MATEMATICKÁ STATISTIKA II. P3

2006-10-16

KORELAČNÍ A REGRESNÍ ANALÝZA – ANALÝZA ZÁVISLOSTÍ:

Odhad a testování korelačního koeficientu:

- ✓ Provádí se za předpokladu, že společné rozdělení obou proměnných lze modelovat dvourozměrným normálním rozdělením nebo – jinak vyjádřeno – rozdělení obou proměnných je normální a jejich vztah je přibližně lineární. Testuje se **hypotéza o nulové hodnotě** korelačního koeficientu základního souboru, tedy H_0: $\rho_{yx} = 0$.
- ✓ Hypotéza předpokládá, že korelace neexistuje, tzn. veličiny X a Y jsou nezávislé. **Alternativní hypotéza** je postavena na existenci korelace, tedy H_1: $\rho_{yx} \neq 0$.
- ✓ Test hypotézy se provádí pomocí testového kritéria $t = \frac{|r|}{\sqrt{1-r^2}} * \sqrt{n-2}$, které má za platnosti H_0 Studentovo t-rozdělení o f = n – 2 stupních volnosti. V případě, že vypočtená hodnota testového kritéria padne do kritického oboru, zamítá se nulová hypotéza a existence lineární korelační závislosti se považuje za prokázanou. $|t| \succ t_{\alpha(n-2)} \Rightarrow H_0$ *se zamítá*

Intervalový odhad korelačního koeficientu

- ✓ V případě, že výběrový soubor má dostatečně velký rozsah (n > 100), lze rozdělení výběrového korelačního koeficientu aproximovat normálním rozdělením.
- ✓ Oboustranný interval spolehlivosti je v daném případě možno psát $P(r - u_\alpha * s_r \leq \rho \leq r + u_\alpha * s_r) = 1 - \alpha$, přičemž $s_r = \frac{1-r^2}{\sqrt{n}}$.

Fischerova transformace:

- ✓ Ve většině případů (především, kdy n < 100) se však využívá Fisherovy transformace, neboť výběrový koeficient korelace neodpovídá kritériím bodového odhadu. Místo výběrového koeficientu korelace r se zavádí transformovaná veličina z_r.

  $$r \rightarrow z_r = \operatorname{arctan} h(r) = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

  Arctan h = arcus tangens hyperbolický.
- ✓ Touto transformací se rozšířil interval hodnot –1 ≤ r ≤ +1 na interval $-\infty \leq z_r \leq +\infty$. Nová proměnná má přibližně průměr μ_{zr} a směrodatnou odchylku s_{zr}.

  $$\mu_{z_r} = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

  $$s_{z_r} = \frac{1}{\sqrt{n-3}}$$
- ✓ Dvoustranný interval spolehlivosti pro transformovanou veličinu základního souboru má vyjádření $P(z_r - t_{\alpha(n-2)} * s_{z_r} \leq \mu_{z_r} \leq z_r + t_{\alpha(n-2)} * s_{z_r}) = 1 - \alpha$.
- ✓ Zpět do měřítka korelačního koeficientu převedeme oba krajní body intervalu pomocí inverzní transformace z_r^{-1}: $r = \frac{e^{2z} - 1}{e^{2z} + 1}$. Získáme tak interval spolehlivosti pro korelační koeficient ρ.

Příklad:

n = 30 r = 0,717078 $t_{0,05(28)} = 2,048$ H_0: $\rho_{yx} = 0$ H_1: $\rho_{yx} \neq 0$

$$t = \frac{0,717078}{\sqrt{1-0,717078^2}} * \sqrt{30-2} = 5,44399$$

$t \succ t_\alpha \Rightarrow H_0$ *se zamítá*

$$z_r = \frac{1}{2}\ln\left(\frac{1+0{,}717}{1-0{,}717}\right) = 0{,}9016$$

$$s_{z_r} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{30-3}} = 0{,}19245$$

$$0{,}9016 - 2{,}048 * 0{,}19245 \le \mu_{z_r} \le 0{,}9016 + 2{,}048 * 0{,}19245$$

$$P\left(0{,}468 \le \mu_{z_r} \le 1{,}29574\right) = 0{,}95$$

$$P\left(0{,}4680 \le \rho \le 0{,}8606\right) = 0{,}95$$

Regresní analýza:

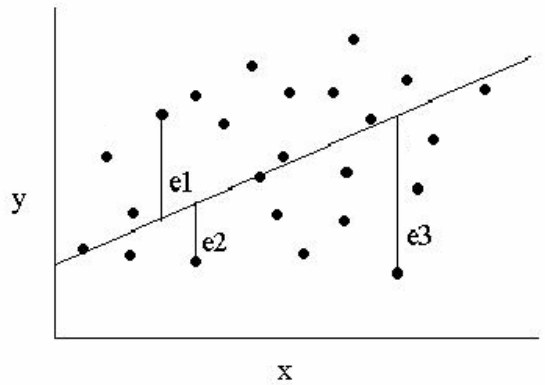
- ✓ Jde o přesnější popis tvaru vztahu mezi proměnnými X a Y a charakterizování jeho vhodnosti pro predikci hodnot závisle proměnné pomocí hodnot nezávisle proměnné.
- ✓ Může jít např. o následující situace:
  - Korelační koeficient i graf prokazují lineární vztah mezi spotřebou zemního plynu v bytě v závislosti na venkovní teplotě. Otázka zní, jak přesně můžeme predikovat spotřebu pomocí teploty.
  - Ve sportovním výzkumu máme např. data o rychlosti skokanů na hraně můstku a dosažené délce skoku. Zajímá nás, jaký je mezi nimi vztah: lze pomocí rychlosti predikovat délku skoku, s jakou přesností, je vztah lineární?
- ✓ V regresní analýze obecně analyzujeme vztah mezi jednou proměnnou zvanou **cílová nebo závislá proměnná** (Y) a několika dalšími, které nazýváme **nezávislé nebo ovlivňující proměnné** (X).
- ✓ Vztah reprezentujeme **matematickým modelem**, což je rovnice, jež svazuje závisle s nezávisle proměnnou a pravděpodobnostní předpoklady, které by měl vztah splňovat.
- ✓ Závisle proměnná se spojena s nezávisle proměnnými funkcí nazývanou **regresní funkcí**, jež obsahuje několik neznámých parametrů. Jestliže tato funkce je lineární v těchto parametrech (nemusí být lineární v proměnných), mluvíme o **lineárním regresním modelu**.
- ✓ Statistické problémy, která nás zajímají v regresní analýze, jsou:
  - Získání statistických odhadů neznámých parametrů regresní funkce.
  - Testování hypotéz o těchto parametrech.
  - Ověřování předpokladů regresního modelu.

Prokládání dat přímkou:

- ✓ Máme k dispozici uspořádané dvojice číselných údajů (x_1, y_1), (x_2, y_2), …, (x_n, y_n) pro proměnné X a Y.
- ✓ Jestliže graf ukáže lineární vztah mezi proměnnými, usilujeme o zachycení vztahu tím, že body proložíme **přímku**. Hledáme přímku, jež je experimentálním bodům co možná nejblíže (žádná přímka neprotne všechny body).
- ✓ Snažíme se určit takovou přímku, která bude co nejlépe predikovat y-hodnoty pomocí x-hodnot.
- ✓ Základní model regresní závislosti s jednou nezávisle proměnnou X vyjadřuje libovolnou hodnotu závisle proměnné Y jako $y_i' = f(x_i) + e_i$, kde $f(x_i)$ je tzv. regresní funkce a e_i je náhodná (reziduální) odchylka i-tého pozorování proměnné Y.

Reziduální odchylka:

- ✓ **Reziduální odchylka (chyba predikce)** – rozdíl mezi naměřenou a očekávanou hodnotou.



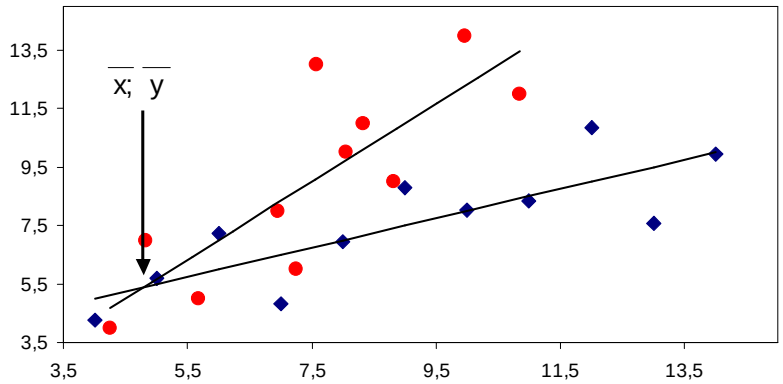
- ✓ Dobře proložená přímka y = a + b*x minimalizuje velikosti reziduálních hodnot pro hodnoty (x_i, y_i), kterými přímku prokládáme.
- ✓ Pro stanovení parametrů se nejčastěji používá metoda nejmenších čtverců. Hodnoty parametrů a, b přímky y = a + b*x získáme touto metodou tak, aby součet druhých mocnin reziduálních hodnot byl minimální vzhledem k parametrům a, b.
  $s_r^2 = \sum e_i^2 = \sum (y_i - a - bx_i)^2$
- ✓ Minimalizuje sečtené čtverce úseček, které vyznačují vzdálenost bodu od proložené přímky ve směru osy Y. Výpočet tohoto minima vede k optimálním hodnotám $a = \bar{y} - b\bar{x}$ a $b = r * \frac{s_y}{s_x}$, kde r je korelace obou proměnných a s_x, s_y jsou směrodatné odchylky naměřených hodnot proměnných X a Y.
- ✓ Hodnota y_i` je odhad cílové proměnné pomocí regresního vztahu ($y_i$` = a + bx_i): **reziduální hodnota** = naměřená hodnota y – predikovaná hodnota y`.
- ✓ Rozptýlenost bodů kolem přímky je charakterizována zbytkovým (reziduálním) rozptylem, případně směrodatnou chybou odhadu při regresi (lze také posoudit přesnost provedených regresních odhadů jako míru chyby) $s_{y.x}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - y_i')^2}{n-2}$.

Metoda nejmenších čtverců:

- ✓ **Metoda nejmenších čtverců** – postup stanovení parametrů u jednoduché lineární závislosti: $y_i' = a + bx_i$, $\sum_{i=1}^{n} (y_i - y_i')^2 = \min$.
- ✓ Z podmínky minimálnosti čtverců jsou vyvozeny normální rovnice, ze kterých se jejich řešením vypočtou neznámé parametry a a b: $f(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = \min$.
- ✓ Má-li tato funkce f(a,b) minimum, musejí se její první parciální derivace podle konstant a a b rovnat nule:
  $\frac{\partial f(a,b)}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(0 - 1 - 0) = -2\sum_{i=1}^{n} (y_i - a - bx_i)$
  $\frac{\partial f(a,b)}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(0 - 0 - x_i) = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i$, tedy platí
  $-2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$
  $-2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0$
- ✓ Vynásobením každé z rovnic –1/2, rozvedením součtů a osamostatněním součtů obsahujících y_i se získá soustava normálních rovnic.
  $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$
  $\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$
- ✓ Řešením soustavy normálních rovnic obdržíme $a = \bar{y} - b * \bar{x}$ a $b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$.

Závislost:

- ✓ **Jednostranná závislost** – proměnná X je nezávisle proměnná a Y pak závisle proměnná.
- ✓ **Oboustranná závislost** – nelze přesně rozhodnout, která proměnná je závislá a která nezávislá.
- ✓ $y'_i = a_{yx} + b_{yx}x_i$, $x'_i = a_{xy} + b_{xy}y_i$
- ✓ Vztahy pro regresi X na Y získáme vhodnou záměnou ve vzorcích (např. $b_{xy} = r*s_x/s_y$, kde r je korelační koeficient).
- ✓ Mezi směrnicemi obou regresních přímek b_{xy} a b_{xy} existuje vztah $r = \sqrt{b_{yx} * b_{xy}}$. Můžeme tedy nalézt dvě regresní přímky, které se budou protínat v bodě $(\bar{x}; \bar{y})$ a tvoří jakési nůžky. Čím větší je korelace, tím více jsou nůžky stisknuty.



Maticové vyjádření regresního problému:

- ✓ Lineární (teoretický) model lze zapsat jako $y = X\beta + \varepsilon$, ve kterém:
  y – n-členný náhodný vektor napozorovaných (zjištěných) hodnot vysvětlované proměnné Y,
  X – nenáhodná matice typu n x (k+1) zvolených n kombinací hodnot vysvětlujících proměnných,
  β – je (k+1)členný vektor neznámých parametrů modelu,
  ε – n-členný vektor nepozorovatelné rušivé (náhodné) složky.
- ✓ Pro lepší představu:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- ✓ Z uvedeného zápisu je vidět, že v n lineárních rovnicích je p = k+1 neznámých regresních parametrů a n hodnot náhodné složky.
- ✓ Soustavu normálních rovnic pro hledanou funkci y = Xb + ε lze pak v maticovém tvaru vyjádřit takto: $\mathbf{X'Xb = X'y}$
- ✓ Za předpokladu, že k matici X`X existuje matice inverzní, dostaneme vektor odhadovaných parametrů podle vztahu $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$
- ✓ Maticově lze stanovit i hodnotu korelačního indexu: $I = \sqrt{\dfrac{\mathbf{b'X'y} - \frac{1}{n}\sum(y_i)^2}{\mathbf{y'y} - \frac{1}{n}\sum(y_i)^2}}$

Předpoklady metody nejmenších čtverců:

- ✓ Regresní parametry β mohou nabývat libovolných hodnot. V technické praxi však často existují omezení parametrů, která vycházejí z jejich fyzikálního smyslu.
- ✓ Regresní model je lineární v parametrech a platí aditivní vztah $y = X\beta + \varepsilon$.
- ✓ Vysvětlující proměnné $X_1, X_2, \ldots, X_k$ jsou **nenáhodné** a neexistuje mezi nimi **funkční** lineární závislost.
- ✓ Pro danou kombinaci hodnot vysvětlujících proměnných jsou hodnoty nepozorovatelné rušivé složky ε_i **normálně rozdělené, nezávislé** náhodné veličiny s nulovými středními hodnotami a se stejným (konstantním)

rozptylem σ^2. Neboli vektor hodnot rušivé složky ε má n-rozměrné normální rozdělení N(0, σ^2) s vektorem středních hodnot E(ε) = 0 a s kovarianční maticí σ^2 E, kde E je jednotková matice.

- ✓ Náhodné chyby εi mají nulovou střední hodnotu E(ε_i) = 0, konstantní a konečný rozptyl E(ε_i^2) = σ^2. Také podmíněný rozptyl D(y/x) = σ^2 je konstantní a jde o homoskedastický případ.
- ✓ Náhodné chyby ε_i jsou vzájemně nekorelované a platí cov (ε_i, ε_j) = E(ε_i, ε_j) = 0. Pokud mají chyby normální rozdělení, jsou nezávislé.

$$\operatorname{cov}(\varepsilon_i\varepsilon_i)=\begin{bmatrix} D(\varepsilon_1) & \operatorname{cov}(\varepsilon_1\varepsilon_2) & \cdots & \operatorname{cov}(\varepsilon_1\varepsilon_n) \\ \operatorname{cov}(\varepsilon_2\varepsilon_1) & D(\varepsilon_2) & \cdots & \operatorname{cov}(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(\varepsilon_n\varepsilon_1) & \operatorname{cov}(\varepsilon_n\varepsilon_2) & \cdots & D(\varepsilon_n) \end{bmatrix}=\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Odhady v regresní analýze:

- ✓ **Interpolace** – předmětem zájmu je některá z použitých kombinací vysvětlujících proměnných
- ✓ **Extrapolace** – pozornost je upřena na hodnotu proměnné Y pro předpokládané budoucí nebo výzkumně zajímavé kombinace hodnot proměnné Y.
- ✓ Je nutné odlišit dva významově zásadně **odlišné případy**:
  - Odhad průměrné hodnoty Y neboli odhad podmíněné střední (očekávané) proměnné Y vzhledem ke zvolené hodnotě (kombinaci hodnot) vysvětlující proměnné.
  - Odhad konkrétní hodnoty $y`_i$ neboli předpověď $y`_i = a + bx_i$ hodnoty proměnné Y vzhledem ke zvolené hodnotě (kombinaci hodnot) vysvětlující proměnné.

Pás spolehlivosti kolem regresní přímky:

- ✓ Z rovnice regresní přímky zkoumaného souboru lze určovat teoretickou hodnotu závisle proměnné příslušející určité skutečné hodnotě nezávisle proměnné. Avšak skutečné konkrétní hodnoty závisle proměnné jsou více méně rozptýleny kolem stanovené regresní přímky.
- ✓ Se zvolenou pravděpodobností je možno určit tzv. pás spolehlivosti, v němž se tyto skutečné (empirické) hodnoty nacházejí jako $y'_i \pm t_{1-\frac{\alpha}{2}} \times s_{y.x}$.

$t_{1-\frac{\alpha}{2}}$ jsou 100 (1-α/2)% kvantily Studentova t-rozdělení s (n-2) stupni volnosti

$s_{y.x}$ je směrodatná chyba, která je rovna $s_{y.x}=\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-y'_i)^2}{n-2}}=\sqrt{\frac{\sum_{i=1}^{n} y_i^2-\sum_{i=1}^{n} y_i y'_i}{n-2}}$, přičemž

$$\sum_{i=1}^{n} y_i y'_i=\sum_{i=1}^{n} y_i\left(a_{yx}+b_{yx}x_i\right)=a_{yx}\sum_{i=1}^{n} y_i+b_{yx}\sum_{i=1}^{n} x_i y_i$$

Příklad:

Pro závislost proměnné Y na proměnné X byla stanovena regresní přímka ve tvaru $y`_i$ = 4,375 + 0,01994 x_i a pomocné výpočty $\sum y_i = 117{,}1$, $\sum y_i^2 = 1162{,}35$, $\sum x_i y_i = 32005{,}4$

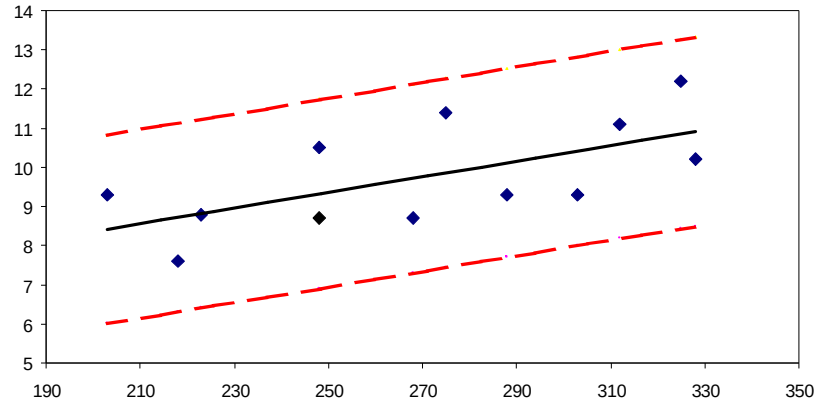
$$s_{y.x}=\sqrt{\frac{1162{,}35-\left(4{,}375\times 117{,}1+0{,}01998\times 32005{,}4\right)}{12-2}}=1{,}082$$

n = 12

$t_{1-\alpha/2\,(10)}$ = 2,228

$y_{i\,(min,\,max)}$ = 4,375 + 0,01994 x_i ± 2,228 * 1,082

Znamená to, že dolní mez pro skutečné hodnoty je $y_{i\,(min)}$ = 1,96456 + 0,01994 x_i a horní mez $y_{i\,(max)}$ = 6,78626 + 0,01994 x_i



Test významnosti regresního koeficientu:

- ✓ Nulová hypotéza předpokládá, že výběrový koeficient regrese je odhadem regresního koeficientu ZS, o němž se předpokládá, že má nulovou hodnotu, tzn. že platí H_0: $\beta_{yx} = 0$.
- ✓ Testové kritérium má tvar $t = \frac{|b_{yx}|}{s_{b_{yx}}}$, kde $s_{b_{yx}} = \frac{s_y}{s_x} \times \sqrt{\frac{1-r^2}{n-2}}$.

  $|t| \succ t_{\alpha(n-2)} \Rightarrow H_0$ *se zamítá*
- ✓ V případě, že se zamítá H_0, je existence lineární závislosti prokázána a odvozenou regresní funkci lze použít k provádění regresních odhadů.

Intervalový odhad regresního koeficientu:

- ✓ Oboustranný interval spolehlivosti pro regresní koeficient je vymezen následujícím způsobem: $P\left(b_{yx} - t_{\alpha(n-2)} \times s_{b_{yx}} \le \beta_{yx} \le b_{yx} + t_{\alpha(n-2)} \times s_{b_{yx}}\right) = 1-\alpha$

Příklad:

$\hat{y}_i = 4{,}375 + 0{,}01994\, x_i$

H_0: $\beta_{yx} = 0$

$t_{0{,}05(10)} = 2{,}228$

$$s_{b_{yx}} = \frac{1{,}33652}{42{,}6027} * \sqrt{\frac{1-0{,}635697^2}{12-2}} = 0{,}0076581$$

$$t = \frac{0{,}0199429}{0{,}0076581} = 2{,}60416$$

$t \succ t_\alpha \Rightarrow H_0$ *se zamítá*

$$P\left(0{,}01994 - 2{,}228 * 0{,}0076581 \le \beta_{yx} \le 0{,}01994 + 2{,}228 * 0{,}0076581\right) = 0{,}95$$

$$P\left(0{,}00288 \le \beta_{yx} \le 0{,}037\right) = 0{,}95$$

Test významnosti regresní přímky:

- ✓ K testování lze použít upravený model analýzy rozptylu.

Variabilita	Součet čtverců	Stupně volnosti	Rozptyl	Testovací kritérium
Regrese		p - 1	$s_1^2 = \frac{S_1}{p-1}$	$F = \frac{s_1^2}{s_r^2}$
Kolem regrese		n - p	$s_r^2 = \frac{Sr}{n-p}$	

p – počet parametrů ověřované funkce

- ✓ Jestliže $F > F_{\alpha\,[(p-1);\,(n-p)]}$, zamítáme H_0.

✓ Příslušné součty čtverců se stanoví následujícím způsobem:

- Pro variabilitu regrese $S_1 = \sum_{i=1}^{n} (y_i' - \bar{y})^2$
- Pro variabilitu kolem regrese $S_r = \sum_{i=1}^{n} (y_i - y_i')^2$
- Pro celkovou variabilitu $S = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Příklad:

Pro závislost proměnné Y na proměnné X byla stanovena regresní přímka ve tvaru $y`_i = 4{,}375 + 0{,}01994\ x_i$.

x_i	y_i	$y`_i$	$y_i' - \bar{y}$	$y_i - y`_i$
268	8,7	9,720109	0,0015	1,0406
312	11,1	10,5976	0,7044	0,2524
223	8,8	8,822679	0,8754	0,0005
203	9,3	8,423821	1,7809	0,7677
248	8,7	9,321251	0,1910	0,3860
328	10,2	10,91668	1,3418	0,5136
303	9,3	10,41811	0,4353	1,2502
325	12,2	10,85685	1,2067	1,8040
275	11,4	9,85971	0,0103	2,3725
218	7,6	8,722964	1,0720	1,2610
248	10,5	9,321251	0,1910	1,3894
288	9,3	10,11897	0,1301	0,6707
celkem	--	--	7,9404	11,7087

$\bar{y} = 9{,}75833$

$S_1 = 7{,}9404$

$S_r = 11{,}7087$

$S = 19{,}6492$

$$s_1^2 = \frac{S_1}{p-1} = \frac{7{,}9404}{2-1} = 7{,}9404$$

$$s_r^2 = \frac{Sr}{n-p} = \frac{11{,}7087}{12-2} = 1{,}17087$$

$$F = \frac{s_1^2}{s_r^2} = \frac{7{,}9404}{1{,}17087} = 6{,}7816$$

$F_{0{,}05\ [(2-1);\ (12-2)]} = 4{,}96$

$F > F_{\alpha\ [(p-1);\ (n-p)]} \Rightarrow$ zamítáme H_0

Intervalový odhad regresní přímky

✓ Interval spolehlivosti, který s danou pravděpodobností pokrývá hledanou regresní přímku základního souboru $y`_j = \alpha_{yx} + \beta_{yx} x_j$, je určen na základě regresní přímky výběrového souboru $y`_i = a_{yx} + b_{yx} x_i$ a je vyjádřen takto:

$$P\left(y_i' - u_{1-\frac{\alpha}{2}} \times s_{\bar{y}} \sqrt{1 + \frac{(x_i - \bar{x})^2}{s_x^2}} \le y_j' \le y_i' + u_{1-\frac{\alpha}{2}} \times s_{\bar{y}} \sqrt{1 + \frac{(x_i - \bar{x})^2}{s_x^2}} \right) = 1 - \alpha$$

$$y_{j\,(H,D)}' = y_i' \pm t_{1-\frac{\alpha}{2}} \times s_{\bar{y}} \sqrt{1 + \frac{(x_i - \bar{x})^2}{s_x^2}}$$

$$s_{\bar{y}} = \frac{s_y}{\sqrt{n}}$$

s_x^2 – rozptyl proměnné X

s_y – směrodatná odchylka proměnné Y

Příklad:

$\bar{x} = 269{,}92$

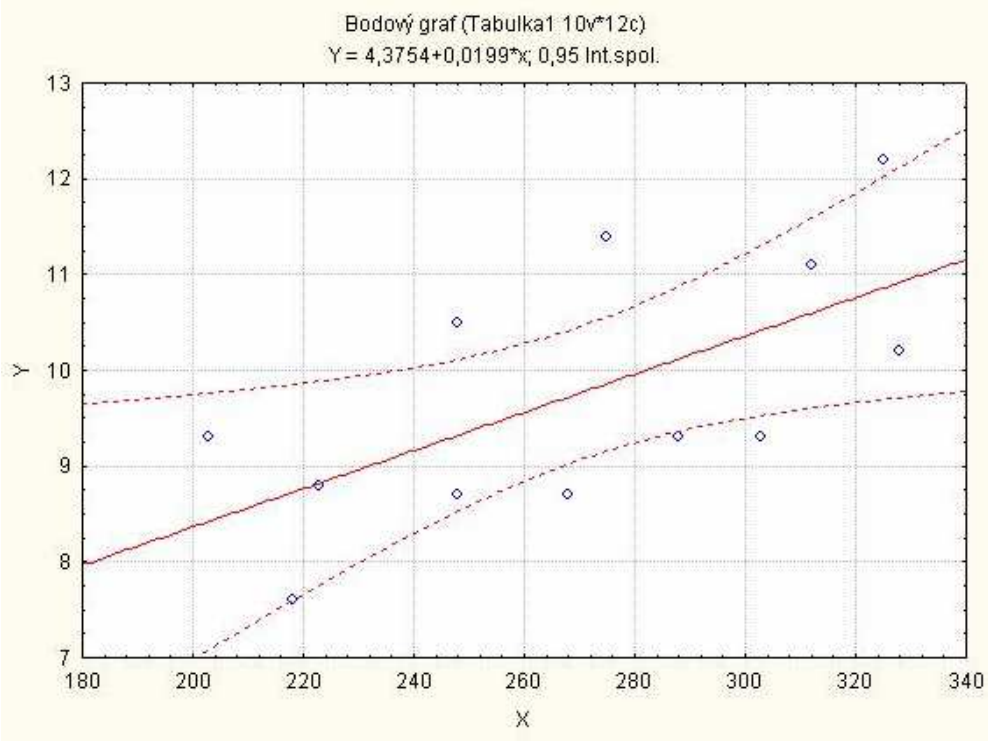
$s_x = 42{,}6027$

$s_y = 1{,}33652$

$$s_{\bar{y}} = \frac{1{,}33652}{\sqrt{12}} = 0{,}38582$$

$$y'_{j(H,D)} = 4{,}3754 + 0{,}1994\, x_i \pm 2{,}228 \times 0{,}38582 \sqrt{1 + \frac{(x_i - 269{,}92)^2}{1814{,}9924}}$$

x_i	$y`_i$	$y`_{j(H)}$	$y`_{j(D)}$
268	9,720	8,860	10,581
312	10,598	9,389	11,806
223	8,823	7,544	10,101
203	8,424	6,823	10,024
248	9,321	8,355	10,288
328	10,917	9,463	12,370
303	10,418	9,330	11,506
325	10,857	9,452	12,262
275	9,860	8,994	10,725
218	8,723	7,368	10,078
248	9,321	8,355	10,288
288	10,119	9,185	11,053



F-test:

- ✓ Standardním výstupem většiny programů regresní analýzy je závěr Fisherova-Snedecorova F-testu o významnosti regresní přímky a výsledky Studentova t-testu o významnosti jednotlivých parametrů vektoru β (vektor regresních parametrů modelu).
- ✓ F-test určuje zároveň simultánní významnost všech složek vektoru β kromě absolutního členu. Mohou tedy nastat tyto případy:
  - F-test vychází nevýznamný, všechny t-testy vychází rovněž jako nevýznamné. Model se pak považuje za nevhodný, protože nevystihuje variabilitu proměnné y.
  - F-test i všechny t-testy vychází významné. Model se považuje za vhodný k vystižení variability proměnné *y*. To však ještě neznamená, že je model navržen správně.
  - F-test vychází významný, ale t-testy nevýznamné u některých regresních parametrů. Model je považován za vhodný a provádí se případné vypouštění těch vysvětlujících proměnných x_i, pro které jsou parametry β_i nevýznamně odlišné od nuly.
  - F-test sice vychází významný, ale t-testy parametrů β indikují nevýznamnost všech vysvětlujících proměnných. To je paradox, protože formálně sice model jako celek vyhovuje, ale žádná z vysvětlujících proměnných není sama o sobě významná. Jde o důsledek multikolinearity.