# Regresná a korelačná analýza (RaKA) resp. Korelačný počet

- **Skúmanie vzťahov medzi kvalitatívnymi znakmi, napr. A⇔B , nazýme *meranie asociácie***

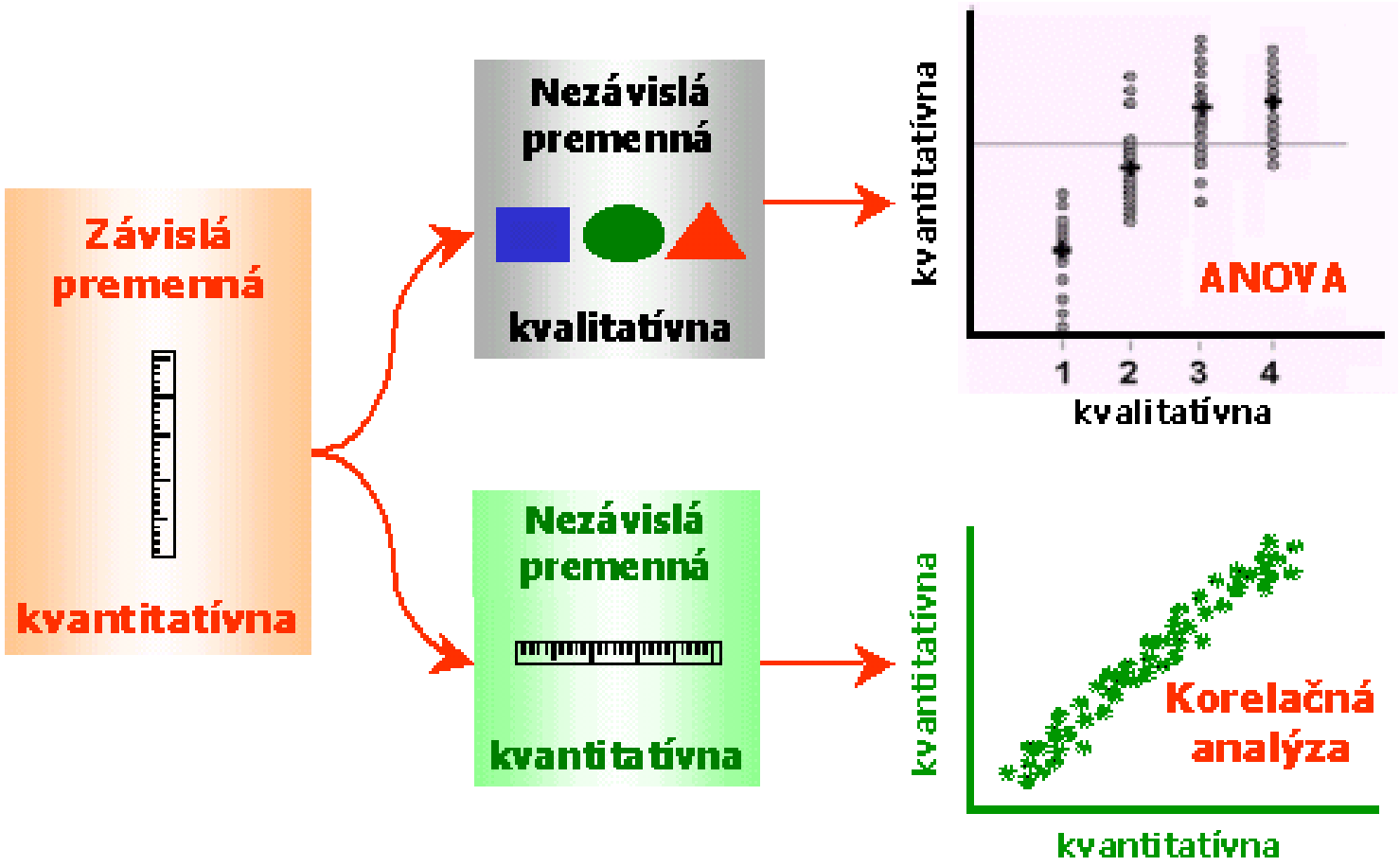
- **skúmanie vzťahov medzi kvantitatívnymi štatistickými znakmi - *regresná a korelačná analýza***

- **Skumanie vzťahov medzi výsledným kvalitatívnym znakom a kvantitatívnymi znakmi logistická regresia**

- **Skúmanie vzťahov medzi výsledným kvantitatívnym znakom a kvalitatívnymi znakmi *AR-analýza rozptylu***

- **Skúmanie závislosti medzi vysledným kvantitatívnym znakom a znakmi kvantitatívnymi a kvalitatívnymi *analýza kovariancie***

# Skúmanie vzťahov medzi štatistickými znakmi:

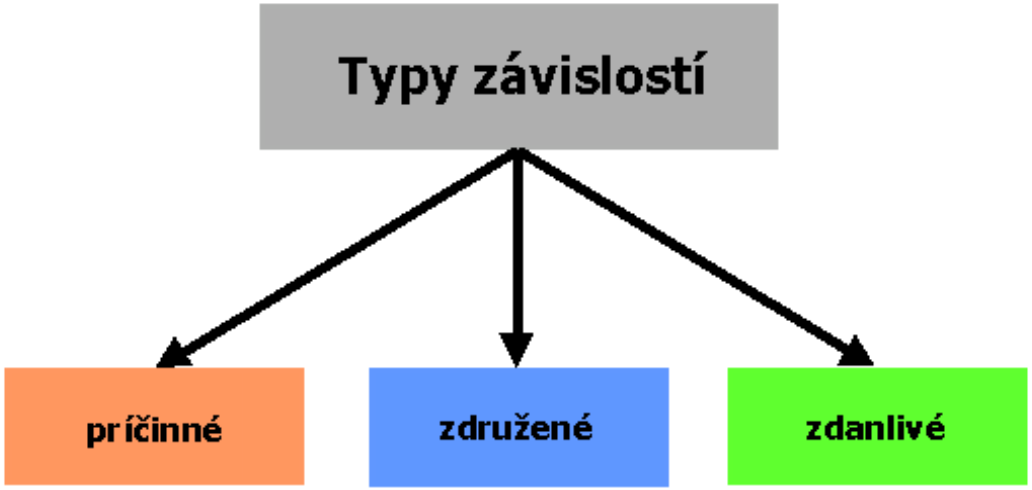


Vieme už ako analyzovať **rozdelenie premenných**, **testovať hypotézy, počítať intervaly spoľahlivosti a podobne.**

Zatiaľ sme sa ale venovali len jednotlivým hromadným javom separátne, no v skutočnosti hromadné javy neexistujú oddelene. Každý jav je výsledkom spolupôsobenia iných javov. Pričom charakter a významnosť vzájomného pôsobenia medzi javmi môžu byť rôzne.

**Predmetom skúmania v štatistike sú príčinné (kauzálne) závislosti**, keď jeden jav alebo skupina javov (**príčina**) vyvoláva iný jav alebo skupinu javov (**účinok**) takou závislosťou je napr. vzťah medzi spotrebou paliva a rýchlosťou.

# Typy závislostí



- príčinné
  - ak jeden jav, alebo skupina javov (príčina) vyvoláva iný jav alebo skupinu javov (účinok)
    - jednostranné - účinok nepôsobí spätne na príčinu
    - obojstranné- účinok a príčina na seba trvalé vzájomne pôsobia
- združené
  - nie sú to príčinné závislosti
    - určitej hodnote, obmene jedného javu spravidla zodpovedá určitá hodnota, obmena iného javu
    - dĺžka ramien - výška jednotlivca
- zdanlivé
  - vzťah medzi určitými javmi nie je dôsledkom ich vzájomnej príčinnej súvislosti
    - je výsledkom pôsobenia ďalšieho javu alebo javov
    - napr. výdavky na ovocie a výdavky na obuv

# Pri regresenej a korelečnej analýze

- skúmanie príčinnej - kauzálnej závislosti,
  skúmanie vzťahov medzi príčinou a účinkom
- kedy jeden resp viac javov (znakov, nezávisle premennných veličín ) vyvoláva účinok - výsledný jav - závisle premennú veličinu

$$Y = f(X_1 X_2 \ldots\ldots X_k, B_o, B_1, \ldots . B_p) + e$$

**Závislé premenná - účinok**

**Nezávislé premenné veličiny - príčiny**

**Neznáme parametre funkčného vzťahu**

**Náhodné, nešpecifikované vplyvy**

Hovoríme tiež o štatistickej alebo voľnej závislosti

# Príklady štatistickej - voľnej -závislosti

- **Skúmanie závislosti spotreby bravčového mäsa od príjmu, ceny mäsa bravčového ceny mäsa hovädzieho a hydiny a od tradície, resp. ďalších nešpecifikovaných, či náhodných vplyvov.**

- **Skúmanie pridanej hodnoty  resp. HDP od vstupov: práce a kapitálu….**

- **Skúmanie závislosti výživy obyvateľstva od stupňa ekonomického rozvoja krajiny….**

# Opakom štatistickej závislosti je funkčná závislosť

$$Y = f(X_1 X_2 \ldots\ldots X_k, B_o, B_1, \ldots, B_p)$$

kedy je závisle prememnná veličina jednoznačne určená funkčným vzťahom,

príklady z fyziky, chémie - takýto druh vzťahov nie je predmetom štatistického skúmania

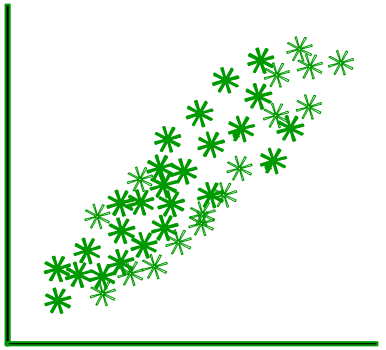
# Regresná a korelačná analýza (RaKA)

- Dve základné úlohy RaKA:
  - regresná úloha (RÚ) jej podstatou je

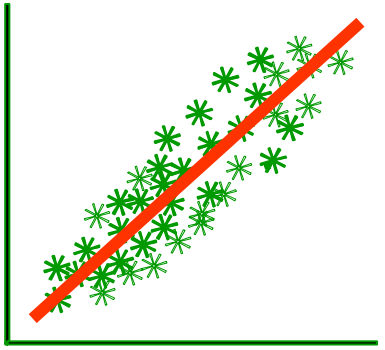
nájsť funkčný vzťah podľa ktorého sa mení závislé premenná so zmenou nezávisle premenných - nájsť vhodnú regresnú funkciu.Súčasne je potrebné odhadnúť parametre regresnej funkcie.

  - korelačná úloha (KÚ)- merať tesnosť - silu skúmanej závislosti.

- **Korelačná úloha**
  - umožňuje kvantifikovať silu lineárnej závislosti
  - vypovedá len o tom, že nejaký typ závislosti medzi premennými existuje
  - nedefinuje jeho presnú podobu



- **Regresná úloha**
  - popisuje tvar závislosti medzi premennými
  - poskytuje odpovede na komplexné otázky o vzájomnom vplyve premenných



- **Regresná analýza**
  - **umožňuje popísať vzťah medzi dvomi alebo viacerými premennými**
  - cieľ regresnej analýzy
    - **vysvetliť adekvátnu časť variability primárnej premennej pomocou jej vzťahu s jednou alebo viacerými sekundárnymi premennými**
    - **odhadnúť hodnoty primárnej premennej pri konkrétnych hodnotách sekundárnych premenných**
  - typy premenných v regresnej analýze
    - **závislé premenné**
      - **sú v centre pozornosti, pretože ich variabilitu sa snažíme vysvetliť**
      - **tzv. vysvetľované premenné, endogénne premenné**
    - **nezávislé premenné**
      - **sú premenné, ktoré používame na vysvetlenie zmien v hodnotách závislej premennej, exogénne premenné**
      - **predpokladáme, že ich hodnoty sa nemenia**
      - **tzv. vysvetľujúce premenné alebo tiež regresory**

- **typy regresnej analýzy podľa počtu premenných**
  - jednoduchá regresia
    - ak popisujeme závislosť jednej kvantitatívnej závislej premennej od jednej kvantitatívnej nezávisle premennej
  - viacnásobná regresia
    - ak popisujeme závislosť jednej kvantitatívnej závislej premennej od viacerých kvantitatívnych nezávislých premenných
  - viacrozmerná regresia
    - ak popisujeme závislosť viacerých kvantitatívnych závisle premenných od viacerých kvantitatívnych nezávisle premenných pomocou viacerých rovníc
- **typy regresnej analýzy podľa typu závislosti**
  - lineárna regresia
    - ak popisujeme závislosť premenných pomocou priamky
  - nelineárna regresia
    - ak popisujeme závislosť premenných pomocou inej krivky ako priamka
  - logistická regresia
    - ak popisujeme závislosť diskrétnej alebo kvalitatívnej závisle premennej od jednej alebo viacerých kvantitatívnych nezávisle premenných

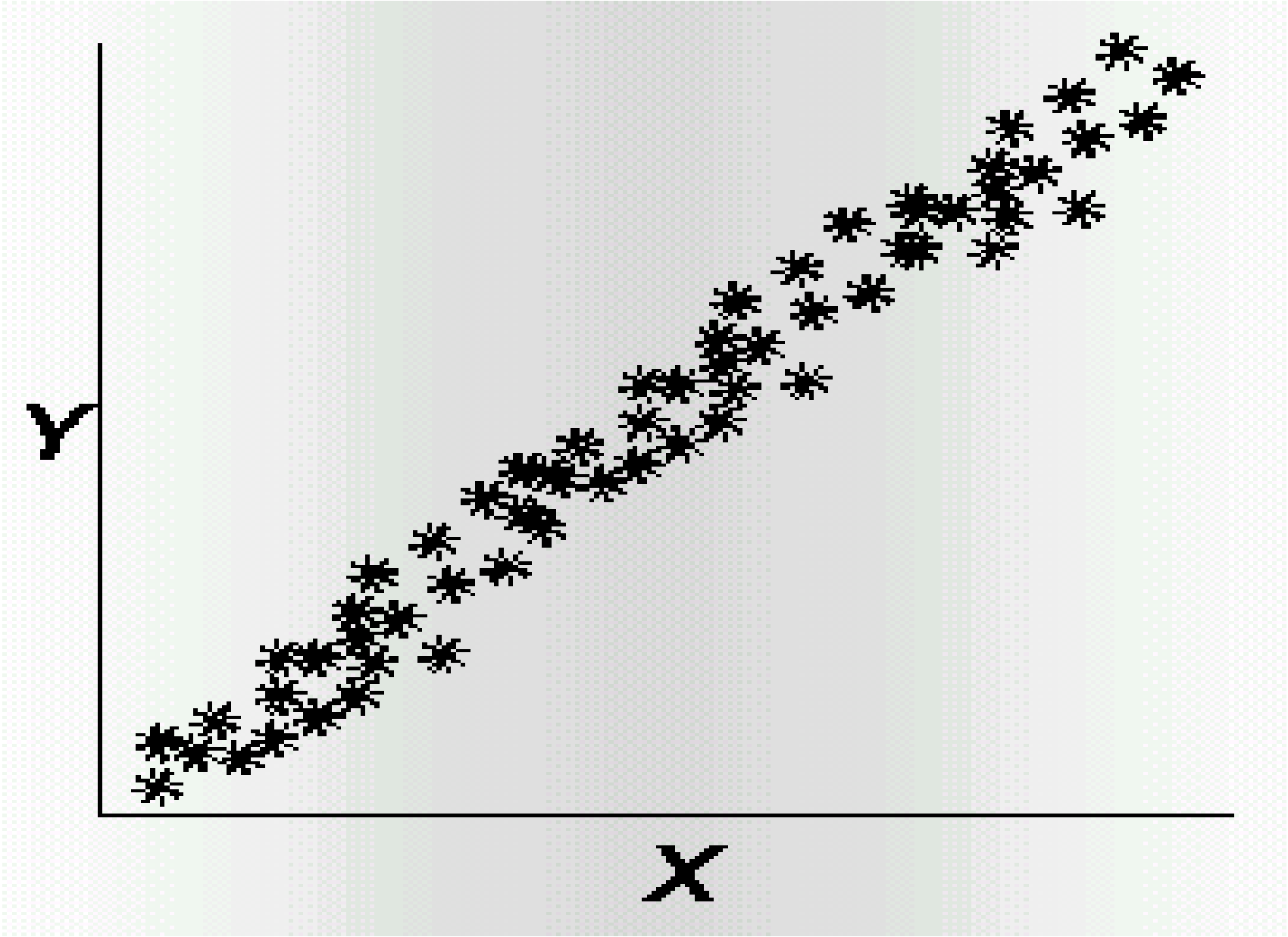
•Na preskúmanie a popísanie závislostí medzi kvantitatívnymi znakmi samozrejme potrebujeme konkrétne nástroje.
Môžeme ich rozdeliť do dvoch skupín:

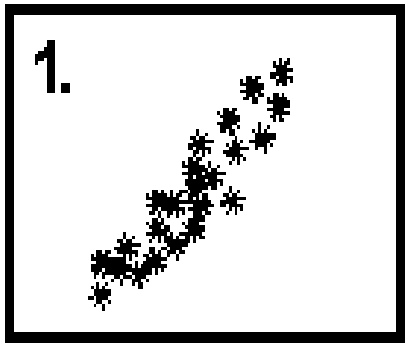
- **grafické –** na úvodné preskúmanie
- **štatistické –** na popísanie sily a tvaru závislosti

Základným grafickým nástrojom je **bodový graf,** v ktorom na os x a y nanesieme hodnoty jednotlivých premenných, ktorých závislosť chceme analyzovať.

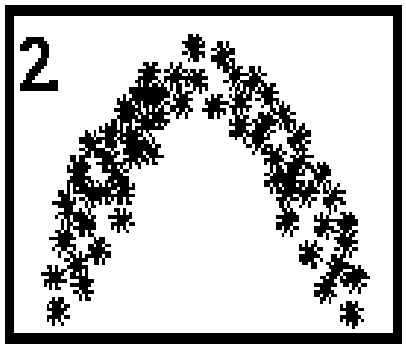
**Bodový graf slúži na:**

- úvodné preskúmanie vzťahov medzi dvomi premennými
- určenie extrémnych alebo typických hodnôt
- určenie možného tvaru závislosti
- porovnanie a prezentáciu výsledkov analýzy


Y
X

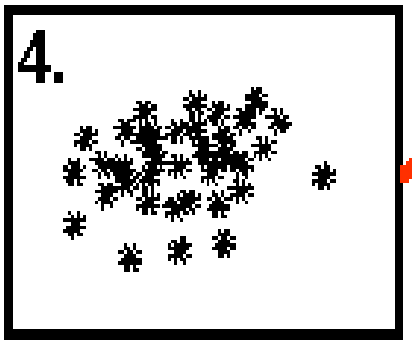



vzťah možno popísať priamkou




vzťah možno popísať krivkou


vzťah možno popísať polynómom




medzi premnnými neexistuje jasný vzťah

Presnú odpoveď poskytne výpočet korelačných štatistík

Bodový graf je dôležitým nástroj už v prvej fáze analýzy závislostí medzi dvomi premennými. Umožňuje totiž:

- **zistiť, či závislosť existuje**
- **zistiť, aký charakter má závislosť – aký tvar má závislosť**

Vyššie sú uvedené príklady, ako môže vyzerať bodový graf pre rôzne typy príčinných závislostí

1.Závislosť medzi premennými existuje a môžeme ju pravdepodobne popísať pomocou priamky
2.Existuje závislosť, ktorú možno popísať pomocou krivky – kvadratickej funkcie
3.Medzi premennými existuje cyklická – sezónna závislosť
4.Medzi premennými neexistuje jednoznačná jasná závislosť.

***Bodové grafy nám vždy slúžia na získanie základnej predstavy. Každý analytik však v grafe môže vidieť niečo iné. Presné potvrdenie našich domnienok nám poskytnú až exaktné štatistické nástroje.***

Po úvodnom grafickom preskúmaní nastupuje fáza hľadania presných štatistík, ktoré potvrdia odhady z grafov. Pre tieto účely používame štatistické nástroje korelačnej analýzy.

Korelačné štatistiky zisťujú **či medzi premennými existuje korelácia aká je sila korelácie. Koreláciou nazývame vzájomný lineárny vzťah – závislosť dvoch premenných.** Tento vzťah môže byť:

**priamy –** s rastúcimi hodnotami jednej premennej rastú aj hodnoty druhej premennej

**nepriamy –** s rastúcimi hodnotami jednej premennej klesajú hodnoty druhej premennej

Ak medzi hodnotami dvoch premenných neexistuje ani priama ani nepriama lineárna závislosť, hovoríme, že sú **nekorelované.**

## Aké štatistiky merajú lineárnu závislosť?

**Prvou mierou, ktorú používame, aby sme potvrdili alebo vyvrátili existenciu lineárnej závislosti (korelácie) je *kovariancia.***
***Kovariania*** **sa vypočíta ako:**

$$\operatorname{cov} xy = \frac{1}{n}\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

Zo spôsobu výpočtu možno odvodiť, kedy potvrdzuje existenciu pozitívnej, negatívnej korelácie a kedy nekorelovanosti.
Ak kovariancia potvrdí neexistenciu lineárneho vzťahu, medzi premennými môže existovať nelineárny vzťah.
Ak kovariancia potvrdí existenciu lineárneho vzťahu, môžeme merať jeho intenzitu

- **cov xy = 0**, medzi premennými ***nie je lineárny vzťah***
- **cov xy > 0,** medzi premennými ***je priamy lineárny vzťah***
- **cov xy < 0,** medzi premennými ***je nepriamy lineárny vzťah***

# Jednoduchá lineárna regresná a korelačná analýza

- Uvažujme štatistický znak ***X*** a ***Y,*** medzi ktorými je v základnom súbore lineárna závislosť

$$Y = B_o + B_1 X + e$$

bodovým odhadom tejto regresnej funkcie je priamka

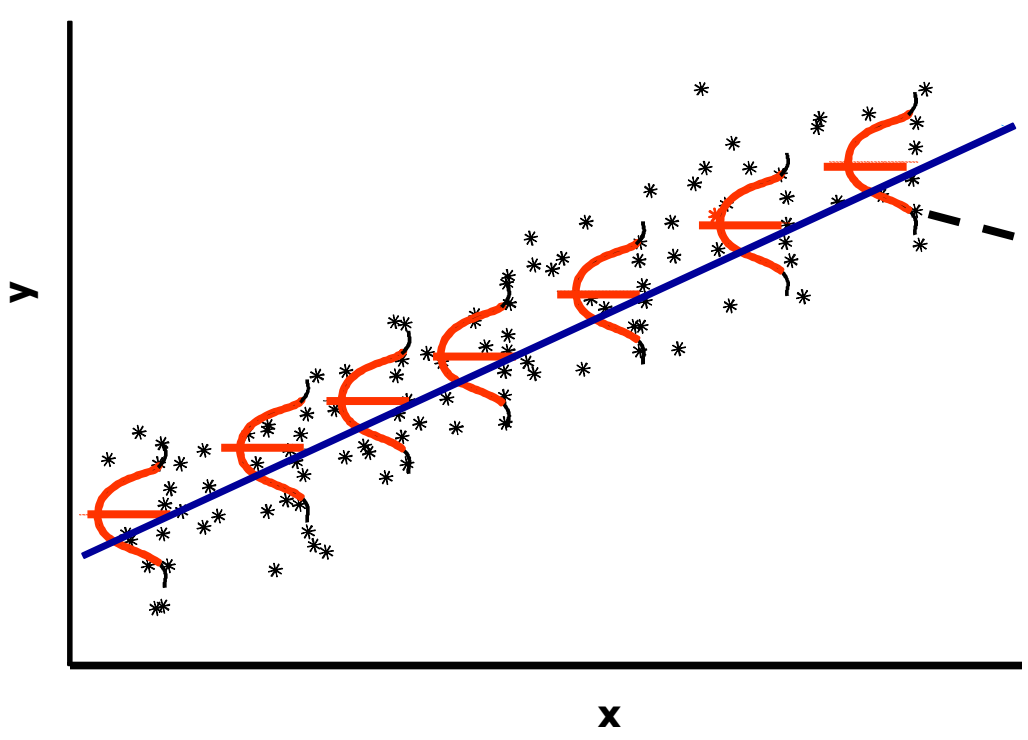
$$y_j = b_0 + b_1 x_j + e_j \ ,$$

ktorej koeficienty vypočítame z výberových údajov

***Akú metódu použiť ???***

- Závislosť v ZS

**Stredné hodnoty závislej premennej v ZS ležia na priamke, ktorú sa snažíme odhadnúť**




**Každej hodnote X zodpovedajú rôzne hodnoty Y. Môžeme vypočítať strednú hodnotu Y = $\mu_Y$ pre dané X.**

**Vzťah medzi premennými nie je funkčne lineárny. Hodnoty (x,y) neležia na priamke, ale majú tendenciu ju vytvárať. Je to štatistická lineárna závislosť**

**Popis závislosti v ZS**

rovnica modelu $Y = \beta_0 + \beta_1 X + \varepsilon$

kde

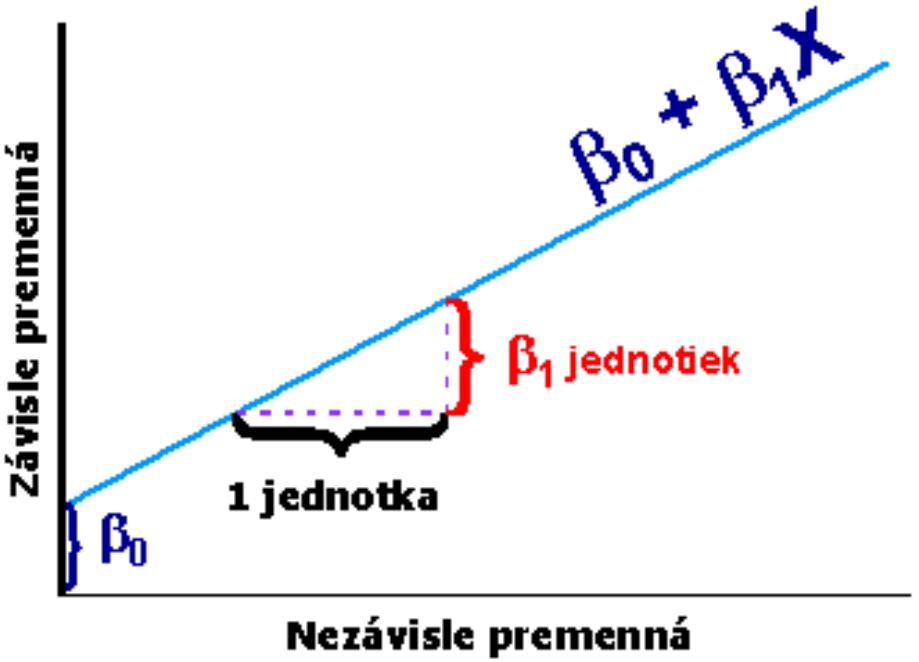
Y je závisle premenná

X je nezávisle premenná

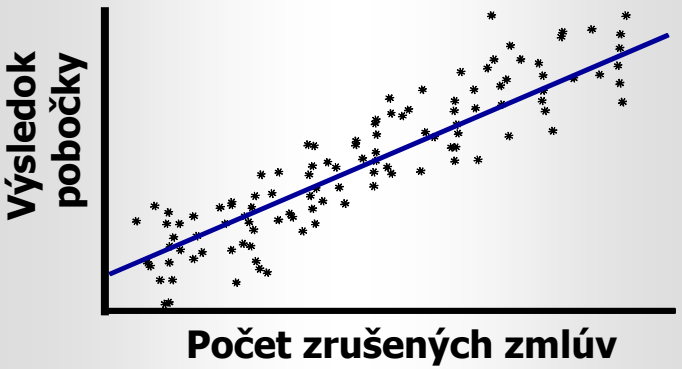
$\beta_0$ je parameter modelu tzv. lokujúca konštanta ktorá vyjadruje akú hodnotu nadobudne premenná Y, ak premenná X bude mať hodnotu 0

$\beta_1$ je parameter modelu tzv. regresný koeficient, ktorý vyjadruje sklon regresnej priamky.

Udáva o koľko jednotiek sa zmení Y, ak sa X zmení o jednotku.



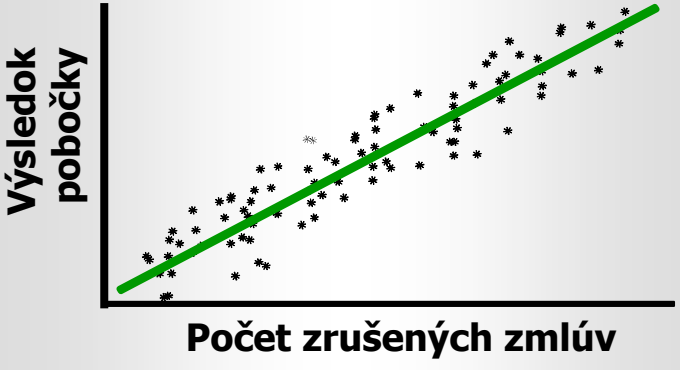
## Základný súbor




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

$$\mu_Y = \beta_0 + \beta_1 X$$

## Výberový súbor

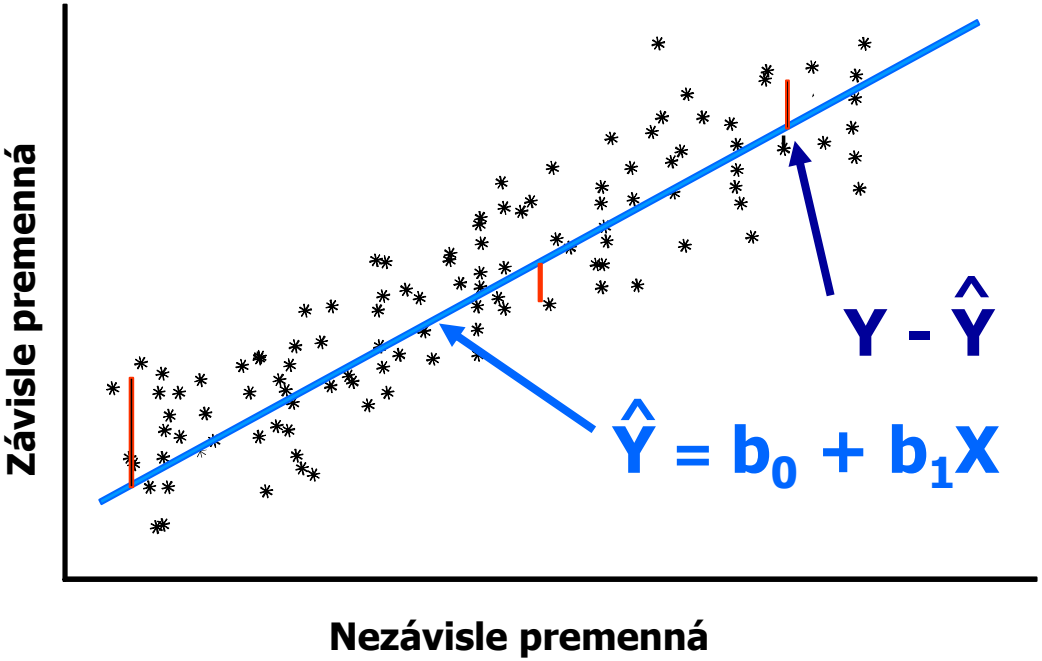



$$\hat{Y} = b_0 + b_1 X$$

$$Y = \text{est}(\mu_Y)$$
$$b_0 = \text{est}(\beta_0)$$
$$b_1 = \text{est}(\beta_1)$$

**Metóda najmenších štvorcov**

metóda odhadu regresného modelu
odhad MNŠ minimalizuje sumu štvorcov reziduálnych odchýlok
= rozdielov medzi skutočnou hodnotou a odhadnutou priamkou
priamka odhadnutá MNŠ je ku všetkým skutočným hodnotám tak blízko ako sa len dá




Vypočítané hodnoty – vyrovnané -- pomocou modelu budeme označovať

$$\hat{Y} = y'$$

# Korelačná úloha korelačného počtu

- Skúmať tesnosť - silu -  závislosti
- k tomu slúžia miery tesnosti závislosti
- požadujeme, aby sa pohybovali v pevne ohraničanom intervale,
- a aby vrámci intervalu rástli s vyššiou silou závislosti

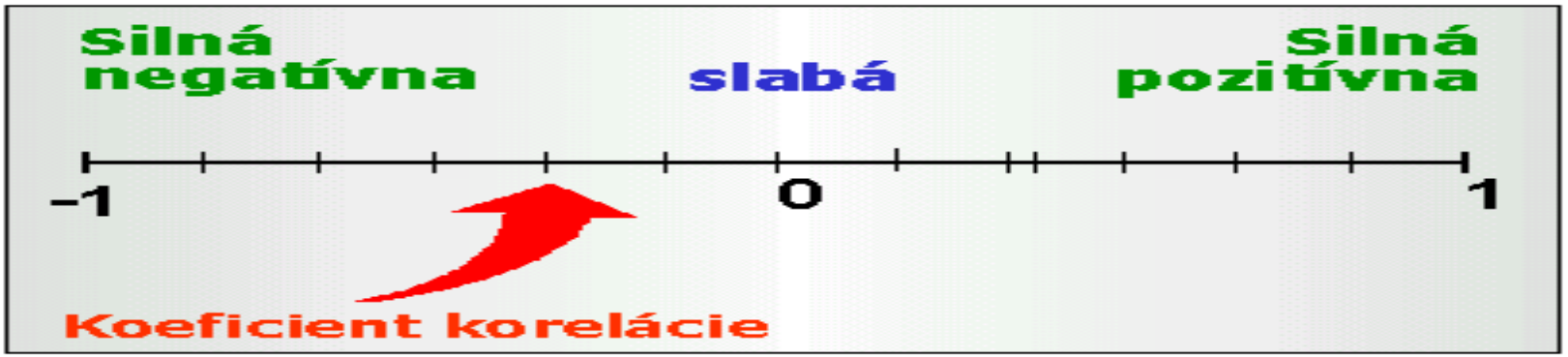
Korelačná analýza predstavuje súhrn metód a postupov, pomocou ktorých overujeme vypovedaciu schopnosť kvantifikovaných regresných modelov ako celku aj jeho častí. Overovanie vypovedacej schopnosti kvantifikovaných regresných modelov vedie k výpočtu číselných charakteristík, ktoré v koncentrovanej forme popisujú kvalitu vypočítaných modelov.

Na zmeranie **intenzity alebo tiež stupňa lineárnej závislosti** medzi dvomi premennými používame koeficient korelácie. Najčastejšie sa používa tzv. **Pearsonov koeficient korelácie**. Hodnoty koeficienta korelácie sú: z intervalu $-1$ až 1

bližšie k jednej z hraníc, ak medzi premennými je vysoký stupeň lineárnej závislosti **blízke 0, ak medzi premennými nie je lineárna závislosť** blízke 1, ak medzi premennými je silná pozitívna lineárna závislosť **blízke –1, ak medzi premennými je silná negatívna závislosť**

Pearsonov koeficient korelácie vypočítame podľa vzorca:

$$r = \frac{\mathrm{cov}\, xy}{s_x . s_y}$$



Je zrejmé, že všeobecne platí vzťah

$$C = V + N$$

$C = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2$ je celkový súčet štvorcov odchýlok

$V = \sum_{j=1}^{n} \left(y'_j - \bar{y}\right)^2$ je vysvetlený súčet štvorcov odchýlok

$N = \sum_{j=n}^{n} \left(y_j - y'_j\right)^2$ je nevysvetlený (reziduálny) súčet štvorcov odchýlok.

# koeficient korelácie $r_{yx}$

$$r_{yx} = \sqrt{\frac{\sum_{j=1}^{n}(y_j' - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}} = \sqrt{\frac{V}{C}}$$

Koeficient determinácie $r_{yx}^2$

$$r_{yx}^2 = \frac{C-N}{C} = 1 - \frac{N}{C} = 1 - \frac{\sum_{j=1}^{n}\left(y_j - y_j'\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}$$

Koeficient determinácie môže nadobúdať hodnoty z intervalu 0 až 1, čím viac sa hodnota koeficienta blíži k jednotke, tým väčšia časť celkovej variability je modelom vysvetlená a naopak, ak sa koeficient determinácie blíži k nule, tým menšia časť celkovej variability je modelom vysvetlená.

Koeficient determinácie sa bežne používa ako kritérium pri rozhodovaní o voľbe konkrétneho tvaru regresnej funkcie. Ak však majú regresné funkcie rôzny počet parametrov je potrebné upraviť koeficient determinácie do korigovanej podoby v tvare:

$$r_{kor}^2 = 1 - \frac{(n-1)\sum_{j=1}^{n}\left(y_j - y'_j\right)^2}{(n-p)\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}$$

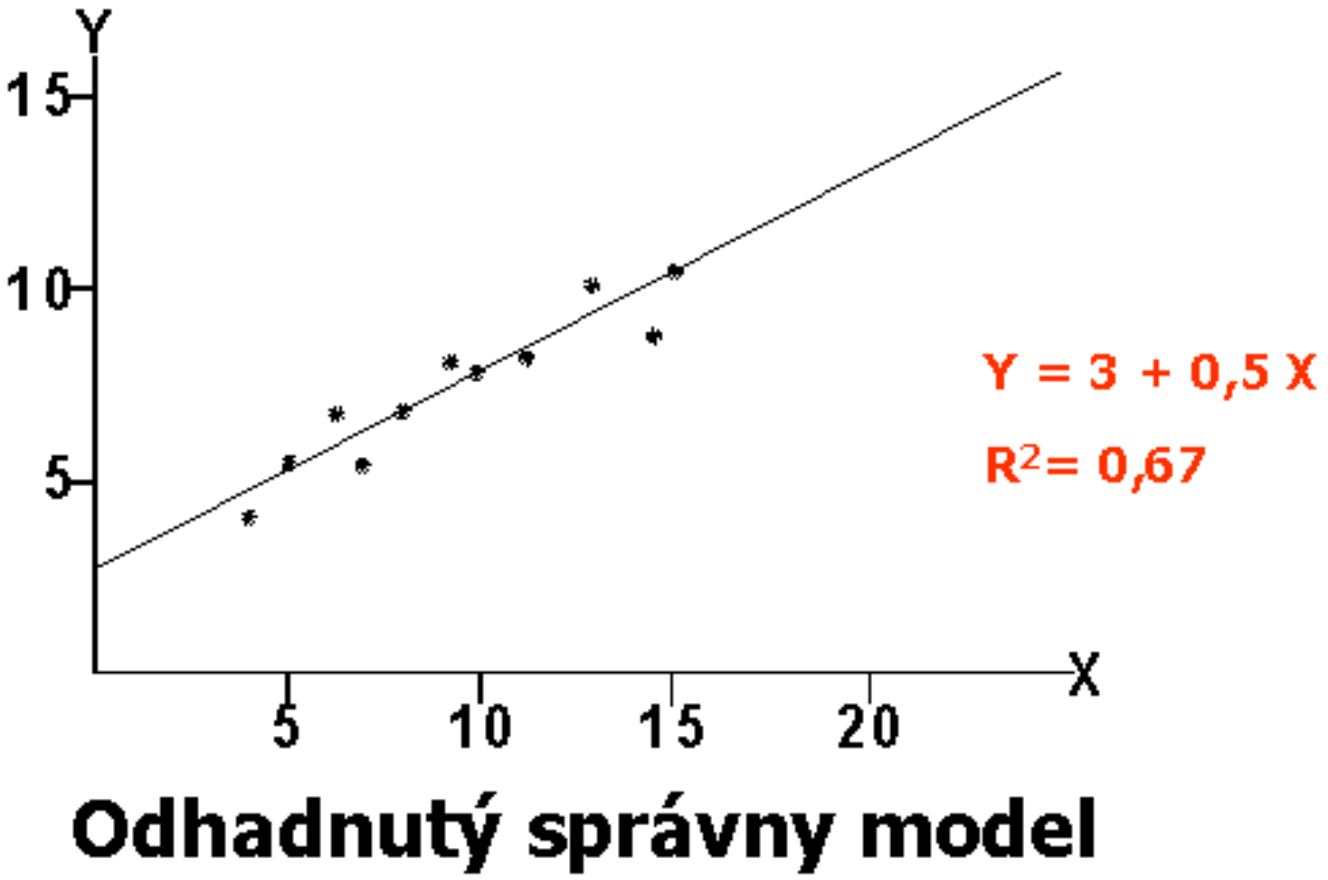
| Variabilita | Súčet štvorcov odchýlok | Stupne voľnosti | Rozptyl | F test |
|---|---|---|---|---|
| **Vysvetlená** | $V = \sum_{j=1}^{n} \left( y'_j - \bar{y} \right)^2$ | $p - 1$ | $s^2_{y'} = \frac{V}{p-1}$ | $F = \frac{s^2_{y'}}{s^2_r}$ |
| **Nevysvetlená** | $N = \sum_{j=n}^{n} \left( y_j - y'_j \right)^2$ | $n - p$ | $s^2_r = \frac{N}{n-p}$ | |
| **Celková** | $C = \sum_{j=1}^{n} \left( y_j - \bar{y} \right)^2$ | $n - 1$ | | |

Testovacie kritérium v tabuľke je možné využiť k súčasnému testovaniu významnosti celého regresného modelu, indexu determinácie aj indexu korelácie. Vypočítanú hodnotu F testu porovnávame s kvantilom F rozdelenia a p-1 a n – p stupňov voľnosti, t.j. $F_\alpha[(p-1),(n-p)]$

ak F $< F_\alpha[(p-1),(n-p)]$ považujeme regresný model za nevýznamný, podobne aj index determinácie a index korelácie.

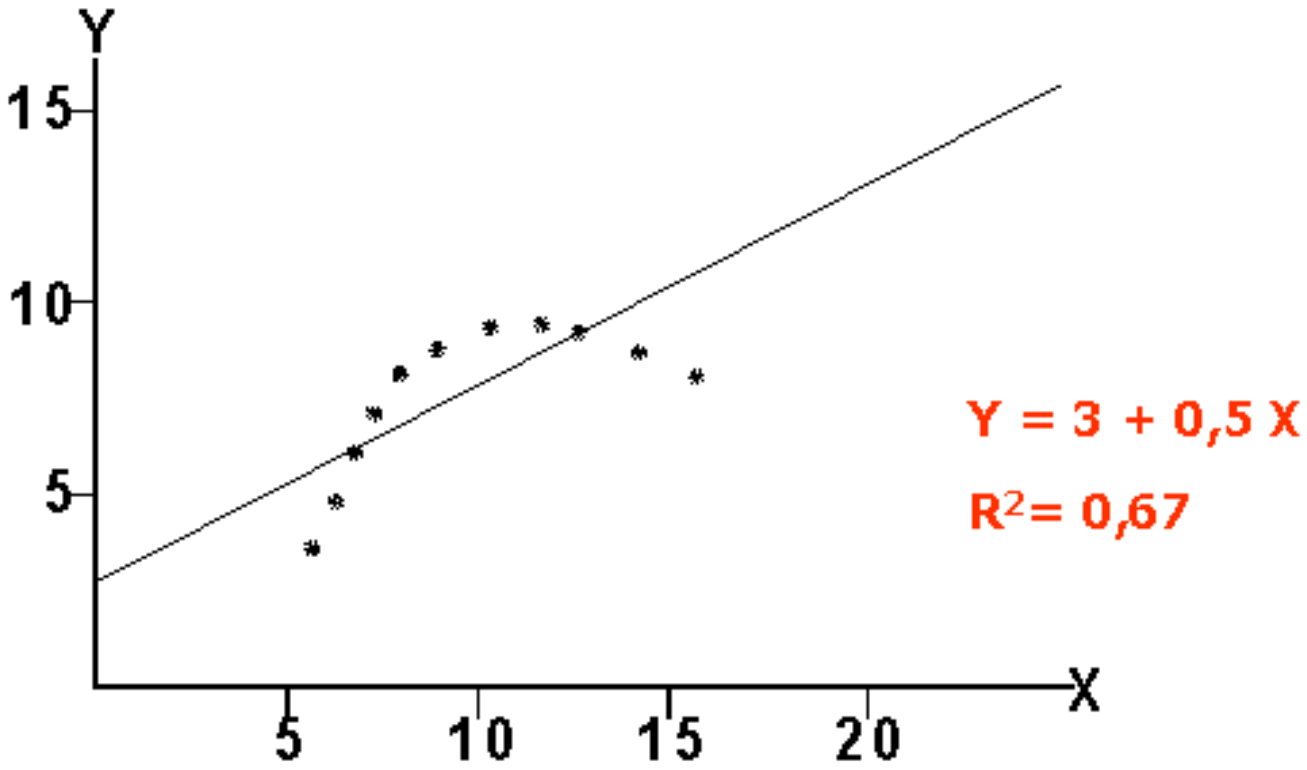
ak F > $F_\alpha[(p-1),(n-p)]$ považujeme regresný model za štatisticky významný, podobne aj index determinácie a index korelácie

# Nedostatky mier kvality odhadu



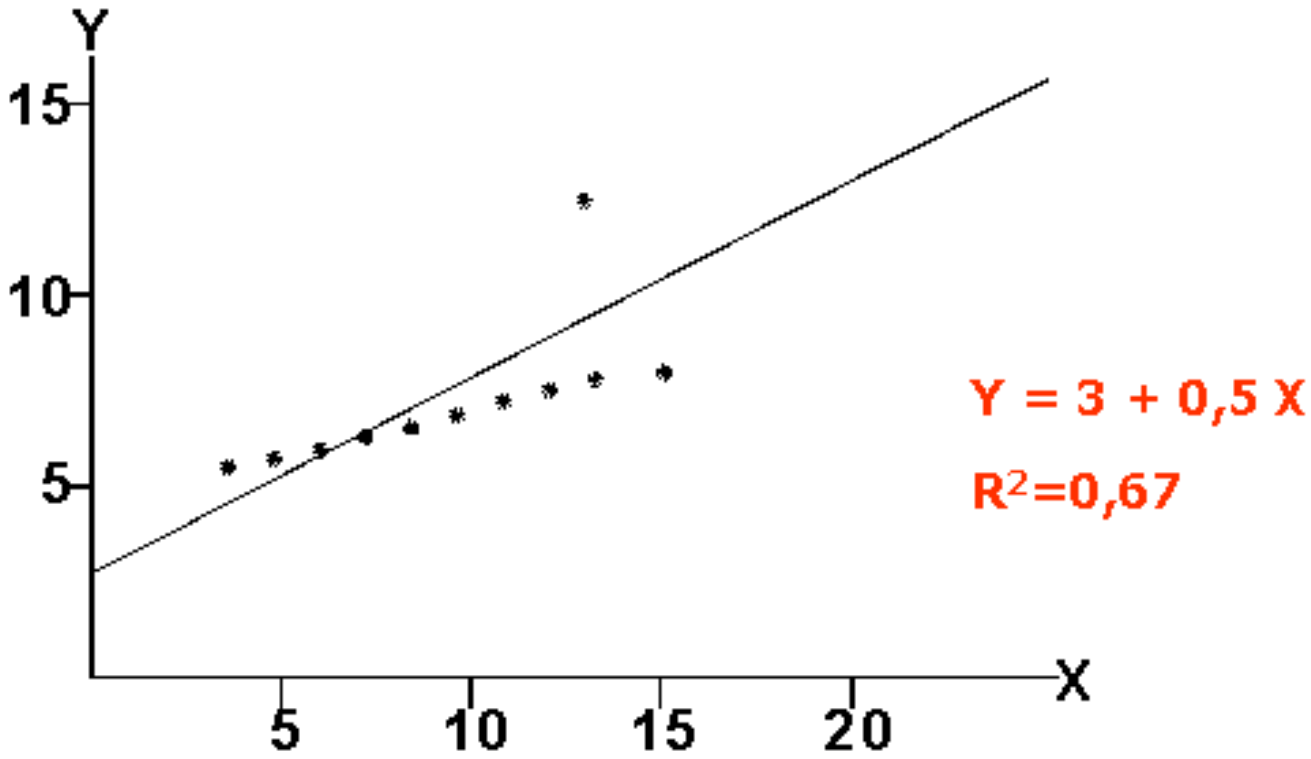

**Odhadnutý správny model**

Predstavme si, že pre dáta v grafe sme odhadli regresnú priamku, ako je uvedená vedľa grafu. **Koeficient determinácie bude 67%.** Ak zobrazíme dáta zistíme, že môžeme byť spokojný, pretože závislosť je skutočne lineárna.



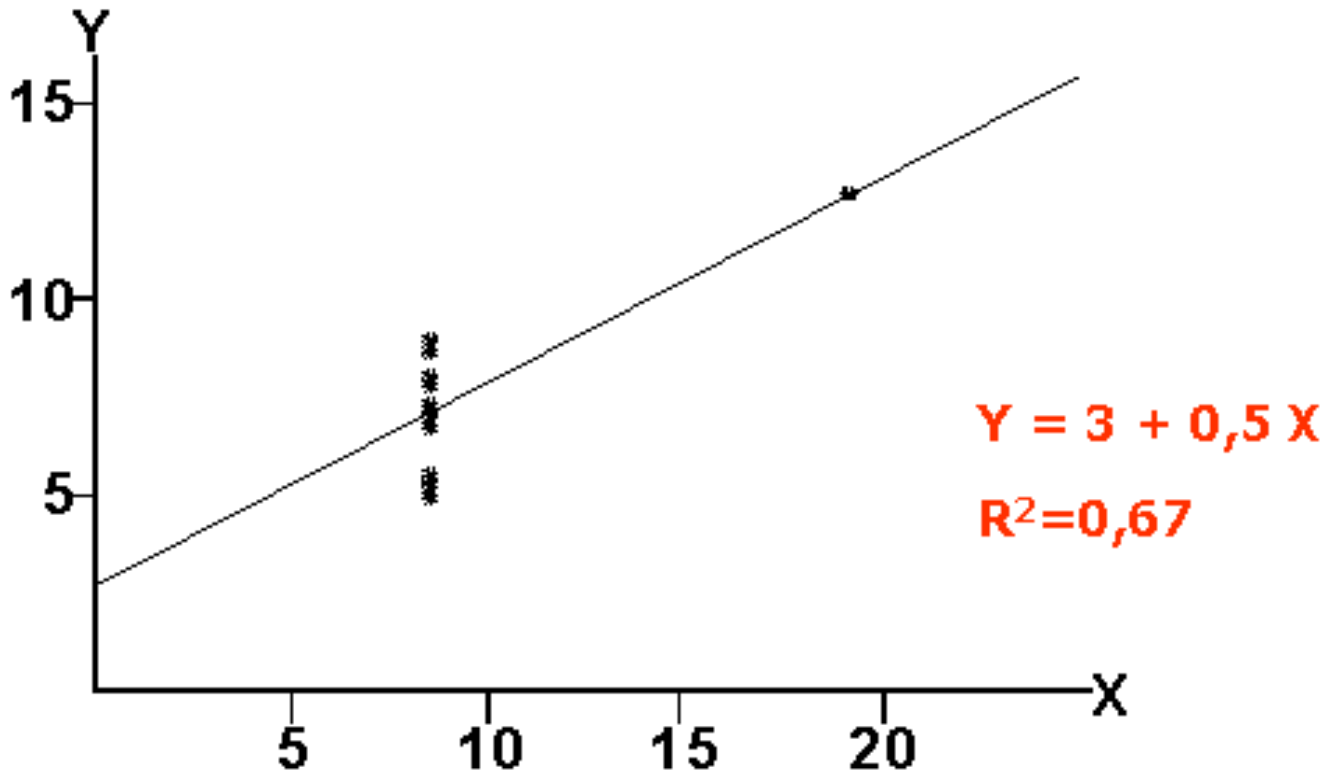

# Odhadnutý nesprávny model

Rovnaký model a koeficient determinácie môže dostať aj pre dáta zobrazené v grafe. V takomto prípade, ale s modelom **nemôžeme byť spokojný**, pretože závislosť nie je lineárna.




# Model s extrémnou hodnotou

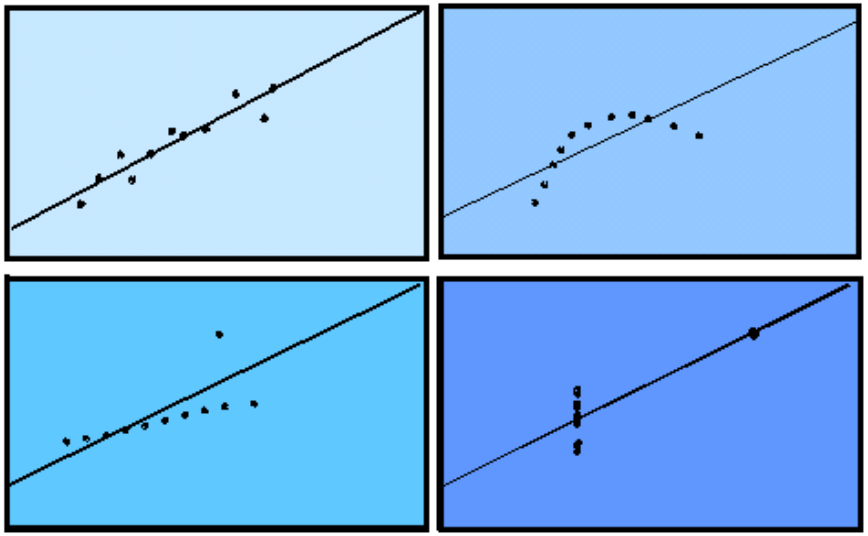
V hore uvedenom grafe máme ďalší prípad, keď vypočítame rovnakú priamku aj rovnaký koeficient determinácie. Problém je však v tom, že ich odhad je **ovplyvnený jednou extrémnou hodnotou** , ktorá mení sklon priamky aj hodnotu koeficienta determinácie.




# Model s rozhodujúcim pozorovaním

V poslednom grafe má **jedno pozorovanie rozhodujúci vplyv** na zistenie významnej lineárnej závislosti. Vedie k dramatickým zmenám v odhade lineárnej regresie, ktorá by bez neho bola nevýznamná.

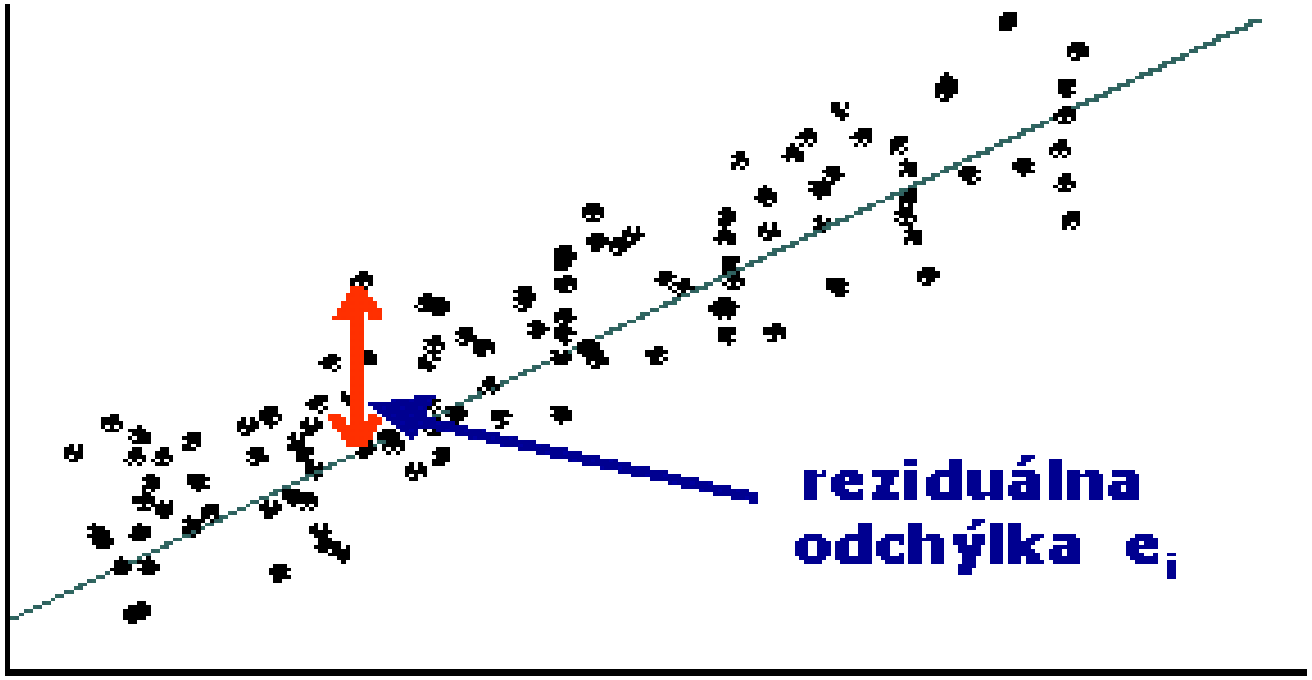
# Nedostatky mier kvality odhadu



**klasické miery kvality odhadu nestačia na odhalenie porušenia predpokladov regresnej analýzy**

Posudzovať kvalitu modelu v oblasti splnenia predpokladov len na základe mier ako je koeficient determinácie a jemu podobným môže byť zavádzajúce. Pre posúdenie splnenia predpokladov je potrebné použiť **špeciálne nástroje**.

## Reziduálne odchýlky




sú základom pre hodnotenie
preto hovoríme
o tzv. analýze reziduálov

Po odhade modelu je dôležité overiť platnosť predpokladov MNŠ pomocou analýzy reziduálov v regresnom modeli

- reziduál je rozdiel medzi skutočnou a odhadnutou hodnotu (získanou dosadením X do rovnice regresnej priamky

- reziduálne odchýlky musia byť náhodné

- **použijeme grafické zobrazenie**

- **bodový graf reziduálov vs. predikovaných hodnôt**

- **bodový graf reziduálov vs. nezávislá premenná**

- reziduálne odchýlky musia byť náhodne rozptýlené okolo nuly

- v grafe nesmie byť žiadny náznak potenciálneho trendu alebo vzoru vývoja

- reziduálne odchýlky možno ohraničiť dvomi priamkami rovnobežnými s x-ovou osou

# Typy reziduálov

- **jednoduché reziduály** $e_i = Y_i - \hat{Y}_i$
  - sú odhadom neznámej náhodnej zložky $e_i = est(\varepsilon_i)$
- **studentizované reziduály**
  - sú vypočítané z jednoduchých reziduálov vydelením štandardnou odchýlkou reziduálov
- **parciálne studentizované reziduály**
  - sú vypočítané rovnako ako studentizované reziduály ale po vylúčení i-teho pozorovania z odhadu regresie

# Overenie kvality modelu

- Testovanie významnosti modelu ako celku
  - **rozklad variability**
    - celková variabilita
      - na koľko sa odchyľujú konkrétne hodnoty premennej Y od celkového priemeru
    - vysvetlená variabilita
      - na koľko sa odchyľujú hodnoty na regresnej priamky od celkového priemeru
    - nevysvetlená variabilita
      - na koľko sa odchyľujú skutočné hodnoty premennej Y od hodnôt odhadnutých regresnou priamkou
  - čím väčšia je vysvetlená variabilita v porovnaní s nevysvetlenou variabilitou, tým lepšie odhadnutá priamka modeluje závislosť premenných

| Variabilita | Súčet štvorcov odchýlok | Stupne voľnosti | Rozptyl | F test |
|---|---|---|---|---|
| **Vysvetlená** | $\mathbf{V} = \sum_{j=1}^{n} \left( y'_j - \bar{y} \right)^2$ | $p - 1$ | $s^2_{y'} = \frac{V}{p-1}$ | $F = \frac{s^2_{y'}}{s^2_r}$ |
| **Nevysvetlená** | $\mathbf{N} = \sum_{j=n}^{n} \left( y_j - y'_j \right)^2$ | $n - p$ | $s^2_r = \frac{N}{n-p}$ | |
| **Celková** | $\mathbf{C} = \sum_{j=1}^{n} \left( y_j - \bar{y} \right)^2$ | $n - 1$ | | |

- Posúdenie kvality vyrovnania
  - **$R^2$ - koeficient determinácie**
    - rovnako ako pri jednoduchom modeli
    - je logické, že čím viac nezávislých premenných použijeme, tým presnejšie vyjadríme variabilitu závislej premennej
      - preto, ak chceme dosiahnuť kompromis medzi kvalitou vyrovnania a jednoduchosťou modelu, musíme počet regresorov korigovať
  - **$R^2_{adj.}$ - upravený koeficient determinácie**

$$R^2_{adj} = 1 - \left( \frac{n-1}{n-p-1} . \left(1 - R^2\right) \right)$$

    - obsahuje korekciu pre počet vysvetľujúcich premenných
      - s pridávaním premenných do modelu sa výraznejšie nemení
      - maximálne nadobúda hodnotu 1
      - môže mať aj zápornú hodnotu
      - výrazný rozdiel medzi $R^2$ a $R^2_{adj.}$ Indikuje, že do modelu bolo zahrnutých príliš veľa premenných