# Korelační analýza – řešený příklad

Máme zadány tyto údaje o proměnných *x* a *y*:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 5 | 5 | 7 | 8 | 9 | 9 | 11 | 12 | 14 | 15 |
| $y_i$ | 25 | 23 | 22 | 22 | 20 | 19 | 19 | 17 | 16 | 16 |

Vypočtěte rovnice sdružených regresních přímek a interpretujte hodnoty obou sdružených regresních koeficientů. Dále vypočítejte hodnotu korelačního koeficientu a koeficientu determinace – obě hodnoty interpretujte. Ověřte významnost koeficientu korelace pomocí vhodného testu na hladině významnosti 5 %.

## *Řešení v SGP (STATGRAPHICS Centurion XVIII):*

Postup řešení je uveden rovnou v programu SGP, neboť výpočty začínají být složitější a časově náročnější a není zde oprávněný předpoklad, že by se příklady podobného typu řešily ručně.

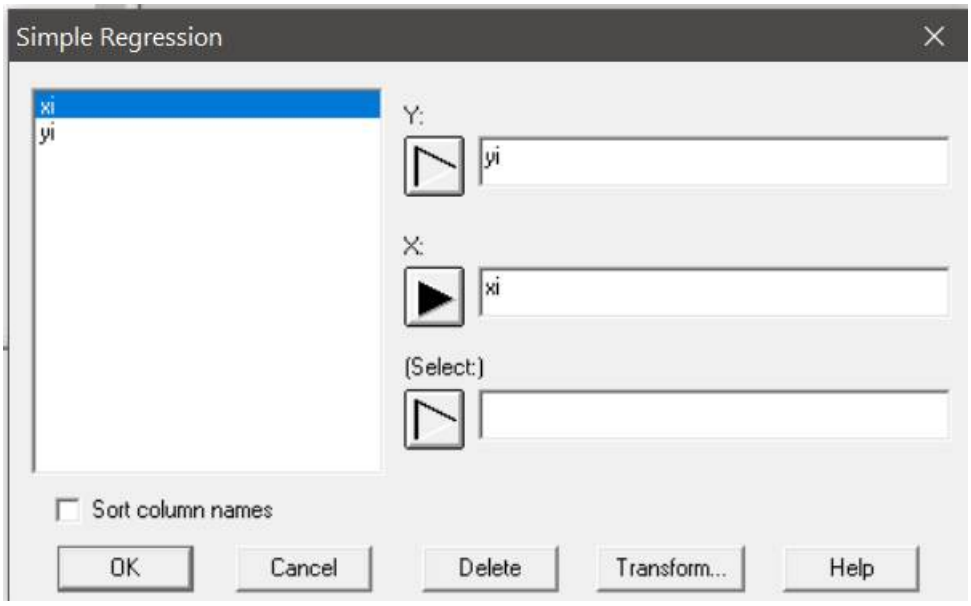
Hodnoty proměnných *x* a *y* zadáme tradičním způsobem do sloupců v DataBook, jak ukazuje Obr. 1.

<untitled>

| | xi | yi |
|---|---|---|
| | Numeric | Numeric |
| 1 | 5 | 25 |
| 2 | 5 | 23 |
| 3 | 7 | 22 |
| 4 | 8 | 22 |
| 5 | 9 | 20 |
| 6 | 9 | 19 |
| 7 | 11 | 19 |
| 8 | 12 | 17 |
| 9 | 14 | 16 |
| 10 | 15 | 16 |

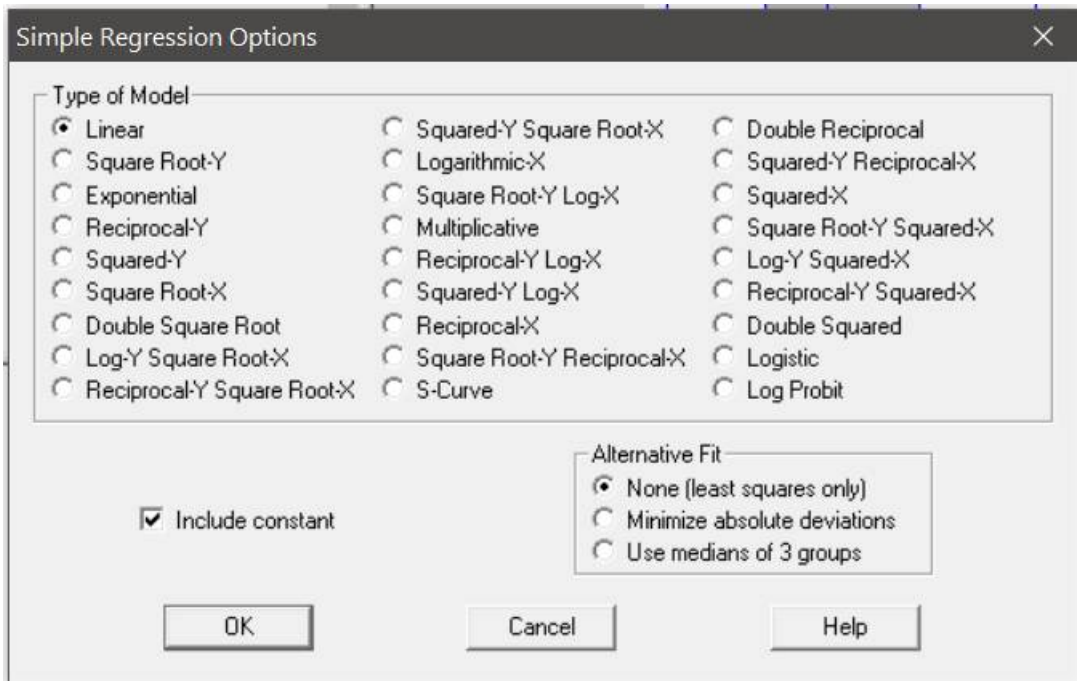
*Obrázek 1 – Zadání dat*

Rovnici přímky lze odvodit v již známé posloupnosti procedur Relate – One Factor – Simple Regression… Do řádku označeného Y nejprve zadáme proměnnou *y* a do řádku označeného X zadáme proměnnou *x*, jak je vidět na Obr. 2.




*Obrázek 2 – Vstupní panel Simple Regression*

Po stisknutí tlačítka OK se objeví nabídka regresních modelů (viz Obr. 3), ve které ponecháme označeno *Linear*.




*Obrázek 3 – Simple Regression – volba regresního modelu*

Po stisknutí tlačítka OK se objeví nabídka *Tables and Graphs*, ve které nemusíme nic měnit. Zaměříme se na výstup Analysis Summary:

**Simple Regression - yi vs. xi**
Dependent variable: yi
Independent variable: xi
Linear model: Y = a + b*X
Number of observations: 10

**Coefficients**

| | *Least Squares* | *Standard* | *T* | |
|---|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Statistic* | *P-Value* |
| Intercept | 27,9991 | 0,838792 | 33,3803 | 0,0000 |
| Slope | -0,852535 | 0,0834216 | -10,2196 | 0,0000 |

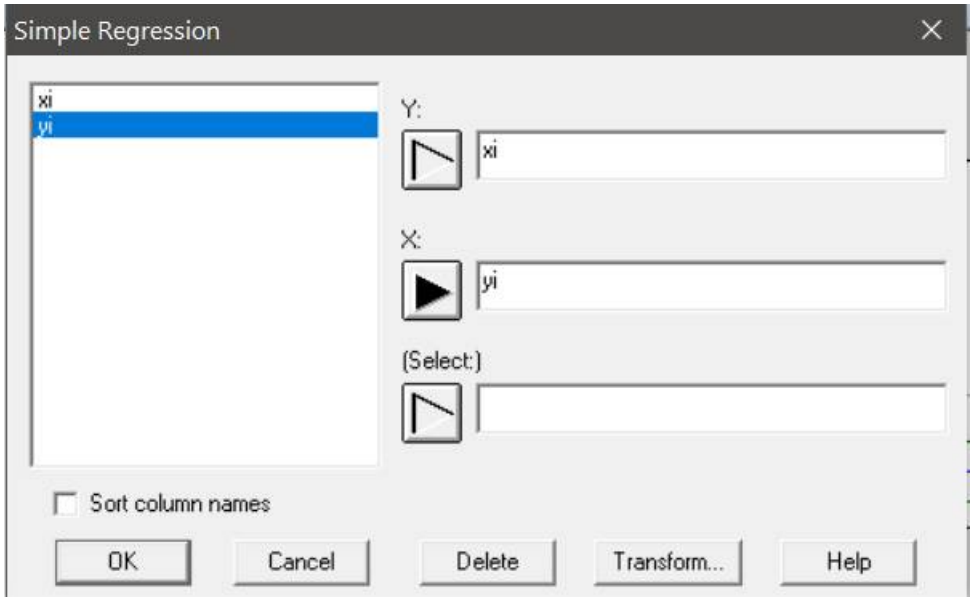
**Analysis of Variance**

| *Source* | *Sum of Squares* | *Df* | *Mean Square* | *F-Ratio* | *P-Value* |
|---|---|---|---|---|---|
| Model | 78,8594 | 1 | 78,8594 | 104,44 | 0,0000 |
| Residual | 6,04055 | 8 | 0,755069 | | |
| Total (Corr.) | 84,9 | 9 | | | |

Correlation Coefficient = -0,963769
R-squared = 92,8851 percent
R-squared (adjusted for d.f.) = 91,9957 percent
Standard Error of Est. = 0,868947
Mean absolute error = 0,650507
Durbin-Watson statistic = 2,15022 (P=0,4355)
Lag 1 residual autocorrelation = -0,258791

Z tabulky Coefficients, z 2. sloupce, odečteme hodnoty parametrů výběrové regresní přímky a zkonstruujeme její rovnici: $Y = 27{,}999 - 0{,}853x$. Provedeme kontrolu vhodnosti modelu a jeho parametrů pomocí celkového F-testu a individuálních t-testů. Už zde tyto testy nejsou uvedeny podrobně, neboť byly detailně probrány v příkladu u Tématu 4. P-Value u celkového F-testu je 0,0000, jak je vidět v posledním sloupci tabulky *Analysis of Variance*, proto zamítáme $H_0$ a přijímáme $H_1$, která tvrdí, že přímka je vhodná pro popis dané závislosti. Vzhledem k tomu, že P-Value u obou t-testů jsou rovněž rovny 0,0000, prokázali jsme, že oba parametry regresní přímky jsou statisticky významné. Tato přímka tedy popisuje závislost $y$ na $x$.

Druhou regresní přímku, která bude popisovat závislost $x$ na $y$, zkonstruujeme ve stejné posloupnosti procedur, tj. Relate – One Factor – Simple Regression… jen s tím rozdílem, že do řádku označeného Y nejprve zadáme proměnnou $x$ a do řádku označeného X zadáme proměnnou $y$, jak je vidět na Obr. 4.




*Obrázek 4 – Vstupní panel Simple Regression*

Opět necháme zvolený model *Linear* a ani v nabídce *Tables and Graphs* není potřeba nic měnit.

Zaměříme se opět na výstup *Analysis Summary*:

**Simple Regression - xi vs. yi**
Dependent variable: xi
Independent variable: yi
Linear model: Y = a + b*X
Number of observations: 10

**Coefficients**

| | *Least Squares* | *Standard* | *T* | |
|---|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Statistic* | *P-Value* |
| Intercept | 31,1814 | 2,14417 | 14,5424 | 0,0000 |
| Slope | -1,08952 | 0,106611 | -10,2196 | 0,0000 |

**Analysis of Variance**

| *Source* | *Sum of Squares* | *Df* | *Mean Square* | *F-Ratio* | *P-Value* |
|---|---|---|---|---|---|
| Model | 100,78 | 1 | 100,78 | 104,44 | 0,0000 |
| Residual | 7,71967 | 8 | 0,964959 | | |
| Total (Corr.) | 108,5 | 9 | | | |

Correlation Coefficient = -0,963769
R-squared = 92,8851 percent
R-squared (adjusted for d.f.) = 91,9957 percent
Standard Error of Est. = 0,982323
Mean absolute error = 0,773145
Durbin-Watson statistic = 2,121 (P=0,4183)
Lag 1 residual autocorrelation = -0,234147

Na základě údajů ve druhém sloupci tabulky *Coefficients* sestavíme rovnici regresní přímky: $X = 31{,}181 - 1{,}090y$. Celkový F-test je významný (P-Value = 0,0000) a individuální t-testy jsou rovněž významné (P-Value = 0,000 v obou případech). Přímka je tedy vhodným modelem i pro popis závislosti proměnné *x* na proměnné *y*.
*Poznámka:* V obou případech jsem výběrové regresní parametry zaokrouhlila na 3 desetinná místa.

*Shrnutí*: Sdružené regresní přímky $Y = 27{,}999 - 0{,}853x$ a $X = 31{,}181 - 1{,}090y$ popisují vzájemnou (oboustrannou) závislost proměnných *x* a *y*.

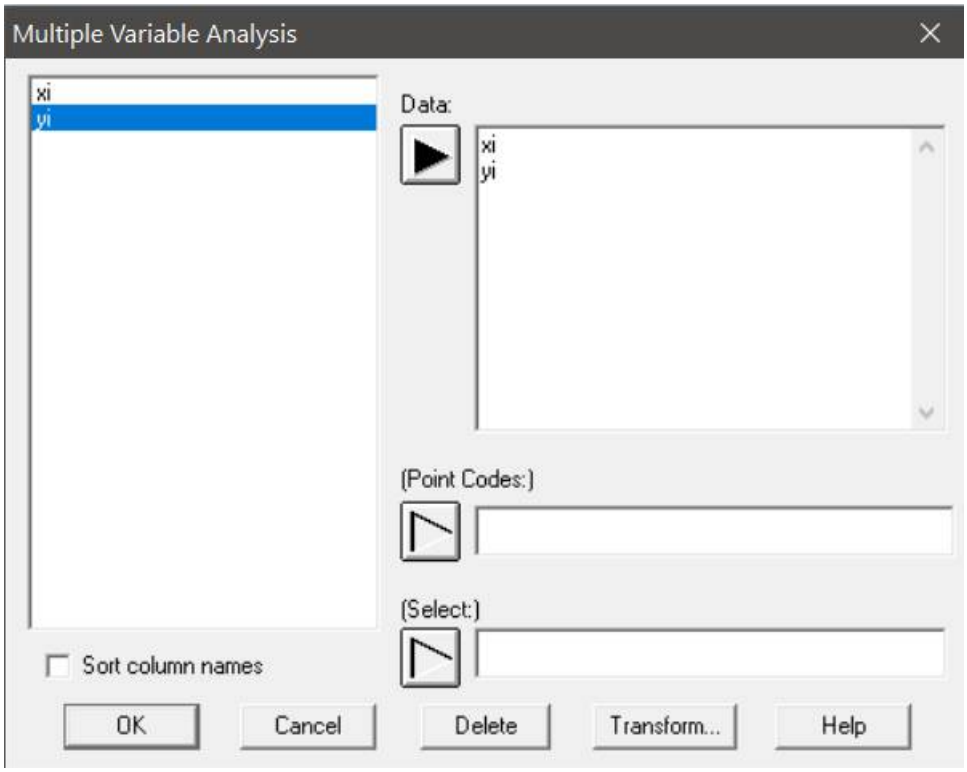
*Interpretace párově sdružených regresních koeficientů:*
Jsou směrnicemi sdružených regresních přímek a udávají, jak se v průměru změní hodnoty závisle proměnné při jednotkové změně nezávisle proměnné.
$b_{yx} = -0{,}853$ …. Říká, že pokud se zvýší hodnota nezávisle proměnné *x* o jednotku, sníží se hodnota závisle proměnné *y* v průměru o 0,853.
$b_{xy} = -1{,}090$ …. Říká, že pokud se zvýší hodnota nezávisle proměnné *y* o jednotku, sníží se hodnota závisle proměnné *x* v průměru o 1,090.

*Výpočet hodnoty koeficientu korelace a determinace:*
Hodnoty obou koeficientů můžeme nalézt v proceduře *Simple Regression* pod tabulkou *Analysis of Variance*:
Correlation Coefficient = -0,963769
R-squared = 92,8851 percent
Hodnota koeficientu korelace je -0,964 (zaokrouhleno) a značí velmi silnou nepřímou lineární závislost proměnných *x* a *y*.
Hodnota koeficientu determinace je 0,929 (zaokrouhleno a neuvedeno v %, nýbrž desetinných číslech). Znamená to, že 92,9 % z celkové variability závisle proměnné je možné vysvětlit příslušnou regresní přímkou.

Abychom mohli výstupy a závěry provedené na základě hodnoty korelačního koeficientu zobecnit, je potřeba provést test významnosti koeficientu korelace:

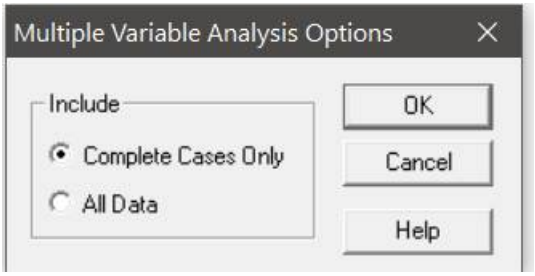
$H_0: \rho_{yx} = 0$
$H_1: \rho_{yx} \neq 0$

Nulová hypotéza říká, že koeficient korelace v základním souboru, který se označuje $\rho_{yx}$ nebo $\rho_{xy}$, je nulový, tzn. mezi proměnnými *x* a *y* není lineární závislost (tj. proměnné jsou nekorelované). Hypotéza alternativní se obvykle uvádí jako oboustranná, jak ji tady máme uvedenou i my, a říká, že proměnné *x* y *y* jsou lineárně závislé (korelované).
A nyní se věnujme posloupnosti procedur, které nám umožní získat výsledek testu. Zvolíme Describe – Multivariate Methods – Multiple-Variable Analysis (Correlation) … Objeví se vstupní panel pro zadání proměnných – viz Obr. 5.




*Obrázek 5 – Vstupní panel Multiple Variable Analysis*

Do řádku *Data* zadáme obě proměnné v libovolném pořadí (protože zkoumáme oboustrannou závislost) a potvrdíme OK. Objeví se dotaz, jaká data zahrnout do analýzy, jak vidíme na Obr. 6. My nemáme žádná chybějící data, proto je jedno, co zaškrtneme. Můžeme jen potvrdit klávesou OK.




*Obrázek 6 – Volba způsobu práce s daty*

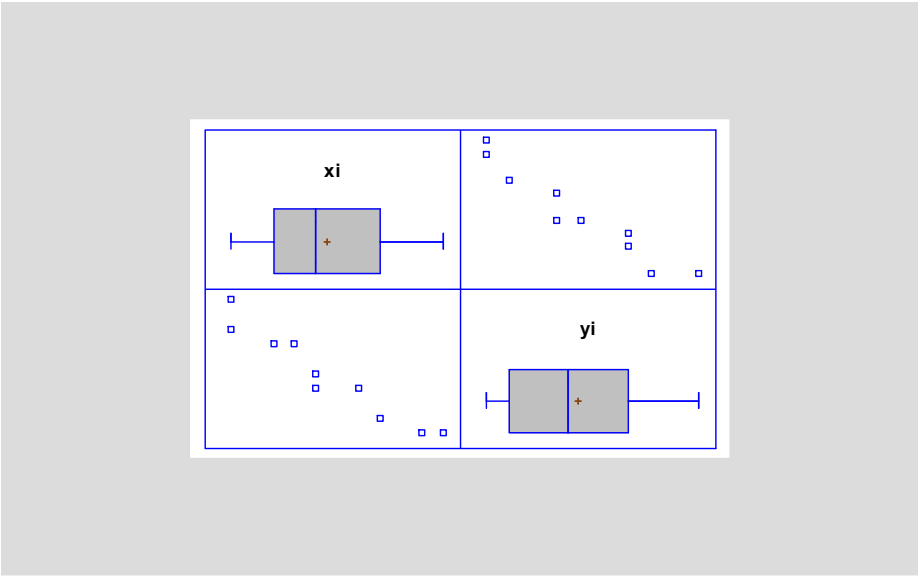
V nabídce *Tables and Graphs* nemusíme nic měnit. Zajímá nás výstup *Correlations*. Výsledek testu i hodnota koeficientu korelace je zaznamenána v korelační matici:

**Correlations**

| | xi | yi |
|---|---|---|
| xi | | -0,9638 |
| | | (10) |
| | | 0,0000 |
| yi | -0,9638 | |
| | (10) | |
| | 0,0000 | |

Correlation
(Sample Size)
P-Value

Jak vidíme v nápovědě pod korelační maticí, prvním údaje v matici je výběrový korelační koeficient, pod ním v závorce rozsah výběru a posledním údajem je P-Value, která se vztahuje k testu významnosti koeficientu korelace.
Vzhledem k tomu, že P-Value je 0,0000, což je méně, než $\alpha = 0{,}05$, zamítáme $H_0$ a přijímáme $H_1$. Na hladině významnosti 5 % jsme tedy prokázali, že proměnné *x* a *y* jsou lineárně závislé. Tato závislost je velmi silná, nepřímá (jak nám ukazuje hodnota koeficientu korelace). Nepřímá závislost vyjadřuje skutečnost, že s růstem hodnot jedné proměnné klesají hodnoty druhé proměnné.
Graficky lze vztah obou proměnných znázornit pomocí bodové matice (Scatterplot Matrix):




*Všimněte si!*

- Pokud jsme při odvození rovnic sdružených regresních přímek zjistili, že F-test je významný a t-testy také, bude významný i test významnosti koeficientu korelace, protože když je prokázáno, že přímka je vhodná k popisu dané závislosti, test významnosti koeficientu korelace tomu nemůže odporovat.
- Koeficient korelace je možné vypočítat také jako $r_{yx} = \pm\sqrt{b_{yx} \cdot b_{xy}}$, proto když mám k dispozici rovnice sdružených regresních přímek, je snadné jej vypočítat. Pozor na to, jaké znaménko pak dáme před odmocninu: Jsou-li oba regresní koeficienty kladné, bude před odmocninou kladná hodnota. Jsou-li oba regresní koeficienty záporné, bude před odmocninou záporná hodnota. Tj. zde: $r_{yx} = \pm\sqrt{b_{yx} \cdot b_{xy}} = -\sqrt{(-0{,}853) \cdot (-1{,}090)} = -\sqrt{0{,}92977} = -0{,}964$. Vyplývá to

z toho, že jak regresní koeficienty, tak i koeficient korelace nás informují o směru závislosti obou proměnných, a tyto dvě informace se musejí shodovat, tj. nelze mít záporný korelační koeficient a kladnou hodnotu některého z regresních koeficientů a naopak.

- Pokud by test významnosti koeficientu korelace vedl k NEZAMÍTNUTÍ $H_0$, neznamená to, že mezi proměnnými neexistuje žádný vztah, pouze to poukazuje na to, že nejsou lineárně závislé. Mohou být ale závislé třeba nelineárně!