Línuleg aðhvarfsgreining

Tölfræði frá grunni

Anna Helga Jónsdóttir og Sigrún Helga Lund

Háskóli Íslands

HÁSKÓLI ÍSLANDS

Helstu atriði:

1. Einfalt línulegt aðhvarf - hluti 1
2. Einfalt línulegt aðhvarf - hluti 2
3. Einfalt línulegt aðhvarf - hluti 3
4. Ályktanir í aðhvarfsgreiningu
5. Fervikagreining í aðhvarfsgreiningu

Hvert erum við komin...

1. Einfalt línulegt aðhvarf - hluti 1
2. Einfalt línulegt aðhvarf - hluti 2
3. Einfalt línulegt aðhvarf - hluti 3
4. Ályktanir í aðhvarfsgreiningu
5. Fervikagreining í aðhvarfsgreiningu

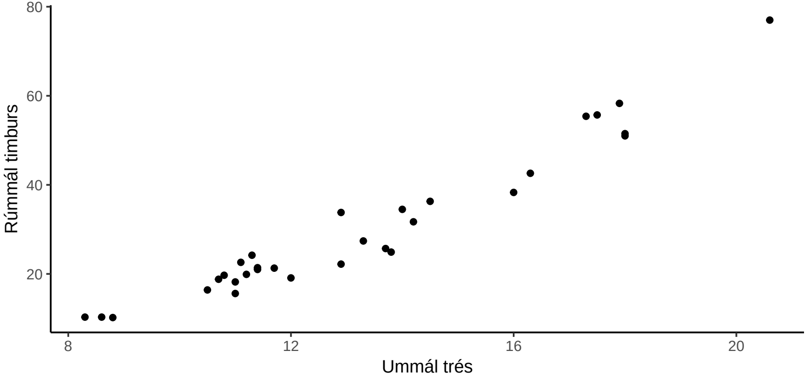
Línuleg aðhvarfsgreining

- Línuleg aðhvarfsgreining er ein algengasta tölfræðiaðferð sem notuð er til að kanna samband tveggja talnabreyta.
- Við gerum ráð fyrir að lýsa megi sambandi breytanna með jöfnu beinnar línu.
- Línuleg aðhvarfsgreining er sér í lagi mikið notuð þegar önnur breytan er **svarbreyta** og hin **skýribreyta**.
- Útvíkka má línulega aðhvarfsgreiningu þannig að nota megi fleiri en eina skýribreytu, sem geta hvort heldur sem er verið flokkabreytur eða talnabreytur.
- Við ætlum þó bara að skoða tilvikið þegar það er einungis ein talnabreyta sem skýribreyta.

Línuleg aðhvarfsgreining

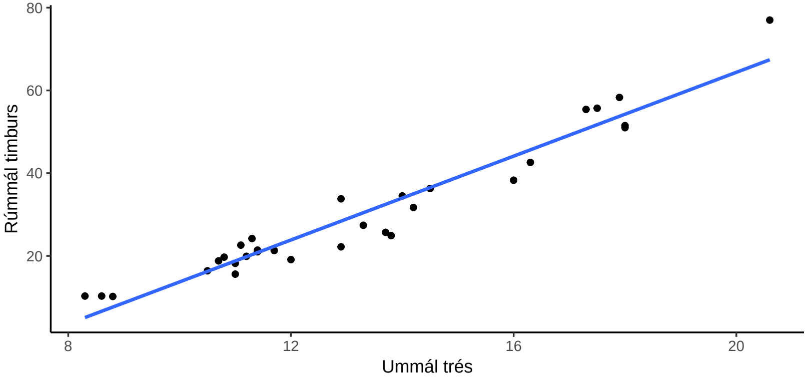
- Í einföldu línulegu aðhvarfi göngum við út frá því að það sé engin óvissa í mælingum okkar á skýribreytunni.
- Þetta er oft orðað þannig að skýribreytan sé **föst**.
- Svarbreytan er hins vegar slembin og er dreifing hennar er háð skýribreytunni.
- Við notum Y til að tákna svarbreytuna og x til að tákna skýribreytuna.

Samband ummáls trés og rúmmáls timbursins sem það gaf



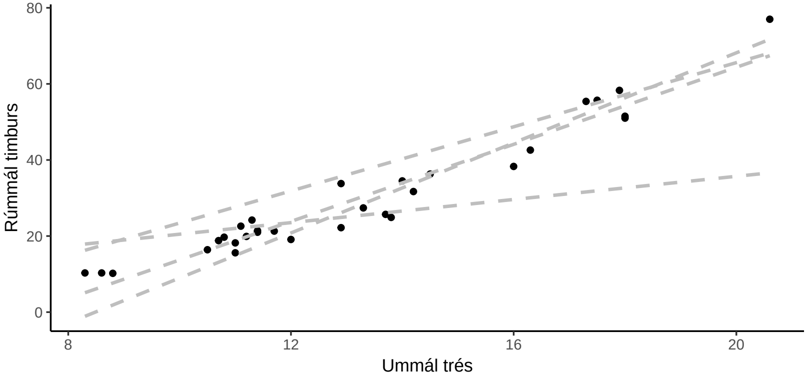
Fyrsta skref í línulegri aðhvarfsgreiningu ætti alltaf að vera að **teikna gögnin**.

Er hægt að lýsa sambandinu með jöfnu beinnar línu?



Er línulegt samband á milli ummáls trés og rúmmáls þess timbur sem það gefur af sér?

Hvernig veljum við línuna?



Aðhvarfsgreiningarlíkanið

Aðhvarfsgreiningarlíkanið

Einfalda aðhvarfsgreiningarlíkanið (simple linear regression model) má skrifa sem

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

þar sem β_0 og β_1 eru óþekktir stikar og ε er normaldreifð slembistærð með meðaltal 0.

Markmið einfalds línulegs aðhvarfs er fyrst og fremst að meta stuðlana β_0 og β_1 með mælingunum á breytunum tveimur, x og Y.

Aðferðin sem við notum kallast aðferð minnstu kvaðrata.

Jafna aðhvarfslínu minnstu kvaðrata

Jafna aðhvarfslínu minnstu kvaðrata

Táknum meðaltal og staðalfrávik x breytunnar með $\bar{x}$ og s_x og y breytunnar með $\bar{y}$ og s_y og fylgnina á milli þeirra með r.

Við metum hallatölu línunnar með

$$b_1 = r\frac{s_y}{s_x}$$

og skurðpunktinn með

$$b_0 = \bar{y} - b_1\bar{x}.$$

Þá er jafna aðhvarfslínu minnstu kvaðrata

$$\hat{y} = b_0 + b_1 x$$

Aðferð minnstu kvaðrata - dæmi

Smíðum aðhvarfslíkan til að meta hversu mikið rúmmál timburs færst af tré með tiltekið ummál. Gefum okkur að eftirfarandi lýsistærðir hafi verið reiknaðar út:

Meðaltal ummáls = 13.25 tommur
Staðalfrávik ummáls = 3.18 tommur
Meðaltal rúmmáls = 30.17 tommur
Staðalfrávik rúmmáls = 16.48 tommur
Fylgni ummáls og rúmmáls = 0.97 tommur.

Aðferð minnstu kvaðrata - dæmi

Smíðum aðhvarfslíkan til að meta hversu mikið rúmmál timburs færst af tré með tiltekið ummál. Gefum okkur að eftirfarandi lýsistærðir hafi verið reiknaðar út:

Meðaltal ummáls = 13.25 tommur
Staðalfrávik ummáls = 3.18 tommur
Meðaltal rúmmáls = 30.17 tommur
Staðalfrávik rúmmáls = 16.48 tommur
Fylgni ummáls og rúmmáls = 0.97 tommur.

$b_1 = r\frac{s_y}{s_x} =$

Aðferð minnstu kvaðrata - dæmi

Smíðum aðhvarfslíkan til að meta hversu mikið rúmmál timburs færst af tré með tiltekið ummál. Gefum okkur að eftirfarandi lýsistærðir hafi verið reiknaðar út:

Meðaltal ummáls = 13.25 tommur
Staðalfrávik ummáls = 3.18 tommur
Meðaltal rúmmáls = 30.17 tommur
Staðalfrávik rúmmáls = 16.48 tommur
Fylgni ummáls og rúmmáls = 0.97 tommur.

$$b_1 = r\frac{s_y}{s_x} = 0.97 \cdot \frac{16.48}{3.18} = 5.03$$

Aðferð minnstu kvaðrata - dæmi

Smíðum aðhvarfslíkan til að meta hversu mikið rúmmál timburs færst af tré með tiltekið ummál. Gefum okkur að eftirfarandi lýsistærðir hafi verið reiknaðar út:

Meðaltal ummáls = 13.25 tommur
Staðalfrávik ummáls = 3.18 tommur
Meðaltal rúmmáls = 30.17 tommur
Staðalfrávik rúmmáls = 16.48 tommur
Fylgni ummáls og rúmmáls = 0.97 tommur.

$$b_1 = r\frac{s_y}{s_x} = 0.97 \cdot \frac{16.48}{3.18} = 5.03$$
$$b_0 = \bar{y} - b_1\bar{x} =$$

Aðferð minnstu kvaðrata - dæmi

Smíðum aðhvarfslíkan til að meta hversu mikið rúmmál timburs færst af tré með tiltekið ummál. Gefum okkur að eftirfarandi lýsistærðir hafi verið reiknaðar út:

Meðaltal ummáls = 13.25 tommur
Staðalfrávik ummáls = 3.18 tommur
Meðaltal rúmmáls = 30.17 tommur
Staðalfrávik rúmmáls = 16.48 tommur
Fylgni ummáls og rúmmáls = 0.97 tommur.

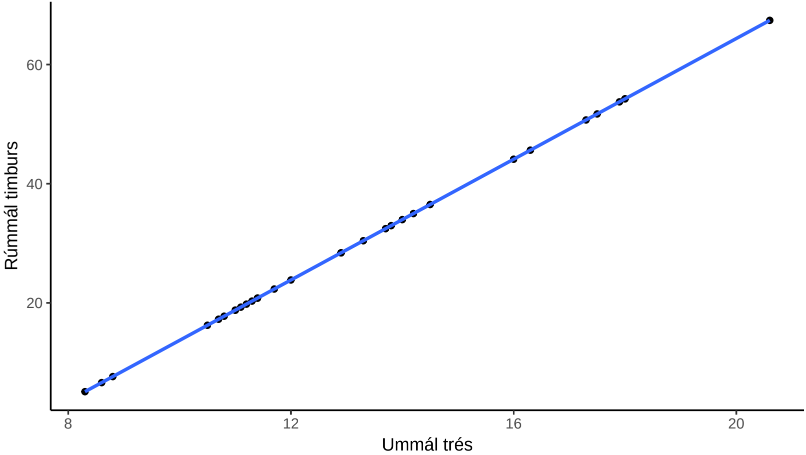
$$b_1 = r\frac{s_y}{s_x} = 0.97 \cdot \frac{16.48}{3.18} = 5.03$$
$$b_0 = \bar{y} - b_1\bar{x} = 30.17 - 5.03 \cdot 13.25 = -36.48$$

Þá er jafnan $\hat{y} = -36.48 + 5.03 \cdot x$

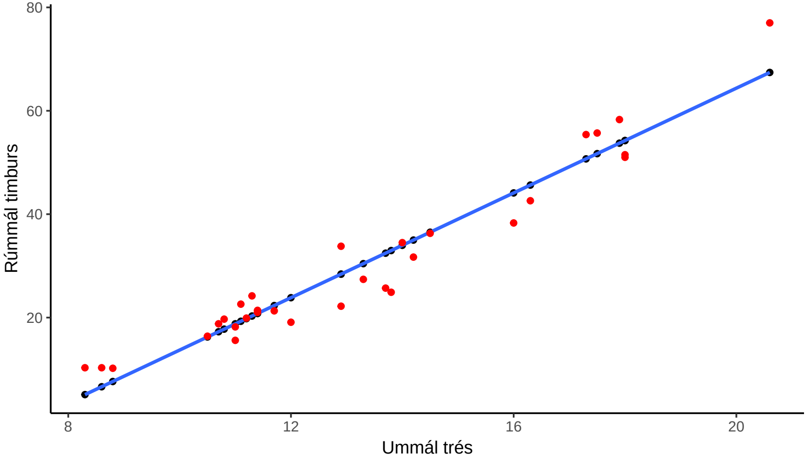
Hvert erum við komin...

1. Einfalt línulegt aðhvarf - hluti 1
2. Einfalt línulegt aðhvarf - hluti 2
3. Einfalt línulegt aðhvarf - hluti 3
4. Ályktanir í aðhvarfsgreiningu
5. Fervikagreining í aðhvarfsgreiningu

Ef það væri engin óvissa...



Raunveruleikinn!



Leifar

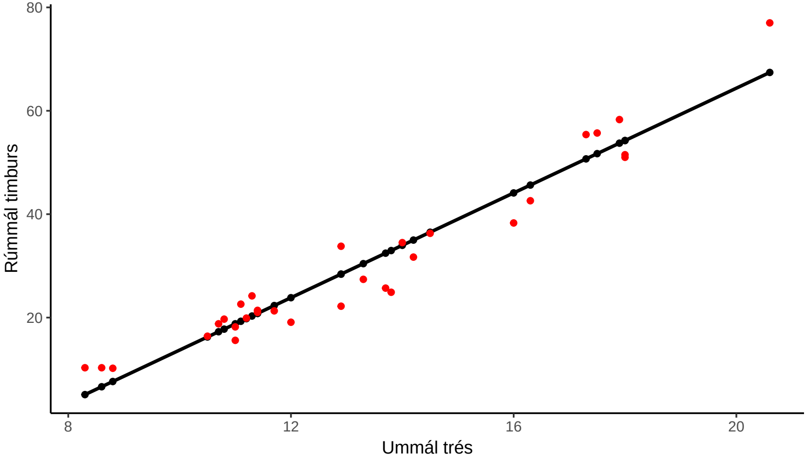
Leifar

Lóðrétta fjarlægðin frá mælingunum okkar að aðhvarfslínunni köllum við leifar og táknum með e. Stærð leifa má reikna með

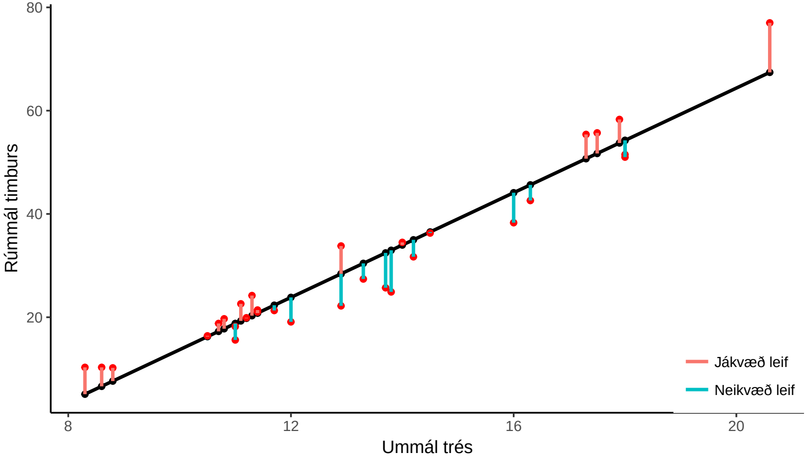
$$e_i = y_i - \hat{y}_i.$$

Punktar ofan aðhvarfslínunnar hafa jákvæða leif en punktar neðan hennar neikvæða.

Leifar

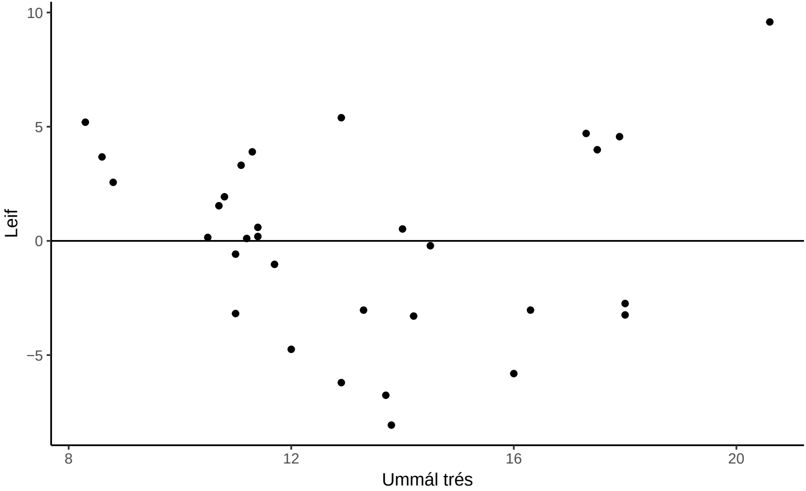


Leifar


80
60
40
20
Rúmmál timburs
8
12
16
20
Ummál trés
Jákvæð leif
Neikvæð leif

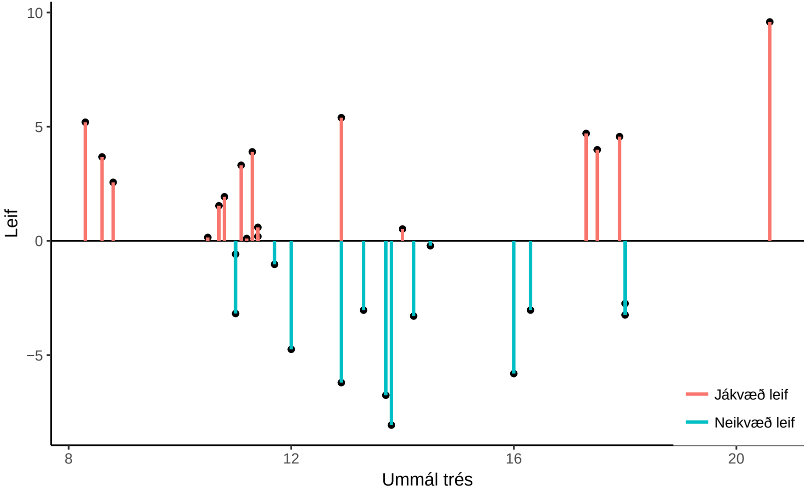
Leifarit

Leifarit sýnir leifarnar á y-ás og skýribreytuna á x-ásnum.



Leifarit

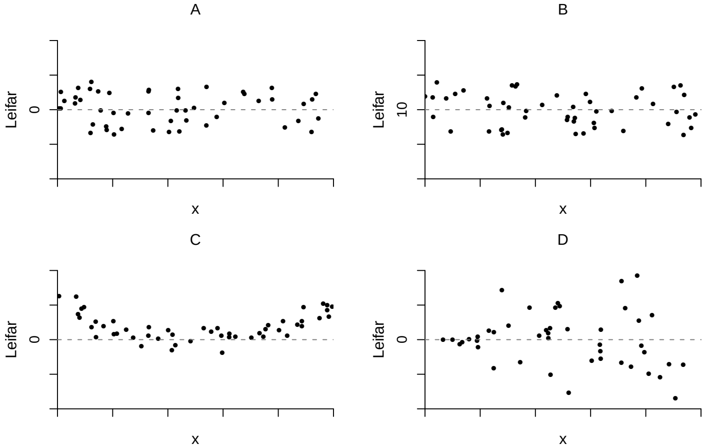
Leifarit sýnir leifarnar á y-ás og skýribreytuna á x-ásnum.



Forsendur líkansins

- Þegar við framkvæmum línulega aðhvarfsgreiningu gerum við ráð fyrir því að leifar líkansins séu **óháðar** og fylgi sömu **normaldreifingunni**.
- Með leifariti getum við kannað hvort leifarnar virðast óháðar (hvort það virðist ekki vera nein regla í þeim).
- Þá reynum við að passa að það sé **ekkert mynstur** í leifunum og að **breytileikinn sé óháður skýribreytunni**.
- Með **normaldreifingarriti** (kafla 5.4.1) getum við svo kannað hvort að dreifing leifanna svipi til normaldreifingar.

Engin regla á að vera í leifariti!



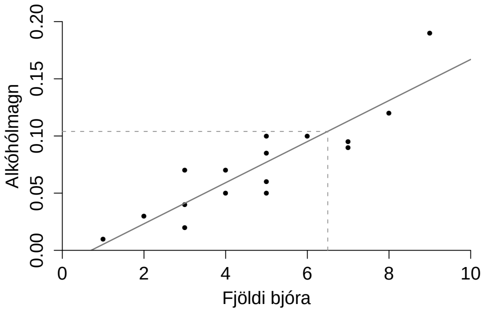
Hvert erum við komin...

1. Einfalt línulegt aðhvarf - hluti 1
2. Einfalt línulegt aðhvarf - hluti 2
3. Einfalt línulegt aðhvarf - hluti 3
4. Ályktanir í aðhvarfsgreiningu
5. Fervikagreining í aðhvarfsgreiningu

Brúun

Brúun

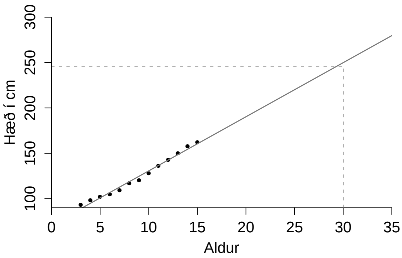
Sé aðhvarfslíkan notuð til að spá fyrir um gildi á Y fyrir eitthvert gildi á x sem er á sama reki og x-gildin sem notuð voru til að meta stikana í líkaninu er talað um að *brúa* (interpolate).



Bryggjun

Bryggjun

Sé aðhvarfslíkan notað til að spá fyrir um gildi á Y fyrir eitthvert gildi á x sem er **fjarri** þeim x-gildum sem notuð voru til að meta stikana í líkaninu er talað um að *bryggja* (exrapolate). Þetta svarar til að lengja aðhvarfslínuna. Það getur verið mjög vafasamt að bryggja!



Skýringarhlutfall

Skýringarhlutfall í einföldu línulegu aðhvarfi

Sé fylgnistuðullinn settur í annað veldi, r^2, er talað um skýringarhlutfall. Oft er notaður stór bókstafur fyrir skýringarhlutfallið, þ.e. R^2.

r^2 stendur fyrir hlutfallslegan breytileika í Y sem er hægt að skýra með breytingum á x.

r^2 getur í minnsta lagi verið 0, en í mesta lagi 1.

Skýringarhlutfall - (uppspunnið) dæmi

Hugsum okkur að það sé línulegt orsakasamband milli BMI stuðuls og blóðþrýstings.

- Hvor breytan ætti að vera x- breytan og hvor ætti að vera Y-breytan?

Skýringarhlutfall - (uppspunnið) dæmi

Hugsum okkur að það sé línulegt orsakasamband milli BMI stuðuls og blóðþrýstings.

- Hvor breytan ætti að vera x- breytan og hvor ætti að vera Y-breytan?

  BMI stuðull ætti að vera skýribreytan (x-breytan) og blóðþrýstingur ætti að vera svarbreytan (Y-breytan).

Skýringarhlutfall - (uppspunnið) dæmi

Hugsum okkur að það sé línulegt orsakasamband milli BMI stuðuls og blóðþrýstings.

- Hvor breytan ætti að vera x- breytan og hvor ætti að vera Y-breytan?

  BMI stuðull ætti að vera skýribreytan (x-breytan) og blóðþrýstingur ætti að vera svarbreytan (Y-breytan).

- Fylgnin mældist 0.8. Hversu hátt hlutfall breytileika í blóðþrýstingi milli einstaklinga má þá útskýra með ólíku BMI?

Skýringarhlutfall - (uppspunnið) dæmi

Hugsum okkur að það sé línulegt orsakasamband milli BMI stuðuls og blóðþrýstings.

- Hvor breytan ætti að vera x- breytan og hvor ætti að vera Y-breytan?

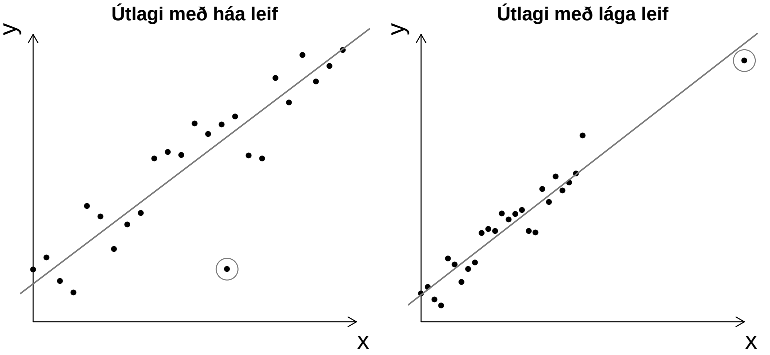
  BMI stuðull ætti að vera skýribreytan (x-breytan) og blóðþrýstingur ætti að vera svarbreytan (Y-breytan).

- Fylgnin mældist 0.8. Hversu hátt hlutfall breytileika í blóðþrýstingi milli einstaklinga má þá útskýra með ólíku BMI?

  Skýringarhlutfallið er $0.8^2 = 0.64$.

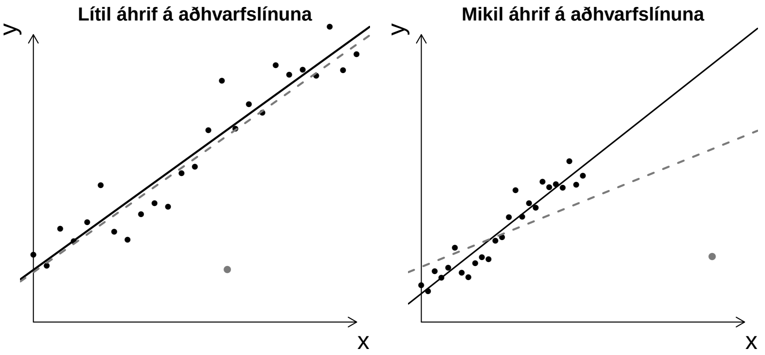
  Það þýðir að 64% breytileika í blóðþrýstingi mátti útskýra með breytileika í BMI.

Útlagar og áhrifamikil mæligildi



Mynd: Útlagar og leifar þeirra.

Útlagar og áhrifamikil mæligildi



Mynd: Áhrifamikil mæligildi.

Meðhöndlun útlaga og áhrifamikilla mæligilda

- Það á alltaf að skoða útlaga og áhrifamikil mæligildi sérstaklega.
- Ef mistök hafa átt sér stað skal fjarlæga mæligildið úr safninu.
- Ef ekki er hægt að sýna fram á að um mistök hafi verið að ræða er oft gott að sýna útreikninga með og án þessara gilda.
- Í sumum tilfellum er eðlilegast að byggja útreikninga á mælisafninu án útlaga/áhrifamikilla mæligilda.
- Í þeim tilfellum verður að taka fram að líkanið gildi ekki fyrir mæligildi utan þess ramma mæligilda sem notuð voru við gerð líkansins.

Hvert erum við komin...

1. Einfalt línulegt aðhvarf - hluti 1
2. Einfalt línulegt aðhvarf - hluti 2
3. Einfalt línulegt aðhvarf - hluti 3
4. Ályktanir í aðhvarfsgreiningu
5. Fervikagreining í aðhvarfsgreiningu

Ályktanir í aðhvarfsgreiningu

- Nú er kominn tími til að setja aðhvarfsgreiningu í samhengi við það sem við höfum áður lært um ályktunartölfræði.
- Mælingarnar okkar eru einhverri slembni háðar og því getum við fengið aðrar niðurstöður ef við endurtökum tilraunina og þar af leiðandi annað mat á aðhvarfslínunni okkar.
- Því er eðlilegt að reikna öryggisbil og tilgátupróf í einföldu línulegu aðhvarfi líkt og þið hafið séð svo mörg dæmi um hingað til.

Aðhvarfsgreiningarlíkanið

Ef við gerum ráð fyrir að við höfum n paraðar mælingar $(x_1, y_1), \ldots, (x_n, y_n)$, má skrifa líkanið sem

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- β_0 er hinn sanni skurðpunktur sem við ekki þekkjum, þýðisskurðpunkturinn.
- β_1 hin sanna hallatala, þýðishallatalan.
- ϵ_i eru frávikin.

β_0 og β_1 eru því lýsistærðir, sem við viljum bæði meta og draga ályktanir um.

Það gerum við með því að beita aðferð minnstu kvaðrata á gögnin okkar.

Slembistærðin ε

ε til að lýsir þeirri óvissu sem er til staðar í mælingum okkar á Y.

Við gerum ráð fyrir að ε_i séu einsdreifðar óháðar slembistærðir sem fylgja normaldreifingu með meðaltal 0 og dreifni σ^2.

Mat á σ^2 í einföldu línulegu aðhvarfi

Mat á σ^2 í einföldu línulegu aðhvarfi táknum við með s_e^2 og reiknum með

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

Þetta er sama og jafna og fyrir „venjulega“ staðalfrávikið, nema nú er deilt með $n-2$ en ekki $n-1$.

Öryggisbil fyrir β_0

Öryggisbil fyrir β_0

Neðra mark $1-\alpha$ öryggisbils fyrir β_0 er:

$$b_0 - t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{s_x^2 \cdot (n-1)}}$$

Efra mark $1-\alpha$ öryggisbils er:

$$b_0 + t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{s_x^2 \cdot (n-1)}}$$

þar sem b_0 er reiknað skv. jöfnu, n er fjöldi paraðra mælinga, $\bar{x}$ er meðaltal skýribreytunnar, s_x er staðalfrávik skýribreytunnar og $t_{1-\alpha/2,(n-2)}$ má finna í t-töflu.

Öryggisbil fyrir β_1

Öryggisbil fyrir β_1

Neðra mark $1-\alpha$ öryggisbils fyrir β_1 er:

$$b_1 - t_{1-\alpha/2,(n-2)} \cdot s_e \frac{1}{\sqrt{s_x^2 \cdot (n-1)}}$$

Efra mark $1-\alpha$ öryggisbils er:

$$b_1 + t_{1-\alpha/2,(n-2)} \cdot s_e \frac{1}{\sqrt{s_x^2 \cdot (n-1)}}$$

þar sem b_1 er reiknað skv. jöfnu, n er fjöldi paraðra mælinga, s_x er staðalfrávik skýribreytunnar og $t_{1-\alpha/2,(n-2)}$ má finna í t-töflu.

Spábil fyrir framtíðarmælingar

Spábil fyrir framtíðarmælingar

Neðra mark $1-\alpha$ spábils fyrir framtíðarmælingu á Y:

$$(b_0 + b_1 x_0) - t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x^2 (n-1)}}$$

Efra mark $1-\alpha$ spábils er:

$$(b_0 + b_1 x_0) + t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x^2 (n-1)}}$$

þar sem b_0 og b_1, má reikna skv. jöfnum, n er fjöldi paraðra mælinga, $\bar{x}$ er meðaltal skýribreytunnar, s_x er staðalfrávik skýribreytunnar og $t_{1-\alpha/2,(n-2)}$ má finna í t-töflu.

Próf á fylgnistuðli

Tilgátupróf fyrir ρ

Núlltilgátan er:

$$H_0 : \rho = 0$$

Prófstærðin er:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Ef núlltilgátan er sönn fylgir prófstærðin t-dreifingu með n-2 frígráður, eða $T \sim t(n-2)$.

Gagntilgáta	Hafna H_0 ef:
$H_1 : \rho < 0$	$T < -t_{1-\alpha}$
$H_1 : \rho > 0$	$T > t_{1-\alpha}$
$H_1 : \rho \neq 0$	$T < -t_{1-\alpha/2}$ eða $T > t_{\alpha/2}$

Hvert erum við komin...

1. Einfalt línulegt aðhvarf - hluti 1
2. Einfalt línulegt aðhvarf - hluti 2
3. Einfalt línulegt aðhvarf - hluti 3
4. Ályktanir í aðhvarfsgreiningu
5. Fervikagreining í aðhvarfsgreiningu

Fervikagreining í aðhvarfsgreiningu

- Nota má fervikagreiningu til að draga ályktanir í línulegri aðhvarfgreiningu.
- Uppsetningin er á margan hátt svipuð og þegar við notum fervikagreiningu til að draga ályktanir um meðaltöl, en fervikasummurnar eru reiknaðar á örlítið annan hátt.

Fyrir sérhverja mælingu y_i gildir að

$$(y_i - \bar{y}_.) = (\hat{y}_i - \bar{y}_.) + (y_i - \hat{y}_i)$$

þar sem $\bar{y}_.$ er meðaltal y-mælinganna og $\hat{y}_i$ er spágildið fyrir y_i.

Fervikagreining í aðhvarfsgreiningu

Þegar aðhvarfsgreiningarlíkan er metið með jöfnu minnstu kvaðrata gildir ennfremur að

$$\sum_{i=1}^{n}(y_i - \bar{y}_.)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_.)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Út frá þessu reiknum við fervikasummur í einföldu línulegu aðhvarfi.

Fervikagreining í aðhvarfsgreiningu

Fervikasummur í einföldu línulegu aðhvarfi

Fervikasummurnar eru reiknaðar með

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y}_.)^2$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_.)^2$$

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Heildarbreytileikanum má skipta upp í breytileika metnu gildanna annars vegar og breytileika leifanna hins vegar eða

$$SS_T = SS_R + SS_E.$$

Fervikagreining í aðhvarfsgreiningu

Líkt og í fervikagreiningu er algengt er að setja kvaðratsummurnar upp í svokallaða *fervikagreiningartöflu* (ANOVA table). Dæmigerða fervikasummutöflu má sjá hér að neðan.

Fervikasummur	Frígráður	Meðalfervikasummur
SS_R	1	$MS_R = SS_R$
SS_E	$N-2$	$MS_E = \frac{SS_E}{N-2}$
SS_T	$N-1$	

Fervikagreining í aðhvarfsgreiningu

Fervikasummurnar má einnig nota til að reikna skýringarhlutfall með öðrum hætti. Sú framsetning útskýrir jafnvel enn betur hvers vegna r^2 stendur fyrir hlutfallslegan breytileika í Y sem er hægt að skýra með breytingum á x.

Skýringarhlutfall

Skýringarhlutfall má reikna með jöfnunni

$$r^2 = \frac{SSR}{SST}$$

Það stendur fyrir hlutfallslegan breytileika í Y sem er hægt að skýra með breytingum á x.

Tilgátupróf í fervikagreiningu fyrir aðhvarfsgreiningu

Tilgátuprófið sem við notum í fervikagreiningu gerir ráð fyrir að frávikin frá aðhvarfslínunni séu i.i.d. normaldreifð. Það skilyrði má kanna með því að teikna normaldreifingarrit af leifunum.

Tilgátupróf í fervikagreiningu fyrir aðhvarfsgreiningu

Tilgátan sem við viljum kanna er

$$H_0 : \beta_1 = 0$$

á móti gagntilgátunni

$$H_1 : \beta_1 \neq 0$$

Tilgátupróf í fervikagreiningu fyrir aðhvarfsgreiningu

Tilgátupróf fyrir einhliða fervikagreiningu, áframhald

Prófstærðin er

$$F = \frac{SS_R/(1)}{SS_E/(N-2)} = \frac{MS_R}{MS_E}.$$

Sé núlltilgátan sönn fylgir prófstærðin F-dreifingu með 1 og $N-2$ fjölda fríráða, eða $F \sim F_{(1,N-2)}$, þar sem N er heildarfjöldi mældra para.

Hafna skal H_0 ef $F > F_{1-\alpha,(1,N-2)}$.

Sé núlltilgátunni hafnað er β_1 frábrugðið núlli.

Þetta tilgátupróf er illframkvæmanlegt í höndunum en má reikna á einfaldan hátt í öllum helstu tölfræðiforritum.