UNIVERZITA PALACKÉHO V OLOMOUCI

PŘÍRODOVĚDECKÁ FAKULTA

KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Závislost nebo korelace? Míry asociace mezi proměnnými

Vedoucí diplomové práce:
RNDr. PhDr. Ivo Müller, Ph.D.
Rok odevzdání: 2013

Vypracovala:
Ivana Janíková
ME, IV. ročník

Prehlásenie

Prehlasujem, že som bakalársku prácu spracovala samostatne pod vedením RNDr. PhDr. Iva Müllera, Ph.D. a že som všetky použité zdroje uviedla.

V Olomouci 24.4.2013

……

Ivana Janíková

Poďakovanie

Veľmi rada by som chcela poďakovať svojmu vedúcemu bakalárskej práce RNDr. PhDr. Ivovi Müllerovi, Ph.D. za ochotu, trpezlivosť, odborné vedenie, cenné rady a pripomienky a za všetok čas venovaný konzultáciám k bakalárskej práci. Ďalej by som chcela poďakovať rodine, blízkym a priateľom za podporu počas štúdia.

Obsah

Úvod

V bakalárskej práci sa budeme zaoberať témou závislosti, jej prítomnosťou, intenzitou a charakterom. Práca pozostáva zo šiestich hlavných kapitol.

V prvej kapitole si vysvetlíme základné pojmy súvisiace s touto témou, z ktorých budeme skloňovať najmä pojmy štatistický znak a závislosť. Štatistické znaky môžeme rozdeliť na kvantitatívne a kvalitatívne, pričom toto delenie ďalej rozvetvíme, a tak sa oboznámime so štatistickými znakmi rôzneho typu, kde ku každému z nich si pre lepšiu predstavu uvedieme príklad. Obdobne tak učiníme aj pri pojme závislosť, ktorá má takisto niekoľko delení.

Druhá kapitola je venovaná kovariancii a jej teoretickým poznatkom.

V tretej kapitole sa budeme zapodievať najprv teoretickými základmi korelácie, korelačným koeficientom $\rho_{X,Y}$ a jeho vlastnosťami, ktoré si postupne dokážeme. Od teoretického koeficienta $\rho_{X,Y}$ prejdeme k výberovému korelačnému koeficientu $r_{X,Y}$, nazývanému aj Pearsonov korelačný koeficient, ktorý sa používa pri praktických výpočtoch. Okrem tohto koeficienta sa zoznámime aj so Spearmanovým korelačným koeficientom r_S .

V ďalšej kapitole sa budeme venovať regresii, ktorá nám pomáha určiť charakter závislosti. Na začiatku sa oboznámime s teóriou lineárnej regresie a potom bude naša pozornosť smerovať k jednoduchej priamkovej regresii a k určeniu odhadov regresných parametrov regresnej funkcie. V priebehu tejto kapitoly si získané vzorce vyskúšame aj na niekoľkých príkladoch.

Piata kapitola je takisto ako predošlé teoretickou kapitolou. V nej sa budeme snažiť odvodiť ďalšie korelačné koeficienty používané v prípadoch, v ktorých nám vystupuje aspoň jedna alternatívna premenná. Týmito koeficientami sú bodovo biseriálny koeficient $r^b_{X,Y}$ a koeficient Φ .

Posledná šiesta kapitola je praktickou časťou bakalárskej práce. V nej budeme na reálnych dátach overovať platnosť vzorcov a porovnávať medzi sebou jednotlivé korelačné koeficienty.

Jedným z cieľov tejto práce, okrem hlbšieho oboznámenia sa s teoretickými poznatkami, ktoré súvisia s touto témou, je nájsť vzťah medzi pojmami nezávislosť a kovariancia, nezávislosť a korelácia, kovariancia a korelácia, zistiť vzťah regresného parametra a korelačného koeficienta, vzťah medzi regresnými parametrami navzájom a pod. Tomuto sa budeme venovať priebežne v jednotlivých kapitolách. Ďalším cieľom, ako už bolo naznačené, je zamerať sa rôzne možnosti definovania korelačného koeficienta, objasniť ich interpretáciu a vzťahy medzi nimi, navzájom ich medzi sebou porovnať a potom vzťahy ilustrovať na dátach.

1 Štatistické znaky a závislosť

Jednou z charakteristických čŕt súčasného obdobia je čoraz širšie uplatňovanie matematických metód v mnohých spoločenských a technických vedeckých disciplínach. Dôsledkom tohto trendu je skúmanie prírodných a technických javov a procesov z pozície matematickej štatistiky. Pomocou jej metód sa odhaľujú štatistické zákonitosti vývoja javov a procesov a ich vzájomné vzťahy.

Nevyhnutným predpokladom každého štatistického skúmania je *hromadnosť* skúmania. Pri hromadnom pozorovaní môže ísť o *jednoduché pozorovanie*, ak do priebehu pozorovaných javov nijakým spôsobom nezasahujeme a neovplyvňujeme ich (údaje získavame pozorovaním, meraním, prostredníctvom štatistického prieskumu a pod.) – typické pre spoločenské javy, alebo o *experiment* (pokus), pri ktorom sa vytvorí súbor kontrolovaných podmienok, v ktorých sa pozorovaný jav opakuje – typické pre prírodné javy. Predmetom štatistického skúmania je teda *hromadný jav*. Hromadným javom rozumieme každý prírodný alebo spoločenský jav (udalosť) (napr. spotreba potravín) pozorovateľný u dostatočne veľkého množstva takých jedincov (prvkov), ktorí majú niektoré podstatné vlastnosti zhodné. Ak presne vymedzíme tieto zhodné vlastnosti u jednotlivých jedincov (prvkov) z hľadiska priestorového, časového a vecného, hovoríme o týchto jedincoch (prvkoch) ako o *štatistických jednotkách* (napr. domácnosť). Množinu štatistických jednotiek nazývame *štatistický súbor* (napr. domácnosti Slovenska) a počtu štatistických jednotiek v tomto súbore hovoríme *rozsah štatistického súboru* (napr. počet domácností na Slovensku). Vlastnosti štatistických jednotiek majú svoje charakteristiky, ktoré nazývame *štatistické znaky* (premenné, náhodné veličiny) (napr. počet členov domácnosti, príjem domácnosti, nároky na kvalitu potravín a pod.) Štatistické znaky sú vonkajšou postrehnuteľnou a merateľnou charakteristikou vlastnosti štatistických jednotiek v štatistickom súbore. Štatistické znaky možno podľa ich charakteru rozdeliť na *kvantitatívne* a *kvalitatívne*.

Kvantitatívne znaky sú merateľné a vyjadrujú vlastnosti štatistickej jednotky reálnymi číslami. Môžu byť:

1.) *spojité*, kedy nadobúdajú ľubovoľnú hodnotu z nejakého intervalu, napr. výška osemročných detí;
2.) *diskrétne*, ktoré nadobúdajú izolované, väčšinou celočíselné hodnoty, napr. počet detí v rodine.

Kvalitatívne znaky popisujú vlastnosti štatistickej jednotky slovne, definíciou alebo symbolmi. Nadobúdajú vždy diskrétne hodnoty a nemusia sa dať jednoznačne merať. Delia sa na:

1.) *ordinálne* (poradové), u ktorých je ich hodnoty možné zmysluplne usporiadať, napr. známka z vyučovacieho predmetu;
2.) *nominálne*, kde o ich dvoch pozorovaných hodnotách je možné rozhodnúť jedine o tom, či sú rovnaké alebo rôzne. Žiadne iné vzťahy sa medzi ich hodnotami neuvažujú. Hodnoty sa nedajú zmysluplne usporiadať, napr. farba očí.

Kvalitatívne štatistické znaky sa podľa počtu kategórií rozlišujú aj na:

1.) *alternatívne* (dichotomické), ktoré nadobúdajú iba dve možné kategórie, napr. pohlavie;
2.) *množné* (multinomické), ktoré môžu nadobúdať viac kategórií, napr. rodinný stav.

Pri sledovaní štatistických znakov (náhodných veličín) často uvažujeme viac znakov zároveň a skúmame, či a akým spôsobom sa ovplyvňujú. Skutočnosť, že sa neovplyvňujú, vystihujeme pomocou pojmu nezávislosť. Oboznámime sa s nasledujúcimi dvoma definíciami.

Definícia 1.1 *Diskrétne náhodné veličiny* $X_1, \dots, X_n$ *sú nezávislé práve vtedy, keď pre pravdepodobnostnú funkciu* p_X *náhodného vektora* $X = (X_1, \dots, X_n)'$ *a marginálne pravdepodobnostné funkcie* p_j *náhodných veličín* X_j , $j = 1, \dots n$ *, platí*

$$p_X(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n) = \prod_{j=1}^{n} P(X_j = x_j) = \prod_{j=1}^{n} p_j(x_j) \quad , \tag{1.1}$$
$$\forall (x_1, \dots, x_n)' \in M_1 \times \cdots \times M_n = M \quad ,$$

kde M_j *je obor hodnôt náhodnej veličiny* X_j , $j = 1, \dots n$.

Definícia 1.2 *Spojité náhodné veličiny* $X_1, \dots, X_n$ *sú nezávislé práve vtedy, keď pre hustotu* $f_X(x_1, \dots, x_n)$ *náhodného vektora* $X = (X_1, \dots, X_n)'$ *a marginálne hustoty* $f_j(x_j)$ *náhodných veličín* X_j , $j = 1, \dots n$ *, platí*

$$f_X(x_1, \dots, x_n) = \prod_{j=1}^{n} f_j(x_j) \quad \text{pre skoro všetky} \quad (x_1, \dots, x_n)' \in R^n \quad . \tag{1.2}$$

V prípade dvoch náhodných veličín platí, že diskrétne náhodné veličiny X a Y sú nezávislé práve vtedy, keď

$$P(X=x, Y=y)=P(X=x)\cdot P(Y=y)$$

pre každé x , y , ktoré X a Y nadobúdajú. Podobne spojité náhodné veličiny X a Y sú nezávislé práve vtedy, keď

$$f_{X,Y}(x, y)=f_X(x)\cdot f_Y(y)$$

pre každé $x, y \in R$.

Štatistická závislosť je opakom (negáciou) nezávislosti. Náhodné veličiny sú závislé, ak nie sú nezávislé. U nominálnych premenných používame tiež pojem asociácia v zmysle závislosti.

Pri skúmaní štatistických znakov sa môžeme stretnúť s rozličnými druhmi závislosti. Sledovaním príčiny a následku delíme závislosť na *príčinnú* a *zdanlivú*:

1.) O *príčinnej (kauzálnej) závislosti* hovoríme vtedy, keď vznik, existencia či zmeny jedných javov (príčin, nezávislých premenných) podmieňujú vznik, existenciu či zmeny iných javov (účinkov, závislých premenných), alebo keď sa javy podmieňujú vzájomne. Preto rozlišujeme príčinnú závislosť na *jednostrannú* a *vzájomnú*. Napr. cena bytu môže závisieť od väčšieho počtu vysvetľujúcich premenných – od lokality, od plochy bytu, od toho, či je panelový alebo tehlový, na ktorom poschodí sa nachádza. V tomto prípade pôjde o jednostrannú závislosť, keďže znaky lokalita, plocha bytu a pod. vystupujú ako príčiny a cena bytu ako následok. Podobne je jednostrannou príčinnou závislosťou i závislosť vreckového dieťaťa ako následku na príčinách, ako je napr. príjem rodičov, známky dieťaťa v škole, vykonávaní domácich prác dieťaťa. Vzájomnú príčinnú závislosť si môžeme všimnúť medzi výškou manžela a výškou manželky. Nemôžeme povedať, že výška manžela je príčinou toho, prečo je manželka práve takto vysoká. A takisto naopak. Hlavnú rolu pri výbere životného partnera hrajú vzájomné sympatie oboch partnerov, vďaka ktorým si navzájom výškovo vyhovujú, a teda nemožno jednoznačne určiť, ktorý zo znakov predstavuje príčinu a ktorý následok, pretože oba môžu vystupovať i ako príčina, i ako následok. U kauzálnej závislosti nazývame premennú, ktorej závislosť od iných premenných zisťujeme, *vysvetľovaná*, resp. *závislá premenná (regresand)*. Premenné, o ktorých predpokladáme, že vyvolávajú zmeny závislej premennej, nazývame *vysvetľujúce*, resp. *nezávislé premenné (regresory)*.
2.) Pri *zdanlivej (klamnej) závislosti* nie je vzťah medzi určitými javmi dôsledkom ich vzájomnej príčinnej súvislosti. Napr. pri skúmaní súvislostí medzi výskytom žuly, ktorá je najčastejšou horninou v Českej republike, a belochov v Českej republike môžeme zistiť silnú mieru závislosti. Avšak logickou úvahou si uvedomujeme, že táto súvislosť je nezmyselná, a teda klamná. V tejto situácii hovoríme o *náhodnej klamnej závislosti* alebo o *náhodnom spoločnom výskyte*. Zdanlivá závislosť (spoločný výskyt) medzi dvoma javmi môže byť tiež výsledkom pôsobenia tzv. *tretieho faktora*. Napr. v štátoch, kde je mnoho televíznych prístrojov, dosahujú obyvatelia vysoký vek. Je však možné zmenou počtu televíznych prístrojov dosiahnuť predĺženie veku v oblastiach sveta, kde je nižšia očakávaná dĺžka života? Táto závislosť je zdôvodnená premennou životná úroveň, ktorá je spoločnou príčinou oboch premenných. Takisto napríklad vzťah medzi telesnou váhou detí a ich zručnosťou (čím majú deti vyššiu telesnú váhu, tým sú zručnejšie) je spôsobený spoločným tretím faktorom - vekom. Pretože staršie deti sú šikovnejšie i ťažšie v pomere k mladším deťom.

Ďalšie delenie závislosti je na *pevnú* a *voľnú*:

1.) *Pevná (funkčná, deterministická) závislosť* medzi premennými je vtedy, keď sa každej hodnote nezávislej premennej priraďuje jednoznačne jediná hodnota závislej premennej. Na závislú premennú pôsobí len vysvetľujúca premenná a žiadne iné činitele. S pevnou závislosťou, ktorá patrí medzi najjednoduchšie formy súvislostí, sa stretávame u niektorých prírodných javov. Napríklad vo fyzike, kde je možné pri sledovaní súvislostí medzi javmi vylúčiť v laboratórnych podmienkach všetky vedľajšie vplyvy, môžeme dosiahnuť, že medzi príčinou a následkom bude existovať jednoznačný vzťah, kedy určitá rovnaká zmena príčiny vyvolá vždy určitú rovnakú zmenu následku. Príkladom môže byť dĺžka kovovej tyče, ktorá je vo funkčnom vzťahu závislá na teplote, alebo závislosť predĺženia pružiny od hmotnosti zaveseného telesa. Takýto druh vzťahov však nie je predmetom štatistického skúmania.
2.) *Voľná (štatistická, stochastická) závislosť* vzniká vtedy, keď rovnakej hodnote nezávislej premennej môže odpovedať viac hodnôt závislej premennej. Na závislú premennú okrem nezávislých premenných pôsobia aj ďalšie činitele a nešpecifikované náhodné vplyvy. Preto je voľná závislosť zložitejšou formou súvislostí, ktorú nachádzame najmä v spoločenských javoch. Napríklad úspory domácností závisia jednak od disponibilného príjmu domácnosti, ale aj od počtu členov domácnosti, veku jednotlivých členov domácnosti a pod. Takisto spotreba vajec v domácnosti závisí na príjme domácnosti, ale aj od aktuálnej ceny vajec alebo od toho, či prichádzajú sviatky alebo oslavy, kedy pečieme viac zákuskov a pod. Zamyslieť sa môžeme aj nad príkladom závislosti predĺženia pružiny od hmotnosti zaveseného telesa. V určitých prípadoch môže dôjsť k tomu, že nepôjde o pevnú závislosť, ale o voľnú. A to v prípade, keď merania uskutočníme mimo laboratórnych podmienok, teda napr. vonku. Zároveň nebudeme predĺženie pružiny merať skutočne presným zariadením, a tak môže nastať, že každej hodnote nezávislej premennej (hmotnosti telesa), nebude jednoznačne priradená jediná hodnota závislej premennej (predĺženie pružiny). Tým, že nepracujeme v laboratórnych podmienkach, môže túto zmenu spôsobiť napr. vonkajšia teplota vzduchu, atmosférický tlak alebo nepresnosť prístroja. Štatistika sa zaoberá predovšetkým skúmaním voľnej závislosti.

Závislosť môžeme posudzovať kvalitatívne, teda v zmysle definícií 1.1 a 1.2 či je alebo nie je prítomná, alebo kvantitatívne. Každá závislosť má dva vzájomne neoddeliteľné kvantitatívne atribúty (vlastnosti), a to mieru (intenzitu) závislosti a priebeh (charakter) závislosti. Podľa matematického tvaru funkcie, ktorá zobrazuje priebeh závislosti, klasifikujeme závislosť na:

1.) *lineárnu*, kedy sú zmeny jednej premennej presne alebo zhruba lineárne závislé na zmenách druhej premennej. Priebeh lineárnej závislosti je možné schematicky popísať priamkou;

2.) *nelineárnu*, kedy zmeny jednej a druhej premennej nie sú na sebe lineárne závislé. Priebeh nelineárnej závislosti dvoch premenných schematicky popisuje nejaká krivka. Podľa tvaru funkcie môže ísť o závislosť kvadratickú, exponenciálnu, logaritmickú a pod.

V oboch prípadoch sa môže jednať o závislosť funkčnú (pevnú) alebo štatistickú (voľnú). V prípade štatistickej závislosti sa metódy pre hľadanie vhodnej funkcie označujú ako *regresia* (*lineárna, nelineárna*).

Čo sa týka miery závislosti, v prípade kvantitatívnych premenných hovoríme o *korelácii* a *korelačnom koeficiente*. V regresii sa používa tzv. *koeficient determinácie*. U ordinálnych premenných môžeme použiť *poradovú koreláciu*. Závislosť medzi kvalitatívnymi premennými nám vyjadruje *asociácia*, ak ide o alternatívne kvalitatívne premenné, a *kontingencia*, ak máme množné kvalitatívne premenné. Ako miery asociácie sa používajú špeciálne prípady korelačného koeficienta, napr. bodovo biseriálny koeficient alebo koeficient Φ. O korelácii, regresii a asociácii bude pojednávané ďalej.

V tejto kapitole bolo čerpané predovšetkým z literatúry: [8], [10], [11].

2 Kovariancia

Kovariancia nám zisťuje prítomnosť a smer lineárnej závislosti dvoch náhodných veličín.

Definícia 2.1 Nech náhodné veličiny X, Y majú konečné druhé momenty. *Kovariancia* $cov(X,Y)$ *náhodných veličín* X, Y je číslo definované vzťahom

$$cov(X,Y)=E\left[(X-E(X))(Y-E(Y))\right] \quad . \tag{2.1}$$

Najčastejšie slúži k výpočtu kovariancie vzorec

$$cov(X,Y)=E(XY)-E(X)E(Y) \quad , \tag{2.2}$$

ktorý odvodíme takto:

$$cov(X,Y)=E\left[(X-E(X))(Y-E(Y))\right]=E\left[XY-X\,E(Y)-Y\,E(X)+E(X)E(Y)\right]=¿$$
$$¿\,E(XY)-E(X)E(Y)-E(X)E(Y)+E(X)E(Y)=E(XY)-E(X)E(Y)\quad .$$

Kovariancia má nasledujúce vlastnosti:

1.) $cov(X,X)=var(X)$,
2.) $cov(X,Y)=cov(Y,X)$ *(symetrickosť),*
3.) *Ak je* $var(X)=0$ *alebo* $var(Y)=0$ *, potom* $cov(X,Y)=0$.

Dôkaz 3. vlastnosti. V prípade, že je aspoň jeden z rozptylov náhodných veličín X, Y rovný nule, napr. $var(X)=0$, potom to znamená, že $X=a$, t. j. náhodná veličina X je konštanta skoro určite. Vieme, že $E(a)=a$. Potom

$$cov(X,Y)=cov(a,Y)=E\left[(a-E(a))(Y-E(Y))\right]=¿$$
$$¿\,E\left[(a-a)(Y-E(Y))\right]=E\left[0\cdot(Y-E(Y))\right]=¿$$
$$¿\,0\cdot E(Y-E(Y))=0\quad .$$

□

Vzťah nezávislosti a kovariancie

Veta 2.1 *Nech sú X, Y nezávislé náhodné veličiny, ktoré majú stredné hodnoty* $E(X)$, $E(Y)$ *a nech existuje* $E(X\cdot Y)$ *. Potom platí*

$$E(X\cdot Y)=E(X)\cdot E(Y)\quad .$$

Dôkaz. a) Pre diskrétne náhodné veličiny X, Y s použitím predpokladu o nezávislosti a existencii stredných hodnôt dostaneme

$$E(X\cdot Y)=\sum_{x_i\in M_1}\sum_{y_j\in M_2}x_i y_j P(X=x_i,Y=y_j)=¿$$

$$¿ \sum_{x_i \in M_1} \sum_{y_j \in M_2} x_i y_j P(X = x_i) P(Y = y_j) = ¿$$
$$¿ \left[\sum_{x_i \in M_1} x_i P(X = x_i) \right] \left[\sum_{y_j \in M_2} y_j P(Y = y_j) \right] = ¿$$
$$¿ E(X) \cdot E(Y) \quad .$$

b) Pre spojité náhodné veličiny X, Y s hustotami f_1, f_2 dostaneme

$$E(X \cdot Y) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x \cdot y f_2(y) dy \right] f_1(x) dx = ¿$$
$$¿ \left[\int_{-\infty}^{\infty} x f_1(x) dx \right] \left[\int_{-\infty}^{\infty} y f_2(y) dy \right] = ¿$$
$$¿ E(X) \cdot E(Y) \quad .$$

□

Lemma 2.1 *Ak sú X a Y nezávislé, potom* $cov(X, Y) = 0$.

Dôkaz. Podľa vzorca (2.2) vieme, že $cov(X, Y) = E(XY) - E(X)E(Y)$, a podľa vety 2.1
$cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$.

□

Lemma 2.2 *Ak je* $cov(X, Y) = 0$ *, náhodné veličiny* X, Y *nemusia byť nezávislé.*

Dôkaz. Tvrdenie dokážeme pomocou protipríkladu. Ak má X rozdelenie, ktoré je symetrické okolo nuly, potom $E(X) = 0$. Takisto $E(X^3) = 0$, lebo funkcia X^3 má tiež rozdelenie symetrické okolo nuly. Vezmeme $Y = X^2$. Y a X sú závislé, pretože Y je funkciou X. Napriek tomu

$$cov(X, Y) = E(XY) - E(X)E(Y) = E(X \cdot X^2) - E(X)E(Y) = ¿$$
$$¿ E(X^3) - E(X)E(Y) = 0 - 0 \cdot E(Y) = 0 \quad .$$

Vidíme teda, že nulová kovariancia nám nezaručuje, že medzi náhodnými veličinami nie je žiadna závislosť.

□

Kovariancia medzi pôvodnou veličinou a jej lineárnou transformáciou vyzerá takto:

$$cov(X, aX+b) = E[X - E(X)][aX+b-E(aX+b)] = E[X - E(X)][aX - a\,E(X)] = a\,E[X-E(X)][X-E(X)] =$$

.

Pri práci s náhodným vektorom $X = (X_1, \ldots, X_n)'$ sa používa nasledujúce zovšeobecnenie pojmu rozptyl a kovariancia.

Definícia 2.2 Nech pre náhodné veličiny $X_1, \ldots, X_n$ existujú rozptyly. Potom výraz

$$C = var(X) = \left(cov(X_i, X_j)\right)_{i,j=1}^{n} = \begin{pmatrix} var(X_1) & cov(X_1, X_2) \cdots & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & \cdots & cov(X_2, X_n) \\ cov(X_n, X_1) & cov(X_n, X_2) \cdots & ¿ \cdots & var(X_n) \end{pmatrix}$$

sa nazýva *kovariančná (variančná) matica* náhodného vektora X .

V tejto kapitole bolo čerpané predovšetkým z literatúry: [8].

3 Korelácia

Veličiny X a Y môžu byť závislé alebo nezávislé. Ak sú závislé, je potrebné túto ich mieru (intenzitu) závislosti nejako kvantitatívne ohodnotiť. Avšak závislosť môže mať najrôznejší priebeh (napr. lineárny, exponenciálny). Najčastejšie sa na meranie závislosti v prípade *lineárneho* vzťahu používa *korelačný koeficient*. Korelácia zisťuje prítomnosť, smer a veľkosť lineárnej závislosti dvoch štatistických znakov.

3.1 Teoretické základy korelácie

Definícia 3.1 Nech X a Y sú náhodné veličiny s konečnými druhými momentmi. *Korelačný koeficient* $\rho_{X,Y}$ náhodných veličín X, Y je číslo definované vzťahmi

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}} \wedge pre\ var(X) > 0,\ var(Y) > 0, \qquad (3.1)$$
$$\rho_{X,Y} = 0 \wedge pre\ var(X) = 0\ alebo\ var(Y) = 0.$$

Veta 3.1 Vlastnosti korelačného koeficienta:

1.) $\rho_{X,X}=1$ *(normovanosť)*;
2.) $\rho_{X,Y}=\rho_{Y,X}$ *(symetrickosť)*.
3.) *Nech a, b, c, d sú reálne čísla, pričom* $bd\neq 0$ *. Potom*

$$\rho_{a+bX,c+dY}=\begin{cases}-\rho_{X,Y} \wedge pre\, bd<0,\\ \rho_{X,Y} \wedge pre\, bd>0.\end{cases}$$

4.) *Pre korelačný koeficient platí* $-1\leq\rho_{X,Y}\leq 1$ *. Rovnosť* $\rho_{X,Y}=1$ *platí práve vtedy, keď* $P(Y=a+bX)=1$ *, pričom* $b>0$ *. Analogicky rovnosť* $\rho_{X,Y}=-1$ *platí práve vtedy, keď* $P(Y=a+bX)=1$ *, pričom* $b<0$ *.*
5.) $\rho_{X,Y}=cov\left(\frac{X-E(X)}{\sqrt{var(X)}},\frac{Y-E(Y)}{\sqrt{var(Y)}}\right)$ *.*

Dôkaz. Prvá a druhá vlastnosť sú zrejmé. Tretia vlastnosť hovorí, že pri lineárnej transformácii sa korelačný koeficient buď nezmení vôbec, alebo len zmení znamienko.

$$\rho_{a+bX,c+dY}=\frac{E\left[a+bX-E(a+bX)\right]\left[c+dY-E(c+dY)\right]}{\sqrt{var(a+bX)\cdot var(c+dY)}}=¿$$

$$¿\frac{E\left[bX-b\,E(X)\right]\left[dY-d\,E(Y)\right]}{\sqrt{E\left[a+bX-E(a+bX)\right]^2 E\left[c+dY-E(c+dY)\right]^2}}=¿$$

$$¿\frac{bd\,E\left[X-E(X)\right]\left[Y-E(Y)\right]}{\sqrt{b^2d^2\,E\left[X-E(X)\right]^2 E\left[Y-E(Y)\right]^2}}=¿$$

$$¿\frac{bd\,E\left[X-E(X)\right]\left[Y-E(Y)\right]}{|bd|\sqrt{var(X)\cdot var(Y)}}=\frac{bd}{|bd|}\rho_{X,Y}\quad .$$

Štvrtú vlastnosť dokážeme pomocou Schwarzovej nerovnosti

$$\left|E\left[X-E(X)\right]\left[Y-E(Y)\right]\right|\leq\sqrt{E\left[X-E(X)\right]^2\left[Y-E(Y)\right]^2}\quad ,$$

ktorá má v našom označení podobu

$$\left|cov(X,Y)\right| \le \sqrt{var(X)\cdot var(Y)} \quad .$$

Z toho nám podľa definície 3.1 plynie, že $-1 \le \rho_{X,Y} \le 1$. Rovnosť vo Schwarzovej nerovnosti je dosiahnutá vtedy, keď platí buď $X - E(X) = 0$, alebo $Y - E(Y) = 0$ skoro určite. To však v našom prípade neprichádza do úvahy vzhľadom k predpokladu $var(X) > 0, var(Y) > 0$ v definícii 3.1. Alebo keď platí $Y - E(Y) = b\left[X - E(X)\right]$ skoro určite pre nejaké $b \neq 0$. Výpočtom sa overí, že v tomto prípade

$$\rho_{X,Y} = \frac{E\left[bX - E(bX)\right]\left[Y - E(Y)\right]}{\sqrt{E\left[bX - E(bX)\right]^2 E\left[Y - E(Y)\right]^2}} = ¿$$
$$¿ \frac{b\,E\left[X - E(X)\right]\left[Y - E(Y)\right]}{\sqrt{b^2}\sqrt{E\left[X - E(X)\right]^2 E\left[Y - E(Y)\right]^2}} = \frac{b}{|b|} \quad ,$$

teda $\rho_{X,Y} = 1$ pre $b > 0$ a $\rho_{X,Y} = -1$ pre $b < 0$.

Piatu vlastnosť dokážeme vypočítaním kovariancie dvoch normovaných náhodných veličín podľa vzorca (2.2).

$$cov\left(\frac{X - E(X)}{\sqrt{var(X)}}, \frac{Y - E(Y)}{\sqrt{var(Y)}}\right) = E\left[\frac{X - E(X)}{\sqrt{var(X)}}\frac{Y - E(Y)}{\sqrt{var(Y)}}\right] - E\left[\frac{X - E(X)}{\sqrt{var(X)}}\right] E\left[\frac{Y - E(Y)}{\sqrt{var(Y)}}\right] = ¿$$
$$¿ E\left[\frac{\left(X - E(X)\right)\left(Y - E(Y)\right)}{\sqrt{var(X)\,var(Y)}}\right] - \frac{E\left[X - E(X)\right]}{\sqrt{var(X)}}\frac{E\left[Y - E(Y)\right]}{\sqrt{var(Y)}} = ¿$$
$$¿ \frac{1}{\sqrt{var(X)\,var(Y)}} E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right] - 0 = ¿$$
$$¿ \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}} = \rho_{X,Y} \quad .$$

□

Pojem korelácie je možné zovšeobecniť. Majme náhodný vektor $X=(X_1,\dots,X_n)'$, ktorého zložky majú konečné druhé momenty a kladné rozptyly. *Korelačnou maticou* vektoru X rozumieme maticu $P=(\rho_{ij})$ typu $n\times n$, kde $\rho_{ij}=\rho_{X_i,X_j}$. Z prvej vlastnosti korelačného koeficientu plynie, že matica P má na diagonále jednotky a z druhej vlastnosti korelačného koeficientu je zrejmé, že P je symetrická.

Nech $X=(X_1,\dots,X_n)'$ a $Y=(Y_1,\dots,Y_m)'$ sú náhodné vektory, ktorých zložky majú konečné druhé momenty a kladné rozptyly. Potom *korelačná matica* týchto vektorov je $cor(X,Y)=(\rho_{X_i,Y_j})$ a má typ $n\times m$. Korelačnú maticu vektorov nazývame tiež *kros-korelačná matica*.

Vzťah kovariancie a korelácie

Lemma 3.1 *Ak je* $cov(X,Y)=0$ *, potom* $\rho_{X,Y}=0$.

Dôkaz.

$$\rho_{X,Y}=\frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}}=\frac{0}{\sqrt{var(X)\cdot var(Y)}}=0 \quad .$$

Ak je kovariancia nulová preto, že buď $var(X)=0$, alebo $var(Y)=0$, potom je korelácia nulová priamo z definície.

□

Lemma 3.2 *Ak je* $\rho_{X,Y}=0$ *, potom* $cov(X,Y)=0$.

Dôkaz.

a) V prípade, že náhodné veličiny *X*, *Y* majú $var(X)>0$ a $var(Y)>0$, potom aj menovateľ korelačného koeficienta $\sqrt{var(X)\cdot var(Y)}>0$, a teda ak $\rho_{X,Y}=\frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}}=0$, potom musí byť $cov(X,Y)=0$.

b) Ak $\rho_{X,Y}=0$ a zároveň je aspoň jeden z rozptylov náhodných veličín *X*, *Y* rovný nule, tak z tretej vlastnosti kovariancie vieme, že $cov(X,Y)=0$.

□

Vzťah nezávislosti a korelácie

Lemma 3.3 *Ak sú X a Y nezávislé, tak* $\rho_{X,Y}=0$.

Lemma 3.4 *Ak je* $\rho_{X,Y}=0$ *, nemusí to znamenať, že X a Y sú nezávislé.*

Dôkaz. Plynie z toho, že nulová korelácia je ekvivalentná s nulovou kovarianciou (lemmata 3.1 a 3.2) a zo vzťahu nezávislosti a kovariancie (lemmata 2.1 a 2.2).

3.2 Výberový korelačný koeficient

Majme náhodný výber

$$(X_1, Y_1)', \ldots, (X_n, Y_n)'$$

z nejakého dvojrozmerného rozdelenia. Označíme výberový priemer $\acute{X}$ a výberový rozptyl S_X^2 ako charakteristiky výberu $X_1, \ldots, X_n$, teda

$$\acute{X}=\frac{1}{n}\sum_{i=1}^{n} X_i \quad , \tag{3.2}$$

$$S_X^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\acute{X}\right)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2-n\acute{X}^2\right) \quad ,$$

a podobne $\acute{Y}$, S_Y^2 ako charakteristiky výberu $Y_1, \ldots, Y_n$. Ďalej definujme *výberovú kovarianciu* ako

$$S_{X,Y}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\acute{X}\right)\left(Y_i-\acute{Y}\right) \quad . \tag{3.3}$$

Pri praktických výpočtoch však často používame vzťah

$$S_{X,Y}=\sum_{i=1}^{n}\left(X_i-\acute{X}\right)\left(Y_i-\acute{Y}\right)=\sum_{i=1}^{n}X_iY_i-n\,\acute{X}\,\acute{Y}\quad .$$

Ak je $S_X^2>0$ a $S_Y^2>0$, definujeme *výberový korelačný koeficient*, nazývaný aj *Pearsonov korelačný koeficient*, výrazom

$$r_{X,Y}=\frac{S_{X,Y}}{\sqrt{S_X^2S_Y^2}}\quad .$$

Ak je niektorá z veličín S_X^2 , S_Y^2 rovná nule, výberový korelačný koeficient definujeme $r_{X,Y}=0$. Po dosadení a úprave dostaneme výpočtový tvar

$$r_{X,Y}=\frac{\sum_{i=1}^{n}\left(X_i-\acute{X}\right)\left(Y_i-\acute{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\acute{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\acute{Y}\right)^2}}=\frac{\sum_{i=1}^{n}\left(X_i-\acute{X}\right)Y_i-\acute{Y}\sum_{i=1}^{n}\left(X_i-\acute{X}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\acute{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\acute{Y}\right)^2}}=\frac{}{\sqrt{\sum_{i=1}^{n}\left(}} \qquad (3.4)$$

.

Výberový korelačný koeficient preberá všetky vlastnosti jeho teoretického náprotivku $\rho_{X,Y}$ a je potrebné pripomenúť, že je rovnako charakteristikou *lineárnej* závislosti medzi štatistickými znakmi *X* a *Y*.

Prirodzeným rozšírením doterajších úvah je situácia, kedy máme výber z rozdelenia *p*-rozmerného náhodného vektora $X=(X_1,\dots,X_p)'$ so strednou hodnotou $\boldsymbol{\mu}$ a variančnou maticou **C**. Označíme ho

$$X_1=\begin{pmatrix} X_{11} \\ \vdots \\ X_{1p} \end{pmatrix},\dots,X_n=\begin{pmatrix} X_{n1} \\ \vdots \\ X_{np} \end{pmatrix} .$$

Obmedzíme sa na prípad, kedy $n>p$, teda rozsah výberu je väčší než počet zložiek vektoru. Zavedieme *výberový priemer* $\acute{X}$ a *výberovú variančnú maticu* $S_X=(s_{ij})_{i,j=1}^{p}$ pomocou vzorcov

$$\acute{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \ ,$$

$$(X_i-\acute{X})(X_i-\acute{X})'=¿\frac{1}{n-1}\left(\sum_{i=1}^{n}X_iX_i'-n\acute{X}\,\acute{X}'\right) .$$

$$S_X=\frac{1}{n-1}\sum_{i=1}^{n}¿$$

Prvkami výberovej variančnej matice sú teda výberové rozptyly (diagonálne prvky) a výberové kovariancie. Ak sú všetky diagonálne prvky matice S_X kladné, definujeme *výberovú korelačnú maticu* vzťahom

$$R_X=(r_{ij})_{i,j=1}^{p}=\left(\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}\right)_{i,j=1}^{p} .$$

Keďže výberový korelačný koeficient preberá všetky vlastnosti jeho teoretického náprotivku $\rho_{X,Y}$, potom pre prvky matice R_X plynie, že jej diagonálne prvky sú vždy rovné jednej, nediagonálne prvky sú výberové korelačné koeficienty zodpovedajúcich zložiek a platí pre ne $-1 \le r_{ij} \le 1$ a matica R_X je symetrická takisto ako korelačná matica P.

3.3 Spearmanov korelačný koeficient

Pri hodnotení štatistickej závislosti môžeme niekedy skonštatovať, že nemožno použiť obyčajný korelačný koeficient. Niekedy totiž v náhodnom výbere $(X_1, Y_1)', \dots, (X_n, Y_n)'$ nie je možné hodnoty uvedených náhodných veličín presne stanoviť, ale máme k dispozícii iba poradie veličín $X_1, \dots, X_n$ a poradie veličín $Y_1, \dots, Y_n$. Ale ak sú poradia *X*-ových a *Y*-ových veličín veľmi podobné, nepochybne to svedčí o istej závislosti medzi X_i a Y_i, $i = 1, \dots, n$. Presným vyjadrením tejto myšlienky je tzv. *Spearmanov korelačný koeficient*, nazývaný aj poradový korelačný koeficient.

Predpokladajme, že $(X_1, Y_1)', \dots, (X_n, Y_n)'$ je výber zo spojitého dvojrozmerného rozdelenia. Veličiny $X_1, \dots, X_n$ sa usporiadajú podľa veľkosti a zistí sa ich poradie $R_1, \dots, R_n$. Potom sa usporiadajú podľa veľkosti veličiny $Y_1, \dots, Y_n$ a stanoví sa ich poradie $Q_1, \dots, Q_n$. Často sa dvojice $(X_1, Y_1)', \dots, (X_n, Y_n)'$ už vopred usporiadajú podľa rastúcich hodnôt $X_1, \dots, X_n$. V takom prípade potom máme priamo $R_i = i$, $i = 1, \dots, n$.

Spearmanov korelačný koeficient r_S sa definuje ako výberový korelačný koeficient počítaný z dvojíc $(R_1, Q_1)', \dots, (R_n, Q_n)'$.

Veta 3.2 *Platí*

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 \quad . \tag{3.5}$$

Dôkaz. Podľa vzorca (3.4) máme

$$r_S = \frac{\sum_{i=1}^{n} R_i Q_i - n \acute{R} \acute{Q}}{\sqrt{\left(\sum_{i=1}^{n} R_i^2 - n \acute{R}^2\right)\left(\sum_{i=1}^{n} Q_i^2 - n \acute{Q}^2\right)}} \quad . \tag{3.6}$$

Pritom

$$\acute{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2} \quad , \qquad \acute{Q} = \acute{R} \quad ,$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} Q_i^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \quad ,$$

$$\sum_{i=1}^{n} R_i Q_i = \frac{1}{2}\left(\sum_{i=1}^{n} R_i^2 + \sum_{i=1}^{n} Q_i^2\right) - \frac{1}{2}\sum_{i=1}^{n} (R_i - Q_i)^2$$

$$¿ \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^{n} (R_i - Q_i)^2 \quad .$$

Tieto výsledky dosadíme do vzorca (3.6) a po úprave dostaneme vzorec (3.5)

$$r_S = \frac{\frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^{n} (R_i - Q_i)^2 - n\left(\frac{n+1}{2}\right)\left(\frac{n+1}{2}\right)}{\sqrt{\left(\frac{n(n+1)(2n+1)}{6} - n\left(\frac{n+1}{2}\right)^2\right)\left(\frac{n(n+1)(2n+1)}{6} - n\left(\frac{n+1}{2}\right)^2\right)}} = ¿ \frac{\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} - \frac{1}{2}\sum_{i=1}^{n} (R_i - Q_i)^2}{\sqrt{\left(\frac{n(n+1)(2n+1)}{6} - n\frac{(n+1)^2}{4}\right)^2}} = ¿$$

□

V tejto kapitole bolo čerpané predovšetkým z literatúry: [7], [8].

4 Regresia

4.1 Úvod do regresie

Regresia sa používa pri skúmaní závislosti dvoch a viacerých kvantitatívnych premenných. Je to súhrn štatistických metód a postupov slúžiacich k odhadu hodnôt alebo stredných hodnôt nejakej premennej odpovedajúcej daným hodnotám jednej alebo väčšieho počtu vysvetľujúcich premenných. Podkladom pre regresiu sú vždy nejaké údaje o týchto premenných, ktoré boli získané pozorovaním (zisťovaním) u *n* jednotiek. Tieto údaje sa považujú za výberové dáta.

Problémy, ktoré sa dajú riešiť využitím regresie, vznikajú v praxi pomerne často. Hľadanie závislosti medzi premennými je dôležité v mnohých oboroch. Zaujíma nás napríklad vplyv priemernej rýchlosti automobilu na spotrebu benzínu, vplyv zmeny teploty na predĺženie medenej rúry či vplyv výšky investícií do reklamy na hodnotu celkových tržieb podniku. V ekonomickej oblasti sa regresia snáď najviac rozšírila pri analýze a prognózovaní spotreby a dopytu, kedy sa konštruovali rôzne regresné modely slúžiace k odhadu strednej (priemernej) spotreby či dopytu domácností s rôznym príjmom, s rôznym počtom členov, s rôznym počtom detí apod.

Ekonomické veličiny často závisia na väčšom počte činiteľov. Z nich je možné pri regresii využiť len tie, ktoré sa dajú merať. Tie potom tvoria okruh vysvetľujúcich premenných. V tomto prípade, t. j. ak sa do odhadov hodnôt alebo stredných hodnôt zapojí viac vysvetľujúcich premenných, hovoríme o *viacnásobnej regresii*. Vo svojej práci sa však budem zaujímať prevažne o *jednoduchú regresiu*, pri ktorej sa využíva iba jedna vysvetľujúca premenná. Jedná sa teda o najjednoduchší prípad, kedy závislá vysvetľovaná premenná je určená jedinou nezávislou vysvetľujúcou premennou. V príkladoch z predchádzajúceho odseka sú týmito vysvetľujúcimi premennými zrejme priemerná rýchlosť, zmena teploty a investície do reklamy.

4.2 Teoretické základy lineárnej regresie

Majme náhodnú veličinu Y a k-ticu náhodných veličín $X_1,\dots,X_k$. Budeme predpokladať, že majú konečné druhé momenty. Hľadajme čo najlepšiu lineárnu aproximáciu veličiny Y pomocou veličín $X_1,\dots,X_k$, inak tiež regresiu Y na $X_1,\dots,X_k$. Používame pre ňu aj pojem *stochastická regresia* a je užitočná najmä v prípadoch, kedy Y je na rozdiel od $X_1,\dots,X_k$ ťažko dostupná a tieto veličiny tak slúžia k jej viac či menej presnému odhadu. Pre ďalšie použitie označme

$$X=\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \quad , \qquad b=\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} \quad ,$$

$$cov(Y,X)=\left(cov(Y,X_1),\dots,cov(Y,X_k)\right) \quad , \quad cov(X,Y)=\left[cov(Y,X)\right]' \quad .$$

Úlohou je teda nahradiť Y lineárnou funkciou $\hat{Y}$ veličín $X_1,\dots,X_k$,

$$\hat{Y}=a+b_1X_1+\cdots+b_kX_k=a+b'X \quad , \tag{4.1}$$

tak, aby stredná kvadratická chyba $E(Y-a-b'X)^2$ bola minimálna. Odpoveď, ako voliť koeficienty $a,\ b_1,\dots,b_k$ (označme túto optimálnu voľbu $\alpha,\ \beta_1,\dots,\beta_k$), uvádza nasledujúca veta.

Veta 4.1 *Nech je variančná matica* var**(X)** *regulárna, a teda pozitívne definitná. Potom platí*

$$E(Y-a-b'X)^2 \geq var(Y)-cov(Y,X)\left[var(X)\right]^{-1}cov(X,Y)$$

a rovnosť je dosiahnutá vtedy a len vtedy, keď

$$\beta=\left[var(X)\right]^{-1}cov(X,Y) \quad , \quad \alpha=E(Y)-\beta' E(X) \quad . \tag{4.2}$$

Dôkaz. Rozpísaním $Y-a-b'X=Y-a-b'X+\left(E(Y)-b'E(X)\right)-\left(E(Y)-b'E(X)\right)$ dostaneme

$$E\left(Y-a-b'X\right)^2 =$$

$$¿\,E\left[\left(Y-E(Y)\right)-\left(b'X-b'E(X)\right)+\left(E(Y)-a-b'E(X)\right)\right]^2=¿\,E\left(Y-E(Y)\right)^2+E\left[b'\left(X-E(X)\right)\right]^2+\left(E(Y)-a-b'E(X\right.$$

Pripočítaním a odčítaním $cov(Y,X)\left[var(X)\right]^{-1}cov(X,Y)$ dostaneme

$$E\left(Y-a-b'X\right)^2=¿$$

$$¿\,var(Y)-cov(Y,X)\left[var(X)\right]^{-1}cov(X,Y)+\left(E(Y)-a-b'E(X)\right)^2+\left(b-\left[var(X)\right]^{-1}cov(X,Y)\right)'var(X)\left(b-\left[var(X\right.$$

Stredná kvadratická chyba je minimálna pri $b=\left[var(X)\right]^{-1}cov(X,Y)=\beta$, ktorá vynuluje posledný sčítanec ($var(X)$ je pozitívne definitná matica), a pri následnej voľbe

$$a=E(Y)-\left(\left[var(X)\right]^{-1}cov(X,Y)\right)'E(X)=\alpha,$$

ktorá eliminuje tretí sčítanec. Minimálna stredná kvadratická chyba sa teda rovná

$$E\left(Y-\alpha-\beta'X\right)^2=var(Y)-cov(Y,X)\left[var(X)\right]^{-1}cov(X,Y).$$

□

O veličine Y pritom vieme len toľko, že má strednú hodnotu $E(Y)$ a jej kolísanie okolo $E(Y)$ je popísané rozptylom $var(Y)$. S informáciou o rozdelení náhodného vektoru **X** možno Y odhadnúť pomocou $\hat{Y}=\alpha+\beta'X$, kde α a β sú dané vzorcom (4.2). Funkciu $\alpha+\beta'X$ nazývame *lineárnou regresnou funkciou* pri regresii Y na **X** a čísla $\alpha, \beta_1,\ldots,\beta_k$ *regresnými koeficientami*. Odchýlenie $\hat{Y}$ od Y meriame pomocou strednej kvadratickej chyby, t. j. číslom

$$E(Y-\hat{Y})^2=var(Y)-cov(Y,X)\left[var(X)\right]^{-1}cov(X,Y) \quad ,$$

ktorému sa hovorí aj *reziduálny rozptyl*. Využitím vzťahu (4.2) pre odhad β ho môžeme ďalej upraviť do tvaru

$$E(Y-\hat{Y})^2=var(Y)-cov(Y,X)\left[var(X)\right]^{-1}var(X)\left[var(X)\right]^{-1}cov(X,Y)=var(Y)-\beta' var(X)\beta.$$

Ako dôležitý špeciálny prípad budeme uvažovať jedinú vysvetľujúcu veličinu *X*. Pre optimálnu lineárnu náhradu náhodnej veličiny *Y*, $\hat{Y}=\alpha+\beta X$, získame z (4.2)

$$\beta=\frac{cov(X,Y)}{var(X)} \quad , \quad \alpha=E(Y)-\beta E(X) \tag{4.3}$$

a reziduálny rozptyl je

$$E(Y-\hat{Y})^2=var(Y)-cov(Y,X)\left[var(X)\right]^{-1}cov(X,Y) \quad .$$

Keďže máme jednu vysvetľujúcu premennú, všetky členy v reziduálnom rozptyle sú čísla, a preto ho môžeme ďalej upraviť ako

$$E(Y-\hat{Y})^2=var(Y)-\frac{cov(Y,X)}{var(X)}cov(X,Y)=var(Y)-\frac{\left[cov(X,Y)\right]^2}{var(X)} \quad , \tag{4.4}$$

lebo vieme, že kovariancia je symetrická. Tento upravený reziduálny rozptyl dáme do vzťahu s korelačným koeficientom, a to tak, že si zo vzorca (3.1) vyjadríme kovarianciu. Dostaneme

$$cov(X,Y)=\rho_{X,Y}\sqrt{var(X)var(Y)} \quad ,$$

ktorú dosadíme do reziduálneho rozptylu. Potom

$$E(Y-\hat{Y})^2=var(Y)-\frac{\left[\rho_{X,Y}\sqrt{var(X)var(Y)}\right]^2}{var(X)}=var(Y)-\frac{\rho_{X,Y}^2 var(X)var(Y)}{var(X)}$$

$$¿\, var(Y)-\rho_{X,Y}^2 var(Y)=var(Y)\left(1-\rho_{X,Y}^2\right) \quad .$$

Odtiaľ vyjadrením si $\rho_{X,Y}^2$ dostaneme vzťah

$$\rho_{X,Y}^2=1-\frac{E(Y-\hat{Y})^2}{var(Y)} \quad . \tag{4.5}$$

Výraz napravo predstavuje tzv. *teoretický index determinácie* a ako je vidieť, rovná sa štvorcu korelačného koeficienta.

Obrátená regresia

Pozornosť budeme venovať aj tzv. *obrátenej lineárnej regresii*, tzn. regresii X na $Y_1,\dots,Y_k$, kde $X=c+d'Y$. Označíme

$$Y=\begin{pmatrix}Y_1\\ \vdots\\ Y_k\end{pmatrix}\quad,\quad d=\begin{pmatrix}d_1\\ \vdots\\ d_k\end{pmatrix}\quad,$$

$$cov(X,Y)=\left(cov(X,Y_1),\dots,cov(X,Y_k)\right)\quad,\quad cov(Y,X)=\left[cov(X,Y)\right]'\quad.$$

V tejto situácii bude mať optimálna lineárna náhrada veličiny X tvar $\widehat{X}=\gamma+\delta'Y$, kde γ a $\delta_1,\dots,\delta_k$ sú odhady regresných koeficientov c a $d_1,\dots,d_k$. Získame ich rovnakým spôsobom ako pri regresii Y na $X_1,\dots,X_k$, preto stačí, ak vo vzorcoch (4.2) zameníme náhodné veličiny, keďže v tomto prípade by sme minimalizovali strednú kvadratickú chybu $E(X-c-d'Y)^2$. Tým pádom dostaneme

$$\delta=\left[var(Y)\right]^{-1}cov(Y,X)\quad,\quad \gamma=E(X)-\delta'E(Y)\quad. \tag{4.6}$$

Stredné štvorcové odchýlenie $\widehat{X}$ od X, t. j. reziduálny rozptyl, bude

$$E(X-\widehat{X})^2=var(X)-cov(X,Y)\left[var(Y)\right]^{-1}cov(Y,X),$$

ktorý opäť môžeme upraviť pomocou vzťahu (4.6) pre odhad δ

$$E(X-\widehat{X})^2=var(X)-cov(X,Y)\left[var(Y)\right]^{-1}var(Y)\left[var(Y)\right]^{-1}cov(Y,X)=var(X)-\delta' var(Y)\delta.$$

V špeciálnom prípade pre jednorozmernú vysvetľujúcu veličinu Y, kedy je optimálna aproximácia $\widehat{X}=\gamma+\delta Y$, platí pre odhady koeficientov

$$\delta=\frac{cov(X,Y)}{var(Y)}\quad,\quad \gamma=E(X)-\delta E(Y)\quad, \tag{4.7}$$

pričom z vlastností kovariancie platí $cov(Y,X)=cov(X,Y)$. Reziduálny rozptyl

$$E(X-\widehat{X})^2=var(X)-\frac{\left[\rho_{X,Y}\sqrt{var(X)var(Y)}\right]^2}{var(Y)}=var(X)\left(1-\rho_{X,Y}^2\right)\ . \tag{4.8}$$

Teraz si na jednoduchej situácii ukážeme postup pri výpočte regresných koeficientov a reziduálneho rozptylu.

Príklad 4.1 Nech je združené rozdelenie diskrétneho náhodného vektora $(X,Y)'$ dané tabuľkou 4.1.

Tabuľka 4.1

$X\backslash Y$	0	1	?
0	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}=P(X=0)$
1	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{5}{15}=P(X=1)$
2	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{7}{15}=P(X=2)$
?	$\frac{6}{15}$	$\frac{9}{15}$	1

$$P(Y=0) \quad P(Y=1)$$

Následne vypočítame

$$E(X)=\sum_{i=0}^{2} x_i P(X=x_i)=\frac{19}{15} \quad ,$$

$$E(X^2)=\sum_{i=0}^{2} x_i^{\,2} P(X=x_i)=\frac{33}{15} \quad ,$$

$$var(X)=E(X^2)-[E(X)]^2=\frac{134}{225} \quad ,$$

$$E(Y)=\sum_{i=0}^{1} y_i P(Y=y_i)=\frac{9}{15}=E(Y^2) \quad ,$$

$$var(Y)=E(Y^2)-[E(Y)]^2=\frac{6}{25} \quad ,$$

$$E(XY)=\sum x_i y_j P(X=x_i, Y=y_j)=\frac{11}{15} \quad ,$$

$$cov(X,Y)=E(XY)-E(X)E(Y)=\frac{-2}{75} \quad ,$$

potom

$$\beta=\frac{cov(X,Y)}{var(X)}=\frac{-3}{67} \quad ,$$

$$\alpha=E(Y)-\beta E(X)=\frac{44}{67} \quad ,$$

a teda optimálnou lineárnou náhradou náhodnej veličiny Y je

$$\hat{Y}=\frac{44}{67}-\frac{3}{67}X$$

a

$$E(Y-\hat{Y})^2=var(Y)-\frac{[cov(X,Y)]^2}{var(X)}=\frac{16}{67} \quad .$$

Taktiež si môžeme vypočítať korelačný koeficient

$$\rho_{X,Y}=\frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}}=\frac{\frac{-2}{75}}{\sqrt{\frac{134}{225}\cdot\frac{6}{25}}}=-0{,}0705 \quad .$$

Na základe vypočítanej hodnoty korelačného koeficienta môžeme usudzovať, že medzi veličinami je veľmi slabá až nulová lineárna závislosť. Nemusí to však znamenať, že veličiny sú nezávislé. Môže medzi nimi existovať iný typ závislosti než lineárny.

Na rovnakom príklade tiež ukážeme, ako bude vyzerať obrátená regresia, tzn. regresia X na Y. Pre optimálnu lineárnu náhradu veličiny X, $\widehat{X}=\gamma+\delta Y$, majú regresné parametre hodnoty

$$\delta=\frac{cov(X,Y)}{var(Y)}=\frac{-1}{9} \quad ,$$

$$\gamma=E(X)-\delta\,E(Y)=\frac{4}{3} \quad ,$$

čiže

$$\widehat{X}=\frac{4}{3}-\frac{1}{9}Y$$

a

$$E(X-\widehat{X})^2=var(X)-\frac{[cov(X,Y)]^2}{var(Y)}=\frac{16}{27} \quad .$$

Podľa druhej vlastnosti korelačného koeficientu vieme, že $\rho_{Y,X}$ má rovnakú hodnotu ako pri predchádzajúcom výpočte. Vidíme však, že lineárna regresia X na Y nedáva rovnakú regresnú priamku ako regresia Y na X. To preto, že regresná analýza vopred očakáva priamy „riadiaci“ vzťah, t. j. v prvom prípade sa predpokladá, že X ovplyvňuje Y, a nie naopak.

Príklad 4.2 V tomto príklade vyjdeme z tabuľky 4.1, ale ešte pred samotným výpočtom normujeme hodnoty náhodných veličín X, Y podľa vzorca

$$X_N=\frac{X-E(X)}{\sqrt{var(X)}} \quad , \quad Y_N=\frac{Y-E(Y)}{\sqrt{var(Y)}} \quad . \qquad (4.9)$$

Normované hodnoty sú uvedené v tabuľke 4.2. Opäť vypočítame odhady regresných parametrov a hodnotu korelačného koeficienta.

Tabuľka 4.2: Normované hodnoty X_N, Y_N

$X_N \{ Y ¿_N$	$\frac{-109}{89}$	$\frac{40}{49}$	?
$\frac{-151}{92}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15} = P\left(X_N = \frac{-151}{92}\right)$
$\frac{-19}{55}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{5}{15} = P\left(X_N = \frac{-19}{55}\right)$
$\frac{19}{20}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{7}{15} = P\left(X_N = \frac{19}{20}\right)$
	$\frac{6}{15}$	$\frac{9}{15}$	
?	$P\left(Y_N = \frac{-109}{89}\right.$	$P\left(Y_N = \frac{40}{49}\right)$	1

Najprv vypočítame charakteristiky

$$E(X_N)=0 \quad , \quad E(X_N^2)=1 \quad , \quad E(Y_N)=0 \quad , \quad E(Y_N^2)=1 \quad ,$$
$$var(X_N)=1 \quad , \quad var(Y_N)=1 \quad ,$$
$$E(X_N Y_N)=\frac{-6}{85} \quad , \quad cov(X_N, Y_N)=\frac{-6}{85} \quad .$$

Tým pádom

$$\beta_N=\frac{cov(X_N, Y_N)}{var(X_N)}=\frac{-6}{85} \quad ,$$
$$\alpha_N=E(Y_N)-\beta_N E(X_N)=0 \quad .$$

Odhad regresnej priamky má preto tvar

$$\hat{Y}_N=\frac{-6}{85} X_N \quad ,$$

čo znamená, že priamka prechádza počiatkom sústavy súradníc. Reziduálny rozptyl a korelačný koeficient majú hodnoty

$$E(Y_N-\hat{Y}_N)^2=\frac{200}{201} \quad ,$$
$$\rho_{X_N,Y_N}=\frac{cov(X_N, Y_N)}{\sqrt{var(X_N)\cdot var(Y_N)}}=\frac{-6}{85}=-0{,}0705 \quad .$$

Na normovaných hodnotách si ukážeme aj obrátenú regresiu, t. j. X_N na Y_N . Odhad regresnej priamky má v tomto prípade tvar

$$\hat{X}_N=\gamma_N+\delta_N Y_N \quad .$$

Výpočtom získame odhady regresných koeficientov

$$\delta_N = \frac{cov(Y_N, X_N)}{var(Y_N)} = \frac{-6}{85} \quad ,$$

$$\gamma_N = E(X_N) - \delta E(Y_N) = 0 \quad ,$$

a tak

$$\widehat{X}_N = \frac{-6}{85} Y_N \quad .$$

Pre reziduálny rozptyl a korelačný koeficient získame rovnaké hodnoty ako v regresii Y_N na X_N . Navyše podľa piatej vlastnosti korelačného koeficienta je korelačný koeficient rovnaký ako pre pôvodné nenormované premenné. Všimnime si, že v prípade normovaných premenných majú obe regresné priamky rovnakú smernicu a tá sa rovná korelačnému koeficientu. Je to náhoda alebo nutnosť? V nasledujúcich odsekoch sa pokúsime na túto otázku odpovedať.

Vzťah medzi parametrami β a δ

Vo vzťahoch (4.3) a (4.7) pre regresné parametre β a δ si môžeme všimnúť, že majú spoločnú kovarianciu veličín *X* a *Y*. Z tohto dôvodu je tu možnosť nájsť medzi nimi určitú rovnosť. Zo vzorca pre parameter δ vyjadríme kovarianciu, t. j.

$$cov(X, Y) = \delta \cdot var(Y) \quad ,$$

ktorú dosadíme do vzorca pre parameter β , a tým dostaneme hľadanú rovnosť

$$\beta = \frac{cov(X, Y)}{var(X)} = \delta \frac{var(Y)}{var(X)} \quad . \qquad (4.10)$$

Pre náhodné veličiny, ktoré majú rovnaké rozptyly, a špeciálne pre normované veličiny, ktorých rozptyly sú jednotkové, platí $\beta = \delta$.

Rovnaké vzťahy platia aj pri počítaní s empirickými premennými, čiže s výberovými dátami,

$$\beta = \delta \frac{S_Y^2}{S_X^2} \tag{4.11}$$

a pre normované veličiny $\beta = \delta$, kde S_X^2 a S_Y^2 sú výberové rozptyly podľa (3.2) a β a δ sú MNČ-odhady, ktorým sa budeme venovať v podkapitole 4.3.

Vzťah regresného parametra a korelačného koeficienta

V priamkovej regresii $Y = \alpha + \beta X$ môžeme odvodiť isté vzťahy medzi regresným parametrom β a korelačným koeficientom $\rho_{X,Y}$. Zo vzorca (4.3) pre parameter β si vyjadríme kovarianciu veličín X, Y

$$cov(X, Y) = \beta \cdot var(X) \quad .$$

Tento vzťah dosadíme do vzorca (3.1) pre korelačný koeficient $\rho_{X,Y}$ a upravíme

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sqrt{var(X) \cdot var(Y)}} = \frac{\beta \cdot var(X)}{\sqrt{var(X) \cdot var(Y)}} = \beta \frac{\sqrt{var(X)}}{\sqrt{var(Y)}} \quad .$$

V situácii, kedy by platilo, že $var(X) = var(Y)$, by bol korelačný koeficient rovný regresnému parametru β , čo je prípad normovaných premenných.

Podobnú rovnosť dostaneme aj pri obrátenej priamkovej regresii $X = \gamma + \delta Y$, a to medzi regresným parametrom δ a korelačným koeficientom $\rho_{Y,X}$. Najprv zo vzorca (4.7) pre parameter δ opäť vyjadríme kovarianciu

$$cov(X, Y) = \delta \cdot var(Y)$$

a následne dosadíme do vzorca (3.1), pretože z vlastností korelačného koeficienta vieme, že $\rho_{Y,X} = \rho_{X,Y}$. A tak

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sqrt{var(X) \cdot var(Y)}} = \frac{\delta \cdot var(Y)}{\sqrt{var(X) \cdot var(Y)}} = \delta \frac{\sqrt{var(Y)}}{\sqrt{var(X)}} \quad . \qquad (4.12)$$

Takisto ako predtým by pre prípad normovaných premenných bol korelačný koeficient rovný regresnému parametru δ .

Z predchádzajúcich odvodení teda platí

$$\rho_{X,Y}=\rho_{Y,X}=\beta\frac{\sqrt{var(X)}}{\sqrt{var(Y)}}=\delta\frac{\sqrt{var(Y)}}{\sqrt{var(X)}} \quad .$$

a pre normované veličiny

$$\rho_{X,Y}=\rho_{Y,X}=\beta=\delta \quad .$$

Odvodené vzťahy medzi teoretickými regresnými parametrami a teoretickým korelačným koeficientom platia takisto aj pri parametroch a koeficiente, ktoré pracujú s empirickými údajmi. To znamená, že platí

$$r_{X,Y}=r_{Y,X}=\beta\frac{S_X}{S_Y}=\delta\frac{S_Y}{S_X} \quad , \tag{4.13}$$

a pre normované veličiny

$$r_{X,Y}=r_{Y,X}=\beta=\delta \quad ,$$

kde $r_{X,Y}$ je nám už známy výberový (Pearsonov) korelačný koeficient a S_X , S_Y nazývame výberové smerodajné odchýlky, ktoré sú druhou odmocninou výberových rozptylov S_X^2 , S_Y^2 podľa vzorca (3.2) a β a δ sú MNČ-odhady z podkapitoly 4.3.

V súvislosti so vzťahom (4.13) môžeme hovoriť o tzv. *koeficiente determinácie*. V prípade jednoduchej lineárnej regresie je druhou mocninou Pearsonovho koeficienta (vzorec (4.5)) a používa sa k popisu presnosti regresného modelu. Platí preň

$$r_{X,Y}^2=r_{X,Y}\cdot r_{Y,X}=\beta\frac{S_X}{S_Y}\cdot\delta\frac{S_Y}{S_X}=\beta\cdot\delta \tag{4.14}$$

a platí ako pre nenormované, tak i pre normované veličiny. Koeficient determinácie porovnáva odhadované a skutočné hodnoty Y a nadobúda hodnoty v rozsahu od 0 do 1. Ak $r^2_{X,Y}=1$, znamená to, že medzi odhadovanými a skutočnými hodnotami Y nie je žiadny rozdiel. Hovoríme vtedy o dokonalej korelácii, a teda jednoduchá lineárna regresia je úplne vyhovujúca. Ak $r^2_{X,Y}=0$, potom regresná rovnica nie je vôbec vhodná na predpovedanie hodnôt Y.

Normované premenné

Štandardizáciou premennej podľa vzťahu (4.9) získame normovanú premennú, ktorej stredná hodnota sa rovná nule a rozptyl je jednotkový. Prechodom k normovaným veličinám sa ich postavenie v regresii stáva symetrickým, formálne sa stráca rozdiel medzi vysvetľujúcou a vysvetľovanou premennou. Na základe tohto dochádza k zjednodušeniu vzťahov pre normované veličiny.

Na základe predchádzajúcich odsekov sa dostávame k odpovedi na otázku v príklade 4.2, kde obe regresné priamky mali rovnakú smernicu a tá bola rovná korelačnému koeficientu. Z uvedených vzťahov pre normované premenné je zjavné, že nešlo o náhodu.

4.3 Regresia s jednou vysvetľujúcou premennou

Závislú premennú je možné považovať za náhodnú veličinu, ktorá má pri danej hodnote (nenáhodnej) vysvetľujúcej veličiny x určité rozdelenie pravdepodobnosti. Predpokladajme, že stredná hodnota veličiny Y pri danej hodnote x, čo označíme $E(Y|x)$, je rovná hodnote známej funkcie g v bode x,

$$E(Y|x)=g(x;a,b_1,\ldots,b_k),$$

kde $a, b_1, \ldots, b_k$, $k \geq 1$, sú neznáme konštanty (*regresné parametre*), na ktorých funkcia g (*regresná funkcia*) závisí. Regresiou s jednou vysvetľujúcou premennou teda rozumieme závislosť medzi strednou hodnotou náhodnej veličiny Y a jedinou premennou x. V ďalšom texte sa budeme zaoberať vybranými prípadmi, kedy je regresná funkcia lineárnou funkciou parametrov (*lineárna regresia*)

$$g(x) = a + b_1 \varphi_1(x) + \ldots + b_k \varphi_k(x) \quad , \tag{4.15}$$

so známymi funkciami $\varphi_i(x)$, $i = 1, \ldots, k$, premennej x.

Pre porovnanie s podkapitolou 4.2, v tomto prípade uvažujeme viac funkcií (regresorov) jednej nenáhodnej vysvetľujúcej premennej, naproti tomu v teoretických základoch lineárnej regresie vo vzorci (4.1) je uvažovaná jedna alebo viac vysvetľujúcich premenných, ktoré sú náhodné.

4.3.1 Jednoduchá regresná priamka

Uvažujme priamkovú regresiu, kedy pre náhodné veličiny $Y_1, \ldots, Y_n$ (opakované pozorovania veličiny Y) a čísla $x_1, \ldots, x_n$ (pevné hodnoty regresoru x) platí

$$E(Y_i | x_i) = g(x_i) = a + b_1 x_i \quad , \quad i = 1, \ldots, n \quad ,$$

t. j.

$$Y_i = a + b_1 x_i + \varepsilon_i \quad , \quad i = 1, \ldots, n \quad , \quad n \geq 2 \quad , \tag{4.16}$$

kde $x_1, \dots, x_n$ sú také známe reálne čísla, že aspoň dve z nich sú rôzne, a $\varepsilon_1, \dots, \varepsilon_n$ sú náhodné veličiny (náhodné odchýlky, chyby merania). V ďalšom texte predpokladáme, že

$$E(\varepsilon_i)=0 \quad , \quad var(\varepsilon_i)=\sigma^2 \quad , \quad cov(\varepsilon_i, \varepsilon_j)=0 \quad , \quad \forall i \neq j \quad , \quad i, j=1, \dots, n \quad .$$

Ide o predpoklad regularity regresného modelu, ktorý zaručí optimálne vlastnosti pre odhad regresných parametrov. My však tento predpoklad v našich úvahách nebudeme potrebovať. Týmito chybami sa spravidla rozumejú chyby vyplývajúce z nepresností pri stanovovaní veličín Y_i . Predpoklad $E(\varepsilon_i)=0$ odpovedá tomu, že pozorovania veličín Y_i nie sú zaťažené systematickými chybami, vzťah $var(\varepsilon_i)=\sigma^2$ hovorí, že merania jednotlivých veličín Y_i sú vykonávané s rovnakou presnosťou, a $cov(\varepsilon_i, \varepsilon_j)=0$ znamená, že chyby meraní rôznych veličín Y_i sú nekorelované. Parameter a je regresnou konštantou, ktorá vyjadruje, akú hodnotu nadobudne veličina Y , ak x bude mať hodnotu 0. Predstavuje teda priesečník regresnej priamky s osou y. Parameter b_1 udáva priemernú zmenu závislej veličiny Y pri jednotkovej zmene nezávislej premennej x . Vyjadruje sklon regresnej priamky, t. j. smernicu. Nadobúda kladné hodnoty, ak je skúmaná závislosť priama, a záporné hodnoty, ak je závislosť nepriama.

Odhady α , β_1 neznámych parametrov a , b_1 určíme tzv. *metódou najmenších štvorcov*. V tejto metóde požadujeme, aby súčet štvorcov odchýlok pozorovaných hodnôt Y_i od odhadnutých hodnôt $\alpha+\beta_1 x_i$ bol minimálny. Matematicky to popisuje nasledujúca definícia.

Definícia 4.1 Náhodné veličiny α , β_1 , ktoré pre $Y_1, \dots, Y_n$ minimalizujú výraz

$$S(a, b_1)=\sum_{i=1}^{n}(Y_i-a-b_1 x_i)^2 \quad , \qquad (4.17)$$

nazývame *odhady parametrov* a , b_1 *určené metódou najmenších štvorcov.*

Z definície odhadov metódou najmenších štvorcov vyplýva, že α , β_1 vyhovujú sústave rovníc

$$\frac{\partial S(a,b_1)}{\partial a}=0 \quad , \quad \frac{\partial S(a,b_1)}{\partial b_1}=0 \quad .$$

Funkciu $S(a,b_1)$ zderivujeme podľa a i podľa b_1

$$\frac{\partial S(a,b_1)}{\partial a}=2\sum_{i=1}^{n}\left(Y_i-a-b_1x_i\right)(-1)=-2\sum_{i=1}^{n}Y_i+2\mathrm{an}+2b_1\sum_{i=1}^{n}x_i \quad ,$$

$$\frac{\partial S(a,b_1)}{\partial b_1}=2\sum_{i=1}^{n}\left(Y_i-a-b_1x_i\right)(-x_i)=-2\sum_{i=1}^{n}Y_ix_i+2\mathrm{a}\sum_{i=1}^{n}x_i+2b_1\sum_{i=1}^{n}x_i^{\,2} \quad .$$

Parciálne derivácie položíme rovné nule. Po vynásobení $\frac{1}{2}$ a osamostatnení členov s Y_i , dostaneme sústavu dvoch rovníc o dvoch neznámych

$$an+b_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}Y_i \quad , \qquad \text{(a)}$$

$$a\sum_{i=1}^{n}x_i+b_1\sum_{i=1}^{n}x_i^{\,2}=\sum_{i=1}^{n}x_iY_i \quad , \qquad \text{(b)}$$

ktorá sa nazýva *sústava normálnych rovníc*, a jej riešením sú odhady

$$\alpha=\acute{Y}-\beta_1\acute{x} \quad , \qquad (4.18)$$

$$\beta_1=\frac{\sum_{i=1}^{n}\left(x_i-\acute{x}\right)Y_i}{\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2} \quad . \qquad (4.19)$$

Ukážeme si, ako sa tieto tvary odvodia riešením normálnych rovníc (a) a (b). Parameter α získame z prvej rovnice sústavy normálnych rovníc nasledujúcimi úpravami. Označme $\acute{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$, $\acute{Y}=\frac{1}{n}\sum_{i=1}^{n} Y_i$. Potom

$$an=\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} x_i$$

$$a=\frac{\sum_{i=1}^{n} Y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$a=\acute{Y}-b_1\acute{x} \quad .$$

Parameter β_1 dostaneme dosadením $\acute{Y}-b_1\acute{x}$ za a do druhej rovnice sústavy normálnych rovníc a jej úpravou. Teda

$$\left(\acute{Y}-b_1\acute{x}\right)\sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} x_i Y_i$$

$$\sum_{i=1}^{n} x_i \acute{Y} - b_1 \sum_{i=1}^{n} x_i \acute{x} + b_1 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} x_i Y_i$$

$$b_1\left(\sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} x_i \acute{x}\right) = \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \acute{Y}$$

$$b_1=\frac{\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \acute{Y}}{\sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} x_i \acute{x}} \quad .$$

Ďalej je potrebné si uvedomiť, že $\sum_{i=1}^{n} x_i = n\acute{x}$, $\sum_{i=1}^{n} Y_i = n\acute{Y}$ a $\sum_{i=1}^{n} x_i \acute{x} = n\acute{x}\acute{x} = n\acute{x}^2$. Potom

$$b_1=\frac{\sum_{i=1}^{n} x_i Y_i - n\acute{x}\frac{\sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} x_i^{\,2} - 2\sum_{i=1}^{n} x_i \acute{x} + \sum_{i=1}^{n} x_i \acute{x}} = \frac{\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} \acute{x} Y_i}{\sum_{i=1}^{n} x_i^{\,2} - 2\sum_{i=1}^{n} x_i \acute{x} + n\acute{x}^2} = ¿$$

$$¿\frac{\sum_{i=1}^{n}\left(x_i-\acute{x}\right)Y_i}{\sum_{i=1}^{n}\left(x_i^{\,2}-2\,x_i\acute{x}+\acute{x}^2\right)}=\frac{\sum_{i=1}^{n}\left(x_i-\acute{x}\right)Y_i}{\sum_{i=1}^{n}\left(x_i-\acute{x}\right)^2}\quad .$$

Odhady je možné vyjadriť i v iných tvaroch, napríklad

$$\alpha=\frac{\left(\sum_{i=1}^{n}Y_i\right)\left(\sum_{i=1}^{n}x_i^{\,2}\right)-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}x_iY_i\right)}{n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2}\quad , \tag{4.20}$$

$$\beta_1=\frac{n\sum_{i=1}^{n}x_iY_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2}\quad . \tag{4.21}$$

V súvislosti zo vzorcami (4.3) pre teoretické odhady regresných parametrov ide v podstate o tie isté vzťahy, keďže jedny obsahujú teoretickú kovarianciu a teoretický rozptyl a druhé výberovú kovarianciu a výberový rozptyl. Rozdiel je len v tom, že vo vzťahoch (4.3) uvažujeme náhodnú premennú *X*, zatiaľ čo v týchto vzťahoch nenáhodnú premennú *x*.

Odhadom hodnoty regresnej funkcie $a+b_1x$ je štatistika

$$\hat{Y}=\alpha+\beta_1x=\acute{Y}-\beta_1\acute{x}+\beta_1x=\acute{Y}+\beta_1\left(x-\acute{x}\right)\quad .$$

Z jej tvaru je zrejmé, že tzv. „vyrovnané" hodnoty $\hat{Y}$ ležia na priamke, ktorá prechádza bodom $\left(\acute{x}\,,\acute{Y}\right)$ a má smernicu β_1 . Odchýlky

$$e_i=Y_i-\hat{Y}_i=Y_i-\alpha-\beta_1x_i=Y_i-\acute{Y}+\beta_1\acute{x}-\beta_1x_i=Y_i-\acute{Y}-\beta_1\left(x_i-\acute{x}\right),\ i=1,\ldots,n\quad ,$$

sa nazývajú *rezíduá* a štatistika

$$S_e=S\left(\alpha\,,\beta_1\right)=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left[Y_i-\acute{Y}-\beta_1\left(x_i-\acute{x}\right)\right]^2 \tag{4.22}$$

sa nazýva *reziduálny súčet štvorcov*, resp. *súčet štvorcov rezíduí*. Na výpočet S_e sa často používa vzorec

$$S(\alpha, \beta_1) = \sum_{i=1}^{n} Y_i^2 - \alpha \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} x_i Y_i \quad . \tag{4.23}$$

K jeho odvodeniu dospejeme takto

$$S_e = \sum_{i=1}^{n} \left[Y_i - \acute{Y} - \beta_1 \left(x_i - \acute{x} \right) \right]^2 = ¿$$

$$¿ \sum_{i=1}^{n} \left[\left(Y_i - \acute{Y} \right)^2 - 2\beta_1 \left(Y_i - \acute{Y} \right) \left(x_i - \acute{x} \right) + \beta_1^2 \left(x_i - \acute{x} \right)^2 \right] = ¿$$

$$¿ \sum_{i=1}^{n} Y_i^{\,2} - n\acute{Y}^2 - \beta_1 \sum_{i=1}^{n} \left(Y_i - \acute{Y} \right) \left(x_i - \acute{x} \right) = ¿$$

$$¿ \sum_{i=1}^{n} Y_i^{\,2} - n\left(\acute{Y} - \beta_1 \acute{x} + \beta_1 \acute{x} \right) \acute{Y} - \beta_1 \sum_{i=1}^{n} \left(x_i - \acute{x} \right) Y_i = ¿$$

$$¿ \sum_{i=1}^{n} Y_i^{\,2} - \alpha \sum_{i=1}^{n} Y_i - \beta_1 n \acute{x} \acute{Y} - \beta_1 \sum_{i=1}^{n} x_i Y_i + \beta_1 n \acute{x} \acute{Y} = ¿$$

$$¿ \sum_{i=1}^{n} Y_i^{\,2} - \alpha \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} x_i Y_i \quad .$$

Reziduálny súčet štvorcov je empirickou variantou strednej kvadratickej chyby (4.4). Ide o rovnaké vzťahy, avšak s tým, že v tomto prípade uvažujeme nenáhodnú premennú *x*.

Predchádzajúce výpočty môžeme zapísať jednoducho i pomocou vektorov a matíc. Označme

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad , \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad , \qquad b = \begin{pmatrix} a \\ b_1 \end{pmatrix} \quad , \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad .$$

Potom vzťah (4.16) je možné zapísať takto

$$Y = X\, b + \varepsilon$$

a predpoklady na chyby ako

$$E(\varepsilon) = o \quad , \qquad var(\varepsilon) = \sigma^2 I_n \quad ,$$

kde o označuje stĺpcový nulový vektor a I_n jednotkovú maticu. Preto

$$E(Y)=E(Xb+\varepsilon)=E(Xb)+E(\varepsilon)=Xb \quad , \qquad var(Y)=var(\varepsilon)=\sigma^2 I_n \quad .$$

Sústava normálnych rovníc má tvar

$$(X'X)b=X'Y$$

a jej riešenie je

$$\beta=(X'X)^{-1}X'Y \quad . \tag{4.24}$$

Výpočtom zistíme, že

$$X'X=\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} ,$$

$$(X'X)^{-1}=\frac{1}{n\sum_{i=1}^{n} x_i^2-\left(\sum_{i=1}^{n} x_i\right)^2}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \tag{4.25}$$

$$X'Y=\begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix} .$$

A teda

$$\beta = \frac{1}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^{\,2} & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix} = ¿$$

$$¿ \frac{1}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i Y_i \\ n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i \end{pmatrix} = ¿$$

$$¿ \begin{pmatrix} \dfrac{\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i Y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \dfrac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta_1 \end{pmatrix},$$

kde α a β_1 sú rovnaké ako v (4.20) a (4.21).

Všetky uvedené vzorce platia aj v prípade obrátenej regresie, kedy stačí vzájomne zameniť premenné.

Príklad 4.3 ([6], str. 55) Firma zaoberajúca sa predajom potravinárskeho a bežného konzumného tovaru znížila v jednej časti svojich predajní cenu 1 kg cukru o 10 %, v druhej časti o 15 % a v tretej časti o 20 %. V každej z týchto troch častí potom sledovala, o koľko percent sa zvýšilo predané množstvo cukru za mesiac po znížení ceny v porovnaní s mesiacom pred jej znížením. V 9 predajniach boli zistené údaje uvedené v tabuľke 4.3.

Tabuľka 4.3: Závislosť zvýšenia predaného množstva cukru na znížení jeho ceny

Predajňa	% zníženia ceny	% zvýšenia predaja	Pomocné výpočty		
	x_i	Y_i	x_i^2	Y_i^2	$x_i Y_i$
1.	10	6,8	100	46,24	68,0
2.	10	8,6	100	73,96	86,0
3.	10	10,4	100	108,16	104,0
4.	15	15,2	225	231,04	228,0
5.	15	16,9	225	285,61	253,5
6.	15	18,6	225	345,96	279,0
7.	20	23,3	400	542,89	466,0
8.	20	25,2	400	635,04	504,0
9.	20	27,1	400	734,41	542,0
Súčty	135	152,1	2175	3003,31	2530,5

Dosadením do vzorcov (4.20) a (4.21) dostaneme

$$\alpha = \frac{152{,}1 \cdot 2175 - 2530{,}5 \cdot 135}{9 \cdot 2175 - 135^2} = -8 \quad ,$$

$$\beta_1 = \frac{9 \cdot 2530{,}5 - 135 \cdot 152{,}1}{9 \cdot 2175 - 135^2} = 1{,}66 \quad .$$

Ak budeme postupovať podľa vzorca (4.24), dosadíme doň maticu hodnôt x_i a vektor hodnôt Y_i

$$X = \begin{pmatrix} 1 & 10 \\ 1 & 10 \\ 1 & 10 \\ 1 & 15 \\ 1 & 15 \\ 1 & 15 \\ 1 & 20 \\ 1 & 20 \\ 1 & 20 \end{pmatrix}, \quad Y = \begin{pmatrix} 6{,}8 \\ 8{,}6 \\ 10{,}4 \\ 15{,}2 \\ 16{,}9 \\ 18{,}6 \\ 23{,}3 \\ 25{,}2 \\ 27{,}1 \end{pmatrix} \quad .$$

Jednotlivé kroky by vyzerali takto

$$X' X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 10 & 10 & 15 & 15 & 15 & 20 & 20 & 20 \end{pmatrix} \begin{pmatrix} 1 & 10 \\ 1 & 10 \\ 1 & 10 \\ 1 & 15 \\ 1 & 15 \\ 1 & 15 \\ 1 & 20 \\ 1 & 20 \\ 1 & 20 \end{pmatrix} = \begin{pmatrix} 9 & 135 \\ 135 & 2175 \end{pmatrix} \quad ,$$

$$(X' X)^{-1} = \frac{1}{1350} \begin{pmatrix} 2175 & -135 \\ -135 & 9 \end{pmatrix} \quad ,$$

$$X'Y=\begin{pmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\ 10 & 10 & 10 & 15 & 15 & 15 & 20 & 20 & 20\end{pmatrix}\begin{pmatrix}6,8\\ 8,6\\ 10,4\\ 15,2\\ 16,9\\ 18,6\\ 23,3\\ 25,2\\ 27,1\end{pmatrix}=\begin{pmatrix}152,1\\ 2530,5\end{pmatrix}\ ,$$

$$\beta=(X'X)^{-1}X'Y=\frac{1}{1350}\begin{pmatrix}2175 & -135\\ -135 & 9\end{pmatrix}\begin{pmatrix}152,1\\ 2530,5\end{pmatrix}=\frac{1}{1350}\begin{pmatrix}-10800\\ 2241\end{pmatrix}=\begin{pmatrix}-8,00\\ 1,66\end{pmatrix}\ .$$

Regresná priamka má teda tvar

$$\hat{Y}=-8+1,66\,x\ .$$

Keby sme chceli odhadnúť, o koľko percent by „v priemere“ mohlo za daných podmienok vzrásť predané množstvo cukru pri znížení jeho ceny o 18 %, dosadíme do tejto rovnice $x=18$ a dostaneme $Y=-8+1,66\cdot 18=21,88$.

K ďalším výpočtom budeme potrebovať výberové priemery, výberové rozptyly a výberovú kovarianciu. Použijeme vzorce z podkapitoly 3.2 o výberovom korelačnom koeficiente.

$$\acute{x}=\frac{1}{n}\sum_1^n x_i=\frac{1}{9}\sum_1^9 x_i=\frac{135}{9}=15\ ,$$

$$\acute{Y}=\frac{1}{n}\sum_1^n Y_i=\frac{1}{9}\sum_1^9 Y_i=\frac{152,1}{9}=16,9\ ,$$

$$s_X^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\acute{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^n x_i^2-n\acute{x}^2\right)=\frac{1}{8}(2175-9\,15^2)=18,75\ ,$$

$$s_Y^2=\frac{1}{n-1}\sum_{i=1}^n(Y_i-\acute{Y})^2=\frac{1}{n-1}\left(\sum_{i=1}^n Y_i^2-n\acute{Y}^2\right)=\frac{1}{8}(3003,31-9\,16,9^2)=54,1\ ,$$

$$s_{X,Y}=\frac{1}{n-1}\sum_{i=1}^n(x_i-\acute{x})(Y_i-\acute{Y})=\frac{1}{n-1}\sum_{i=1}^n x_iY_i-n\acute{x}\acute{Y}=\frac{1}{8}(2530,5-9\,15\,16,9)=31,125\ .$$

Reziduálny súčet štvorcov vypočítame podľa vzorca (4.23)

$$s_e = s(\alpha, \beta_1) = \sum_{i=1}^{n} Y_i^2 - \alpha \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} x_i Y_i = ¿$$

$$¿\, 3\,003{,}31 + 8\,152{,}1 - 1{,}66\, 2530{,}5 = 19{,}48 \quad .$$

Podľa vzorca (3.4) vypočítame výberový korelačný koeficient

$$r_{X,Y} = \frac{\sum_{i=1}^{n} x_i Y_i - n \acute{x} \acute{Y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n \acute{x}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n \acute{Y}^2\right)}} = \frac{s_{X,Y}}{\sqrt{s_X^2 s_Y^2}} = \frac{31{,}125}{\sqrt{18{,}75\, 54{,}1}} = 0{,}977 \quad .$$

Vypočítaná hodnota korelačného koeficienta vyjadruje, že medzi premennými je silná priama lineárna závislosť. Overme si túto hodnotu podľa vzorca (4.13)

$$r_{X,Y} = \beta \frac{s_X}{s_Y} = 1{,}66 \frac{\sqrt{18{,}75}}{\sqrt{54{,}1}} = 0{,}977 \quad .$$

Určme si takisto koeficient determinácie, ktorý je druhou mocninou korelačného koeficienta.

$$r_{X,Y}^2 = 0{,}977^2 = 0{,}95 \quad .$$

Koeficient je blízky jednej, preto považujeme model lineárnej regresie za veľmi dobrý. Ak vynásobíme koeficient číslom 100, dostaneme v percentách tú časť rozptylu závislej premennej Y, ktorú sa podarilo vysvetliť regresnou priamkou, t. j. 95%.

Vypočítame ešte parameter δ použitím vzorca (4.21), v ktorom zameníme premenné, keďže ide o parameter obrátenej regresie, a použijeme ho opäť k výpočtu koeficienta determinácie, tentokrát však podľa vzorca (4.14)

$$\delta = \frac{n \sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2} = \frac{9 \cdot 2530{,}5 - 135 \cdot 152{,}1}{9 \cdot 3003{,}31 - 152{,}1^2} = 0{,}575 \quad ,$$

$$r^2_{X,Y} = \beta \cdot \delta = 1{,}66\,0{,}575 = 0{,}95 \quad .$$

4.3.2 Regresná priamka prechádzajúca počiatkom

V tomto prípade má regresný model tvar

$$Y_i = b\,x_i + \varepsilon_i \quad , \quad i = 1, \ldots, n \quad . \tag{4.26}$$

Odhadom parametra b je

$$\beta = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^{\,2}} \quad , \tag{4.27}$$

ktorý získame minimalizáciou výrazu

$$S(b)=\sum_{i=1}^{n}\left(Y_i-b\,x_i\right)^2 \quad .$$

Funkciu $S(b)$ zderivujeme podľa b

$$\frac{\partial S(b)}{\partial b}=2\sum_{i=1}^{n}\left(Y_i-b\,x_i\right)\left(-x_i\right) \quad ,$$

položíme rovnú nule a upravíme

$$-2\sum_{i=1}^{n}x_i\left(Y_i-b\,x_i\right)=0$$

$$-2\sum_{i=1}^{n}x_i\,Y_i+2\mathrm{b}\sum_{i=1}^{n}x_i^{\,2}=0 \quad .$$

Vynásobíme $\frac{1}{2}$ a osamostatníme b

$$b=\frac{\sum_{i=1}^{n}x_i\,Y_i}{\sum_{i=1}^{n}x_i^{\,2}} \quad .$$

Odhadom hodnoty regresnej funkcie $b\,x$ je teda štatistika

$$\hat{Y}=\beta\,x \quad .$$

Ak regresný model zapíšeme vo vektorovom tvare, uvidíme, že vektor b má jediný prvok b a že pre maticu X , ktorá je typu $n\times 1$, a pre Y platí

$$X=\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix} \quad , \quad Y=\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix} \quad .$$

Odhad regresného parametra b je v tvare

$$\beta=(X'X)^{-1}X'Y \quad , \tag{4.28}$$

kde

$$X'X=\sum_{i=1}^{n}x_i^2 \quad , \qquad (X'X)^{-1}=\frac{1}{\sum_{i=1}^{n}x_i^2} \quad \text{a} \quad X'Y=\sum_{i=1}^{n}x_iY_i \quad . \tag{4.29}$$

Tieto vzorce majú súvislosť so vzorcami v (4.25). Vidíme však, že zatiaľ čo predtým sme násobením dvoch matíc získali maticu, v tomto prípade sme dostali číslo. Je to tým, že v predošlých vzorcoch ide o maticu **X** typu $n \times 2$ a maticu (vektor) **Y** typu $n \times 1$. V týchto vzorcoch majú obe matice typ $n \times 1$, preto výsledkom ich násobenia je číslo, a nie matica.

Príklad 4.4 Model regresie prechádzajúci počiatkom aplikujeme na príklad 4.3 o predaji cukru. Ak znížime cenu cukru o 0%, predpokladáme, že množstvo predaného cukru sa nezvýši, teda dôjde k 0%-nému zvýšeniu predaja. Táto úvaha nás vedie k tomu, že použitie modelu bez absolútneho člena má zmysel. Dosadením údajov z tabuľky 4.3 do vzorca (4.27) získame odhad parametra b , a to

$$\beta = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \frac{2530{,}5}{2175} = 1{,}163 \quad .$$

Rovnaký výsledok získame dosadením do vzorca (4.28). Najprv však podľa (4.29)

$$(X'X)^{-1} = \frac{1}{\sum_{i=1}^{n} x_i^2} = \frac{1}{2175} \qquad \text{a} \qquad X'Y = \sum_{i=1}^{n} x_i Y_i = 2530{,}5 \quad ,$$

a teda

$$\beta = (X'X)^{-1} X'Y = \frac{1}{2175} \cdot 2530{,}5 = 1{,}163 \quad .$$

Tvar regresnej priamky je

$$\hat{Y} = 1{,}163\, x \quad .$$

Aj v tomto prípade vypočítame, o koľko percent by „v priemere“ mohlo za daných podmienok vzrásť predané množstvo cukru pri znížení jeho ceny o 18 %. Dosadíme $x = 18$ a dostaneme hodnotu $Y = 1{,}163 \cdot 18 = 20{,}93$, ktorá je dosť podobná hodnote v príklade 4.3. Reziduálny súčet štvorcov získame výpočtom

$$s_e = s(\beta) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta x_i\right)^2 = \sum_{i=1}^{n} \left(Y_i^2 - 2\beta x_i Y_i + \beta^2 x_i^2\right) = ¿$$

$$¿ \sum_{i=1}^{n} Y_i^2 - 2\beta \sum_{i=1}^{n} x_i Y_i + \beta^2 \sum_{i=1}^{n} x_i^2 = 3003{,}31 - 2\,1{,}163\,2530{,}5 + 1{,}163^2\,2175 = 59{,}2 \quad .$$

V porovnaní s predchádzajúcim reziduálnym súčtom štvorcov s hodnotou 19,48 v príklade 4.3 je jeho hodnota podstatne väčšia, čo znamená, že v tomto prípade je regresnou priamkou nevysvetlená časť celkového súčtu štvorcov väčšia. Regresná priamka prechádzajúca počiatkom nie je v tomto príklade dostatočne vhodná, čo môže byť spôsobené tým, že závisí len od jedného neznámeho parametra b . Pri vyššom počte parametrov, od ktorých regresná funkcia závisí, získané odhadované hodnoty budú presnejšie, odchýlky od skutočných hodnôt, t. j. reziduá, budú menšie, tým pádom bude menší aj reziduálny súčet štvorcov.

Vypočítame výberový korelačný koeficient podľa vzťahu (3.4) i podľa (4.13). Výberové priemery, výberové rozptyly a výberová kovariancia sú rovnaké ako v príklade 4.3, keďže dáta sa nijako nezmenili.

$$r_{X,Y} = \frac{\sum_{i=1}^{n} x_i Y_i - n\acute{x}\acute{Y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\acute{x}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\acute{Y}^2\right)}} = \frac{s_{X,Y}}{\sqrt{s_X^2 s_Y^2}} = \frac{31{,}125}{\sqrt{18{,}75\,54{,}1}} = 0{,}977 \quad ,$$

$$r_{X,Y} = \beta \frac{s_X}{s_Y} = 1{,}163 \frac{\sqrt{18{,}75}}{\sqrt{54{,}1}} = 0{,}685 \quad .$$

Hodnota 0,977 vypočítaná podľa vzorca (3.4) je zhodná s hodnotou v príklade 4.3, pretože sa na výpočte nič nezmenilo. Avšak pri výpočte podľa vzťahu (4.13) sme získali rozdielne hodnoty. V príklade 4.3 to bola hodnota 0,977, v tomto príklade nám vyšla hodnota 0,685. V predchádzajúcom príklade regresná priamka závisela aj od regresného parametra a, ktorého odhad $\alpha=-8$. Regresná priamka prechádzajúca počiatkom nezávisí od parametru a, keďže $\alpha=0$. Môžeme si však predstaviť, že na grafe by bol priesečník priamky s osou y v tejto situácii vyššie ako v predchádzajúcom príklade. A aby bolo možné preložiť dáta priamkou, ktorá by zároveň prechádzala počiatkom, musí sa zmeniť, resp. zmenšiť sklon regresnej priamky, t. j. smernica, ktorú predstavuje parameter b. Z tohto dôvodu nám vyšla hodnota odhadu parametra menšia ($\beta=1,163$) ako v príklade 4.3 ($\beta=1,66$), ktorá mala zase vplyv na hodnotu korelačného koeficienta podľa vzorca (4.13). Týmto konštatujeme, že vzorec (4.13) platí len pre lineárnu regresiu s absolútnym členom. Pre lineárnu regresiu bez absolútneho člena je potrebné tento vzorec upraviť, skúsme takto

$$r^0_{X,Y}=\beta\frac{s^0_X}{s^0_Y} \quad ,$$

kde $s^0_X=\sqrt{\sum_{i=1}^{n} x_i^2}$, $s^0_Y=\sqrt{\sum_{i=1}^{n} Y_i^2}$. Potom pre naše dáta dostaneme takto upravený korelačný koeficient (nie Pearsonov)

$$r^0_{X,Y}=1,163\frac{\sqrt{2175}}{\sqrt{3003,31}}=0,990 \quad .$$

Vidíme, že sa líši od $r_{X,Y}$ podľa vzorca (3.4), ale ukážeme, že jeho druhá mocnina $\left(r^0_{X,Y}\right)^2=0,980$ sa dá opäť interpretovať ako koeficient determinácie. Vyjdeme zo vzťahu (4.5) pre teoretický index determinácie. Ten bude mať v tomto prípade pre výberové dáta takúto podobu

$$1-\frac{s_e}{\sum_{i=1}^{n} Y_i^2}=\frac{1}{\sum_{i=1}^{n} Y_i^2}\left(\sum_{i=1}^{n} Y_i^2-\sum_{i=1}^{n} Y_i^2+2\beta\sum_{i=1}^{n} x_i Y_i-\beta^2\sum_{i=1}^{n} x_i^2\right)=¿$$

$$¿\frac{1}{\sum_{i=1}^{n} Y_i^2}\left(2\beta^2\sum_{i=1}^{n} x_i^{\,2}-\beta^2\sum_{i=1}^{n} x_i^{\,2}\right)=\beta^2\frac{\sum_{i=1}^{n} x_i^{\,2}}{\sum_{i=1}^{n} Y_i^2}=\left(r_{X,Y}^0\right)^2$$

a hodnotu

$$\left(r_{X,Y}^0\right)^2=1{,}163^2\frac{2175}{3003{,}31}=0{,}980 \quad .$$

Ďalej vypočítame parameter δ pre obrátenú regresiu pomocou (4.27), kde zameníme premenné, a použijeme opäť k výpočtu koeficienta determinácie, ale podľa vzorca (4.14).

$$\delta=\frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} Y_i^{\,2}}=\frac{2530{,}5}{3003{,}31}=0{,}843 \quad ,$$

$$\left(r_{X,Y}^0\right)^2=\beta\cdot\delta=1{,}163\,0{,}843=0{,}980 \quad .$$

Hodnota koeficientu vyjadruje, že regresný model je dostatočne presný a vhodný, keďže regresná priamka prechádzajúca počiatkom vysvetlila 98 % rozptylu závislej premennej.

V tejto kapitole bolo čerpané predovšetkým z literatúry: [6], [7], [8].

5 Asociácia

Asociácia skúma závislosť v situáciách, v ktorých vystupuje alternatívna premenná. Ako vieme, alternatívna (dichotomická) premenná môže nadobúdať iba dve hodnoty, napr. odpovede áno - nie, dobrý - zlý, muž - žena. V prípadoch, v ktorých sa vyskytuje jedna alternatívna premenná, meriame asociáciu tzv. *bodovo biseriálnym korelačným koeficientom*. Ak sú obe premenné alternatívne, asociácia sa meria tzv. *korelačným koeficientom Φ*. Pritom oba koeficienty sú pôvodne odvodené zo vzorca pre Pearsonov výberový korelačný koeficient

$$r_{X,Y}=\frac{\sum_{i=1}^{n} X_i Y_i - n\acute{X}\acute{Y}}{\sqrt{\sum_{i=1}^{n}\left(X_i-\acute{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\acute{Y}\right)^2}} \quad .$$

5.1 Bodovo biseriálny korelačný koeficient

Majme situáciu s dvoma premennými *X*, *Y*, kedy znak X $Alt(p)$, a teda nadobúda iba dve hodnoty. Môže nadobudnúť buď hodnotu $X_i=0$ s pravdepodobnosťou $1-p$, alebo hodnotu $X_i=1$, a to s pravdepodobnosťou p . Je teda zrejmé, že pre alternatívnu premennú platí vzťah $X_i^2=X_i$, ktorý využijeme nižšie pri výpočte. Pre alternatívne rozdelenie ďalej platí, že $E(X)=p$ a $var(X)=p(1-p)$. Výberové veličiny $X_1,\ldots,X_n$ sú nezávislé rovnako rozdelené náhodné veličiny s rozdelením $Alt(p)$, a tak odhad $E(X)=\hat{p}=\acute{X}$ a $var(X)=\acute{X}(1-\acute{X})$. Znak Y je kvantitatívna premenná alebo kvalitatívna premenná, ktorá nadobúda viac než dve hodnoty, a výberové veličiny $Y_1,\ldots,Y_n$ sú takisto nezávislé rovnako rozdelené.

K odvodeniu hľadaného koeficienta budeme potrebovať pomocný výpočet

$$\sum_{i=1}^{n}\left(X_i-\acute{X}\right)^2=\sum_{i=1}^{n}\left(X_i^2-2X_i\acute{X}+\acute{X}^2\right)=\sum_{i=1}^{n}X_i^2-2\acute{X}\sum_{i=1}^{n}X_i+n\acute{X}^2=\sum_{i=1}^{n}X_i-2\mathrm{n}\acute{X}^2+n\acute{X}^2=n\acute{X}-n\acute{X}^2=n\acute{X}(1-\acute{X}$$

.

Potom môžeme výberový korelačný koeficient prepísať do tvaru

$$r_{X,Y}=\frac{\sum_{i=1}^{n} X_i Y_i - n\acute{X}\,\acute{Y}}{n\acute{X}(1-\acute{X})}\,\frac{\sqrt{n\acute{X}(1-\acute{X})}}{\sqrt{\sum_{i=1}^{n}(Y_i-\acute{Y})^2}} \quad . \qquad (5.1)$$

Ďalej keďže platí $1=X_i+(1-X_i)$, potom

$$\sum_{i=1}^{n} Y_i=\sum_{i=1}^{n}\left(X_i+(1-X_i)\right)Y_i=\sum_{i=1}^{n} X_iY_i+\sum_{i=1}^{n}(1-X_i)Y_i \quad .$$

Tento pomocný výpočet použijeme pri samostatnej úprave prvého súčiniteľa vo vzťahu (5.1)

$$\frac{\sum_{i=1}^{n} X_iY_i-n\acute{X}\acute{Y}}{n\acute{X}(1-\acute{X})}=\frac{\sum_{i=1}^{n} X_iY_i(1-\acute{X})+\sum_{i=1}^{n} X_iY_i\acute{X}-n\acute{X}\acute{Y}}{n\acute{X}(1-\acute{X})}=¿$$

$$¿\frac{\sum_{i=1}^{n} X_iY_i}{n\acute{X}}+\frac{\acute{X}\left(\sum_{i=1}^{n} X_iY_i-n\acute{Y}\right)}{n\acute{X}(1-\acute{X})}=\frac{\sum_{i=1}^{n} X_iY_i}{n\acute{X}}+\frac{\sum_{i=1}^{n} X_iY_i-\sum_{i=1}^{n} Y_i}{n(1-\acute{X})}=¿$$

$$¿\frac{\frac{1}{n}\sum_{i=1}^{n} X_iY_i}{\acute{X}}-\frac{\sum_{i=1}^{n} X_iY_i-\left(\sum_{i=1}^{n} X_iY_i+\sum_{i=1}^{n}(1-X_i)Y_i\right)}{n(1-\acute{X})}=¿$$

$$¿\frac{\frac{1}{n}\sum_{i=1}^{n} X_iY_i}{\acute{X}}-\frac{\frac{1}{n}\sum_{i=1}^{n}(1-X_i)Y_i}{(1-\acute{X})} \quad .$$

Pri poslednej úprave sme získali dva zlomky, ktoré označíme ako $M_1=\frac{\frac{1}{n}\sum_{i=1}^{n} X_iY_i}{\acute{X}}$ a

$M_0=\frac{\frac{1}{n}\sum_{i=1}^{n}(1-X_i)Y_i}{(1-\acute{X})}$. Podľa vzťahu (3.2) vieme, že $\sum_{i=1}^{n}(Y_i-\acute{Y})^2=\sum_{i=1}^{n} Y_i^2-n\acute{Y}^2$. Dosadením do vzťahu (5.1) nakoniec dostaneme

$$r_{X,Y}=(M_1-M_0)\frac{\sqrt{n\acute{X}(1-\acute{X})}}{\sqrt{\sum_{i=1}^{n} Y_i^2-n\acute{Y}^2}} \quad .$$

Definícia 5.1 Korelačný koeficient

$$r^{b}_{X,Y}=\left(M_1-M_0\right)\frac{\sqrt{n\acute{X}\left(1-\acute{X}\right)}}{\sqrt{\sum_{i=1}^{n}Y_i^2-n\acute{Y}^2}}\quad , \tag{5.2}$$

kde $M_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i}{\acute{X}}$ predstavuje priemer tých Y_i, pre ktoré $X_i = 1$, a

$M_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} (1 - X_i) Y_i}{(1 - \acute{X})}$ predstavuje priemer tých Y_i, pre ktoré $X_i = 0$, sa nazýva sa *bodovo biseriálny korelačný koeficient*.

5.2 Korelačný koeficient Φ

Pozorujeme dva kvalitatívne alternatívne štatistické znaky X a Y vo forme dvojrozmerného náhodného výberu $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ o rozsahu n prvkov. Pre odhad rozptylov použijeme, podobne ako v podkapitole 5.1,

$$\bar{var}(X) = \acute{X}(1 - \acute{X}) = S_X^2 ,$$
$$\bar{var}(Y) = \acute{Y}(1 - \acute{Y}) = S_Y^2 .$$

Keďže alternatívny znak nadobúda len dve hodnoty, môžeme všetky prvky náhodného výberu rozdeliť do štyroch skupín podľa znakov X a Y a hodnôt, ktoré nadobudli. Takto získame nasledujúce početnosti

n_{11} … počet dvojíc X, Y, kde $X = 1$ a $Y = 1$,
n_{01} ... počet dvojíc X, Y, kde $X = 0$ a $Y = 1$,
n_{10} … počet dvojíc X, Y, kde, $X = 1$ a $Y = 0$,
n_{00} … počet dvojíc X, Y, kde $X = 0$ a $Y = 0$,

ktoré môžeme zapísať do štvorpoľnej kontingenčnej tabuľky 2x2

$X \backslash Y$	0	1	?
0	n_{00}	n_{01}	$n_{0.}$

1	n_{10}	n_{11}	$n_{1\cdot}$
?	$n_{\cdot 0}$	$n_{\cdot 1}$	n

K odvodeniu koeficienta pre takúto situáciu použijeme nasledujúce pomocné výpočty:

$$\acute{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}\left(n_{10}+n_{11}\right)=\frac{1}{n}n_{1\cdot}\quad,$$

$$\acute{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1}{n}\left(n_{01}+n_{11}\right)=\frac{1}{n}n_{\cdot 1}\quad,$$

$$1-\acute{X}=1-\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}\sum_{i=1}^{n}\left(1-X_i\right)=\frac{1}{n}\left(n_{00}+n_{01}\right)=\frac{1}{n}n_{0\cdot}\quad,$$

$$1-\acute{Y}=1-\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1}{n}\sum_{i=1}^{n}\left(1-Y_i\right)=\frac{1}{n}\left(n_{00}+n_{10}\right)=\frac{1}{n}n_{\cdot 0}\quad.$$

Na odvodenie koeficienta použijeme už získaný vzorec (5.2) pre bodovo biseriálny korelačný koeficient. Keďže v súčasnom prípade má nielen X, ale aj Y alternatívne rozdelenie, takisto ako v predchádzajúcej podkapitole 5.1 platí, že $\sum_{i=1}^{n}\left(Y_i-\acute{Y}\right)^2=n\acute{Y}\left(1-\acute{Y}\right)$. Po dosadení a niekoľkých úpravách dostaneme

$$r_{X,Y}=\left[\frac{\frac{n_{11}}{n}}{\frac{n_{1\cdot}}{n}}-\frac{\frac{n_{01}}{n}}{\frac{n_{0\cdot}}{n}}\right]\sqrt{\frac{n\frac{n_{1\cdot}}{n}\frac{n_{0\cdot}}{n}}{n\frac{n_{\cdot 1}}{n}\frac{n_{\cdot 0}}{n}}}=¿$$

$$¿\left(\frac{n_{11}}{n_{1\cdot}}-\frac{n_{01}}{n_{0\cdot}}\right)\sqrt{\frac{n_{1\cdot}n_{0\cdot}}{n_{\cdot 1}n_{\cdot 0}}}=¿$$

$$¿\frac{n_{11}\left(n_{01}+n_{00}\right)-n_{01}\left(n_{10}+n_{11}\right)}{n_{1\cdot}n_{0\cdot}}\sqrt{\frac{n_{1\cdot}n_{0\cdot}}{n_{\cdot 1}n_{\cdot 0}}}=¿$$

$$¿\frac{\left[n_{11}\left(n_{01}+n_{00}\right)-n_{01}\left(n_{10}+n_{11}\right)\right]}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}=\frac{n_{11}n_{00}-n_{01}n_{10}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}.$$

Definícia 5.2 Výraz

$$\Phi=\frac{n_{11}n_{00}-n_{01}n_{10}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}} \qquad (5.3)$$

nazývame *korelačný koeficient* Φ .

Ako zaujímavé doplnenie tejto podkapitoly si odvodíme vzorec pre smernicu β . Zo vzťahu (4.13) vieme, že

$$r_{X,Y}=\beta\frac{S_X}{S_Y} \quad .$$

Vyjadrením parametra β a nahradením koeficienta $r_{X,Y}$ koeficientom Φ získame vzťah pre smernicu regresie Y na X, kde $X \quad Alt$, $Y \quad Alt$, t. j.

$$\beta=\Phi\frac{S_Y}{S_X} \quad .$$

Do vzorca dosadíme odpovedajúce vzťahy, pričom

$$S_X=\sqrt{S_X^2}=\sqrt{\acute{X}(1-\acute{X})}=\sqrt{\frac{1}{n}n_{1\bullet}\frac{1}{n}n_{0\bullet}}=\sqrt{\frac{n_{1\bullet}}{n}\frac{n_{0\bullet}}{n}} \quad ,$$

$$S_Y=\sqrt{S_Y^2}=\sqrt{\acute{Y}(1-\acute{Y})}=\sqrt{\frac{1}{n}n_{\bullet 1}\frac{1}{n}n_{\bullet 0}}=\sqrt{\frac{n_{\bullet 1}}{n}\frac{n_{\bullet 0}}{n}} \quad ,$$

upravíme

$$\beta=\frac{n_{11}n_{00}-n_{01}n_{10}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 1}n_{\bullet 0}}}\frac{\sqrt{\frac{n_{1\bullet}}{n}\frac{n_{0\bullet}}{n}}}{\sqrt{\frac{n_{\bullet 1}}{n}\frac{n_{\bullet 0}}{n}}}=¿$$

$$¿\frac{n_{11}n_{00}-n_{01}n_{10}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 1}n_{\bullet 0}}}\frac{\sqrt{n_{1\bullet}n_{0\bullet}}}{\sqrt{n_{\bullet 1}n_{\bullet 0}}}\frac{\sqrt{n_{\bullet 1}n_{\bullet 0}}}{\sqrt{n_{\bullet 1}n_{\bullet 0}}}=¿$$

$$¿\frac{n_{11}n_{00}-n_{01}n_{10}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 1}n_{\bullet 0}}}\frac{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 1}n_{\bullet 0}}}{n_{\bullet 1}n_{\bullet 0}} \quad ,$$

a teda

$$\beta=\frac{n_{11}n_{00}-n_{01}n_{10}}{n_{\bullet 1}n_{\bullet 0}}. \qquad (5.4)$$

6 Použitie korelačných koeficientov

Posledná kapitola je venovaná použitiu korelačných koeficientov na získaných dátach. V nasledujúcich príkladoch budeme pracovať s rôznymi typmi dát, či už s kvantitatívnymi, alebo kvalitatívnymi, na ktoré, ak to bude možné, aplikujeme všetky naše korelačné koeficienty. Zhrňme si, akými koeficientami sme sa zaoberali v predchádzajúcich kapitolách. V kapitole 3 sme sa venovali Pearsonovmu korelačnému koeficientu $r_{X,Y}$ (3.4), ktorý je základom pre ďalšie koeficienty a používa sa v prípade kvantitatívnych premenných. Na konci tejto kapitoly sme spomenuli takisto Spearmanov korelačný koeficient r_S (3.5), ktorý používame pri práci s ordinálnymi premennými. V kapitole 5 sme sa oboznámili s ďalšími dvoma koeficientami, a to s bodovo biseriálnym korelačným koeficientom $r^b_{X,Y}$ (5.2), ktorý využívame v prípadoch, kedy máme jednu alternatívnu premennú, a koeficientom Φ (5.3), ak obe premenné sú alternatívneho typu. V každom príklade najprv určíme, ktorý z korelačných koeficientov je pre výpočet najvhodnejší a potom s jeho hodnotou porovnáme vypočítané hodnoty ostatných koeficientov. V niektorých príkladoch taktiež vypočítame regresné parametre, určíme regresnú priamku a overíme vzťahy regresných parametrov a korelačného koeficienta, ktoré sme odvodili v kapitole 4.

Príklad 6.1 V tabuľke 6.1 máme údaje ([5], str. 40) o 27 vybraných pozemkoch, na ktorých poľnohospodárske družstvá v určitej oblasti pestujú ozimný jačmeň. Na týchto dátach budeme skúmať intenzitu i priebeh závislosti hektárového výnosu jačmeňa (t/ha) od nadmorskej výšky pozemku (m).

Najprv prevedieme potrebné numerické výpočty

$$\sum_{i=1}^{27} X_i = 9379, \quad \bar{X} = \frac{\sum_{i=1}^{27} X_i}{27} = 347{,}3704,$$

$$\sum_{i=1}^{27} X_i^2 = 3443403,$$

$$\sum_{i=1}^{27} Y_i = 124{,}6 \quad , \quad \bar{Y} = \frac{\sum_{i=1}^{27} Y_i}{27} = 4{,}6148 \quad ,$$

$$\sum_{i=1}^{27} Y_i^2 = 595{,}48 \quad ,$$

$$\sum_{i=1}^{27} X_i Y_i = 41477{,}4 \quad .$$

Tabuľka 6.1

i	Nadmorská výška (m) X_i	Hektárový výnos (t/ha) Y_i	i	Nadmorská výška (m) X_i	Hektárový výnos (t/ha) Y_i	i	Nadmorská výška (m) X_i	Hektárový výnos (t/ha) Y_i
1	215	6,3	10	301	5,2	19	395	4,1
2	220	6,5	11	315	4,9	20	408	3,9
3	228	5,9	12	332	4,6	21	420	4,2
4	246	5,8	13	340	4,4	22	437	3,7
5	256	5,5	14	346	3,9	23	445	3,5
6	260	5,6	15	355	5,1	24	460	4,1
7	272	4,8	16	360	4,3	25	468	3,6
8	281	4,9	17	372	4,4	26	475	3,3
9	295	4,6	18	388	4,1	27	489	3,4

Teraz už môžeme vypočítať Pearsonov korelačný koeficient, keďže máme dve kvantitatívne premenné, ktorý predstavuje mieru závislosti medzi nadmorskou výškou a hektárovým výnosom. Použijeme vzorec (3.4)

$$r_{X,Y}=\frac{\sum_{i=1}^{27} X_i Y_i - n\acute{X}\acute{Y}}{\sqrt{\left(\sum_{i=1}^{27} X_i^2 - n\acute{X}^2\right)\left(\sum_{i=1}^{27} Y_i^2 - n\acute{Y}^2\right)}}$$

$$¿\frac{41477{,}4-27\cdot 347{,}3704\cdot 4{,}6148}{\sqrt{\left(3443403-27\cdot 347{,}3704^2\right)\left(595{,}48-27\cdot 4{,}6148^2\right)}}=-0{,}9262 \quad .$$

Hodnota koeficienta značí o silnej nepriamej korelácii, čo znamená, že so zvyšujúcou sa nadmorskou výškou pozemku klesá hektárový výnos jačmeňa. Zaujímať nás teda bude tvar regresnej priamky $\hat{Y}=\alpha+\beta_1 X$. Jej regresné parametre získame použitím vzorcov (4.20) a (4.21), t. j.

$$\alpha=\frac{\left(\sum_{i=1}^{27} Y_i\right)\left(\sum_{i=1}^{27} X_i^2\right)-\left(\sum_{i=1}^{27} X_i\right)\left(\sum_{i=1}^{27} X_i Y_i\right)}{n\sum_{i=1}^{27} X_i^2-\left(\sum_{i=1}^{27} X_i\right)^2}=\frac{124{,}6\cdot 3443403-9379\cdot 41477{,}4}{27\cdot 3443403-9379^2}=7{,}9963 \quad ,$$

$$\beta_1=\frac{n\sum_{i=1}^{27} X_i Y_i-\left(\sum_{i=1}^{27} X_i\right)\left(\sum_{i=1}^{27} Y_i\right)}{n\sum_{i=1}^{27} X_i^2-\left(\sum_{i=1}^{27} X_i\right)^2}=\frac{27\cdot 41477{,}4-9379\cdot 124{,}6}{27\cdot 3443403-9379^2}=-0{,}00973 \quad .$$

Regresná priamka má tvar $\hat{Y}=7{,}9963-0{,}00973\,X$. Vypočítame si aj reziduálny súčet štvorcov podľa vzorca (4.23)

$$S(\alpha,\beta_1)=\sum_{i=1}^{27} Y_i^2-\alpha\sum_{i=1}^{27} Y_i-\beta_1\sum_{i=1}^{27} X_i Y_i$$

$$¿595{,}48-7{,}9963\cdot 124{,}6+0{,}00973\cdot 41477{,}4=2{,}7161 \quad .$$

Rovnaký postup zopakujeme aj v prípade obrátenej regresie, t. j. závislosti nadmorskej výšky od hektárového výnosu. V tejto situácii hľadáme regresné parametre priamky $\widehat{X}=\gamma+\delta_1 Y$. Táto obrátená regresia má význam napríklad v prípade, že poznáme hektárové výnosy, ale nepoznáme nadmorské výšky, v ktorých boli dosiahnuté jednotlivé hektárové výnosy. Pearsonov korelačný koeficient nemusíme opäť počítať, pretože vieme, že je symetrický, a teda jeho hodnota je takisto $-0{,}9262$. Zámenou premenných vo vzorcoch dostaneme

$$\gamma=\frac{\left(\sum_{i=1}^{27} X_i\right)\left(\sum_{i=1}^{27} Y_i^2\right)-\left(\sum_{i=1}^{27} Y_i\right)\left(\sum_{i=1}^{27} X_i Y_i\right)}{n\sum_{i=1}^{27} Y_i^2-\left(\sum_{i=1}^{27} Y_i\right)^2}=\frac{9379\cdot 595{,}48-124{,}6\cdot 41477{,}4}{27\cdot 595{,}48-124{,}6^2}=754{,}202 \quad ,$$

$$\delta_1=\frac{n\sum_{i=1}^{27} X_i Y_i-\left(\sum_{i=1}^{27} Y_i\right)\left(\sum_{i=1}^{27} X_i\right)}{n\sum_{i=1}^{27} Y_i^2-\left(\sum_{i=1}^{27} Y_i\right)^2}=\frac{27\cdot 41477{,}4-124{,}6\cdot 9379}{27\cdot 595{,}48-124{,}6^2}=-88{,}1577$$

a

$$S(\gamma,\delta_1)=\sum_{i=1}^{27} X_i^2-\gamma\sum_{i=1}^{27} X_i-\delta_1\sum_{i=1}^{27} X_i Y_i=¿$$

$$¿\,3443403-754{,}202\cdot 9379+88{,}1577\cdot 41477{,}4=26294{,}628 \quad .$$

Regresná priamka je $\widehat{X}=754{,}202-88{,}1577\,Y$.

Keďže máme vypočítané parametre oboch regresií, ukážeme, že platia vzťahy medzi nimi navzájom a medzi parametrami a Pearsonovým korelačným koeficientom. Medzi parametrami β a δ má platiť rovnosť (4.11), t. j.

$$\beta=\delta\frac{S_Y^2}{S_X^2} \quad .$$

Vypočítame výberové rozptyly

$$S_X^2=\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2-n\acute{X}^2\right)=\frac{1}{27-1}\left(3443403-27\cdot 347{,}3704^2\right)=7131{,}3746 \quad ,$$

$$S_Y^2=\frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2-n\acute{Y}^2\right)=\frac{1}{27-1}\left(595{,}48-27\cdot 4{,}6148^2\right)=0{,}7876 \quad ,$$

dosadíme

$$\beta = -88{,}1577 \cdot \frac{0{,}7876}{7131{,}3746} = -0{,}00973$$

a vidíme, že nám vyšla rovnaká hodnota parametra β ako predtým. Ďalší vzťah je (4.13)

$$r_{X,Y} = r_{Y,X} = \beta \frac{S_X}{S_Y} = \delta \frac{S_Y}{S_X} \quad .$$

Stačí len dosadiť, keďže $S_X = \sqrt{S_X^2}$, $S_Y = \sqrt{S_Y^2}$, a skutočne platí

$$r_{X,Y} = r_{Y,X} = -0{,}00973 \cdot \frac{\sqrt{7131{,}3746}}{\sqrt{0{,}7876}} = -88{,}1577 \cdot \frac{\sqrt{0{,}7876}}{\sqrt{7131{,}3746}} = -0{,}926 \quad .$$

Nakoniec potvrdíme rovnosť vzťahu (4.14), kde

$$r_{X,Y}^2 = r_{X,Y} \cdot r_{Y,X} = \beta \frac{S_X}{S_Y} \cdot \delta \frac{S_Y}{S_X} = \beta \cdot \delta \quad .$$

Najprv si vypočítame koeficient determinácie ako druhú mocninu Pearsonovho koeficienta

$$r_{X,Y}^2 = (-0{,}9262)^2 = 0{,}8578$$

a teraz ako súčin parametrov β a δ

$$r_{X,Y}^2 = \beta \cdot \delta = -0{,}00973 \cdot (-88{,}1577) = 0{,}8578 \quad .$$

Rovnosť je jednoznačná.

Teraz skúsme vypočítať i ďalšie koeficienty. Ako prvý použijeme Spearmanov koeficient, kvôli ktorému však potrebujeme zistiť poradia hodnôt. Keďže sa v tabuľke 6.1 opakujú niektoré hodnoty hektárového výnosu, je nutné v ich prípade určiť priemerné poradie. Poradia sú uvedené v tabuľke 6.2

Tabuľka 6.2

i	Poradie nadmorskej výšky R_i	Poradie hektárového výnosu Q_i	i	Poradie nadmorskej výšky R_i	Poradie hektárového výnosu Q_i	i	Poradie nadmorskej výšky R_i	Poradie hektárového výnosu Q_i
1	1	26	10	10	21	19	19	9
2	2	27	11	11	18,5	20	20	6,5
3	3	25	12	12	15,5	21	21	11
4	4	24	13	13	13,5	22	22	5
5	5	22	14	14	6,5	23	23	3
6	6	23	15	15	20	24	24	9
7	7	17	16	16	12	25	25	4
8	8	18,5	17	17	13,5	26	26	1
9	9	15,5	18	18	9	27	27	2

Najprv vypočítame

$$\sum_{i=1}^{n}\left(R_i-Q_i\right)^2=6332$$

a dosadením do vzorca (3.5) dostaneme

$$r_S=1-\frac{6}{n\left(n^2-1\right)}\sum_{i=1}^{n}\left(R_i-Q_i\right)^2=1-\frac{6\cdot 6332}{27\cdot\left(27^2-1\right)}=-0{,}933 \quad .$$

Vyšla nám hodnota blízka hodnote Pearsonovho koeficienta (-0,9262), keďže Spearmanov koeficient je definovaný ako Pearsonov výberový korelačný koeficient.

Ďalej použijeme na dáta biseriálny koeficient podľa vzorca (5.2), hoci vieme, že nemáme žiadnu alternatívnu premennú. Zistíme M_1 a M_0, pričom $\sum_{i=1}^{n}\left(1-X_i\right)Y_i=-41352{,}8$,

$$M_1=\frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i}{\acute{X}}=\frac{\frac{1}{27}41477{,}4}{347{,}3704}=4{,}422 \quad ,$$

$$M_0=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(1-X_i\right)Y_i}{\left(1-\acute{X}\right)}=\frac{\frac{1}{27}\left(-41352{,}8\right)}{1-347{,}3704}=4{,}422 \quad .$$

Získali sme rovnaké hodnoty, čo má za následok, že koeficient bude mať hodnotu

$$r_{X,Y}^{b}=0\frac{\sqrt{n\acute{X}\left(1-\acute{X}\right)}}{\sqrt{\sum_{i=1}^{n}Y_i^2-n\acute{Y}^2}}=0 \quad ,$$

a teda je v tomto príklade nepoužiteľný.

Koeficient Φ nie je možné podľa vzorca (5.3) vypočítať, keďže sú potrebné početnosti podľa kontingenčnej tabuľky 2x2, ktoré pri týchto dátach nemôžeme získať.

Príklad 6.2 Zo známej internetovej filmovej databázy [www.imdb.com] sme vybrali 20 filmov, ktoré sa nachádzajú ako v rebríčku najobľúbenejších filmov hodnotených ženskými užívateľmi, tak aj v rebríčku najobľúbenejších filmov hodnotených mužskými užívateľmi tejto internetovej stránky. Určili sme poradie filmov podľa ich ratingov najprv podľa hodnotenia ženami, a potom podľa hodnotenia mužmi. Obe poradia sú zapísané v tabuľke 6.3.

Tabuľka 6.3

Film	*Ženy*	*Muži*	*Film*	*Ženy*	*Muži*
i	R_i	Q_i	i	R_i	Q_i
1	1	2	11	11	12
2	2	5	12	12	6
3	3	1	13	13	14
4	4	8	14	14	17
5	5	11	15	15	19
6	6	10	16	16	15
7	7	7	17	17	13
8	8	9	18	18	20
9	9	4	19	19	16
10	10	3	20	20	18

Na začiatok si vypočítame

$$Q_i = ¿\,210$$

$$R_i = ¿\sum_{i=1}^{20} ¿ \quad ,$$

$$\sum_{i=1}^{20} ¿$$

$$R_i^2 = \sum_{i=1}^{20} Q_i^2 = ¿\,2870$$

$$\sum_{i=1}^{20} ¿ \quad ,$$

$$R_i Q_i = ¿\,2743$$

$$\sum_{i=1}^{20} ¿ \quad ,$$

$$\acute{R} = \frac{\sum_{i=1}^{20} R_i}{20} = \frac{\sum_{i=1}^{20} Q_i}{20} = \acute{Q} = 10{,}5 \quad ,$$

$$\sum_{i=1}^{n} \left(R_i - Q_i\right)^2 = 254 \quad .$$

Ako prvé zistíme hodnotu Spearmanovho korelačného koeficienta, ktorý sa používa práve v prípadoch, ak máme k dispozícii iba poradie hodnôt skúmaných premenných. Použijeme vzorec (3.5)

$$r_S = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} \left(R_i - Q_i\right)^2 = 1 - \frac{6 \cdot 254}{20 \cdot \left(20^2-1\right)} = 0{,}809 \quad .$$

Hodnota koeficienta je pomerne vysoká, čo svedčí o silnej priamej korelácii medzi poradiami zostavenými ženami a mužmi, a teda medzi obľúbenosťou filmov u žien a u mužov. Všimnúť si to môžeme aj v tabuľke, kde vysokému poradiu v stĺpci ženy (t. j. 1, 2, 3) odpovedá takisto vysoké poradie v stĺpci muži (t. j. 2, 5, 1). Pokračujeme výpočtom regresných parametrov priamky $\hat{Y} = \alpha + \beta_1 X$. Pritom $R_i = X_i$ a $Q_i = Y_i$. Používame rovnaké vzorce (4.20), (4.21) a (4.23) pre reziduálny súčet štvorcov

$$\alpha=\frac{\left(\sum_{i=1}^{20}Y_i\right)\left(\sum_{i=1}^{20}X_i^2\right)-\left(\sum_{i=1}^{20}X_i\right)\left(\sum_{i=1}^{20}X_iY_i\right)}{n\sum_{i=1}^{20}X_i^2-\left(\sum_{i=1}^{20}X_i\right)^2}=\frac{210\cdot 2870-210\cdot 2743}{20\cdot 2870-210^2}=2{,}0052\quad ,$$

$$\beta_1=\frac{n\sum_{i=1}^{20}X_iY_i-\left(\sum_{i=1}^{20}X_i\right)\left(\sum_{i=1}^{20}Y_i\right)}{n\sum_{i=1}^{20}X_i^2-\left(\sum_{i=1}^{20}X_i\right)^2}=\frac{20\cdot 2743-210\cdot 210}{20\cdot 2870-210^2}=0{,}809\quad ,$$

$$S(\alpha,\beta_1)=\sum_{i=1}^{20}Y_i^2-\alpha\sum_{i=1}^{20}Y_i-\beta_1\sum_{i=1}^{20}X_iY_i=¿$$

$$¿2870-2{,}0052\cdot 210-0{,}809\cdot 2743=229{,}821\quad .$$

A tak sme dostali priamku tvaru

$$\widehat{Y}=2{,}0052+0{,}809\,X\quad .$$

Tak ako v príklade 6.1, tak aj v tomto sa pozrieme na obrátenú regresiu a hľadanie priamky $\widehat{X}=\gamma+\delta_1 Y$. Ak sa však zamyslíme, uvedomíme si, že priamka bude mať rovnaké hodnoty regresných parametrov, ako mali parametre α a β_1 . Vidieť je to na numerických výpočtoch, ktoré sme vykonali na začiatku príkladu (platí rovnosť súčtu poradia R_i a súčtu poradia Q_i , rovnosť súčtov ich druhých mocnín, a takisto rovnosť ich priemerov). Tým pádom by sme síce vo vzorcoch (4.20), (4.21) a (4.23) navzájom zamenili X_i a Y_i , ale do vzorcov by sme vlastne dosadili rovnaké hodnoty ako predtým. Preto môžeme hneď napísať

$$\gamma=\alpha=2{,}0052\quad ,$$
$$\delta_1=\beta_1=0{,}809\quad ,$$
$$S(\alpha,\beta_1)=S(\gamma,\delta_1)=229{,}821\quad ,$$
$$\widehat{X}=2{,}0052+0{,}809\,Y\quad .$$

Spearmanov korelačný koeficient $r_S = 0{,}809$ sa taktiež nezmení. Overenie vzťahu $\beta = \delta \frac{S_Y^2}{S_X^2}$ je jednoduché. Keďže platí $S_X^2 = S_Y^2$, tak $\beta = \delta$.

Pre dáta v tabuľke 6.3 skúsme spočítať Pearsonov korelačný koeficient podľa vzorca (3.4), pričom opäť $R_i = X_i$ a $Q_i = Y_i$. Jeho hodnota je

$$r_{X,Y} = \frac{\sum_{i=1}^{n} X_i Y_i - n \acute{X} \acute{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n \acute{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n \acute{Y}^2\right)}} = \frac{2743 - 20\,10{,}5\,10{,}5}{\sqrt{\left(2870 - 20\,10{,}5^2\right)\left(2870 - 20\,10{,}5^2\right)}} = 0{,}809 \quad ,$$

ktorá je zhodná s hodnotou Spearmanovho koeficienta, pretože ten, ako sme dokázali v podkapitole 3.3, je definovaný ako výberový korelačný koeficient počítaný z dvojíc $\left(R_1, Q_1\right)', \ldots, \left(R_n, Q_n\right)'$.

Takisto skúsme zistiť hodnotu biseriálneho koeficienta, kde znova $R_i = X_i$ a $Q_i = Y_i$. Najprv

$$M_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i}{\acute{X}} = \frac{\frac{1}{20} 2743}{10{,}5} = 13{,}062 \quad ,$$

$$M_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} \left(1 - X_i\right) Y_i}{\left(1 - \acute{X}\right)} = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i Y_i\right)}{\left(1 - \acute{X}\right)} = \frac{\frac{1}{20}\left(210 - 2743\right)}{1 - 10{,}5} = 13{,}332 \quad ,$$

potom

$$r_{X,Y}^b = \left(M_1 - M_0\right) \frac{\sqrt{n \acute{X}\left(1 - \acute{X}\right)}}{\sqrt{\sum_{i=1}^{n} Y_i^2 - n \acute{Y}^2}} = \left(13{,}062 - 13{,}332\right) \frac{\sqrt{20\,10{,}5\left(1 - 10{,}5\right)}}{\sqrt{2870 - 20\,10{,}5^2}} \quad .$$

Pod odmocninou v čitateli nám vychádza záporné číslo, preto nie je možné tento koeficient v tomto prípade použiť.

Koeficient Φ opäť nie je možné vypočítať podľa vzorca (5.3).

Príklad 6.3 Údaje pre tento príklad bolo získané položením dvoch otázok náhodne vybraným 20 študentom bez ohľadu na pohlavie. Otázky zneli: „Aká je tvoja hmotnosť v kilogramoch?“ a „Zvykneš jesť večer ťažšie jedlá alebo viac, ako by si mal/a?“ Tieto dve otázky si skrátene označíme ako hmotnosť (znak *Y*) a jedenie večer (znak *X*). Získané údaje sú zapísané v tabuľke 6.4. Pre dáta vypočítame hodnotu bodovo biseriálneho korelačného koeficienta, keďže hmotnosť je kvantitatívnym znakom a jedenie večer je kvalitatívnym znakom s odpoveďami áno a nie, ktoré si ohodnotíme číslami 1 a 0. V tabuľke 6.4 sa nachádzajú aj niektoré pomocné výpočty.

Pred samotným výpočtom koeficienta budeme ešte potrebovať priemer X_i a priemer Y_i a tie sú

$$\acute{X}=\frac{9}{20}=0{,}45 \quad , \quad \acute{Y}=\frac{1387}{20}=69{,}35 \quad .$$

Najprv si vypočítame M_1 a M_0 podľa definície 5.1

$$M_1=\frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i}{\acute{X}}=\frac{\frac{1}{20}672}{0{,}45}=\frac{224}{3} \quad ,$$

$$M_0=\frac{\frac{1}{n}\sum_{i=1}^{n} (1-X_i) Y_i}{(1-\acute{X})}=\frac{\frac{1}{20}715}{1-0{,}45}=65 \quad .$$

Potom hodnota bodovo biseriálneho koeficienta bude

$$r^b_{X,Y}=(M_1-M_0)\frac{\sqrt{n\acute{X}(1-\acute{X})}}{\sqrt{\sum_{i=1}^{n} Y_i^2-n\acute{Y}^2}}=\left(\frac{224}{3}-65\right)\frac{\sqrt{20\,0{,}45(1-0{,}45)}}{\sqrt{100321-20\,69{,}35^2}}=0{,}335 \quad .$$

Tabuľka 6.4

i	*Jedenie večer*	*Ohodnotenie* X_i	*Hmotnosť* Y_i	Y_i^2	$X_i Y_i$	$(1-X_i)Y$
1	nie	0	75	5625	0	75
2	nie	0	50	2500	0	50
3	nie	0	62	3844	0	62

4	áno	1	46	2166	46	0
5	áno	1	78	6084	78	0
6	nie	0	86	7396	0	86
7	nie	0	75	5625	0	75
8	áno	1	83	6889	83	0
9	nie	0	63	3969	0	63
10	nie	0	67	4489	0	67
11	áno	1	94	8836	94	0
12	áno	1	54	2916	54	0
13	nie	0	65	4225	0	65
14	áno	1	94	8836	94	0
15	áno	1	72	5184	72	0
16	nie	0	49	2401	0	49
17	nie	0	61	3721	0	61
18	áno	1	61	3721	61	0
19	nie	0	62	3844	0	62
20	áno	1	90	8100	90	0
□	-	9	1387	100321	672	715

Táto hodnota nám značí, že medzi hmotnosťou študentov a ich jedením večer je mierna priama korelácia.

Skúsme pre naše dáta spočítať hodnotu Pearsonovho korelačného koeficienta podľa vzorca (3.4)

$$r_{X,Y}=\frac{\sum_{i=1}^{n} X_i Y_i - n\acute{X}\acute{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\acute{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\acute{Y}^2\right)}}=\frac{672-20\,0{,}45\,69{,}35}{\sqrt{(9-20\,0{,}45^2)(100321-20\,69{,}35^2)}}=0{,}335 \quad .$$

Vidíme, že oba koeficienty majú rovnakú hodnotu, pretože ako vieme, bodovo biseriálny koeficient je odvodený od Pearsonovho korelačného koeficienta pre prípad, že jedna z premenných je alternatívna.

Pre výpočet Spearmanovho koeficienta zistíme poradia hodnôt a zapíšeme do tabuľky 6.5

Tabuľka 6.5

i	*Poradie hmotnosti* R_i	*Poradie jedenia večer* Q_i	*i*	*Poradie hmotnosti* R_i	*Poradie jedenia večer* Q_i

1	1	16	11	11	6
2	2	6	12	12	16
3	3	6	13	13,5	6
4	4	16	14	13,5	6
5	5,5	6	15	15	16
6	5,5	16	16	16	16
7	7,5	6	17	17	6
8	7,5	6	18	18	16
9	9	6	19	19,5	16
10	10	6	20	19,5	16

Najprv si vypočítame

$$\sum_{i=1}^{n}\left(R_i-Q_i\right)^2=838$$

a dosadíme do vzorca (3.5)

$$r_S=1-\frac{6}{n\left(n^2-1\right)}\sum_{i=1}^{n}\left(R_i-Q_i\right)^2=1-\frac{6\cdot 838}{20\cdot\left(20^2-1\right)}=0{,}370 \quad .$$

Hodnota Spearmanovho koeficienta sa približuje k hodnote $r^b_{X,Y}=0{,}335$, čiže vzorec (3.5) funguje i v tomto prípade, kedy máme iba dve rôzne hodnoty pre poradie Q_i , a teda je takisto použiteľný.

Skúsime vypočítať aj koeficient Φ a to tak, že z hmotnosti spravíme alternatívnu premennú Z . Jednotlivé hodnoty hmotnosti rozdelíme do dvoch skupín. Hodnoty, ktoré budú menšie ako $\acute{Y}=69{,}35$, ohodnotíme číslom 0 a hodnoty väčšie ako tento priemer ohodnotíme číslom 1. Takto sme dosiahli, že obe premenné, jedenie večer (znak X) a hmotnosť (znak Z), sú alternatívneho typu. Tým pádom môžeme zistiť jednotlivé početnosti, ktoré uvedieme do tabuľky 6.6.

Tabuľka 6.6

X\Z	nie	áno	?
nie	8	3	11
áno	3	6	9
?	11	9	20

Koeficient Φ podľa (5.3) bude mať hodnotu

$$\Phi = \frac{n_{11} n_{00} - n_{01} n_{10}}{\sqrt{n_{1.} n_{0.} n_{.1} n_{.0}}} = \frac{68 - 33}{\sqrt{9\,11\,9\,11}} = 0{,}394 \quad .$$

Vypočítaná hodnota je relatívne blízko k hodnote $r^{b}_{X,Y} = 0{,}335$, z čoho usudzujeme, že vzťah (5.3) pre koeficient Φ môžeme k výpočtu taktiež použiť.

Príklad 6.4 Údaje pre tento príklad boli opäť získané položením dvoch otázok náhodne vybraným 20 študentom bez ohľadu na pohlavie. Otázky zneli: „Je súčasťou tvojho každodenného jedálnička i zelenina (aspoň 2-krát, 3-krát denne)?“ a „Zvykneš sa prejedať?“ Znovu si otázky skrátene označíme ako zelenina (znak *X*) a prejedanie (znak *Y*). Získané údaje sú zapísané v tabuľke 6.7. Pre dáta vypočítame hodnotu korelačného koeficienta Φ , keďže máme dva kvalitatívne znaky s odpoveďami áno a nie.

Najprv si však zistíme jednotlivé početnosti a zapíšeme ich do tabuľky 6.8, ktorá je kontingenčnou tabuľkou 2x2.

Tabuľka 6.7

i	*Zelenina*	*Ohodnotenie* X_i	*Prejedanie*	*Ohodnotenie* Y_i	$X_i Y_i$
1	áno	1	nie	0	0
2	áno	1	nie	0	0
3	áno	1	nie	0	0
4	áno	1	nie	0	0
5	nie	0	áno	1	0
6	nie	0	nie	0	0
7	nie	0	áno	1	0
8	nie	0	áno	1	0
9	áno	1	nie	0	0
10	áno	1	nie	0	0
11	nie	0	nie	0	0
12	áno	1	áno	1	1
13	nie	0	nie	0	0
14	áno	1	nie	0	0
15	nie	0	áno	1	0
16	áno	1	nie	0	0

17	áno	1	áno	1	1
18	nie	0	nie	0	0
19	áno	1	nie	0	0
20	áno	1	áno	1	1
□	-	12	-	7	3

Tabuľka 6.8

X\Y	nie	áno	?
nie	4	4	8
áno	9	3	12
?	13	7	20

Potom podľa vzorca (5.3) dostaneme hodnotu

$$\Phi = \frac{n_{11} n_{00} - n_{01} n_{10}}{\sqrt{n_{1\cdot} n_{0\cdot} n_{\cdot 1} n_{\cdot 0}}} = \frac{34 - 49}{\sqrt{12\,8\,7\,13}} = -0{,}257 \quad .$$

ktorá nám vypovedá o slabej nepriamej korelácii medzi prejedaním sa a jedením zeleniny.

Pre dáta skúsme spočítať tiež hodnotu Pearsonovho korelačného koeficienta podľa vzorca (3.4). K tomuto výpočtu potrebujeme ohodnotenia odpovedí podľa oboch znakov, ktoré už máme pripravené v tabuľke 6.7. Ďalej budeme potrebovať priemer X_i a priemer Y_i, ktoré sú

$$\acute{X} = \frac{12}{20} = 0{,}6 \quad , \quad \acute{Y} = \frac{7}{20} = 0{,}35 \quad .$$

Ako bolo povedané v podkapitole 5.1, platí $X_i^2 = X_i$, preto $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i$ a podobne i pre Y_i^2. Potom Pearsonov korelačný koeficient

$$r_{X,Y} = \frac{\sum_{i=1}^{n} X_i Y_i - n \acute{X} \acute{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n \acute{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n \acute{Y}^2\right)}} = \frac{3 - 20\,0{,}6\,0{,}35}{\sqrt{\left(12 - 20\,0{,}6^2\right)\left(7 - 20\,0{,}35^2\right)}} = -0{,}257 \quad .$$

Pre oba koeficienty nám vyšla rovnaká hodnota, pretože k odvodeniu koeficienta Φ v podkapitole 5.2 sme síce použili vzorec pre bodovo biseriálny koeficient, ten je však odvodený od Pearsonovho korelačného koeficienta.

Ďalej vyskúšajme použiť Spearmanov korelačný koeficient. Samozrejme je potrebné určiť poradia hodnôt R_i a Q_i. Tie sme zapísali do tabuľky 6.9.

Tabuľka 6.9

i	*Poradie zeleniny*	*Poradie prejedania*	*i*	*Poradie zeleniny*	*Poradie prejedania*

	R_i	Q_i		R_i	Q_i
1	14,5	7	11	4,5	7
2	14.5	7	12	14,5	17
3	14,5	7	13	4,5	7
4	14,5	7	14	14,5	7
5	4,5	17	15	4,5	17
6	4,5	7	16	14,5	7
7	4,5	17	17	14,5	17
8	4,5	17	18	4,5	7
9	14,5	7	19	14,5	7
10	14,5	7	20	14,5	17

Ako prvé vypočítame

$$\sum_{i=1}^{n}\left(R_i-Q_i\right)^2=1175$$

a dosadíme do (3.5)

$$r_S=1-\frac{6}{n\left(n^2-1\right)}\sum_{i=1}^{n}\left(R_i-Q_i\right)^2=1-\frac{6\cdot 1175}{20\cdot\left(20^2-1\right)}=1{,}117 \quad .$$

Vidíme, že hodnota Spearmanovho koeficienta vôbec nesúhlasí s hodnotou koeficienta $\Phi=-0{,}257$, preto je Spearmanov koeficient úplne nepoužiteľný v našom príklade, v ktorom sú obe premenné alternatívne.

Teraz skúsme na dáta použiť bodovo biseriálny koeficient. Najprv si vypočítame M_1 a M_0 podľa definície 5.1

$$M_1=\frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i}{\acute{X}}=\frac{\frac{1}{20}3}{0{,}6}=0{,}25 \quad ,$$

$$M_0=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(1-X_i\right)Y_i}{\left(1-\acute{X}\right)}=\frac{\frac{1}{20}4}{1-0{,}6}=0{,}5 \quad .$$

Potom hodnota bodovo biseriálneho koeficienta bude

$$r^b_{X,Y}=\left(M_1-M_0\right)\frac{\sqrt{n\acute{X}\left(1-\acute{X}\right)}}{\sqrt{\sum_{i=1}^{n}Y_i^2-n\acute{Y}^2}}=\left(0{,}25-0{,}5\right)\frac{\sqrt{20\,0{,}6\left(1-0{,}6\right)}}{\sqrt{7-20\,0{,}35^2}}=-0{,}257 \quad .$$

Znovu sme dostali rovnakú hodnotu, pretože podľa teórie sme koeficient Φ odvodili od bodovo biseriálneho koeficienta.

Pre zaujímavosť si ďalej vypočítajme smernicu β podľa vzťahu (5.4)

$$\beta = \frac{n_{11} n_{00} - n_{01} n_{10}}{n_{\bullet 1} n_{\bullet 0}} = \frac{34 - 49}{7\,13} = -0{,}264 \quad .$$

Smernica má zápornú hodnotu, t. j. skúmaná závislosť je nepriama, čo korešponduje so zápornou hodnotou koeficienta Φ .

Záver

V bakalárskej práci sme sa venovali téme závislosti premenných. Zisťovali sme, či medzi premennými existuje vôbec nejaká závislosť a ak áno, chceli sme ju kvantitatívne ohodnotiť pomocou korelačného koeficienta, prípadne následne zistiť jej charakter pomocou regresnej priamky. K tomu všetkému sme samozrejme potrebovali naštudovať a porozumieť teoretickým poznatkom, odvodeniam, rôznym vlastnostiam a vykonať ich dôkazy, čo sa nám aj podarilo.

Cieľom bolo takisto zistiť vzťahy medzi určitými pojmami, napr. aký je vzťah regresného parametra a korelačného koeficienta, tieto vzťahy dokázať a nakoniec overiť na reálnych dátach. Tento cieľ sa nám podarilo splniť. V druhej kapitole sme dokázali vzťah nezávislosti a kovariancie, v tretej kapitole vzťah nezávislosti a korelácie a vzťah kovariancie a korelácie a vo štvrtej kapitole zase vzťah medzi regresnými parametrami β a δ a medzi regresným parametrom a korelačným koeficientom. Dokonca sme sa okrajovo zaoberali aj normovanými veličinami, kedy sme zistili, že v ich prípade dochádza k zjednodušeniu vzťahu medzi regresnými parametrami β a δ a vzťahu regresného parametra a korelačného koeficienta. Celkovo nám všetky tieto vzťahy môžu uľahčiť a urýchliť výpočty.

Ďalším cieľom bolo nájsť rôzne možnosti ako definovať korelačný koeficient. U tohto cieľa sme dospeli k štyrom korelačným koeficientom, pričom každý je vhodný pre inú dvojicu premenných. Pearsonov koeficient používame pri kvantitatívnych premenných, ale v šiestej kapitole sme sa presvedčili, že je možné ho použiť aj pri iných typoch premenných, napr. pri ordinálnych alebo alternatívnych, keďže je v podstate základom, od ktorého sú potom odvodené zvyšné koeficienty. Spearmanov koeficient je vhodný pre ordinálne premenné, koeficient Φ pre dve alternatívne premenné a bodovo biseriálny koeficient je vhodný v prípade, ak jedna z premenných je alternatívna. Použitím týchto troch koeficientov v príkladoch s premennými, pre ktoré nebol daný koeficient úplne vhodný, sme dostali buď hodnotu koeficienta blízku hodnote najvhodnejšieho koeficienta, alebo sme zistili, že daný koeficient je nepoužiteľný. Môžeme teda skonštatovať, že na výpočet intenzity závislosti premenných je najlepšie použiť koeficient, ktorý je na daný prípad určený, alebo použiť Pearsonov korelačný koeficient.

Spracovávanie tejto témy bolo pre mňa zo začiatku trocha ťažšie, keďže najprv bolo potrebné zorientovať sa v celej teórii a pochopiť jednotlivé súvislosti. Postupne som však začala tomu viac a viac rozumieť, videla som svoj posun v práci, čo spôsobilo, že ma to začalo viac baviť. Najzaujímavejšie pre mňa bolo použitie vzorcov na reálnych dátach a porovnávanie jednotlivých výsledkov.

Zoznam použitej literatúry

[1] Anděl, J.: Matematická statistika. SNTL/Alfa, Praha, 1978.

[2] Anděl, J.: Statistické metody. MATFYZPRESS, Praha, 1998.

[3] Anděl, J.: Základy matematické statistiky. MATFYZPRESS, Praha, 2007.

[4] Cyhelský, L., Hustopecký, J., Závodský, P.: Příklady k základům statistiky. SNTL/Alfa, Praha, 1988.

[5] Cyhelský, L.: Statistika v příkladech. SNTL, Praha, 1967.

[6] Hindls, R., Kaňoková, J., Novák, I.: Metody statistické analýzy pro ekonomy. Management Press, Praha, 1997

[7] Hron, K., Kunderová, P.: Základy počtu pravděpodobnosti a metod matematické statistiky. Vydavatelství Univerzity Palackého, Olomouc, 2013.

[8] Kunderová, P.: Základy pravděpodobnosti a matematické statistiky. Vydavatelství Univerzity Palackého, Olomouc, 2004.

[9] Základy statistiky v příkladech [online], dostupné z: http://www1.osu.cz/home/Gajda/Zaklady%20statistiky%20v%20prikladech.pdf

[10] Štatistika [online], dostupné z: http://www.ff.unipo.sk/kvdsp/files/predmety/Statistika/Statistika4SP%2BA.pdf

[11] Štatistická analýza závislosti [online], dostupné z: http://www.sjf.tuke.sk/transferinovacii/pages/archiv/transfer/14-2009/pdf/246-248.pdf

[12] Matematická statistika [online], dostupné z: http://pef-info.wz.cz/download/MSIIb_prednasky.pdf

[13] Zdanlivá korelácia [online], dostupné z: http://akce.fs.vsb.cz/2000/KonfFS04/Proceedings/papers/26.pdf

[14] Aplikovaná statistika [online], dostupné z: http://www.vscht.cz/ktk/www_324/lab/texty/statistika/as.pdf

[15] Základné poznatky štatistiky [online], dostupné z: http://www.svit.luba.szm.com/html/zaklad_poznat_stat.htm#%C5%A0tatistick%C3%BD_znak_a_jeho_delenie_

[16] Úvod do štatistiky [online], dostupné z: http://eldum.phil.muni.cz/mod/resource/view.php?id=1715

[17] Měření závislosti ve statistice [online], dostupné z: fsi.uniza.sk/kkm/old/zamestnanci/novak/p_09.doc