3. STATISTIKE Z DVEMA SPREMENLJIVKAMA

- Bivariatne metodo obravnavajo dve spremenljivki hkrati, zato so podatki zapisani:

x_1 y_1

x_2 y_2

: :

x_n y_n

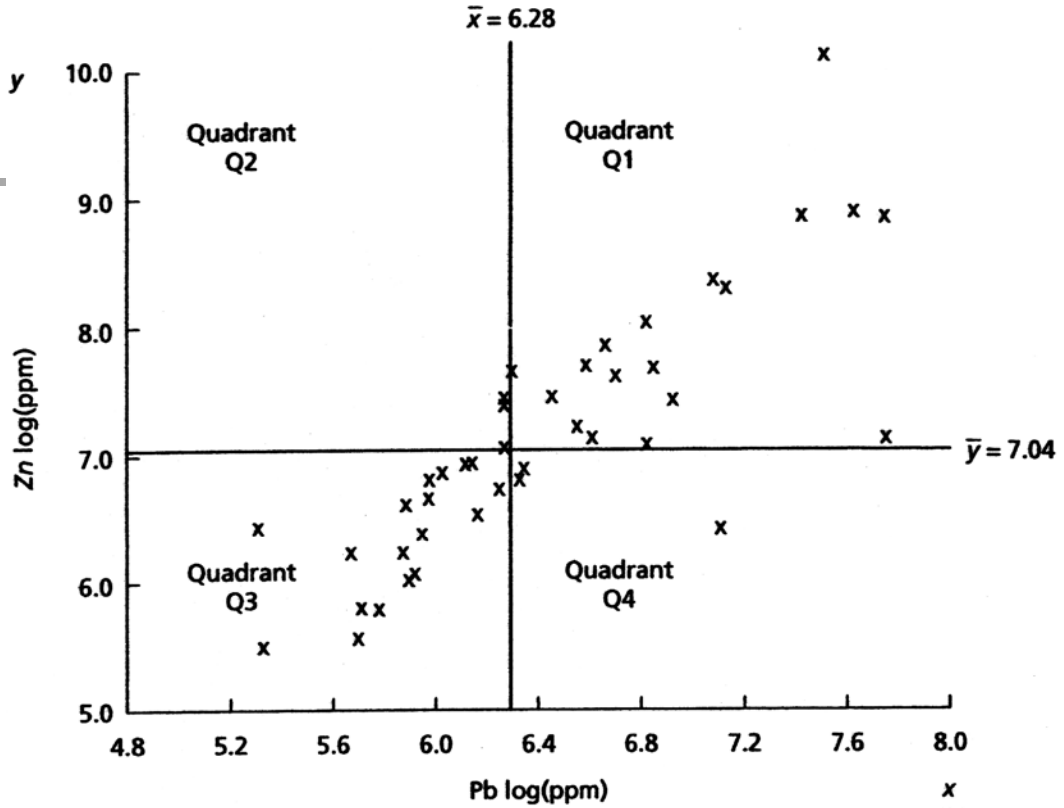
3.1. KORELACIJSKI KOEFICIENT

- Mera stopnje linearne povezanosti je korelacijski koeficient ρ, njegova ocena, korelacijski koeficient vzorca pa r.
- Temelji na popravljeni vsoti zmnožkov VP (CSP - Corrected sum of Products):

$$VP_{xy} = CSP_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$

3.1. KORELACIJSKI KOEFICIENT

- Grafična porazdelitev VP omogoča razumeti lastnosti podatkov, iz katerih izhaja korelacijski koeficient.
- Predznak VP zavisi od tega, v katerem kvadrantu je točka oz. ali je njena vrednost višja ali nižja od povprečja spremenljivke.



	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
Q1	+	+	+
Q2	-	+	-
Q3	-	-	+
Q4	+	-	-

3.1. KORELACIJSKI KOEFICIENT

- VP je odvisna od velikosti vzorca – n. Učinek podecenjene vrednosti populacije zmanjšamo z deljenjem z (n-1). Dobljeno količino imenujemo kovarianca dveh spremenljivk (cov_{xy}):

$$\mathrm{cov}_{xy} = \frac{VP_{xy}}{n-1}$$

3.1. KORELACIJSKI KOEFICIENT

- Vpliv enot odpravimo s standardizacijo, tako da razlike spremenljivk delimo s pripadajočim standardnim odklonom, kar je enako deljenju kovariance z zmnožkom standardnih odklonov $s_x s_y$.
- Parameter je brez enot in se imenuje Pearsonov produkt – momentni korelacijski koeficient r.

$$r_{xy} = \frac{\text{cov}_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})}{(n-1) \cdot s_x \cdot s_y} = \frac{VP_{xy}}{\sqrt{VK_x \cdot VK_y}}$$

$$VK_x = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = s_x^2 \cdot (n-1)$$

3.1. KORELACIJSKI KOEFICIENT

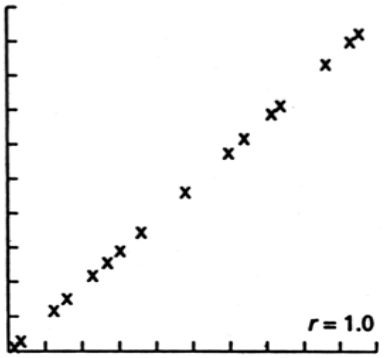
- Determinacijski koeficient D pove, koliko se spremeni x, zaradi spremembe y.
- Izračunamo ga:

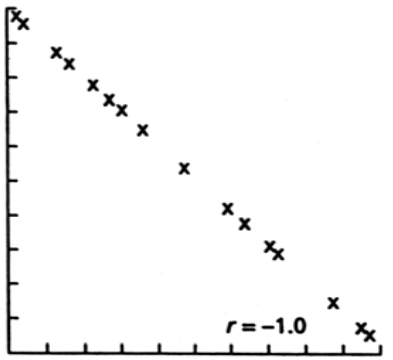
$$D = r^2$$

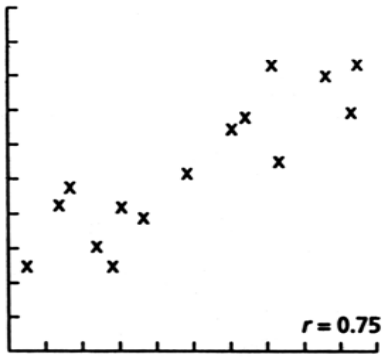
3.1. KORELACIJSKI KOEFICIENT

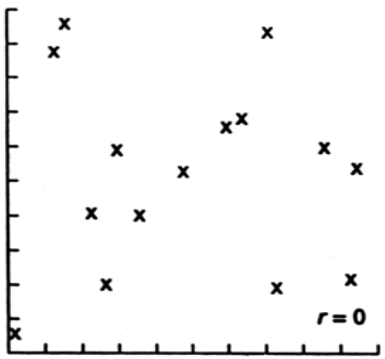
- Opozorila:

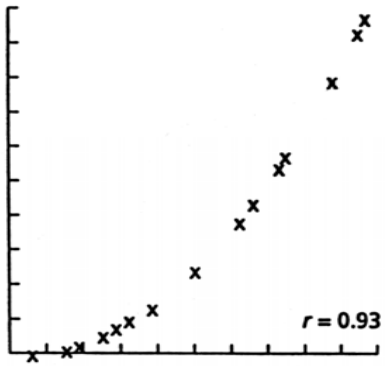
1. Korelacijski koeficient ni splošna mera odnosa med dvema spremenljivkama, temveč le mera tendence k ravni črti. Vedno moramo skupaj z r preveriti tudi bivariatno sipanje.

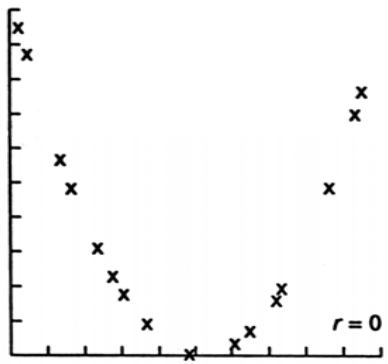






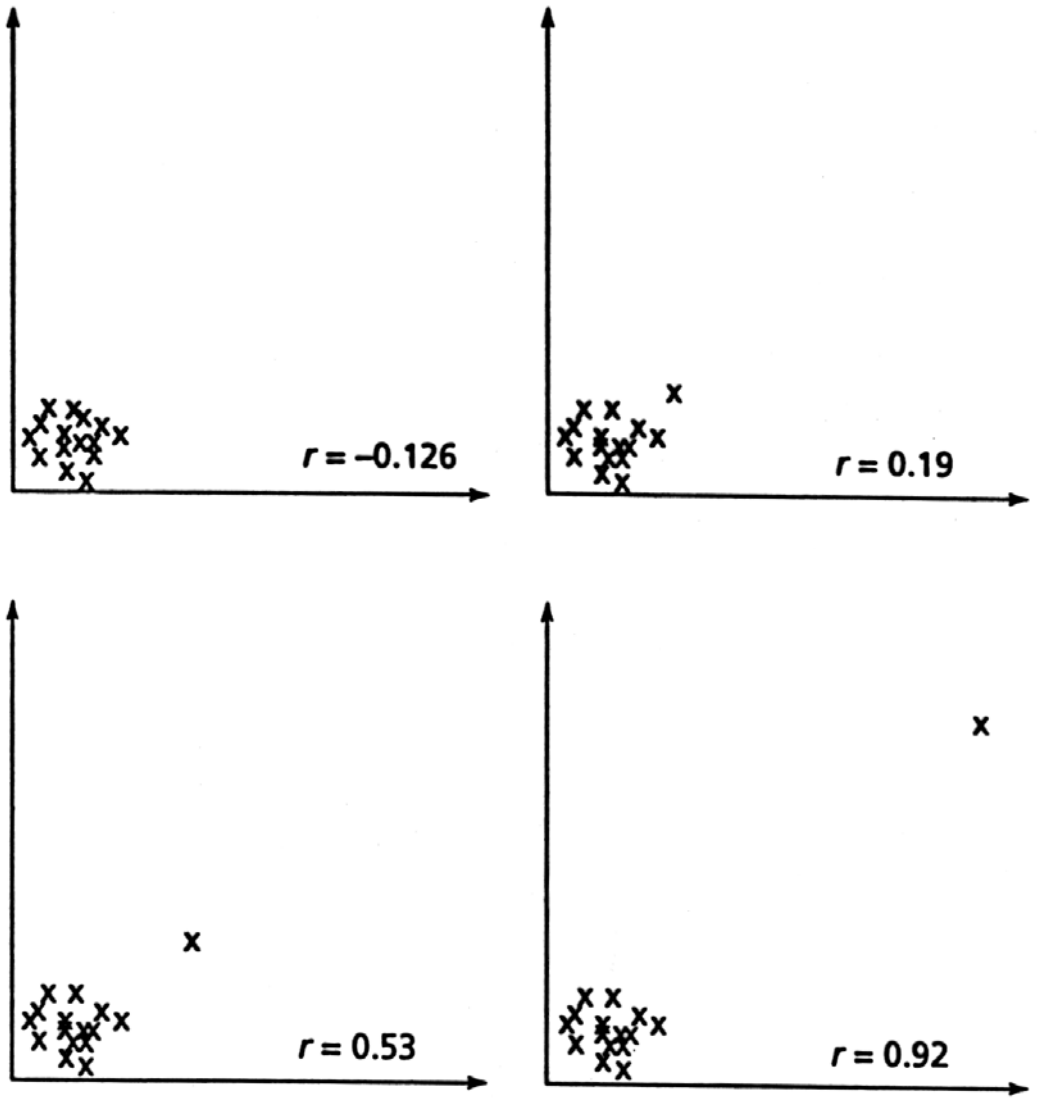






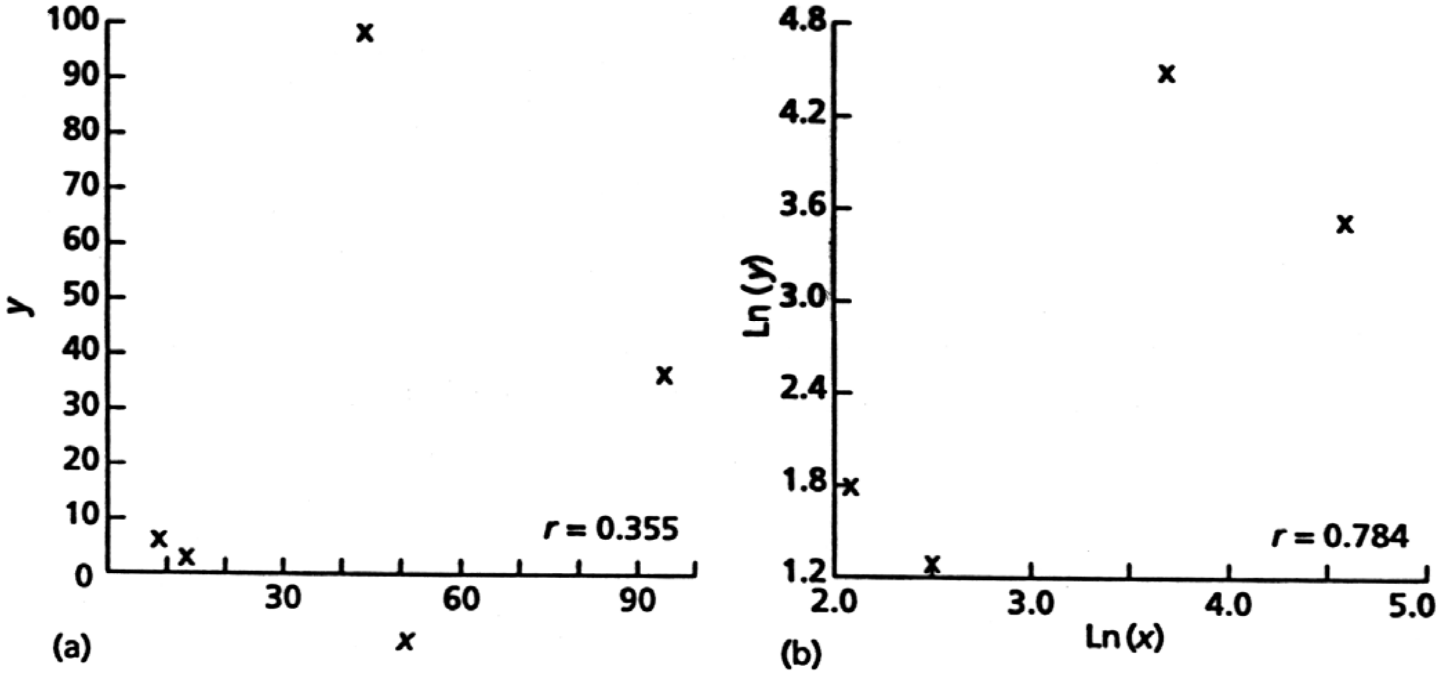
3.1. KORELACIJSKI KOEFICIENT

2. Korelacijski koeficient je občutljiv na neobičajno visoke vrednosti – outlierje.



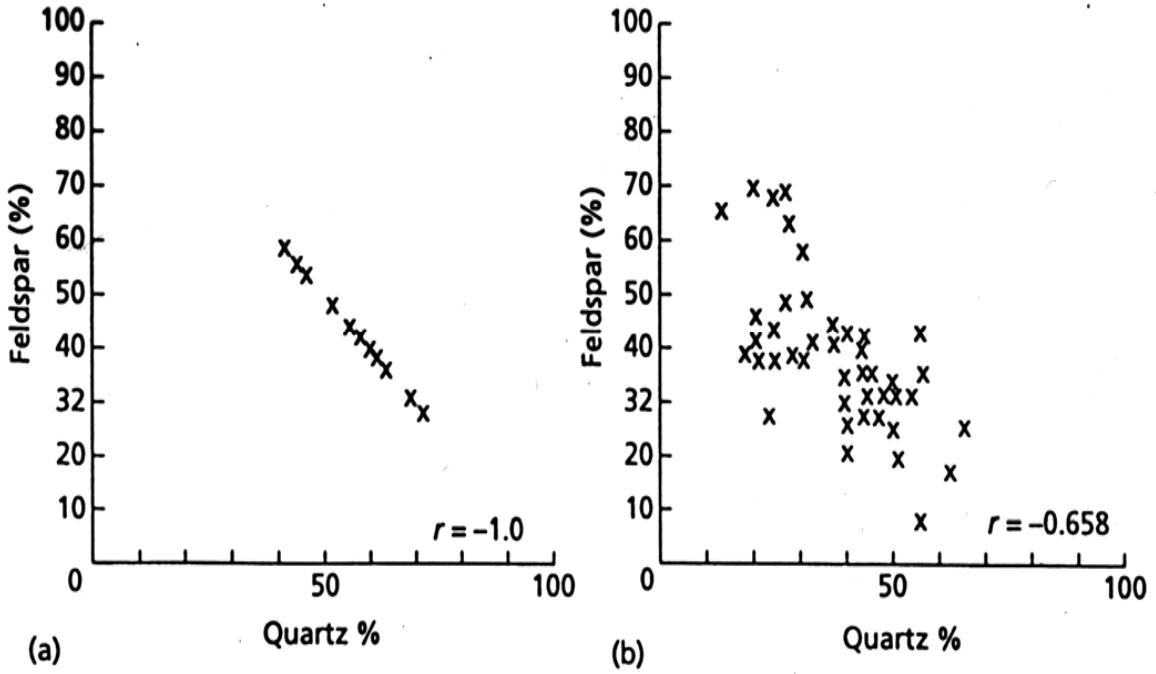
3.1. KORELACIJSKI KOEFICIENT

3. Korelacijski koeficient dveh logaritmiranih spremenljivk je višji, kot če uporabimo njune naravne vrednosti.



3.1. KORELACIJSKI KOEFICIENT

4. Korelacijski koeficient zaprtih nizov podatkov je pogosto lažen. Problem rešimo tako, da namesto pravih vrednosti uporabimo:
   - Razmerja
   - Preostali delež
   - Glavne komponente
   - Hipotetično odprto množico
   - Log-razmerje transformacijo.



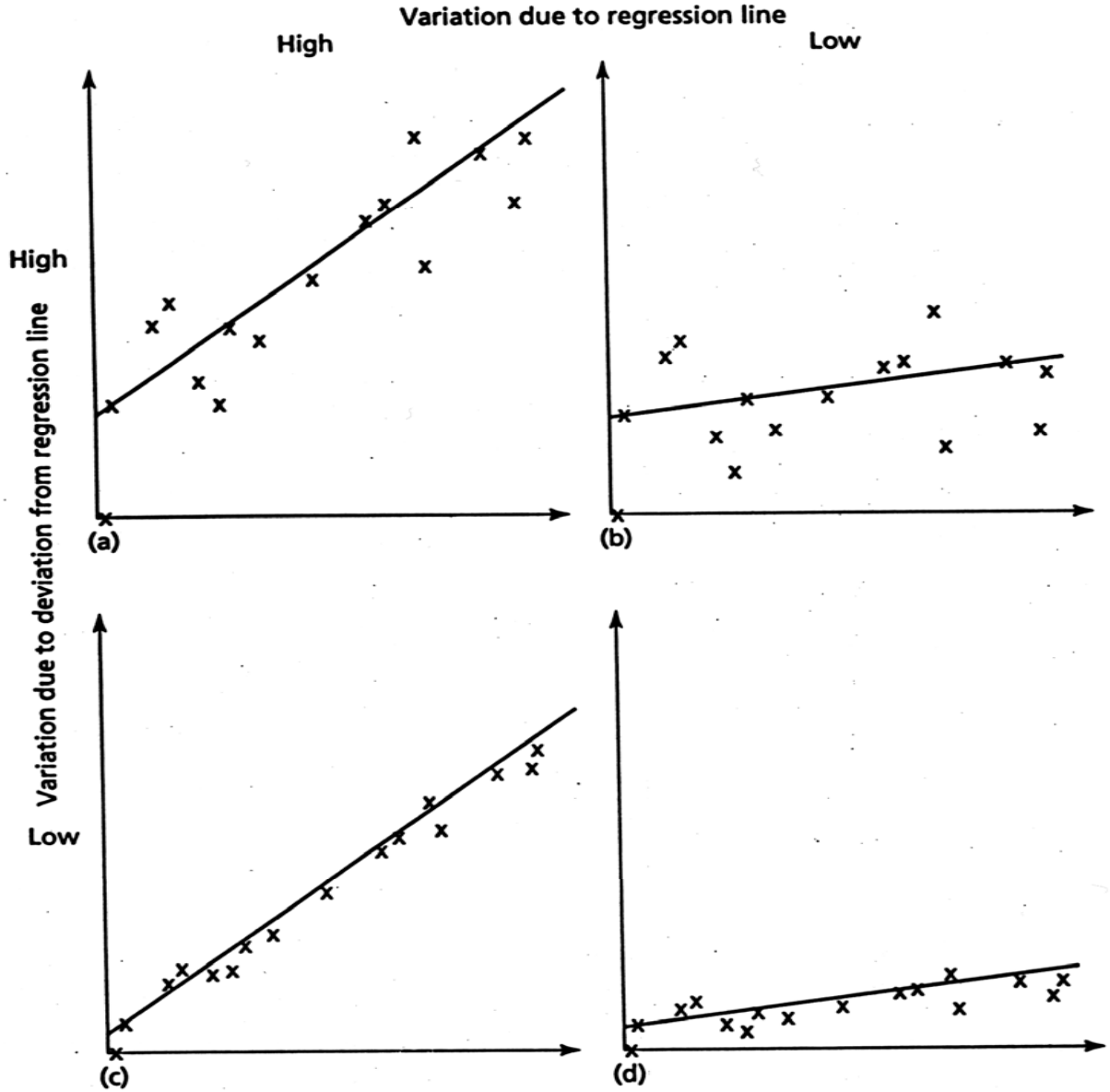
3.1. KORELACIJSKI KOEFICIENT

- Visoka vrednost korelacijskega koeficienta ne pomeni nujno, da je linearna povezava med spremnljivkama tudi statistično značilna. Značilnost ugotovimo s testom:
  - H_0: $\rho=0$ ni korelacije
    H_1: $\rho\neq0$ korelacija obstaja
  - Testna statistika $$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$
  - Zavrni H_0, kadar je $t_{izračunana} > t_{\alpha, n-2}$.
  - Test je veljaven le, če so podatki zajeti nepristransko iz populacije z normalno porazdelitvijo za obe spremenljivki, kar je v geologiji redko. Zato uporabimo neparametrični koeficient ali linearno regresijo.

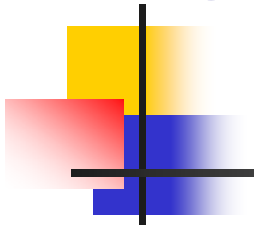
3.2. BIVARIATNA REGRESIJA

- Prednosti regresije pred korelacijskim koeficientom:
  - Lahko jo statistično testiramo kljub nenormalno porazdeljenim podatkom.
  - Testiramo lahko različne ukrivljene odnose.
  - Naravo bivariatnega odnosa natančneje definiramo z ustrezno enačbo.

3.2. BIVARIATNA REGRESIJA



3.2. BIVARIATNA REGRESIJA

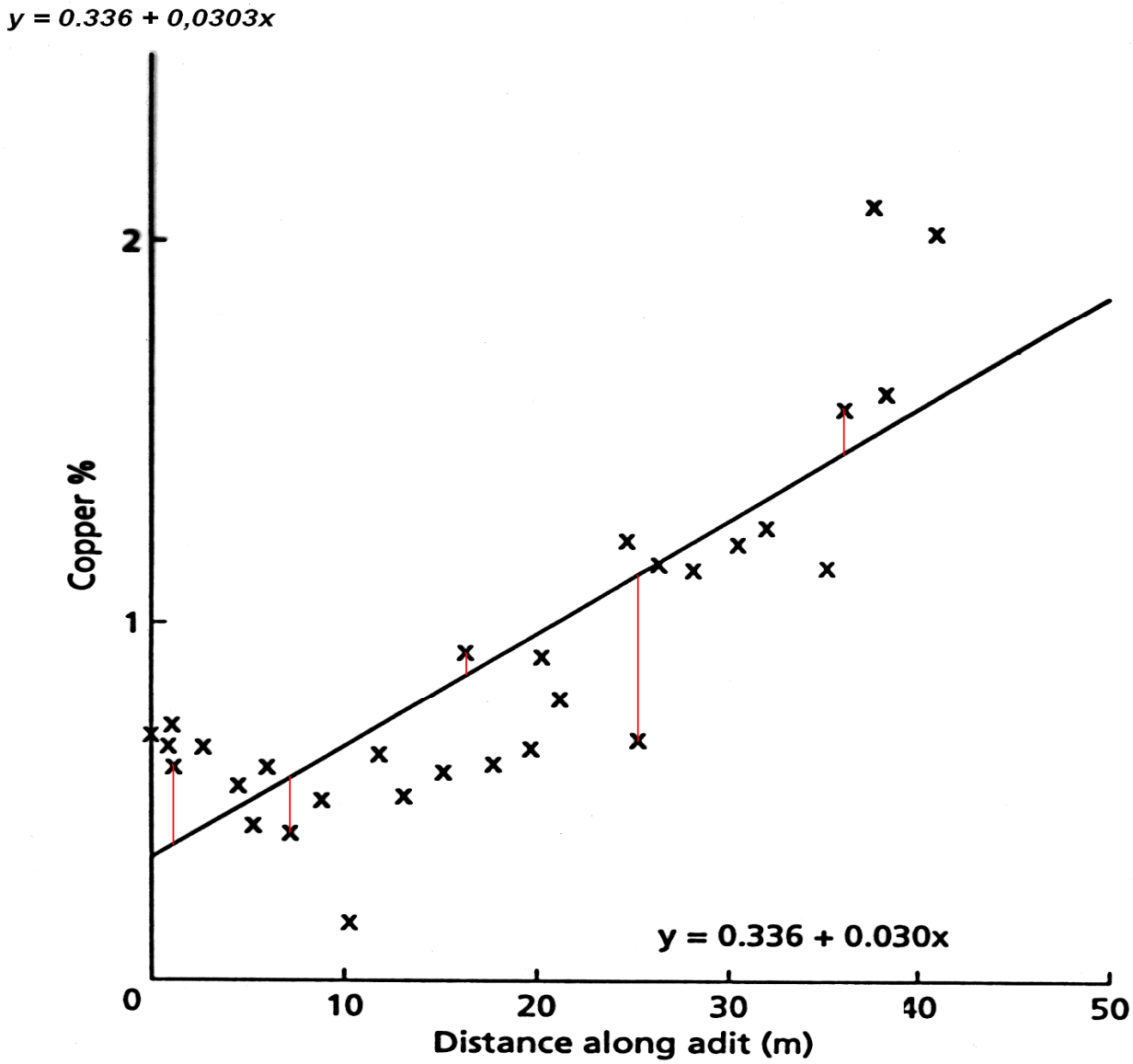


- Vsi postopki skušajo regresijo izračunati tako, da je odstopanje od regresijske krivulje ter napaka regresije čim manjša.
- Napako lahko definiramo različno:
  - Napaka izhaja le iz ene od obeh spremenljivk – klasična regresija.
  - Napaka izhaja iz obeh spremenljivk – strukturna (sestavljena) regresija.
  - Napako, ki izhaja iz ene od spremenljivk, zaradi ponovljenih meritev vnaprej poznamo – metoda tehtanih najmanjših kvadratov.

3.2.1. Klasična regresija

- Spremenljivki sta si zelo različni:
  - Neodvisna ali regresorska ali napovedovalna
    - Označena je z x
    - Nanašamo jo na absciso (vodoravno os)
    - Je točno izmerjena in ne vsebuje napake
    - Lahko jo določi uporabnik (npr. razdalje – x, vzdolž katerih merimo vrednost spremnljivke y).
  - Odvisna ali regresirana ali napovedana
    - Označena je z y
    - Nanašamo jo na ordinato (navpično os)
    - Meritve niso zanesljive in vsa odstopanja točk od regresorske linije (napaka) naj bi bile le v navpični smeri

3.2.1. Klasična regresija



3.2.1.1. Klasična linearna regresija

- Regresijsko enačbo za populacijo zapišemo:

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

kjer koeficient β_0 podaja sečišče premice z osjo y pri $x = 0$ in β_1 njen naklon. ε_i podaja odstopanje točk od premice.

3.2.1.1. Klasična linearna regresija

- Za vzorec velja:

  $y_i = b_0 + b_1 x_i$

  kjer sta b_0 in b_1 iz vzorca izhajajoči oceni β_0 in β_1.
- Iščemo taka b_0 in b_1 , da je celotna napaka $\Sigma\varepsilon_i^2$ čim manjša.
- Mera napake je vsota kvadratov zaradi odklona (SS_D):

$$SS_D = \sum_{i=1}^{n} \left(\hat{y}_i - y \right)^2$$

kjer je $\hat{y}_i$ ocena y – vrednost y, kjer navpičnica, ki poteka skozi točko, seka regresijsko premico.

3.2.1.1. Klasična linearna regresija

- Problem je rešljiv z dvema enačbama:

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i$$

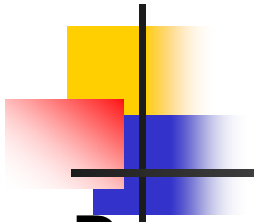
$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

- Koeficienta regresijske premice dejansko izračunamo:

$$b_1 = r \frac{s_y}{s_x} = \frac{VP_{xy}}{VK_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

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

3.2.1.1. Klasična linearna regresija



- Regresijo lahko pojmujemo kot razdelitev celotne variance podatkov na dva dela:
  - Varainca zaradi regresijske premice – pripisujemo jo nekemu (geološkemu) procesu.
  - Varianca zaradi odstopanja od regresijske premice - pripisujemo jo naključju.

- Obe kvantificiramo iz podatkov in izračunane regresijske premice:

3.2.1.1. Klasična linearna regresija

- Vsota kvadratov zaradi regresije:

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

- Vsota kvadratov zaradi odstopanja:

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

Kar običajno izračunamo iz $VK_D = VK_C - VK_R$

- Pri čemer je celotna vsota kvadratov:

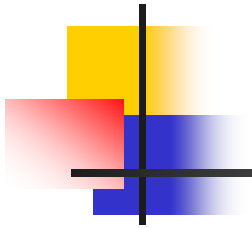
$$VK_C = SS_T = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

3.2.1.1. Klasična linearna regresija

- Kako dobra je regresija preverimo s statistiko ustreznosti prileganja R^2:

$$R^2 = \frac{VK_R}{VK_C}$$

3.2.1.1. Klasična linearna regresija



- Statistična značilnost regresijske premice je odvisna od razmerja varianc iz obeh virov.
- Značilnost je višja, če je varianca zaradi regresije večja in varianca zaradi napake manjša.
- Značilnost regresije opravimo z AVAR regresije, kjer postavimo hipotezi:

H_0: $\beta_1 = 0$ premica nima naklona

H_1: $\beta_1 \neq 0$ premica ima naklon – regresija obstaja

3.2.1.1. Klasična linearna regresija

- Regresija je statistično značilna na ravni zaupanja α, kadar izračunana vrednost F presega tabelirano.
- Tabela AVAR linearne regresije:

Vir	ν	VK	s^2	F	F_{tab}
regresija	1	VK_R	$s_R^2 = VK_R$	s_R^2 / s_D^2	$F_{\alpha,1,n-2}$
napaka	n-2	VK_D	$s_D^2 = VK_D/(n-2)$		
celotna	N-1	VK_C			

3.2.1.1. Klasična linearna regresija

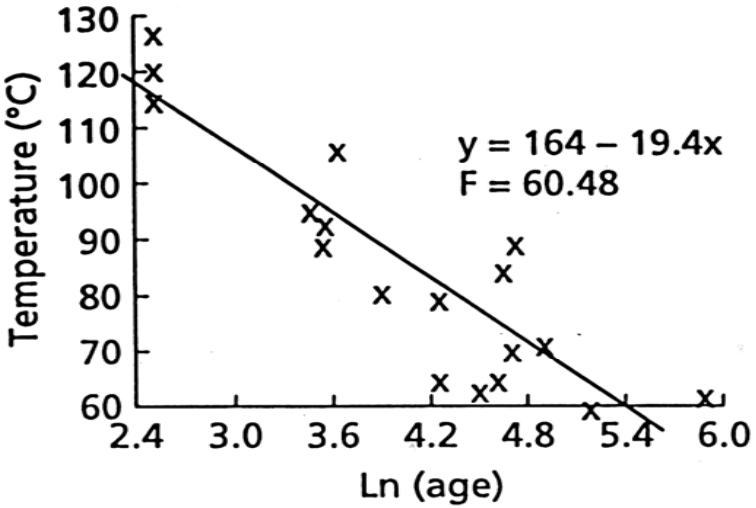
- Opozorila:

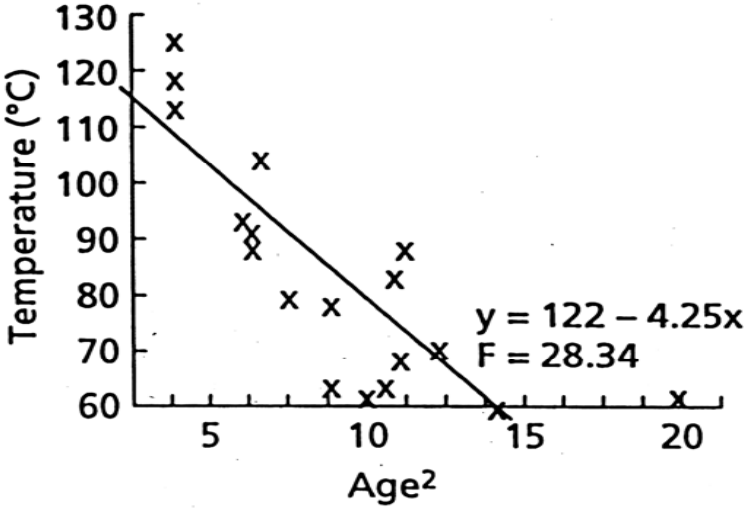
1. Točke, ki so outlierji (zlasti glede na neodvisno spremenljivko x) lahko prekomerno vplivajo na regresijo. Vpliv takih vrednosti preverimo tako, da jih odstranimo in ponovno izračunamo regresijo.
2. Razlika med točko in regresijsko premico v smeri y je preostanek. Veljavnost regresijskih statistik je odvisna od porazdelitve preostankov, ki :
    a. morajo biti normalno porazdeljeni
    b. morajo biti homoscedastični (brez trenda)
    c. Ne smejo biti avtokorelirani
3. Ne glede potrjeno statistično značilnost regresije, moramo paziti, kako uporabimo rezultat za napovedovanje. Ekstrapoliranje izven meja prvotnih podatkov ni priporočljivo, razen če poznamo (geološki) proces in vemo, da je to smiselno.

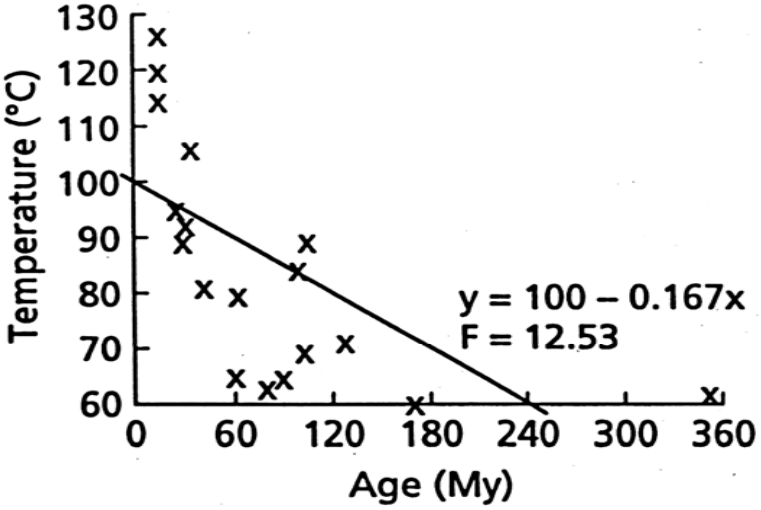
3.2.1.2. Krivuljna regresija

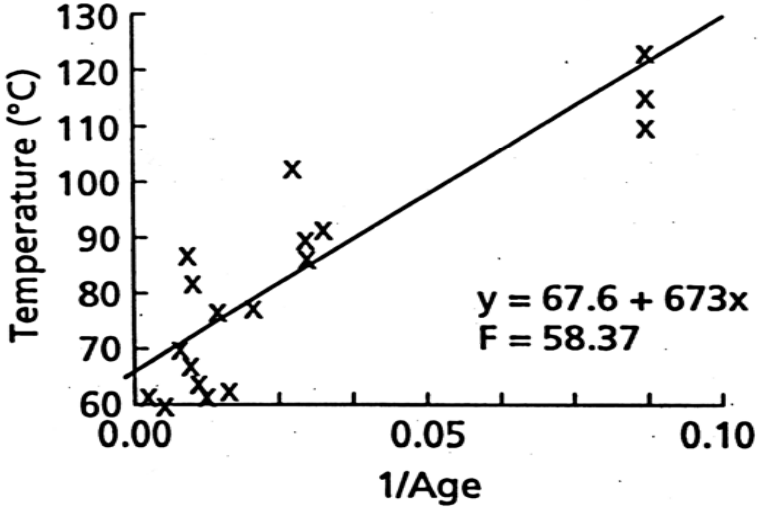
- Spremenljivki sta med seboj lahko povezani tudi s kako drugo krivuljo in ne le linearno premico.
- Prilagajanje takim krivuljam lahko ugotavljamo na dva načina:
  - Podatke transformiramo v želeno obliko krivulje (npr. logaritmiranje, kvadriranje...) in zanje izračunamo linearno regresijo
  - Uporabimo polinomno regresijo – izračunamo prilagajanje izbrani polinomni funkciji.

3.2.1.2. Krivuljna regresija









3.2.1.2. Krivuljna regresija

- Najpreprostejša polinomna regresija je kvadratna, ki opisuje parabolično ukrivljenje. Linerarni enačbi dodamo člen x^2:

  $y = b_0 + b_1x + b_2x^2$

- Značilnost kvadratne enačbe testiramo:

  H_0: $\beta_1 = 0$ in $\beta_2 = 0$

  H_1: $\beta_1 \neq 0$ in/ali $\beta_2 \neq 0$

3.2.1.2. Krivuljna regresija

- Kadar smo za iste podatke ugotovili značilnost linearne regresije, smo že zavrnili H_0 in avtomatično zavrnemo tudi H_0 kvadratne regresije. V tem primeru testiramo le koeficient β_2.
- Dodajanje člena vodi do enačbe kubične regresije:

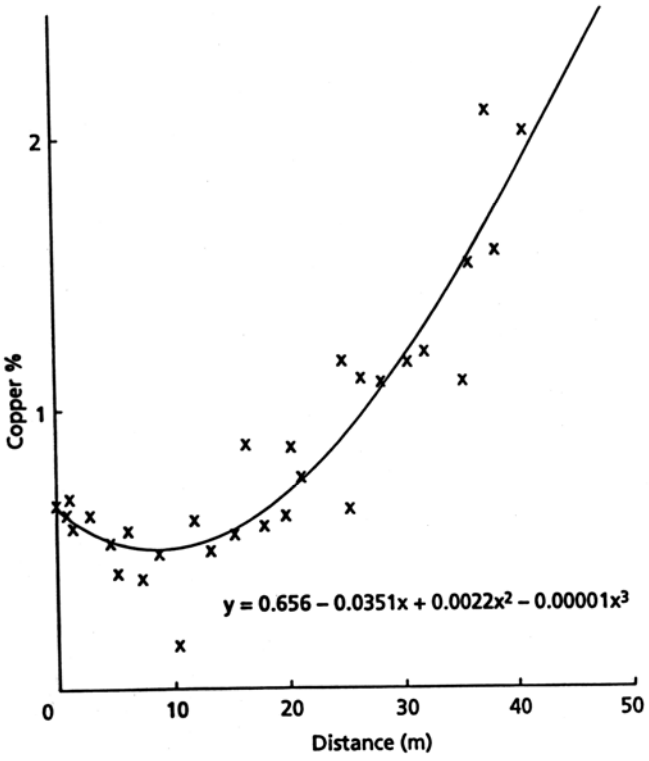
$$y = b_0 + b_1x + b_2x^2 + b_3x^3$$

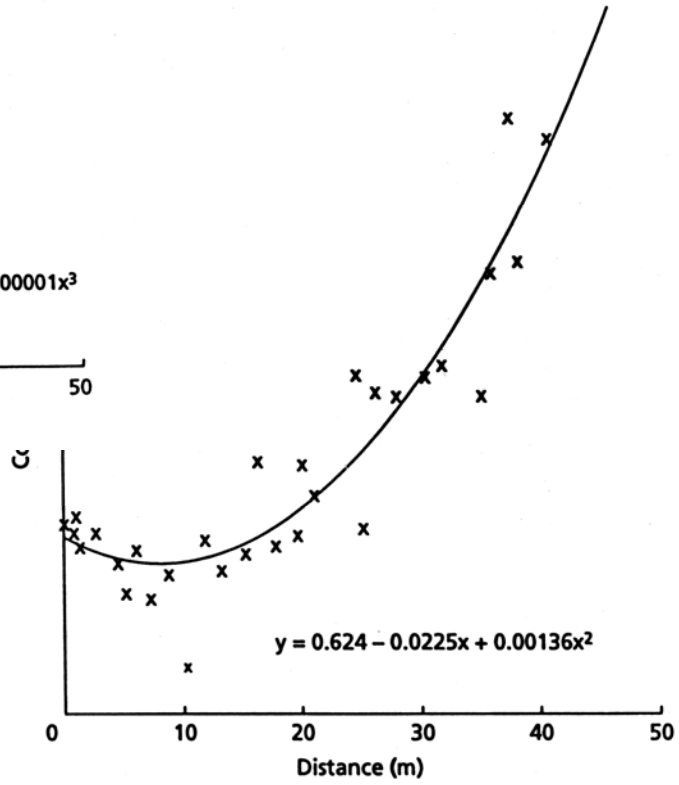
- In regresije na četrto potenco:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + b_4x^4$$

3.2.1.2. Krivuljna regresija

- Vsak dodatni člen da krivulji dodaten pomen. Kvadratna regresija nima točke prevoja, kubična dovoljuje eno, itd.
- Če zaporedno preverjamo enačbe, testiramo vedno le dodatni člen.
- Navadno hitro pridemo do točke, ko z dodajanjem člena nič ne pridobimo.





3.2.1.2. Krivuljna regresija

- Značilnost regresije preverimo z AVAR. H_0 zavrnemo, kadar izračunana vrednost presega kritično.
- Za polinomno regresijo reda k velja splošna tabela:

Vir	ν	VK	s^2	F	F_{tab}
k-ta regresija	k	VK_{Rk}	$s_{Rk}^2=VK_{Rk}/k$	s_{Rk}^2/ s_{Dk}^2	$F_{\alpha,k,n-k-1}$
(k-1)-ta regresija	k-1	VK_{Rk-1}			
zvišanje prileganja	1	$VK_{RZ}=$ $VK_{Rk}- VK_{Rk-1}$	$s_{RZ}^2=VK_{RZ}$	s_{RZ}^2/ s_{Dk}^2	$F_{\alpha,1,n-k-1}$
K-ta napaka	n-k-1	VK_{Dk}	$s_{Dk}^2=VK_{Dk}/(n-k-1)$		
celotna	n-1	VK_C			

3.2.1.2. Krivuljna regresija

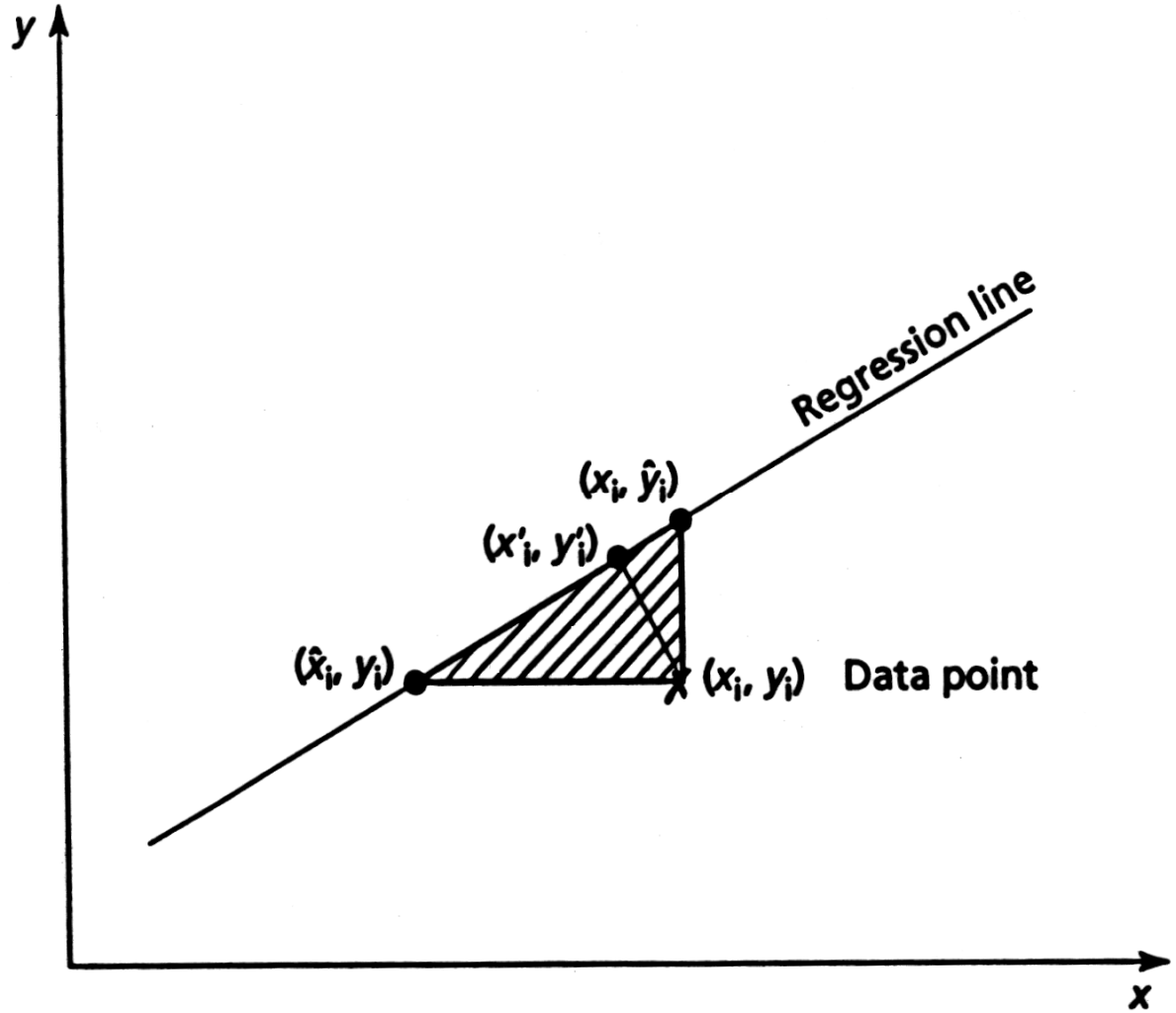
- Opozorila:
  - Veljajo vsa opozorila kot za linearno regresijo
  - Linearni in kvadratni členi so lastni številnim fizikalnim procesom, višje stopnje polinomov pa ne, zato je dobro prileganje višjim regresijam pogosto zgolj slučajno.
  - Visoke stopnje polinomov imajo lahko skrajne naklone, ki pri ekstrapolaciji vodijo do neverjetnih ocen.

3.2.2. Strukturna regresija

- Spremenljivki sta enakovredni – napaka izvira iz obeh (kar je v geologiji pogosto).
- Možna sta dva pristopa:
  - Glavne osi
  - Zmanjšane glavne osi (Reduced Major Axis – RMA)
- Oba sta v paleontologiji uporabna za ugotavljanje izometrije in alometrije.

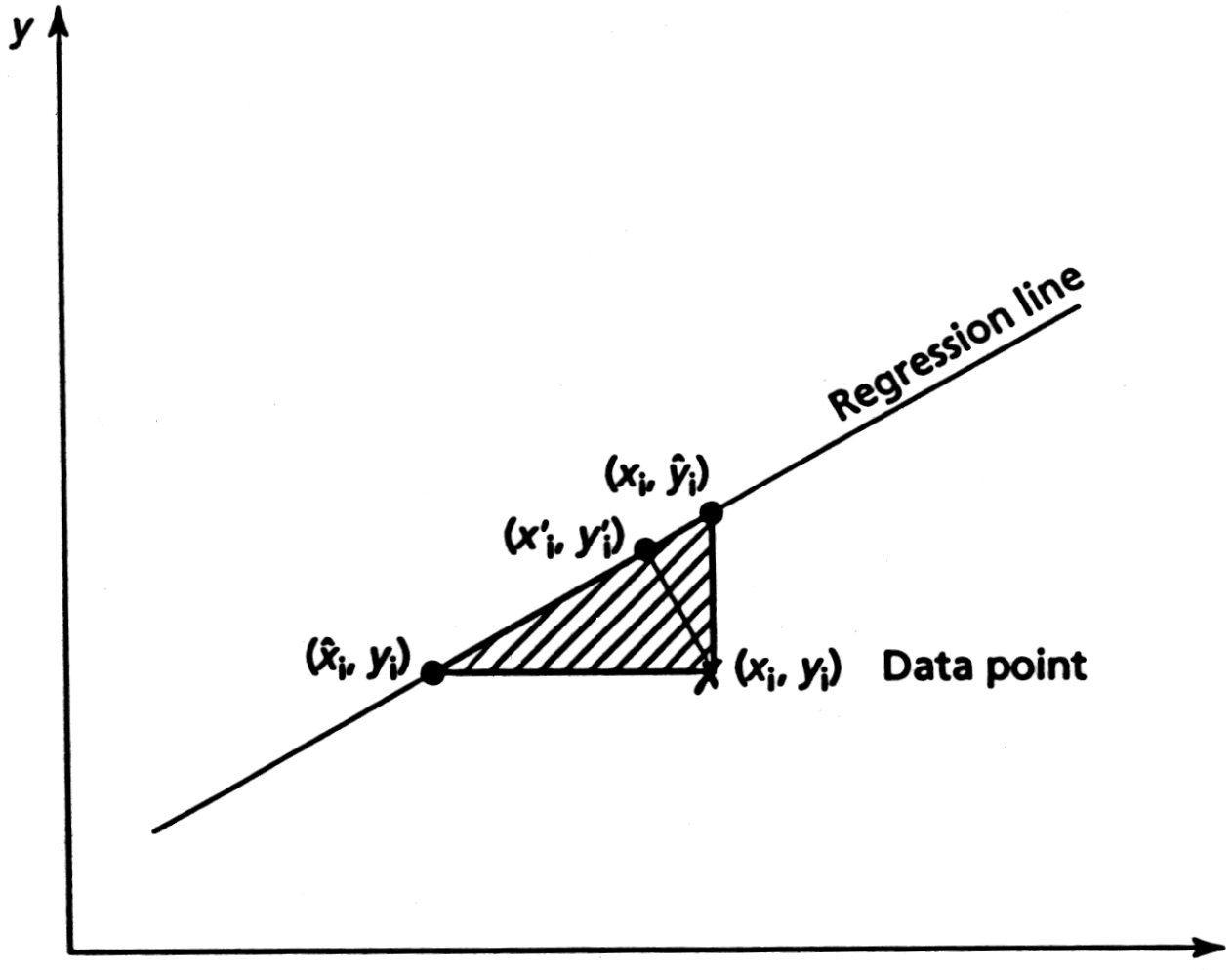
3.2.2.1. Glavne osi

- Črta glavne osi, ki poteka skozi roj podatkov, minimizira vsoto oddaljenosti med premico in vsako točko.
- Metoda je enakovredna multivariatni metodi glavnih komponent (PC).



3.2.2.2. Zmanjšane glavne osi (RMA)

- Regresijska premica je izračunana tako, da je vsota površin trikotnikov med točkami podatkov in premico čim manjša.



3.2.2.2. Zmanjšane glavne osi (RMA)

- Naklon b_1 izračunamo iz:

  $b_1 = s_x/s_y$

- Regresijska linija mora sekati skupno srednjo vrednost x in y , tako da za presečiščni koeficient b_0 velja:

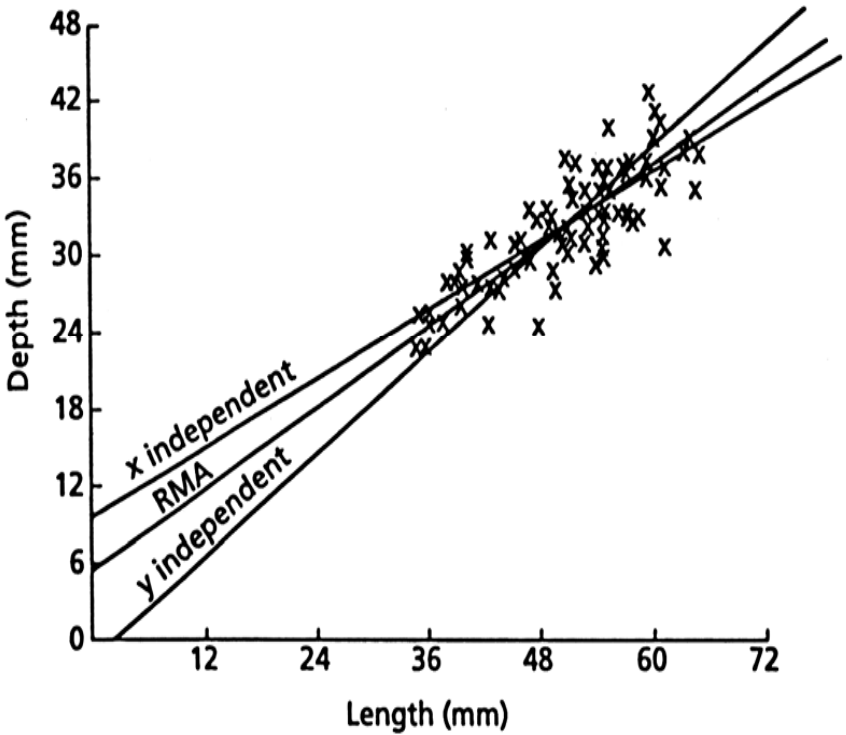
  $b_0 = y - b_1x$

3.2.2.2. Zmanjšane glavne osi (RMA)

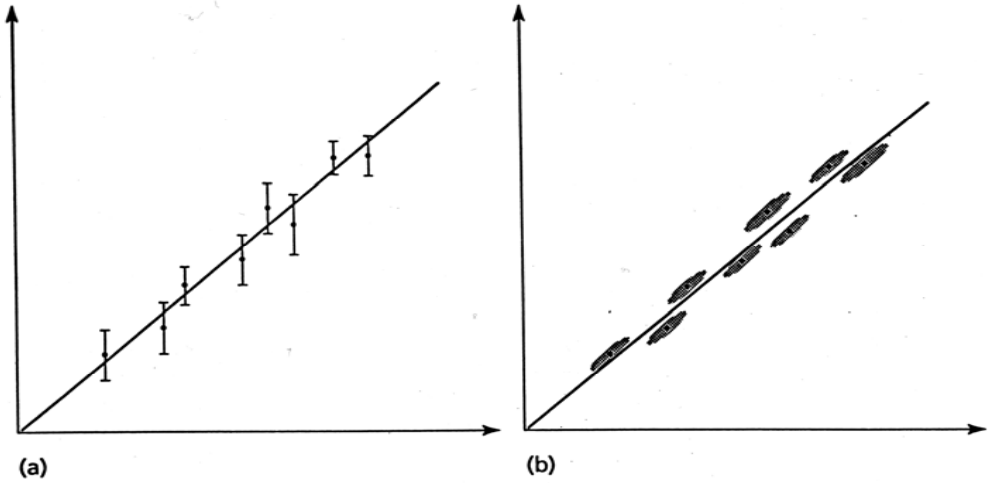
- Pri eliptičnem sipanju podatkov bo koeficient b_1 povezan z razmerjem osi elipse, ne glede na to, če dejansko ni naklona regresijske premice.
- Rezultat zato testiramo s standardno napako b_1 - s_e:

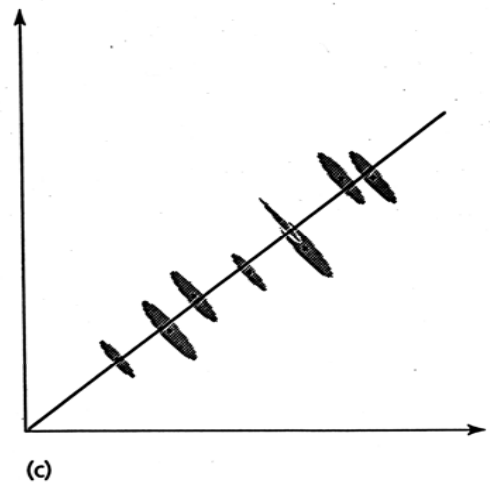
$$s_e = \sqrt{\frac{1 - r^2}{n}}$$

- Če razpon $b_1 \pm 1{,}96 \cdot s_e$ vključuje vrednost 0, to pomeni da premica nima naklona in da regresija na ravni zaupanja 5% ni značilna.



3.2.3. Metoda tehtanih najmanjših kvadratov





- Neodvisno moramo poznati napako v x in/ali y smeri:
  - Ponovljene meritve
  - Poznavanje natančnosti merilne naprave
- Regresijsko premico prilagajamo tako, da poteka čim bližje točkam, ki izkazujejo najmanjšo napako.
- Napaka x in y je lahko pozitivno ali negativno korelirana.