

Aðfallsgreining hlutfalla (logistic regression)

Fyrirlestur í Tölfræði III (10.02.62)

Tvískipt fylgibreyta

Þegar við höfum flokkabreytu sem frumbreytu en fylgibreytan er meginndleg, notum við dreifigreiningu.

Stundum er þessu öfugt farið. Hvað gerum við þegar fylgibreytan er flokkabreyta?

Gegnum tíðina hefur aðfallsgreining verið notuð þrátt fyrir að forsendur bresti. Nýrri aðferð sem nýtur vaxandi vinsælda er hlutfallsgreining (*logistic analysis*).

Við fáum matsmenn til að meta bragðgæði osts. Fylgibreytan er samtala bragðgæða yfir alla matsmennina. Ef bragðgæðin ná 37 á þessum kvarða, telst osturinn viðunandi.

Til samanburðar mælum við sýrustig ostsins. Við gerum ráð fyrir því að hærra mat tengist betri bragðgæðum.

Við viljum geta sagt til um hvort osturinn er nægjanlega góður út frá sýrustigi. Því búum við til nýja fylgibreytu (TasteOK) sem fær gildið 0 ef bragðgæðin eru undir 37 en 1 ef bragðgæðin er 37 eða hærra.

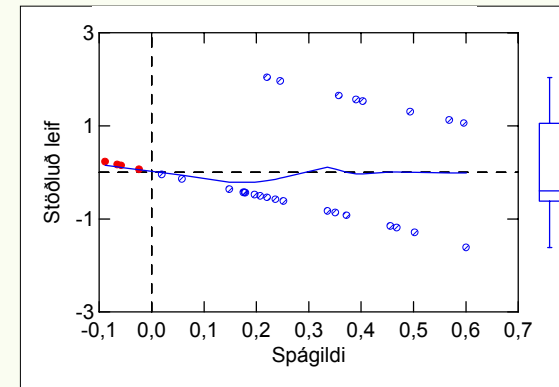
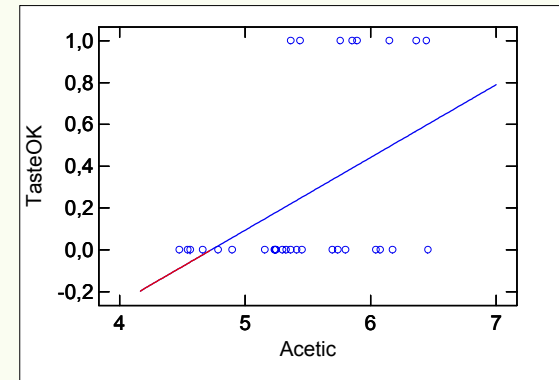
En hvernig vinnum við úr gögnum þar sem *fylgibreytan* er eigindleg og frumbreytan meginndleg.

Vandi línulegrar aðfallsgreiningar

Notkun línulegrar aðfallsgreiningar leiðir til tvenns konar vanda.

Efri myndin sýnir að fylgibreytan tekur aðeins tvö gildi. Spágildin má túlka sem það hlutfall osta sem myndu teljast nógu bragðgóðir. Línan sýnir að hlutfallið verður *neikvætt* fyrir ákveðið sýrustig.

Neðri myndin sýnir að leifin hefur afbrigðilega dreifingu. Því er hætt við því að öll marktektarpróf verði röng.



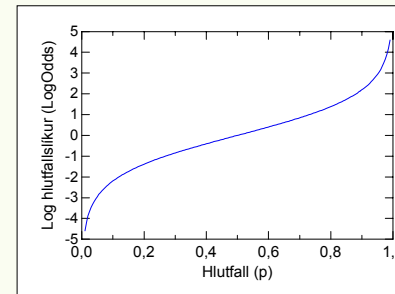
Hlutfallslíkur (*odds*)

Vandinn er sá að hlutföll takmarkast við talnabilið 0,0 til 1,0 og því erfitt að nota þau í aðfallsgreiningu.

Hlutföllum má hins vegar breyta í hlutfallslíkur, sem eru aðeins takmarkaðar í neðri endann; þær ná frá 0,0 upp í $+\infty$ (óendanlega hátt).

Lógariþminn af 0,0 er jafnt og $-\infty$ og lógariþminn af $+\infty$ er jafnt og $+\infty$. Því gefur lógariþmi hlutfallslíkinda okkur ótakmarkaða breytu.

$$Odds = \frac{\hat{p}}{1 - \hat{p}}$$
$$LogOdds = \ln\left(\frac{\hat{p}}{1 - \hat{p}}\right)$$



p	Odds	Log(Odds)	
0,00	–	0,0	$-\infty$
0,01	1:90	0,0	-4,60
0,10	1:9	0,1	-2,20
0,20	1:4	0,3	-1,39
0,30	3:7	0,4	-0,85
0,40	2:3	0,7	-0,41
0,50	1:1	1,0	0,00
0,60	3:2	1,5	0,41
0,70	7:3	2,3	0,85
0,80	4:1	4,0	1,39
0,90	9:1	9,0	2,20
0,95	19:1	19,0	2,94
0,96	24:1	24,0	3,18
0,97	97:3	32,3	3,48
0,98	49:1	49,0	3,89
0,99	99:1	99,0	4,60
1,00	–	∞	∞

Líkanið í aðfallsgreiningu hlutfalla

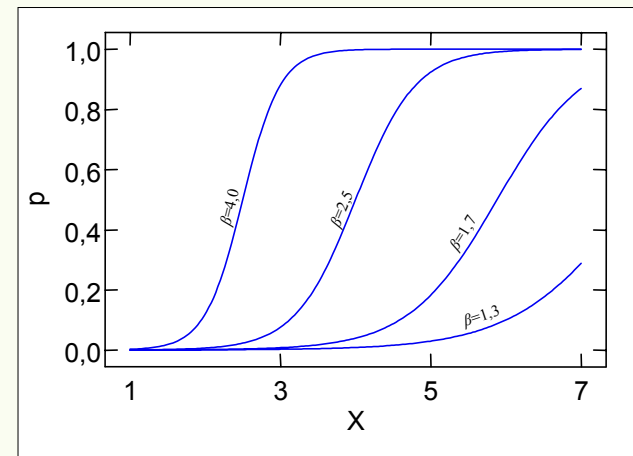
Líkanið er nákvæmlega eins og í línulegri aðfallsgreiningu en spágildið er lógariþmi hlutfallslíkinda.

Ef við breytum LogOdds í hlutföll, sjáum við að β_1 stjórnar því hversu hratt hlutfallið breytist. Því hærri sem hallastuðullinn er því meiri áhrif hefur frumbreytan á hlutfallið.

Fastinn β_0 og β_1 stjórna til samans hvar ferillinn er staðsettur, þ.e. hvenær ferillinn byrjar að færast úr 0 yfir í 1,0.

$$\text{LogOdds} = \beta_0 + \beta_1 x$$

$$\text{Odds} = \exp(\beta_0 + \beta_1 x)$$



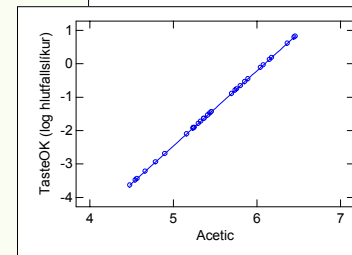
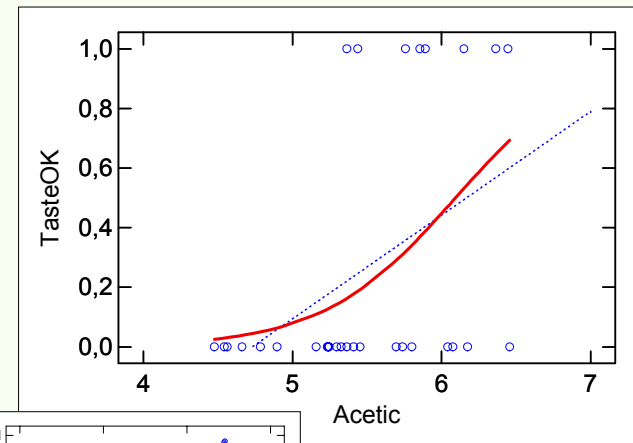
Tengsl sýru og bragðgæða

Textareiturinn sýnir mat á líkaninu fyrir bragðgæði osts. Hallatalan 2,25 tilgreinir áhrif sýrustigs á líkur þess að osturinn sé bragðgóður. Talan segir okkur ekki mikið en því hærri sem hún er því meiri áhrif.

Líkanið er línulegt, sbr. litlu myndina, þ.e. ef við vinnum í lógariþma.

Á stóru myndinni er líkanið á formi hlutfalla. Niðurstaðan er greinilega trúverðugri heldur en beina línan og er auk þess hvergi neikvæð.

$$\text{LogOdds} = -13,71 + 2,25 \times \text{Acetic}$$



$$p = \frac{\text{Odds}}{1 + \text{Odds}} = \frac{\exp(\text{LogOdds})}{1 + \exp(\text{LogOdds})}$$

Áhættuhlutfall (odds ratio)

Áhættuhlutfall leysir vandann við hallastuðlana. Í stað þess að túlka þá, athugum við hvað hlutfallslíkur aukast mikið við breytingu um eina einingu á frumbreytunni.

Hlutfallið 1,0 þýðir að engin breyting verði, lægra en 1,0 þýðir minnkun og hærra en 1,0 þýðir aukning.

Áhættuhlutfallið fyrir kynferði er 1,43 þegar athugað er hve oft er dottið í það. Það þýðir að hlutfallslíkurnar eru 43% hærri fyrir karla heldur en konur.

$$OR = \frac{Odds_1}{Odds_2}$$

$$OR = \exp(\beta_1)$$

Dottið í það eftir kyni

Af körlum detta 22,7% reglulega í það en 17,0% kvenna.

$$Odds_{kk} = \frac{0,227}{1 - 0,227} = 0,227 / 0,773 = 0,294$$

$$Odds_{kvk} = \frac{0,170}{1 - 0,170} = 0,170 / 0,830 = 0,205$$

$$OR = \frac{0,294}{0,205} = 1,43$$

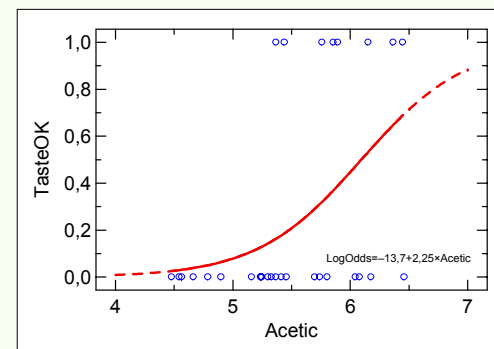
Áhrif sýrustigs á bragðgæði

Áhættuhlutfall sýrustigs er um 9,5. Það þýðir að hvar sem litið er á ferilinn, þá nífaldast hlutfallslíkur við hverja einingu sem sýrustigið eykst.

Ef litið er á hlutfallið, áttfaldast það milli 4 og 5, nær sexfaldast milli 5 og 6 og tæplega tvöfaldast milli 6 og 7.

Áhættuhlutfallið er hins vegar alls staðar það sama, þ.e. hlutfallslíkur rúmlega nífaldast við hverja einingu sem frumbreytan hækkar. Þetta er kosturinn við að nota áhættuhlutfall.

$$OR = \exp(2,249) \\ = 9,479 \approx 9,5$$



Acetic	LogOdds	Odds	p
1	-11,5	0,00	0,00
2	-9,2	0,00	0,00
3	-7,0	0,00	0,00
4	-4,7	0,01	0,01
5	-2,5	0,09	0,08
6	-0,2	0,81	0,45
7	2,0	7,68	0,88

Öryggisbil og próf Walds

Eins og í línulegri aðfallsgreiningu fáum við staðalvillu fyrir hallatöluna. Við getum prófað hvort hallatalan víki frá núlli með Wald-prófi.

Áhrifin túlkum við út frá hlutfalli hlutfallslíkinda (OR) en það táknar SPSS með Exp(B).

Öryggisbilið fyrir áhættuhlutfallið hjálpar okkur við að túlka niðurstöður nánar. Sams konar viðmið má nota og við önnur öryggisbil.

Í SPSS er Wald gefið upp sem kíkvaðrat!

	B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I. for EXP(B)	
							Lower	Upper
Sjæp 1	acetic	2,249	1,027	4,795	,029	9,479	1,266	70,959
	Constant	-13,705	5,932	5,338	,021	,000		

Wald - próf

$$z = \frac{b_1}{SE_{b_1}}$$

Wald-prófið er hliðstætt t -prófi og túlkast eins og það. Stundum—svo sem í SPSS—er það gefið upp miðað við kíkvaðratdreifingu en þá er kvaðratrótin jafnt og z .

Niðurstaðan gefur til kynna að Acetic hafi áhrif í þýði. Áhættuhlutfallið gefur til kynna að áhrifin sé mjög mikil.

Ef litið er á öryggisbilið sést þó að mikil óvissa er um stærð áhrifa. Hlutfallslíkur gætu verið að aukast um 27% fyrir hverja einingu á Acetic eða jafnvel sjötugfaldast fyrir hverja einingu.

Sennileikahlutfall (*likelihood ratio*)

Traustasta prófið á líkanið felst í sennileikahlutfallinu (*likelihood ratio*). Það er mælikvarði á það hversu sennilegt líkanið er miðað við gögnin.

Sennileikahlutfallið er gefið upp í lógariþma. Við athugum breytinguna við hvert skref í líkaninu og prófum hana tölfræðilega.

Ef munurinn er tvöfaldaður, dreifir hann sér sem kíkvaðrat. SPSS gefur sennileikann alltaf upp tvöfaldaðan.

Skef 0: Við setjum inn fastann β_0 .

LL= -17,397; 2LL= $2 \times -17,397 = -34,794$;
df=1 (SPSS gefur *ekki* þessar upplýsingar)

Skef 1: Við setjum inn β_1 .

LL= -14,113; 2LL= -28,227; df=2

Marktektarpróf: Við athugum breytinguna í 2LL við að bæta inn β_1 .

Breyting á 2LL: $34,794 - 28,227 = 6,568$

Breyting á df: $2 - 1 = 1$

Marktekt: $\chi^2(1) = 6,568$, $p \approx 0,01$ (Tafla F í ISP).

Það fékkst önnur niðurstaða með Wald-prófi en *sú* niðurstaða er almennt talið ónákvæmari heldur en prófun á sennileikahlutfallinu.

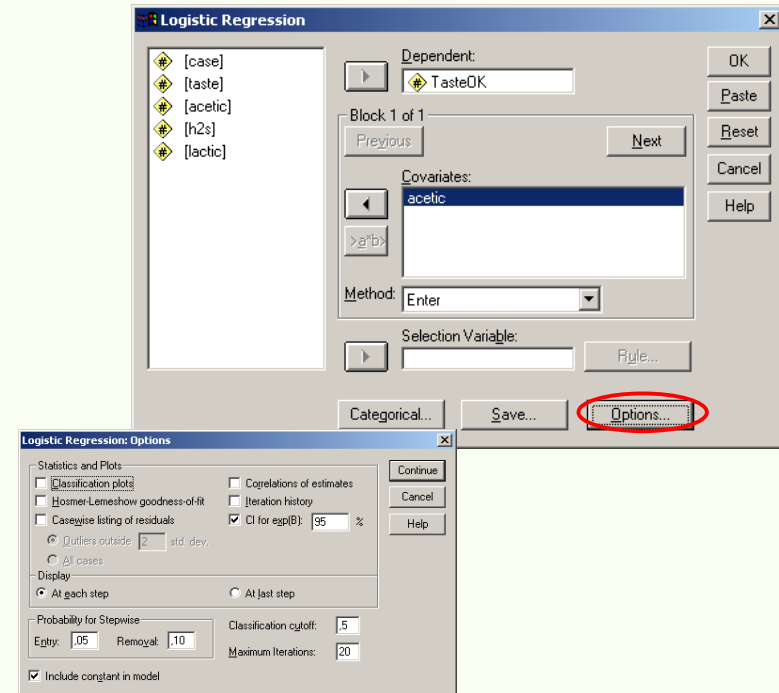
Aðfallsgreining hlutfalla í SPSS

Í SPSS förum við í Analyze / Regression / Binary Logistic....

Þá sprettur fram valglugginn hér til hliðar. Við færum fylgibreytuna í efsta textareitinn og frumbreytuna í neðri reitinn (*Covariates*).

Við smellum á Options og hökum þar við CI for exp(B).

Að síðustu smellum við á OK til að framkvæma úrvinnsluna eða Paste til að afrita yfir í skipanagluggann.



Niðurstöður í SPSS

Við skoðum aðeins **Block 1**. Fyrst kemur allsherjarpróf á líkanið. Þetta er kíkvaðratpróf á breytinguna í 2LL.

Síðan kemur yfirlit yfir líkanið. Þar er gefið upp $-2LL$ en almennt séð þurfum við lítið að huga að þessu.

Að lokum koma hallatölur, OR og öryggisbil fyrir breyturnar í líkaninu.

Við höfum rætt þetta á glærunum á undan; hér er ekkert nýtt.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	6,568	1	,010
	Block	6,568	1	,010
	Model	6,568	1	,010

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	28,227 ^a	,197	,286

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1	acetic	2,249	1,027	4,795	1	,029	9,479	1,266	70,959
	Constant	-13,705	5,932	5,338	1	,021	,000		

a. Variable(s) entered on step 1: acetic.

Vísibreytur (*indicator variables*)

Við getum haft eiginlegar breytur sem frumbreytur en þá þarf að kóða þær sem vísibreytur.

Tvískiptar breytur, eins og kynferði í Dottið í það gagnasafninu, má kóða sem 0 og 1.

Við látum konur hafa mæligildið 0 og karla fá 1 og köllum breytuna GenderM til að minna okkur á að hún er kóðuð fyrir karlkyn (**Male**).

Þessi tegund kóðunar nefnist staðgengilskóðun (*dummy coding*).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	GENDERM	0.362	0.039	86.611	1	0.000	1.435
	Constant	-1.587	0.027	3520.069	1	0.000	0.205

a Variable(s) entered on step 1: GENDERM

Figure 16-2a
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

95% öryggisbil fyrir OR

$$\exp(b_1 - 1,96 \times SE_{b_1}) \leq OR \leq \exp(b_1 + 1,96 \times SE_{b_1})$$

$$\exp(0,362 - 1,96 \times 0,039) \leq OR \leq \exp(0,362 + 0,076)$$

$$\exp(0,286) \leq OR \leq \exp(0,438)$$

$$1,33 \leq OR \leq 1,55$$

Niðurstaðan segir okkur að það sé kynjamunur á því að detta í það hjá háskólanemum. Að jafnaði er hlutfallslíkurnar um 43% hærrí fyrir karla heldur en konur.

Öryggisbilið gefur til kynna að þetta sé tiltölulega nákvæm niðurstaða þar sem OR liggur sennilega á mjög þröngu talnabili.

Túlkun staðgengilskóðunar

Staðgengilskóðaðar vísibreytur geta skapað vandamál í túlkun.

Mikilvægt er að átta sig á því að fastinn á við þegar vísibreytan tekur 0, í okkar tilviki ef þátttakandi er kona. $\text{GenderM} \times \beta_1$ verður 0 ef GenderM er 0.

Ef vísibreytan er 1, þátttakandinn er karl, bætist hallastuðullinn við fastann, þar sem $\text{GenderM} \times \beta_1 = \beta_1$ ef $\text{GenderM} = 1$.

$$\text{LogOdds} = -1,587 + 0,362 \times \text{GenderM}$$

Ef Karl þá $\text{GenderM} = 1$, en ef kona þá $\text{GenderM} = 0$.

Fyrir konur : $\text{LogOdds} = -1,587$

$$\text{Odds}_{\text{konur}} = \exp(-1,587) = 0,205 \text{ eða um } 1 : 5$$

Fyrir karla : $\text{LogOdds} = -1,587 + 0,362 \times 1$

$$= -1,587 + 0,362 = -1,225$$

$$\text{Odds}_{\text{karlar}} = \exp(-1,225) = 0,293 \text{ eða um } 1 : 3$$

Þetta verður flóknara ef vísibreytur eru margar. Þá gildir fastinn ef allar vísibreytur eru 0. Hægt er að fá út öll hlutfallslíkindi með því að leggja saman hallatölur í réttum samsetningum.

Ef megindleg frumbreyta er með vísibreytum, verður þetta enn flóknara þar sem þá þarf að taka tillit til þeirrar breytu, þ.e. hlutfallslíkindin breytast eftir því hvaða gildi hún tekur.