Regrese a korelace

Regrese *versus* korelace

Regrese (*regression*)*

popisuje vztah = závislost dvou a více kvantitativních (popř. ordinálních) proměnných formou funkční závislosti

Korelace *(correlation)*

měří těsnost vztahu = závislosti mezi dvěma proměnnými

Liší se chápání proměnných u obou metod?

<u>Regrese</u>:

lze rozlišit, která proměnná závisí na které (= příčinnost) – rozlišujeme nezávislou *(independent;* není zatížena „chybou“) a závislou (*dependent, response; prediktor,* je zatížena chybou) proměnnou (*variable*) *(Pozn.: lze řešit i případy, kdy obě proměnné jsou zatíženy chybou…)*

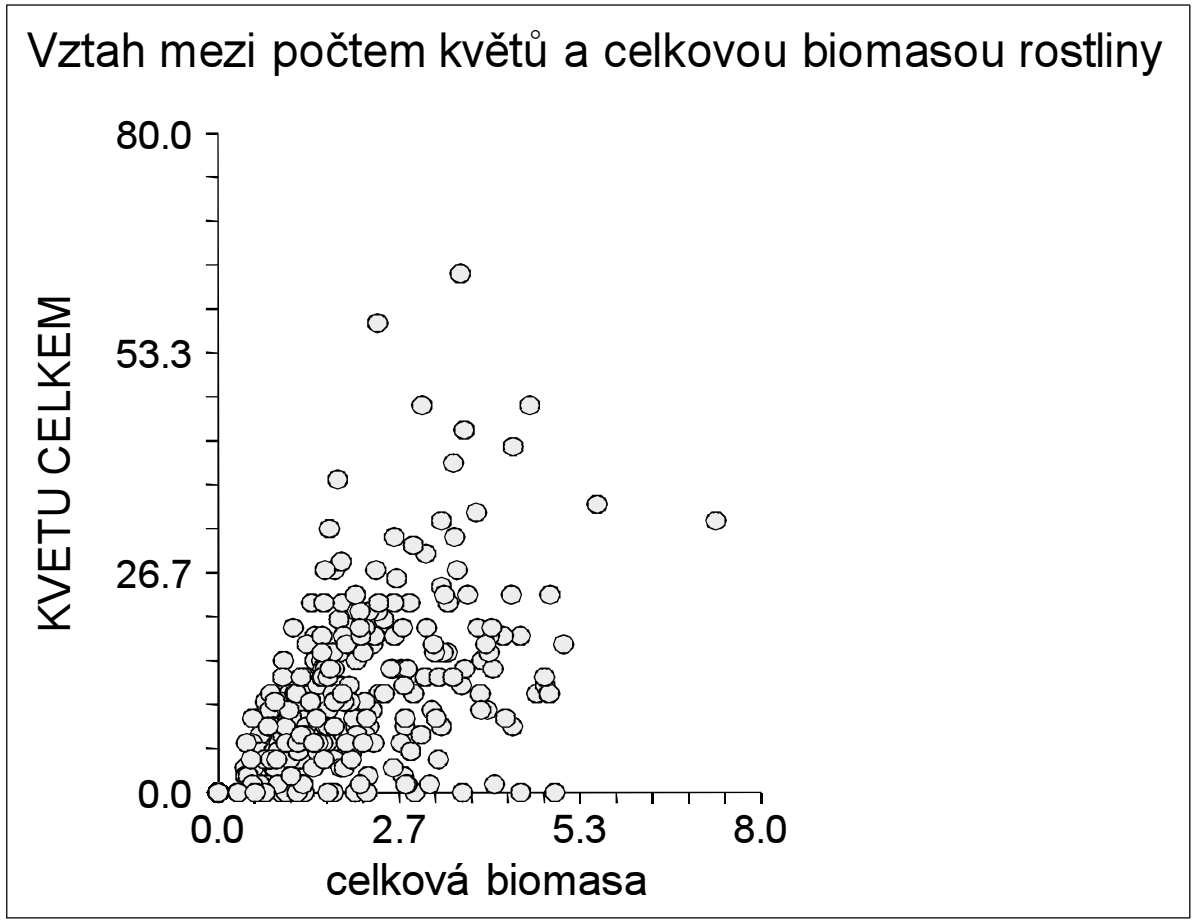
<u>Korelace</u>:

nelze rozlišit proměnné na závislou a nezávislou (obě jsou zatíženy chybou)

** v užším slova smyslu*

Graf závislosti
(*Scatter plot* = *x-y* graf)

Osa ***y*** = závislá proměnná



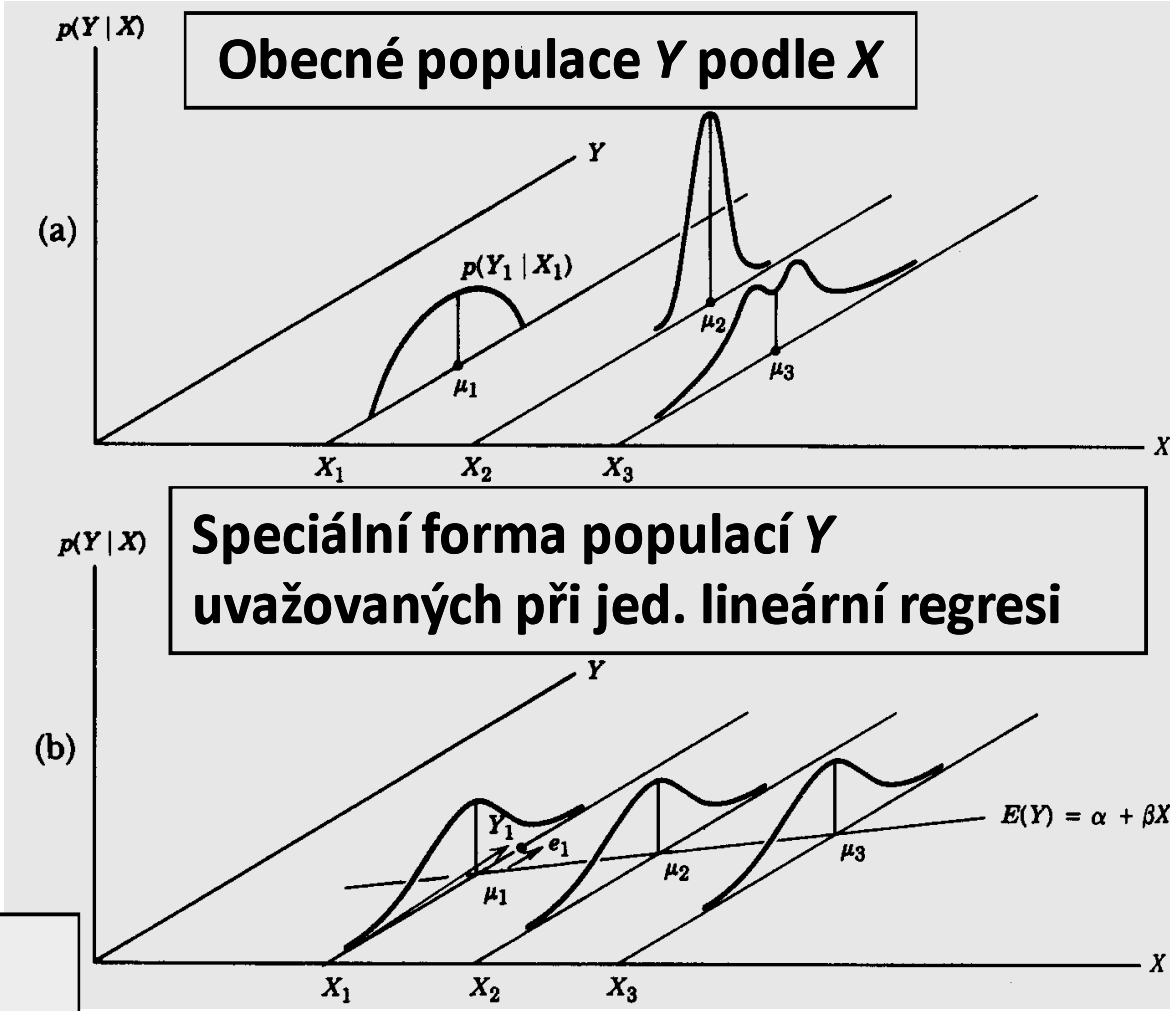
Osa ***x*** = nezávislá proměnná

Jednoduchá regrese - lineární model

- nejjednodušší případ regrese: „jednoduchá“ = pouze 1 nezávislá proměnná; „lineární“ = závislost *y* na *x* vyjadřujeme přímkou

Předpoklady lineární regrese:

1. **homogenní rozptyl:** všechna *Y* mají stejnou rozptýlenost
2. **linearita:** střední hodnota y_i leží na regresní přímce populace
3. **nezávislost**
4. **proměnná *X* je měřena bez chyby** (pokud ne, pak užít Model II regrese*)



Náhodné veličiny $Y_1, Y_2, .., Y_i$ jsou nezávislé se střední hodnotou $\alpha+\beta x_i$ a rozptylem σ^2

Jednoduchá lineární regrese

(simple linear regression)

Obecný předpis

$$y_i = \alpha + \beta x_i$$

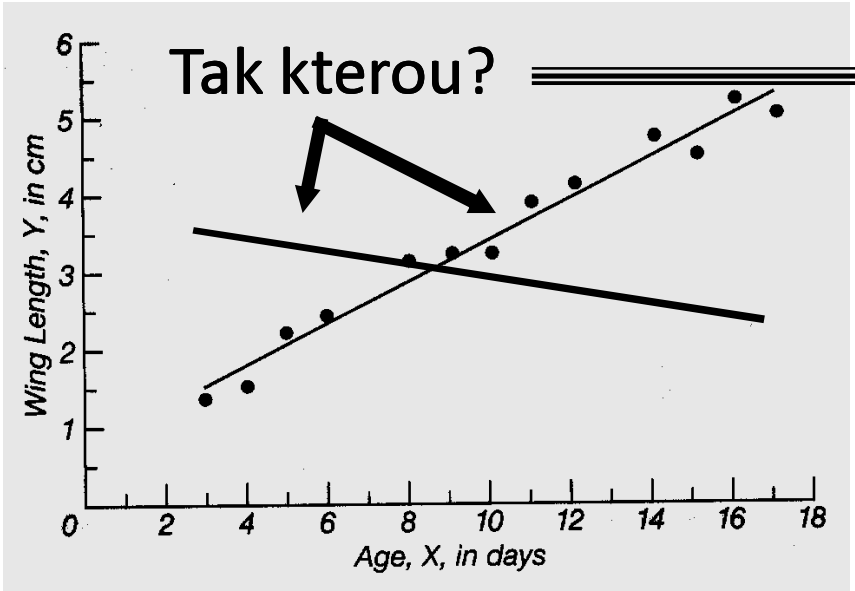
Jak to vypadá ale v populaci?

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

… chyba = reziduál (*residuum*)=odchylka očekávaného od skutečného y_i

Nezávislá proměnná *x* je na horizontální ose *x*, závislá proměnná *y* je na vertikální ose *y*.

Jak ale nalézt přímku, která bude „nejlépe“ fitovat = = prokládat naše data?



Musí platit, aby (reziduální) součet čtverců odchylek skutečných od očekávaných hodnot (RSS) byl minimální

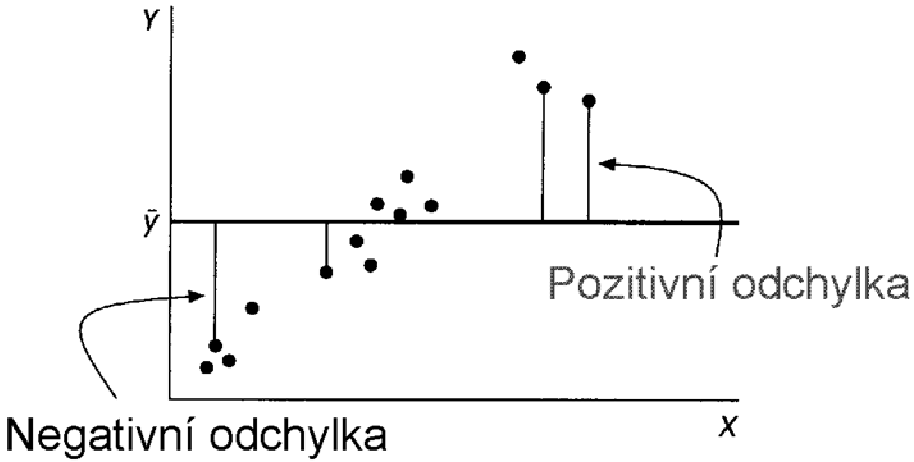
$$RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \min.$$

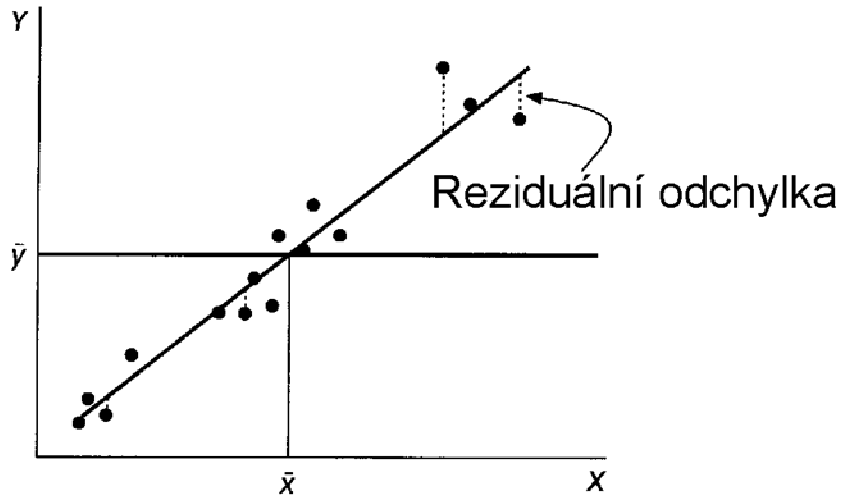
Jak vybrat nejlepší přímku?

1) spočítat celkový průměr všech hodnot *y* a vynést ho do grafu
2) přímka prochází bodem $(\bar{x}, \bar{y})$
3) rotuj přímku kolem bodu (2) přičemž $SS_Y = SS_R + SS_E$ a musí být splněno

$$SS_Y = RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{min.}$$

4) pak byla nalezena nejlepší možná přímka





(Grafen a Hails 2002, str. 24)

Výpočet parametrů regresní rovnice

$y_i = \alpha + \beta x_i$ ➡ $y = a + bx$ **... „odhad“**

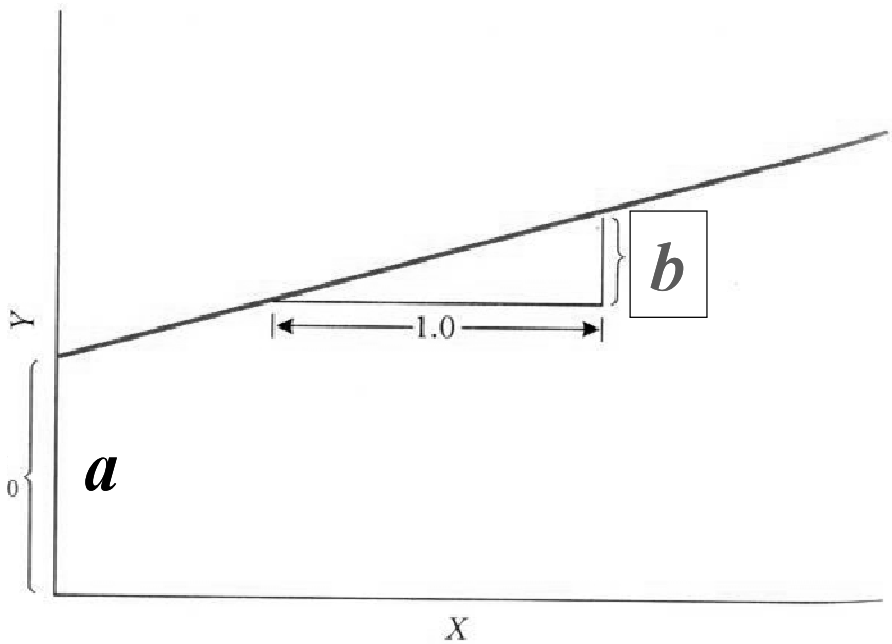
(pracujeme s výběrem)

β *(b)* = regresní koeficient (směrnice přímky, *slope*)
může nabývat všech hodnot na reálné ose
směrnice *b* = změna *Y*, která je způsobena změnou *X* o jednotku

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

α *(a)* = absolutní člen (průsečík s osou *y*; *intercept*)
- bod daný koordinátami $(\bar{x}, \bar{y})$ leží vždy na přímce,
pak

$$a = \bar{y} - b\bar{x}$$



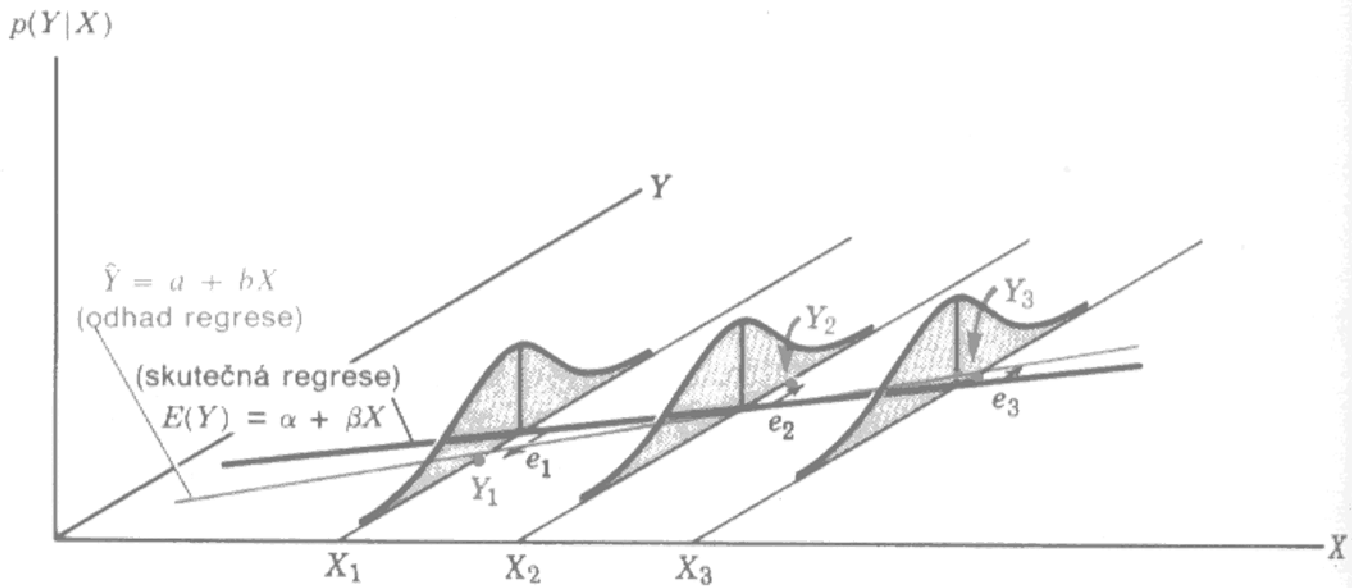
Jak blízko je odhadovaná přímka skutečné regresní přímce populace?

Odhad *b* je zatížen chybou (*SE*$_b$; s$_b$; *standard error of b*): ta má normální rozdělení

$$SE = \frac{s_y}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\frac{1}{n-2}\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Jak „zlepšit“ odhad *b*?

1. **snížením rozptylu *Y***
2. **zvýšením *n***
3. **zvýšením rozptylu *X***



(Wonnacot & Wonnacot 1993, str. 406)

Dělení variability v regresi

Celková variabilita proměnné *y* je

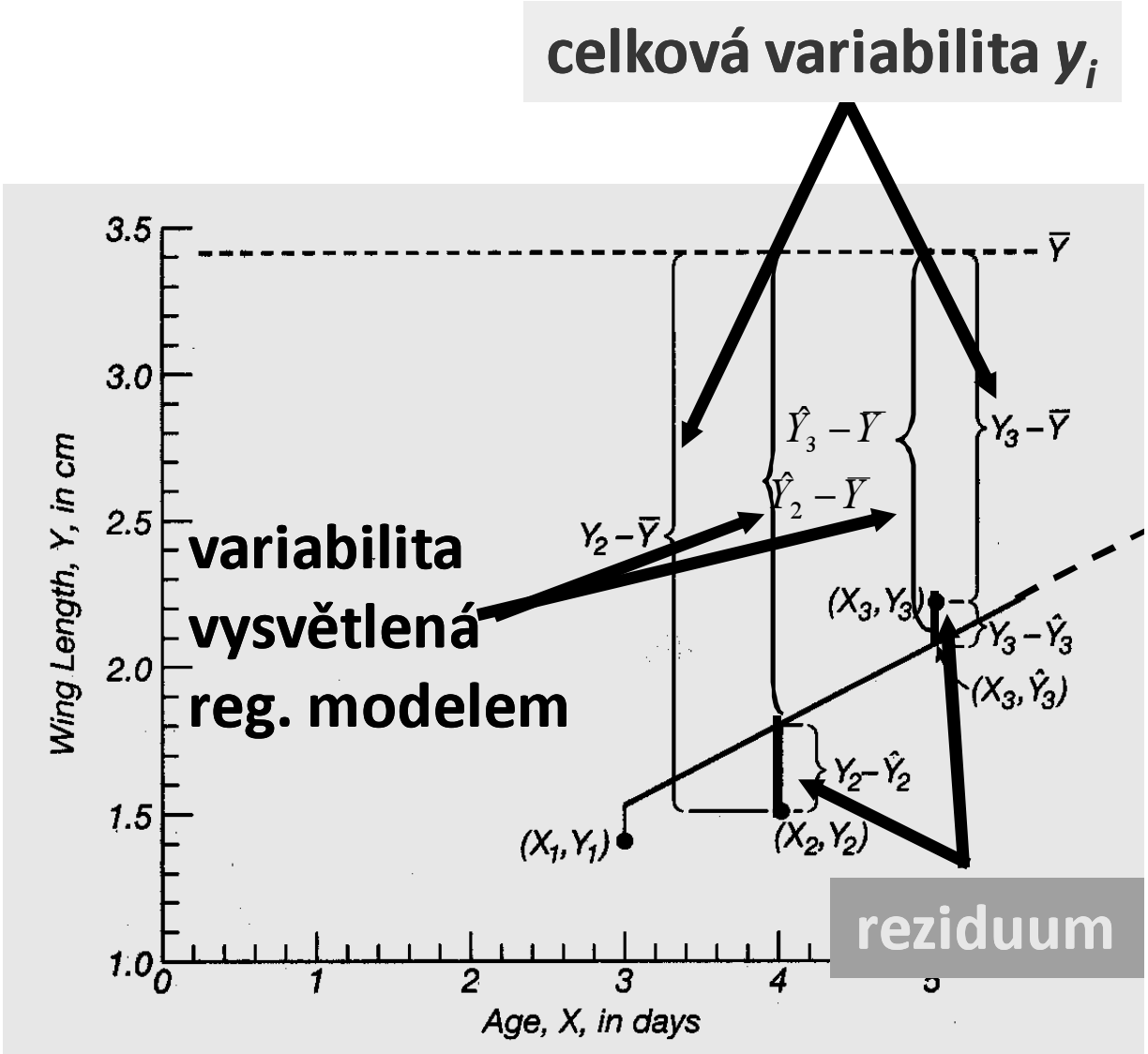
$$SS_{TOTAL} = \sum (y_i - \bar{y})^2$$

Tuto variabilitu lze rozdělit na variabilitu vysvětlenou regresním modelem:

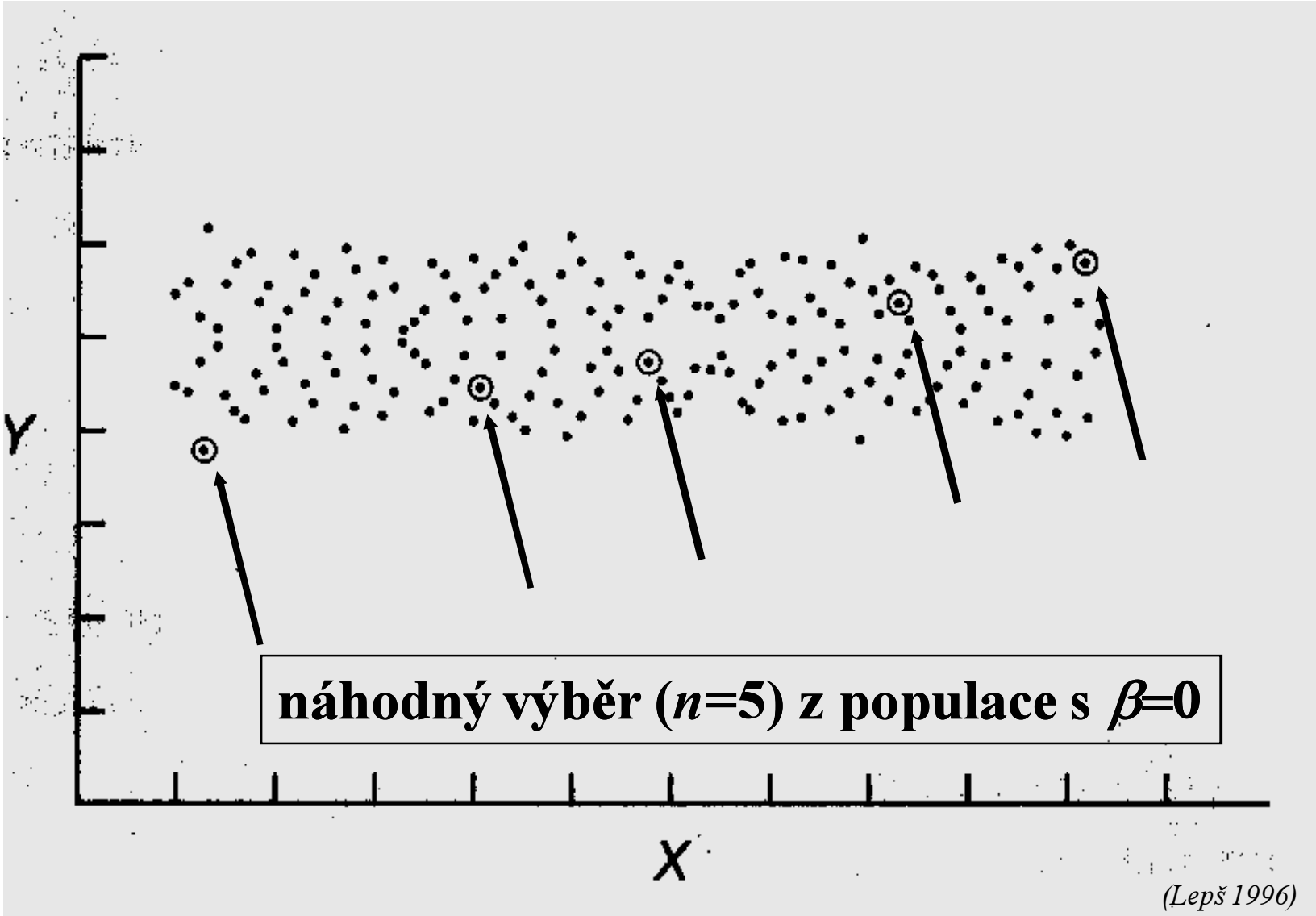
$$SS_{REG} = \sum (\hat{y}_i - \bar{y})^2$$

a zbytkovou, nevysvětlenou variabilitu:

$$SS_{ERROR} = \sum (y_i - \hat{y}_i)^2$$



Testovat regresi? Proč? A co?



(Lepš 1996)

Testování signifikance regrese

H_0: b=0 [$y_i = \alpha + \varepsilon_i$]
H_A: b≠0 [$y_i = \alpha + \beta_1 x_i + \varepsilon_i$]

Máme 2 možnosti: ANOVA a t-test

porovnáme obě dílčí variability:
čím menší bude SS_{ERROR} , tím těsnější bude vztah mezi skutečnými hodnotami a přímkou

DF_{REG}=1
DF_{ERR}=*n*-2

$$MS_{REG} = \frac{SS_{REG}}{1} \qquad MS_{ERR} = \frac{SS_{ERR}}{n-2}$$

pak

$$\boxed{F = \frac{MS_{REG}}{MS_{ERR}}}$$

je-li $F \geq F_{\alpha(1),\,1,\,n-2}$
... zamítáme H_0

alternativně:

$$t = \frac{b - \beta_0}{s_b}$$

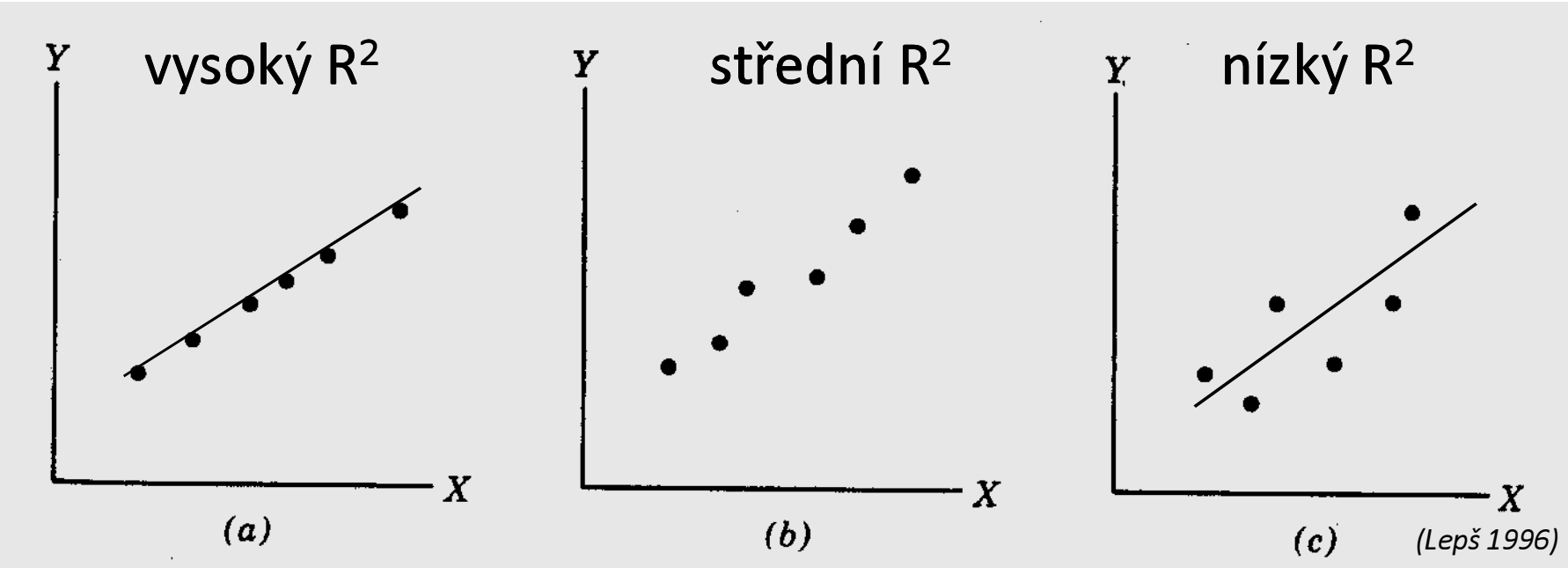
je-li $t \geq t_{\alpha(2),\,n-2}$
... zamítáme H_0

Jak „úspěšná“ byla regrese?

Koeficient determinace (*coefficient of determination*)

- udává (*rozsah <0;1>,* popř. *100 v % <0;100>*), jaká část variability závisle proměnné je vysvětlena regresním modelem

$$r^2 = R^2 = \frac{SS_{REG}}{SS_{TOT}}$$

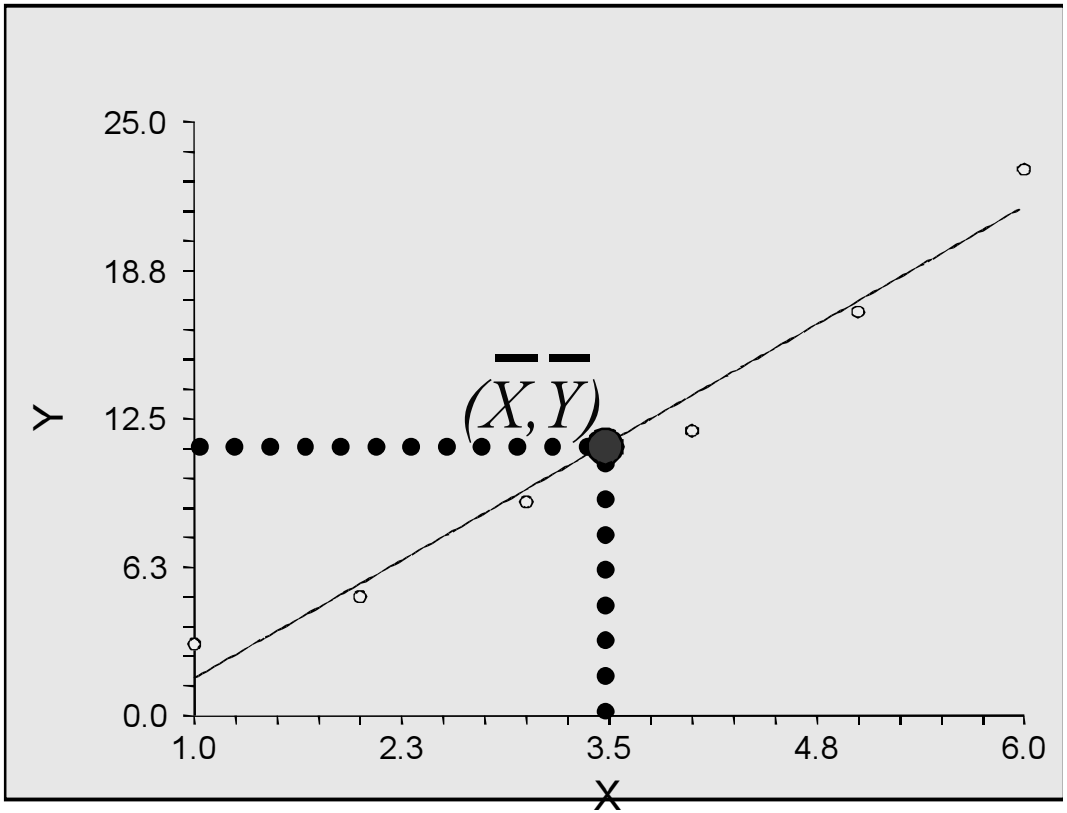


(Lepš 1996)

Lineární regrese - příklad

X	Y
1	3
2	5
3	9
4	12
5	17
6	23

	Y	X
n	6	6
Průměr	11.5	3.5
s	7.5299	1.8708



$$Y = -2.4 + 3.97 * X$$

Analysis of Variance Section

Source	*DF*	*Sum of Squares*	*Mean Square*	*F-Ratio*	*Prob Level*
Slope	*1*	*276.0143*	*276.0143*	*147.4885*	*0.0003*
Error	*4*	*7.485714*	*1.871429*		

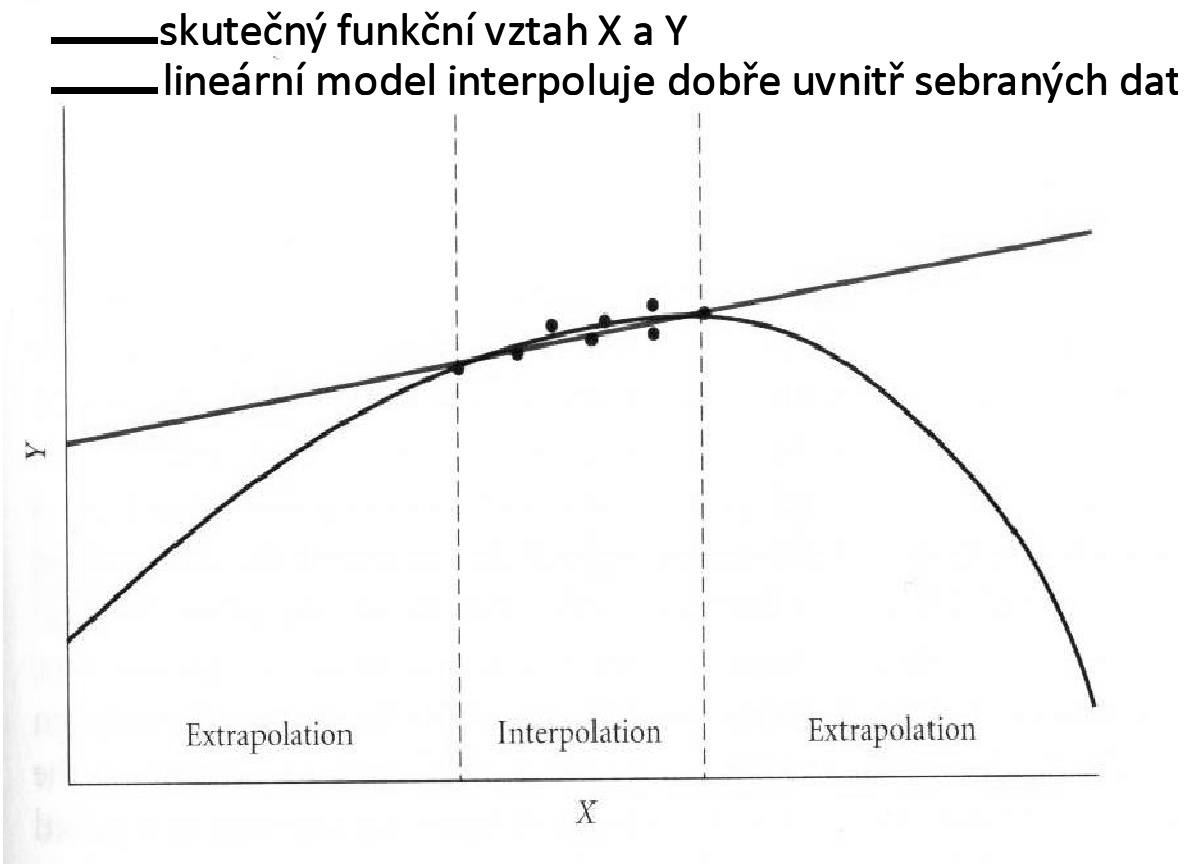
Interpolace, extrapolace

Interpolace = stanovení nových hodnot ležících uvnitř rozmezí sebraných dat

Extrapolace = stanovení nových hodnot ležících mimo rozmezí sebraných dat

Lineární modely mohou aproximovat nelineární závislosti v omezeném rozsahu nezávislé proměnné *X*.

- interpolace uvnitř těchto limitů bude akceptovatelná, i když mimo ně (extrapolace) nepopisuje skutečnou závislost *Y* na *X*



(Gotelli & Ellison 2004, str. 241)

Konfidenční a predikční intervaly

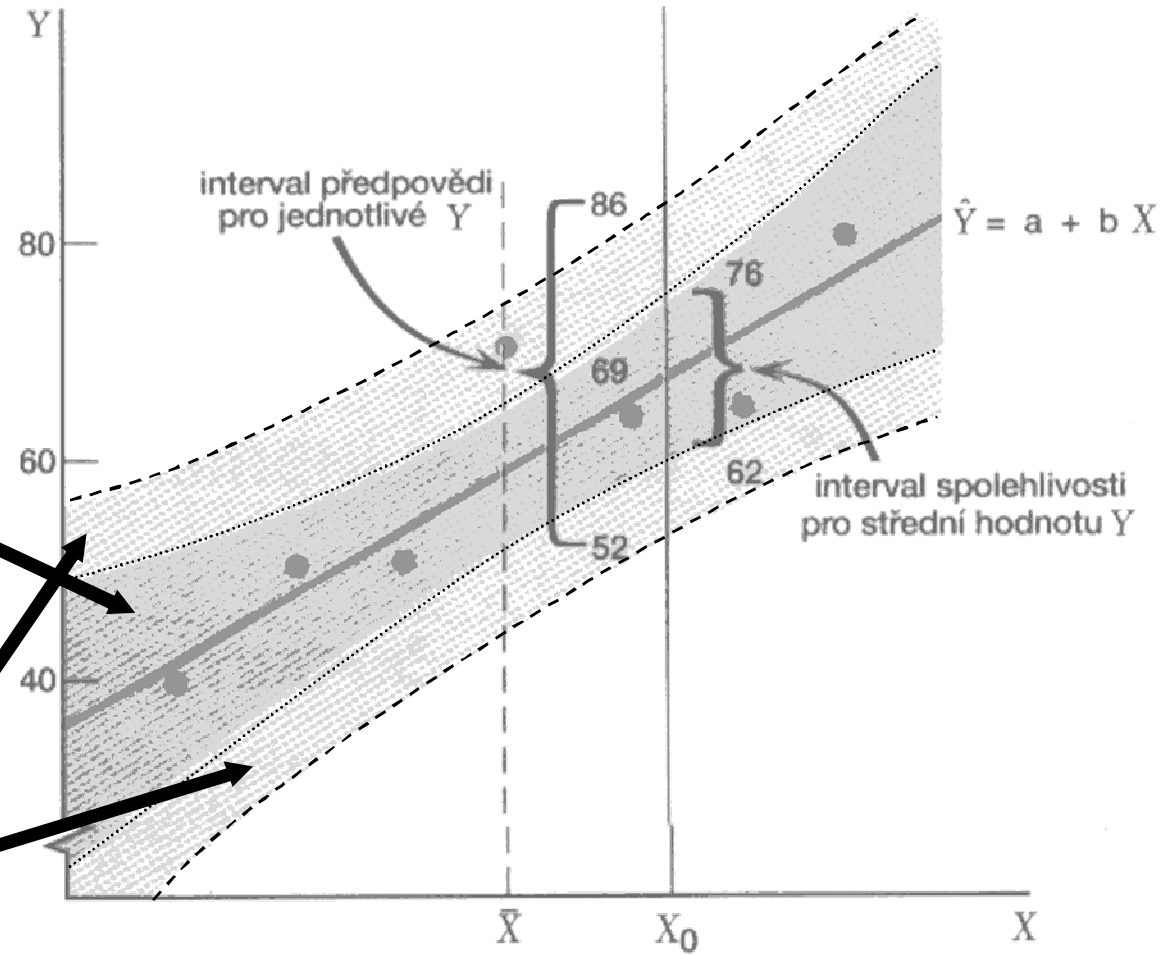
Cílem regresní analýzy není pouze predikce hodnoty závisle proměnné na nezávislé proměnné, ale i určení přesnosti takové predikce!

- **Pro β:**

$$b \pm t_{\alpha(2),n-2} * SE_b$$

- **Pro střední hodnotu y_i při daném x_i: konfidenční interval (*confidence bands*)**

- **Pro individuální (jednotlivé) hodnoty y_i při daném x_i: predikční interval (*prediction intervals*)**

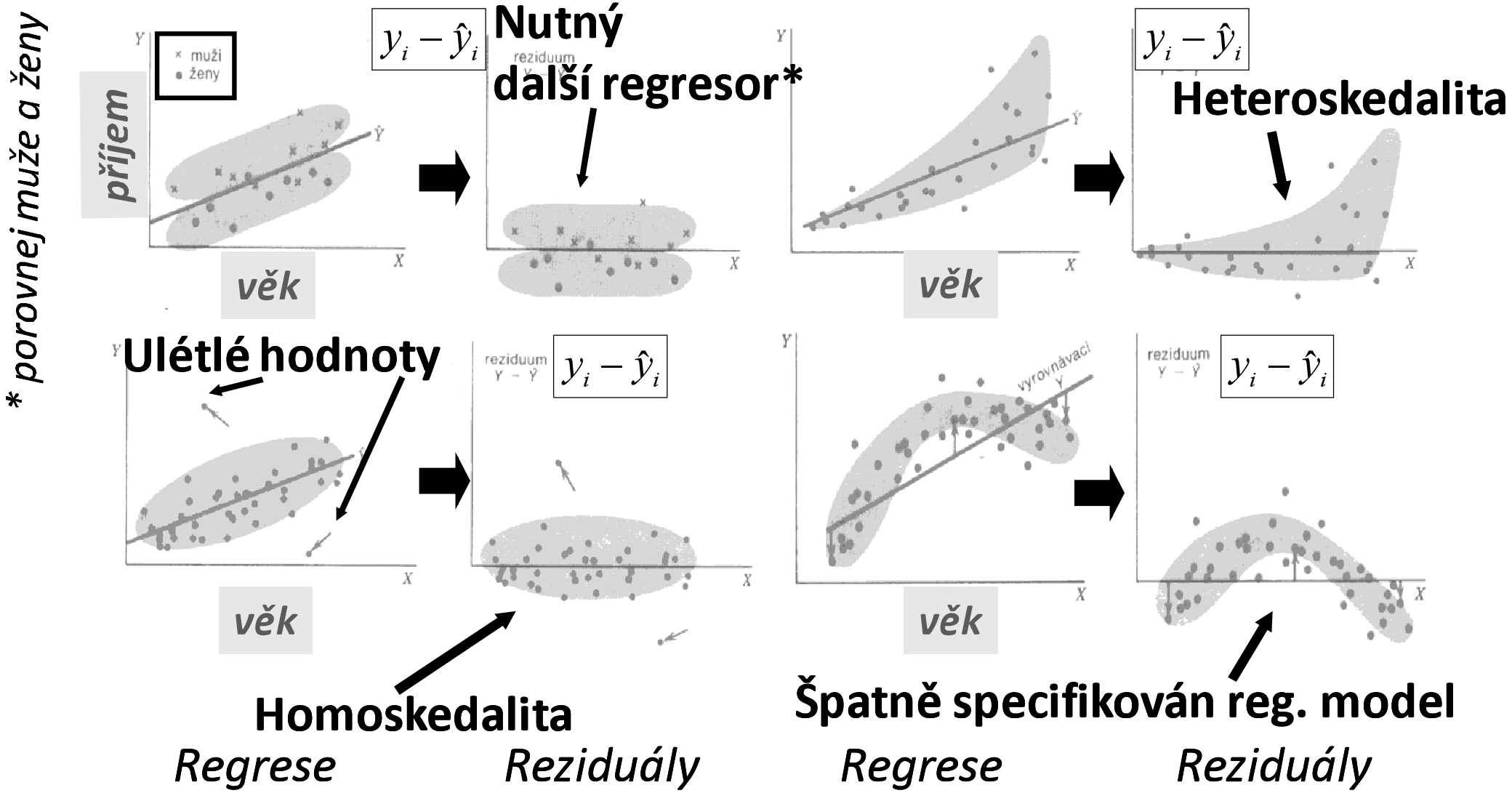


Diagnostické testy: kdy je použití lineární regrese chybné?

Reziduály *y* jako funkce jejich příslušného *x*

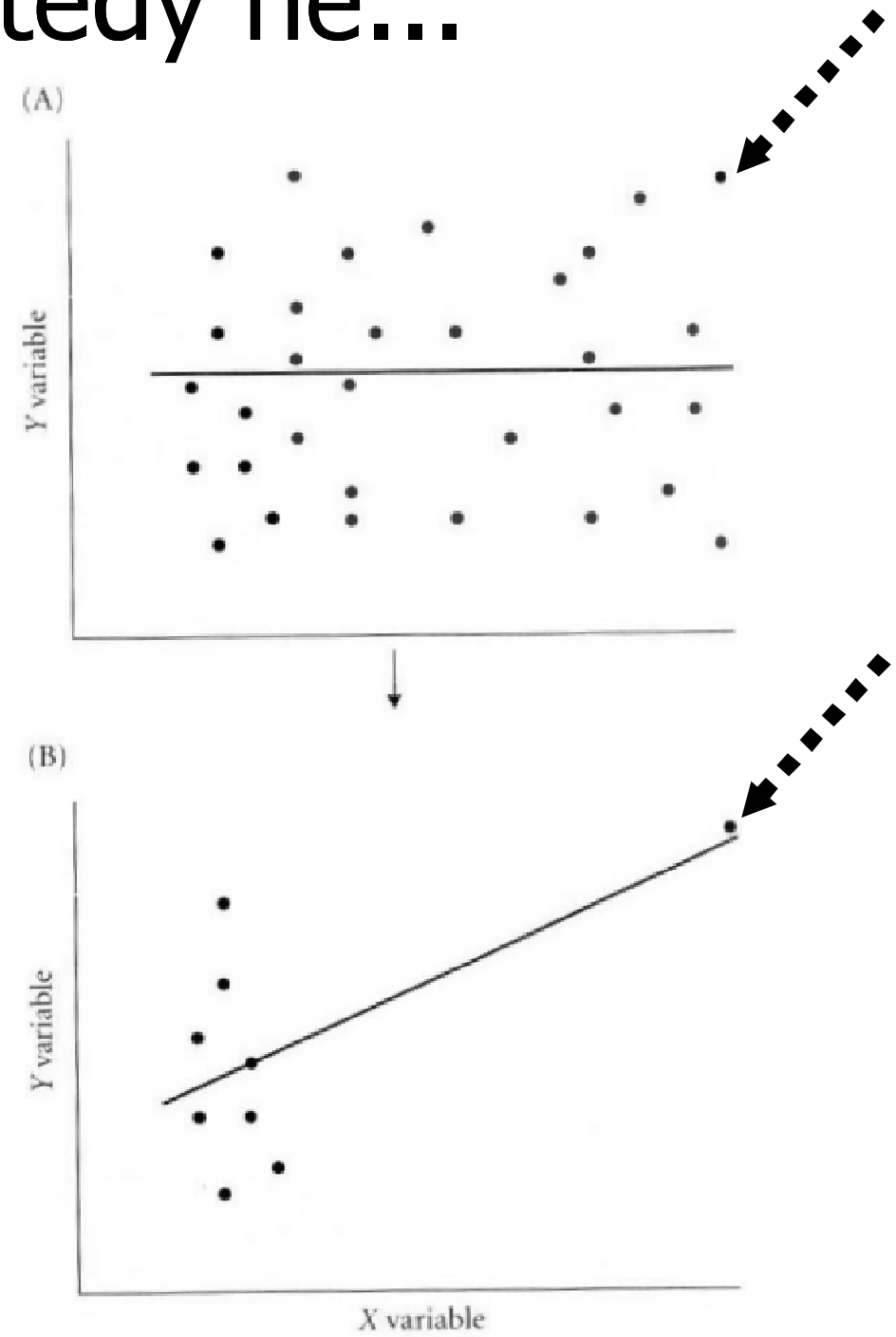
(graf residuálů = residual plot)

** porovnej muže a ženy*



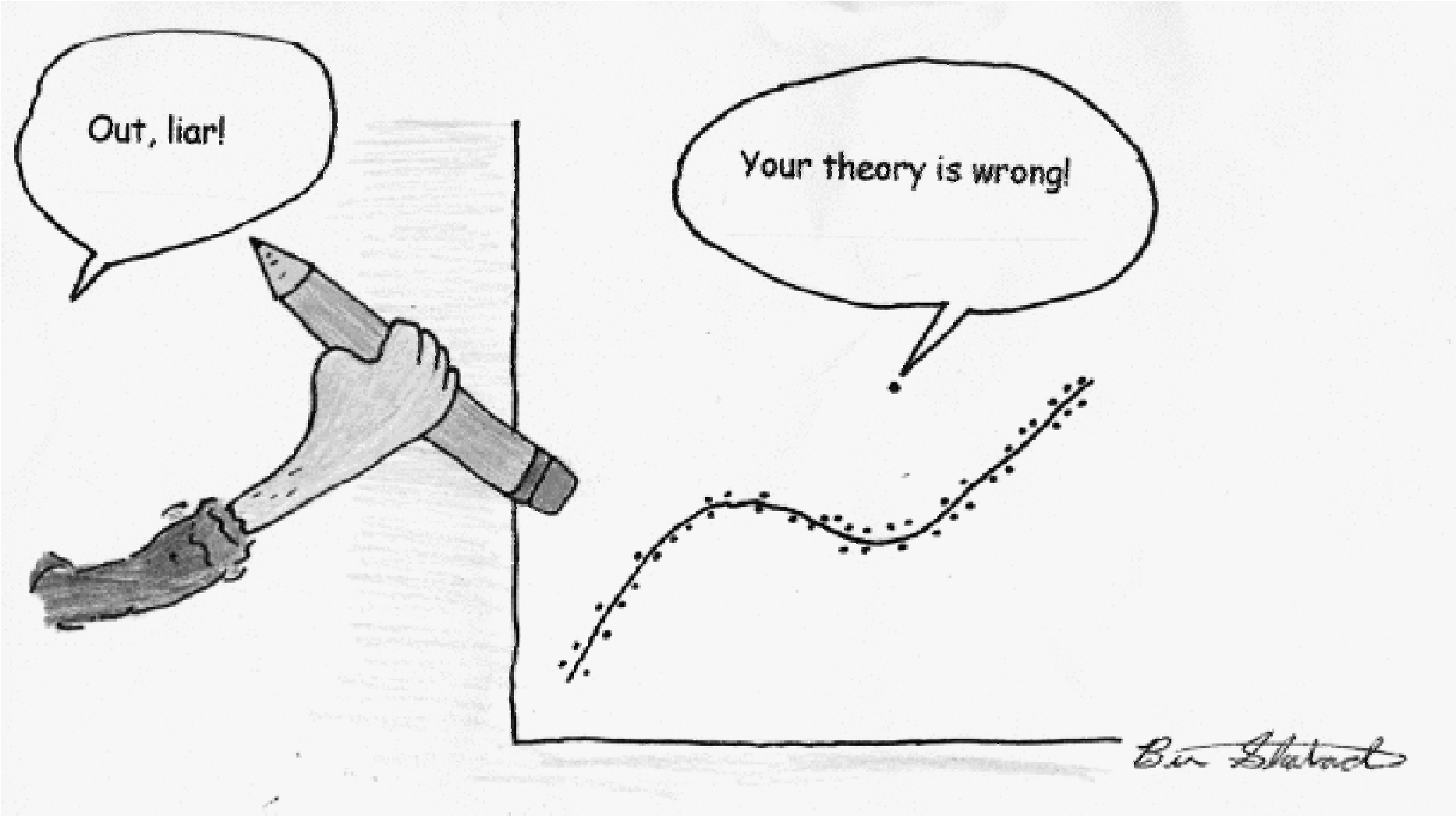
Tak takhle tedy ne...

Nepřiměřený vliv 1 bodu může zcela změnit závislost *y* na *x*, pokud není dodržena podmínka zahrnutí celého (resp. většího) rozmezí proměnné *x* do analýzy...



(Gotelli & Ellison 2004)

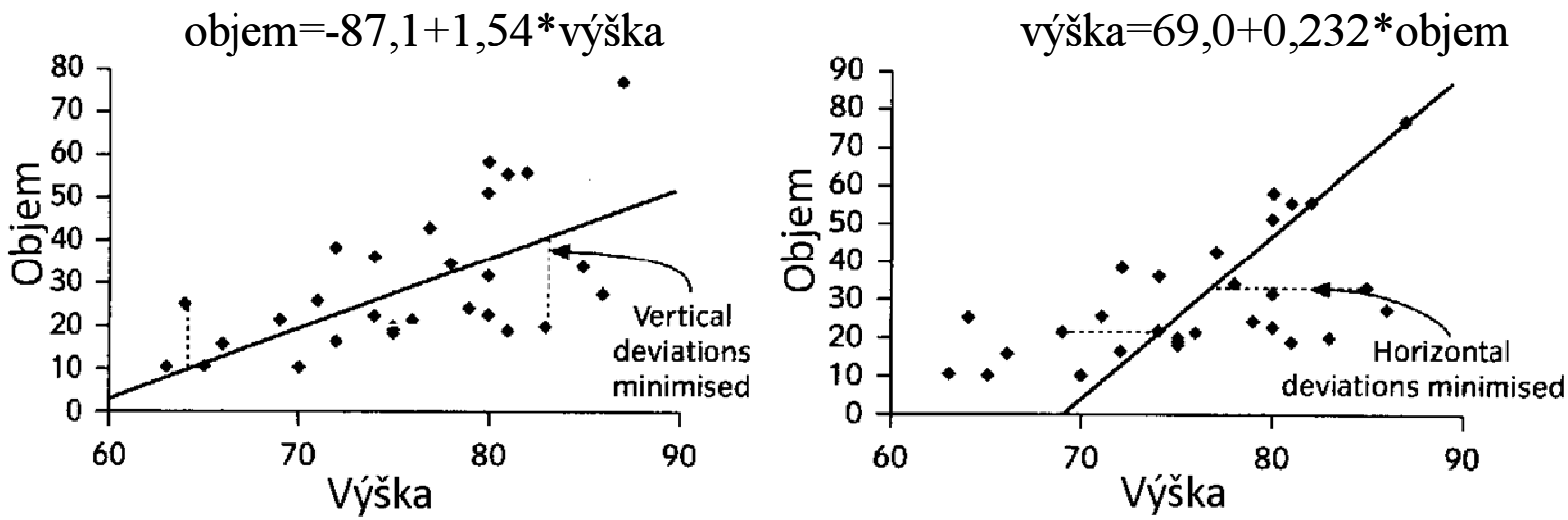
Ty ulétlé hodnoty...



Role X a Y: záleží které je které?

Model I regrese: jsou minimalizovány vertikální odchylky
→ jiný pohled: „fitovaná přímka popisuje vztah mezi naměřenými hodnotami X a očekávanými hodnotami Y“ → výsledný vztah je podmíněný použitým souborem hodnot X

Model II regrese: uvažujeme existenci chyby jak pro X, tak pro Y.
→ lze využít **pouze tehdy**, když zároveň
(a) jedinou příčinou odchylek naměřených hodnot od přímky je chyba měření (*measurement error*)
(b) tuto chybu lze přesně rozdělit mezi X a Y proměnnou



(Grafen a Hails 2002, str. 43)

Jiné typy regresní analýzy I.

Robustní regrese (*robust regression*):
- reziduály se počítají ne pomocí čtverce (rozdíl^2), ale jako např. absolutní odchylky, ale existuje řada jiných typů (např. *M-estimators, least-trimmed*,...)
- méně náchylná na vliv ulétlých hodnot (zvl. jsou-li skutečné)
- nelze použít klasického postupu, nutno užít **iterační procedury** pro nalezení parametrů přímky

Kvantilová regrese (*quantile regression*):
- minimalizuje odchylky od fitované regresní přímky, ale minimalizující funkce je asymetrická – pozitivní a negativní odchylky jsou váženy rozdílně

Jiné typy regresní analýzy II.

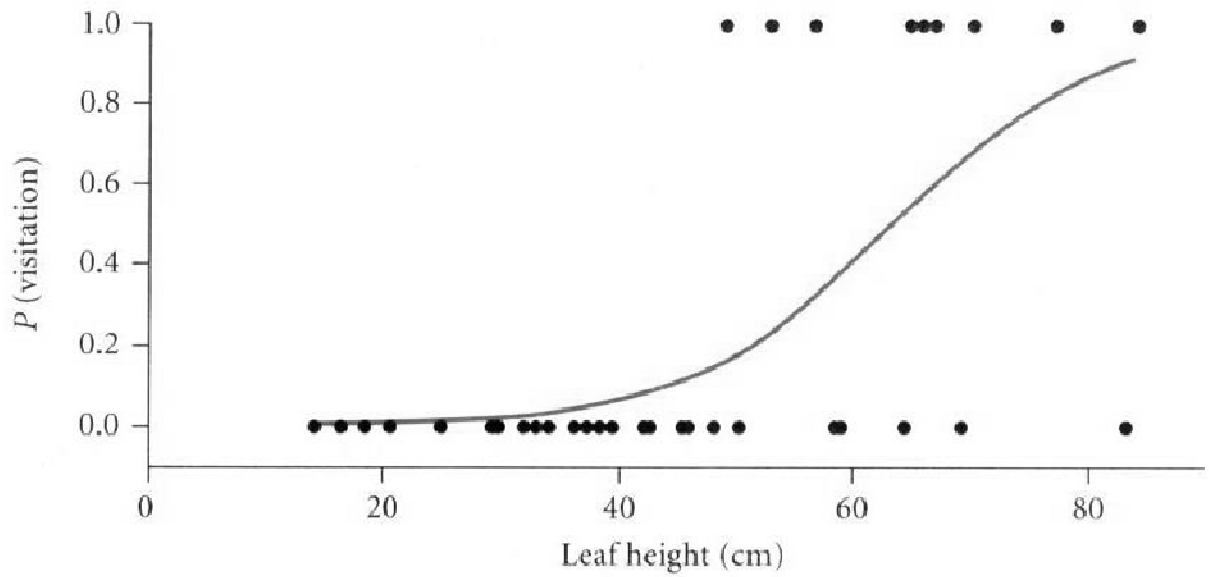
Logistická regrese (*logistic regression*):

- speciální forma regrese, kde **proměnná *y* je kategoriální** (nejčastěji 2 kategorie, méně často více)
- závislost je vyjádřena tzv. S-křivkou, tedy logistickou křivkou, která stoupá od jisté minimální hodnoty (0 a vyšší) do maximální asymptoty (max.=1)
- výpočet parametrů je prováděn tzv. ***maximum likelihood approach*** (princip maximální věrohodnosti)

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

β_0 = pravděpodobnost úspěchu (y=1), když je x=0

β_1= určuje, jak rychle bude křivka stoupat k hodnotě p = 1



Jiné typy regresní analýzy III.

Nelineární regrese *(Non-linear regression)*

- jde o typ regrese, kde regresní funkce není lineární v parametrech (α, β, ε), tj. není-li jejich lineární kombinací

$$y = ab^x$$

exponenciální růst

$$y = ab^{-x}$$

exponenciální pokles

$$y = \frac{a}{1 + bc^x}$$

logistický růst

Výpočet parametrů regresní rovnice je složitý a provádí se tzv. iterační procedurou (*iteration*)

Mnohonásobná lineární regrese

(Multiple linear regression)

- studuje závislost jedné závislé proměnné na 2 a více nezávislých proměnných

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \ldots + \varepsilon_j$$

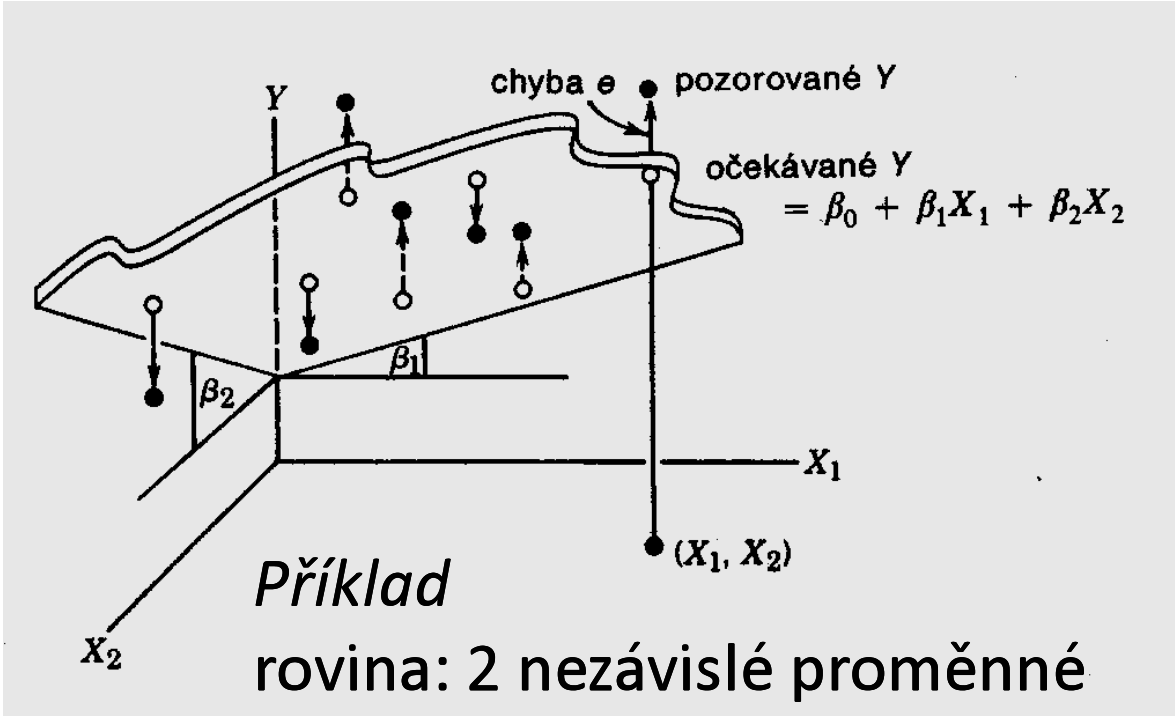
parciální regresní koeficienty ***(partial regression coefficients)***

testuje se obecná hypotéza F-testem:

$$\mathbf{H_0}: \beta_1 = \beta_2 = \ldots = \beta_m = 0$$

v případě zamítnutí H_0 se testují dílčí parciální koeficienty t-testem:

$$\mathbf{H_0}: \beta_i = 0$$



Příklad

rovina: 2 nezávislé proměnné

Jak vypadá výstup mnohonásobné regrese z PC?

Závislá proměnná (*Dependent v.*): počet pacibulek česneku domácího

Nezávislé proměnné

Regression Equation Section

Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level	Decision (5%)	Power (5%)
Intercept	12.54699	18.82904	0.6664	0.507279	Accept Ho	0.100829
výška rost.	1.017877	0.3341421	3.0462	0.003224	Reject Ho	0.852213
počet listů	-9.270854	4.083927	-2.2701	0.026155	Reject Ho	0.610345
R-Squared	0.152349					

testování parciálních koeficientů t-testem

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1	130227.8	130227.8			
Model	2	8336.867	4168.434	6.5602	0.002398	0.898402
Error	73	46385.29	635.4149			
Total	75	54722.16	729.6288			

ANOVA (F-test)

R-Squared	0.1523
Adj R-Squared	0.1291

Polynomická regrese

(Polynomial regression)

- jedná se o zvláštní typ mnohonásobné regrese

- pouze jedna proměnná *X*, ale v rovnici se vyskytují její 1-*n* mocniny

$$y=a+b_1x_i+b_2x_i^2+\ldots+b_mx_i^m+\varepsilon_i$$

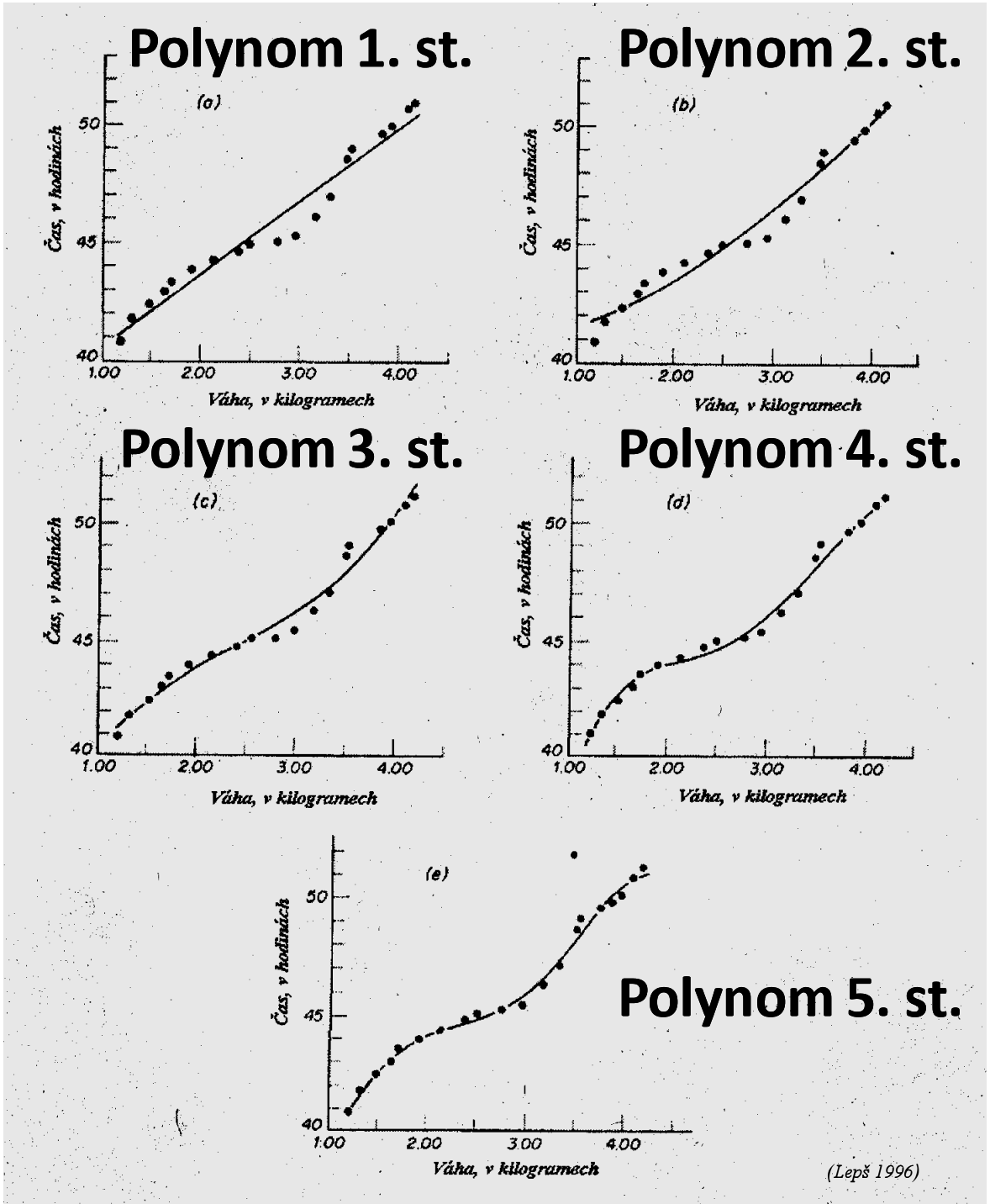
Kolik členů použít?

Testujeme postupně b_m stále vyšších mocnin na H_0: $b_m=0$ a podle výsledku určujeme počet členů rovnice

(nejužívanější je kvadratická regrese)

Polynomická regrese- příklad:

Který je nejvhodnější?



Analýza kovariance (ANCOVA)

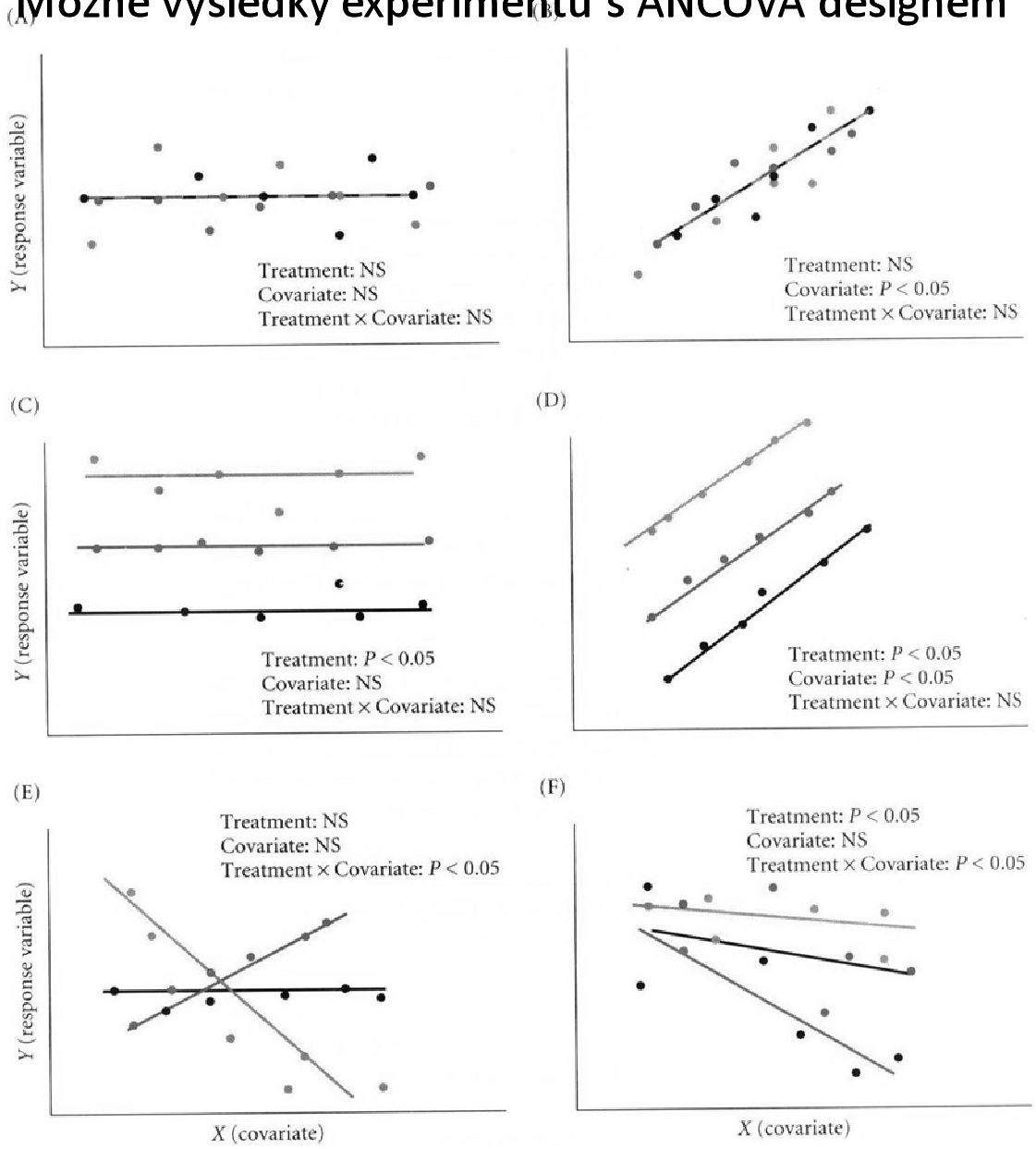
-„hybrid“ regrese a ANOVy
- užívá se v případě, kdy v analýze pomocí ANOVA zohledňuji **dodatečnou kvantitativní proměnnou** (**kovariáta**) měřenou pro každé opakování
- hypotéza je, že také kovariáta přispívá k variabilitě závisle proměnné

Model:*

$$Y_{ij} = \mu + A_i + \beta_i(x_{ij} - \bar{x}_i) + \varepsilon_{ij}$$

Pokud má kovariáta vliv, pak reziduály budou výrazně menší a test rozdílů mezi zásahy bude výrazně silnější

Možné výsledky experimentů s ANCOVA designem



*nejkomplexnější model, kdy každá hladina faktoru A má vlastní regresní přímku

(Gotelli et Ellison 2004, str. 334)

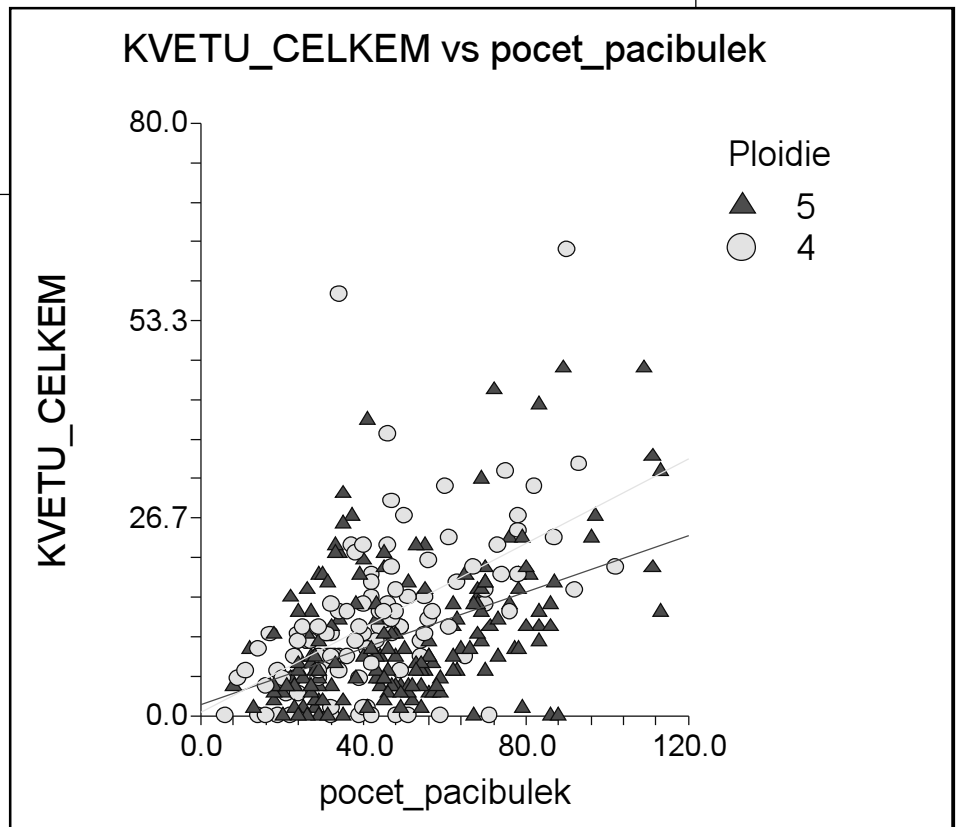
ANCOVA - příklad

<u>Př.:</u> Liší se počet květů v květenství česneku domácího mezi cytotypy (4x, 5x) s odstraněním vlivu počtu pacibulek (kovariáta)?

Analysis of Variance Table

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
X(pocet_pacibulek)	**1**	**6554.282**	**6554.282**	**86.72**	**0.000000***
A: Ploidie	1	761.285	761.285	10.07	0.001671*
S	282	21313.13	75.57848		
Total (Adjusted)	284	28147.87			
Total	285				

* Term significant at alpha = 0.05



Kovariance *(covariance)*

- **jedním číslem vyjadřuje vztah mezi dvěma (kvantitativními) proměnnými**
- její hodnoty závisí na jednotkách, ve kterých jsou měřeny proměnné (viz čitatel vzorce)
- teoreticky se pohybuje od -∞ do + ∞

$$COV = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Variance-covariance matrix: matice *n* x *n* proměnných (čtverec), kde na úhlopříčce leží variance jednotlivých proměnných, a nad a pod diagonálou leží kovariance párů proměnných x_i a x_j

Parametrická korelace (*correlation*)

- měří stupeň neboli <u>těsnost lineární</u> závislosti dvou kvantitativních proměnných

Pearsonův korelační koeficient
(correlation coefficient)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- standardizovaná kovariance

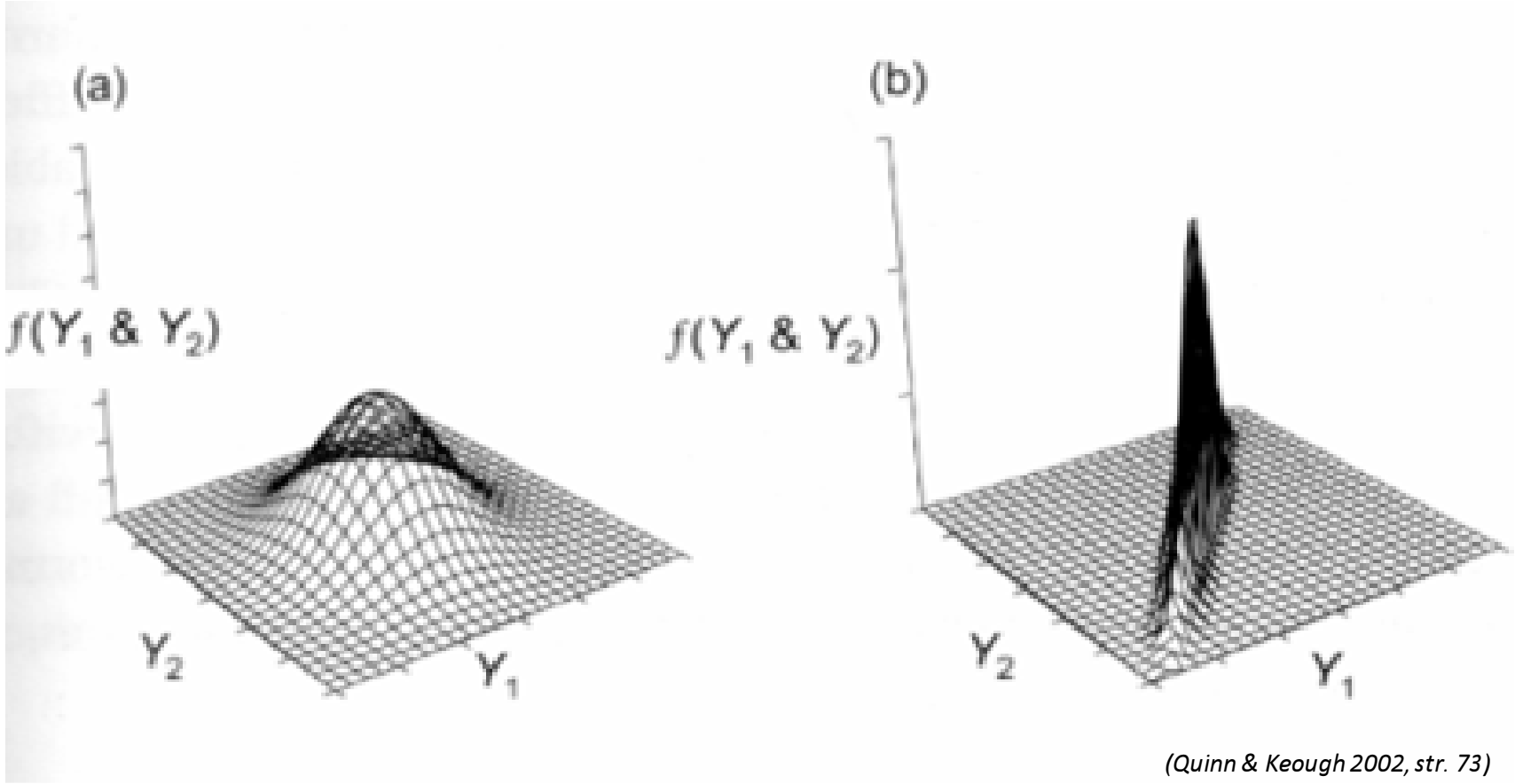
-nelze rozlišit závislou a nezávislou proměnnou (obě jsou zatíženy chybou)

-parametrický výběrový korelační koeficient + je odhadem parametru r

- je bezrozměrný, nabývá hodnot <-1;+1>

Předpoklad užití: dvourozměrná normální populace (!!!)

Dvourozměrné normální rozdělení



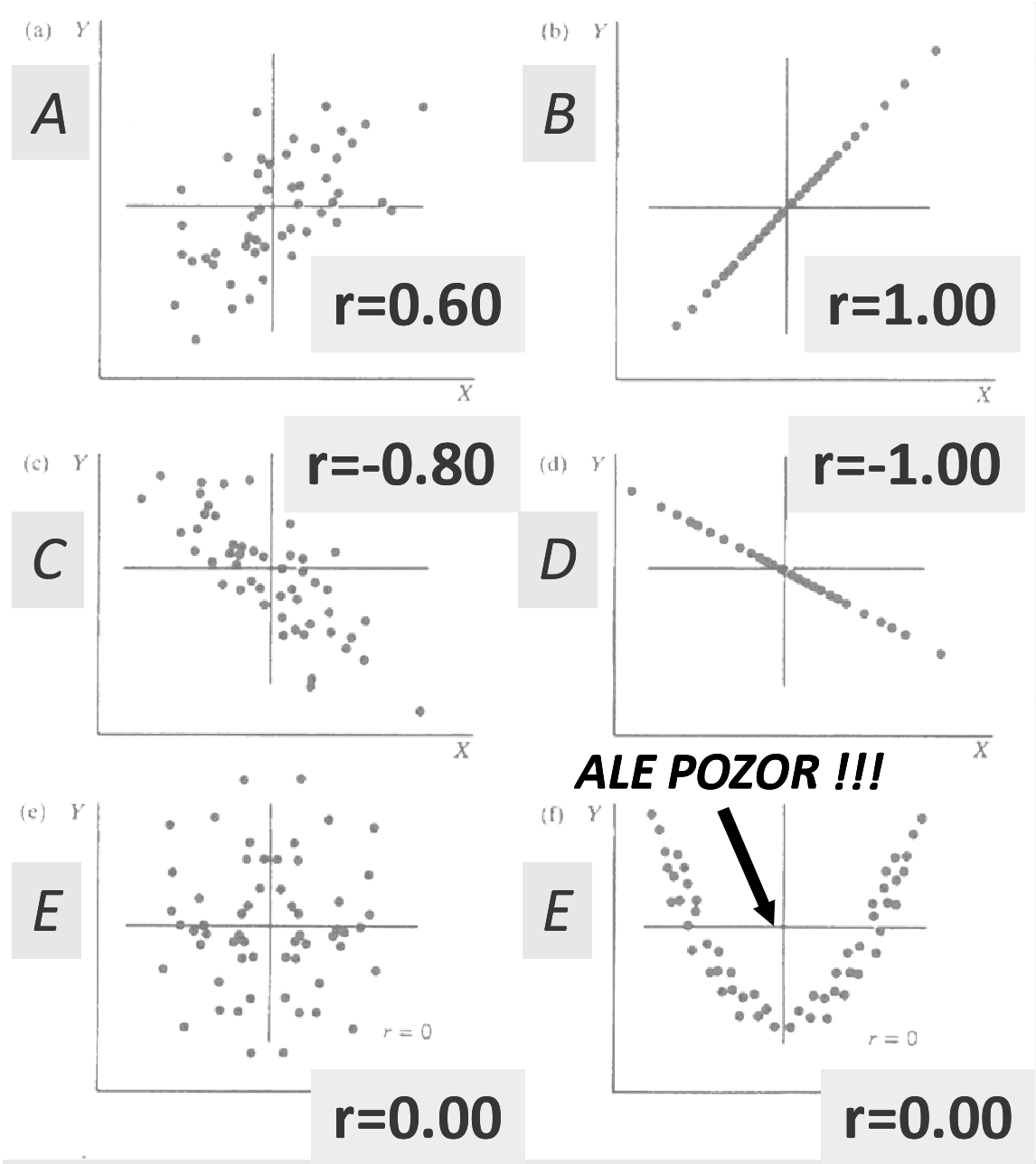
(Quinn & Keough 2002, str. 73)

Dvourozměrné normální rozdělení pro dvě proměnné Y_1 a Y_2, které jsou **slabě korelované.**

Dvourozměrné normální rozdělení pro dvě proměnné Y_1 a Y_2 které jsou **silně pozitivně korelované.**

Jaký význam má korelační koeficient?

(A) **r>0 ...pozitivní korelace**
- stoupá *X*, stoupá *Y*
(obráceně je to zcela totožné)
(*(B)* r=1 ... úplná pozitivní (deterministická) korelace)
(C) **r<0 ... negativní korelace**
- stoupá *X*, klesá *Y*
(= stoupá Y, klesá X)
(*(D)* r=-1 ... úplná negativní (deterministická) korelace)
(E) **r=0 ... nulová korelace (proměnné jsou nekorelované)**
- není lineární vztah mezi proměnnými



r je odhadem: je nutno ho testovat !

H_0: r=0
H_A: r≠0

$$t = \frac{r}{s_r}$$

kde

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

= standardní chyba korelačního koeficientu

je-li $|t| \geq t_{\alpha(2),n-2}$ **... zamítáme** H_0

A co když nemám splněnu podmínku pro užití param. korelace?

Spearmanův pořadový koeficient korelace

(*Spearman rank correlation coefficient*)

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

kde d_i^2 = rozdíl pořadí mezi x_i a y_i

-pracujeme s pořadím

- nevyžaduje dvourozměrné normální rozdělení

-interpretace jako u parametrického *r*

- testování se provádí porovnáním s kritickými hodnotami na zvolené hladině významnosti při příslušné velikosti souboru (v tabulkách)

Parciální korelace
(Parcial correlation)

-vyjadřuje těsnost závislosti dvou proměnných za předpokladu, že další proměnná (proměnné) se nemění

- umožňuje odfiltrovat interakce

-mají úzkou souvislost s parciálními regresními koeficienty