Regresní analýza

Úlohy a otázky

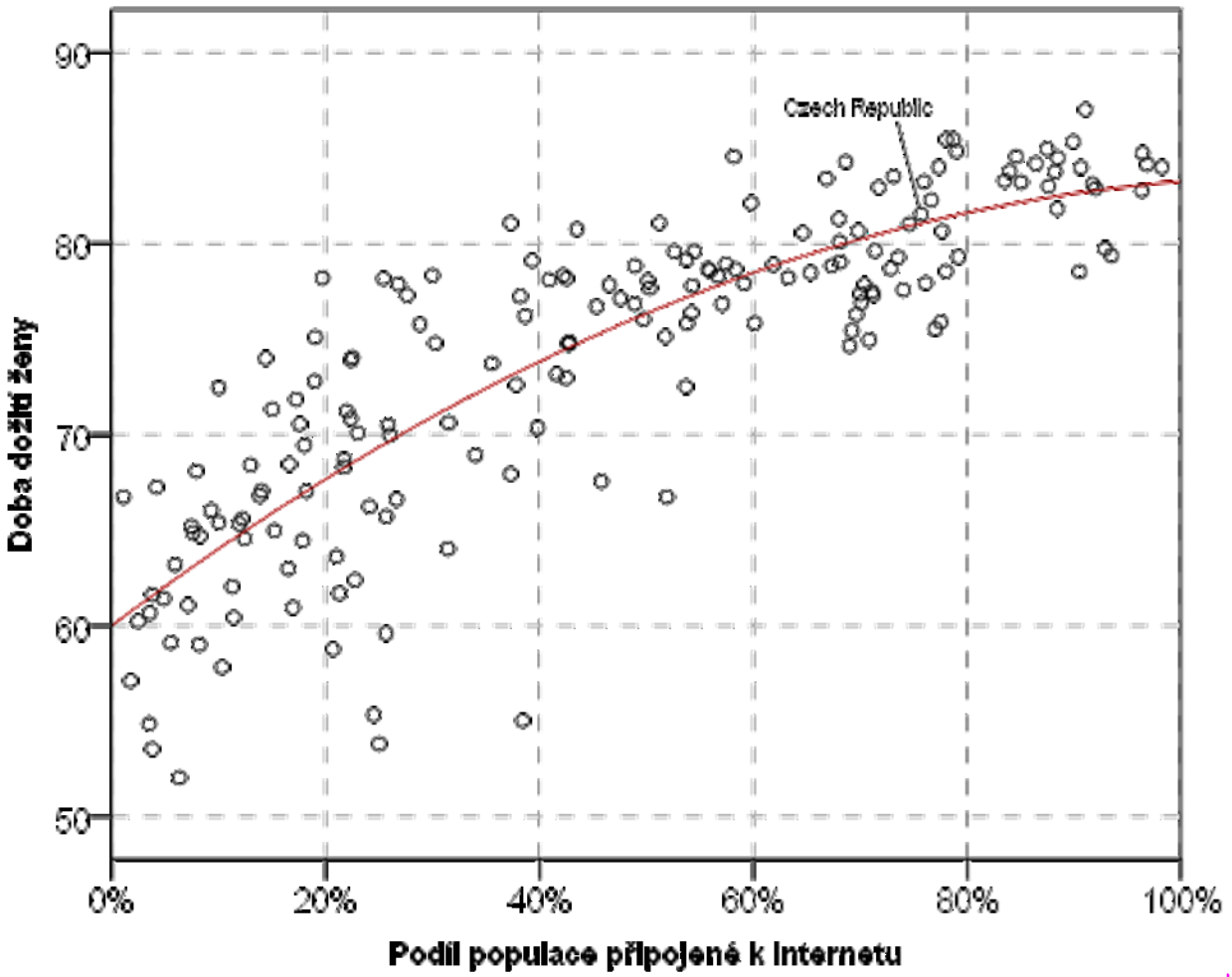
- Lze vyjádřit vztah proměnné X (nebo množiny proměnných $X_1, \ldots, X_k$) a proměnné Y pomocí vhodně volené rovnice?
- Má tento vztah explanační charakter?

- hodnota Y důsledkem hodnoty X (hodnot $X_1, \ldots, X_k$)?
- Reprezentují proměnné X_k příčiny pro důsledek Y?

- Obsahuje X (nebo množina proměnných $X_1, \ldots, X_k$) nějakou informaci o Y a jak vyjádřit přenos takové informace?
- Můžeme tuto informaci použít pro predikci?

Internet prodlužuje život

- nesmyslná interpretace
- jevy spolu nesouvisí odvozeným způsoben
- společná příčina – obecný rozvoj země
- nesmyslnost odhalena jen na základě logické úvahy



Regresní analýza: jednosměrný vztah

Směr vztahu je volba uživatelská volba

nezávislá proměnná –> závislá proměnná <– chyba

- **široká terminologie**
  - nezávislá – závislá
  - vysvětlující – vysvětlovaná
  - vstupní – výstupní (cílová)
  - prediktor – predikant
  - určující – určená

u korelace: jednosměrný nebo symetrický vztah

u regrese: jednosměrný vztah

Popis vztahu rovnicí

model – rovnice závisející na neznámých koeficientech

model vztahu

$$Y = f(X) + \varepsilon$$

- **chyba rovnice**
  - zahrnuje neznámé vlivy na Y
  - náhodné číslo s průměrem 0 a rozptylem σ_ε^2

směr vztahu

proměnných X může být více

Model (rovnici) volí uživatel. Regresní metoda určuje koeficienty (parametry) rovnice.

rovnice rozloží hodnotu Y na dvě části:

a) **model = převod z X**

b) **náhodná chyba/zbytek, který se neúčastní převodu**

Chyba rovnice

- **chování *Y* nevysvětlené modelem**
- **náhodné vlivy**
  - při měření, zjišťování
  - při chování
  - ze své podstaty nelze v modelu odstranit
- **tvar funkce**
  - závislost je tvořena jinou funkcí (se stejnými proměnnými)
  - např. logaritmická nebo kvadratická funkce místo přímky
  - špatný tvar lze teoreticky zjistit a opravit
- **neznámé vlivy**
  - všechny další veličiny, které mají vliv na Y
  - nemáme v datech
  - teoreticky lze odstranit zjištěním chybějících proměnných a jejich doplněním do modelu
- **mezi náhodnými a neznámými vlivy nezle v praxi rozhodnout => v regresní teorii se vše zahrnuje pod náhodnou odchylku**

Náhodná chyba

- **má nulový průměr pro každou hodnotu *X***
  - zajistí vhodná funkce odpovídající *skutečnému modelu* – obvykle neznámý
  - nulový průměr odhadu chyb na celém souboru zajistí vždy konstantní člen v modelu β_0
- **má stejný rozptyl pro každou hodnotu *X* – kontroluje se v datech**

Konstantní člen vždy do modelu zahrňte !

Model přímky – obvyklá volba

rovnice přímky

chyba

rovnice datového procesu

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

predikční rovnice

$$\hat{Y} = \beta_0 + \beta_1 X$$

koeficienty

očekávaná hodnota Y pro dané X

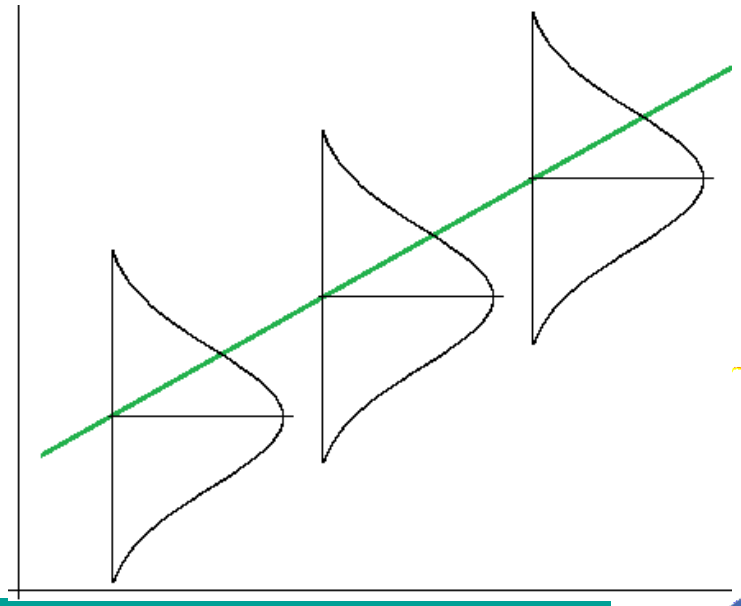
$$\hat{Y} = E(Y|X)$$

očekáváno pro dané X

není obsaženo v X

$$Y = \hat{Y} + \varepsilon$$

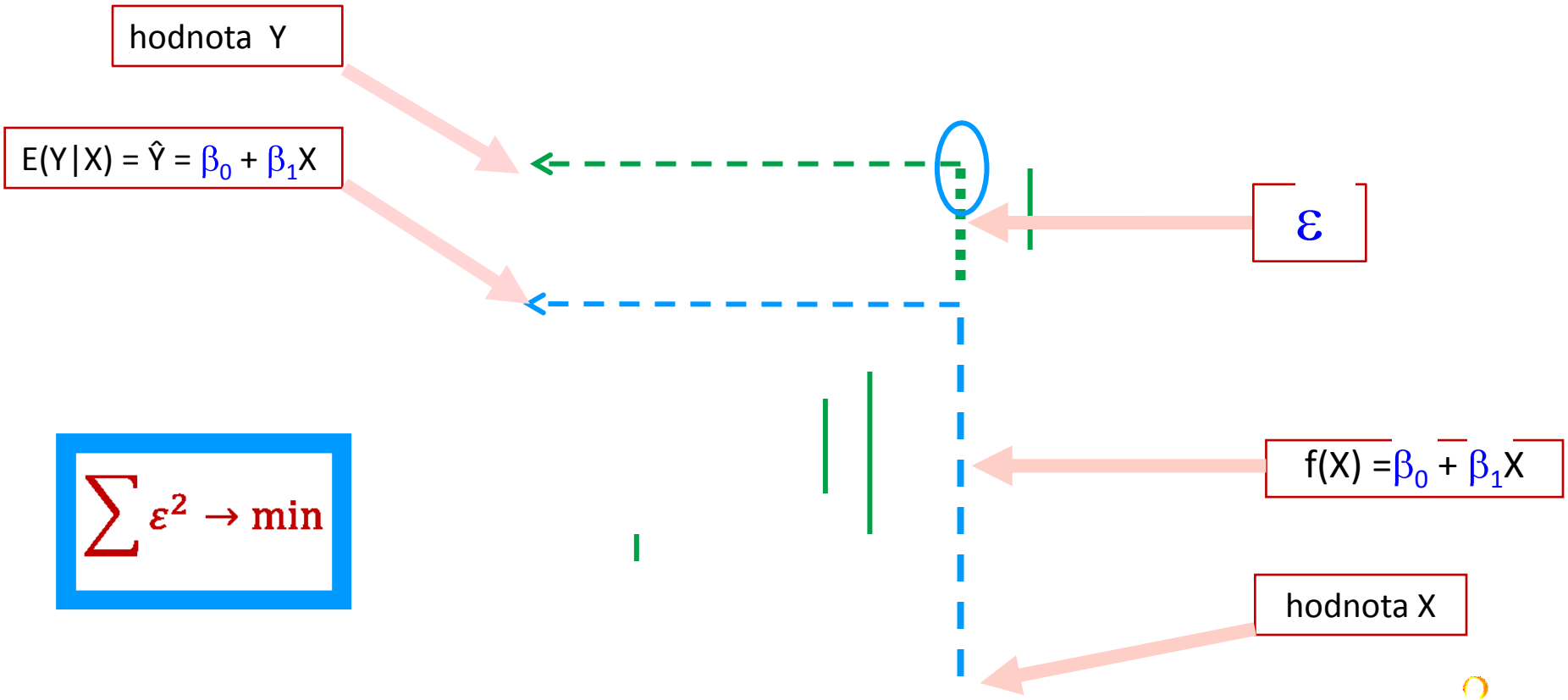
očekáváme, že chyba je v průměru nulová



Význam koeficientů přímky

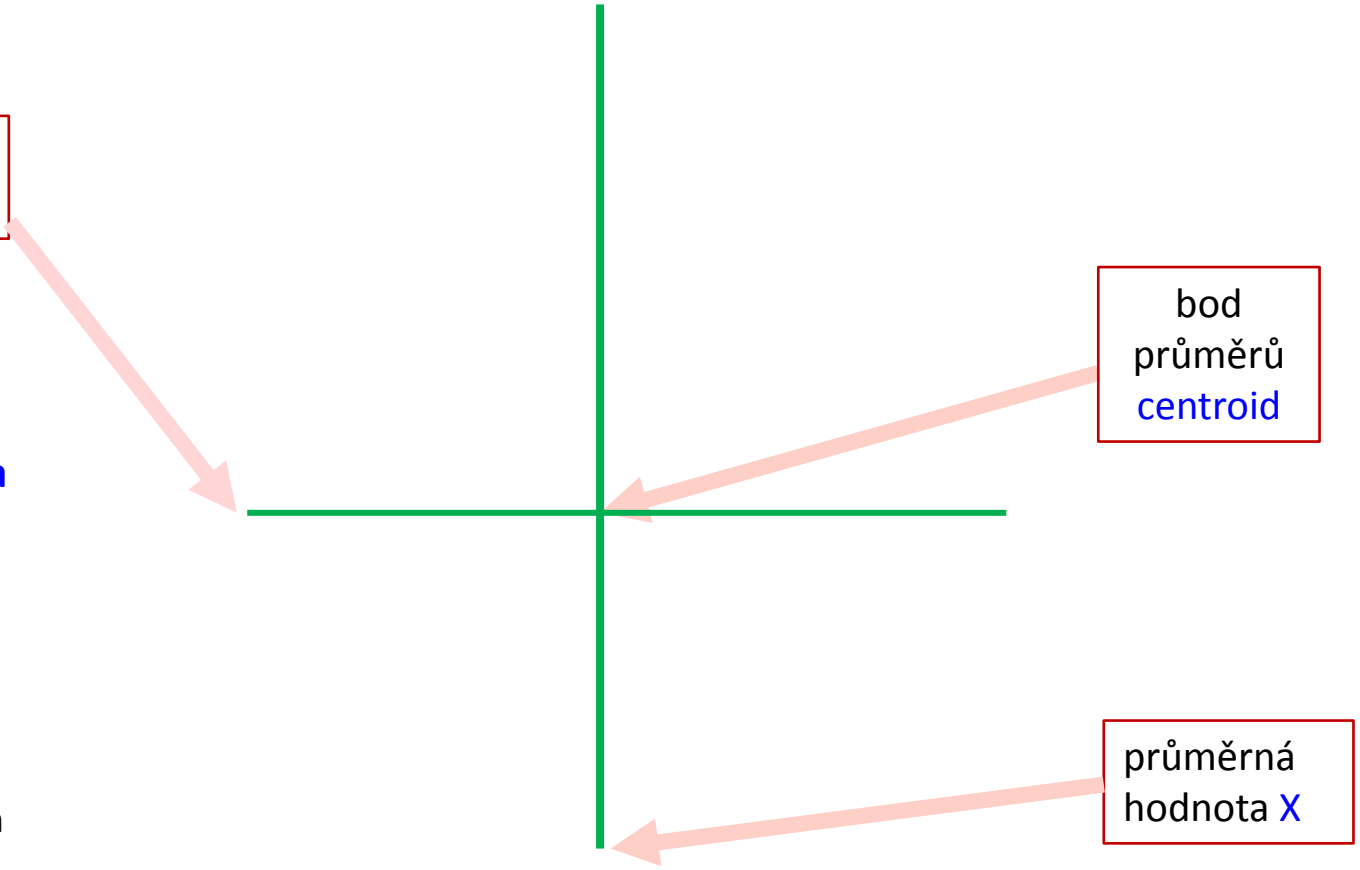
- **β_1 = regresní koeficient – koeficient úměry vlivy *X* na *Y* u každého jednoho případu**
  - $\beta_1 > 0$ – přímka má růstový/stoupavý trend
    - kladný trend
    - s rostoucím *X* roste *Y*
  - $\beta_1 < 0$ – pří
  - přímka má ztrátový/klesavý trend
    - záporný trend
    - s rostoucím *X* klesá *Y*
  - $\beta_1 = 0$ – přímka je rovnoběžná s osou *X*, absence trendu
    - s rostoucím *X* se *Y* nemění: nulový trend
    - hodnota *Y* na *X* nezávisí
- **β_0 = konstantní člen (posunutí)– hodnota *Y* pro nulové *X* nebo koeficient rovnoměrné změny pro každý případ bez ohledu na jeho *X* hodnotu**

Metoda nejmenších čtverců (MNČ)



Vlastnosti přímky získané MNČ

- **přímka minimalizuje součet čtvercových odchylek**
- **součet residuí je nula**
- **přímka prochází centroidem**
  - bod, kde všechny proměnné jsou rovné svým průměrům
  - průměr skutečných a vyrovnaných hodnot je stejný

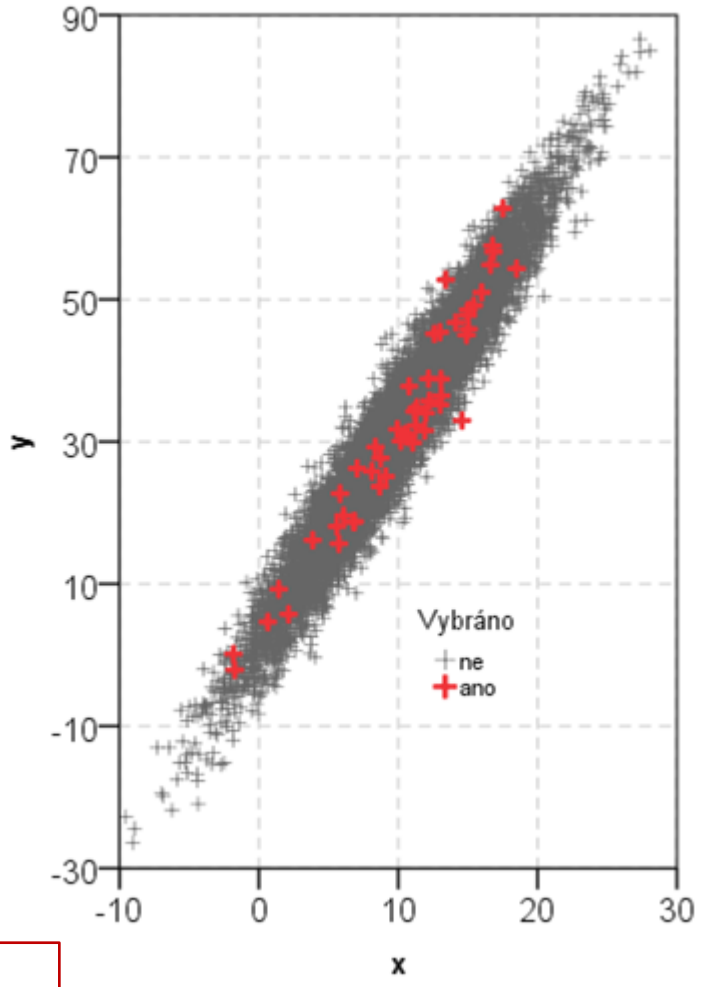


$$\bar{\hat{Y}} = \bar{Y}$$

$$b_0 = \bar{Y} - b_1\bar{X} \qquad b_1 = Cov(Y,X)/{s_X}^2$$

Odhad na základě výběru

- získané koeficienty a vše z nich vyplývající jsou jen **odhadem** skutečných koeficientů
- skutečné koeficienty se týkají **základního souboru** (často hypotetický a nedosažitelný)
- pracujeme s **výběrem** ze základního soubor => získáme jen odhad koeficientů
- jiný výběr by vedl jinému odhadu
- odhad chyby se nazývá **residuum**
  - v teorii je termíny nutno důsledně rozlišovat



Vše je jen odhad!

$$Y = \beta_0 + \beta_1 X + \varepsilon \Rightarrow Y = b_0 + b_1 X + e$$

skutečný ale neznámý vztah v základním souboru

odhad vztahu na základě výběru

Vlastnosti odhadu MNČ

- **lineární odhad**
  - výpočetně výhodné
  - odhad koeficientů i vyrovnaná hodnota se dá vyjádřit jako vážený součet hodnot *Y* pevně danými koeficienty
  - zajišťuje přibližnou normalitu i pro nenormální data
  - odhad závisí na každé hodnotě Y

$$b = \Sigma c_i Y_i \qquad \hat{Y} = \Sigma d_i Y_i$$

- **nevychýlený a konzistentní odhad parametrů rovnice**
  - odhad je rozptýlen kolem skutečných parametrů
  - s růstem parametrů se odhad blíží ke skutečným hodnotám
- **nejlepší odhad**
  - MNČ dává odhad s nejmenším rozptylem
  - pro daný výběr a model nelze odhad spočítat lépe
  - velikost rozptylu je úměrná $\sigma_\varepsilon/\sqrt{n}$ – je závislá na schopnosti uživatele najít dobrý model a získat dostatek případů pro odhad

Best **L**inear **U**nbiased **E**stimator

Koeficient determinace

- **vychází se z rozkladu rozptylu *Y***
- **ukazuje, jakou část rozptylu *Y* vysvětluje rozptyl $\hat{Y}$ neboli model**
  - zbytek rozptylu *Y* je rozptyl residuí
- **popisuje sílu vztahu modelu a závislé proměnné – $\hat{Y}$ a *Y***
  - je-li vysvětlujících proměnných *X* více, popisuje jejich společné působení na *Y*
- **R^2 = čtverec korelačního koeficientu $r(Y, \hat{Y}) \Rightarrow R^2 = r(Y, \hat{Y})^2$**
  - v modelu s jednou proměnnou *X* platí také $R^2 = r(Y, X)^2$
- **často se vyjadřuje v procentech**
- **nezávisí na počtu případů, ale na kvalitě vztahu v základním souboru**

$$R^2 = \sigma_{\hat{Y}}^2 / \sigma_Y^2$$
$$= 1 - \sigma_\varepsilon^2 / \sigma_Y^2$$
$$= 1 - ESS / TSS = MSS/TSS$$

Testování významnosti modelu: ANOVA

- **vychází se z rozkladu rozptylu *Y***
- **F - test – kritérium pro zjištění existence vztahu**
- **testuje existenci vztahu modelu a závislé proměnné – $\hat{Y}$ a *Y***
  - je-li vysvětlujících proměnných *X* více, testuje jejich společné působení na *Y*
- **závisí na počtu případů a na kvalitě vztahu v základním souboru**
  - čím více případů tím spíše se H_0 zamítne
- **velmi mírný – H_0 zamítnuta takřka vždy**
  - např. pro přímku a 50 případů je významný vztah s R^2=7,8%

H_0: $\beta_i = 0$ pro všechna *X*

H_1: $\beta_i \neq 0$ alespoň pro jedno *X*

$F(p-1, n-p) = [MSS/(p-1)]/[ESS/(n-p)] =$

$R^2 / (1- R^2) * (n-p)/(p-1)$

dosažená významnost F = α^*

n = počet případů

p = počet regresních koeficientů

F(p-1,n-p) – rozdělení F se stupni volnosti p-1 a n-p

Testy významnosti koeficientů

- **jsou založeny na směrodatné odchylce residuí se**
- **testuje existenci vztahu proměnné nezávislé a závislé proměnné – *X* a *Y***
  - každá proměnná *X* se testuje zvlášť
- **závisí na počtu případů a na kvalitě vztahu v základním souboru**
  - čím více případů tím spíše se H_0 zamítne
- **nezamítnutá H_0 znamená slabý (neexistující) vztah – proměnnou *X* z modelu vyloučíme**

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

pro rovnici s jedním prediktorem

$t^2 = F$

$t\,(n\text{-}p) = b_i / s_{bi} = b_i / [s_e * c_n(X)]$

dosažená významnost t = α^*

n = počet případů

p = počet regresních koeficientů

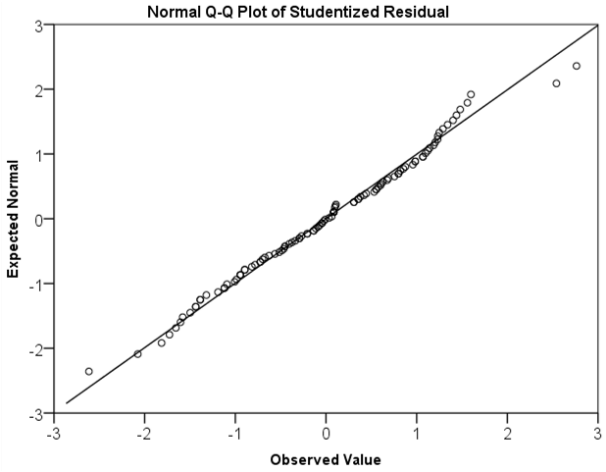
s_{bi} = směrodatná odchylka odhadu b

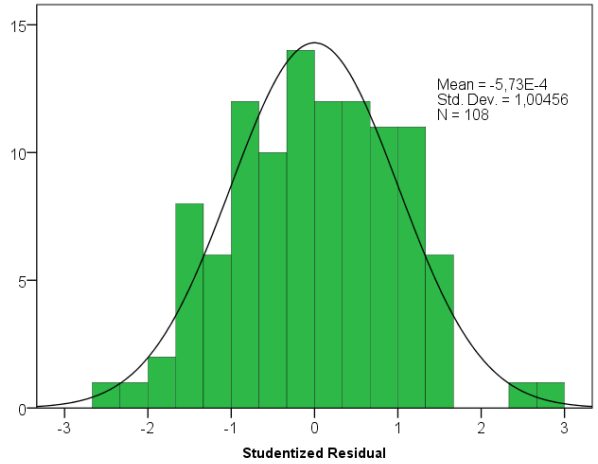
$c_n(x)$ = hodnota pevně určená proměnnými X a počtem případů

t(n-p) – rozdělení t s n-p stupni volnosti

Normalita residuí

- **je podstatná jen pro testování a intervaly spolehlivosti**
- **není kritická, pro větší soubory (>50) je normalita odhadu b zaručena na základě centrálního limitního teorému**
  - testy a intervaly pro parametry jsou v pořádku, i když residua nejsou normálně rozložena
  - intervaly pro individuální hodnoty jsou ale zkreslené
- **možnost otestovat – nejvhodnější jsou studentizovaná residua (stejný rozptyl)**
  - histogram
  - Q-Q, P-P graf
  - testy normality – s rostoucím počtem případů zamítají i nepatrné odchylky





Testy významnosti koeficientů – ukázka

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**97,017**	15,558		6,236	,000	66,173	127,862
	Výška otce	**,452**	,088	,445	5,122	,000	,277	,627

a. Dependent Variable: Výška syna

- **vztah výšky otce a syna je statisticky významný (Signifikance = 0,00)**
  - průměrná výška synů dvou otců, jejichž výška se liší o 1 cm, se liší o 0,45 cm
- **úrovňová konstanta je statisticky významná (Signifikance = 0,00)**
  - průměrná výška syna otce, který by měřil 0 cm, by byla 97 cm
  - v tomto případě není konstanta věcně smysluplná
- **skutečný koeficient vztahu výšek leží v intervalu (0,277;0,627) s pravděpodobností 95%**
  - interval neobsahuje 0 – ekvivalentní zamítnutí hypotézy o nulovosti koeficientu

- **u jednoduché lineární regrese s jedním prediktorem je Beta = r**
- **standardizovaný koeficient regrese je roven korelačnímu koeficientu obou proměnných**

Lineární regresní analýza – více prediktorů

- **přímé zobecnění jednoduché regrese**
- **další členy jsou přidány prostým přičtením, každá člen má svůj koeficient β_k**
- **mohou se přidávat i libovolné pevně dané funkce proměnných *X***
  - X^2, X^3, $1/X$, $ln(X)$, X_1X_2, atd.
  - modelem je křivka obecnější než přímka (rovina)
  - speciální variantou jsou proměnné typu 0 - 1
- **linearita – model je součtem jednotlivých komponent $\beta_k f(X_k)$**
- **Interpretace analogická jako u jednoduché regrese**

$$Y = f(X_1, X_2, X_3, \ldots X_k) + \varepsilon$$

$$E(Y|X) = \hat{Y} = f(X_1, X_2, X_3, \ldots X_k)$$

R^2,R a R^2_{adj}

- **ukazuje, jakou část rozptylu *Y* vysvětluje rozptyl $\hat{Y}$ neboli model**
  - zbytek rozptylu *Y* je rozptyl residuí
- **koeficient vícenásobné korelace R – korelační koeficient mezi *Y* a $\hat{Y}$ (lineární kombinace nezávislých proměnných *X*)**
  - lineární kombinace (odhadnutá rovnice) získaná MNČ maximalizuje korelační koeficient s *Y*
- **R^2 – čtverec vícenásobného korelačního koeficientu $R^2 = R(Y, \hat{Y})^2$**
- **R^2 vždy roste s přidáním nové proměnné nebo další funkce existujících proměnných (zvětšení modelu)**
  - řídit se pouze R^2 by vedlo k nesmyslně velkým modelům
- **R^2_{adj} – modifikované R^2**
  - samotné přidání proměnné je penalizováno snížením koeficientu
  - penalizace je slabá, R^2_{adj} po přidání proměnné téměř vždy vroste

$$R^2_{adj} = R^2 - (1- R^2)(p-1)/(n-p)$$

n = počet případů

p = počet regresních koeficientů

Výběr proměnných

- **obvykle ne všechny proměnné *X* v datech lze použít v modelu**
- **proměnné *X* mohou být korelovány – nelze je obě použít v jednom modelu, jejich vliv se vzájemně oslabuje (vysoká hodnota signifikance)**
- **často lze vytvořit více podobně kvalitních modelů**
- **metody pro automatické budování modelů**
  - postupné budování modelů podle kritérií založených na testech vlivu
  - sekvenční metody – další krok je závislý na předcházejících
  - zdaleka neprozkoumávají všechny možnosti
  - nalezený model není optimální (nepoužívá se kritérium optimality)
  - různé metody mohou vést k různým modelům
  - při větším počtu proměnných (asi >20) nemusejí vést ke smysluplným modelům
  - přijetí nebo modifikace nalezeného modelu je vždy volba uživatele

Automatický výběr proměnných

- **FORWARD – postupné zařazování prediktorů**
  - začíná s modelem obsahujícím jen konstantu
  - postupné zařazování prediktorů podle schopnosti snížit residuální rozptyl modelu – je požadována určitá míra snížení (volba uživatele)
  - vstup proměnné do modelu je silně závislý na proměnných dříve do modelu přidaných
  - přesnost modelu roste
  - není zaručeno, že všechny proměnné v modelu jsou signifikantní
- **BACKWARD – postupné vyřazování prediktorů**
  - začíná s plným modelem
  - postupně jsou odstraňovány proměnné, jejichž odstranění zvýší residuální rozptyl nejméně – je stanovena mez, kterou nesmí zvýšení překročit (volba uživatele)
  - přesnost modelu klesá
  - není vhodná, pokud je výchozí model příliš veliký
- **STEPWISE – kombinace obou**
  - začíná s modelem obsahujícím jen konstantu
  - přidává proměnné metodou FORWARD
  - po každém přidání zkouší metodou BACKWARD odstranit dříve přidané proměnné
  - nejkomplexnější

Forward – ukázka

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	52,125	,965		54,004	,000
	Logaritmus podílu nízkopříjmového obyvatelstva	-28,739	,909	-,815	-31,627	,000
2	(Constant)	22,886	3,552		6,443	,000
	Logaritmus podílu nízkopříjmového obyvatelstva	-22,302	1,138	-,633	-19,597	,000
	Průměrný počet místností na dům	3,598	,423	,275	8,512	,000
3	(Constant)	36,912	3,936		9,377	,000
	Logaritmus podílu nízkopříjmového obyvatelstva	-20,255	1,125	-,575	-18,000	,000
	Průměrný počet místností na dům	3,268	,406	,250	8,041	,000
	Počet žáků na učitele	-,762	,108	-,179	-7,026	,000
4	(Constant)	**41,456**	3,988		10,396	**,000**
	Logaritmus podílu nízkopříjmového obyvatelstva	**-24,005**	1,377	**-,681**	-17,438	**,000**
	Průměrný počet místností na dům	**2,743**	,415	**,210**	6,610	**,000**
	Počet žáků na učitele	**-,784**	,106	**-,185**	-7,362	**,000**
	Podíl budov starších než 1940	**,044**	,010	**,135**	4,559	**,000**

a. Dependent Variable: Medián ceny domu

roste

klesá

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,815	**,665**	,664	**5,329**
2	,841	**,707**	,706	**4,987**
3	,856	**,733**	,732	**4,763**
4	**,863**	**,744**	**,742**	**4,672**

FORWARD

Backward – ukázka

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42,560	4,053		10,501	,000
	Průměrný počet místností na dům	2,796	,416	,214	6,721	,000
	Podíl budov starších než 1940	,038	,011	,115	3,539	,000
	Počet žáků na učitele	-,824	,110	-,194	-7,506	,000
	Logaritmus podílu nízkopříjmového obyvatelstva	-24,068	1,376	-,683	-17,496	,000
	Podíl velkých domů	-,017	,011	-,043	-1,470	,142
2	(Constant)	**41,456**	3,988		10,396	**,000**
	Průměrný počet místností na dům	**2,743**	,415	**,210**	6,610	**,000**
	Podíl budov starších než 1940	**,044**	,010	**,135**	4,559	**,000**
	Počet žáků na učitele	**-,784**	,106	**-,185**	-7,362	**,000**
	Logaritmus podílu nízkopříjmového obyvatelstva	**-24,005**	1,377	**-,681**	-17,438	**,000**

a. Dependent Variable: Medián ceny domu

klesá

roste

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,863	**,745**	,743	**4,667**
2	**,863**	**,744**	,742	**4,672**

BACKWARD

Doporučení při budování modelu

- **využijte všech teoretických znalostí modelované problematiky**
- **prozkoumejte datovou situaci**
  - korelační matice nezávislých proměnných
  - bodové grafy všech nezávislých proměnných mezi sebou i se závislou proměnnou
- **z korelovaných nezávislých proměnných zvolte do modelu jednu, případně proměnné vhodně zkombinujte (např. vážený průměr)**
- **proměnné slabě korelované se závislou můžete z modelování vyloučit**
- **není-li vysvětlený rozptyl přijatelný, vyzkoušejte i funkce vysvětlujících proměnných (pozor na smysluplnost)**
- **kontrolujte statistickou významnost proměnných v modelu**
- **při automatickém budování modelu vyzkoušejte více metod a vždy zhodnoťte věcnou smysluplnost modelu, nalezené modely případně upravte**
- **vyzkoušejte více variant modelu a vyberte nejvhodnější i s ohledem na interpretovatelnost modelu**