# Doplňkový materiál k přednášce z Biostatistiky 21.11.2007

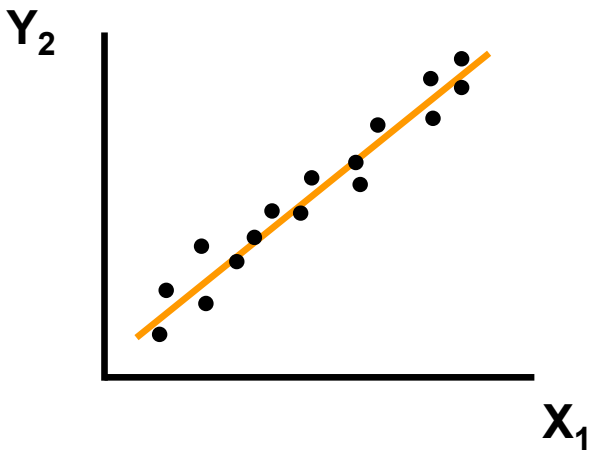
Regrese a korelace

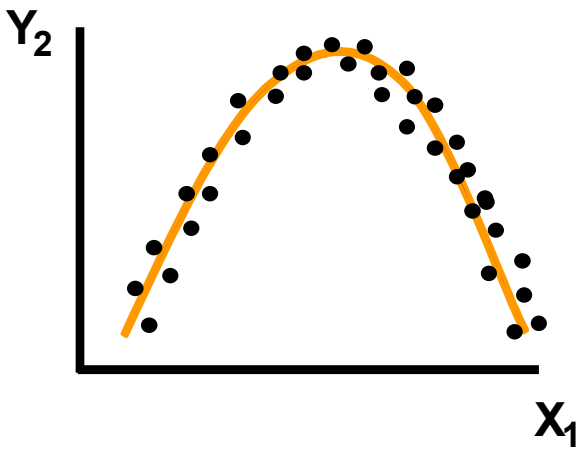
# Korelační koeficient

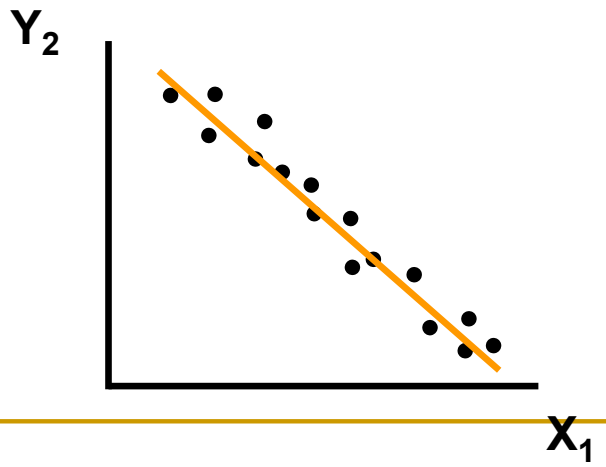
Jedna a více nezávisle proměnných

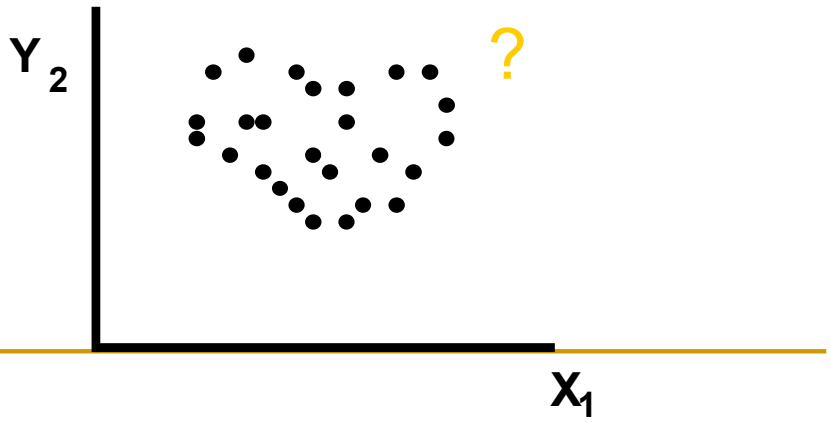
# Základy korelační analýzy - I.

## Korelace – *vzájemný* vztah dvou znaků (parametrů)



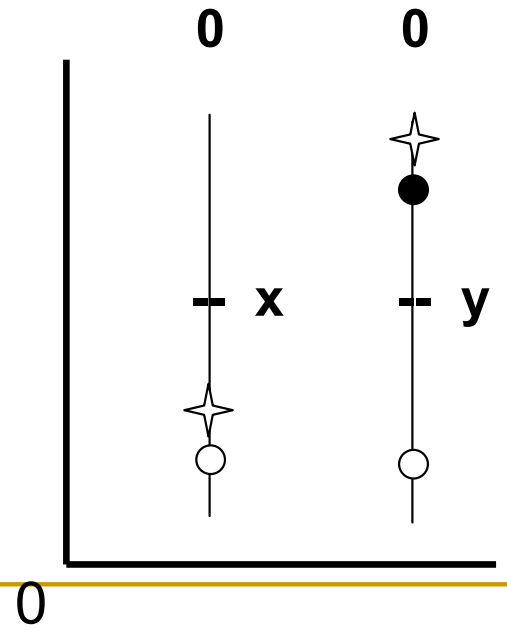



# Základy korelační analýzy - II.

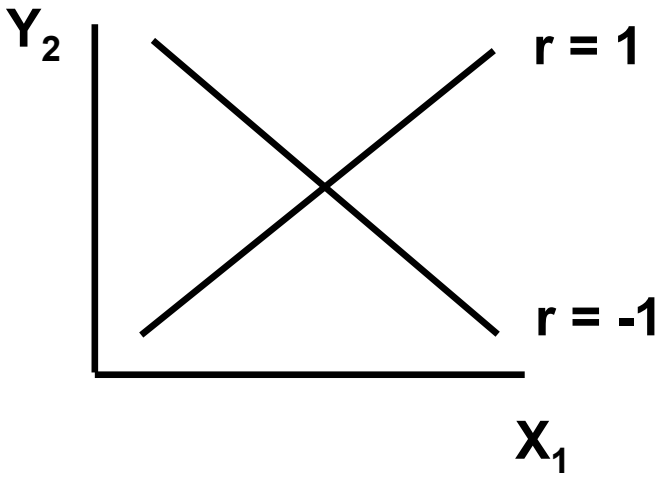
## Parametrické míry korelace

Kovariance

$$Cov(x, y) = E(x_i - \bar{x}).(y_i - \bar{y})$$




Pearsonův koeficient korelace= normovaná kovariance



# Pearsonův korelační koeficient

$$r = \frac{Cov(x,y)}{S_x.S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}}$$

- **Pearsonův korelační koeficient**
  **postižení lineárního vztahu mezi veličinami**

- **R=1 … přímá úměra, kladná korelace**
- **R=-1… záporná korelace**
- **R=0… mezi veličinami není žádná spojitost, žádná korelace, není lineární vztah mezi proměnnými**

- **Předpoklady: dvourozměrné normální rozdělení**

- **http://www.causeweb.org/repository/statjava/**
  **(statistical application -> correlation)**

# Jednovýběrový test I.

| $P_I$ (zem) | 10 | 14 | 15 | 32 | 40 | 20 | 16 | 50 |
|---|---|---|---|---|---|---|---|---|
| $P_I$ (rostl.) | 19 | 22 | 26 | 41 | 35 | 32 | 25 | 40 |

$$I = 1,.....,n; n = 8; v = 6$$

$$r = \frac{Cov\,(x,y)}{S_x.S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}} = 0,7176$$

$$H_0 : \rho = \phi : \alpha = 0,05$$

$$\textbf{tab}: r(v = 6) = 0,7076$$

# Jednovýběrový test II.

| $P_I$ (zem) | 10 | 14 | 15 | 32 | 40 | 20 | 16 | 50 |
|---|---|---|---|---|---|---|---|---|
| $P_I$ (rostl.) | 19 | 22 | 26 | 41 | 35 | 32 | 25 | 40 |

$$I = 1,.....,n; n = 8; v = 6$$

$$r = \frac{Cov\,(x,y)}{S_x.S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}} = 0,7176$$

$$H_0 : \rho = \phi \qquad t = \left[\frac{r}{\sqrt{1 - r^2}}\right] \cdot \sqrt{n - 2} \qquad v = n - 2$$

$$\left.\begin{array}{l} t = \dfrac{0,7176}{0,6965} \cdot \sqrt{6} = 2,524 \\ \textbf{tab}: \; t_{0,975}^{(n-2)} = 2,447 \end{array}\right\} \quad P \leq 0,05$$

# Dvouvýběrový test

**1.** $n_1 = 1258$

$r_1 = 0{,}682$

**2.** $n_2 = 462$

$r_2 = 0{,}402$

Krevní tlak **x** koncentrace kysl. radikálů

$$Z_i = 1.1513 \cdot \log \frac{(1 + r_i)}{(1 - r_i)}$$

$Z_1 = 0{,}833$ $\quad$ $Z_2 = 0{,}426$

**Test:** $H_0 : \rho_1 = \rho_2$ ; $\alpha = 0{,}05$

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0{,}407}{0{,}0545} = 7{,}461$$

**tabulky** : $Z_{0{,}975} = 1{,}96$

**7,461 >> 1,96 => P << 0,01**

# Spearmanův pořadový koeficient korelace

- **Není nutný předpoklad normality veličin**
- **Interpretace stejná jako u parametrického r**
- **Výpočet založený na práci s pořadími hodnot**

| Data X | 0 | 6,9 | 3,3 | 100 | 5,8 |
|---|---|---|---|---|---|
| Pořadí X | 1 | 4 | 2 | 5 | 3 |
| Data Y | 10,1 | 9,8 | 4,2 | 3,2 | -1 |
| Pořadí Y | 5 | 4 | 3 | 2 | 1 |
| Di=Pořadí X-Pořadí Y | -4 | 0 | -1 | 3 | 2 |

$$r_s = 1 - \frac{6 \cdot \sum di^2}{n\left(n^2 - 1\right)}$$

Kde $d_i^2$ = rozdíl pořadí mezi $x_i$ a $y_i$

# Spearmanův korelační koeficient

| $P_I$ v půdě | 1 | 2 | 3 | 6 | 7 | 5 | 4 | 8 |
|---|---|---|---|---|---|---|---|---|
| $P_I$ v rostl. | 1 | 2 | 4 | 8 | 6 | 5 | 3 | 7 |
| $d_I$ | 0 | 0 | 1 | 2 | -1 | 0 | -1 | -1 |

**i = 1, ….. n;  n = 8  => v = 6**

$$r_s = 1 - \frac{6 \cdot \sum di^2}{n\left(n^2 - 1\right)} = 0,9048$$

**tab :** $r_s(v = 6) = 0,89$

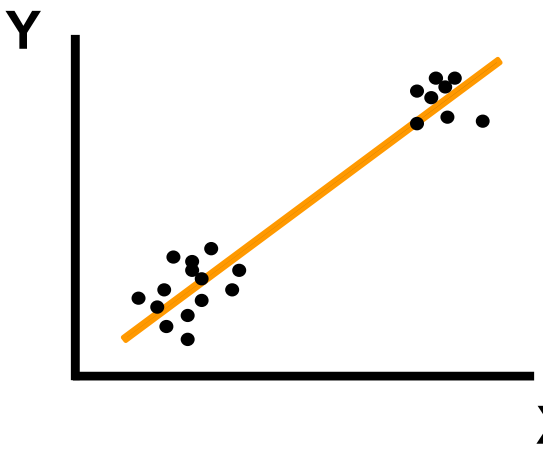
| Pacient č. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Lékař 1 | 4 | 1 | 6 | 5 | 3 | 2 | 7 |
| Lékař 2 | 4 | 2 | 5 | 6 | 1 | 3 | 7 |
| $d_I$ | 0 | -1 | 1 | -1 | 2 | -1 | 0 |

$$r_s = 1 - \frac{6 \cdot 8}{7\left(49 - 1\right)} = 0,857$$

**P = 0,358**

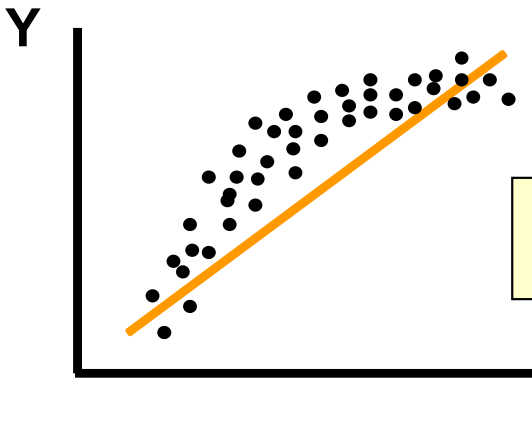
# Pasti a pastičky (Pearsonův k.k.)

## Problém rozložení hodnot



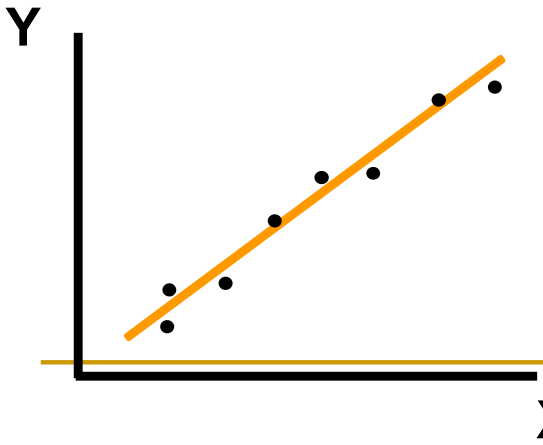

r = 0,981
(p < 0,001)

## Problém typu modelu

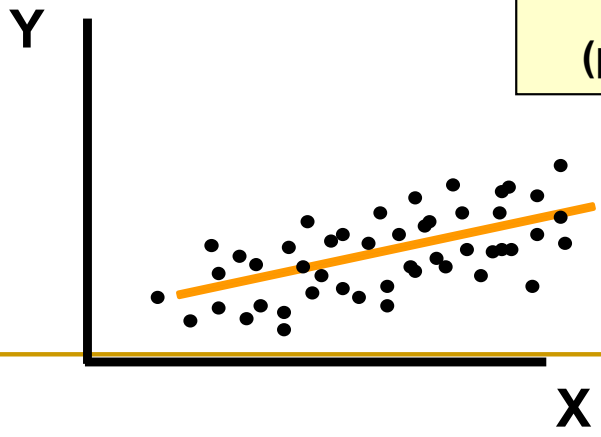



r = 0,761
(p < 0,032)

## Problém velikosti vzorku




r = 0,891
(p < 0,214)




r = 0,212
(p < 0,008)

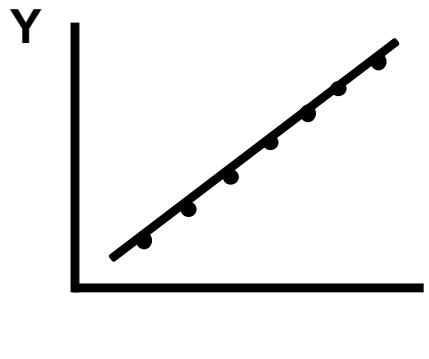
# Regrese

**Regrese - funkční vztah dvou nebo více proměnných**
**závislost jedné veličiny na druhé**

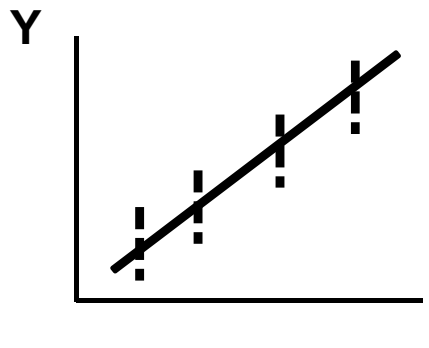
| Jednorozměrná | Vícerozměrná |
|---|---|
| y = f(x) | y = f(x1, x2, x3, ……xp) |

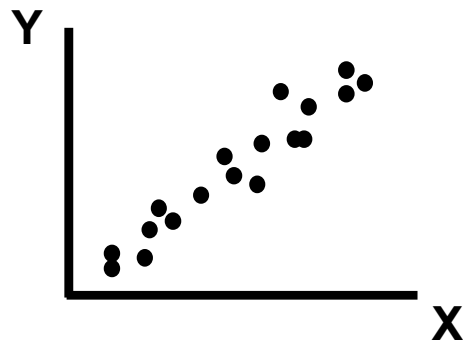
**Vztah x, y**

- **Deterministický**




- **Regresní, stochastický**




**Pro každé x existuje pravděpodobnostní rozložení y**

# Lineární regresní model

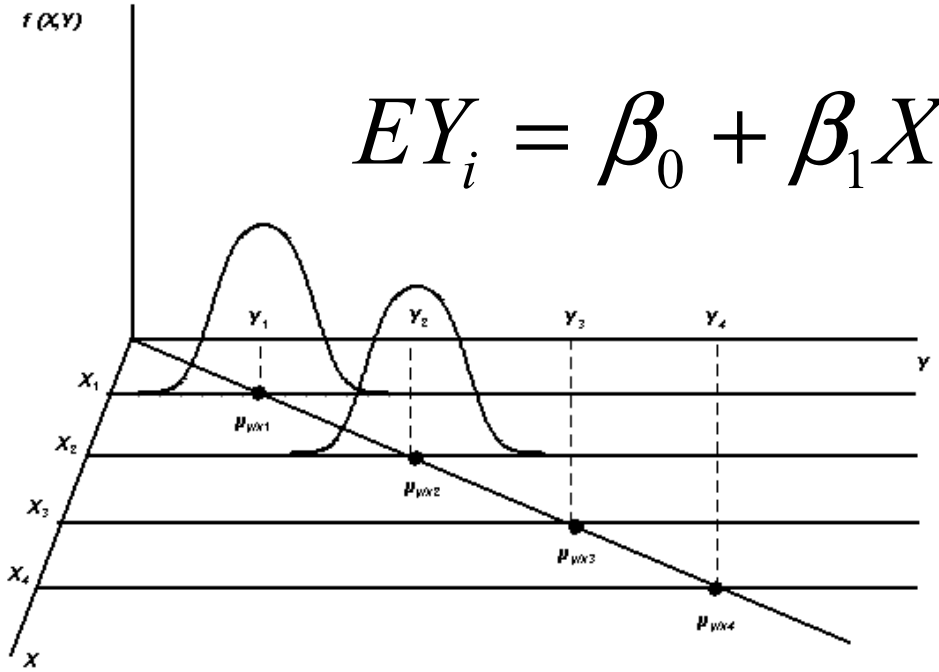
Jedna a více nezávisle proměnných

- *n* objektů
- Pro každý objekt: pozorované veličiny X a Y - spojité
- Pozorování, objekty – navzájem nezávislá
- Zajímá nás závislost veličiny Y na X – POZOR! – nutná podmínka je, že závislost je stejná pro všechny zkoumané objekty.

- Příklad: V egyptské vesnici Kalama se studoval vliv výživy na zdravotní stav dětí. Při této příležitosti se měřily průměrné výšky dětí (v cm) ve věku od 18 měsíců do 29 měsíců.
- ? Jaká je závislost výšky dítěte na jeho věku?

- X,Y – náhodné veličiny (střední hodnota, rozptyl)
- Existuje souvislost mezi středními hodnotami N.V.?

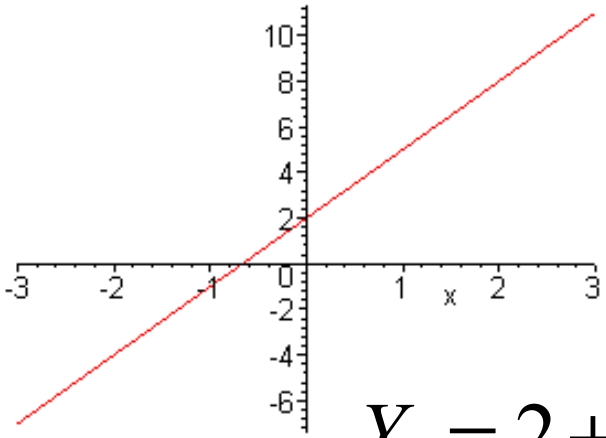
$$EY_i = \beta_0 + \beta_1 X_i, i = 1,...,12.$$



# Opakování z gymnázia – analytická geometrie

- Analytické vyjádření přímky, rovnice

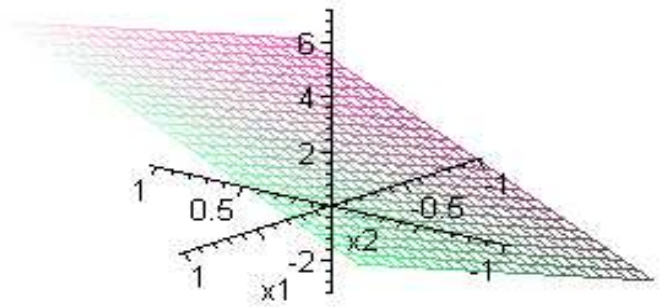
$$Y = \beta_0 + \beta_1 X$$




$$Y = 2 + 3X$$

- Analytické vyjádření roviny, rovnice

$$Y = \beta_0 + \beta_1 X1 + \beta 2 X2$$




$$Y = 2 + 3X1 + 2X2$$

# Nejjednodušší typ závislosti - **lineární**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,\ldots,n$$

- Systematická část modelu
- Náhodná část, složka modelu (náhodné chyby, *random error*)

# Regresní rovnice - proměnné

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,\ldots,n$$

- **Závisle proměnná**
- ***Dependent variable***
- Jedná se o veličinu, kterou zkoumáme a chtěli bychom najít její popis pomocí dalších měřených veličin.

- **Nezávisle proměnná**
- ***Independent variable***
- **Kovariáta *(covariate)***
- **Prediktor**
- **Regresor**
- Jedná se o veličiny, které nám slouží pro popis závisle proměnné**.**

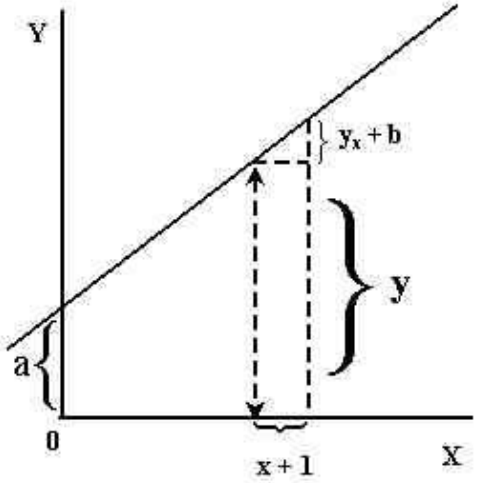
Příklad - Kalama: Věk = nezávisle proměnná(X), horizontální osa
Výška = závisle proměnná(Y), vertikální osa

# Regresní rovnice, přímka? - parametry

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,\ldots,n$$

- Průsečík s osou Y
- *Intercept*

- Směrnice

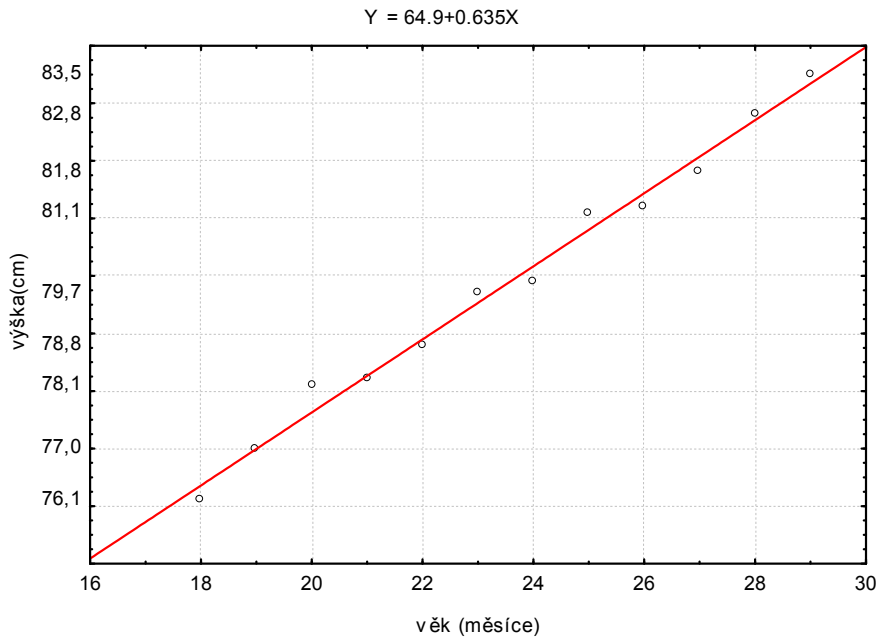



Interpretace parametrů:

**Směrnice:** o kolik se změní hodnota závisle proměnné, jestliže hodnota nezávisle proměnné vzroste o 1 jednotku.

**Průsečík:** udává hodnotu závisle proměnné, jestliže hodnota nezávisle proměnné je rovna 0.

# Příklad: Kalama




- Lineární závislost – přímka
- Height=64.9+0.635Age
- **Průsečík: 64,9**
- Interpretace ad absurdum: výška dítěte ve věku 0 měsíců (tj. při porodu). *Ale to by byla **extrapolace**, tedy rozšíření modelu na oblast, kde jsme data neměřili.*
- **Směrnice: 0,635**
- Dítě starší o jeden měsíc je v průměru větší o 0,635 cm.

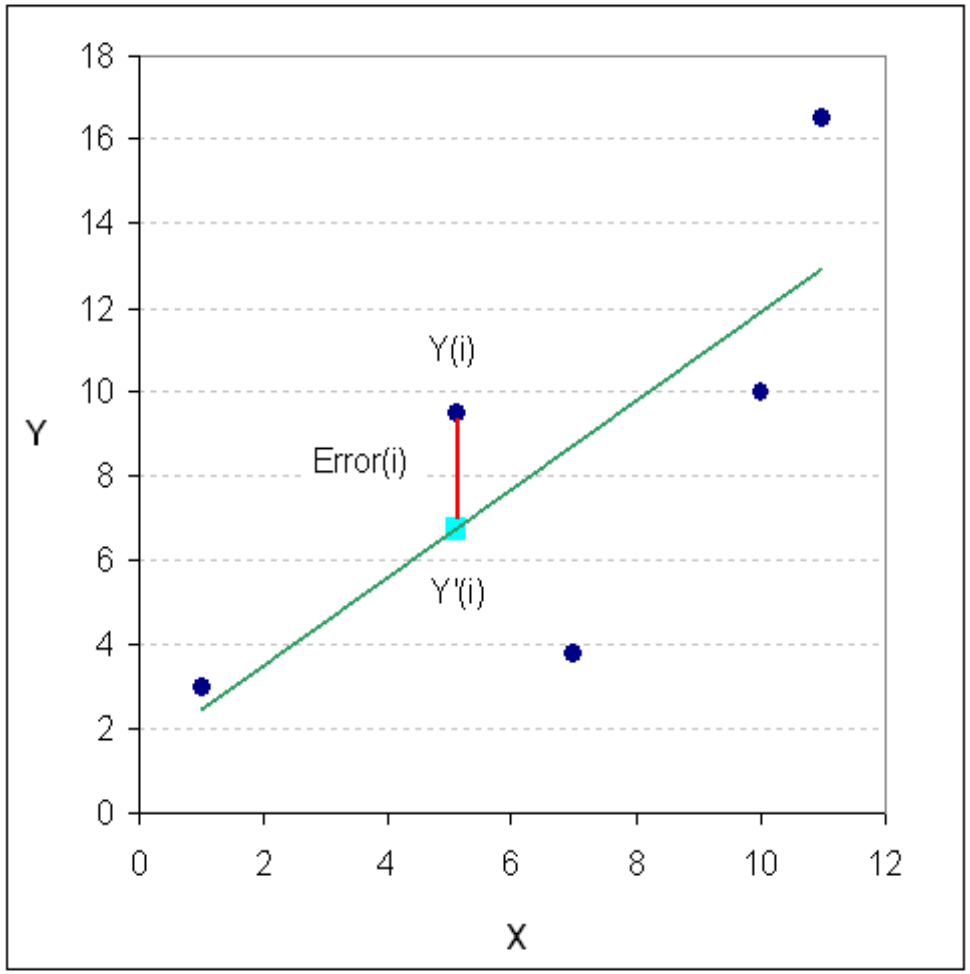
# Tvorba lineárního regresního modelu

- Je-li závisle proměnná spojitá a nezávisle proměnné jsou spojité nebo diskrétní (podmínkou je, že alespoň jedna nezávisle proměnná je spojitá) a jsou-li splněny jisté předpoklady, o kterých budeme hovořit později, můžeme přistoupit k budování lineárního regresního modelu.

- Při tvorbě modelu (obecně, nejen lineárního) postupujeme následujícím způsobem:

1. Odhadneme parametry modelu
2. Hledáme významné (signifikantní) prediktory
3. Na závěr hodnotíme vhodnost námi vytvořeného modelu, jak dobře popisuje funkcionální závislost mezi závisle proměnnou a nezávisle proměnnými.

# Residua




- Svislé odchylky naměřených hodnot od regresní přímky nazýváme **residua.**
- *i*-té residuum vypočteme jako rozdíl skutečně naměřené hodnoty Y a hodnoty $\hat{Y}$ predikované regresním modelem

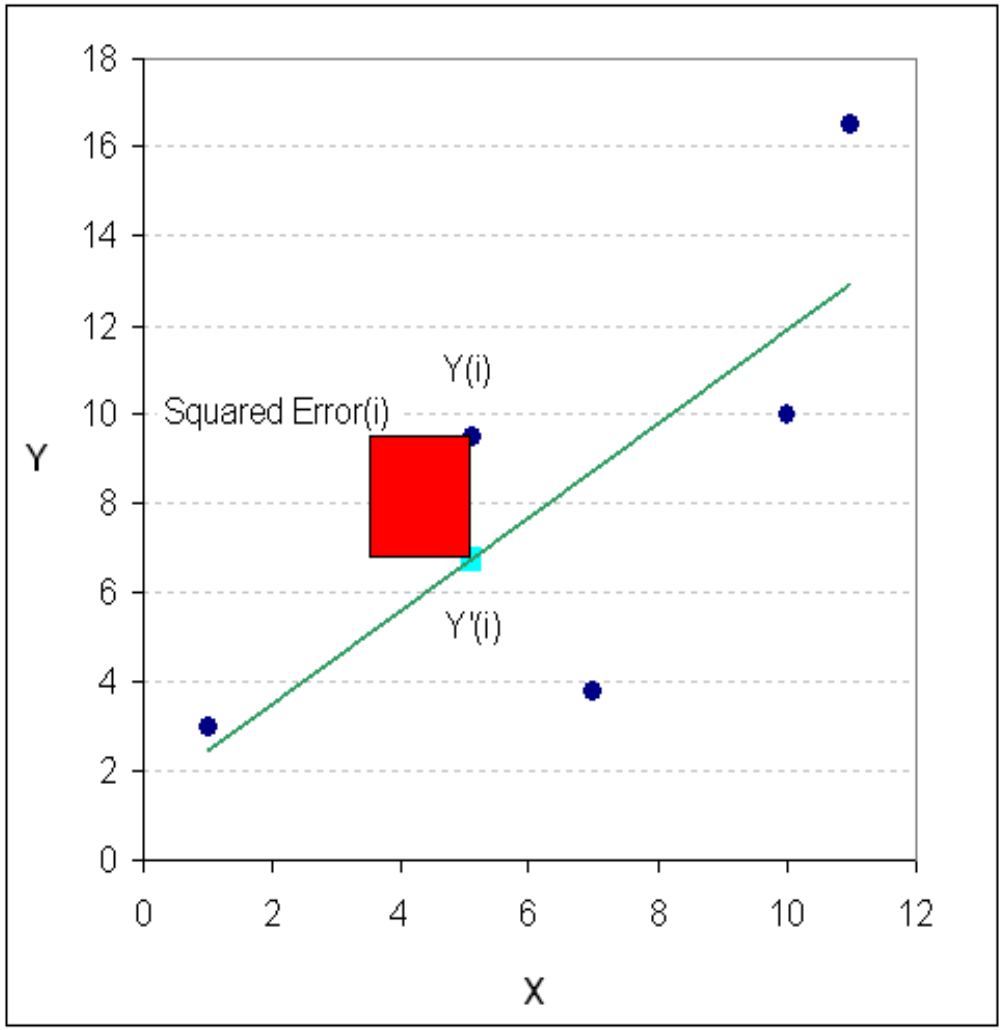
$$\text{Re}siduum_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \beta_0 - \beta_1 X_i$$

# Pozn.: Residuální součet čtverců

- Výsledný minimální součet čtverců residuí (pro b0 a b1) nazýváme **residuální součet čtverců** (*residual sum of squares*), $S_e$

$$S_e = \sum_{i'=1}^{n} (Y - b_0 - b_1 X_i)^2$$

# Metoda nejmenších čtverců




Interaktivní hrátky:

- http://hadm.sph.sc.edu/COURSES/J716/demos/LeastSquares/LeastSquaresDemo.html
- http://ite.pubs.informs.org/Vol1No1/ErkutIngolfsson/ErkutIngolfsson.php
- http://www.causeweb.org/repository/statjava/ (statistical application -> regression)

# Metoda nejmenších čtverců *(least squares method)* - odhad parametrů modelu

- Metoda nejmenších čtverců spočívá v minimalizaci přes $\beta_0$ a $\beta_1$ součtu čtverců reziduí.

$$\sum_{i'=1}^{n}(Y - \beta_0 - \beta_1 X_i)^2$$

- Výsledné hodnoty $\beta_0$ a $\beta_1$, pro které je součet čtverců minimální označujeme $b_0$ a $b_1$
- Odhadnutá regresní rovnice má tvar

$$Y = b_0 + b_1 X$$

# Vzorce pro odhad parametrů regresní přímky – metoda nejmenších čtverců

Odhad b je zatížený chybou:

$$S_b^2 \sim \sigma_\beta^2: \quad \frac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$$

$$S_{y \cdot x}^2 = \frac{\sum d_{y \cdot x}^2}{n-2} = \frac{\sum Y_i^2 - \frac{\sum Y_i^2}{n} - b^2 \cdot \sum (X_i - \bar{X})^2}{n-2}$$

I. $$b \sim \beta: \quad b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

---

II. $$a \sim \alpha: \quad a = \bar{Y} - b \cdot \bar{X} \qquad S_a^2 \sim \sigma_\alpha^2 \qquad S_\alpha^2 = \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum X^2} \right] \cdot S_{y \cdot x}^2$$

**intercept**

---

III. $\hat{Y}$ : **modelová hodnota**

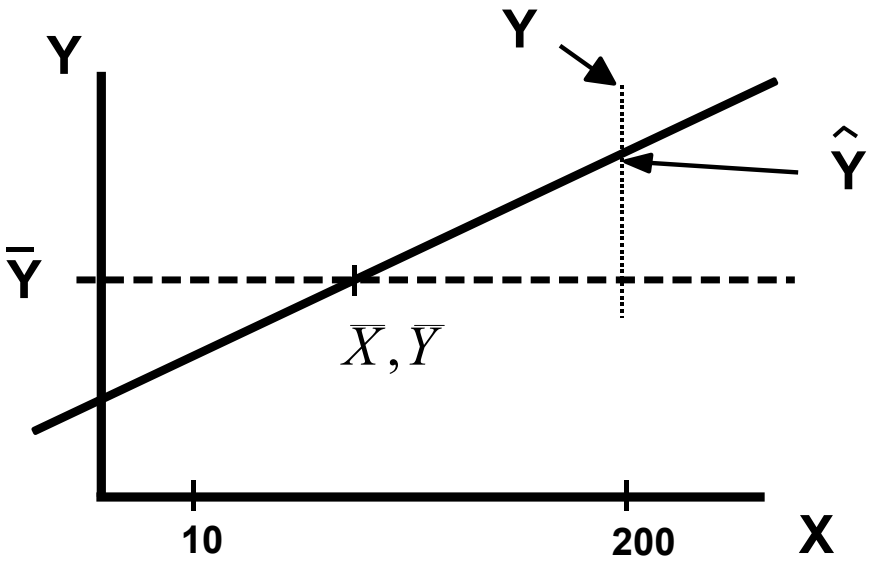
$$\hat{Y}_i = a - b \cdot X_i \qquad S_{\hat{y}_i} = (S_{y \cdot x}) \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum X^2}}$$

# Příklad: Spalování odpadu

**X: Množství spáleného odpadu (tuny)**
**Y: Koncentrace kovu ve vzduchu(ng/m3)**

**Platí: X = 0; 10; 100; 150; 200; 250; 300 tun**

**Model: Y = a + b . X**




**Výsledek:** $\hat{Y} = 14 + 0{,}123 \cdot X; \quad \hat{Y} \rightarrow \left[\frac{\text{ng kov}}{m^3}\right]$

**Např. :** Skutečná data pro X = 200 t:

$Y_i$ = 16; 25; 41; 28; 31; 20 => $Y_i$ = 26.8

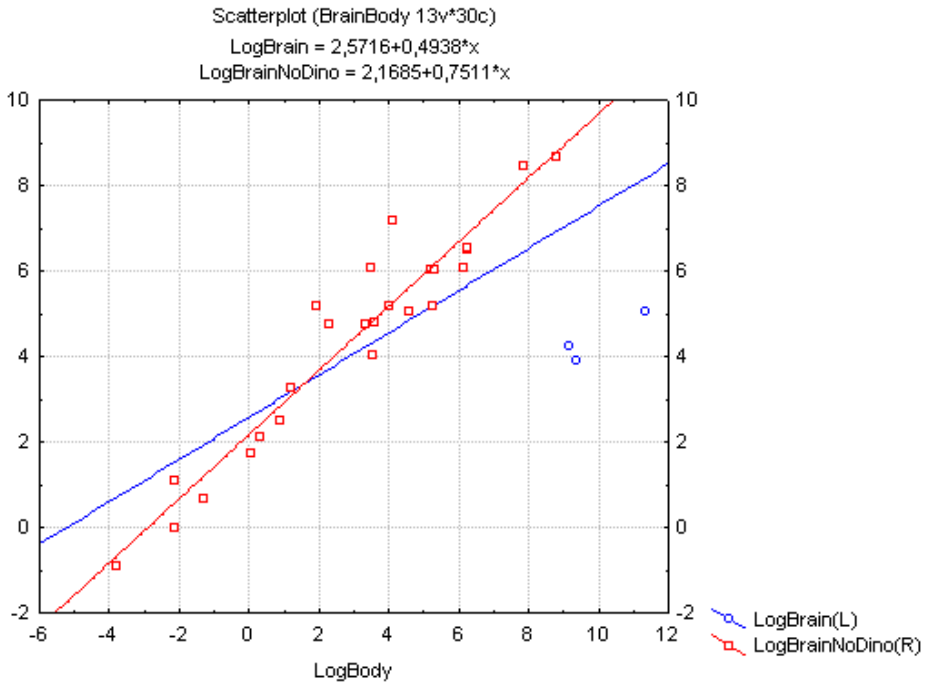
$$\left.\begin{aligned} \hat{Y} &= \bar{Y} + b \cdot (X - \bar{X}) \\ \hat{Y} &= a + b \cdot X \end{aligned}\right\} a = \bar{Y} - b \cdot \bar{X}$$

**Odhadnuto z modelu pro X = 200 t:**

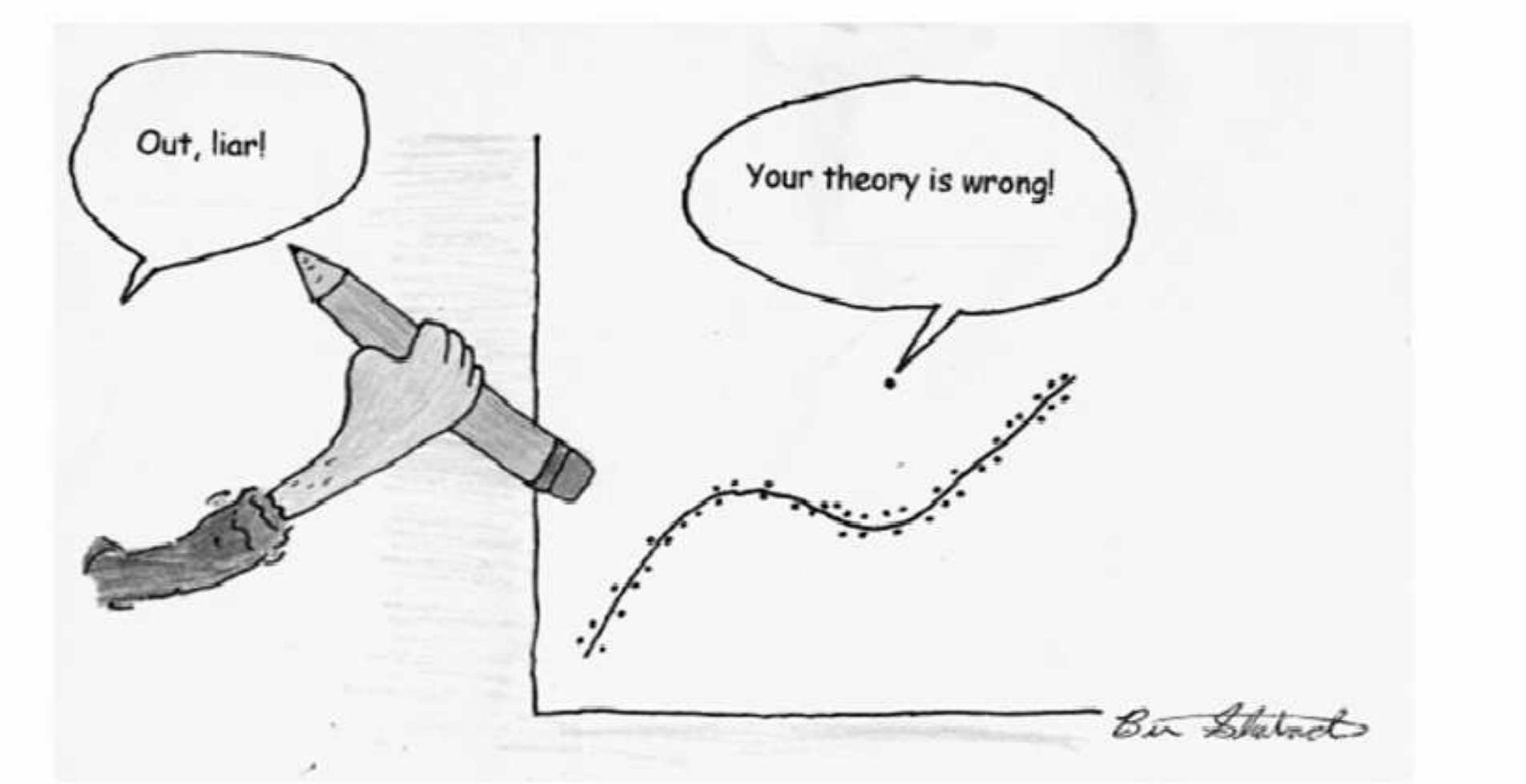
$\hat{Y}$ **= 14 + 0,123 . 200 = 38,6**

# Odlehlá pozorování - Nebezpečí *(outliers)*

- Závislost velikosti mozku(g) na váze těla (kg) (pro různé živočichy),log.transform ace
- Modrá - model pro všechna zvířata.
- Červená - model bez dinosauru.
- Dinosauři zkreslili výsledný model.



# Outliers (http://botany.upol.cz/prezentace/duch (soubor statistika4.pdf))



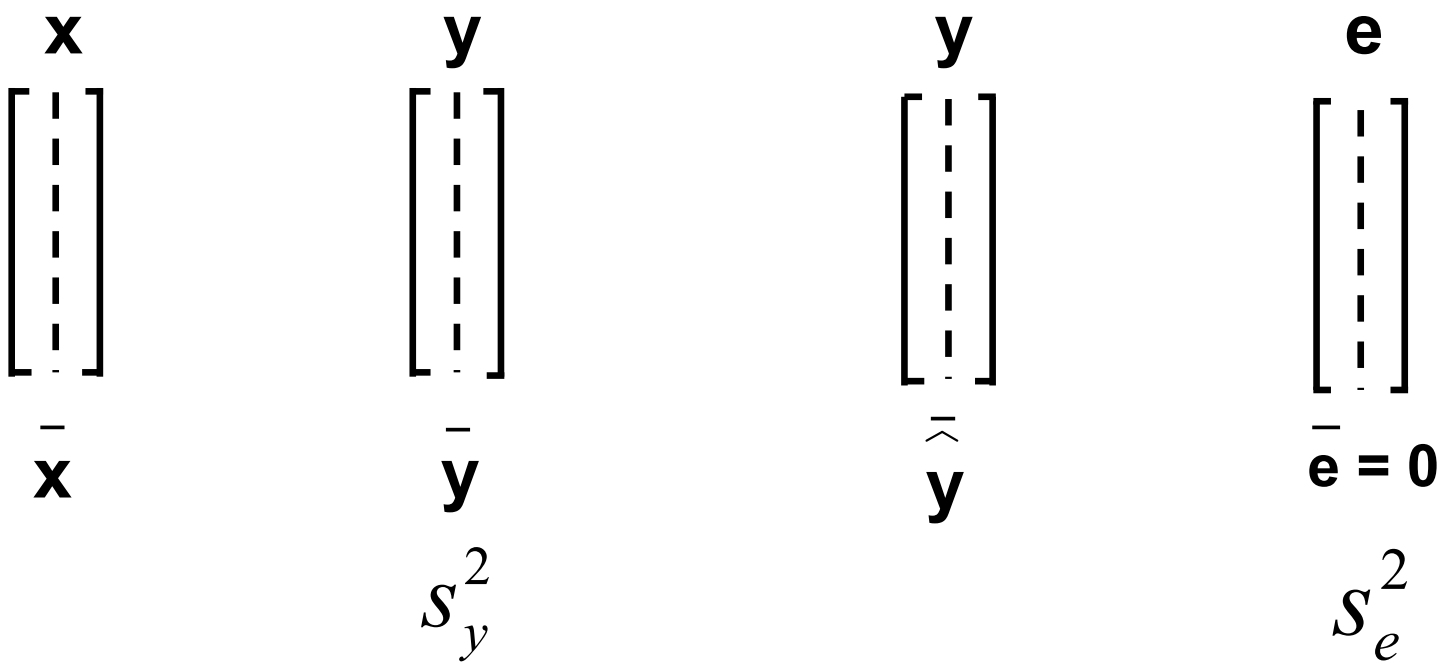
# Hledáme významné (signifikantní) prediktory

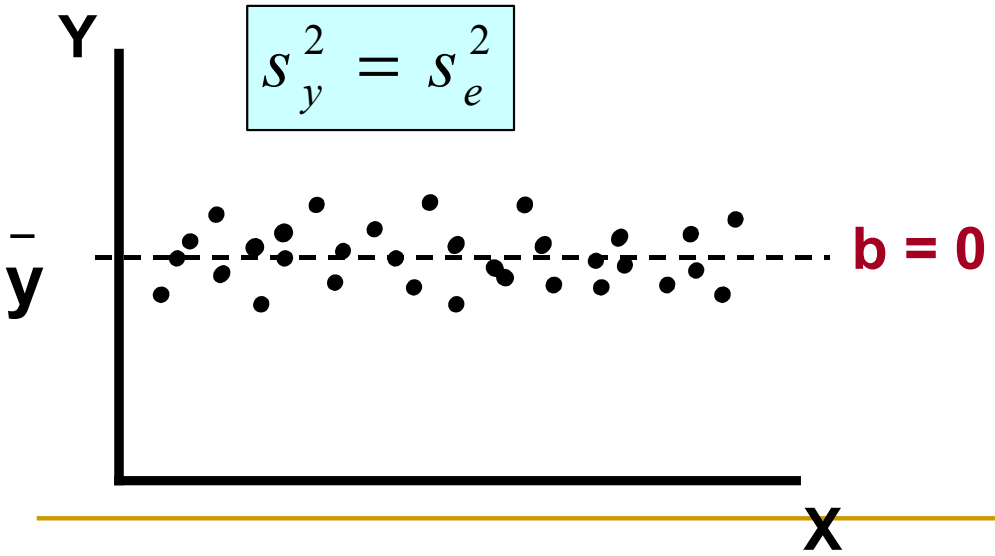
Při konstrukci regresního modelu bychom chtěli prokázat, že závislá veličina skutečně závisí na nezávisle proměnné.Tuto závislost na *X* prokazujeme testováním nulové hypotézy

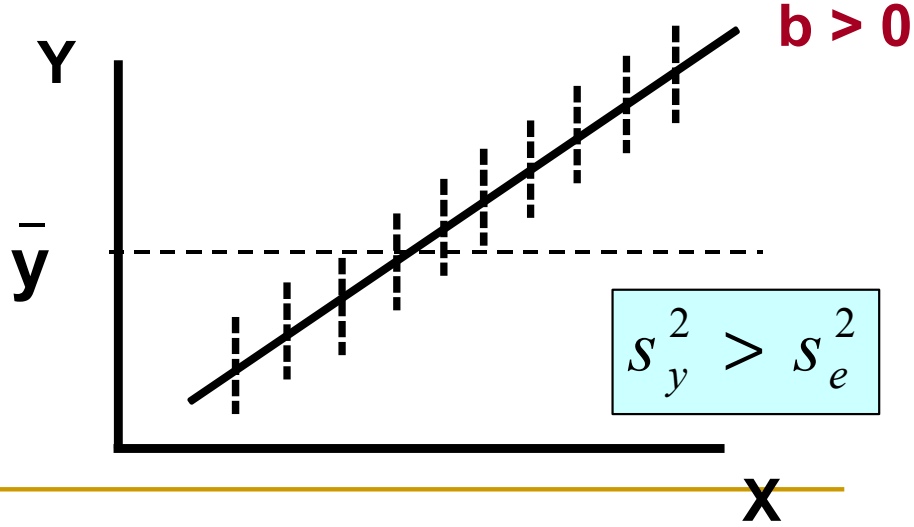
$$H_0 : \beta_1 = 0$$

proti alternativní hypotéze

$$H_A : \beta_1 \neq 0.$$


x
y
y
e
$\bar{x}$
$\bar{y}$
$\hat{y}$
$\bar{e} = 0$
$s_y^2$
$s_e^2$


Y
$s_y^2 = s_e^2$
$\bar{y}$
b = 0
X


b > 0
Y
$\bar{y}$
$s_y^2 > s_e^2$
X

# T-test

- Nezamítneme-li nulovou hypotézu, pak střední hodnota $Y_i$ nezávisí na *X*, tj. střední hodnota $Y_i$ je pro všechny hodnoty *X* stejná a má hodnotu $b_0$.
- Nulovou hypotézu $H_0$ testujeme pomocí testové statistiky $$T = \frac{b_1}{S.E.(b_1)}$$
- a zamítáme ji v případě, že $$|T| \geq t_{n-2}(1-\alpha/2),$$

kde $t_{n-2}(1-\alpha/2)$ je kvantil t-rozdělení s *n-2* stupni volnosti; *n* je počet pozorování, pro které konstruujeme regresní model.

# Příklad

**X: Koncentrace drogy: 0; 2; 6; 8; 10; 12; 15 mg/ml krve**
**Y: Koncentrace volných metabolitů**

**Pro každé X: 3 opakování Y, n=21**

---

**Model: Y = a + b . x** ➡ **Y = 0,11 + 0,092 . X**

$$t_{0,975}^{(v=19)} = 2,093$$

**I.** $H_0 : \beta = 0; \alpha = 0,05$
$b = 0,092 ; s_b = 0,023$ $\Big\} t = \dfrac{b}{S_b} = 4,00$

**P < 0,01**

$$\beta : b \pm t_{1-\alpha/2}^{(n-2)} \cdot S_b$$

$$P(0,044 \le \beta \le 0,140) = 0,95$$

---

**II.** $H_0 : \alpha = 0; \alpha = 0,05$
$a = 0,11; s_a = 0,029$ $\Big\} t = \dfrac{a}{S_a} = 3,793$

$$t_{0,975}^{(v=19)} = 2,093$$

**P < 0,05**

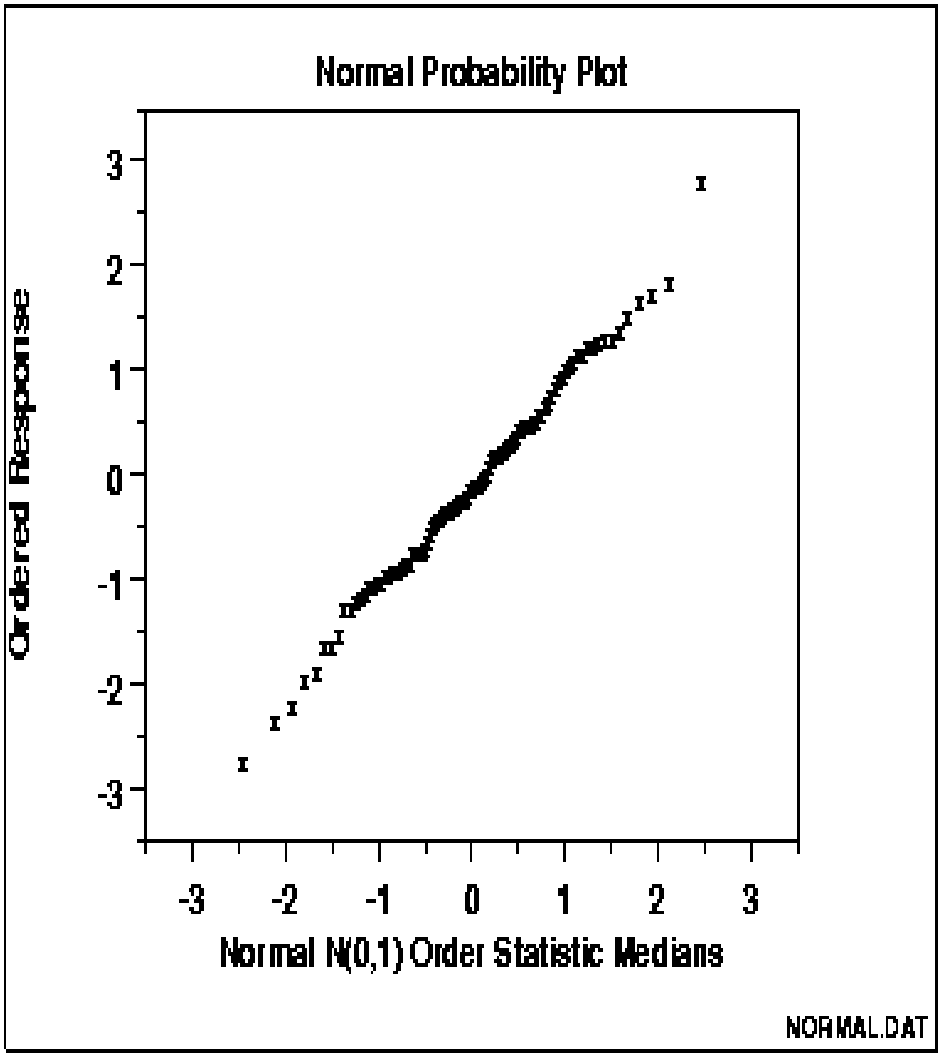
$$\alpha : \alpha \pm t_{1-\alpha/2}^{(n-2)} \cdot S_a$$

$$P(0,049 \le \alpha \le 0,171) = 0,95$$

# Předpoklady

- Nutný předpoklad potřebný ke všem testům spojeným s regresním modelem je **normalita residuí.**
- Residua mají mít normální rozdělení s **nulovou střední hodnotou a konstantním rozptylem** $\sigma^2$
- Dále předpokládáme, že všechna **pozorování** jsou **navzájem nezávislá.**

# Normalita residuí – graficky
# Q-Q plot *(Quantile-Quantile plot)*




- Grafická metoda pro srovnání rozdělení dvou výběrů.
- Vodorovná osa – empirické kvantily rozdělení 1. výběru. (jestliže vynášíme teoretické kvantily normovaného normálního rozdělení – **normal probability plot**)
- Svislá osa – empirické kvantily rozdělení 2. výběru (např. reziduí).
- Jsou-li obě rozdělení totožná, leží body (odpovídající si kvantily) na diagonální přímce

# Q-Q plot
## další vlastnosti

- http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

# Normalita residuí - testy

- Testy normality:

1. Kolmogorov-Smirnov
2. Shapiro-Wilks

Není-li splněn předpoklad normality – mohou pomoci **transformace** (později, dříve).

- Autokorelace residuí

1. Durbin-Watsonův test

# Diagnostika residuí

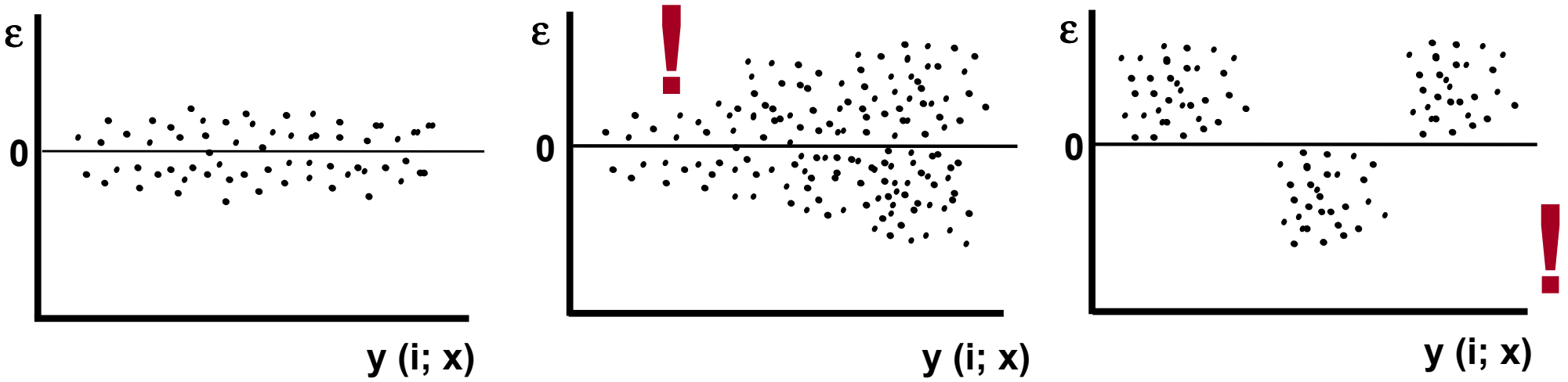
- Je námi zvolená závislost (lineární) vhodná?
- Pomoc grafické znázornění – **grafy závislosti hodnot residuí na hodnotách** $\hat{Y}_i$ **nebo** $X_i$

- V případě, že zvolený tvar závislosti byl vhodný, jsou residua

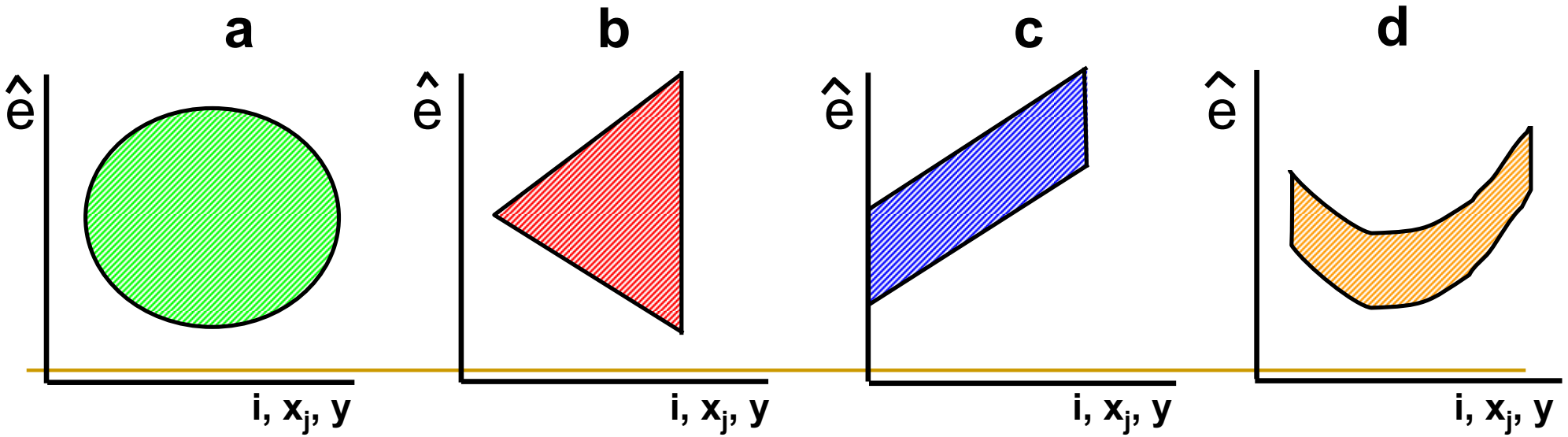
1. umístěna náhodně kolem nulové střední hodnoty
2. nevykazují žádný systematický trend
3. jejich rozptyl je homogenní

# Diagnostika residuí

## 3) Grafy residuí modelů (příklady)

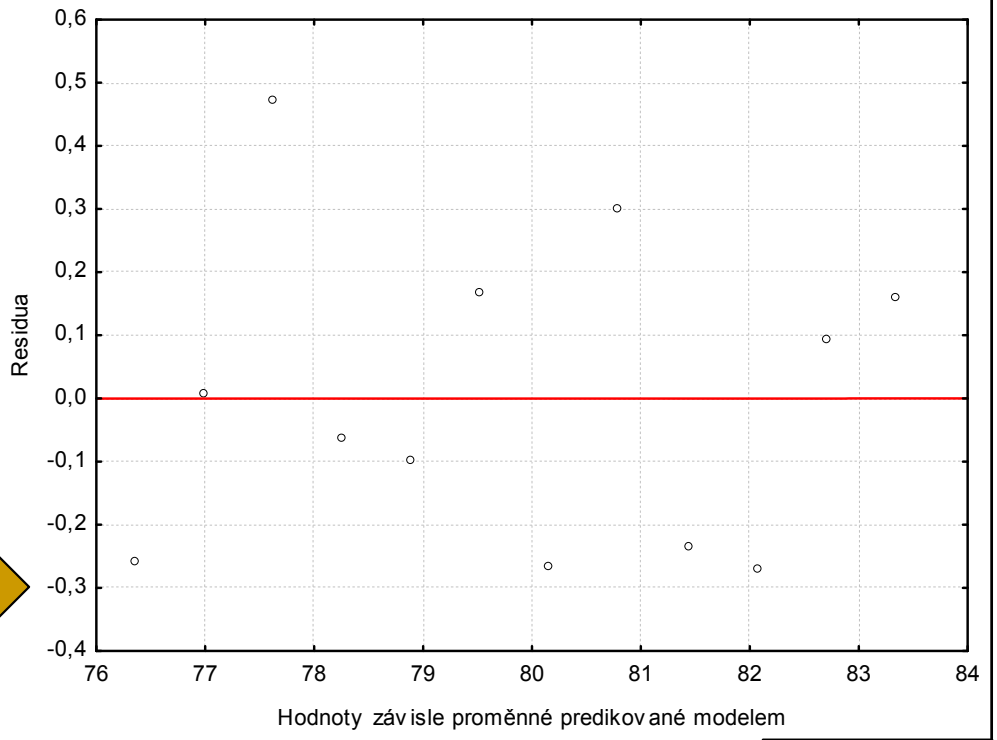



## Obecné tvary residuí modelů (schéma)



# Diagnostika residuí - obrázky

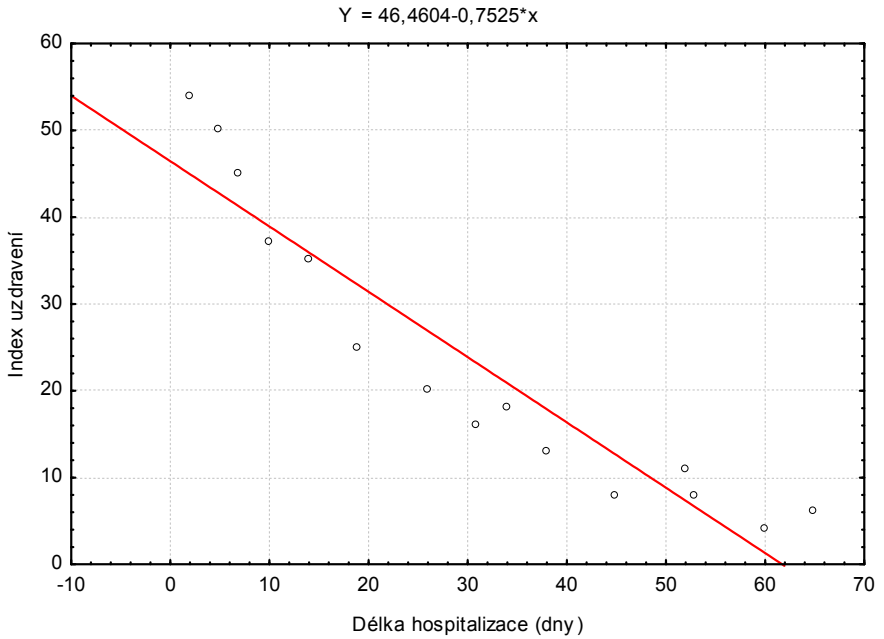
Příklad Kalama: Hodnota testové statistiky T=29.66, Nulovou hypotézu zamítáme na hladině 0,05 (p-hodnota =0,00). Výška dětí závisí na jejich věku. Koeficient determinace je $R^2 = 0{,}98$

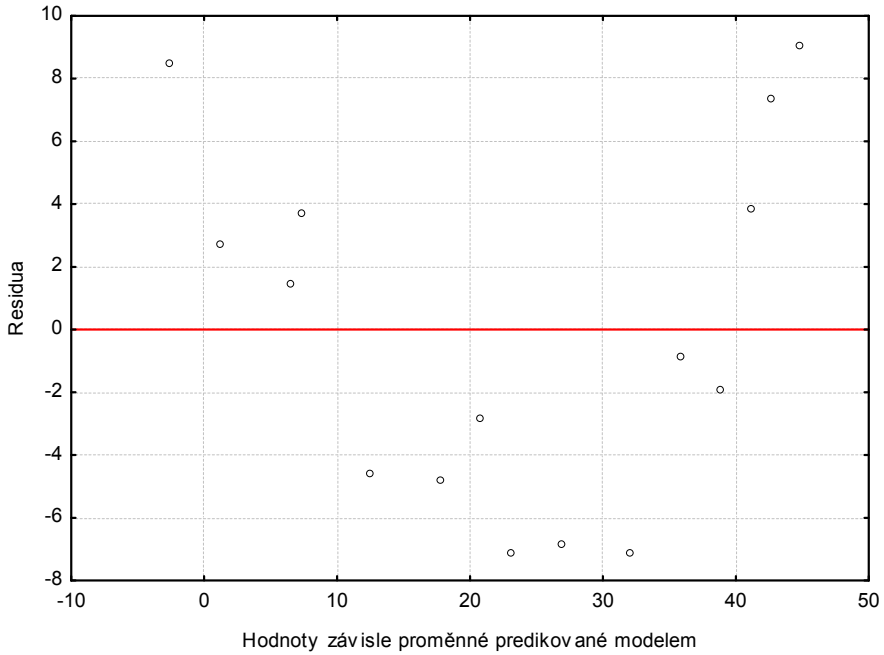



Bodový graf, ve kterém jsou vykresleny hodnoty residuí proti hodnotám $\hat{Y}_i$.
Residua náhodně fluktuují kolem nulové hodnoty, v závislosti na hodnotách $\hat{Y}_i$ nevykazují žádný systematický trend a ani jejich rozptyl není závislý na hodnotách $\hat{Y}_i$
Námi zvolený lineární tvar závislosti je vhodný.

# Příklad: Index uzdravení

- Existuje závislost mezi délkou hospitalizace pacienta v nemocnici (X, uvedeno ve dnech) a tzv. Indexem uzdravení (Y)?
- Y = 46,5 – 0.75X.
- Koeficient determinace tohoto lineárního modelu je poměrně vysoký, $R^2 = 0{,}88$
- Residua vs. Hodnoty predikované modelem $\hat{Y}_i$ vidíme, že residua jsou seřazena do tvaru písmene U.



# Transformace závisle a nezávisle proměnné

- Cíle

1. Odstranění nelineární závislosti mezi závisle a nezávisle proměnnou
2. Stabilizace rozptylu

- **„Žebřík transformací“:**

. . . ,.1/x2,.1/x,.1/√x, log x,√x, x, x2, . . . .

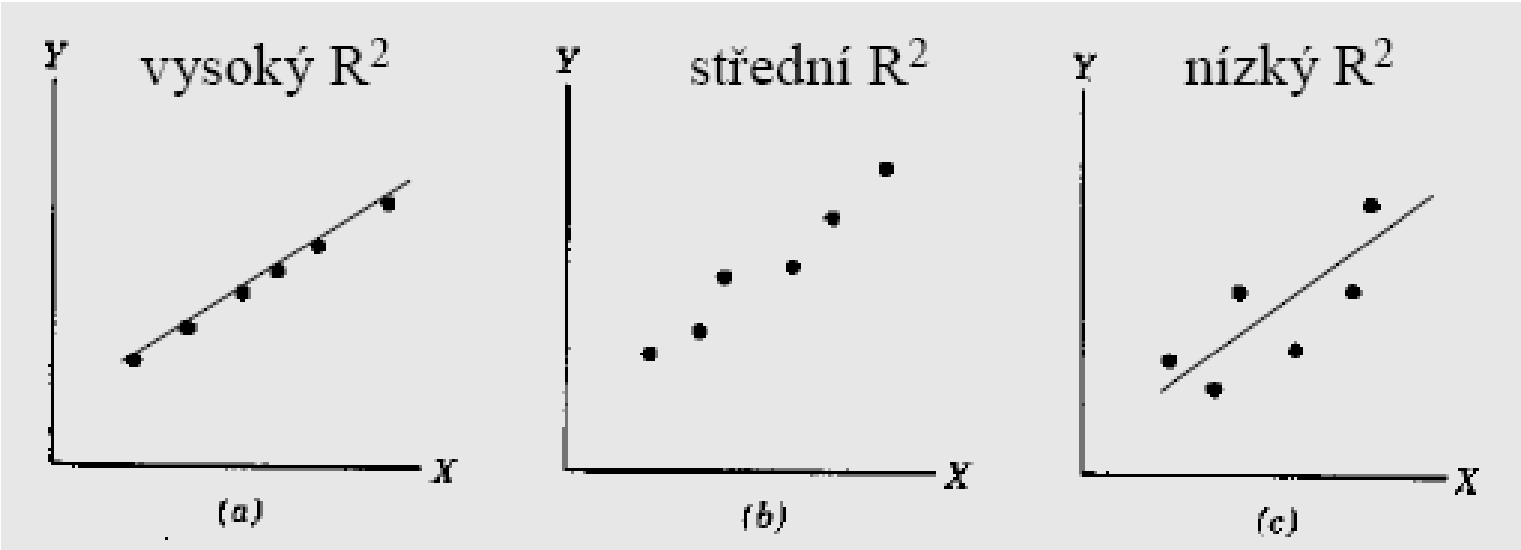
- Po tomto žebříku transformací se můžeme pohybovat buď nahoru (k vyšším mocninám) nebo dolů. Cílem je především linearizace závislosti.
- Když dosáhneme pohybem po zvoleném žebříku (na ose x nebo ose y) přibližně lineární závislosti, potom současným pohybem po obou žebřících se pokusíme také o stabilizaci rozptylu.

# Koeficient determinace

Jak úspěšná byla regrese?

- **Koeficient determinace** je definován jako podíl celkové variability závislé veličiny, která je vysvětlena závislostí.
- Jedná se o podíl vysvětlené a celkové variability náhodné veličiny Y.

$$R^2 = \frac{\text{variabilita vysvetlena modelem}}{\text{celkova variabilita } Y} = 1 - \frac{\text{residualní variabilita}}{\text{celkova variabilita } Y} = 1 - \frac{S_e}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$




(Lepš 1996)

# Koeficient determinace - vlastnosti

- Koeficient determinace udává relativní velikost variability závisle proměnné, kterou se uvažovanou závislostí podařilo vysvětlit.
- Koeficient determinace nabývá hodnot od 0 do 1.
- Čím vyšší je hodnota koeficientu determinace, tím je náš regresní model lepší.
- V případě regrese s jedinou nezávisle proměnnou je hodnota koeficientu determinace rovna kvadrátu Pearsonova korelačního koeficientu mezi veličinami X a Y.

$$R^2 = corr(X,Y)^2$$

# Nelineární regresní model
# Exponenciální závislost

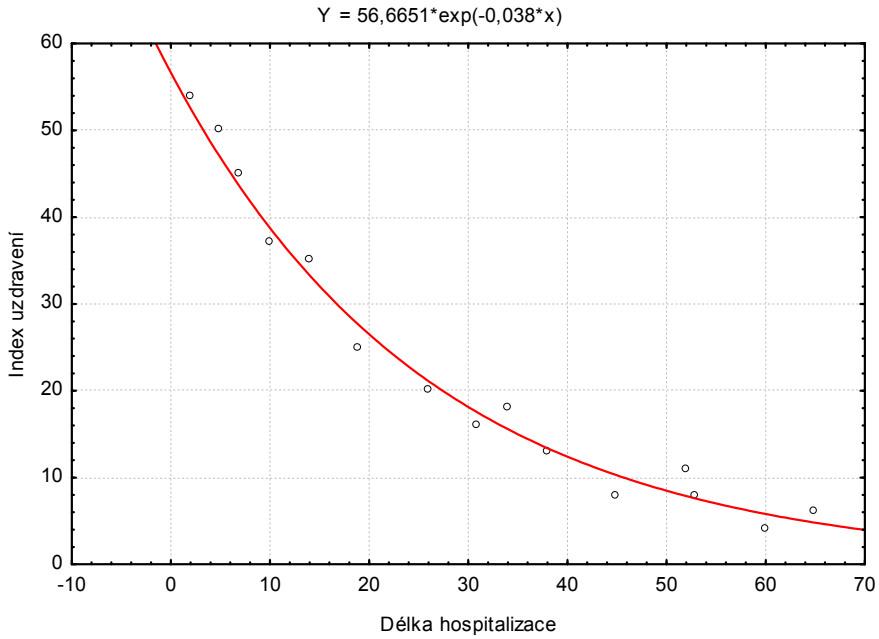
- Obecný tvar exponenciální závislosti je

$$Y = \beta_0 + \exp(\beta_1 + \beta_2 X)$$

- Je-li parametr $\beta_2$ kladný, pak s rostoucími hodnotami *X* rostou i hodnoty *Y*. Je-li parametr $\beta_2$ záporný, pak s rostoucími hodnotami *X* klesají hodnoty *Y*. Parametr $\beta_2$ charakterizuje strmost nárustu resp. poklesu, parametry $\beta_0$ a $\beta_1$ „mají na starost“ umístění křivky. Bude-li například hodnota $\beta_0 = 0$ a $\beta_2$=-2, pak při nárustu hodnoty *X* o jednu jednotku, dojde ke snížení hodnoty závisle proměnné $\exp^2 = 2{,}71^2 = 7{,}3$ krát. Křivka bude klesající a její hodnota se bude se vyrůstající hodnotou *X* blížit nule.

# Příklad: Index uzdravení Exponenciální závislost

- Existuje závislost mezi délkou hospitalizace pacienta v nemocnici (X, uvedeno ve dnech) a tzv. Indexem uzdravení (Y)?
- Y = 0 + 56,6*exp(-0,038X) = 0 + exp( 4.036-0,038X)



# Exponenciální závislost v přírodě

- Počet buněk se zvyšuje exponenciálně. Z každé buňky vzniknou dělením dvě nové buňky. V každé nové generaci je dvojnásobně více buněk než v té předchozí. **Podíl** počtu buněk v po sobě následujících generacích **je konstantní**. (V případě lineární závislosti by byl **rozdíl** počtu buněk mezi po sobě následujícími generacemi **konstantní**).

# Exponenciální závislost

- Arabský matematik **Ibn Kallikan** v roce 1256 popsal jeden z prvních šachovnicových hlavolamů. Na první pole šachovnice je umístěno zrnko rýže a na každé následující pole je umístěn dvojnásobek zrnek z pole předchozího. Kolik bude celkem zrnek rýže na šachovnici?

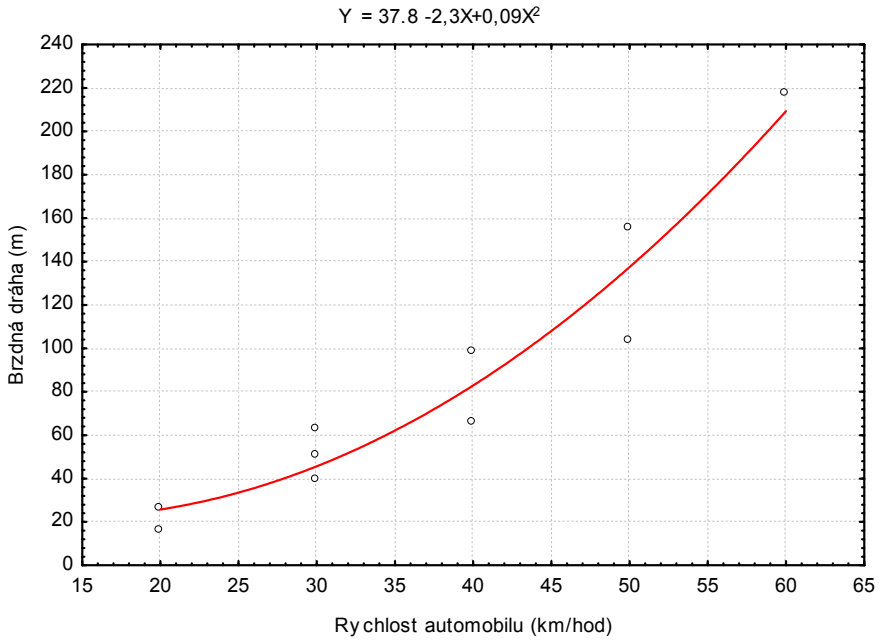
# Nelineární regresní model
# Polynomiální závislost

- Závislost brzdné dráhy automobilu na jeho rychlosti.
- Regresní rovnice obsahuje polynom druhého stupně (má kvadratický člen).

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$Y = 37.8 - 2{,}3X + 0{,}09X^2$$

- Grafem závislosti brzdné dráhy na rychlosti je část paraboly.



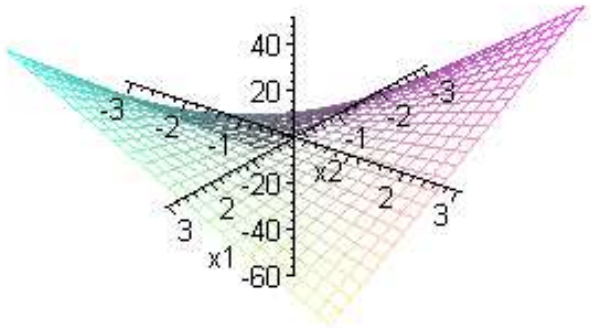
# Více nezávisle proměnných *(Multiple regression model)*

- Dvě nezávisle proměnné:
- Model: $Y_i = \beta_0 + \beta_1 X1_i + \beta 2 X2_i + \varepsilon_i$
- Koeficient beta1 lze interpretovat jako střední změnu Y při jednotkové změně X1 a nezměněné hodnotě X2.
- Nulová hypotéza $H_0 : \beta_1 = 0$ znamená, že populační průměr Y závisí nejvýše na X2.
- Tj. platí, že $Y_i = \beta_0 + \beta 2 X2_i + \varepsilon_i$
- Další interpretace $H_0 : \beta_1 = 0$ je, že proměnná X1 nepřináší žádnou informaci o střední hodnotě Y nad tu, která je již obsažena v X2.
- Snaha o co nejjednodušší model, obsahující jenom významné prediktory (nezávisle proměnné)

# Regresní plocha
## *(Response surface, regression surface)*

- Model s interakcemi

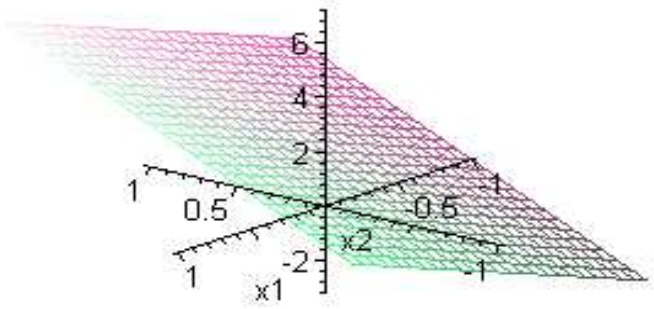
$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X1 \cdot X2$$




$$Y = 2 + 3X1 + 2X2 - 5X1X2$$

- Model bez interakcí – regresní rovina (*plane*)

$$Y = \beta_0 + \beta_1 X1 + \beta 2 X2$$




$$Y = 2 + 3X1 + 2X2$$

# T-test, F-test

- t-test: $H_0 : \beta_1 = 0$ nebo $H_0 : \beta_2 = 0$
- F test: $H_0 : \beta_1 = \beta_2 = 0$
- Upozornění: opakovaný t-test a F-test mohou dávat nekonzistentní výsledky

- **Podmodel** = jednodušší model obsahující pouze některé nezávisle proměnné (signifikantní) původního regresního modelu.
- S každou mocninou veličiny musí být v modelu všechny mocniny nižšího stupně, se součinem veličin musí být v modelu také všechny složky tohoto součinu.

# Opakování ANOVA

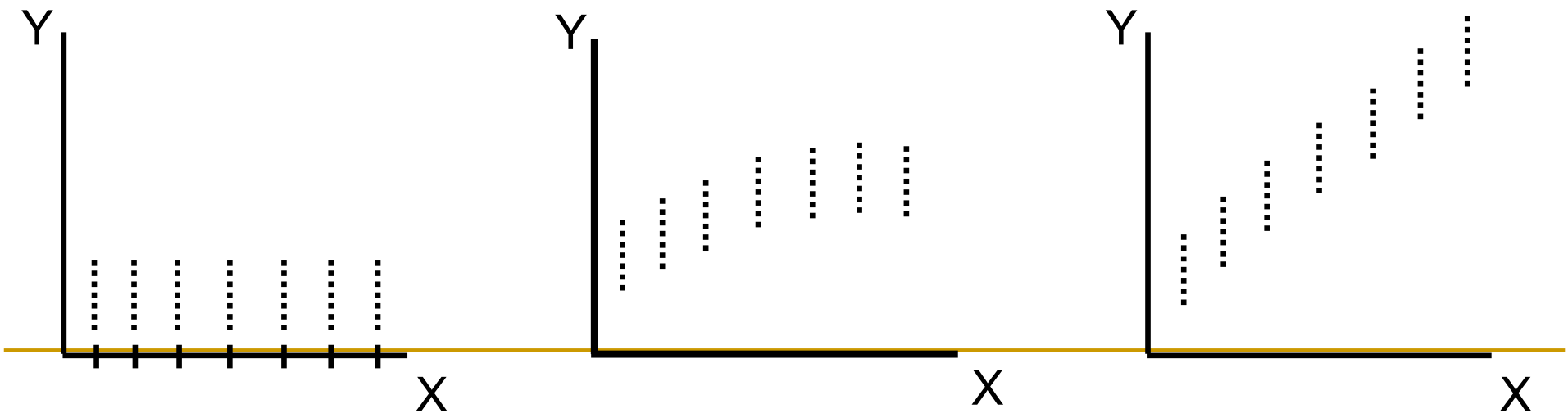
## 1) Experimentální data

| $y_1$ | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| $y_n$ | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
| | $s_0^2$ | $s_1^2$ | $s_2^2$ | $s_3^2$ | $s_4^2$ |

## 2) Celková ANOVA "one way"

| Zdroj rozptylu | St.v. | SS | MS | F |
|---|---|---|---|---|
| Mezi skupinami | a-1 | $SS_B$ | $SS_B/(a-1)$ | MSB/MSE |
| Uvnitř skupin | na-a | $SS_E$ | $SS_E/(na-a)$ | |
| Celkem | na-1 | $SS_T$ | $s_y^2$ | |

$$= \frac{SS_T}{na-1}$$



# ANOVA jako nástroj analýzy regresních modelů - příklad na modelu přímky

**3)** **Celková ANOVA** → $SS_B/SS_T$ (variance ratio)

$MS_B/MS_E = F$

**4)** **Analýza rozptylu regresního modelu (zde přímky)**

| Zdroj rozptylu | st.v. | SS | MS | F |
|---|---|---|---|---|
| Model (přímka) | 1 | $SS_{MOD}$ | $MS_{MOD}$ | $MS_{MOD} / MS_R$ |
| Residuum | na - 2 | $SS_R$ | $MS_R$ | |
| celkem | na - 1 | $SS_T$ | | |

$(SS_{MOD}/SS_T) . 100 =$ % rozptylu Y "vyčerpaného" přímkou = koeficient determinace ($R^2$)

# Příklad

**X:** konc.Cd: 1,2,3,4,5,6 ng/ml
**Y:** absorb: 0,23; 0,49; 0,72; 0,90; 1,16; 1,39

**b=0,228** $S_b = 4{,}99.10^{-3}$ **P = 0,000**
**a=0,016** $S_a = 0{,}019$ **P = 0,457**

r = 0,999

$R_2$ = 99,81%

**St. Error of est: 0,021**

**ANOVA**

| Source | D.f. | SS | MS | F | P |
|---|---|---|---|---|---|
| Model | 1 | 0,912 | 0,912 | 2086,3 | 0 |
| Residual | 4 | 0,0017 | 0,000425 | | |
| Total ( c ) | 5 | 0,9138 | | | |

$s^2_{y.x} = 4{,}25 \cdot 10^{-4}$

$s^2_y = 0{,}18275$

# Strategie hledání vhodného podmodelu
# Sekvenční postupy

- **Sestupný výběr** - Nejprve se spočítá nejbohatší model, pak se jednotlivé regresory postupně z modelu vylučují. V každém kroku se vylučuje takový regresor, který v daném modelu nejméně přispívá k vysvětlení.
- **Vzestupný výběr** – opak sestupného výběru. Vyjde se z prázdné množiny regresorů, do níž se pak v každém kroku přidá vždy ten z ještě nezařazených regresorů, který v daném kroku co možná nejlépe zlepší vysvětlení závisle proměnné.
- **Kroková (stepwise) regrese** - kombinuje oba předešlé postupy. Vzestupný výběr je v každém kroku kombinován s pokusem o zjednodušení pomoci sestupného výběru.
- Každá z popsaných metod může dát jiný výsledný model, kromě jiného závisí také na volbě hladin testů.
- Zejména u krokové regrese se doporučuje najít několik téměř optimálních modelů a pokusit se najit mezi nimi ten,který má nejlepší interpretaci.

# Umělé proměnné *(Dummy variables, dummies)*

- Vyjádření nominální veličiny s více než 2 hodnotami
- j úrovní faktoru -> j-1 umělých proměnných (v modelu buďto všech j-1 umělých proměnných nebo žádná)

| Proměnná | Umělé proměnné (stačí 3) | | | |
|---|---|---|---|---|
| Rodinný příslušník (4 úrovně) | Otec (0/1) | Matka (0/1) | Strýc (0/1) | Dědeček (0/1) (zbytečná) |
| „otec“ | 1 | 0 | 0 | 0 |
| „matka“ | 0 | 1 | 0 | 0 |
| „strýc“ | 0 | 0 | 1 | 0 |
| „dědeček“ | 0 | 0 | 0 | 1 |