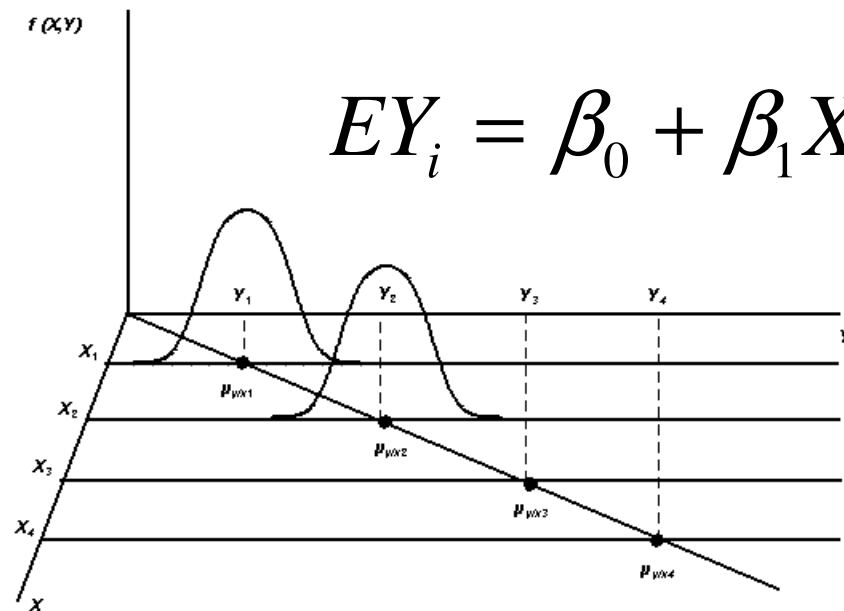

Lineární regresní model

Jedna a více nezávisle proměnných

Lineární regresní model s jednou nezávisle proměnnou (*Simple regression*)

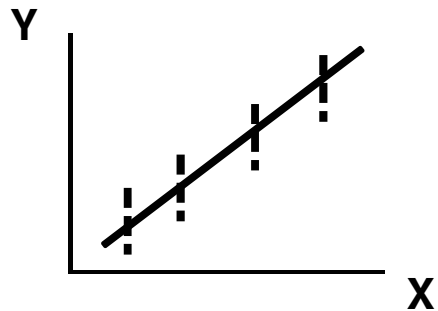
- n objektů
 - Pro každý objekt: pozorované veličiny X a Y - spojité
 - Pozorování, objekty – navzájem nezávislá
 - Zajímá nás závislost veličiny Y na X – POZOR! –
nutná podmínka je, že závislost je stejná pro všechny zkoumané objekty.
 - Příklad: V egyptské vesnici Kalama se studoval vliv výživy na 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.?



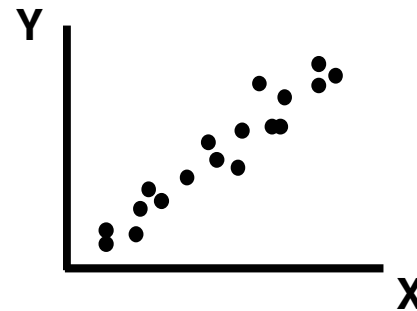
$$EY_i = \beta_0 + \beta_1 X_i, i = 1, \dots, 12.$$

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



Designed experiments

Máme kontrolu na X

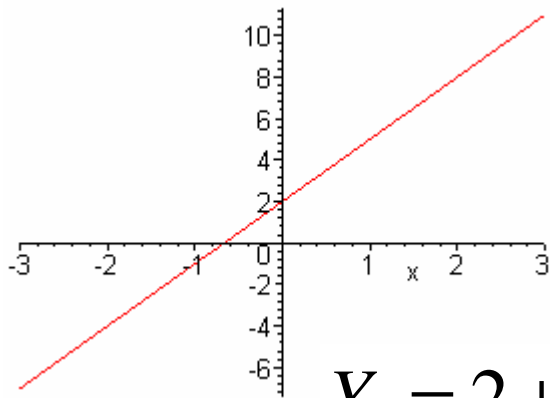


Observational experiments

Opakování z gymnázia – analytická geometrie

- Analytické vyjádření
přímky, rovnice

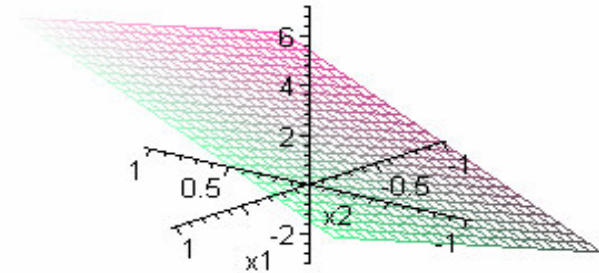
$$Y = \beta_0 + \beta_1 X$$



$$Y = 2 + 3X$$

- Analytické vyjádření
roviny, rovnice

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



$$Y = 2 + 3X_1 + 2X_2$$

Nejjednodušší typ závislosti - lineární

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



- Systematická část modelu



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

Oběcně hledáme takové beta, tak aby systematická část modelu vysvětlovala co nejvíce Y (závisle proměnnou).

Regresní rovnice - proměnné

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



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



- **Nezávisle proměnná (*Independent variable*)**
- **Vysvětlující proměnná (*Explanatory variable*)**
- **Kovariáta (*covariate*)**
- **Prediktor (*predictor*)**
- **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

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



- Průsečík s osou Y

- *Intercept*

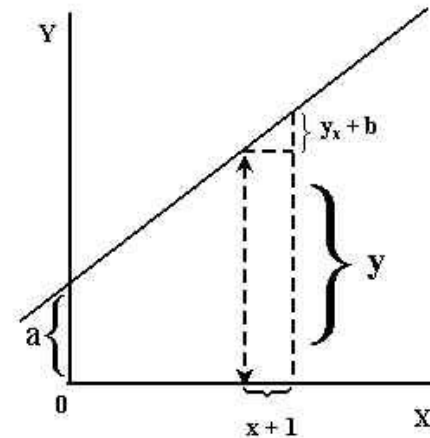


- Směrnice

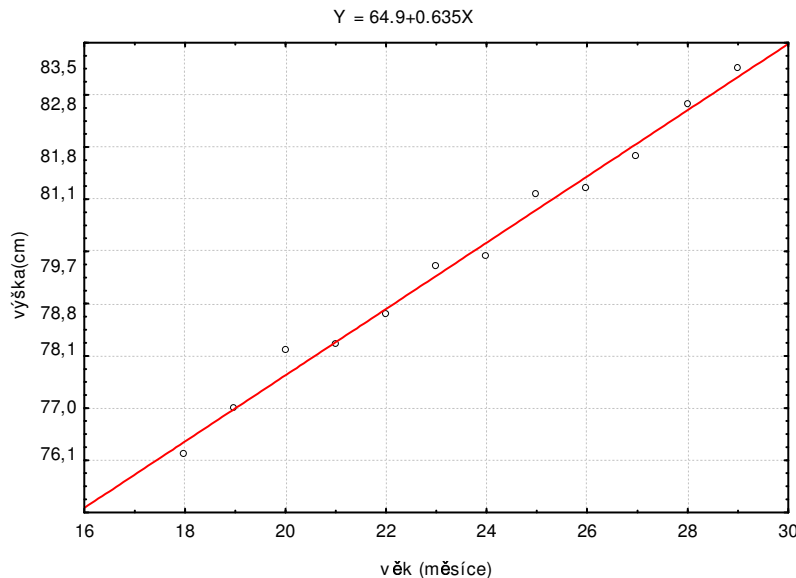
Interpretace parametrů:

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

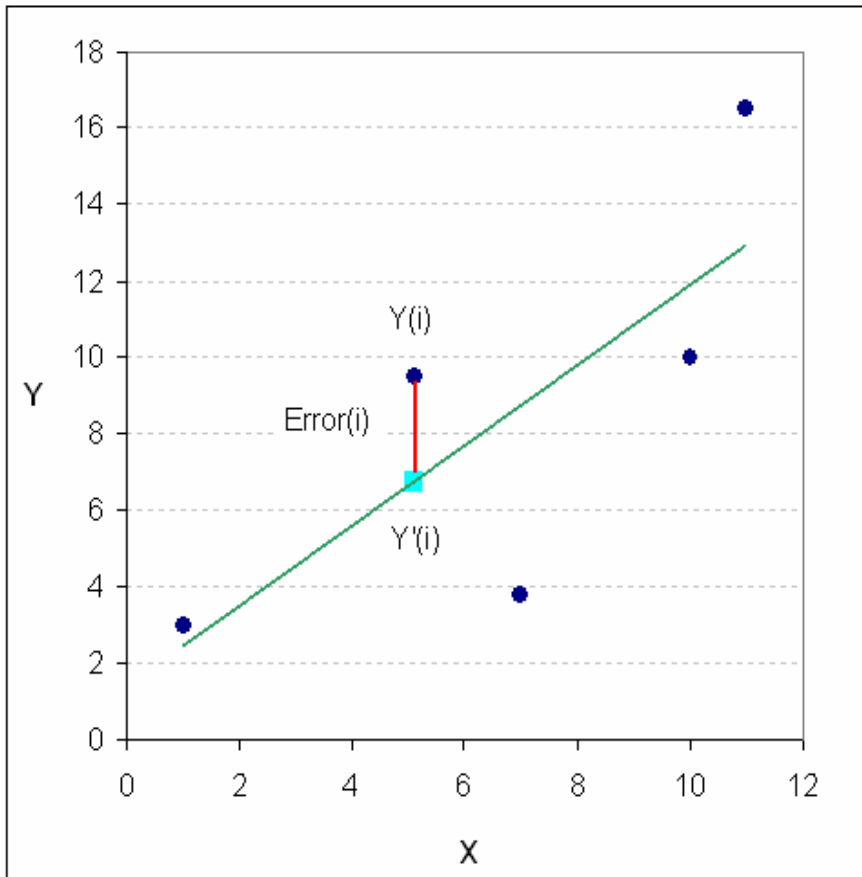


- Lineární závislost – přímka
- $\text{Height} = 64.9 + 0.635 \text{Age}$
- **Průsečík: 64,9**
- Interpretace ad absurdum: výška dítěte ve věku 0 měsíců (tj. při porodu). *Ale to by byla **extrapolace**, tedy rozšíření modelu na oblast, kde jsme data neměřili.*
- **Směrnice: 0,635**
- Dítě starší o jeden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

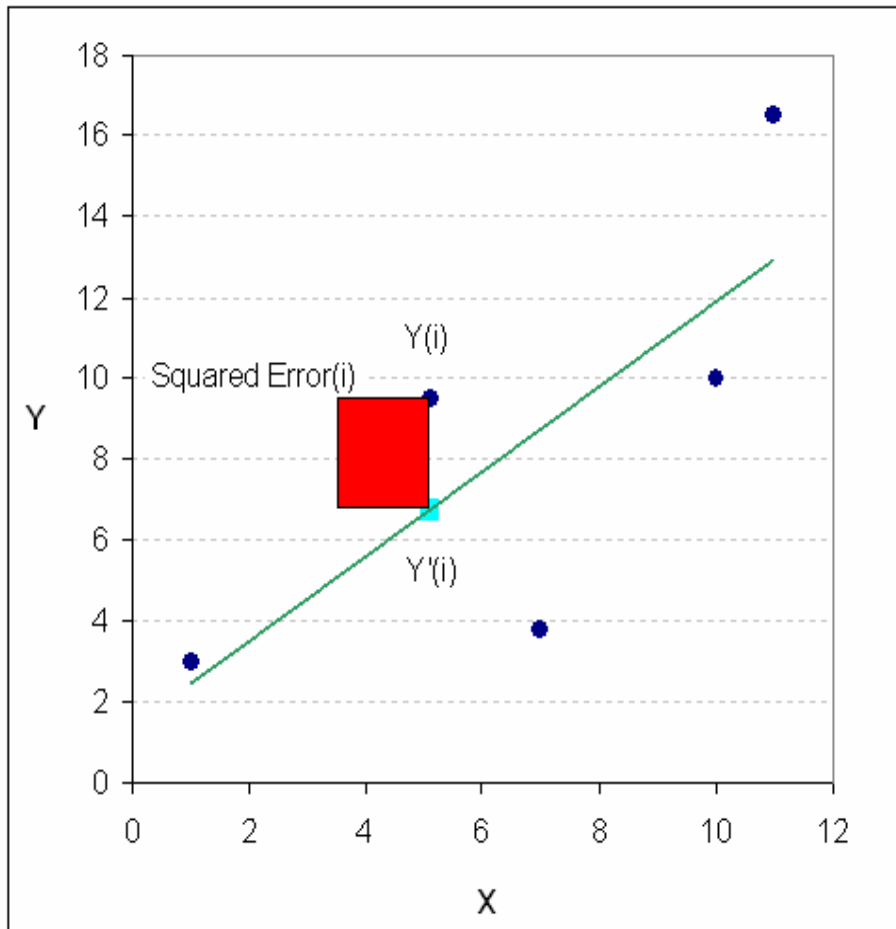
$$\text{Residuum}_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \beta_0 - \beta_1 X_i$$

Pozn.: Residuální součet čtverců

- Výsledný minimální součet čtverců residuí (pro b_0 a b_1) 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/COURSE/S/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 β_0 a β_1 součtu čtverců reziduí.

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

- Výsledné hodnoty β_0 a β_1 pro které je součet čtverců minimální označujeme b_0 a b_1
- Odhadnutá regresní rovnice má tvar

$$Y = b_0 + b_1 X$$

- Maticový zápis

$$b = \hat{\beta} = (X^T X)^{-1} X^T y$$

Gauss-Markovův teorém

- Odhad získaný metodou nejmenších čtverců je nejlepší lineární, nestranný odhad.
 - Ale existují situace, kdy vhodný není:
 - Chyby jsou korelované, nemají stejný rozptyl -> GLM
 - Rozdělení chyb má těžké chvosty -> robustní odhady
 - Prediktory jsou vysoce korelované (kolinearita) -> hřebenová regrese (*ridge regression*)
-

Vzorce pro odhad parametrů regresní přímky – metoda nejmenších čtverců

Odhad b je zatížený chybou:

$$\text{I. } b \sim \beta : b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$
$$S_b^2 \sim \sigma_\beta^2 : \frac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$$
$$S_{y \cdot x}^2 = \frac{\sum d_{y \cdot x}^2}{n-2} = \frac{\sum Y_i^2 - \frac{\sum Y_i^2}{n} - b^2 \cdot \sum (X_i - \bar{X})^2}{n-2}$$

$$\text{II. } a \sim \alpha : a = \bar{Y} - b \cdot \bar{X}$$

intercept

$$S_a^2 \sim \sigma_\alpha^2$$
$$S_\alpha^2 = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum X^2} \right] \cdot S_{y \cdot x}^2$$

$$\text{III. } \hat{Y} : \text{modelová hodnota}$$

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

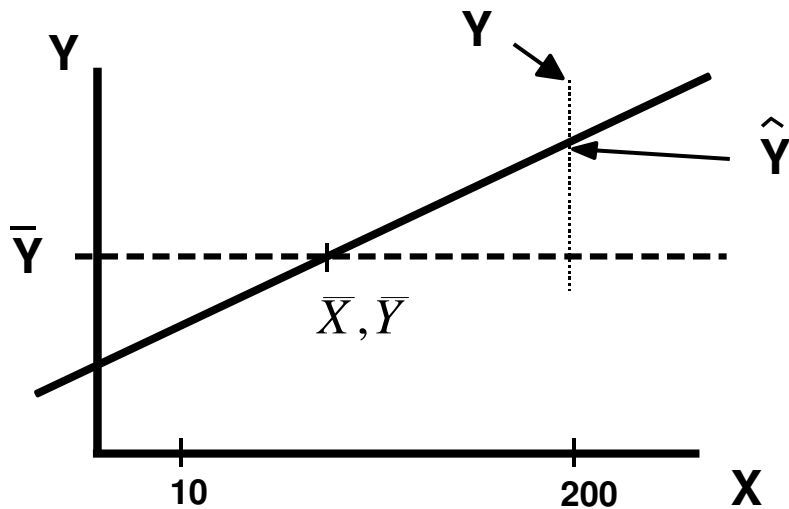
Příklad: Spalování odpadu

X: Množství spáleného odpadu (tuny)

Y: Koncentrace kovu ve vzduchu(ng/m³)

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

Model: $Y = a + b \cdot X$



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



Např. : Skutečná data pro X = 200 t:

$Y_i = 16; 25; 41; 28; 31; 20 \Rightarrow Y_i = 26.8$

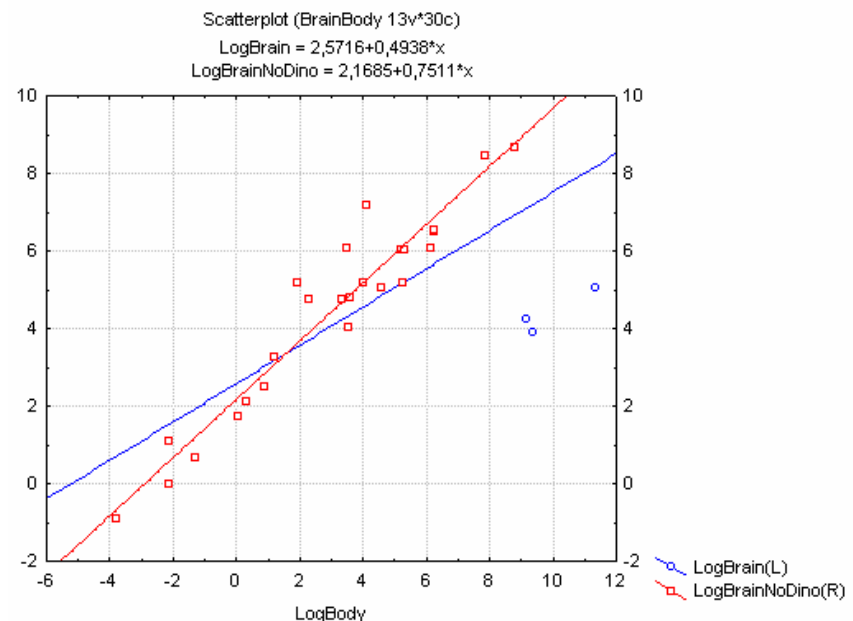
Odhadnuto z modelu pro X = 200 t:

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

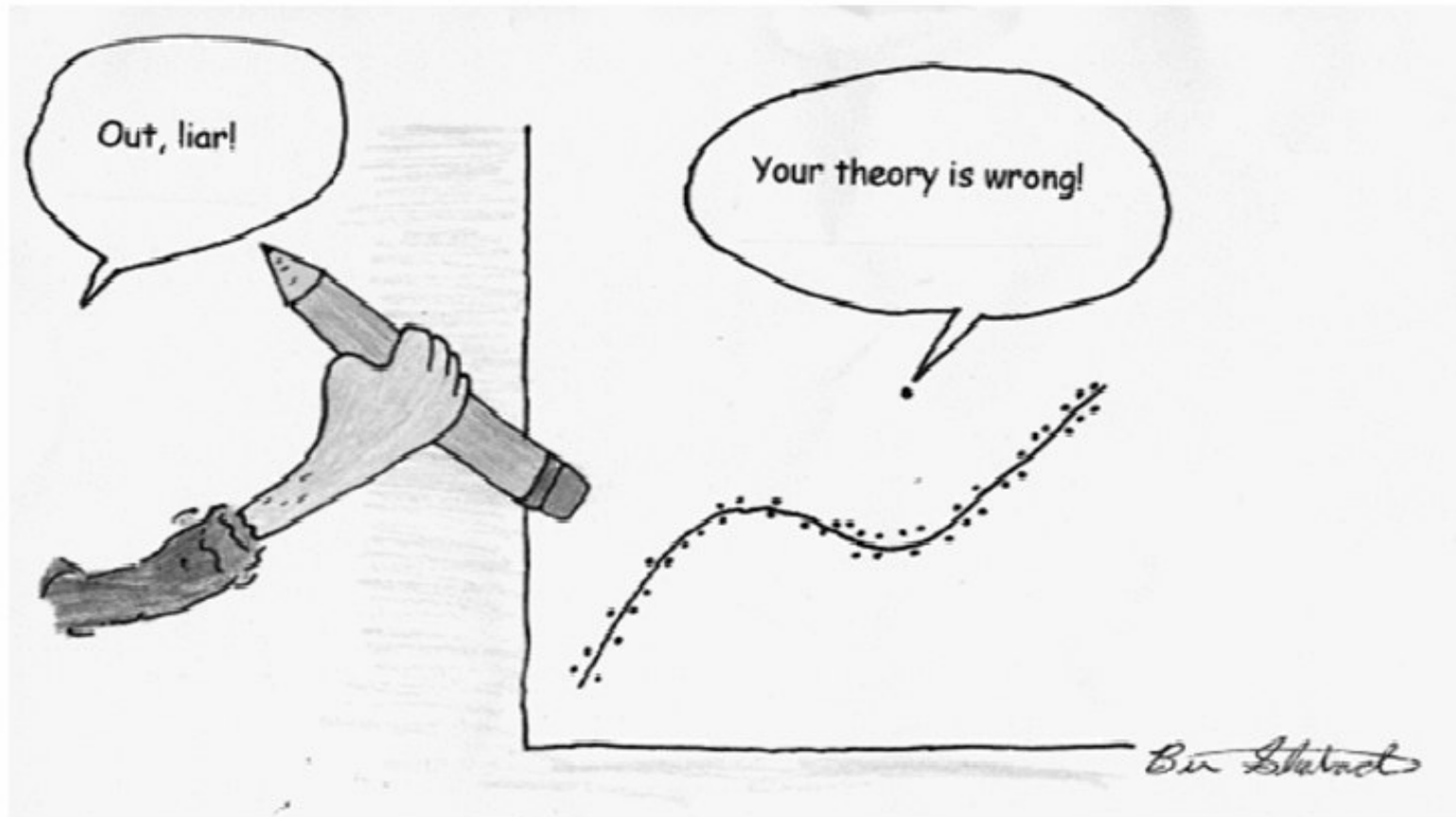
$$\left. \begin{aligned} \hat{Y} &= \bar{Y} + b \cdot (X - \bar{X}) \\ \hat{Y} &= a + b \cdot X \end{aligned} \right\} a = \bar{Y} - b \cdot \bar{X}$$

Odlehlá pozorování - Nebezpečí (*outliers*)

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



Outliers ([http://botany.upol.cz/prezentace/duch\(soubor statistika4.pdf\)](http://botany.upol.cz/prezentace/duch(soubor%20statistika4.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

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

$$\mathbf{x}$$

$$\begin{bmatrix} | \\ | \\ | \\ | \\ | \end{bmatrix}$$

$$\bar{\mathbf{x}}$$

$$\mathbf{y}$$

$$\begin{bmatrix} | \\ | \\ | \\ | \\ | \end{bmatrix}$$

$$\bar{\mathbf{y}}$$

$$s_y^2$$

$$\mathbf{y}$$

$$\begin{bmatrix} | \\ | \\ | \\ | \\ | \end{bmatrix}$$

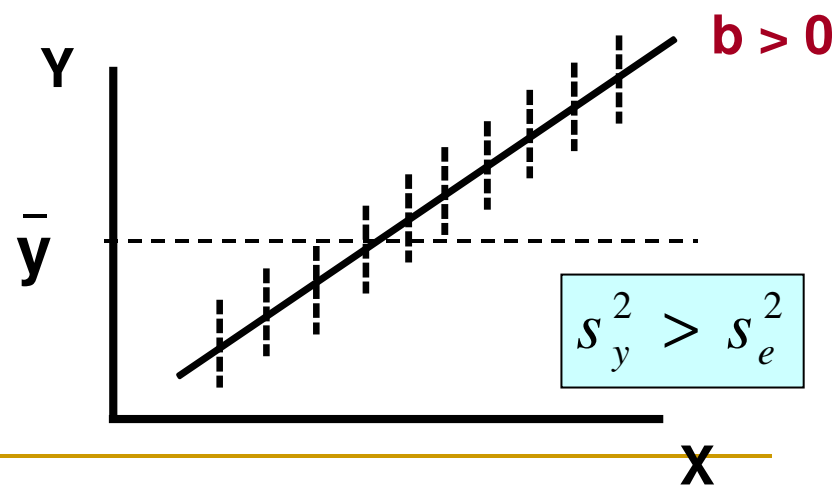
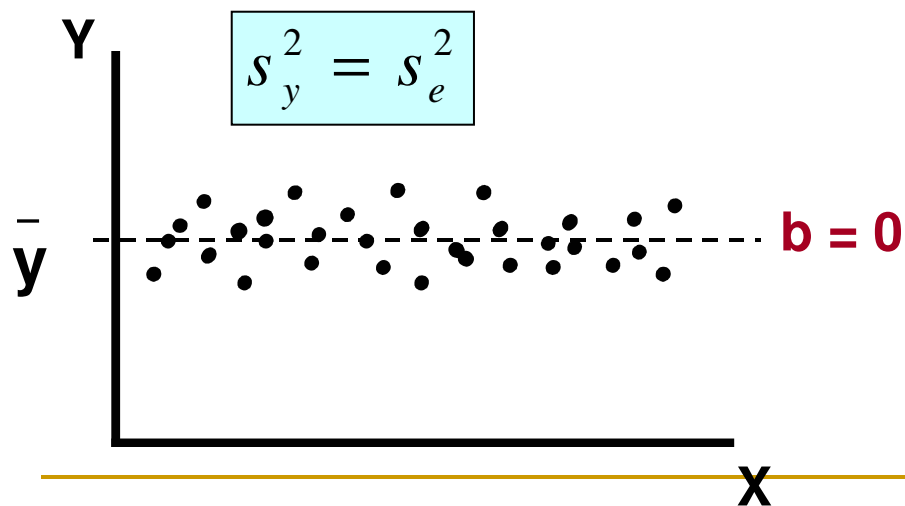
$$\hat{\bar{\mathbf{y}}}$$

$$\mathbf{e}$$

$$\begin{bmatrix} | \\ | \\ | \\ | \\ | \end{bmatrix}$$

$$\bar{\mathbf{e}} = 0$$

$$s_e^2$$



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
- a zamítáme ji v případě, že

$$T = \frac{b_1}{S.E.(b_1)}$$

$$|T| \geq t_{n-2}(1 - \alpha/2),$$

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: $Y = a + b \cdot x \longrightarrow Y = 0,11 + 0,092 \cdot X$

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

$$\text{I. } \left. \begin{array}{l} H_0 : \beta = 0; \alpha = 0,05 \\ b = 0,092 ; s_b = 0,023 \end{array} \right\} t = \frac{b}{S_b} = 4,00$$

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

P < 0,01

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

$$\text{II. } \left. \begin{array}{l} H_0 : \alpha = 0; \alpha = 0,05 \\ a = 0,11; s_a = 0,029 \end{array} \right\} t = \frac{a}{S_a} = 3,793$$

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

P < 0,05

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

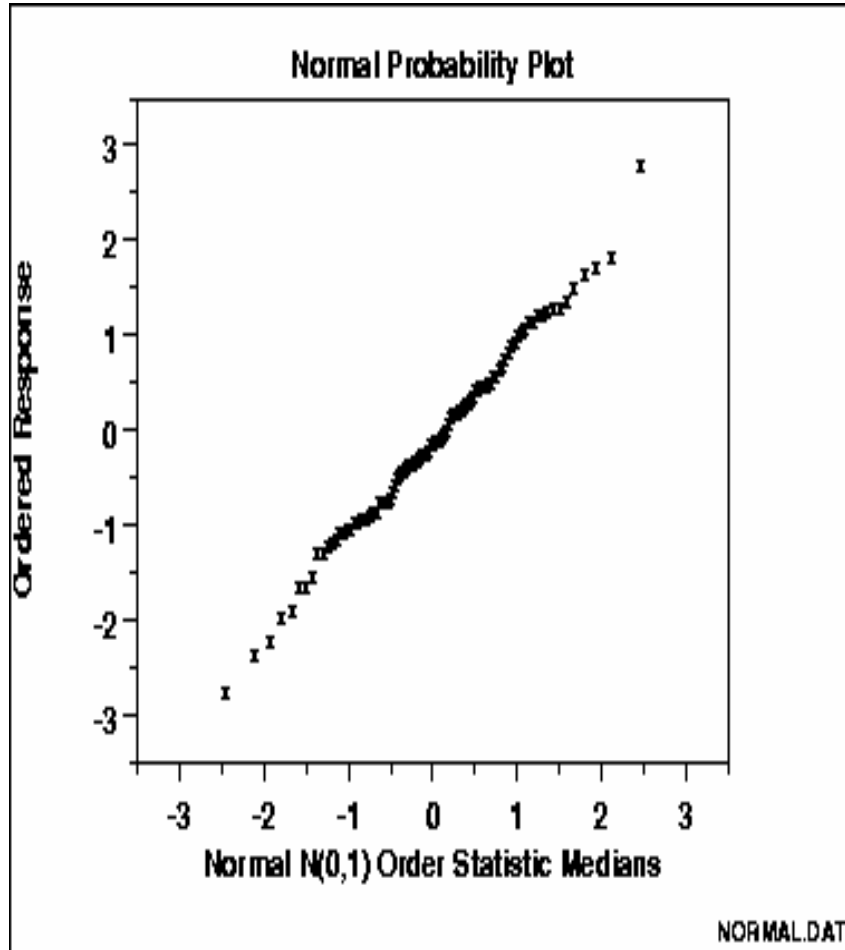
$$P(0,049 \leq \alpha \leq 0,171) = 0,95$$

Předpoklady

- Nutný předpoklad potřebný ke všem testům spojeným s regresním modelem je **normalita residuů**.
 - Residua mají mít normální rozdělení s **nulovou střední hodnotou a konstantním rozptylem σ^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).

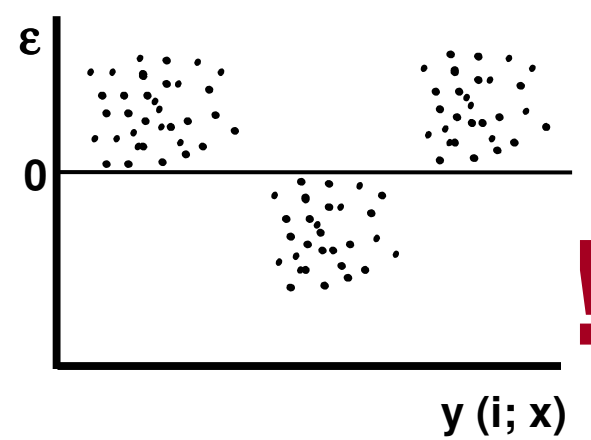
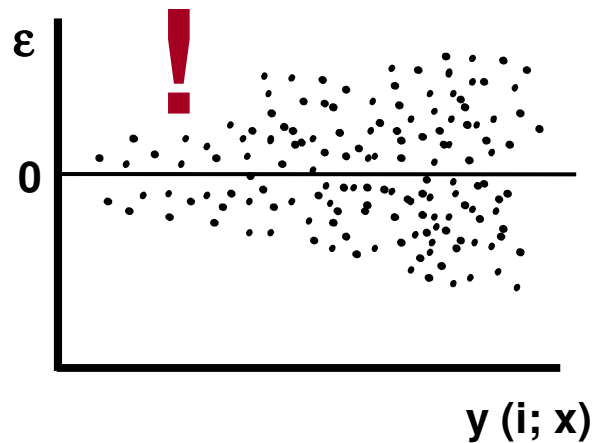
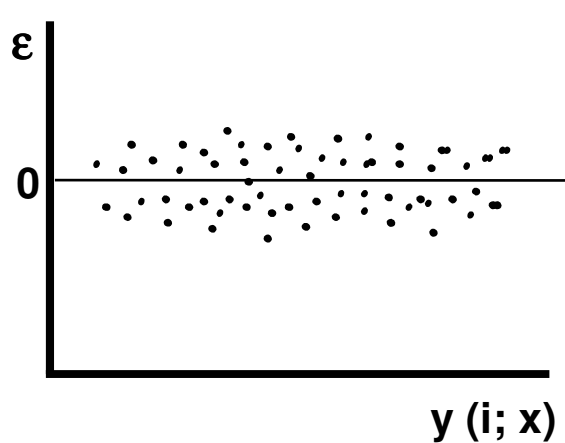
- Autokorelace residuí
 1. Durbin-Watsonův test
-

Diagnostika residuí

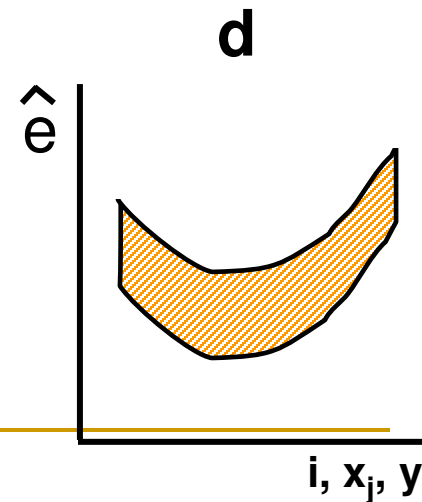
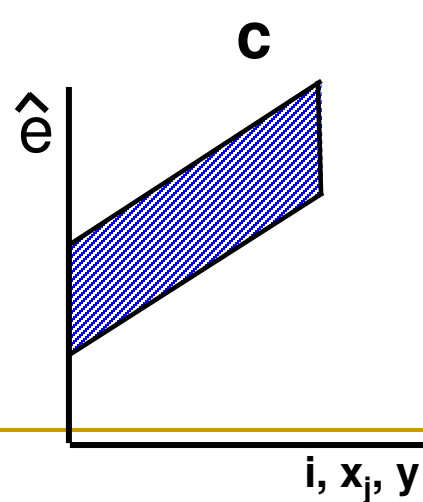
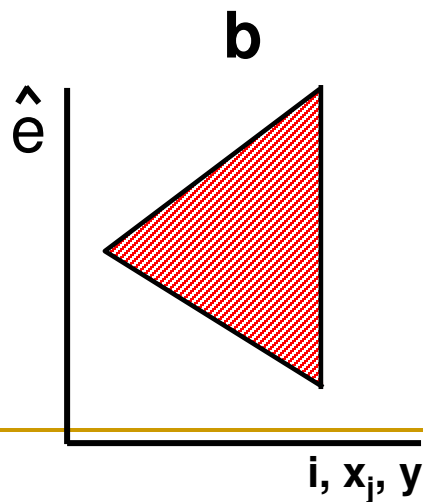
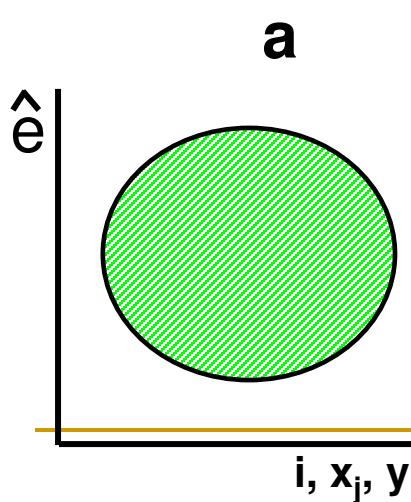
- Je námi zvolená závislost (lineární) vhodná?
 - Pomoc grafické znázornění – **grafy závislosti hodnot residuí na hodnotách \hat{y}_i nebo x_i**
 - V případě, že zvolený tvar závislosti byl vhodný, jsou residua
 1. umístěna náhodně kolem nulové střední hodnoty
 2. nevykazují žádný systematický trend
 3. jejich rozptyl je homogenní
-

Diagnostika residuí

3) Grafy residuí modelů (příklady)

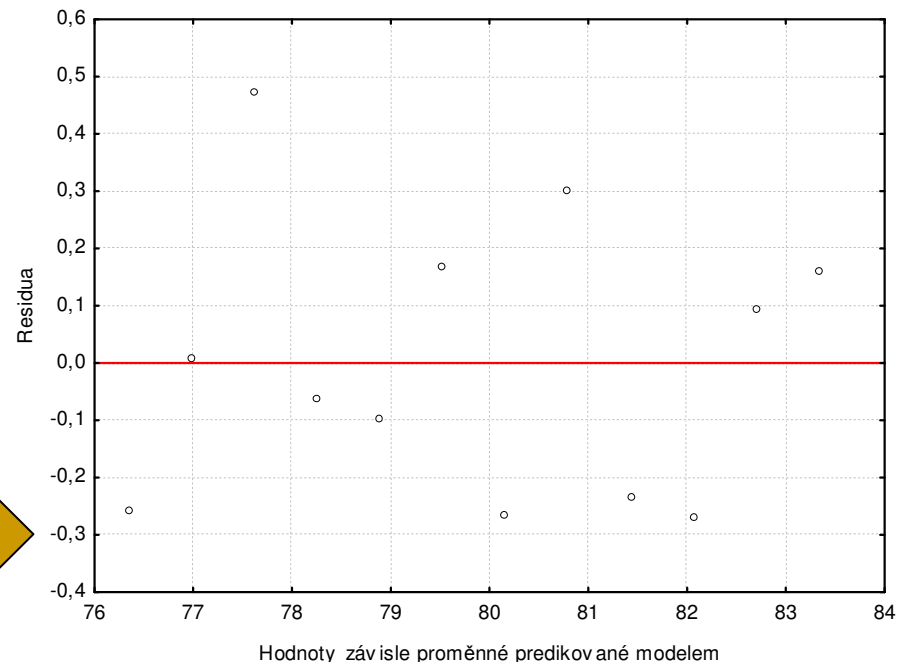
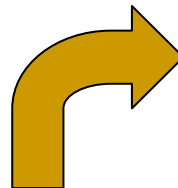


Obecné tvary residuí modelů (schéma)



Diagnostika residuí - obrázky

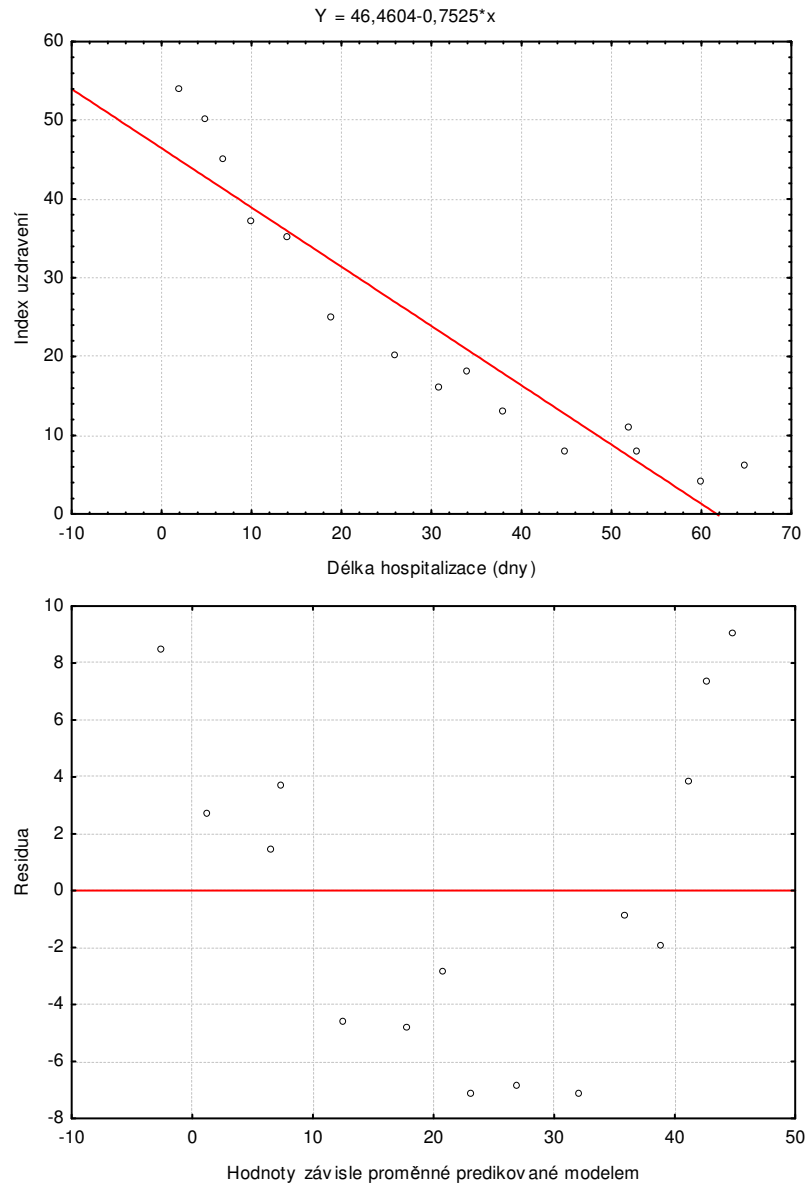
Příklad Kalama: Hodnota testové statistiky $T=29.66$, Nulovou hypotézu zamítáme na hladině 0,05 (p-hodnota =0,00). Výška dětí závisí na jejich věku. Koeficient determinace je $R^2 = 0,98$



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“:

... , $1/x^2$, $1/x$, $1/\sqrt{x}$, $\log x$, \sqrt{x} , x , x^2 , ...

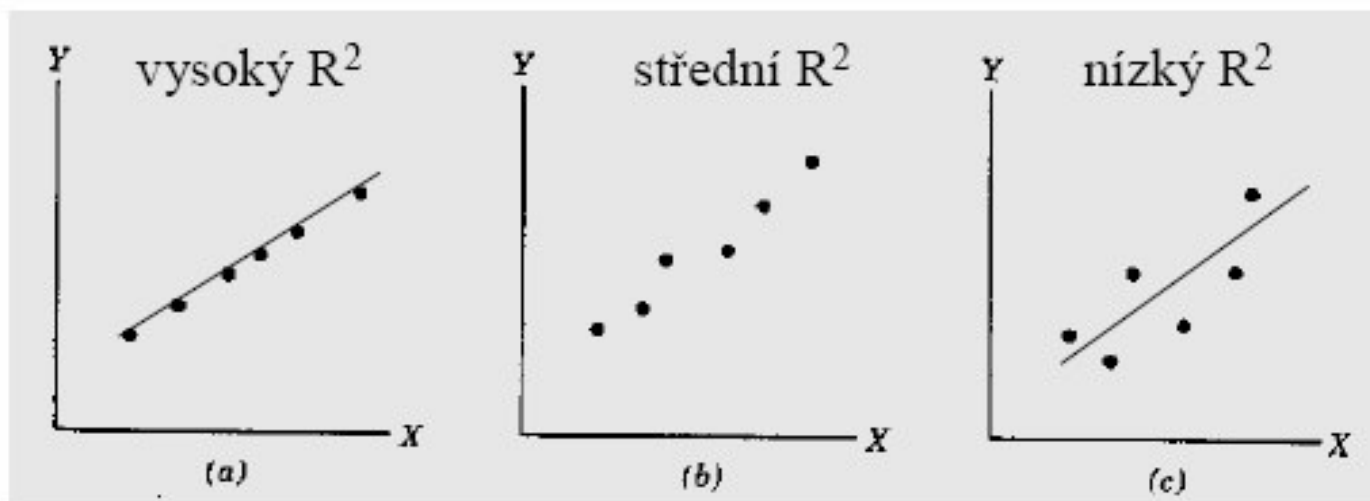
- Po tomto žebříku transformací se můžeme pohybovat buď nahoru (k vyšším mocninám) nebo dolů. Cílem je především linearizace závislosti.
- Když dosáhneme pohybem po zvoleném žebříku (na ose x nebo ose y) přibližně lineární závislosti, potom současným pohybem po obou žebřících se pokusíme také o stabilizaci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

$$R^2 = \frac{\text{variabilita vysvětlena modelem}}{\text{celková variabilita } Y} = 1 - \frac{\text{residuální variabilita}}{\text{celková variabilita } Y} = 1 - \frac{S_e}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$



(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 .

$$R^2 = \text{corr}(X, Y)^2$$

Nelineární regresní model

Exponenciální závislost

- Obecný tvar exponenciální závislosti je

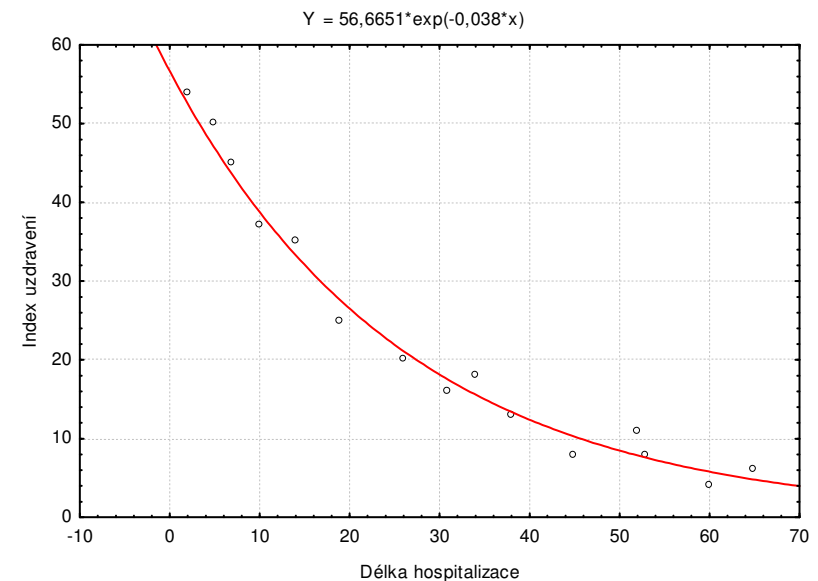
$$Y = \beta_0 + \exp(\beta_1 + \beta_2 X)$$

- Je-li parametr β_2 kladný, pak s rostoucími hodnotami X rostou i hodnoty Y . Je-li parametr β_2 záporný, pak s rostoucími hodnotami X klesají hodnoty Y . Parametr β_2 charakterizuje strmost nárustu resp. poklesu, parametry β_0 a β_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 blíží nule.
-

Příklad: Index uzdravení

Exponenciální závislost

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



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 zrněk z pole předchozího. Kolik bude celkem zrněk rýže na šachovnici?
-

Nelineární regresní model

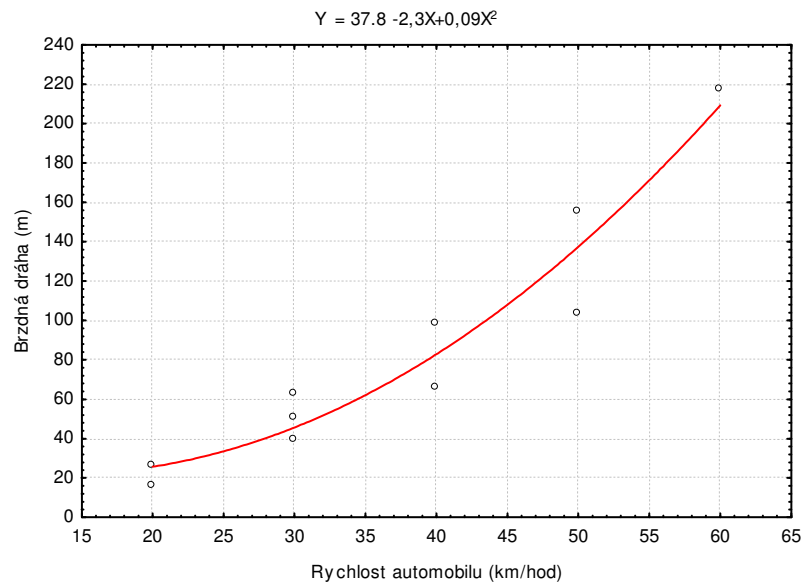
Polynomiální závislost

- 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.



Křivky dávka-odpověď (Dose response curves)

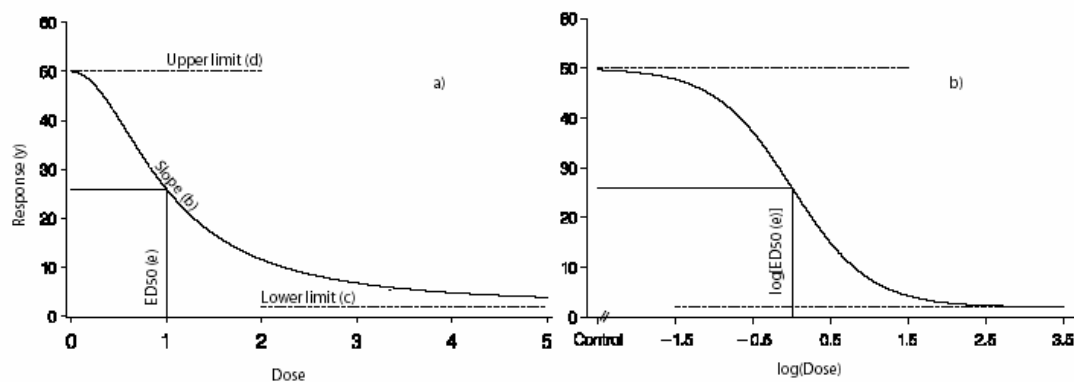
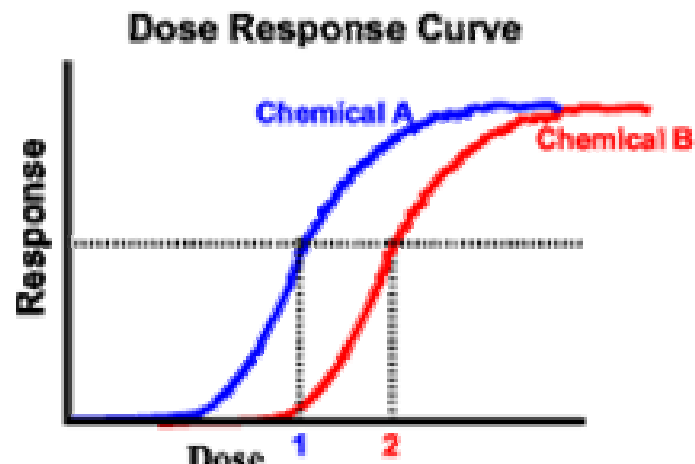
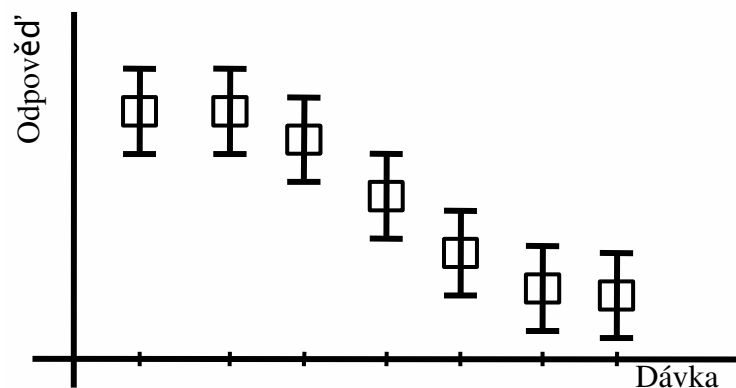


Figure 2: The logistic dose-response curve a) on non-logarithmic dose scale; b) on logarithmic dose scale. The parameters in this example are: $b=2$, $c=2$, $d=50$, $e=1$.

Více nezávisle proměnných

(Multiple regression model)

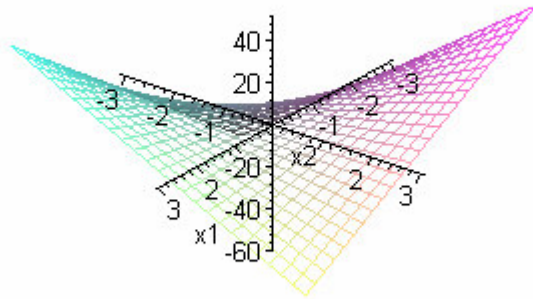
- Dvě nezávisle proměnné:
- Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
- Koeficient β_1 lze interpretovat jako střední změnu Y při jednotkové změně X_1 a nezměněné hodnotě X_2 .
- Nulová hypotéza $H_0 : \beta_1 = 0$ znamená, že populační průměr Y závisí nejvýše na X_2 .
- Tj. platí, že $Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$
- Další interpretace $H_0 : \beta_1 = 0$ je, že proměnná X_1 nepřináší žádnou informaci o střední hodnotě Y nad tu, která je již obsažena v X_2 .
- Snaha o co nejjednodušší model, obsahující jenom významné prediktory (nezávisle proměnné)

Regresní plocha

(Response surface, regression surface)

- Model s interakcemi

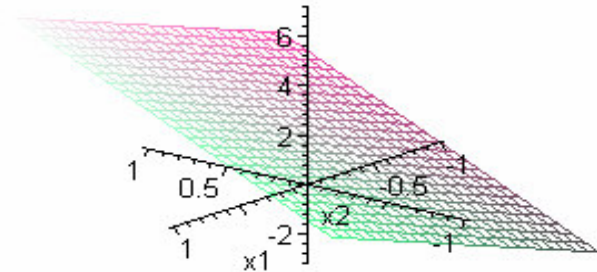
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$$



$$Y = 2 + 3X_1 + 2X_2 - 5X_1X_2$$

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



$$Y = 2 + 3X_1 + 2X_2$$

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.
-

F-test

- Větší model, počet parametrů $p \geq \Omega$
- Menší model, počet parametrů $q < p$ ω
- n počet pozorování
- Intuitivně: je-li jmenovatel následujícího výrazu malý, pak menší model je skoro stejně dobrý jako větší model

$$\frac{RSS_{\omega} - RSS_{\Omega}}{RSS_{\Omega}}$$

- Testová statistika

$$F = \frac{(RSS_{\omega} - RSS_{\Omega})/(p - q)}{RSS_{\Omega}/(n - p)} \approx F_{p-q, n-p}$$

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 pomocí 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 najít 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