8. SKÚMANIE ZÁVISLOSTI DVOCH KVANTITATÍVNYCH ZNAKOV

8.1 Štatistická závislosť

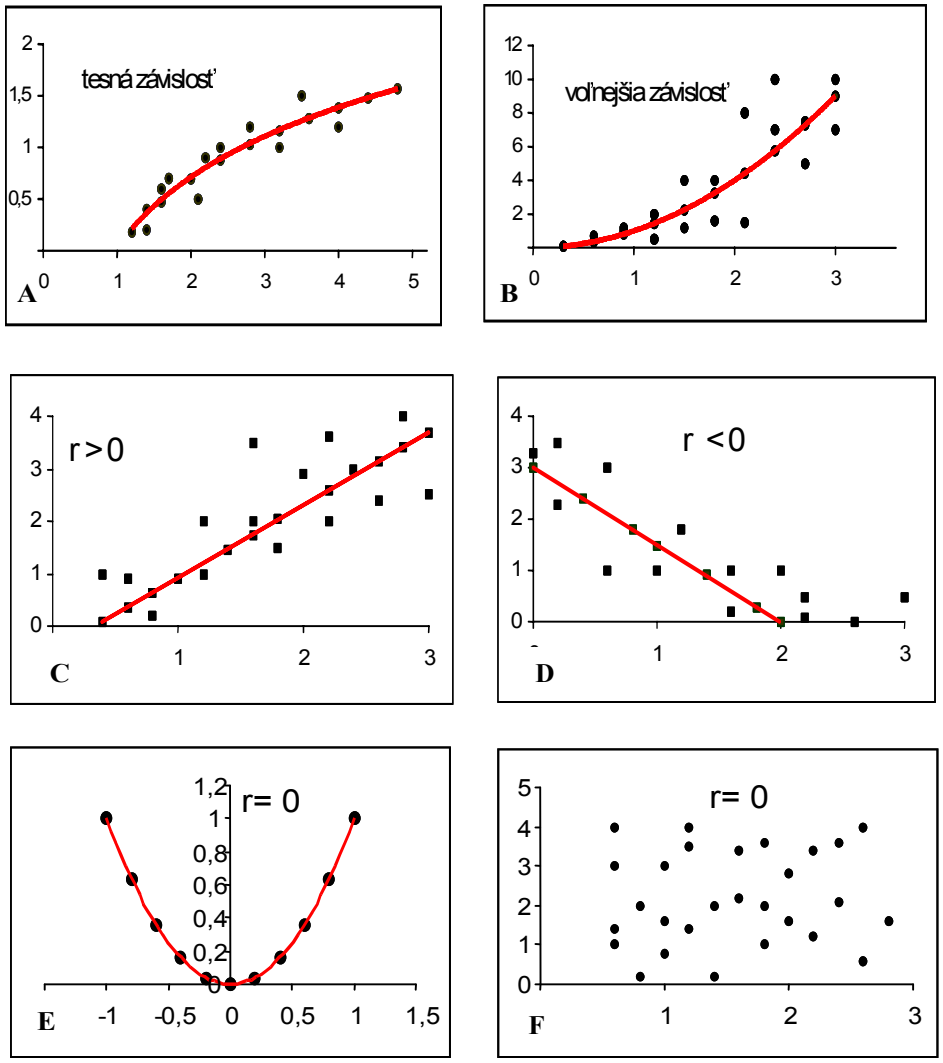
Dôležitá úloha všetkých technických, ekonomických i sociálnych oborov je hľadať a skúmať závislosť medzi premennými. Doteraz sme pracovali s funkčnými vzťahmi, kde závislá premenná y je jednoznačne určená funkciou $y = f(x)$ alebo $y = f(x_1, x_2, ..., x_n)$.

Často však, v dôsledku pôsobenia náhodných faktorov, alebo nezohľadňovania nejakého faktora, či v dôsledku nepresnosti merania má závisle premenná Y a jej pozorované hodnoty $y_1, y_2, ..., y_n$ povahu náhodnej veličiny, ktorá má isté rozdelenie pravdepodobnosti. Takáto závislosť sa volá **stochastická (štatistická) závislosť**. Nezávislé premenné môžu byť nenáhodné (fixné) alebo tiež náhodné veličiny. V tejto časti sa budeme zaoberať **jednoduchou (párovou) regresiou,** kde uvažujeme len jednu nezávislú premennú X s hodnotami $x_1, x_2, ..., x_n$.

Uvažujme závislosť ceny ojazdeného auta v autobazáre od veku auta. Zistíme, že autá s rovnakým vekom majú rôznu cenu. Preto cenu napríklad štvorročného auta považujeme za náhodnú premennú, jej rozdelenie sa volá podmienené rozdelenie. Kedy teda považujeme náhodné veličiny za štatisticky závislé? Rozdelenie početností jednej veličiny Y (kvantitatívneho znaku), ktoré zodpovedá istej, konkrétnej hodnote druhej veličiny X (kvantitatívneho znaku) sa volá **podmienené rozdelenie početností.** Ak pri zmenách hodnôt jedného znaku dochádza ku zmenám podmieneného rozdelenia početností druhého znaku, považujeme znaky za štatisticky **závislé.** A naopak, ak pri zmenách jedného znaku sa nemení rozdelenie druhého znaku, považujeme ich za **nezávislé.** O štatistickej závislosti možno hovoriť aj u kvalitatívnych znakov.

Elementárny spôsob grafického znázornenia závislosti dvoch kvantitatívnych znakov je **bodový diagram**. Zo znázornenia bodov $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

v rovine, kde (x_i, y_i) sú konkrétne hodnoty premenných X, Y namerané na i-tej štatistickej jednotke, možno zistiť charakteristické rysy závislosti. Obr. 8.1A ukazuje, že s narastajúcimi hodnotami premennej X rastú aj hodnoty premennej Y a navyše, že sa tento rast postupne spomaľuje. Schematicky znázorňuje túto tendenciu krivka preložená medzi bodmi. Voláme ju **regresná krivka**. Na Obr.8.1B s narastajúcim X rastú aj hodnoty Y, ale rast sa postupne zrýchľuje. Závislosti znázornené na obrázkoch majú teda rôzny priebeh.



Obr. 8.1 Rôzne druhy závislostí

Obrázky sa líšia ešte z iného hľadiska. Na Obr.8.1B sú jednotlivé body rozptýlené okolo regresnej krivky oveľa viac ako na Obr.8.1A. Medzi *X* a *Y* na Obr.8.1B je **voľnejšia závislosť** ako na Obr.8.1A. Obe závislosti sa líšia **silou** závislosti.
Pri skúmaní závislosti teda treba riešiť dve úlohy, ktoré spolu úzko súvisia:

- Posúdiť tesnosť závislosti pomocou nejakej charakteristiky, ktorá popisuje do akej miery premenná *X* vysvetľuje variabilitu premennej *Y* (**korelačná analýza)**.
- Charakterizovať priebeh tejto závislosti, to znamená, odhadnúť funkčný vzťah, podľa ktorého sa mení závislá premenná pri zmenách nezávisle premennej (**regresná analýza)**.

Podľa toho, koľko nezávislých premenných berieme do úvahy pri riešení týchto úloh, hovoríme

- o jednoduchej (párovej) korelácii a regresii, ak pracujeme len s jednou nezávislou premennou,
- o viacnásobnej (mnohonásobnej) korelácii a regresii, ak je počet nezávislých premenných väčší ako jeden.

Použitie viacnásobnej regresie síce vedie k presnejším odhadom, ale veľký počet premenných sťažuje analýzu úlohy i interpretáciu výsledkov. Preto v modeli treba uvažovať len tie premenné, ktoré majú zásadný vplyv na závislú premennú.
V celej tejto kapitole ide len o zisťovanie matematických súvislostí, ktoré nemôžeme zamieňať za vzťah príčiny a následku, lebo ani vysoký stupeň štatistickej závislosti nehovorí nič o príčinnej súvislosti javov. Väčšinou túto zdanlivú súvislosť spôsobuje tretí faktor, na ktorom sú oba pôvodné javy závislé. Pri zlej interpretácii môžeme dostať komické tvrdenia. Napríklad zistený vzťah medzi nízkou augustovou spotrebou plynu v kotolniach a vysokým predajom opaľovacích krémov ovplyvňuje tretí faktor - počasie.

8.2 Korelačná analýza

Vzťah medzi *X, Y* môže mať rôznu intenzitu, od úplnej nezávislosti až po úplnú funkčnú závislosť. Stupeň štatistickej závislosti sa dá popísať rôznymi mierami, my sa budeme venovať len **kovariancii a korelačnému koeficientu premenných *X, Y*.** Obe charakteristiky sú miery **lineárnej závislosti** premenných *X, Y*. **Kovariancia medzi *X, Y*** vo výberovom súbore s rozsahom *n* je číslo

$$\operatorname{cov} xy = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right). \tag{8.1}$$

Vzťah sa dá upraviť na jednoduchší tvar

$$\operatorname{cov} xy = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - \bar{x}\, y_i - x_i\, \bar{y} + \bar{x}\, \bar{y}\right) = \frac{1}{n}\sum x_i y_i - \frac{\bar{x}}{n}\sum y_i - \frac{\bar{y}}{n}\sum x_i + \frac{\bar{x}\,\bar{y}}{n}\sum 1 =$$

$$= \overline{xy} - \bar{x}\,\bar{y}. \tag{8.2}$$

Vlastnosti kovariancie:

- $\operatorname{cov} xy$ môže nadobudnúť ľubovoľnú reálnu hodnotu.
- $\operatorname{cov} xy = \operatorname{cov} yx$.
- Ak $\operatorname{cov} xy > 0$, premenné *X, Y* sú **priamo** lineárne závislé (Obr. 8.1C).
- Ak $\operatorname{cov} xy < 0$, premenné *X, Y* sú **nepriamo** lineárne závislé (Obr. 8.1D).
- Ak *X,Y* sú nezávislé, potom $\operatorname{cov} xy = 0$ (Obr. 8.1F).
- Kovariancia je mierou lineárnej závislosti, nehovorí nič o iných typoch závislosti. To, že $\operatorname{cov} xy = 0$ (hovoríme aj, že *X, Y* sú nekorelované) ešte neznamená, že *X, Y* sú nezávislé. Aj v prípade nulovej kovariancie môžu byť znaky **nelineárne** funkčne závislé (Obr. 8.1E).
- Nevýhodou kovariancie je, že jej hodnoty sú závislé na mierke, v ktorej sú vyjadrené *X , Y*. Preto vznikla veličina, ktorá tento nedostatok nemá, a to korelačný koeficient.

Korelačný koeficient je v základnom súbore označovaný $\rho_{x,y}$ a definovaný

$$\rho_{x,y} = \frac{\operatorname{cov} xy}{\sigma_x . \sigma_y}, \tag{8.3}$$

Ak použijeme namiesto základného súboru výberový súbor a kovarianciu výberového súboru a štandardné odchýlky výberového súboru

$$s_x = \sqrt{\frac{1}{n}\sum\left(x_i - \overline{x}\right)^2} \text{ a } \quad s_y = \sqrt{\frac{1}{n}\sum\left(y_i - \overline{y}\right)^2},$$

dostaneme bodový odhad (ale skreslený) korelačného koeficientu, ktorý sa volá **výberový korelačný koeficient** $r_{x,y}$:

$$r_{x,y} = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum\left(x_i - \overline{x}\right)^2} . \sqrt{\sum\left(y_i - \overline{y}\right)^2}} \tag{8.4}$$

Vlastnosti korelačného koeficientu:

- $\left|r_{xy}\right| \leq 1$.
- $r_{xy} = r_{yx}$, preto sa používa stručné označenie len r (alebo len ρ).
- Ak $r_{xy} > 0$, premenné *X,Y* sú priamo lineárne závislé (Obr. 8.1C).
- Ak $r_{xy} < 0$, premenné *X, Y* sú nepriamo lineárne závislé (Obr. 8.1D).
- Korelačný koeficient je mierou sily **lineárnej závislosti**, nehovorí nič o iných typoch závislosti. V prípade nulového korelačného koeficientu znaky sú **lineárne nezávislé**, môžu byť ale až nelineárne funkčne závislé, čo ilustruje Obr. 8.1E.
- Keď medzi premennými *X , Y* je funkčný **lineárny** vzťah $Y = B_0 + B_1 X$ ($B_1 \neq 0$), potom $r_{xy} = 1$ pre $B_1 > 0$, ($r_{xy} = -1$ pre $B_1 < 0$).
- Interpretácia konkrétnej hodnoty korelačného koeficientu závisí od povahy experimentálnych údajov a od rozsahu výberového súboru. Absolútna hodnota

korelačného koeficientu blízka jednotke znamená silnú závislosť, blízka nule slabú závislosť.

- Hodnota korelačného koeficientu je nezávislá na merných jednotkách.

Ak je výberový korelačný koeficient blízky nule, chceme overiť, či je nenulový len v dôsledku náhodného výberu. Uvedieme len jeden z mnohých testov pre testovanie korelačného koeficientu.

T-test **lineárnej nezávislosti premenných *X, Y*** overuje platnosť $H_0 : \rho = 0$ oproti alternatívnej hypotéze $H_1 : \rho \neq 0$.

Ekvivalentne možno formulovať test takto:

H_0 : Znaky sú lineárne nezávislé.

H_1 : Znaky sú lineárne závislé.

Tab. 8.1 T-test lineárnej nezávislosti

Hypotézy	**Použité rozdelenie**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
H_0: $\rho = 0$ H_1: $\rho \neq 0$	Studentovo	$T = r\sqrt{\frac{n-2}{1-r^2}}$	$\lvert t \rvert > t_{1-\alpha/2,}$ $d.f. = n-2$

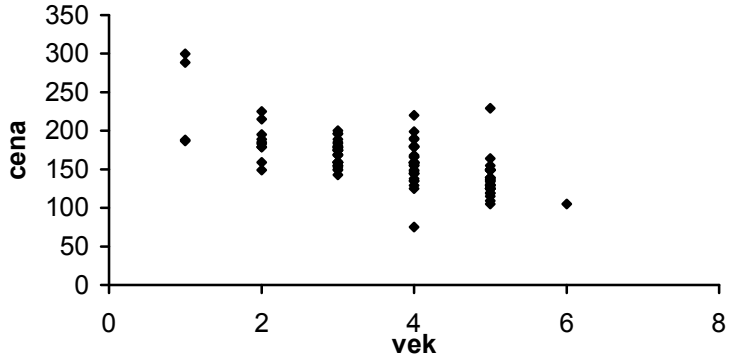
Príklad 8.1

V súbore **Autobazár** sú údaje o veku a cenách áut z 3 predajní autobazáru. Znázornite bodovým diagramom závislosť ceny od veku. Vyšetrite pomocou korelačného koeficientu a kovariancie závislosť ceny auta od veku, použite údaje zo všetkých 3 predajní. Na hladine spoľahlivosti $\alpha = 0{,}05$ otestujte nulovú hypotézu $H_0 : \rho = 0$, oproti alternatívnej hypotéze $H_1 : \rho \neq 0$.

EXCEL

Použitie EXCELu pri riešení korelačnej úlohy budeme ilustrovať na riešení Príkladu 8.1.

Po voľbe ***Vložiť' graf/ Závislosť'*** vytvoríme bodový diagram (Obr. 8.2). Z grafu vidieť, že s narastajúcim vekom mierne klesá cena áut. Po voľbe ***Nástroje/Analýza údajov/Korelácia*** a zadaní údajov sa objaví výstupná **korelačná matica.** Na jej uhlopriečke sú $r_{xx} = 1$ a $r_{yy} = 1$, a okrem toho výberový korelačný koeficient $r_{xy} = -0{,}6748$, čo predstavuje nepriamu miernu lineárnu závislosť, t.j. s narastajúcim vekom klesá cena auta.



Obr. 8.2 Bodový graf závislosti ceny áut od veku

Po voľbe ***Nástroje/Analýza údajov/Kovariancia*** ako výstup dostaneme ***kovariančnú maticu***, na jej uhlopriečke sú hodnoty $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ a $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ a $\operatorname{cov} xy = -25{,}044$. Rovnaké výsledky sa dajú získať aj postupom ***Prilepiť' funkciu/štatistické/*** **CORREL (COVAR)**.

Tab. 8.2 Korelačná matica

	cena	**vek**
cena	1	
vek	-0,67487	1

Tab. 8.3 Kovariančná matica

	cena	**vek**
cena	1080,48391	
vek	-25,044	1,27456

Na záver testujme hypotézu $H_0: \rho = 0$ oproti alternatívnej hypotéze $H_1: \rho \neq 0$.

Hodnota testovacej štatistiky je $t = -0{,}67487\sqrt{\frac{106-2}{1-(-0{,}67487)^2}} = -9{,}3265$. Porovnáme ju s kvantilom $t_{0{,}975;104} = 1{,}983035$ Studentovho rozdelenia. Platí $-9{,}3265 < -1{,}9830$, preto zamietame nulovú hypotézu a tvrdíme, že na hladine významnosti $\alpha = 0{,}05$ je $\rho \neq 0$, alebo že lineárna závislosť znakov je štatisticky významná.

8.3 Regresná analýza

Jednoduchá (párová) lineárna regresia

Úlohou regresnej analýzy pri skúmaní štatistickej závislosti *Y* na *X* je nájsť vhodný matematický model (funkciu), v ktorom je vyjadrená predstava o tejto závislosti. Ak by sa nám podarilo odstrániť spolupôsobenie vedľajších vplyvov na vzťah medzi *X* a *Y,* ležali by všetky body (x_i, y_i) na krivke s rovnicou $y = \eta(x)$, čo je deterministický model. Na premennú *Y* však vplývajú okrem *X* aj iné faktory, preto body (x_i, y_i) neležia na krivke, ale kolíšu okolo nej. To sa snažíme zachytiť aj v matematickom modeli. Preto každú hodnotu závisle premennej *Y* rozložíme na dve zložky, na deterministickú a náhodnú, t.j.

$$y_i = \eta(x_i) + \varepsilon_i, \quad i = 1,2,....,n.$$

Funkcia $\eta = \eta(x)$ sa volá **regresná funkcia.** Môže to byť napr. priamka $y = B_0 + B_1 x$, parabola $y = B_0 + B_1 x + B_2 x^2$ a iné známe funkcie. Náš model, ktorý zachytáva **lineárnu závislosť** *X*, *Y* bude **lineárna funkcia – regresná priamka.** Lineárny vzťah medzi *Y* a *X* v základnom súbore možno vyjadriť modelom

$$y_i = B_0 + B_1 x_i + \varepsilon_i \quad i = 1, 2, ... \tag{8.5}$$

kde y_i – *i*-ta hodnota premennej *Y* v základnom súbore,

B_0 – priesečník osi *y* s regresnou priamkou,

B_1 – **regresný koeficient** v základnom súbore, ktorý udáva o koľko sa zmení y, ak sa x zmení o jednu jednotku (je to smernica regresnej priamky),

x_i – *i*-ta hodnota premennej *X* v základnom súbore,

ε_i – *i*-ta **náhodná chyba** premennej *Y*.

Časť $B_0 + B_1 x_i = \eta_i$ je **deterministická časť modelu,** voláme ju **regresná funkcia**. Je to nám nedostupná teoretická priamka - **regresná priamka** v základnom súbore, okolo ktorej kolíšu skutočné hodnoty *Y* pre dané hodnoty *X*. Pretože k dispozícii máme len výberový súbor s rozsahom *n*, preložíme bodmi výberového súboru **vyrovnávajúcu regresnú priamku**, ktorú môžeme považovať za bodový odhad regresnej priamky v základnom súbore. Označíme ju vzťahom

$$\tilde{y}_i = b_0 + b_1 x_i, \; i = 1, 2, ..., n \tag{8.6}$$

kde $\tilde{y}_i$ - **očakávaná (vyrovnaná) hodnota premennej *Y*** pre danú hodnotu premennej *X,*

x_i - *i*-ta hodnota premennej X,

b_0 - bodový odhad koeficientu B_0,

b_1 - bodový odhad koeficientu B_1, volá sa **výberový regresný koeficient.**

Na výpočet neznámych koeficientov b_0 a b_1 v rovnici vyrovnávajúcej regresnej priamky sa používa **metóda najmenších štvorcov**. Označme rozdiely (chyby) medzi nameranými hodnotami y_i a medzi vyrovnanými hodnotami $\tilde{y}_i$, t.j. $y_i - \tilde{y}_i = e_i$ ako **rezíduá (reziduálne odchýlky)**. Sú to bodové odhady náhodných chýb ε_i regresného modelu. „ Najlepšie preložená“ priamka medzi bodmi (x_i, y_i) je tá, ktorá minimalizuje súčet štvorcov reziduálnych odchýlok

$$\sum_{i=1}^{n} e_i^2 = \sum (y_i - \tilde{y}_i)^2 \,. \tag{8.7}$$

To je podstata **metódy najmenších štvorcov**. Pri hľadaní koeficientov b_0 a b_1 využijeme skutočnosť, že hľadáme minimum funkcie dvoch premenných

$$f(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \tilde{y}_i)^2 = \sum [y_i - (b_0 + b_1 x)]^2 \,. \tag{8.8}$$

Vieme, že extrém funkcie tohto typu môže existovať len v stacionárnom bode funkcie, t.j. musí platiť

$$\frac{\partial f}{\partial b_0} = 0 \quad \text{a} \quad \frac{\partial f}{\partial b_1} = 0 \,.$$

Teda

$$\sum -2(y_i - b_0 - b_1 x_i) = 0 \tag{8.9}$$

$$\sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0 \tag{8.10}$$

Po úprave rovnice (8.9) dostaneme

$$\sum y_i - b_1 \sum x_i = b_0 n$$

odtiaľ $$b_0 = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} = \bar{y} - b_1 \bar{x} \,.$$

Úpravou rovnice (8.10) dostaneme

$$\sum x_i y_i - b_0 \sum x_i = b_1 \sum x_i^2$$

$$\sum x_i y_i - (\bar{y} - b_1 \bar{x}) \sum x_i = b_1 \sum x_i^2$$

$$\sum x_i y_i - \bar{y}\sum x_i = b_1(\sum x_i^2 - \bar{x}\sum x_i).$$

Po vynásobení poslednej rovnice výrazom $1/n$ a úprave

$$\frac{\sum x_i y_i}{n} - \bar{x}.\bar{y} = b_1\left[\frac{\sum x_i^2}{n} - \left(\bar{x}\right)^2\right] \quad \Rightarrow \quad b_1 = \frac{\overline{xy} - \bar{x}.\bar{y}}{s_x^2} = \frac{\mathrm{cov}\, xy}{s_x^2} = r\frac{s_y}{s_x}$$

$$b_1 = r\frac{s_y}{s_x} \tag{8.11}$$

$$b_0 = \bar{y} - b_1\bar{x}\ . \tag{8.12}$$

Vyrovnávajúca regresná priamka má rovnicu $\tilde{y} = b_0 + b_1 x$

$$\tilde{y} = \left(\bar{y} - \bar{x}\, r\frac{s_y}{s_x}\right) + \left(r\frac{s_x}{s_y}\right)x$$

čo po úprave je

$$\tilde{y} - \bar{y} = r\frac{s_y}{s_x}\left(x - \bar{x}\right) \tag{8.13}$$

$$\tilde{y} - \bar{y} = b_1\left(x - \bar{x}\right) \tag{8.14}$$

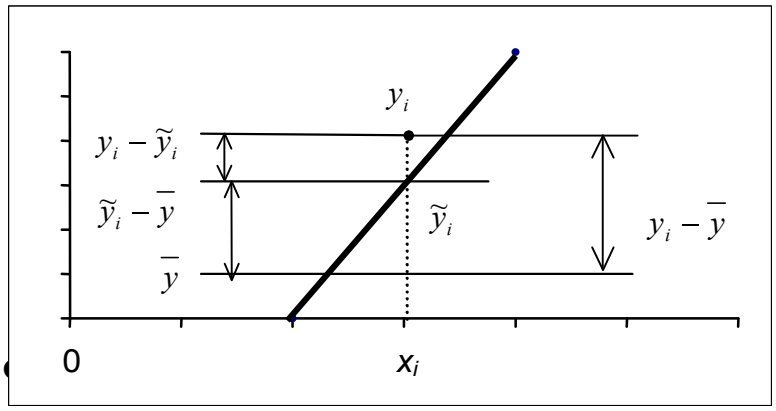
Nebudeme sa zdržiavať dôkazom, že v tomto stacionárnom bode má funkcia skutočne lokálne minimum. Teoretickú regresnú priamku sme odhadli priamkou $\tilde{y} = b_0 + b_1 x$, ktorú považujeme za **bodový odhad** neznámej regresnej priamky .

Poznámka 8.1

Na konštrukciu koeficientov b_0 a b_1 nemôžeme použiť len súčet chýb e_i, lebo vždy platí $\sum_{i=1}^{n} e_i = 0$, aj pre zle zvolenú regresnú priamku. Všimnime si ešte dve vlastnosti regresnej priamky. Regresná priamka prechádza bodom $\left(\bar{x}, \bar{y}\right)$ a regresný koeficient má vždy rovnaké znamienko ako korelačný koeficient.

8. 4 Skúmanie štatistickej významnosti modelu

Po nájdení rovnice regresnej priamky treba overiť, či tento model je „kvalitný“, či dobre vystihuje závislosť medzi *X, Y*. Pri riešení regresnej úlohy prichádza často do úvahy viacero typov regresných funkcií (kvadratická, logaritmická), preto sa skúma, ktorá z týchto funkcií „lepšie prilieha“ výberovým údajom. To sa dá merať rôznymi charakteristikami: **reziduálny súčet štvorcov, reziduálny rozptyl, štandardná odchýlka reziduí, koeficient determinácie** alebo preveriť rôznymi **testami**.



Obr. 8.3 Rozklad celkovej variability premennej Y

Na Obr. 8.3 je jasný vzťah:

$$\left(y_i - \bar{y}\right) = \left(\tilde{y}_i - \bar{y}\right) + \left(y_i - \tilde{y}_i\right), \tag{8.15}$$

t.j. odchýlka od celkového priemeru = odchýlka vysvetlená regresiou + odchýlka nevysvetlená regresiou (reziduálna). Prekvapivo platí aj

$$\sum \left(y_i - \bar{y}\right)^2 = \sum_i \left(\tilde{y}_i - \bar{y}\right)^2 + \sum_i \left(y_i - \tilde{y}_i\right)^2, \tag{8.16}$$

$$SSY = SSR + SSE \tag{8.17}$$

SSY - je **celková variabilita** premennej *Y* (celkový súčet štvorcov, sum of squares total),

SSR - je **variabilita vysvetlená regresným modelom** (sum of squares due to regression),

SSE - je **variabilita nevysvetlená regresným modelom, reziduálny súčet štvorcov** (sum of squares due to error).

Dokážeme vlastnosť (8.16). Po umocnení výrazu (8.15) a sčítaní pre všetky $i = 1,2,...,n$ dostaneme

$$\sum \left(y_i - \bar{y}\right)^2 = \sum_i \left(\tilde{y}_i - \bar{y}\right)^2 + \sum_i \left(y_i - \tilde{y}_i\right)^2 + 2\sum_{i=1} \left(\tilde{y}_i - \bar{y}\right)\left(y_i - \tilde{y}_i\right) .$$

Hodnota posledného sčítanca je nula, lebo

$$\sum \left(y_i - \tilde{y}_i\right)\tilde{y}_i - \bar{y}\sum \left(y_i - \tilde{y}_i\right) = \sum \left(y_i - b_0 - b_1 x_i\right)\left(b_0 + b_1 x_i\right) - \bar{y}\sum \left(y_i - b_0 - b_1 x_i\right)$$

$$= b_0 \sum \left(y_i - b_0 - b_1 x_i\right) + b_1 \sum \left(y_i - b_0 - b_1 x_i\right)\left(x_i\right) - \bar{y}\sum \left(y_i - b_0 - b_1 x_i\right) = 0 ,$$

pričom sme použili vzťahy (8.9) a (8.10), t.j. parciálne derivácie

$$\frac{\partial f}{\partial b_0} = 0 \text{ a } \frac{\partial f}{\partial b_1} = 0 .$$

Porovnanie zložiek SSY, SSR, SSE je jedna možnosť, ako posúdiť štatistickú významnosť modelu ako celku:

- Pri funkčnej závislosti je *SSE* = 0, *SSY* = *SSR*, lebo všetky body y_i ležia na vyrovnávajúcej priamke.
- Pri nezávislosti je *SSR* = 0, *SSY* = *SSE*, lebo vyrovnávajúca priamka je rovnobežná s osou *x* a prechádza napríklad bodom $\left(x_1, \bar{y}\right)$.
- Závislosť *X, Y* je tým silnejšia, čím je väčší podiel variability *SSR* na celkovej variabilite *SSY*. Sila tejto lineárnej závislosti sa meria **výberovým koeficientom determinácie,** ktorý je definovaný

$$r^2 = \frac{SSR}{SSY}; \qquad r^2 \in \langle 0,1 \rangle . \tag{8.18}$$

Lineárny vzťah medzi *X,Y* je tak vysvetlený na $\frac{SSR}{SSY}.100$ %, preto je z viacerých modelov „kvalitnejší“ model s vyšším koeficientom determinácie. Výberový koeficient determinácie r^2 je bodovým odhadom **koeficientu determinácie** ρ^2 v

základnom súbore, ale skresleným. Neskreslený odhad dáva **korigovaný koeficient determinácie**

$$r_{adj}^2 = 1 - \left(1 - r^2\right)\frac{n-1}{n-2}. \tag{8.19}$$

Koeficient determinácie $r^2 = \frac{SSR}{SSY}$ (8.18) je druhá mocnina korelačného koeficientu r (8.4), ktorý bol definovaný v časti 8.2. Dokážeme toto tvrdenie. Využijeme rovnicu vyrovnávajúcej regresnej priamky (8.14)

$$\tilde{y}_i - \bar{y} = b_1\left(x_i - \bar{x}\right).$$

Po umocnení a sčítaní pre $i = 1,2,...,n$ platí

$$\sum\left(\tilde{y}_i - \bar{y}\right)^2 = b_1^2 \sum\left(x_i - \bar{x}\right)^2.$$

Po dosadení tohto vzťahu do $SSY = SSR + SSE$ dostaneme:

$$\sum\left(y_i - \bar{y}\right)^2 = b_1^2 \sum\left(x_i - \bar{x}\right)^2 + \sum\left(y_i - \tilde{y}_i\right)^2. \tag{8.20}$$

Podľa (8.11), kde r je korelačný koeficient, platí $b_1 = r\frac{s_y}{s_x}$ t.j.

$$b_1^2 = r^2\frac{s_y^2}{s_x^2} = r^2\frac{\sum\left(y_i - \bar{y}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}$$

a po dosadení do (8.20)

$$SSY = r^2\sum\left(y_i - \bar{y}\right)^2 + SSE$$

$$SSY = r^2 SSY + SSE$$

$$r^2 = \frac{SSY - SSE}{SSY} = \frac{SSR}{SSY}.$$

Cieľom metódy najmenších štvorcov bolo minimalizovať variabilitu nevysvetlenú regresným modelom, hodnotu $SSE = \sum\left(y_i - \tilde{y}_i\right)^2$, ktorá sa volá aj **reziduálny súčet štvorcov**. Z dvoch modelov, ktoré by teoreticky prichádzali do úvahy, je lepší ten, kde je menší SSE. Mierou variability hodnôt y_i okolo vyrovnávajúcej regresnej priamky je **štandardná odchýlka rezíduí**

$$s_{rez} = \sqrt{\frac{\sum (y_i - \tilde{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}. \qquad (8.21)$$

Je to neskreslený bodový odhad štandardnej odchýlky náhodných chýb ε_i v základnom súbore. Jej druhá mocnina s_{rez}^2 sa nazýva **reziduálny rozptyl**.

Poznámka 8.2

Koeficient determinácie, štandardná odchýlka rezíduí, korigovaný koeficient determinácie tvoria výstup EXCELu po procedúre ***Regresia***.

8.5 Testy hypotéz používané pri voľbe regresnej funkcie

a) test linearity (celkový F-test)

Na začiatku našich úvah sa pýtame, či vôbec medzi premennými *X* a *Y* existuje lineárna závislosť. Ak empirické údaje zobrazíme bodovým diagramom a body $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ ležia v páse, ktorý sa dá približne ohraničiť dvomi priamkami, ktoré nie sú rovnobežné s osou *x*, môžeme predpokladať lineárnu závislosť medzi *X* a *Y*. Preto sformulujeme nulovú a alternatívnu hypotézu takto:

H_0 : Lineárny model nie je štatisticky významný (t.j. *X,Y* nie sú lineárne závislé).

H_1 : Lineárny model je štatisticky významný (t.j. *X,Y* sú lineárne závislé).

Na overenie platnosti H_1 použijeme známu analýzu rozptylu tak, že odhad s_y^2 celkového rozptylu σ^2 závisle premennej Y rozložíme na dve zložky:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{SSE}{n-1} = \frac{1}{n-1}\left[\sum_i (y_i - \tilde{y}_i)^2 + \sum_i (\tilde{y}_i - \bar{y})^2 \right]$$

$$= \frac{1}{n-1}\left[SSE + SSR \right], \text{ t.j.}$$

$$(n-1)s_y^2 = SSE + SSR$$

Náhodné premenné $\frac{(n-1)s_y^2}{\sigma^2}$, $\frac{SSE}{\sigma^2}$, $\frac{SSR}{\sigma^2}$ majú χ^2 – rozdelenia postupne s $(n-1)$, $(n-2)$ a 1 stupňom voľnosti. Podiel rozptylov $F = \frac{SSR/1}{SSE/n-2} = \frac{MSR}{MSE}$ má Fisherovo rozdelenie s $(1, n-2)$ stupňami voľnosti, kde

MSR - priemerný štvorec regresie (mean square of regression),

MSE - priemerný štvorec chýb (mean square of errors).

Podstata testu je v tom, že sme našli náhodnú premennú, ktorá je funkciou *SSR* a *SSE* a ktorej rozdelenie poznáme. Model je tým lepší, čím je väčšie číslo *F*, preto veľké hodnoty testovacej štatistiky *F* hovoria v prospech alternatívnej hypotézy, teda padnú do oboru zamietnutia H_0.

Záver: F – test je len jednostranný test (pravostranný). Nulovú hypotézu zamietame, ak pri zvolenej hladine významnosti α je hodnota testovacej štatistiky $F > F_{1-\alpha(1,n-2)}$, kde $F_{1-\alpha}$ je príslušný kvantil F-rozdelenia s $(1, n-2)$ stupňami voľnosti. V tomto prípade teda **prijímame alternatívnu hypotézu o lineárnom vzťahu medzi *X* a *Y***. Nájdená regresná priamka je vhodný typ funkcie na vyjadrenie priebehu závislosti.

Tab. 8.4 Celkový F-test

Hypotézy	Použité rozdelenie	Testovacia štatistika	Oblasť zamietnutia H_0
H_0 : *X,Y* sú lineárne nezávislé. H_1 : *X,Y* sú lineárne závislé.	Fisherovo	$F = \frac{SSR/1}{SSE/n-2}$	$F > F_{1-\alpha}$ *d.f.* $= (1, n-2)$

b) t-test o lineárnej nezávislosti *X, Y*

Tento test je založený na nasledujúcej myšlienke. Regresný koeficient B_1 je smernica regresnej priamky a vyjadruje priemernú zmenu *Y* pri zmene *X* o jednu jed-

notku. Ak $B_1 = 0$, regresná priamka je rovnobežná s osou *x*, teda aj po zmene nezávisle premennej *X* sa nemenia hodnoty *Y* (presnejšie podmienené stredné hodnoty). Preto sa nedá hovoriť o lineárnej závislosti *X, Y*.

Ak je výberový regresný koeficient b_1 blízky nule, treba overiť hypotézu, či koeficient B_1 je rôzny od nuly, t.j. overiť hypotézu, či medzi *X* a *Y* existuje lineárna závislosť.

H_0: $B_1 = 0$ (t.j. *X,Y* sú lineárne nezávislé.)

H_1: $B_1 \neq 0$ (t.j. *X,Y* sú lineárne závislé.)

Na testovanie použijeme testovaciu štatistiku $T = \frac{b_1 - B_1}{s(b_1)}$, ktorá má Studentovo rozdelenie s $(n-2)$ stupňami voľnosti a $s(b_1) = \frac{s_{rez}}{\sqrt{\sum_i (x_i - \overline{x})^2}}$ je štandardná odchýlka koeficientu b_1. Ak platí nulová hypotéza, vypočítame hodnotu testovacej štatistiky $T = \frac{b_1}{s(b_1)}$.

Záver: Nulovú hypotézu zamietame, ak pri zvolenej hladine významnosti α je hodnota testovacej štatistiky $|t| > t_{1-\alpha/2,(n-2)}$, kde $t_{1-\alpha/2}$ je kvantil Studentovho rozdelenia s $(n-2)$ stupňami voľnosti. V tomto prípade teda **prijímame alternatívnu hypotézu o lineárnom vzťahu medzi *X* a *Y***.

Tab. 8.5 T-test o lineárnej nezávislosti

Hypotézy	**Použité rozdelenie**	**Testovacia štatistika**	**Oblasť zamietnutia H_0**
H_0: $B_1 = 0$ H_1: 1. $B_1 \neq 0$ 2. $B_1 > 0$ 3. $B_1 < 0$	Studentovo	$T = \frac{b_1}{s(b_1)}$ $s(b_1) = \frac{s_{rez}}{\sqrt{\sum_i (x_i - \overline{x})^2}}$	1. $\|t\| > t_{1-\alpha/2,}$ 2. $t > t_{1-\alpha}$ 3. $t < -t_{1-\alpha}$ $d.f. = n-2$

Podobne sa dá testovať hypotéza $H_0 : B_0 = 0,\ H_1 : B_0 \neq 0$.

Test lineárnej závislosti vieme urobiť tromi ekvivalentnými spôsobmi, posledné dva sú aj výstupom EXCELu :

- testovať korelačný koeficient
- celkový F-test
- testovať regresný koeficient

EXCEL

Na riešenie regresnej úlohy ponúka EXCEL nasledujúce prostriedky.

Po voľbe ***Nástroje/ Analýza údajov / Regresia*** a zadaní údajov najskôr pre závislú *Y*, potom pre nezávislú premennú *X* sa v tabuľkách objavia údaje Tab. 8.6, pričom niektoré sú zle pomenované.

Tab. 8.6 Regresná štatistika

pomenovanie	skutočný význam
Násobné R	$\lvert r \rvert$ - absolútna hodnota r
Hodnota spoľahlivosti	r^2 - koef. determinácie
Nastavená hodnota spoľahlivosti	upravený koef. determinácie
Chyba strednej hodnoty	s_{rez}
Pozorovania	n

Tabuľka ANOVA poskytuje rozklad celkového rozptylu na dve zložky a celkový F-test .

Tab. 8.7 ANOVA

	stupne voľnosti	*SS*	*MS*	F	významnosť F p- hodnota
Regresia	1	*SSR*	*MSR*=*SSR*/1	hodnota testovacej štatistiky	H_0 : Lineárny model nie je štatisticky významný.
Rezíduá	*n*-2	*SSE*	*MSE*=*SSE*/*n*-2 s^2_{rez}		
Celkom	*n*-1	*SSY*			

V poradí tretia Tab. 8.8 okrem koeficientov regresnej priamky obsahuje aj t-test pre nulovosť regresného koeficientu B_1 (druhý riadok) a koeficientu B_0 (prvý riadok).

Tab. 8.8 Testovanie koeficientov regresnej priamky

	koeficienty	chyba strednej hodnoty	t-stat	p-hodnota	dolný 95%	horný 95%	dolný 99%	horný 99%
hranice	b_0	$s(b_0)$	$b_0/s(b_0)$	H_0: B_0=0	intervaly spoľahlivosti pre B_0			
X	b_1	s(b_1)	$b_1/s(b_1)$	H_0: B_1=0	intervaly spoľahlivosti pre B_1			

Posledná tabuľka obsahuje aj pre každý prvok x_i výberového súboru vypočítanú očakávanú hodnotu $\tilde{y}_i$ a aj rezíduum $e_i = y_i - \tilde{y}_i$.

Príklad 8.2

V súbore **Autobazár** sú údaje o veku a cenách 106 áut z 3 predajní autobazáru. Vyšetrite lineárnu závislosť ceny auta od veku, použite údaje zo všetkých 3 predajní, nájdite rovnicu regresnej priamky, na hladine významnosti $\alpha = 0{,}05$ otestujte štatistickú významnosť lineárneho modelu.

V príklade po voľbe ***Nástroje/ Analýza údajov / Regresia*** dostaneme nasledujúce výstupné tabuľky:

Regresní statistika	
Násobné R	0,675
Hodnota spolehlivosti R	0,455
Nastavená hodnota spolehlivosti F	0,450
Chyba stř. hodnoty	24,489
Pozorování	106

	Koeficienty	Chyba stř. hodnoty	t stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 99%	Horní 99%
Hranice	233,951	8,279	28,257	1,325E-50	217,533	250,369	212,226	255,676
vek	-19,649	2,107	-9,327	2,1494E-15	-23,827	-15,471	-25,178	-14,121

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	52163,363	52163	86,98364	2,14942E-15
Rezidua	104	62367,931	599,7		
Celkem	105	114531,294			

Z tabuliek vyplýva:

- Absolútna hodnota korelačného koeficientu je $|r| = 0{,}675$, regresný koeficient je (–19,649). Korelačný koeficient má rovnaké znamienko ako regresný koeficient, preto je korelačný koeficient $r = -0{,}675$, čo interpretujeme ako nepriamu, miernu lineárnu závislosť.
- Koeficient determinácie je $r^2 = 0{,}455$, tzn. len 45,5 % variability ceny áut sa dá vysvetliť lineárnym vzťahom s vekom áut.
- Neskreslený odhad koeficientu determinácie v základnom súbore je číslo 0,4502.
- p-hodnota pre celkový F-test je $2{,}14 \cdot 10^{-15}$, čo je veľmi malé číslo. Na všetkých bežných hladinách významnosti zamietame nulovú hypotézu, prijímame alternatívnu hypotézu, že daný model je štatisticky významný, t.j. premenné sú lineárne závislé.
- **Rovnica vyrovnávajúcej regresnej priamky** je $y = -19{,}649x + 233{,}95$.

- $s_{rez} = 24{,}489$, tzn. skutočné ceny áut sa odchyľujú od hodnôt regresnej priamky približne o $\pm 24{,}5$ tisíc korún.
- p-hodnota pri t-teste hypotézy $H_0 : B_1 = 0$ oproti $H_1 : B_1 \neq 0$ je to isté malé číslo $2{,}14 \cdot 10^{-15}$, preto na každej bežnej hladine významnosti prijímame alternatívnu hypotézu, že premenné sú lineárne závislé.

8.6 Použitie regresnej priamky

Regresnú priamku $\tilde{y} = b_0 + b_1 x$ považujeme za **bodový odhad** strednej hodnoty závisle premennej Y a môžeme ju použiť na

- bodový odhad hodnoty Y pre jednu konkrétnu hodnotu X,
- bodový odhad priemernej hodnoty Y pre istú úroveň znaku X, ale len na intervale $\langle x_{min}, x_{max} \rangle$.

Napríklad, môžeme očakávať, že cena jedného 2 ročného auta je

$$y = 233{,}95 - 19{,}65.2 = 194{,}65 \text{ tisíc Sk.}$$

Je to zároveň priemerná cena všetkých dvojročných áut.

Poznámka 8.3

Vieme však určiť aj :

- 100(1-α)% interval spoľahlivosti pre koeficient B_0 :

$$b_0 - t_{1-\alpha/2,(n-2)}.s(b_0) < B_0 < b_0 + t_{1-\alpha/2,(n-2)}.s(b_0) ,$$

- 100(1-α)% interval spoľahlivosti pre koeficient B_1:

$$b_1 - t_{1-\alpha/2,(n-2)}.s(b_1) < B_1 < b_1 + t_{1-\alpha/2,(n-2)}.s(b_1),$$

- 100(1-α)% interval spoľahlivosti pre **priemernú hodnotu *Y* v základnom súbore pre danú konkrétnu hodnotu** x_i, označíme ju $\mu(y/x_i)$ a platí

$$(b_0 + b_1 x_i) - t_{1-\alpha/2,(n-2)}.s_i < \mu(y/x_i) < (b_0 + b_1 x_i) + t_{1-\alpha/2,(n-2)}.s_i ,$$

kde $s_i = s_{rez}.\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}$ je štandardná odchýlka vyrovnaných hodnôt $\tilde{y}_i = b_0 + b_1 x_i$. Šírka tohto intervalu je iná pre každé x_i a rozširuje sa so vzďaľovaním x_i od $\bar{x}$.

- 100(1-α)% interval spoľahlivosti pre **individuálnu hodnotu *Y* v základnom súbore pre danú konkrétnu hodnotu** x_i **,** označíme ju $Y(x_i)$ a platí

$$(b_0 + b_1 x_i) - t_{1-\alpha/2,(n-2)}.s_i < Y(x_i) < (b_0 + b_1 x_i) + t_{1-\alpha/2,(n-2)}.s_i ,$$

kde $s_i = s_{rez}.\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}+1}$ je štandardná odchýlka individuálnych hodnôt premennej *Y*. Šírka tohto intervalu je väčšia ako pre odhad stabilnejšej priemernej hodnoty.

EXCEL

Z výstupných tabuliek pre ***Regresiu*** sa dá pre náš Príklad 8.2 určiť :

- 95% interval spoľahlivosti pre B_0 : $217{,}53 < B_0 < 250{,}36$,
- 95% interval spoľahlivosti pre B_1 : $-23{,}83 < B_1 < -15{,}47$,
- Podobne z tabuliek sú známe 99% intervaly spoľahlivosti pre B_0 a B_1.

95% interval spoľahlivosti pre priemernú cenu 2 – ročného auta musíme dopočítať bez EXCELu.

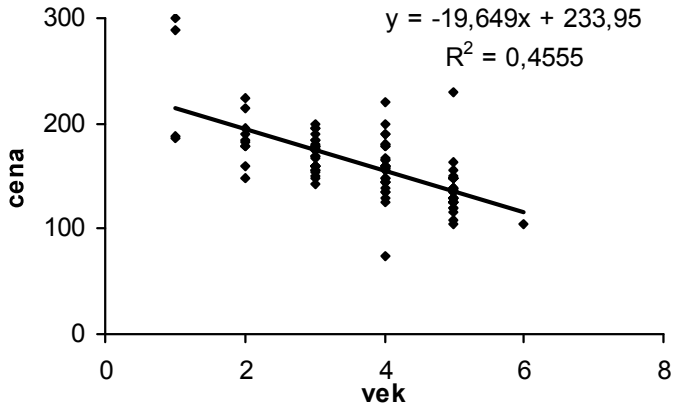
$$\bar{x} = 3{,}764\,,\ (2-\bar{x})^2 = 3{,}112\,,\ \sum_i (x_i-\bar{x})^2 = 135{,}104\ \sqrt{\frac{1}{n}+\frac{(2-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}} = 0{,}180$$

$$s_{rez} = 24{,}489\,,\quad t_{1-\alpha/2,(n-2)} = 1{,}983\,,\quad \tilde{y}_2 = b_0 + b_1.2 = 194{,}652$$

maximálne prípustná chyba odhadu je teda $t_{1-\alpha/2,(n-2)}.s_i = 8{,}750$, preto 95% interval spoľahlivosti pre priemernú cenu 2 – ročného auta je $194{,}652 \pm 8{,}750$ tisíc Sk.

Poznámka 8.4

Úpravami bodového diagramu v EXCELi sa dá do grafu vložiť regresná krivka, rovnica regresnej krivky i koeficient determinácie. Klikneme pravým tlačidlom na niektorý bod grafu, zvolíme ***Pridať trendovú čiaru***, na záložke ***Typ/lineárny*** a na záložke ***Možnosti/zobraziť v grafe rovnicu regresnej priamky a R^2***.



Obr. 8.4 Úpravy bodového grafu

8.7 Predpoklady pre použitie metódy najmenších štvorcov

Metóda najmenších štvorcov dáva neskreslený odhad regresnej priamky pri splnení istých predpokladov o rozdelení pravdepodobnosti náhodných chýb ε_i v modeli $y_i = B_0 + B_1 x_i + \varepsilon_i$.

Sú to tieto predpoklady:

- Stredná hodnota náhodných chýb je nula.
- Rozptyl náhodných chýb je konštantný.
- Rozdelenie pravdepodobnosti náhodných chýb je normálne .
- Náhodné chyby sú medzi sebou vzájomne nezávislé.

Splnenie týchto predpokladov sa dá overiť až po zvolení regresného modelu, lebo až vtedy sú známe rezíduá, ktoré sú odhadmi náhodných chýb. Ich splnenie sa približne overí graficky zostrojením histogramu rozdelenia rezíduí a z bodového diagramu hodnôt $(\tilde{y}_i , e_i)$. Podrobnejšie sa tomuto problému nebudeme venovať (pozri [6]).

8.8 Iné typy regresných funkcií

Lineárna regresná funkcia je vďaka ľahkej interpretácii preferovaná pred inými typmi, ale niekedy z povahy problému vyplýva, že pre popis danej závislosti by bola vhodnejšia iná regresná funkcia. Uvedieme niektoré iné modely.

Parabolická regresia Regresná funkcia je tvaru $\eta = B_0 + B_1 x + B_2 x^2$, bodové odhady koeficientov získame priamo použitím metódy najmenších štvorcov, t.j. hľadaním minima funkcie troch premenných $\sum_{i=1}^{n} e_i^2 = \sum (y_i - \tilde{y}_i)^2$, kde

$$\tilde{y}_i = b_0 + b_1 x_i + b_2 x_i^2 .$$

Zovšeobecnením môže byť polynomická regresia vyššieho stupňa, v praxi sa stretávame s polynómami maximálne 3. a 4. stupňa.

Hyperbolická regresia Regresná funkcia je tvaru $\eta = B_0 + \frac{B_1}{x}$. Bodové odhady koeficientov získame tiež priamo použitím metódy najmenších štvorcov.

Logaritmická regresia Regresná funkcia je tvaru $\eta = B_0 + B_1 \log x$. Bodové odhady koeficientov získame priamo použitím metódy najmenších štvorcov.

Exponenciálna regresia Regresná funkcia je tvaru $\eta = B_0 \cdot B_1^x$. Bodové odhady koeficientov sa nedajú získať priamo použitím metódy najmenších štvorcov. Vhodnou úpravou (transformáciu) regresnej funkcie ju upravíme na taký tvar, kde funkcie jej parametrov sa dajú odhadnúť metódou najmenších štvorcov. V tomto prípade logaritmovaním dostaneme :

$$\ln \eta = \ln B_0 + x \ln B_1$$

a budeme hľadať minimum funkcie $\sum (\ln y_i - \ln b_0 - x_i \ln b_1)^2$.

Známym spôsobom použijeme parciálne derivácie tejto funkcie a ako riešenie sústavy získame $\ln b_0$ a $\ln b_1$.

Poznámka 8.5

Podobne postupujeme aj v prípade iných typov funkcií, ktorým sa však nebudeme venovať.

Pri výbere vhodného typu regresnej funkcie sa v EXCELi orientujeme podľa hodnoty koeficientu determinácie, ktorý je definovaný nezávisle na type regresnej funkcie.

Príklad 8.3

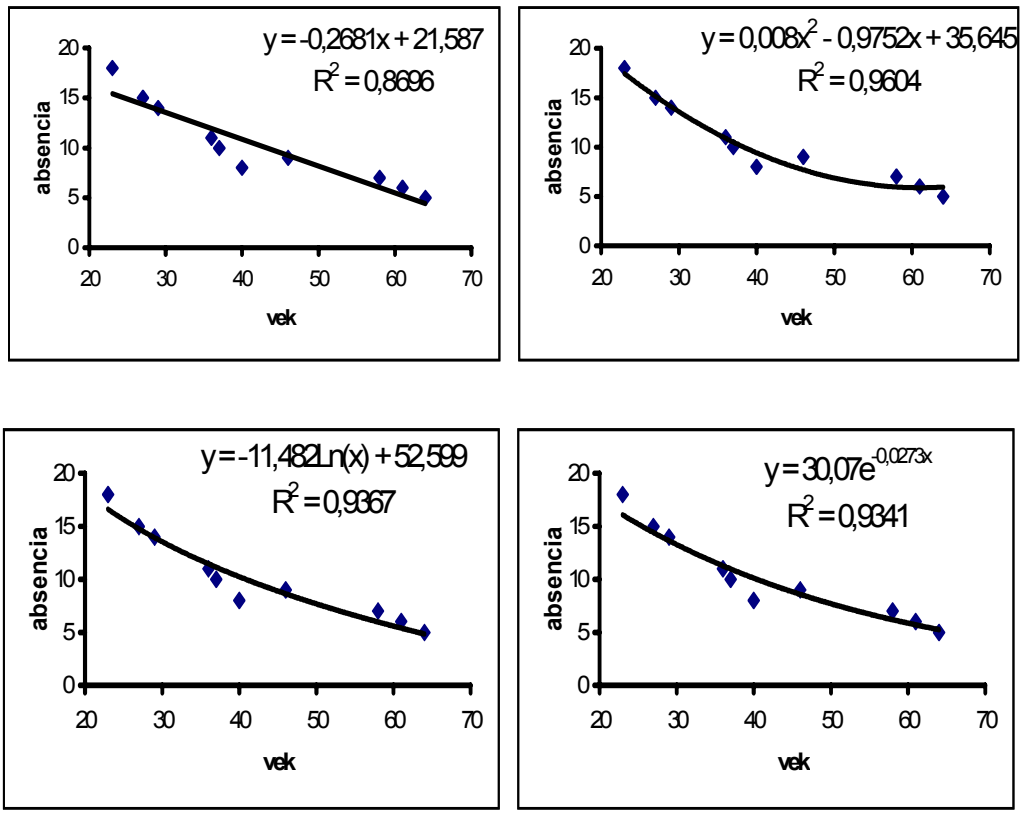
Pracovník personálneho oddelenia cíti, že existuje vzťah medzi počtom dní absencie v práci a vekom pracovníka. Vyšetrite túto závislosť na základe údajov v Tab. 8.9. Pre názornosť príkladu sme použili nevyhovujúci, veľmi malý rozsah výberu.

Tab. 8.9

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

EXCEL

S vyžitím EXCELu vieme vložiť do bodových diagramov graf regresnej čiary, EXCEL poskytuje na výber lineárnu, logaritmickú, exponenciálnu, polynomickú (ľubovoľného stupňa) alebo mocninovú krivku s jej analytickým vyjadrením i koeficientom determinácie. V tomto príklade vidíme, že z viacerých možných riešení je najvhodnejšia parabola s najväčším koeficientom determinácie. Sám riešiteľ úlohy sa musí rozhodnúť, ktorý z týchto možných modelov je pre jeho potreby vyhovujúci.



Obr. 8.5 Rôzne modely regresných funkcií k Príkladu 8.3

Ak sa podrobnejšie zaujímame aj o celkový F-test štatistickej významnosti modelu, t-testy nulovosti koeficientov regresnej funkcie, o intervalové odhady koeficientov regresnej funkcie použijeme v EXCELi funkciu ***Regresia***, kde do vstupnej tabuľky pre

nezávislú premennú vložíme viac stĺpcov . Výstupné tabuľky pre náš príklad sú uvedené pre parabolickú regresiu, interpretácia tabuliek je rovnaká ako pri lineárnej regresii.

Tab. 8.10 Kvadratická regresia k Príkladu 8.3

abs	vek	vek na druhú
15	27	729
6	61	3721
10	37	1369
18	23	529
9	46	2116
7	58	3364
14	29	841
11	36	1296
5	64	4096
8	40	1600

VÝSLEDEK

Regresní statistika	
Násobné R	0,9800
Hodnota spolehlivosti R	0,9604
Nastavená hodnota spolehlivosti R	0,9491
Chyba stř. hodnoty	0,9516
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	153,7611	76,8805	84,8979	1,23507E-05
Rezidua	7	6,3389	0,9056		
Celkem	9	160,1000			

	Koeficienty	*Ch.stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	35,645	3,638	9,799	0,0000	27,043	44,247
vek	-0,975	0,178	-5,484	0,0009	-1,396	-0,555
vek na druhu	0,008	0,002	4,006	0,0052	0,003	0,013

Príklady na precvičenie

8.4 Zostrojte bodový diagram závislosti znakov Y na X pre dané výberové súbory a dané premenné, vložte do grafu rovnicu regresnej priamky, hodnotu koeficientu determinácie, vytvorte korelačnú maticu, pomocou t-testu na

hladine významnosti $\alpha = 0{,}05$ otestujte lineárnu závislosť premenných, do grafu vložte iný typ regresnej krivky.

a) PRIJÍMACIE SKÚŠKY/1.fakulta, *X*-matematika, *Y*- fyzika

Určite, aké priemerný počet bodov dosiahne študent z fyziky, ak z matematiky získal 15 bodov (21 bodov).

b) PRIJÍMACIE SKÚŠKY/2.fakulta, *X*-matematika, *Y*- fyzika.

Určite, aké priemerný počet bodov dosiahne študent z fyziky, ak z matematiky získal 15 bodov (21bodov).

c) PRIJÍMACIE SKÚŠKY/1.fakulta, *X*-matematika (upr.) *Y*- priemer

Určite, aké priemernú známku dosiahne študent po 2.semestri, ak zo skúšky z matematiky (upír) mal známku 3,2 (2).

d) PRIJÍMACIE SKÚŠKY/2.fakulta, *X*-matematika, *Y*- priemer.

Určite, aké priemernú známku dosiahne študent po 2.semestri, ak zo skúšky z matematiky (upr) mal známku 3,2 (2).

8.5 Pomocou nástroja ***Regresia*** vyšetrite lineárnu závislosť znakov *X, Y* v základných súboroch, určite korelačný koeficient, koeficient determinácie, nájdite rovnicu regresnej priamky, na hladine významnosti $\alpha = 0{,}05$ pomocou celkového F-testu, alebo t-testu posúďte významnosť štatistického modelu, nájdite príslušné intervaly spoľahlivosti pre koeficienty regresnej priamky.

a) PRIJÍMACIE SKÚŠKY/1.fakulta, *X*-matematika, *Y*- fyzika,

b) PRIJÍMACIE SKÚŠKY/2.fakulta, *X*-matematika, *Y*- fyzika.