PRAVDEPODOBNOSŤ a ŠTATISTIKA

Korelačná a regresná analýza

doc. RNDr. Štefan Peško, CSc.

Katedra matematických metód a operačnej analýzy, FRI ŽU

13. mája 2019

V praxi sa stretávame s problematikou skúmania a hodnotenia závislosti medzi premennými, ktoré predstavujú poznatky technických, ekonomických, informatických či sociálnych odborov. No nemôžeme spoľahnúť na odvodené jednoznačné funkčné vzťahy v tvare $y = f(x_1, x_2, \dots x_n)$, nakoľko buď nevieme presne získať hodnoty nezávislých premenných alebo pozorovaná závislá premenná nadobúda rôzne hodnoty pri tých istých hodnotách nezávislých premenných.

Pozorované hodnoty závislej premennej y potom modelujeme ako realizácie náhodnej premennej $\mathbf{Y}$ s hodnotami $y_1, y_2, \dots, y_n$ a pozorované hodnoty nezávislých premenných $x_1, y_2, \dots, x_n$ môžeme interpretovať buď ako realizácie náhodných premenných $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ s príslušnými pravdepodobnostnými rozdeleniami alebo ako nenáhodné premenné nadobudajúce práve jednu z možných hodnôt.

Príkladmi takých premenných sú

- spotreba, vek, typ, hmotnosť, rok výroby, najazdené kilometre vozidla,
- príjmy, výdavky na jedlo, výdavky na bývanie, počet členov rodín,
- prospech, chorobnosť študentov vo vybraných ročníkoch, lokalita a typ škôl,
- počet dní absencie, vek, vzdelanie, pracovné zaradenie zamestnancov podniku.

Ak sledujeme hodnoty štatistických znakov $Y, X_1, X_2, \ldots, X_n$ a určitým spôsobom sa mení podmienené rozdelenie znaku Y pri zmenách štatistických znakov $Y, X_1, X_2, \ldots, X_n$ hovoríme o štatistickej závislosti znaku Y na $X_1, X_2, \ldots, X_n$.

Špeciálnym typom takej závislosti je tzv. korelačná závislosť.

Koreláciou rozumieme vzájomný lineárny vzťah (závislosť) dvoch náhodných premenných.

V teórii pravdepodobnosti bola ako miera lineárnej závislosti náhodných premenných **X** a **Y** zavedený Pearsonov koeficient korelácie:

$$\rho = \rho(\mathbf{X},\mathbf{Y}) = \begin{cases} \dfrac{cov(\mathbf{X},\mathbf{Y})}{\sqrt{D(\mathbf{X})\mathbf{D}(\mathbf{Y})}} & \text{ak } D(\mathbf{X}) > 0, D(\mathbf{Y}) > 0, \\ 0 & \text{ináč} \end{cases} \tag{1}$$

pomocou kovariancie $cov(\mathbf{X},\mathbf{Y}) = E(\mathbf{XY}) - E(\mathbf{X})E(\mathbf{Y})$ a rozptylov náhodných premenných.

Vlastnosti koeficientu korelácie ρ

Dohoda: Keď nebude hroziť nedorozumenie budeme premenné $\mathbf{X}$ a $\mathbf{Y}$ vynechávať a písať len ρ miesto $\rho(\mathbf{X},\mathbf{Y})$.

- $-1 \leq \rho \leq 1$, pričom k rovnosti dochádza ak je medzi premennými lineárna závislosť,
- ak sú premenné $\mathbf{X},\mathbf{Y}$ nezávislé, potom $\rho = 0$,
- ak je $\rho = 0$, potom hovoríme, že $\mathbf{X},\mathbf{Y}$ sú nekorelované náhodné premenné,
- ak je $\rho > 0$, potom hovoríme, že $\mathbf{X},\mathbf{Y}$ sú pozitívne korelované (s rastúcim $\mathbf{X}$ rastie $\mathbf{Y}$),
- ak je $\rho < 0$, potom hovoríme, že $\mathbf{X},\mathbf{Y}$ sú negatívne korelované (s rastúcim $\mathbf{X}$ klesá $\mathbf{Y}$),

Výberový koeficient korelácie

Pearsonov koeficient korelácie ρ vieme určiť len ak poznáme združené rozdelenie náhodného vektora $(\mathbf{X}, \mathbf{Y})$. V praxi však máme k dispozícii len realizácie náhodného výberu z nejakého dvojrozmerného rozdelenia $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
Potom má zmysel definovať výberový korelačný koeficient **r** pomocou výberových charakteristík – analogicky vzťahu (1) – takto:

$$\mathbf{r} = r(\mathbf{X}, \mathbf{Y}) = \frac{S_{\mathbf{X},\mathbf{Y}}}{S_{\mathbf{X}} S_{\mathbf{Y}}} \text{ pre } S_{\mathbf{X}} > 0, S_{\mathbf{Y}} > 0 \tag{2}$$

kde $S_{\mathbf{X}}$ a $S_{\mathbf{Y}}$ sú výberové smerodajné odchylky premených **X** a **Y** a $S_{\mathbf{X},\mathbf{Y}}$ je výberová kovariancia definovaná vzťahom:

$$S_{\mathbf{X},\mathbf{Y}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}}). \tag{3}$$

Po úpravách dostaneme realizáciu **r**:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - (\overline{x})^2}\sqrt{\overline{y^2} - (\overline{y})^2}} \tag{4}$$

Koeficient determinácie

Stupeň príčinnej závislosti premennej **Y** od premennej **X** vyjadruje koeficient determinácie, ktorý je definovaný ako druhá mocnina korelačného koeficientu r^2.

Interpretácia koeficienta determinácie vychádza z analýzy variability (rozptylu) závisle premennej **Y**, ktorú by mala do značnej miery vysvetliť variabilita nezávisle premennej **X** za predpokladu, že od nej lineárne závisí veľkosť hodnôt **Y**.

Ak napr. $r = 0.7$, potom $r^2 = 0.49$ čo znamená, že iba 49% variability premennej **Y** sa dá vysvetliť lineárnym vzťahom s premennou **X** (regresnou priamkou). Pretože 51% variability premennej **Y** zostalo nevysvetlenej lineárnym vzťahom s premennou **X**, je zrejmé, že model bol zvolený nevhodne (namiesto lineárnej závislosti sa mala uvažovať nelineárna závislosť).

Vlastnosti koeficientu korelácie ρ sa prenášajú aj na korelačný koeficient r. Keď zistíme, že výberový korelačný koeficient $r \neq 0$, potom nás zaujíma či je táto korelácia štatisticky významná.

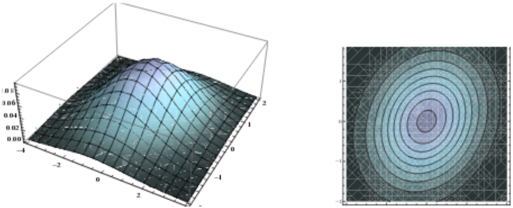
Môžeme testovať nulovú hypotézu $H_0 : \rho = 0$ proti alternatívnym hypotézam $H_A : \rho \neq 0$ alebo $H_A : \rho < 0$ alebo $H_A : r > 0$.

Nech je $\mathbf{X} = ((\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, \mathbf{Y}_n))$ náhodný výber z dvojrozmerného normálneho rozdelenia. Potom má za predpokladu platnosti nulovej hypotézy, testovaciu štatistika

$$T(\mathbf{X}) = \frac{\mathbf{r}\sqrt{n-2}}{\sqrt{1-\mathbf{r}^2}} \sim t(n-2) \tag{5}$$

Studentovo rozdelenie s $n-2$ stupňami voľnosti. Rozhodnutie výsledku testu urobíme pomocou štandartnej p-hodnoty.

Poznámky k dvorozmernému normálnemu rozdeleniu



Obr. 1: Hustotota a vrstevnice dvojrozmerného normálneho rozdelenia.

- Ak má náhodný vektor $(\mathbf{X}, \mathbf{Y})$ dvojrozmerné normálne rozdelenie, potom jeho zložky majú tiež normálne rozdelenie $\mathbf{X} \sim N(\mu_{\mathbf{X}}, \sigma^2_{\mathbf{X}}), \mathbf{Y} \sim N(\mu_{\mathbf{Y}}, \sigma^2_{\mathbf{Y}})$. Predpoklad o združenom rozdelení $(\mathbf{X}, \mathbf{Y})$ sa veľmi ťažko overuje, normalita zložiek je len nutná no nie postačujúca podmienka, ale pre prax stačí.
- Ak sú zložky náhodného vektora $(\mathbf{X}, \mathbf{Y})$ nekorelované, potom sú aj nezávislé a tak je nekorelovanosť ekvivalentná nezávislosti, čo však všeobecne neplatí.

Príklad 11.1

Vyučujúci má k dispozícii výsledky 1. a 2. semestrálnej písomky 10-tich študentov. Na hladine významnosti 0.05 testujte hypotézu, či sú vysledky kladne korelované.

1.	7	8	10	4	14	9	6	2	13	5
2.	9	7	12	6	15	6	8	2	11	8

Na overenie, či sú výsledky realizácie dvojrozmerného normálneho rozdelenie použijeme Shapiro-Wilk test normality, ktorý ponúka Python, p-hodnota$_{\mathbf{X}} = 0.9305 > 0.1$ a p-hodnota$_{\mathbf{Y}} = 0.9398 > 0.1$ a tak nulovú hypotézu nezamietame.

Budeme testovat $H_0 : \rho = 0$ proti $H_A : \rho > 0$. Z výberových charakteristík $\overline{x} = 7.8, \overline{y} = 8.6, s_{\mathbf{X}}^2 = 14.6, s_{\mathbf{Y}}^2 = 10.7, s_{\mathbf{X},\mathbf{Y}} = 10.6$ máme $r = 0.845$ a po dosadení do (5) $t_{OBS} = 4.47$ odkiaľ p-hodnota $= 1 - F_0(t_{OBS}) = 0.001$ a na hladine významnosti 0.05 zamietame nulovú hypotézu v prospech alternatívy t.j. výsledky písomiek môžeme považovať za kladne korelované.

použijeme ak nie je splnený predpoklad o normalite nameraných dvojíc údajov alebo aspoň jeden údaj je ordinárny (poradie).

Na náhodný výber $(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ sa budeme pozerať ako na dva výbery dané jeho zložkami. Realizácie náhodného vektora $\mathbf{X}$ aj náhodného vektora $\mathbf{Y}$ usporiadame do neklesajúcej postupnosti. Nech p_i, q_i je poradie hodnoty x_i, y_i. Dosadením p_i, q_i za x_i, y_i do vzťahu (2) dostaneme Spearmanov koeficient poradovej korelácie

$$r_S(\mathbf{P}, \mathbf{S}) = \frac{\sum_{i=1}^{n}(p_i - \overline{p})(q_i - \overline{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})}\sqrt{\sum_{i=1}^{n}(q_i - \overline{q})}} \tag{6}$$

kde $\mathbf{P}, \mathbf{Q}$ sú náhodné permutácie poradí množiny $1, 2, \ldots, n$ a priemer poradí $\overline{p} = \overline{q} = \frac{n+1}{2}$.
Spearmanov koeficient korelácie nadobúda hodnoty z intervalu $< -1, 1 >$ a silu vzťahu medzi štatistickými znakmi interpretujeme rovnako ako pri Pearsonovom koeficiente korelácie.

Príklad 11.2

V tabuľke sú uvedené umiestnenia hokejistov v kanadskom bodovaní po skončení základnej časti a ich príjmy (mil. dolárov).

Poradie	1	2	3	4	5	6	7	8	9	10
Príjem	3.75	2.85	2.1	3.2	2.25	1.9	3.21	2.15	1.8	2.65

Zistite, či existuje závislosť medzi výkonnosťou a príjmami hokejistov v NHL.

P – poradie hráča v kanadskom bodovaní, **S** – príjem hráča. Koeficient poradovej korelácie vypočítame podľa vzťahu (5)

$$r_S(\mathbf{P}, \mathbf{S}) = \frac{36.5}{\sqrt{82.5}\sqrt{82.5}} \approx 0.4424$$

Koeficient determinácie $r_S(\mathbf{P}, \mathbf{S})^2 = 0.1957$ znamená, že len 19.57% rozptylu možno vysvetliť lineárnym vzťahom medi poradím a príjmom hokejistov.

Test významnosti r_S

Rovnako ako v prípade Pearsonovho koeficienta korelácie, aj test významnosti Spearmanovho koeficienta umožňuje testovať, či medzi pozorovanými znakmi existuje štatisticky významná korelácia. Používa sa testovacia štatistika $T(\mathbf{X})$ ako v (5) ale len pre $n > 30$

$$r_S(\mathbf{P}, \mathbf{S}) = 1 - \frac{6}{n^2 - 1} \sum_{i=1}^{n} (p_i - q_i)^2 \tag{7}$$

Testuje sa $H_0 : \rho(\mathbf{P}, \mathbf{S}) = 0$ proti H_A :

- $\rho(\mathbf{P}, \mathbf{S}) \neq 0 : W^* = (-\infty, -t^*_\alpha\rangle \cup \langle t^*_\alpha, \infty)$
- $\rho(\mathbf{P}, \mathbf{S}) < 0 : W^* = (-\infty, -t^*_\alpha\rangle$
- $\rho(\mathbf{P}, \mathbf{S}) > 0 : W^* = \langle t^*_\alpha, \infty)$

Pre $n \leq 30$ sú kritické hodnoty t_α pre testovanie významnosti Spearmanovho korelačného koeficienta r_S tabelované.

Tabuľka kritických hodnôt pre testovanie r_S pre $n \leq 30$

n	$\alpha = 0,05$	$\alpha = 0,01$
5	0,9000	
6	0,8286	0,9429
7	0,7450	0,8929
8	0,6905	0,8571
9	0,6833	0,8167
10	0,6364	0,7818
11	0,6091	0,7545
12	0,5804	0,7273
13	0,5549	0,6978
14	0,5341	0,6747
15	0,5179	0,6536
16	0,5000	0,6324
17	0,4853	0,6152

n	$\alpha = 0,05$	$\alpha = 0,01$
18	0,4716	0,5975
19	0,4579	0,5825
20	0,4451	0,5684
21	0,4351	0,5545
22	0,4251	0,5426
23	0,4150	0,5306
24	0,4061	0,5200
25	0,3977	0,5100
26	0,3894	0,5002
27	0,3822	0,4915
28	0,3749	0,4828
29	0,3685	0,4744
30	0,3620	0,4665

Príklad 11.2 – pokračovanie

Budeme testovať $H_0 : \rho(\mathbf{P}, \mathbf{Q}) = 0$ proti $H_A : \rho(\mathbf{P}, \mathbf{Q}) \neq 0)$ na hladine významnosti $\alpha = 0.05$.

Máme $r_S(\mathbf{X}, \mathbf{Y}) = 0.4424$. V tabuľke nájdeme $r_S(\alpha) = 0.6364$. Zisťujeme, že

$$r_S(\mathbf{X}, \mathbf{Y}) \in W^* = (-\infty, -r_S(0.05)\rangle \cup \langle r_S(0.05), \infty)$$

a tak nulovú hypotézu zamietame.

Test významnosti nezamietol nulovú hypotézu, čo môžeme interpretovať tak, že korelácia medzi umiestnením hokejistov v kanadskom bodovaní a príjmom nie je štatisticky významná.

Ak existuje medzi premennými **X** a **Y** významná lineárna závislosť t.j. koeficient korelácie je štatisticky významne nenulový, bude nás zrejme zaujímať aj rovnica priamky, ktorá túto závislosť medzi premennými **X** a **Y** najlepšie vyjadruje.

V matematike poznáme lineárny vzťah ako súbor usporiadných dvojíc (x, y) ktoré ležia na priamke $y = ax + b$. V praktických úlohách, ktorými sa zaoberá štatistika, nie je medzi náhodnými premennými funkčne lineárny vzťah, pretože namerané hodnoty (x_i, y_i) neležia na priamke, ale môžu mať tendenciu ju vytvárať.

V prípade existencie štatistickej závislosti medzi znakmi X a Y bude stredná hodnota znaku **Y** závislá od toho, akú hodnotu nadobudol štatistický znak **X**. Ide o podmienenú strednú hodnotu $E(\mathbf{Y}|\mathbf{X} = x)$.

Regresnou funkciou štatistického znaku Y závislého od štatistického znaku X rozumieme funkciu $y(x) = E(\mathbf{Y}|\mathbf{X} = x)$.

Predpokladajme, že testujeme lineárnu závislosť medzi dvoma fyzikálnymi veličinami **X** a **Y**, medzi ktorými existuje lineárna závislosť

$$\mathbf{Y} = \beta_1 \mathbf{X} + \beta_0, \tag{8}$$

pričom parametre β_1 a β_2 sú neznáme. Preto urobíme experiment, kde meriame n dvojíc $(x_i, y_i), i = 1, 2, \ldots, n$. Meranie hodnôt x_i prebieha temer presne, zatiaľ čo y_i sa meria s chybou. Preto sa zavádzame štatistický model

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i, \text{ pre } i = 1, 2, \ldots, n, \tag{9}$$

kde

- y_i je i-ta hodnota premennej **Y** vo výberovom súbore,
- β_0 je veľkosť premennej **Y**, ak v základnom súbore $Y = 0$,
- β_1 je regresný koeficient udávajúci o koľko sa zmení y_i ak sa x_i zmení o 1 mernú jednotku,
- x_i je i-ta hodnota premennej **X** vo výberovom súbore,
- ε_i je náhodná chyba premennej **Y**, má rozdelenie $N(0, 1)$

Bodové odhady regresných koeficientov

Ak nahradíme regresné koeficienty β_0, β_1 (regresory) ich bodovými odhadmi b_0, b_1 získame odhad regresnej priamky v tvare vyrovnávacej priamky

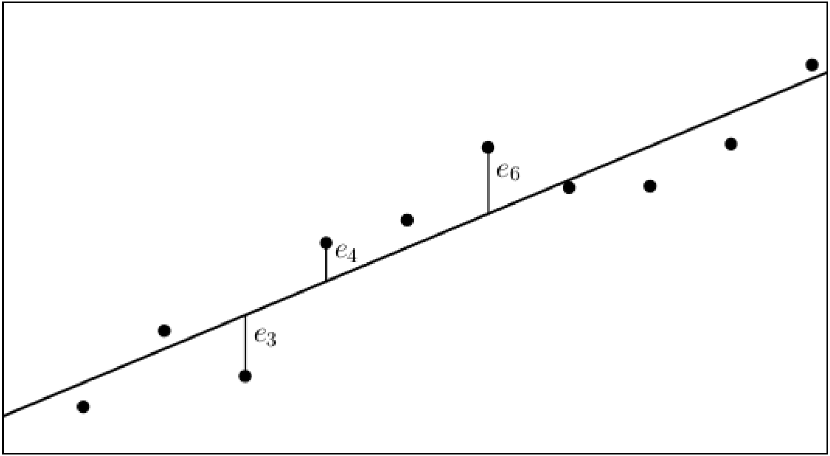
$$\hat{Y} = b_1 x + b_0. \tag{10}$$

Odhady musia byť také, aby vyrovnávacia priamka čo najlepšie aproximovala pozorované hodnoty závislej premennej $\mathbf{Y}$.

Na ich hľadanie sa najčastejšie používa metóda najmenších štvorcov, ktorá hľadá také riešenie, ktoré minimalizuje súčet štvorcov chýb nájdeného riešenia.

Chyby nájdeného riešenia sa nazývajú rezíduá získame takto:

$$e_i = y_i - \hat{Y}_i, \quad \text{pre } i = 1, 2, \ldots, n \tag{11}$$



Obr. 2: $e_i = y_i - \hat{Y}_i$

Metóda najmenších štvorcov

Hľadáme minimum funkcie

$$\psi = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Najskôr určíme sústavu rovníc

$$\frac{\partial \psi}{b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0,$$

$$\frac{\partial \psi}{b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0.$$

Po úpravách dostaneme bežne uvádzaný tvar rovníc

$$\sum_{i=1}^{n} y_i = n b_0 - b_1 \sum_{i=1}^{n} x_i, \tag{12}$$

$$\sum_{i=1}^{n} y_i x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \tag{13}$$

Z rovnice (12) dostanme b_0 pre odhad β_0 a po dosadení do (13) a úpravách b_1 pre odhad β_1 v tvare

$$b_0 = \overline{y} - b_1\overline{x} \tag{14}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{15}$$

Za povšimnutie stojí, že odhad regresnej priamky možno napísať v tvare

$$\hat{Y} = b_0 + b_1 x = \overline{y} - b_1\overline{x} + b_1 x = \overline{y} + b_1(x - \overline{x}),$$

odkiaľ vidíme, že regresná priamka prechádza bodom $(\overline{x}, \overline{y})$.

Overovanie modelu

Pri aplikácii metódy najmenších štvorcov platí vzťah

$$SS_Y = SS_{\hat{Y}} + SS_e,$$

kde

- $SS_Y = \sum_{i=1}^{n} (y_i - \overline{y})^2$ je celkový súčet štvorcov,
- $SS_{\hat{Y}} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{y})^2$ je súčet štvorcov modelu,
- $SS_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2$ je reziduálny súčet štvorcov.

Na posúdenie presnosti modelu potrebujeme odhad rozptylu náhodnej zložky σ^2 ktorej nestranným odhadom je štatistika reziduálny rozptyl

$$S_e = \frac{SS_e}{n-2}. \tag{16}$$

Celkový F-test

Celkový F-test nám umožňuje zistiť, či sme zvolili správny typ regresnej funkcie – regresnú priamku. Testuje hypotézu, či platí lineárny vzťah (8) medzi premennými.

Testujeme $H_0 : \beta_1 = 0$ proti $H_A : \beta_1 \neq 0$.

Testovaná štatistika má F rozdelenie

$$\frac{SS_{\hat{Y}}}{S_e} \sim F(1, n-2)$$

a p-hodnota $= 1 - F_0(x_{OBS})$.

Príklad 11.3

Pracovník personálneho oddelenia cíti, že existuje vzťah medzi počtom dní absencie v práci a vekom pracovníka. Náhodne vyberie pracovné záznamy 10 pracovníkov a získa údaje o ich veku v rokoch počte dní, v ktorých nenastúpili do práce.

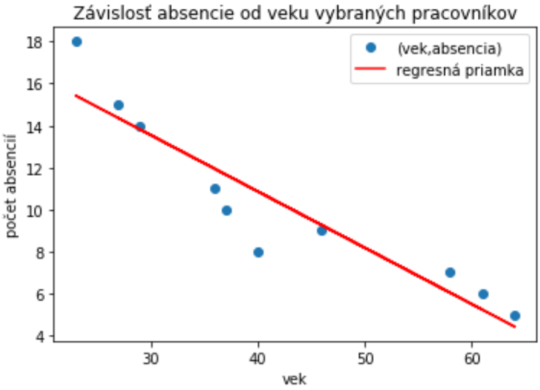
vek	27	61	37	23	46	58	29	36	64	40
absencie	15	6	10	18	9	7	14	11	5	8

Zobrazte regresnú priamku závislosti počtu dní absencie od veku pracovníkov a vypočítajte jej základné charakteristiky v jazyku Python.

```python
from scipy import stats

vek = np.array([27, 61, 37, 23, 46, 58, 29, 36, 64 , 40])
absencie = np.array([ 15, 6, 10, 18, 9, 7, 14, 11, 5, 8])
dataXY = {'X':vek,'Y':absencie}
df = pd.DataFrame(data = dataXY)
df.describe()
```

	X	Y
count	10.000000	10.000000
mean	42.100000	10.300000
std	14.670075	4.217688
min	23.000000	5.000000
25%	30.750000	7.250000
50%	38.500000	9.500000
75%	55.000000	13.250000
max	64.000000	18.000000



Obr. 3: Regresna priamka: absencie = 21.587 -0.268 * vek

Koeficient determinácie: 0.870
p-hodnota: 0.000
Štandartná chyba: 0.037
$b_1, b_0, r_{value}, p_{value}, std_{err} = stats.linregress(vek, absencie)$