérhæf áhrif frumbreytu

Hér er önnur rannsókn sem kannar tengsl sjálfsmyndar og einkunna.

Taktu eftir að áhrif sjálfsmyndar eru nær helmingi minni þegar greind er einnig í líkaninu.

Nærtækasta skýring er sú að þeir sem hafi góða sjálfsmynd séu gjarnan vel gefnir. Ef við berum saman nemendur með misgóða sjálfsmynd, erum við því einnig að bera saman misjafnlega vel gefna nemendur. Líkan 1 ofmetur því áhrif sjálfsmyndar á frammistöðu í námi.

	Líkan 1	Líkan 2
Fasti	2,23	–3,89
Sjálfsmynd	0,09	0,05
Greind	—	0,08

Í líkani tvö eru áhrif sjálfsmyndar leiðrétt fyrir áhrif greindar. Í því felst að við erum að meta áhrif sjálfsmyndar á einkunnir miðað við að greind sé haldið fastri.

Líkan tvö metur sérhæf áhrif sjálfsmyndar, þ.e. þau áhrif sjálfsmyndar á frammistöðu í námi sem ekki er deilt með áhrifum greindarfars.

Dæmi um sérhæf og ósérhæf áhrif

Karlar eru betri en konur í að skipta um ljósaperur

> Réttara er að hávaxnir eiga auðveldara með að teygja sig og karlar eru að jafnaði hærri en konur.

Það að vera karl leiðir til hærri launa en konur fá

> Hluti áhrifanna er meiri menntun karla en kvenna—ef ungt fólk er undanskilið. Ef leiðrétt er fyrir menntun, minnkar launamunurinn en hverfur þó ekki.

Aðfallsgreining *leiðréttir* fyrir áhrif utan að komandi breytu með því t.d. að hafa (a) bæði kyn *og* líkamshæð eða (b) bæði kyn og menntun.

Slík tölfræðileg leiðrétting gefur mat á áhrifum kyns óháð hinni breytunni. Matið er óskekkt ef hin breytan er mæld án villu en annars er um vanleiðréttingu að ræða.

Sérhæf áhrif—tölfræðileg leiðrétting—er dæmi um stjórn (*control*). Sambærilega stjórn má fá með rannsóknarsniðinu, t.d. athuga áhrif kyns fyrir jafnstórt fólk eða áhrif kyns fyrir fólk með sömu menntun.

Þannig fæst oftast betri stjórn á ytri breytum en slík snið takmarkast af fjölda utan að komandi breyta, kostnaði, fyrirhöfn, o.s.frv.

Athugun á forsendum

Forsendur aðfallsgreiningar snúa flestar að villunni, þær helstu eru:

- Villan er normaldreifð.
- Villan er óháð.
- Villan hefur sama staðalfrávik óháð gildum frumbreyta.
- Línuleg jafna lýsir tengslum frumbreyta við fylgibreytu.
- Enga mikilvæga breytu vantar.

Þýðisgildin eru óþekkt en við skoðum hvort leifin samræmist forsendum.

Við athugum normaldreifingu villunnar með því að skoða leifina.

Við skoðum leif á móti númerum þátttakenda til að kanna hvort villan sé óháð.

Leifarrit gerir kleift að meta hvort villan sé einsleit, með sameiginlegt staðalfrávik.

Leifarrit gefur vísbendingu um frávik frá línulegum tengslum.

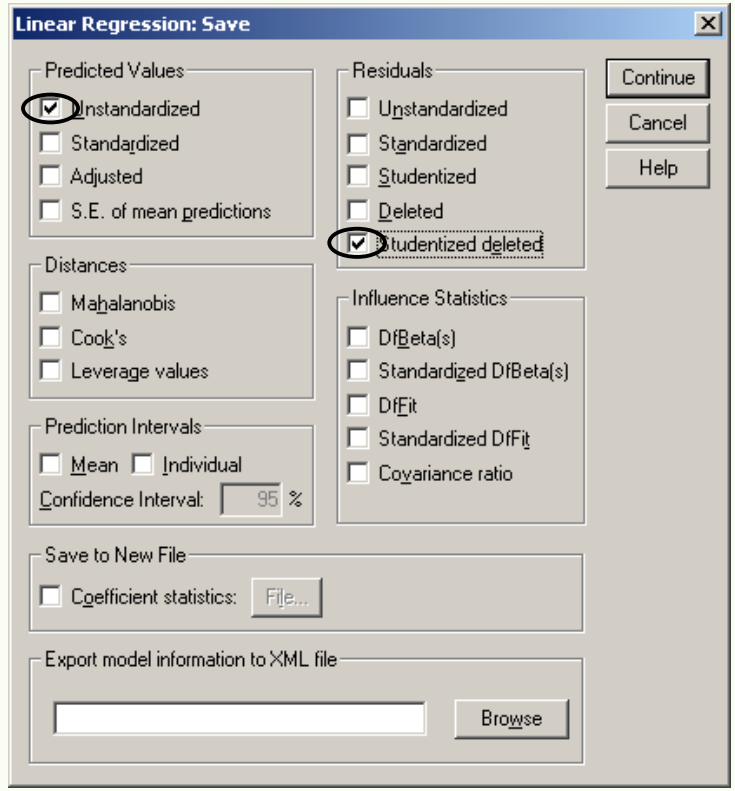
Engin ein aðferð til að kanna hvort mikilvægar frumbreytur vanti.

Greiningarstuðlar

Í marghliða aðfallsgreiningu eru spágildi og stöðluð leif oftast valin til skoðunar.

Upplýsingarnar birtast sem nýjar breytur í gagnaglugganum. Óstaðlað spágildi birtist sem breyta sem byrjar á PRE og stöðluð leif sem breyta með nafn sem byrjar á SDR. Gott er að gefa breytunum nýtt nafn til að forðast rugling.

Breyturnar eru skoðaðar myndrænt samanber næstu glærur.



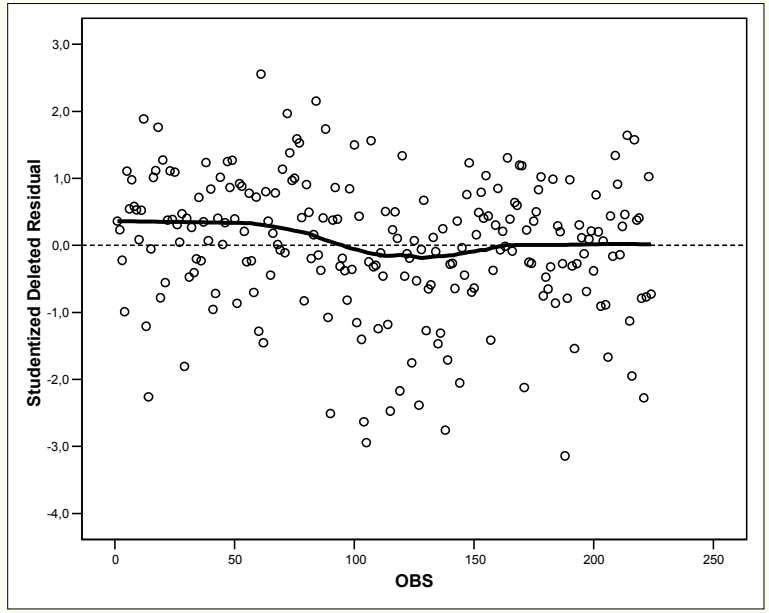
Athuga hvort leifin sé óháð

Hér höfum við teiknað fylgnirit þar sem þátttakendanúmerið er á x-ás og stöðluð leif á y-ás.

Ef allt er með felldu ættum við ekki að finna neitt mynstur í punktaskýinu.

Trega línan sýnir einhverjar sveiflur. Það getur bent til þess að villan sé ekki fyllilega óháð.

Frávikin eru þó tiltölulega væg. Það er óvíst að þetta sé brot á forsendunni um óháða villu en þá væri frávikið fremur lítið.



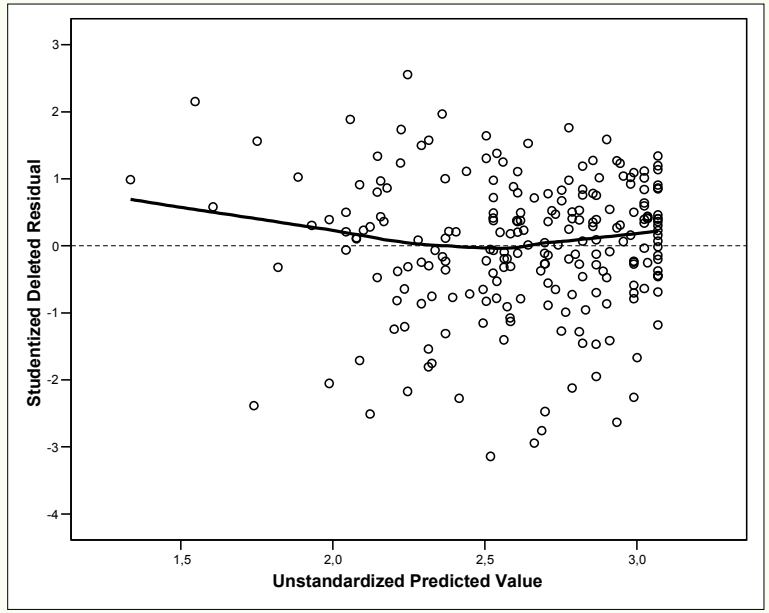
Leifarrit

Leifarrit sýnir staðlaða leif á móti spágildi. Gott er að setja inn trega línu til að sjá betur leitni gagnanna.

Hér er fátt nýtt að sjá. Rjáfurhrif sjást hægra megin á myndinni og gisin dreifing vinstra megin: Þetta er ekki brot á forsendum.

Leifin er of há vinstra megin en dreifingin er gisin og því vitum við ekki hvort það er að marka.

Dreifing leifar virðist mjög svipuð óháð stærð spágildanna.



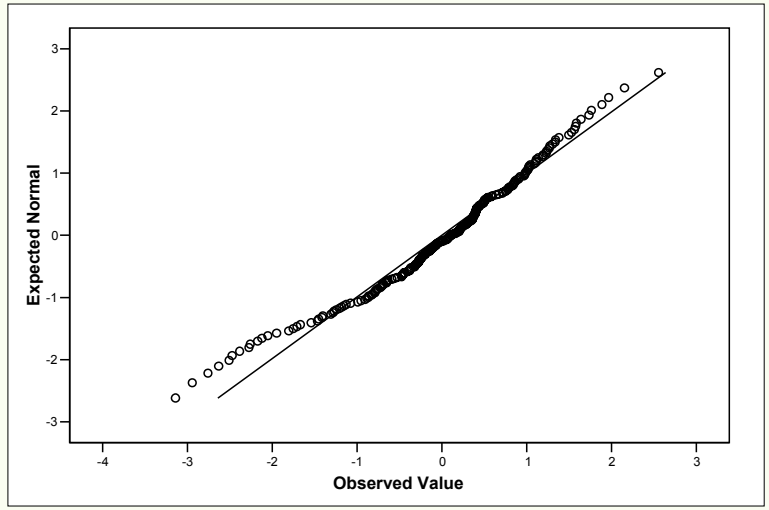
Athugun á leif

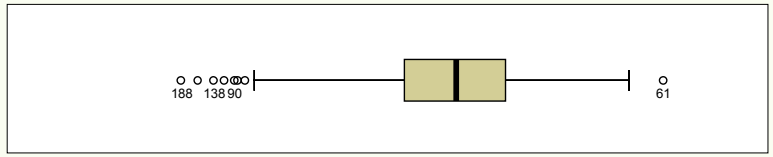
Leifin er með þykka hala (*heavy tails*), þ.e. tilhneigingu til fráviksgilda, sem er vandamál fyrir aðfallsgreiningu. Þess vegna lítum við alltaf eftir fráviksgildum í leifinni.

Best er að nota kassarit en normalrit gefur einnig nánari upplýsingar.

Í okkar tilviki er greinileg tilhneiging til fráviksgilda í neðri endanum, nóg til stuðla að varkárni í túlkun.

Niðurstöður eru afgerandi og því óvíst að þetta skipti máli.





Fækkun frumbreyta

Ef HSS er fellt brott breytist R^2 ekki og forspárhæfni líkansins því óbreytt. Ef HSE er einnig fjarlægt verður forspárhæfnin litlu minni. Staðalvillan fyrir HSM minnkar eftir því sem færri breytur eru í líkaninu.

Ef markmiðið er forspá, veljum við einfaldasta líkanið með nægjanlega forspárhæfni, þ.e. HSM eina og sér.

Merking hallatalnanna breytist við það að frumbreytum fækkar í líkaninu.

R^2= 0,205

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	,590	,294		2,005	,046	,010	1,170
	HSM	,169	,035	,354	4,749	,000	,099	,239
	HSS	,034	,038	,075	,914	,362	-,040	,108
	HSE	,045	,039	,087	1,166	,245	-,031	,121

a. Dependent Variable: GPA

R^2= 0,202

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	,624	,292		2,140	,033	,049	1,199
	HSM	,183	,032	,384	5,716	,000	,120	,246
	HSE	,061	,035	,117	1,747	,082	-,008	,129

a. Dependent Variable: GPA

R^2= 0,191

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	,908	,244		3,727	,000	,428	1,388
	HSM	,208	,029	,436	7,229	,000	,151	,264

a. Dependent Variable: GPA

Forspá eða tilteknar spurningar

Í forspá reynum við að finna einfalda jöfnu sem gefur okkur sem besta spá um fylgibreytuna.

Við lítum því til R^2 og staðalvillu spágildis. Reynt er að fækka frumbreytum án þess að minnka forspárhæfnina umtalsvert.

Lokalíkanið er fundið með því að velja frumbreytur sem skapa góða forspárhæfni en reynt að hafa þær sem fæstar.

Ef leitað er svara við tilteknum spurningum, höldum við mikilvægum frumbreytum inni í líkaninu jafnvel þótt sérhæf áhrif séu ómarktæk og framlag þeirra lítið til forspárhæfni.

Frumbreytur eru fjarlægðar ef það breytir litlu fyrir hallastuðla hinna breytanna og fræðilegt mikilvægi þeirra er takmarkað.

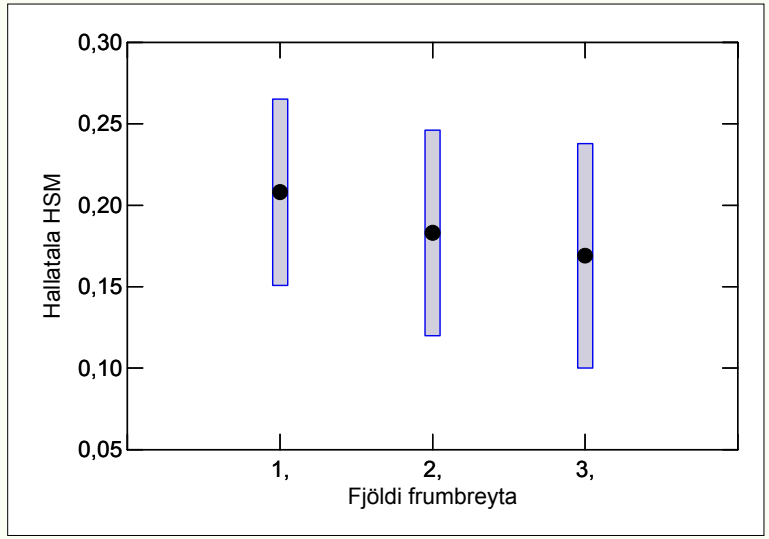
Slík grisjun skilar oft nákvæmara mati á hallastuðlum breytanna sem eftir verða.

Rangt en betra líkan?

Með því að sleppa frumbreytu erum við að ákveða að hallatala hennar sé 0,0 í þýði. Ef við höfum rangt fyrir okkur, skekkjast hinar hallatölurnar.

Matið verður hins vegar oft betra við það að einfalda líkanið, svo við viljum oft taka áhættuna af skekktu líkani til þess að fá nákvæmara mat.

Myndin sýnir þrjár ólíkar niðurstöður fyrir HSM, allar á svipuðu talnabili og erfitt að gera upp á milli. R^2 er nánast það sama í öllum tilvikum.



Öll öryggisbilin gætu metið sömu þýðistölu. Einfaldasta líkanið gefur nákvæmasta hallatölu fyrir HSM. Það sakar sennilega ekki að gera ráð fyrir að hinar breyturnar hafi lítil eða engin sérhæf áhrif í þýði.