

# Einföld aðfallsgreining

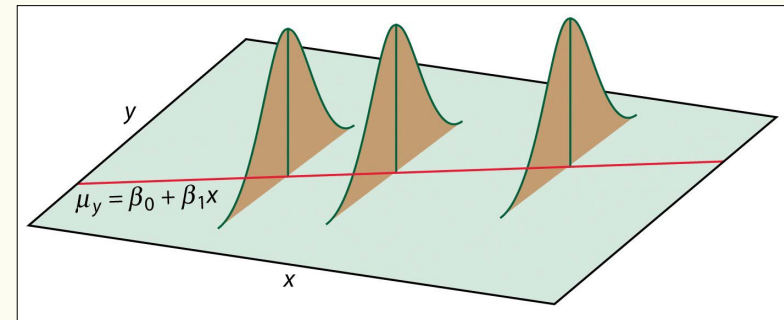
Fyrirlestur í Tölfræði II (SÁL203G)

# Einföld aðfallsgreining

Við höfum upplýsingar um tvær samfelldar breytur og viljum gera líkan af áhrifum frumbreytunnar á fylgibreytuna.

Við teljum að þegar frumbreytan hækkar um eina einingu muni fylgibreytan hækka (eða lækka) ákveðið mikið *að jafnaði*.

Við viljum vita hvernig tengslin eru, hversu vel tengslin lýsa gögnum og hversu vel er hægt að spá á grunni tengslanna.



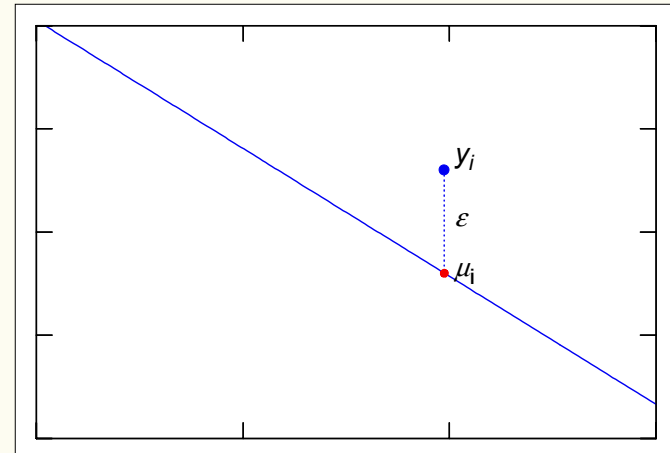
Myndin sýnir þrjú gildi frumbreytu, línu sem lýsir tengslunum og dreifingu fylgibreytu kringum línunnar.

Línan lýsir meðaltali fylgibreytu fyrir gildin þrjú á frumbreytunni. Normalferlarnir sýna dreifingu fylgibreytunnar í kringum meðaltölin.

# Aðfallsjafnan

Jafnan gefur okkur þýðismeðaltalið fyrir hvert gildi frumbreytunnar. Yfirleitt þekkjum við ekki jöfnuna nákvæmlega.

Gildi hvers einstaklings,  $y_i$ , er ólíkt þýðismeðaltalinu. Frávikið er villan sem við táknum með  $\varepsilon$ . Athugið að villan er ákveðin tala fyrir hvern einstakling, þ.e. frávik hans frá meðaltalinu.



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

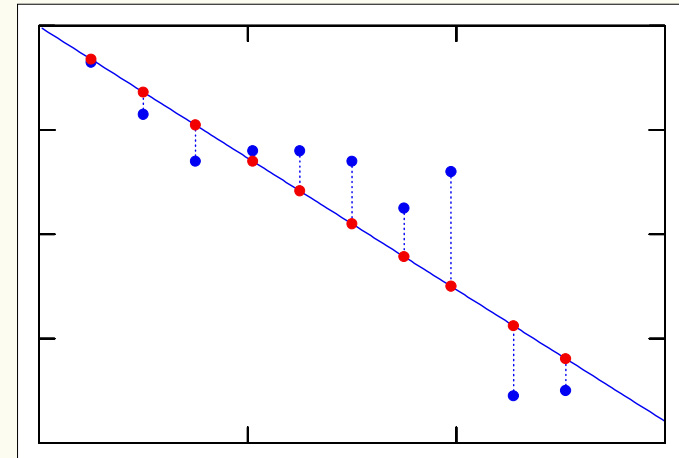
$$\mu_y = \beta_0 + \beta_1 x$$

# Aðfallsjafnan sem röð meðaltala

Hér eru nokkur þýðismeðaltöl sýnd með rauðum punktum. Aðfallslínan er í reynd röð slíkra meðaltala.

Bláu punktarnir eru gildin sem einstaklingarnir hafa. Þau víkja að jafnaði frá línunni. Frávikið er villan og táknuð með  $\varepsilon$ .

Línuleg aðfallsgreining gerir ráð fyrir að rauðu þýðismeðaltölin falli á beina línu, frávik bláu punktanna séu að jafnaði jafnstór alls staðar á línunni og normaldreifist.

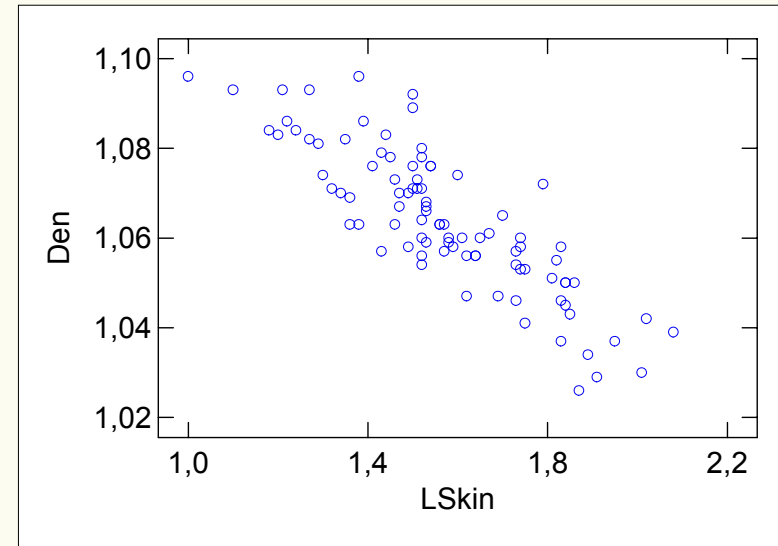


$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \mu_y + \varepsilon_i\end{aligned}$$

# Tengslin í gögnunum

Myndin sýnir tengsl eðlisþyngdar líkamans og þykktar fitufellinga. Hér vitum við ekki hver tengslin eru milli breytanna tveggja. Gögnin gefa hins vegar möguleika á að meta þau, þ.e. komast að einhverri niðurstöðu sem við höldum að lýsi tengslunum í þýði.

Við þekkjum ekki þýðismeðaltölin né villuna fyrir hvern einstakling. Í besta falli getum við fengið bestu spá fyrir jöfnuna og ákvarðað þannig spágildi einstaklingsins og leif hans.



Breytt tákun!

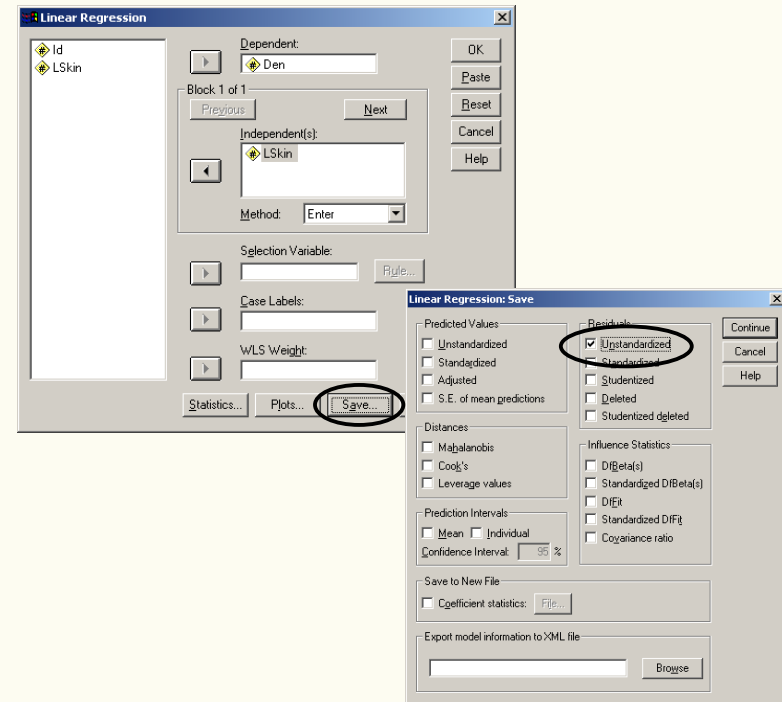
$$y_i = b_o + b_1 x_i + e_i$$

# Einhliða aðfallsgreining í SPSS

Ég fer í Analyze / Regression / Linear... og færi þar fylgibreytuna í Dependent-reitinn og frumbreytuna í Independent-listann.

Ef ég vil sjá leifin smelli ég á Save og haka þar við Unstandardized undir Residuals. Þá vistast leifin fyrir hvern þátttakanda sem sérstök breyta með nafn sem byrjar á Res\_, t.d. Res\_1.

Að lokum er smelt á OK.



# Mat á jöfnunni í SPSS

SPSS gefur margvíslegar upplýsingar um tengslin í úrtakinu. Við skoðum þessar töflur betur síðar.

Neðsta taflan gefur upp matið á línunni. Samkvæmt henni er jafna línunnar:  $y_i = 1,163 - 0,063 \cdot x_i + e_i$ .

Þessi niðurstaða er óviss og gæti verið töluvert ólík raunverulegum tengslum breytanna. Þetta er mat á tengslum breytanna, ekki tengslin eins og þau eru í raun og veru í þýðinu.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,849 <sup>a</sup>	,720	,717	,00854

a. Predictors: (Constant), LSkin

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,017	1	,017	231,894	,000 <sup>a</sup>
	Residual	,007	90	,000		
	Total	,023	91			

a. Predictors: (Constant), LSkin

b. Dependent Variable: Den

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,163	,007		177,296	,000
	LSkin	-,063	,004	-,849	-15,228	,000

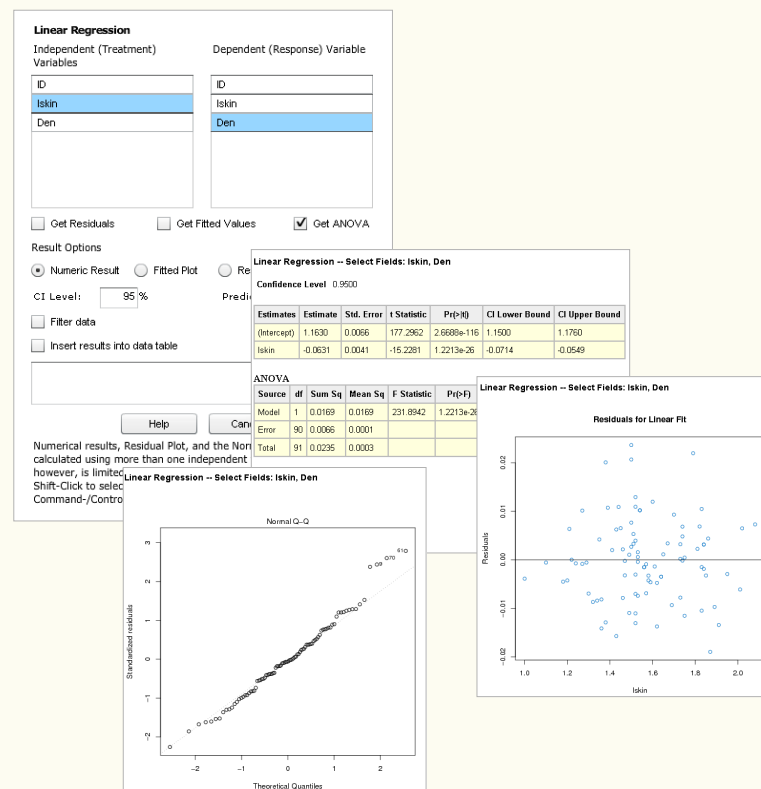
a. Dependent Variable: Den

# Mat á jöfnunni í CrunchIt

Ég opna Lskin og Den.txt, vel Statistics/ Regression/ Linear og vel frum- og fylgibreytu. Ég haka við Get ANOVA og vel Numeric Results og vel OK.

Til að fá leifarrit og normalrit þarf að fara aftur í valmyndina og velja þar, eftir atvikum, Residual Plot eða QQ Plot.

Normalritið snýr eins og í bókinni en er „upp á kant“ miðað við SPSS.





# Ólík form á jöfnunni

Ef við höfum allar upplýsingar í þýðinu getum við notað efstu tvær jöfnunnar. Þá getum við ýmist spáð fyrir um mæligildi hvers einstakling,  $y_i$ , eða reiknað þýðismeðaltalið sem samsvarar gildi frumbreytunnar.

Ef við höfum aðeins upplýsingar úr úrtaki vitum við ekki hvernig línan er né frávikið frá línuna, villuna, fyrir hvern einstakling. Við getum hins vegar fundið spágildi fyrir hvern einstakling og frávík hans frá því.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\mu_y = \beta_0 + \beta_1 x$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$y_i = b_0 + b_1 x_i + e_i$$

# Helstu tákni

Við viljum lýsa línunni í þýði en höfum aðeins spágildin fyrir fasta og hallatölu. Við notum því  $b_0$  og  $b_1$  sem mat okkar á samsvarandi þýðistölum.

Rétta línan í þýði lýsir meðaltalinu  $\mu_i$  fyrir hvert gildi frumbreytunnar. Við þekkjum hins vegar aðeins *reiknuðu* línuna sem gefur okkur spágildi línunnar.

Villan er frávik hvers einstaklings frá réttu línunni en við getum aðeins metið hana með leifinni  $e_i$ .

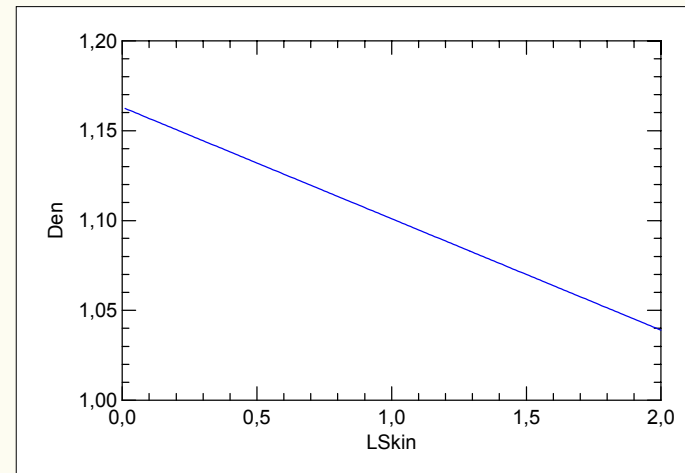
Lýsing	Þýðistala	Spátala
Fasti	$\beta_0$	$b_0$
Hallatala	$\beta_1$	$b_1$
Meðaltal	$\mu_y$	$\hat{\mu}_y$
Mæligildi	$y_i$	$\hat{y}_i$
Villa / leif	$\varepsilon_i$	$e_i$

# Túlkun hallastuðla

Hallatölurnar í töflunni gefa upp áhrif þykktar húðfellingar á eðlisþyngd líkamans, þ.e. hlutfall fitu.

Fastinn 1,163 segir að eðlisþyngdin sé 1,163 þegar frumbreytan er 0,0. Oftast er fastinn ekki að segja okkur mikið og því ekki túlkaður

Hallatalan  $-0,063$  fyrir LSkin segir okkur hver halli línunnar er. Hann gefur til kynna hvað eðlisþyngdin breytist mikið við hverja einingu sem frumbreytan breytist.



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,163	,007		177,296	,000
	LSkin	-,063	,004	-,849	-15,228	,000

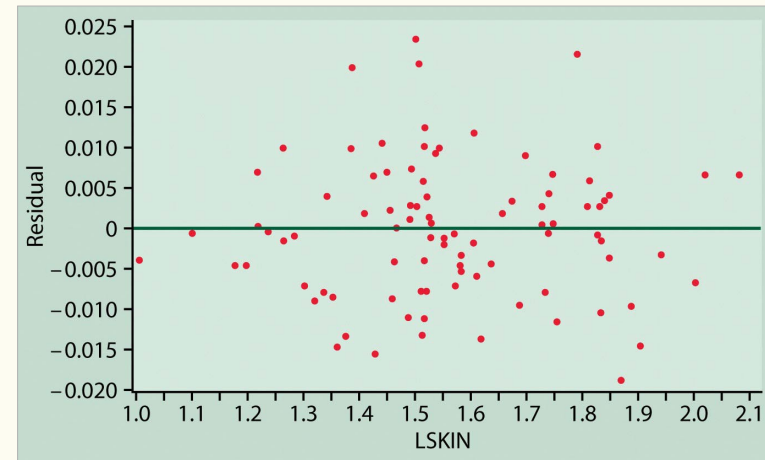
a. Dependent Variable: Den

# Mat á leifinni

Leifin er frávik hvers einstaklings frá reiknuðu línunni. Við vitum ekki hvar rétta línan liggur, getum því ekki reiknað  $\mu_i$  og vitum þar með ekki villuna fyrir hvern einstakling.

Leifin gefur okkar besta mat á villunni. Með því að skoða hana getum við ályktað um sennilegustu eiginleika villunnar.

Myndin sýnir að engin regla er í leifinni sem er í samræmi við að villan sé alls staðar jafnmikil.



$$\text{Leif: } e_i = y_i - \hat{y}_i$$

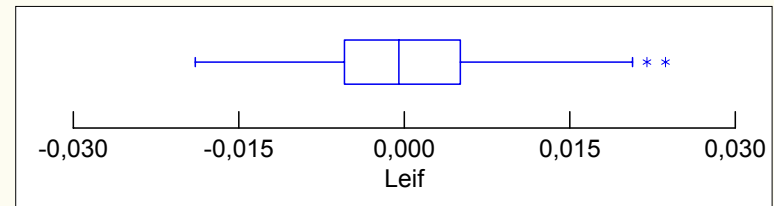
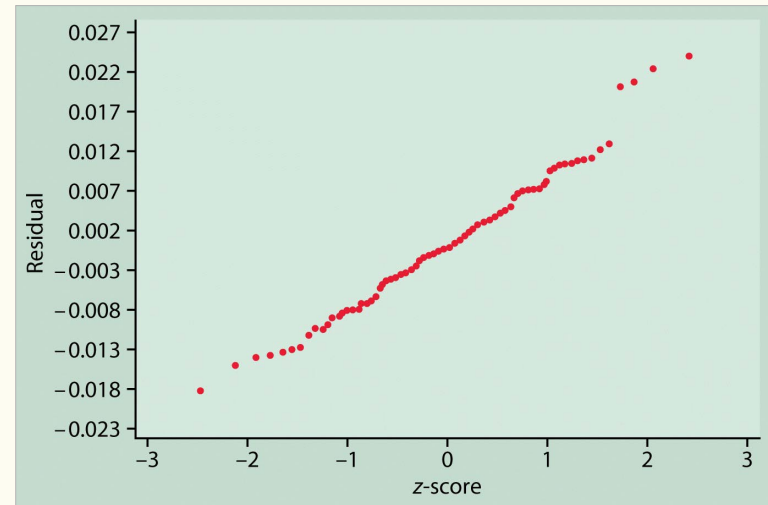
# Normalrit af leif

Aðfallsgreining miðast við að villan sé normaldreifð.

Normalrit af leif gefur færi á að meta þessa forsendu. Ef normalritið myndar því sem næst beina línu, gefur það til kynna normaldreifingu.

Kassaritið gefur svipaðar upplýsingar, takið eftir að engin stór fráviksgildi eru í leifinni.

Almennt séð þolir aðfallsgreining töluverð frávik frá normaldreifingu en illa frávillinga.



# Marktektarpróf fyrir hallatölur

Marktektarprófið í SPSS prófar hvort hallatalan sé 0,0 í þýðinu.

Venjulega höfum við ekki áhuga á fastanum. Hallatalan fyrir  $LSk_i$  metur hins vegar hvort það séu tengsl á milli þykktar húðfellinga og fitu mældrar sem eðlisþyngd líkamans.

Við getum reiknað prófið í höndunum en einfaldast er að trúa SPSS. Prófið er marktækt og því ályktum við að þykkt húðfellinga tengist eðlisþyngd, þ.e. fitu, líkamans.

Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1,163	,007		177,296	,000
	LSkin	-,063	,004	-,849	<b>-15,228</b>	<b>,000</b>

a. Dependent Variable: Den

$$t = \frac{b_1}{SE_{b_1}}$$
$$df = n - 2$$

# Öryggisbil fyrir hallatölur

Við getum séð óvissuna í hallatölunni með því að biðja um 95% öryggisbil.

Í okkar tilviki er hallatalan  $-0,063$  og staðalvillan  $0,004$ . Við getum ýmist reiknað öryggisbilið í höndunum með formúlunni eða beðið SPSS að gefa það upp. Oftast nægir að einfaldlega reikna í huganum tvær staðalvillur í hvora átt.

Samkvæmt öryggisbilinu er það trúverðugt að rétta hallatalan sé á bilinu  $-0,07$  til  $-0,06$ .

Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,163	,007		177,296	,000
	LSkin	-,063	,004	-,849	-15,228	,000

a. Dependent Variable: Den

Öryggisbil hallatölu

$$b_1 \pm t^* SE_{b_1}$$
$$df = n - 2$$

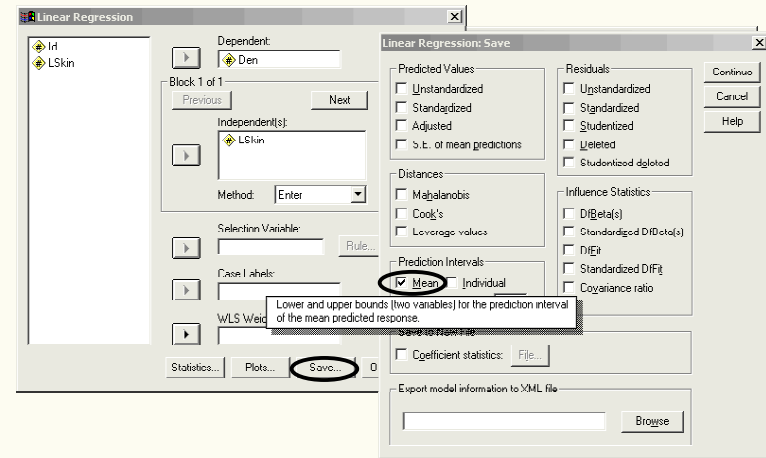
95% öryggisbil

$$-0,071 \leq \beta_1 \leq -0,055$$

# Óvissa um meðaltöl línunnar

Líta má á aðfallslínuna sem röð meðaltala. Öryggisbil þessara meðaltala gefur okkur upplýsingar um óvissuna í staðsetningu þeirra.

Við getum reiknað staðalvilluna með formúlunni og búið til öryggisbil fyrir nokkra staði á línunni. Við getum einnig beðið SPSS um að gefa efri og neðri mörk öryggisbilsins fyrir hvert spágildi í gögnunum. Þá birtast tvær nýjar breytur, LMCI\_1 og LUCI\_1 í gagnaglugganum með þessum upplýsingum.



$$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

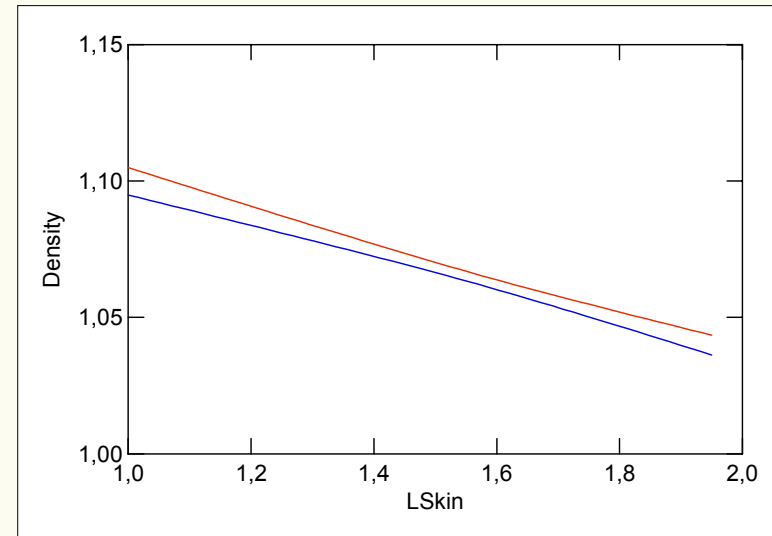


# Öryggisbil meðaltalanna

Myndin sýnir að efri og neðri öryggismörk meðaltalanna.

Nákvæmin er greinilega mest nálægt miðju frumbreytunnar og minnkar eftir þeim sem nær dregur jöðrum dreifingarinnar.

Þetta endurspeglar það að ef halli línunnar ( $b_1$ ) er metinn rangt, hefur það mest áhrif til endanna en fremur lítil á miðri línunni.



$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$$

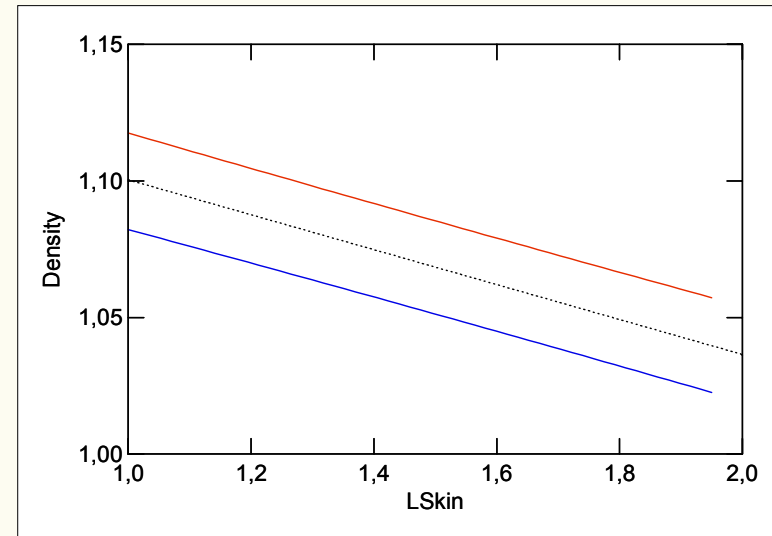
# Óvissa í forspá nýrra staka

Ef ég ætla að nota aðfallsjöfnuna til að segja fyrir um ný stök, þarf ég að reikna víðara öryggisbil.

Hér er tekið tillit til þess að það er ekki aðeins óvissa um meðaltal línunnar heldur einnig um það hversu nálægt línunni nýja gildið lendir.

Þessi víðu öryggisbil þýða að oftast er erfitt að nota aðfallsjöfnuna í forspá sökum ónákvæmni í forspánni.

Formúlan er eilítið breytt en einnig er hægt að láta SPSS reikna þetta.



$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

# Sundurgreining mæligilda

Mæligildi hvers einstaklings ( $y_i$ ) má skipta niður í spágildi og leif.

Spágildið fæst með aðfallsjöfnunni og er mat okkar á viðkomandi þýðistölu. Leifin endurspeglar einstaklingsmun og er mat okkar á villunni í þýði.

Breytileiki mæligilda skiptist á sama hátt í breytileika spágildanna og breytileika leifarinnar. Þannig má sundurgreina breytileikann á sama hátt og mæligildi hvers og eins einstaklings (staks).

$$\text{Mæligildi} = \text{Spágildi} + \text{Leif}$$

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) = \hat{y}_i + e_i$$

$$\text{Var}(y) = \text{Var}(\hat{y}) + \text{Var}(e)$$

# Summa kvaðrata

Summa kvaðrata er mælikvarði á breytileika mæligilda, spágilda og leifar.

Heildarsumma kvaðrata ( $SS_T$ ) gefur breytileika allra mæligildanna og samsvarar dreifitölu fylgibreytunnar. Summa kvaðrata fyrir aðfallslíkanið ( $SS_M$ ) gefur breytileika spátalnanna en summa kvaðrata leifarinnar ( $SS_E$ ) gefur breytileika leifarinnar.

Skiptingin samsvarar reglunni, því sem við getum skýrt, og ruddanum, því sem aðfallslínan skýrir ekki.

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,017	1	,017	231,894	,000(a)
	Residual	,007	90	,000		
	Total	,023	91			

a Predictors: (Constant), LSkin  
b Dependent Variable: Den

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SS_T = SS_M + SS_E$$

$$\text{Gögn} = \text{Regla} + \text{Ruddi}$$

Regla: Hvernig gögnin skipast niður, þ.e. mynstrið í þeim.  
Ruddi: Lélegt hrakið hey, þ.e. „draslið“ í gögnunum.

# Meðalsumma kvaðrata

Ef deilt er í summu kvaðrata með samsvarandi frígráðum, fæst meðalsumma kvaðrata.

Meðalsumma kvaðrata leifarinnar ( $MS_E$ ) samsvarar dreifitölu hennar og metur því breytileika villunnar ( $\sigma^2$ ).

Ef hallatalan í þýði ( $\beta_1$ ) er 0,0, metur meðalsumma kvaðrata fyrir líkanið ( $MS_M$ ) einnig breytileika villunnar. Ef hins vegar hallatalan er ekki 0 ( $\beta_1 \neq 0$ ), metur  $MS_M$  stærð sem er meiri en breytileiki villunnar ( $\sigma^2$ ).

ANOVA <sup>b</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,017	1	,017	231,894	,000(a)
	Residual	,007	90	,000		
	Total	,023	91			

a Predictors: (Constant), LSkin  
b Dependent Variable: Den

Í okkar tilviki er  $MS_M$  mun stærra en  $MS_E$ . Við spyrjum því hvort kvaðratsummurnar séu báðar að meta  $\sigma^2$  en fyrir tilviljun verði  $MS_M$  miklu hærri. Hinn möguleikinn er að  $\beta_1$  sé ekki 0 og sé því að meta meiri stærð heldur en  $MS_E$ .

# F-prófið

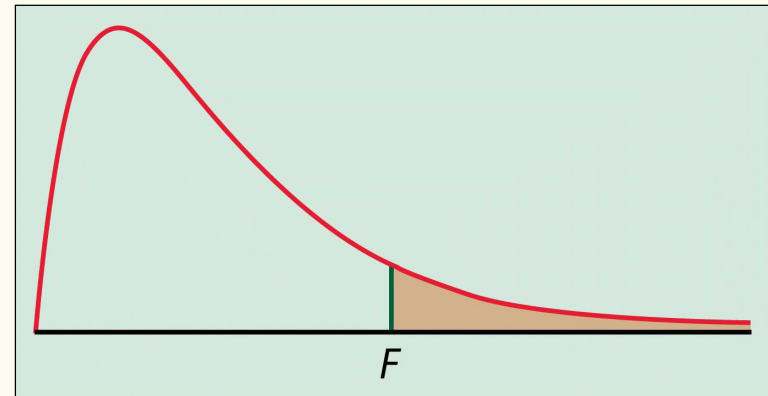
$F$ -prófið ber saman  $MS_M$  og  $MS_E$  og metur hvort munurinn sé líklegur ef hallatalan væri 0,0.

Myndin sýnir hvernig niðurstaða prófsins dreifist undir núlltilgátunni  $H_0: \beta_I=0,0$ .

Í okkar tilviki er  $F$  það hátt og við því komin það langt til hægri á myndinni að líkurnar á þetta hárra eða hærri niðurstöðu ef  $H_0$  er rétt er minni en 0,1%. Við höfnum því núlltilgátunni og ályktum að þykkt húðfellinga tengist eðlisþyngd líkamans.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,017	1	,017	<b>231,894</b>	<b>,000<sup>a</sup></b>
	Residual	,007	90	,000		
	Total	,023	91			

a Predictors: (Constant), LSKin  
b Dependent Variable: Den



# Forspárhæfni

Ef við viljum nota líkanið í forspá, þarf það að hafa nægjanlega forspárhæfni.  $R^2$  er hlutfall  $SS_M$  og  $SS_T$  og gefur upp hvað frumbreytan skýrir af dreifingu fylgibreytu. Í okkar tilviki skýrast 72% af breytileika eðlisþyngdar af þykkt húðfellinga.

Staðalvilla spágildis sýnir dreifingu mæligilda í kringum línuna. Hún metur því forspárhæfnina á kvarða fylgibreytunnar. Athugið að hún gefur *ekki* nákvæmni í forspá nýrra staka.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,849 <sup>a</sup>	,720	,717	,00854

a Predictors: (Constant), LSkin  
b Dependent Variable: Den

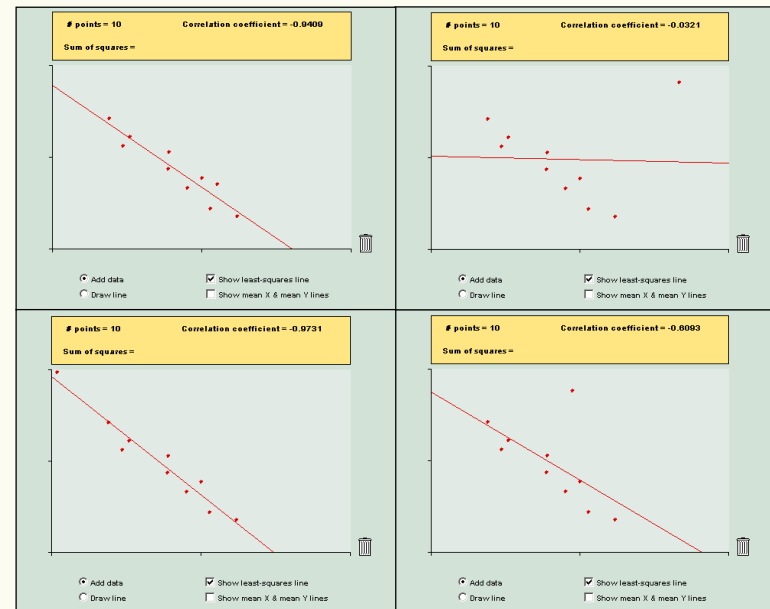
Það eru engin algild viðmið um stærð  $R^2$ . Í okkar tilviki er valkosturinn að sökkva fólki niður í vatnstank og meta eðlisþyngd þannig. Að geta skýrt 72% af niðurstöðunni með einfaldri mælingu er því alveg frábær forspárhæfni. Í sumum tilvikum gæti þetta jafnvel verið of lítil forspárhæfni.

Sumt er þess eðlis að forspá er erfið og  $R^2$  því mjög lágt. Ef markmiðið er ekki forspá, kemur það ekki að sök.

# Áhrif frávillinga

Leifarrit, normalrit eða kassarit af leif geta gefið til kynna eitt eða fleiri fráviksgildi. Ef fráviksgildi eru áberandi, geta þau haft áhrif á niðurstöðuna; jafnvel eitt fráviksgildi getur gerbreytt niðurstöðum.

Myndin efst til vinstri sýnir tengsl án fráviksgilda. Myndin efst til hægri er með frávilling sem gjörbreytir halla línunnar. Neðri myndirnar tvær hafa frávilling sem hefur lítil áhrif á línuna en hefur áhrif á ýmist  $MS_E$  eða  $R^2$ .



[bcs.whfreeman.com/ips6e/content/cat\\_010/applets/CorrelationRegression.html](http://bcs.whfreeman.com/ips6e/content/cat_010/applets/CorrelationRegression.html)