Aðhvarfsgreining

Tölfræði frá grunni - kafli 17

Anna Helga Jónsdóttir
Sigrún Helga Lund

Háskóli Íslands

Helstu atriði:

1. Punktarit
2. Jafna beinnar línu
3. Fylgni og orsakasamband
4. Einfalt línulegt aðhvarf
5. Ályktanir í aðhvarfsgreiningu

Hvert erum við komin...

Punktarit

Punktarit

Við notum *punktarit* (scatter plot) til að skoða samband milli tveggja talnabreyta.

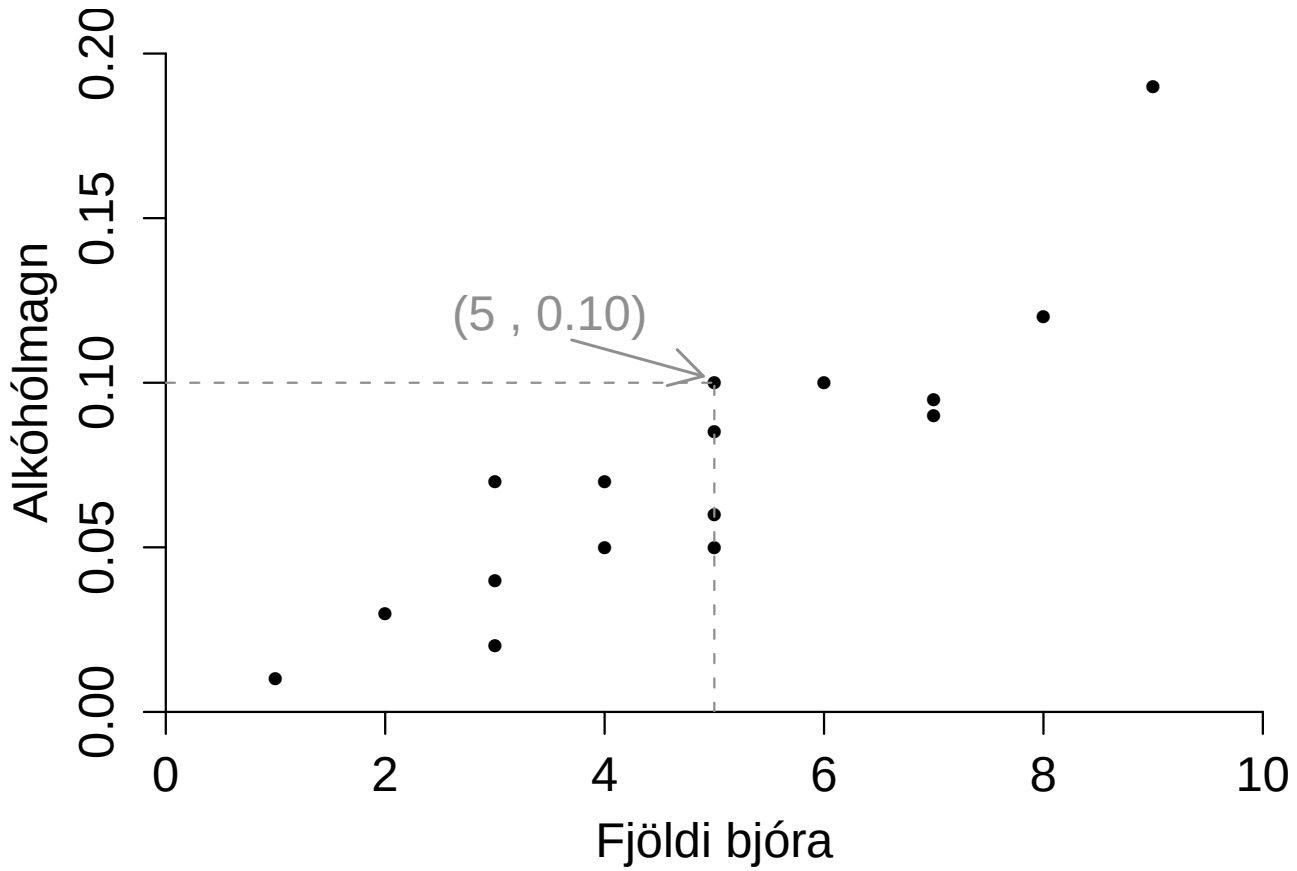
Gildi annarrar breyturnar eru á y-ásnum (lóðréttur) og hinnar á x-ásnum (láréttur).

Þegar önnur breytan er skýribreyta og hin er svarbreyta er svarbreytan alltaf á y-ásnum og skýribreytan á x-ásnum.

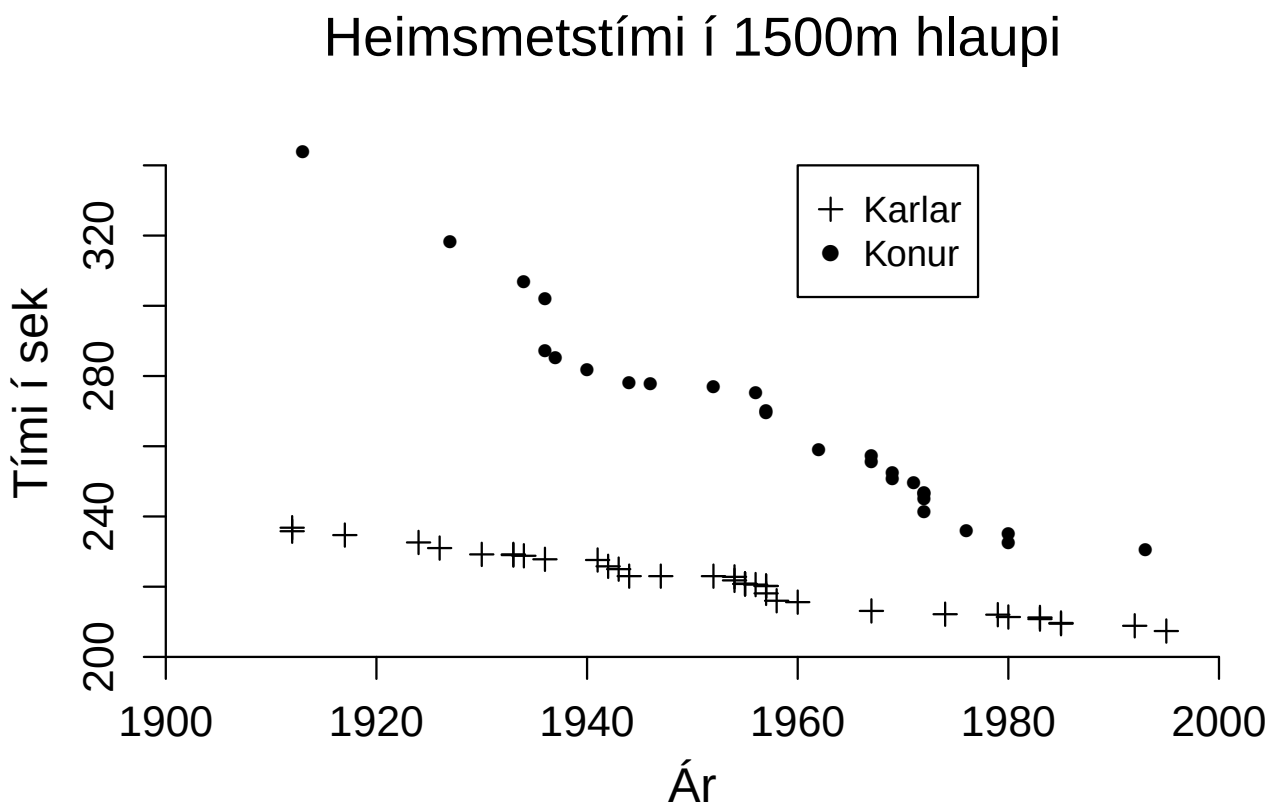
Svarbreytur og skýribreytur

Fyrir sérhvert viðfangsefni mun gildi *skýribreytu* (explanatory variable) þess hafa áhrif á það hvaða gildi *svarbreytan* (response variable) mun taka.

Punktarit - samfelldar breytur



Punktarit með flokkabreytu



Hvert erum við komin...

Jafna beinnar línu

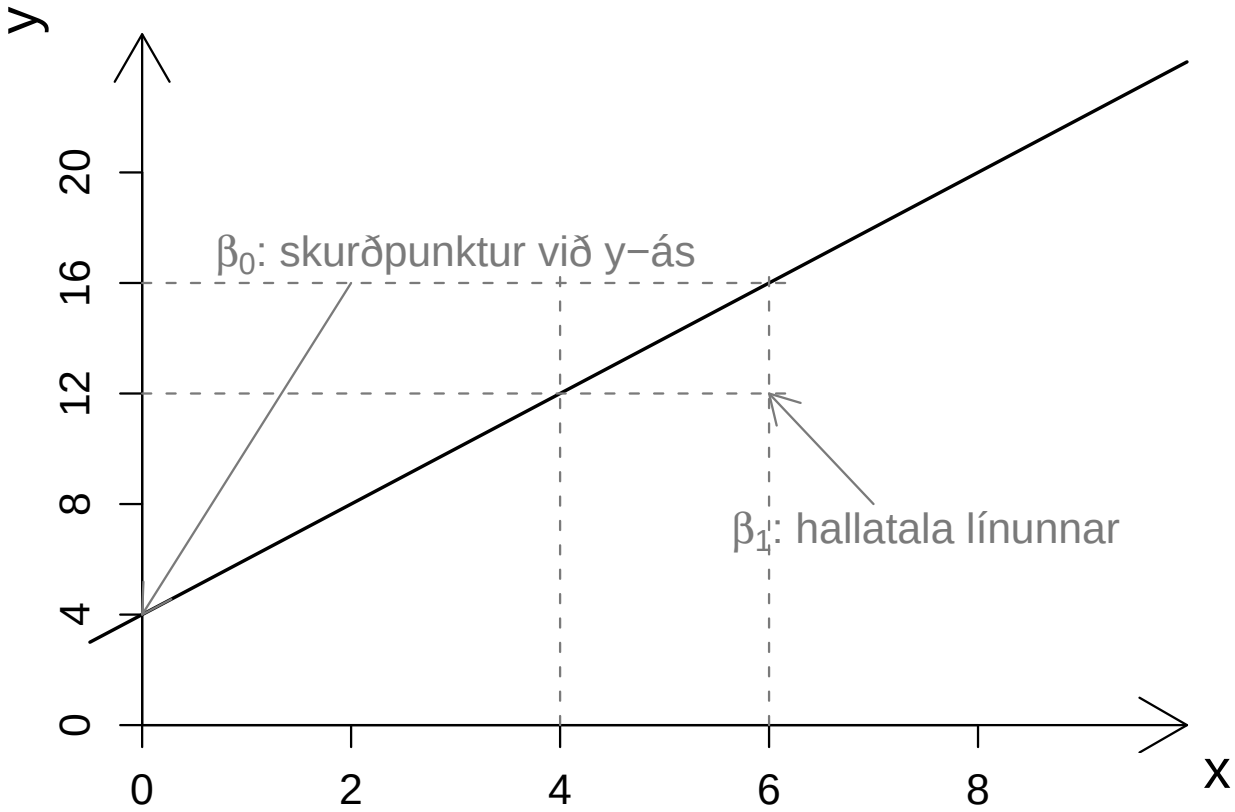
Jafna beinnar línu

Jafna beinnar línu lýsir línulegu sambandi tveggja breyta, y og x. Jöfnuna má skrifa sem

$$y = \beta_0 + \beta_1 x$$

þar sem β_0 er *skurðpunktur* (intercept) línunnar við y-ás og β_1 er *hallatala* (slope) línunnar.

Jafna beinnar línu



Mynd: Jafna beinnar línu.

Hvert erum við komin...

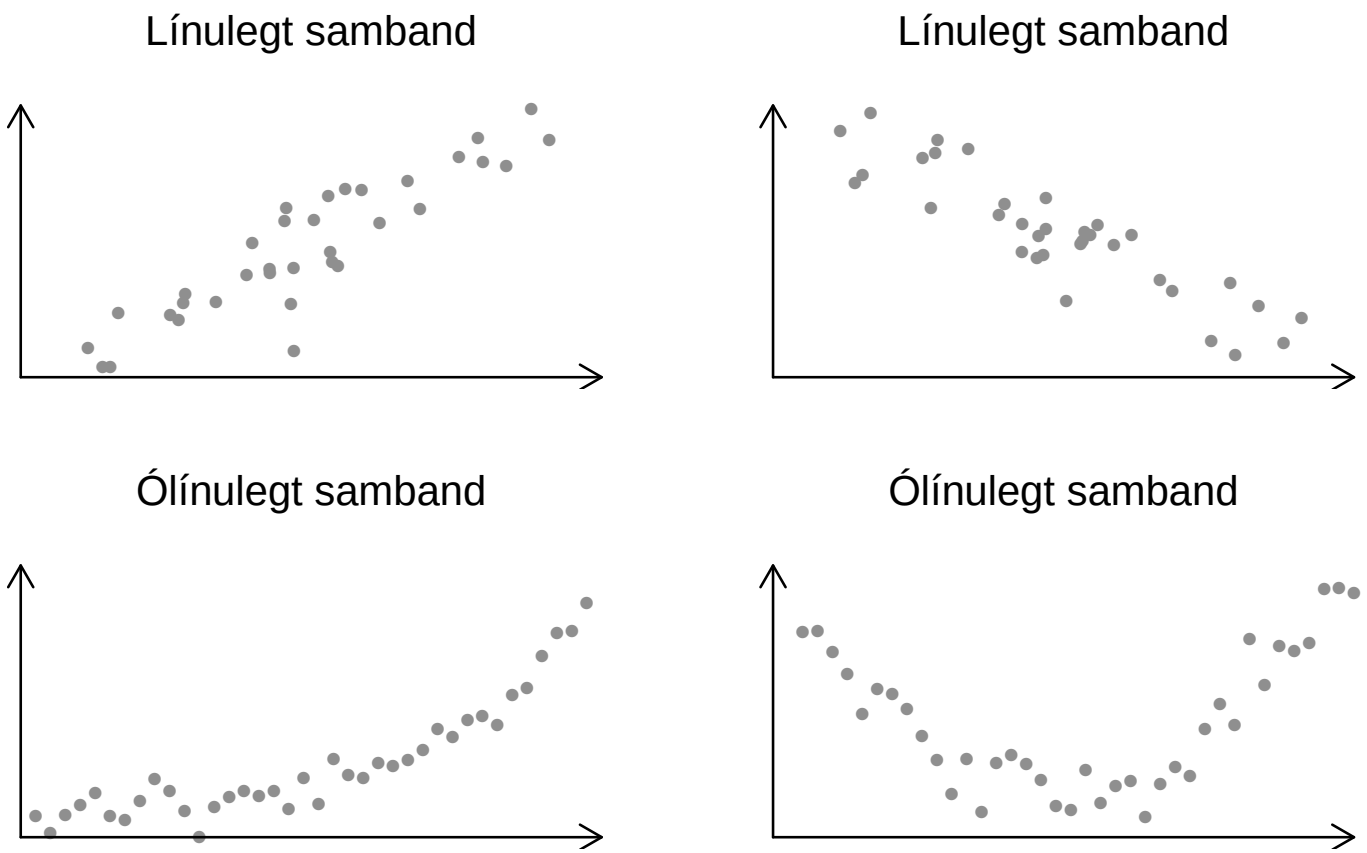
Línulegt samband

Línulegt samband

Við segjum að samband tveggja breyta sé *línulegt* (linear) ef nota má jöfnu beinnar línu til spá fyrir um gildi háðu breytunnar út frá gildi óháðu breytunnar.

Athugið að það geta verið margs konar aðrar gerðir af samböndum milli tveggja breyta. Til dæmis ef lýsa má sambandinu með fleygboga, veldisvísisfalli og svo framvegis. Þau sambönd eru einu orði nefnd ólínuleg og eru utan efni þessa námskeiðs.

Línulegt og ólínulegt samband



Mynd: Punktarit þar sem samband breyta er línulegt (að ofan) og ólínulegt (að neðan).

Fylgnistuðull úrtaks

Fylgnistuðull úrtaks

Gerum ráð fyrir að við höfum n mælingar á tveimur breytum x og y.

Táknum meðaltal og staðalfrávik x breytunnar með $\bar{x}$ og s_x og meðaltal og staðalfrávik y breytunnar með $\bar{y}$ og s_y.

Fylgnistuðul úrtaksins (sample coefficient of correlation) reiknum við með

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

Gætið ykkar að við notum fylgnisstuðul eingöngu til að meta **línulegt** samband!

Stefna og styrkleiki línulegs sambands

Stefna línulegs sambands

Formerki fylgnistuðulsins segir til um það hver *stefna* línulegs sambands er. Hún er annað hvort jákvæð eða neikvæð.

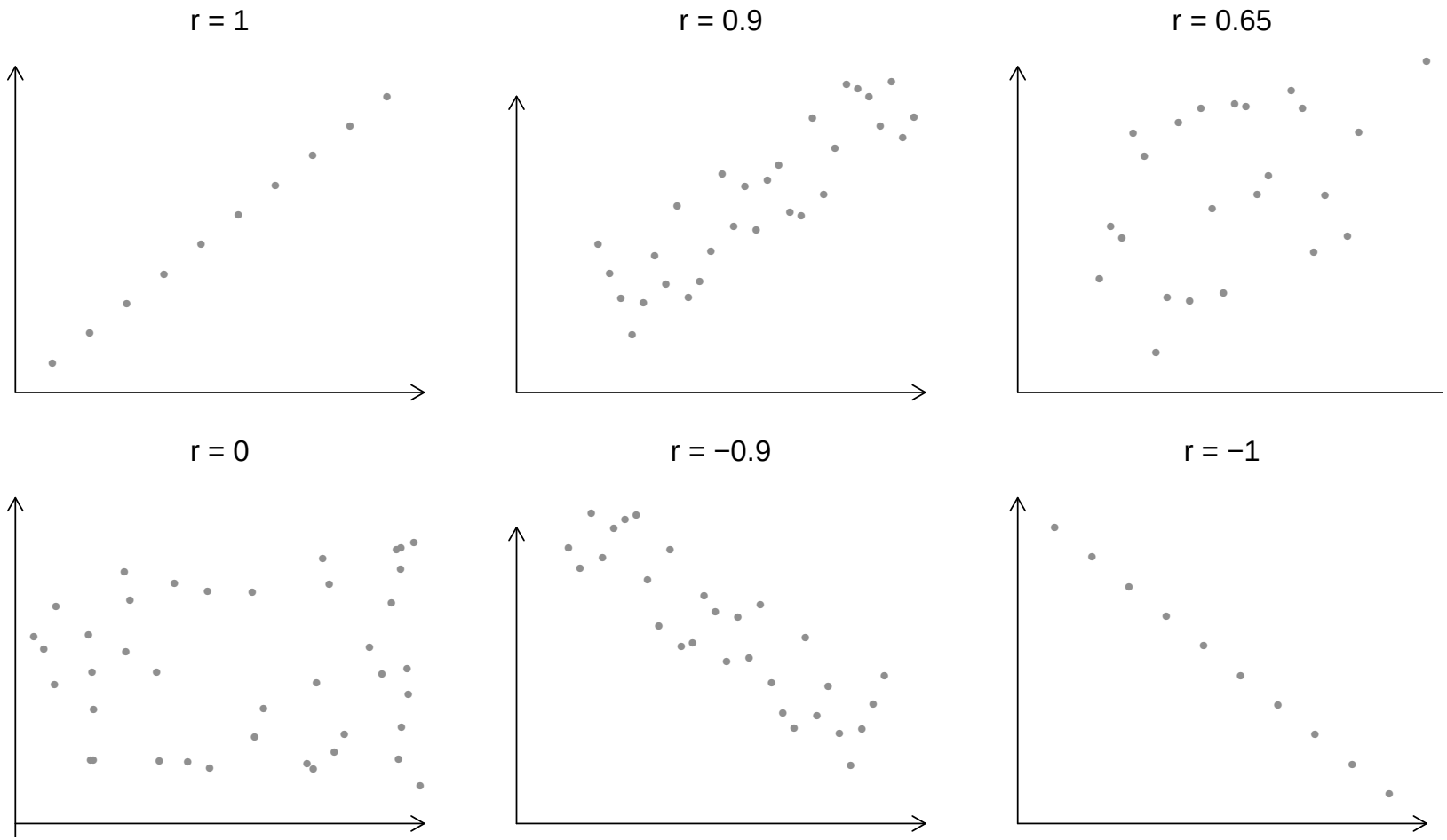
- Ef fylgnistuðull tveggja breyta er jákvæður, þá segjum við að fylgni þeirra sé *jákvæð*.
- Ef fylgnistuðull tveggja breyta er neikvæður, þá segjum við að fylgni þeirra sé *neikvæð*.

Styrkleiki línulegs sambands

Algildi (absolute value) fylgnistuðuls lýsir *styrkleika* línulega sambandsins sem gildir milli breytanna.

Hann segir okkur hversu vel við getum ákvarðað gildi svarbreytunnar út frá gildi skýribreytunnar.

Stefna og styrkleiki línulegs sambands



Mynd: Punktarit fyrir mismunandi gildi á r.

Fylgni og orsakasamband

- *Orsakasamband* (causation) er til staðar þegar breyting á annarri breytunni **veldur** breytingu í hinni breytunni.
- Oft má finna sterka fylgni á milli breyta þó svo að orsakasamband sé ekki til staðar.
- Í mörgum tilfellum eru breyturnar þá undir áhrifum þriðju breytunnar sem þá er kölluð *dulin breyta* (lurking variable).
- Því dugar há fylgni aldrei ein og sér til að fullyrða að orsakasamband sé á milli tveggja breyta.

Dæmi

Drukknanir og ísát í Ohio er eitt dæmi. Getið þið nefnt fleiri?

Hvert erum við komin...

Aðhvarfsgreiningarlíkanið

Aðhvarfsgreiningarlíkanið

Einfalda aðhvarfsgreiningarlíkanið (simple linear regression model) má skrifa sem

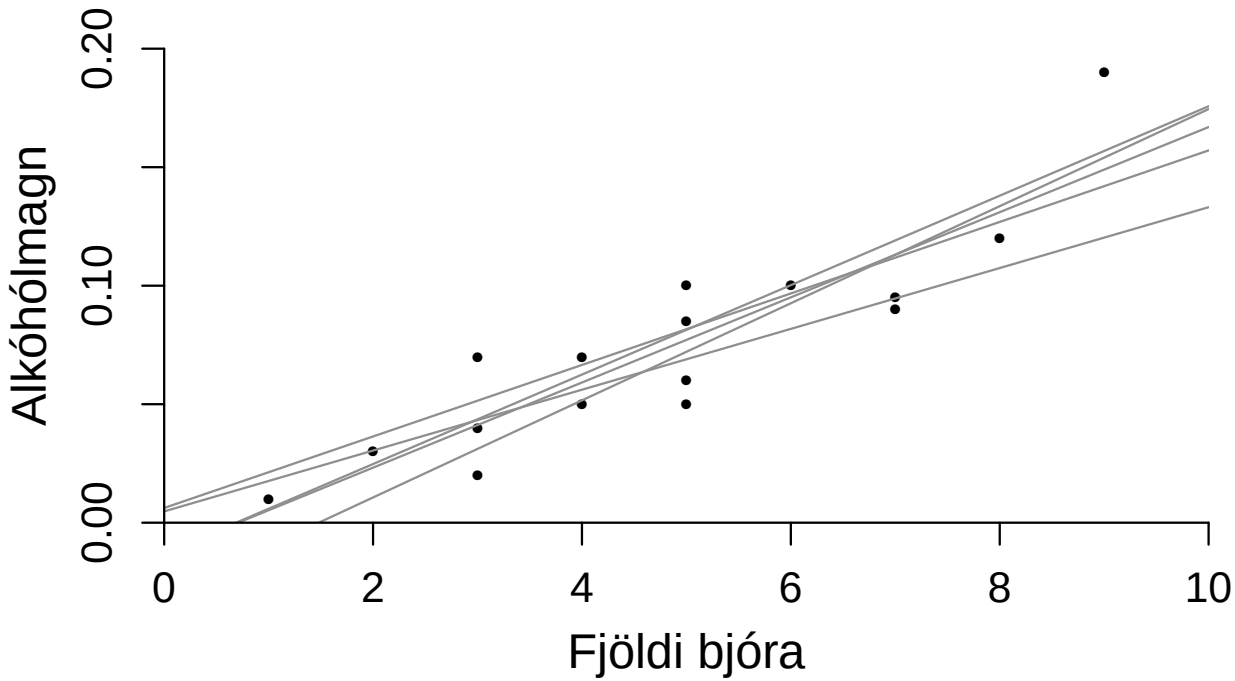
$$Y = \beta_0 + \beta_1 x + \varepsilon$$

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.

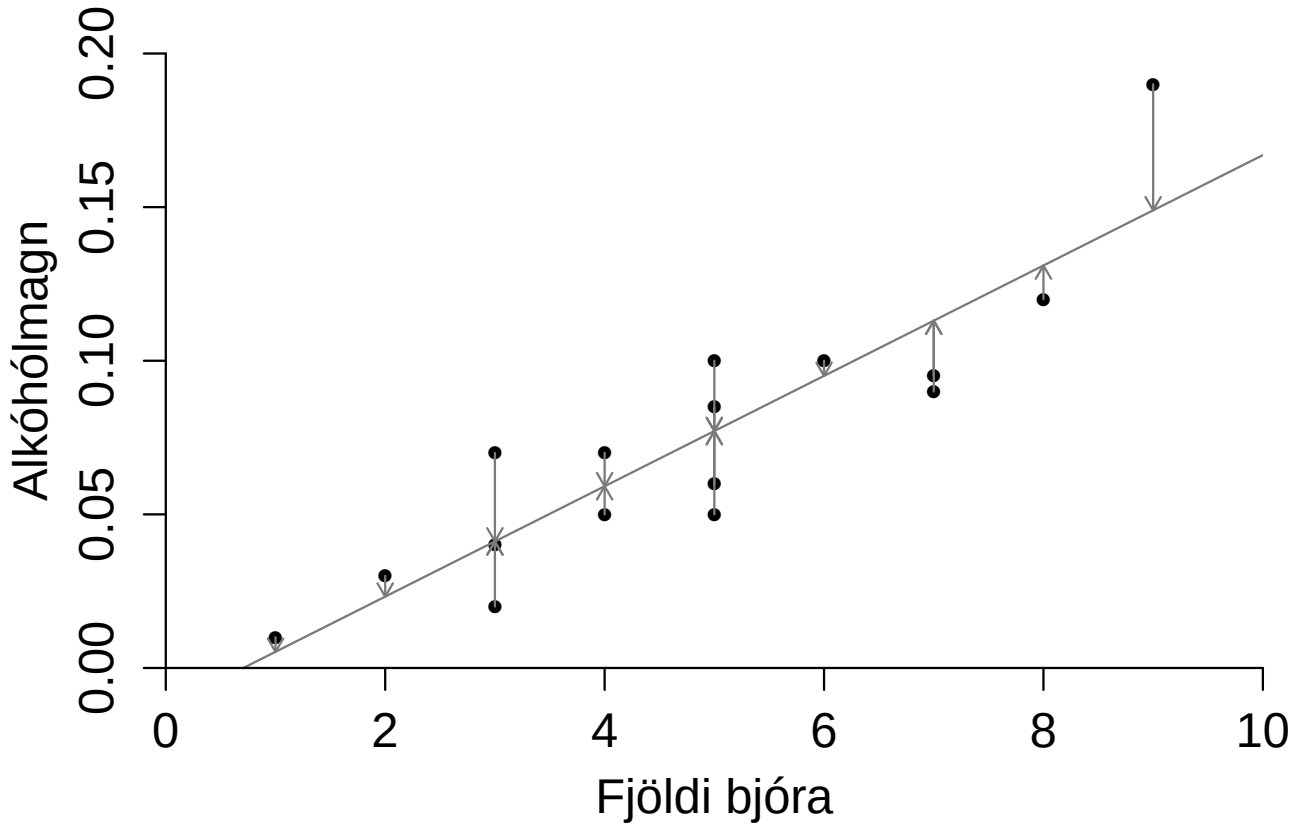
Aðferðin sem við notum kallast aðferð minnstu kvaðrata.

Aðferð minnstu kvaðrata



Mynd: Margar línur, en hvaða lína er best?

Aðferð minnstu kvaðrata



Mynd: 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.

Notum b_0 til að tákna mat á β_0 og b_1 til að tákna mat á β_1. Þá reiknum við b_0 og b_1 með

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

og

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

Við notum stikana til að *spá* (predict) gildi á y fyrir þekkt gildi á x með jöfnu aðhvarfslínu minnstu kvaðrata

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

Dæmi

Dæmi

Sambandi hraða bíla og stöðvunarvegalengdar þeirra má lýsa með línulegu líkani. Eftirfarandi R skipanir voru notaðar til að reikna lýsistærðir fyrir gögnin:

```
 > mean(hradi)
[1] 24.7786
> sd(hradi)
[1] 8.50782
> mean(vegalengd)
[1] 13.10030
> sd(vegalengd)
[1] 7.854506
> cor(hradi,vegalengd)
[1] 0.8068949
```

Finnið jöfnu aðhvarfslínu fyrir samband hraða og stöðvunarvegalengdar bíla. Hvað myndum við spá að stöðvunarvegalengd bíls sem ekur á 20 km hraða á klukkustund sé löng?

Leifar

Leifar

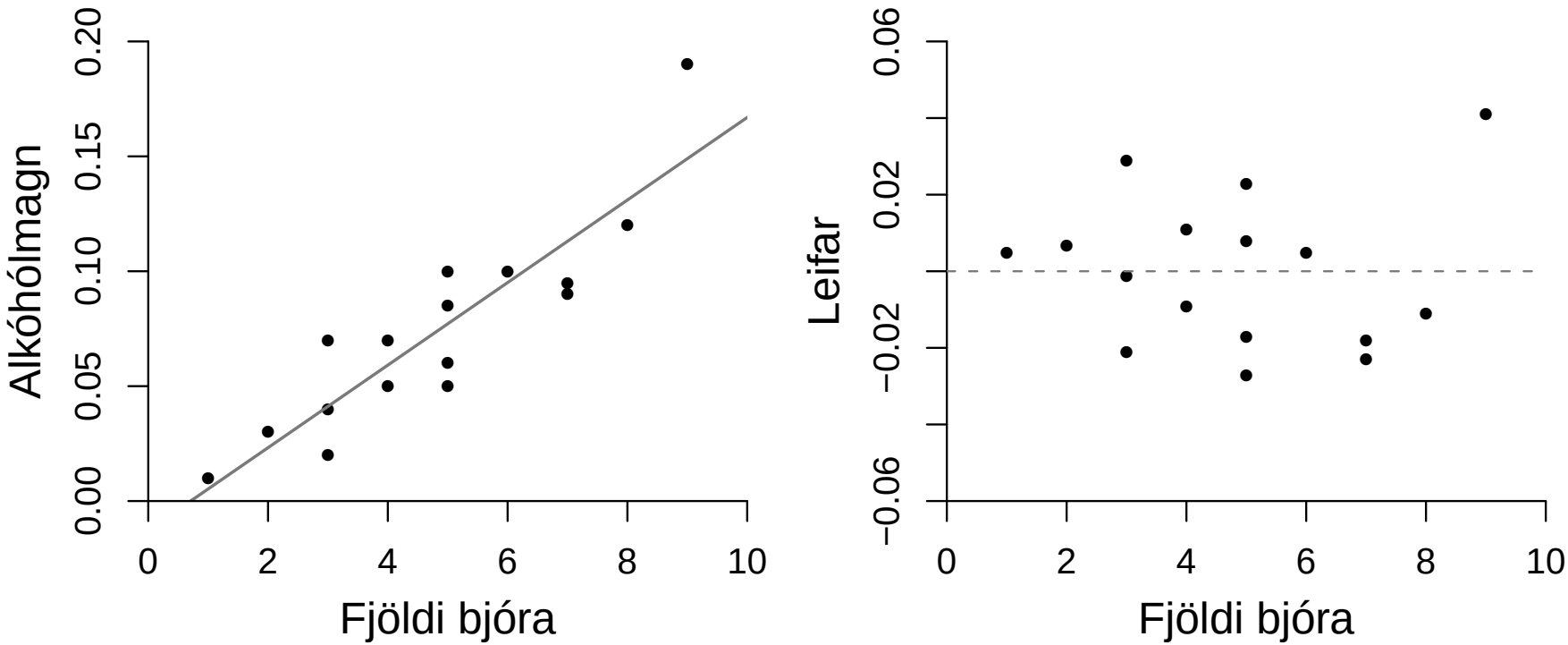
Lóðrétta fjarlægð frá mælingunum okkar að aðhvarfslínunni köllum við *leifar* (residuals) og táknum með e. Stærð leifa má reikna með

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

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

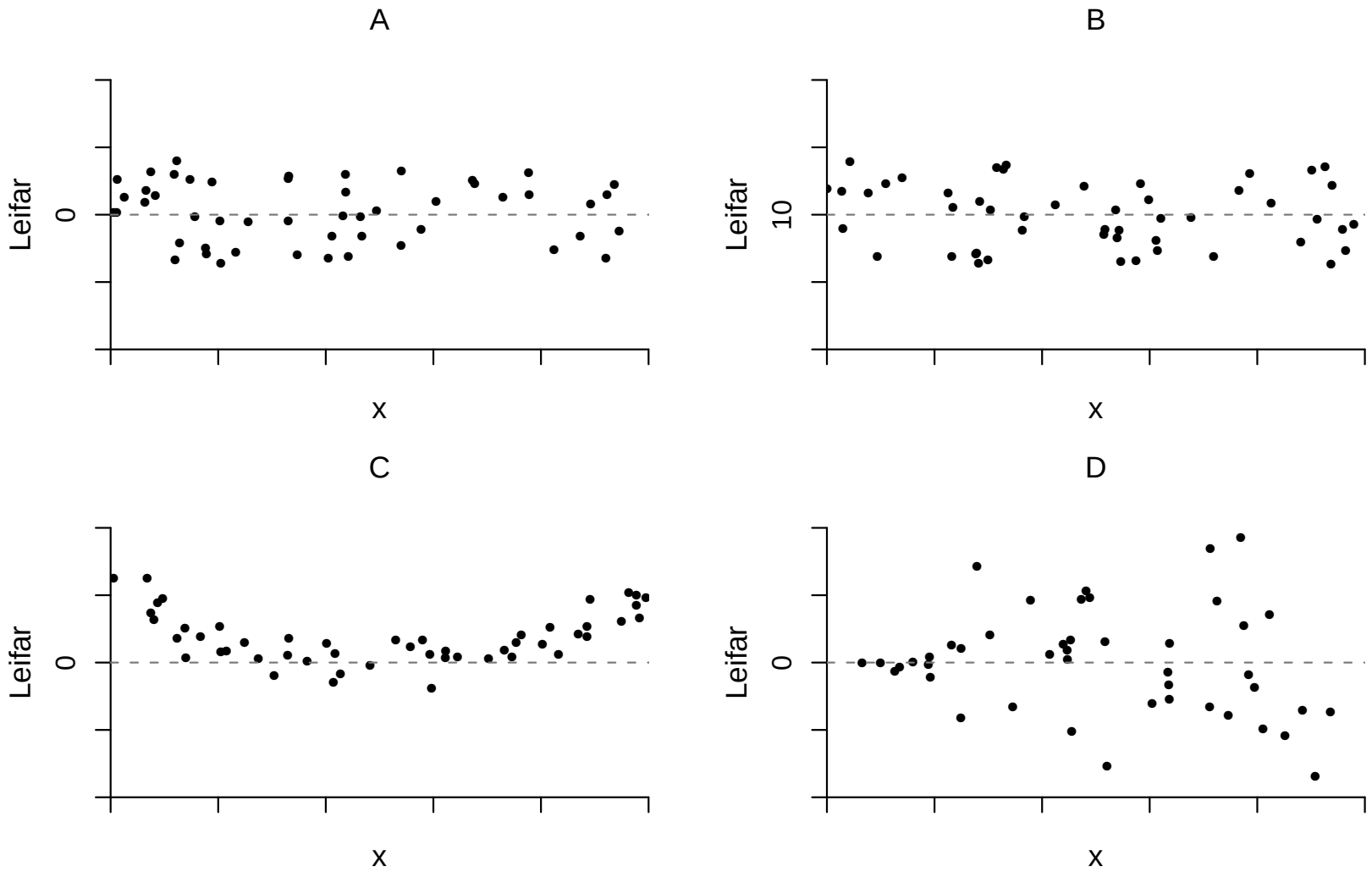
Leifarit

Leifaritið sýnir leifarnar á y-ásnum og skýribreytuna á x-ásnum.



Mynd: Punktarit af gögnum og leifarit

Leifarit

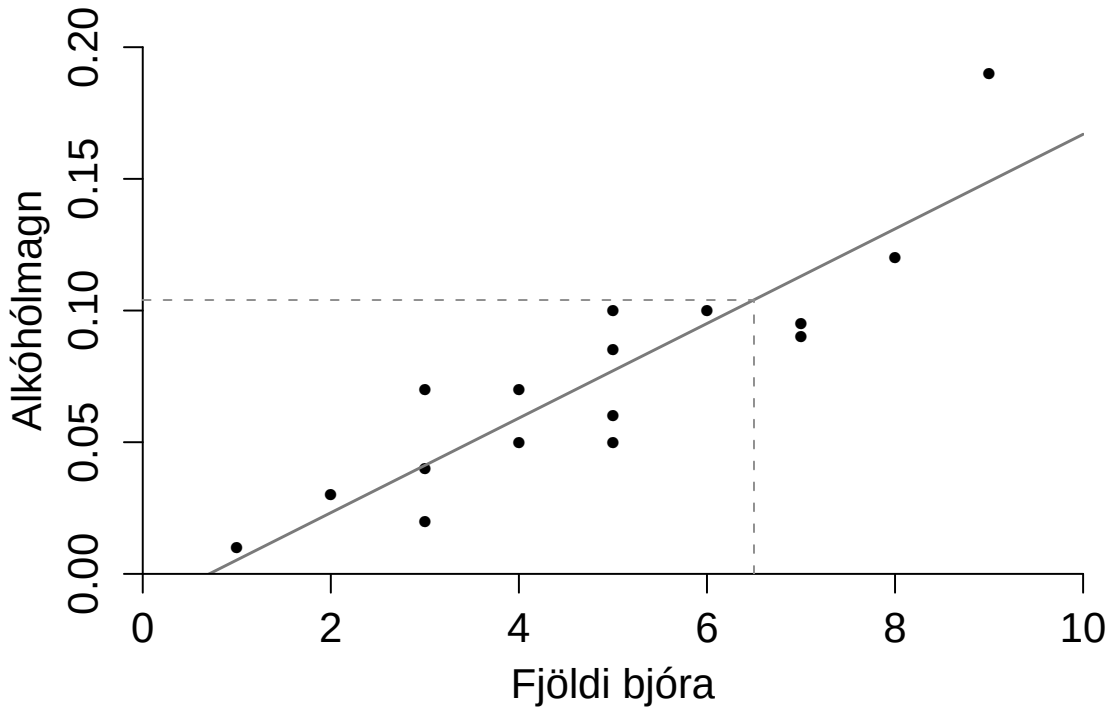


Mynd: Leifarit

Brúun

Brúun

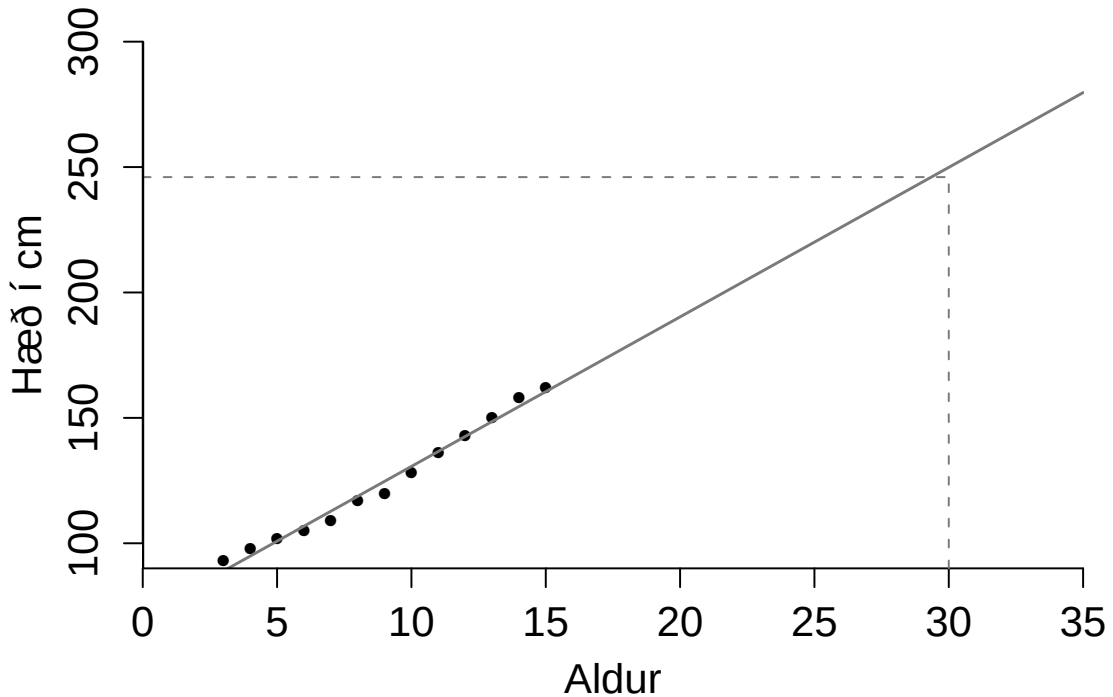
Sé aðhvarfslíkan notuð til að spá fyrir um gildi á Y fyrir eitthvert gildi á x sem er á sama reki og x-gildin sem notuð voru til að meta stikana í líkaninu er talað um að *brúa* (interpolate).



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

r^2 í aðhvarfsgreiningu

Sé fylgnistuðullinn settur í annað veldi, r^2, er talað um skýringarhlutfall.

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

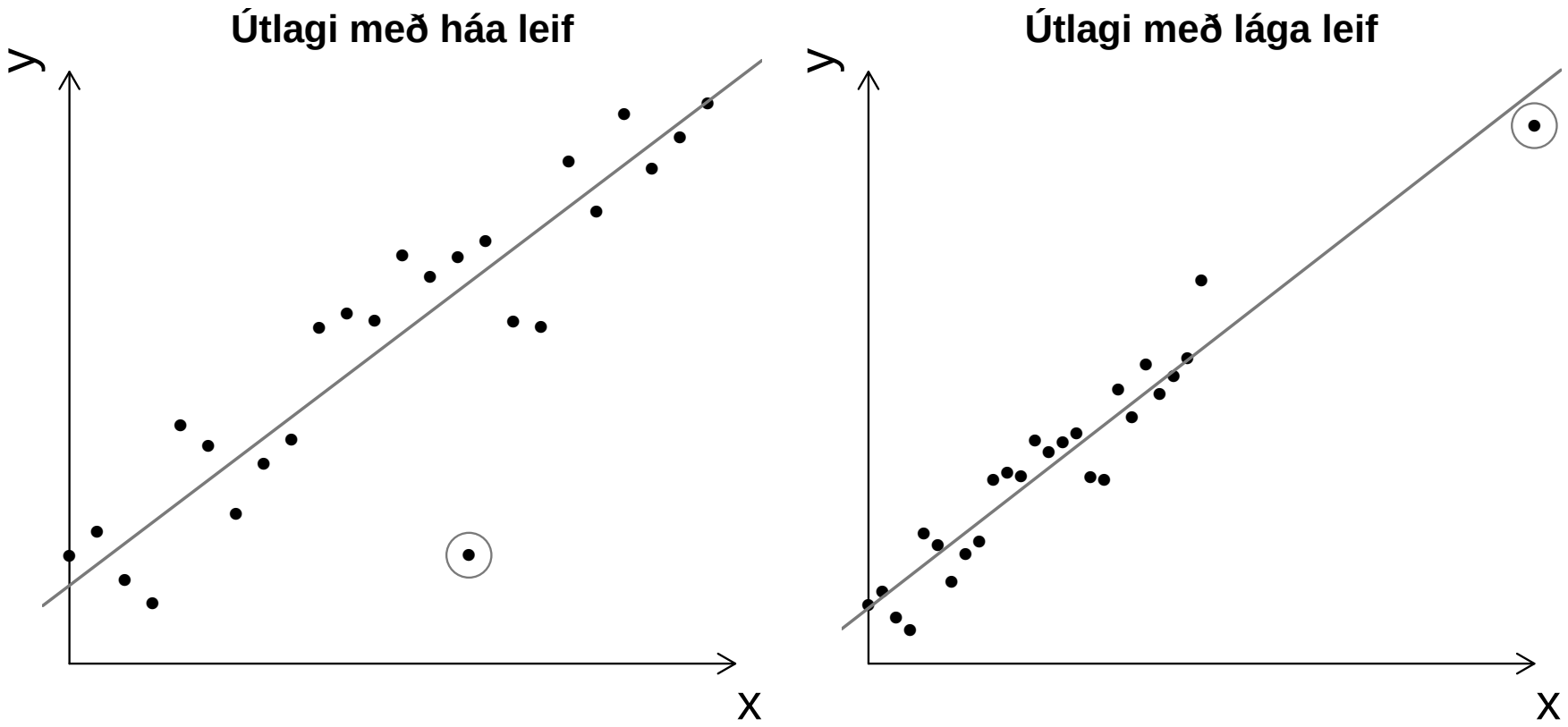
Dæmi

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?

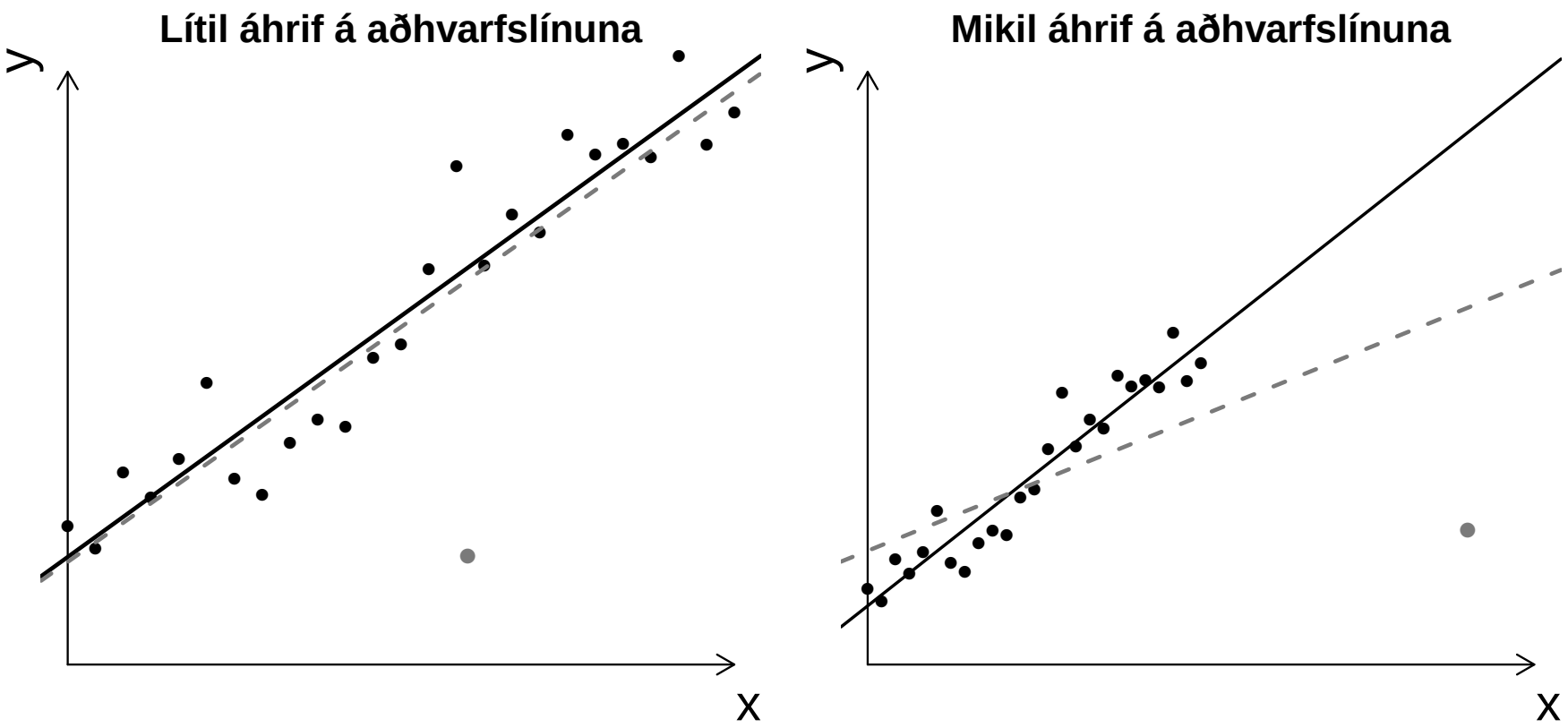
Fylgnin mældist 0.8. Hversu hátt hlutfall breytileika í blóðþrýstingi milli einstaklinga má þá útskýra með ólíku BMI?

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...

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.

Dæmi

Dæmi

s_e í línulega aðhvarfinu á stöðvunarvegalengd bíla eftir hraða þeirra reyndist 3.19. Að sama skapi byggðu útreikningarnir á 50 mælingum.

Reiknið 95% öryggisbil fyrir β_0 og β_1 og reiknið einnig 95 % spábil fyrir stöðvunarvegalengd bíls sem ekur á 30 km hraða á klukkustund.

Væri við hæfi að spá fyrir um stöðvunarvegalengd bíls sem ekur á 80 km hraða á klukkustund út frá þessum gögnum?

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 $T > t_{\alpha/2}$

Dæmi

Dæmi

Getum við fullyrt að það sé jákvæð fylgni á milli stöðvunarvegalengdar bíla og hraða þeirra?