Univerzita Pardubice

Fakulta ekonomicko-správní

Aplikace regrese a korelace v ekonomii

Zbyněk Černovský

Bakalářská práce
2013

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2012/2013

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Zbyněk Černovský**
Osobní číslo:	**E09441**
Studijní program:	**B6208 Ekonomika a management**
Studijní obor:	**Management podniku: Management malých a středních podniků**
Název tématu:	**Aplikace regrese a korelace v ekonomii**
Zadávající katedra:	**Ústav matematiky a kvantitativních metod**

Zásady pro vypracování:

Cílem práce bude popis metod regrese a korelace a rozbor použití v ekonomické praxi.
Součástí bude:
- rešerše odborné literatury
- rozbor možností použití metod regrese a korelace v ekonomické praxi
  řešení problémů s reálnými ekonomickými daty

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 30 stran**

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.
KUBANOVÁ, J. Matematická statistika. Statistické metody pro ekonomickou a technickou praxi. 3. vyd. Bratislava: Statis, 2008. ISBN 978-80-85659-47-4.
SYNEK, M., KISLINGEROVÁ, E. a kol. Podniková ekonomika 5. vyd. Praha: C.H. Beck, 2010. ISBN 978-80-7400-336-3.

Vedoucí bakalářské práce: **Ing. Kateřina Seinerová**
Ústav matematiky a kvantitativních metod

Datum zadání bakalářské práce: **30. září 2012**
Termín odevzdání bakalářské práce: **30. dubna 2013**

doc. Ing. Renáta Myšková, Ph.D.
děkanka

L.S.

doc. Ing. Marcela Kořenová, Ph.D.
vedoucí ústavu

V Pardubicích dne 3. října 2012

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 15. 8. 2013 Zbyněk Černovský

PODĚKOVÁNÍ:

Tímto bych rád poděkovala svému vedoucímu práce Ing. Kateřině Seinerové za její ochotu a odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování bakalářské práce. Dále bych rád poděkoval celé své rodině za projevenou podporu.

ANOTACE

Cílem bakalářské práce je vymezení základních pojmů regresní a korelační analýzy a využití těchto znalostí při praktickém využití. Teoretická část obsahuje informace z odborné literatury. V praktické části jsou příklady řešené metodami regresní a korelační analýzy a následuje vytvoření regresních funkcí

KLÍČOVÁ SLOVA

Regresní analýza, Korelační analýza, Metoda nejmenších čtverců, Index determinace, Index korelace

TITLE

Application of Regression and Correlation in Economics

ANNOTATION

The aim of this work is to explain of basic expressions related to regression and correlation analysis and application of this knowledge in practical solutions. The theoretical part includes information the literature. In practical part of the bachelor work there are examples solved by methods of regression and correlation analysis and then are regression functions created.

KEYWORDS

Regression analysis, Correlation analysis, Method of the least squares, Index determination, Index correlation.

Obsah

SEZNAM TABULEK

SEZNAM ILUSTRACÍ

SEZNAM ZKRATEK A ZNAČEK

X	Náhodná veličina
x	Hodnota statistického znaku
$\bar{x}$	Aritmetický průměr
$\hat{y}$	Odhad hodnoty statistického znaku
I^2	Index determinace
I	Index korelace
ε	Náhodná chyba
R	korelační koeficient

ÚVOD

S regresí a korelací se v běžném životě setkáváme v nejrůznějších oblastech téměř každý den. Setkáváme se s nimi všude, kde platí vztah příčiny a následku, protože vztah příčiny a následku je vlastně závislost mezi dvěma jevy, kterou se zabývá právě regresní a korelační analýza.

Tato práce by se dala rozdělit na část teoretickou a část praktickou.

V teoretické části jsou vysvětleny jednotlivé pojmy regresní a korelační analýzy i matematické vztahy související s tímto problémem. Všechny popsané vztahy a definice jsou zpracované pomocí odborné literatury uvedené v použitých zdrojích. Teoretická část obsahuje a vysvětluje základní pojmy regresní a korelační analýzy, metody pro měření kvality regresní funkce a intenzity závislosti. Dále nám teoretická část popisuje i používanou metodu nejmenších čtverců, která nám slouží k výpočtu základních parametrů regresní funkce.

V teoretické části je zahrnut i jeden z hlavních cílů této práce a to rozbor použití metod regrese a korelace v ekonomické praxi. Pro tento rozbor použití jsem si vybral jednu oblast z ekonomické praxe a to podnikání. V této oblasti se zaměřuji jak se dá využít korelační a regresní analýzy při vedení podniku. Zvláštní pozornost jsem věnoval převážně využití při modelování nákladových funkcí, protože řízení nákladů v podniku je podle mého názoru jednou z nejdůležitějších složek úspěšného podnikání.

V praktické části jsou tyto teoretické dovednosti využity k aplikaci na konkrétních příkladech. Nejprve jsou tyto metody použity na příkladu z ekonomické praxe, poté na příkladech s reálnými daty z konkrétního podniku. K výpočtům jsem použil tabulkový program MS Excel. Nejprve byl pomocí tohoto programu vytvořen regresní model. V příkladu z ekonomické praxe byla jen pro zajímavost provedena předběžná predikce porovnání vstupních závislých proměnných y a očekávaných yi vytvořených analýzou. Pomocí T – testu zjistíme p-hodnotu pro beta koeficient. Pomocí této hodnoty zjistíme, zda existuje lineární závislost mezi x a y a poté zjistíme její těsnost pomocí indexu determinace. Poté byla provedena korelační analýza pomocí testu významnosti pro koeficient korelace.

1 ZÁKLADNÍ POJMY REGRESNÍ ANALÝZY

1.1 Pevná a volná závislost

Závislosti je vhodné rozlišovat z hlediska zkoumání na tzv. pevné a volné závislosti. Pevnou závislostí označujeme případ, kde výskyt jednoho jevu nutně odpovídá výskytu druhého jevu. Z hlediska pravděpodobnosti jde o vztah, který se jistě projeví, s pravděpodobností rovnou jedné. Pevná závislost se charakterizuje tím, že se opakuje v jednotlivých případech, v nichž je pozorována, tzn., že ji charakterizuje i jediné pozorování. Uvažujeme - li, že výskyt jednoho jevu ovlivňuje výskyt druhého jevu, tím že při nastoupení prvního jevu se zvýšila pravděpodobnost nastoupení druhého jevu, hovoříme tedy o závislosti volné. Podmínkou u volné závislosti je tedy nastoupení prvního jevu. *[3]*

Při rozšíření na statistické znaky, které bude vhodnější nazývat proměnnými, znamená pevná závislost vztah, kdy každé hodnotě jedné proměnné odpovídá jedna a jen jedna hodnota jiných proměnných i naopak. Podle Hindlse volnou závislostí označujeme vztah, kdy hodnotám (např. jedné proměnné) odpovídá více různých hodnot jiné proměnné, ale kdy lze hovořit o jakési „obecné tendenci", která se projevuje při změnách hodnot těchto proměnných. Z toho vyplývá, že když označíme jednu proměnnou x a druhou proměnnou jako y. Poté můžeme očekávat, při volné závislosti mezi x a y, změnu hodnot y při změnách hodnot x a naopak při změnách hodnot y můžeme očekávat změnu hodnot x. *[3]*

1.2 Funkční závislost

Funkční závislost nám vyjadřuje závislost hodnot jedné proměnné na hodnotách druhé proměnné. Tuto závislost můžeme zapsat funkčním vztahem $y=f(x)$. Pokud tedy známe konkrétní hodnoty x, pak dokážeme přesně určit, jaké hodnoty nabude proměnná y. V praktických úlohách to není vždy tak lehké, neboť na sledovanou veličinu obvykle nepůsobí pouze jedna náhodná veličina X, ale je jich více. V takovémto případě nelze hovořit o tom, že mezi veličinami X a Y existuje funkční závislost, ale přesto se dá říci, že veličiny jsou závislé. Nemluvíme pak o funkční závislosti, ale o závislosti stochastické. *[3]*

1.3 Stochastická závislost

Stochastickou závislost nejlépe vyjadřuje definice stochasticky závislých veličin *[4]*: Nechť X, Y jsou dvě náhodné veličiny. Jestliže změna hodnoty jedné náhodné veličiny vyvolá změnu rozdělení pravděpodobností druhé veličiny, říkáme, že náhodné veličiny X, Y jsou stochasticky závislé.

Závislostí stochastickou nazýváme volnou závislost, která se týká kvantitativních znaků. S pevnými závislostmi se setkáme zejména v teoretické oblasti. Tímto způsobem byly sepsány různé teoretické zákony z ekonomické oblasti. Jedná se například o závislosti množství peněz v ekonomice na úrocích. V reálných situacích se naopak setkáváme většinou pouze s volnými závislostmi. K poznání a matematickému popisu stochastických závislostí slouží metody regresní analýzy.

Pro stochastickou závislost je charakteristické, že:

- změny závislé proměnné jsou vysvětlovány jen některými činiteli těchto změn,
- bereme v úvahu působení náhodných vlivů,
- při zjišťování údajů připouštíme možnost chyb.

Jednou ze zvláštností stochastické závislosti je, že se projevuje ve změně střední hodnoty jedné náhodné veličiny, která souvisí se změnou hodnot druhé náhodné veličiny. Z toho vyplývá, že se projevuje prostřednictvím středních hodnot.

Za příklad stochastické závislosti můžeme brát počet členů domácností a výdaje domácností na nákup potravin. Tedy můžeme tvrdit, že určitému počtu domácností odpovídá určité rozdělení výdajů na potraviny. Výdaje na potraviny jsou ovlivněny různými vlivy, které nedokážeme kontrolovat. Tyto vlivy nazýváme náhodné vlivy, které jsou např. oslava jubilea, návštěva, nemoc, nepřesné zaznamenání údajů o výdajích apod. *[4]*

2 REGRESNÍ FUNKCE

Definice *[4]*: Nechť X a Y jsou náhodné veličiny. Podmíněnou střední hodnotu E (Y|x), považovanou za funkci proměnné x, budeme nazývat regresní funkcí náhodné veličiny Y vzhledem k X. Regresní funkce vyjadřuje změny podmíněné střední hodnoty jedné náhodné veličiny při změně hodnot druhé náhodné veličiny. Graf regresní funkce nazýváme regresní křivka.

Regresní funkce nám umožňuje předpovědět, jaké hodnoty nabude jedna náhodná veličina, když známe hodnotu druhé veličiny. Protože Y je náhodná veličina, tak nám nemusí vždy při dané hodnotě x náhodné veličiny X nabýt hodnoty E (Y|x), ale bude nabývat hodnoty rozptýlené okolo ní. Za hlavní úkol regresní analýzy se považuje, zjištění tvaru stochastické závislosti a parametrů regresní funkce. V regresní analýze se budeme zabývat závislostí náhodné veličiny Y na veličině x (nezávisle proměnné), která nebude náhodná a může být obecně m-rozměrná. *[4]*

2.1 Jednoduchý model lineární regrese

Jednoduchým modelem lineární regrese se nazývá takový lineární model, kdy grafem regresní funkce je přímka. Abychom získali větší přehled, tak pro parametry β_0 a β_1 použijeme tradiční značení α a β. Předpokládejme, že $Y_1, Y_2, \ldots, Y_n$ je n-tice nekorelovaných náhodných veličin s vlastnostmi $EY_i = \alpha + \beta x$, $DY_i\ \sigma^2$, $i = 1,2,\ldots,n$, kde α, β, σ^2 jsou neznámé parametry a $x_1, x_2, \ldots, x_n$ je n-tice známých hodnot. Pak jednoduchým modelem lineární regrese budeme nazývat model $Y_i = \alpha + \beta x_i + \varepsilon_i$, kde složky ε_i jsou nezávislé náhodné veličiny, pro které platí $E\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$, $i = 1,2,\ldots, n$. Tyto složky obsahují působení náhodných vlivů, které nejsou v modelu zahrnuty. *[4]*

Regresní přímkou nazýváme přímku $y = \alpha + \beta x$, kde β je její směrnice, podle které se dá určit sklon regresní přímky, a také jestli je přímka rostoucí (směrnice je kladná) nebo klesající (směrnice je záporná). α nám určuje, v jaké vzdálenosti od počátku nám přímka vede, přičemž může být i záporná. Pro získání platného regresního modelu, musíme odhadnout neznámé parametry α, β, σ^2 daného modelu. Tyto odhady si označíme A, B, S^2. Bodové odhady parametrů α, β získáme metodou nejmenších čtverců.

2.2 Modely lineární vzhledem k parametrům

Těmito modely rozumíme modely, které mají regresní funkci tvaru $g(x, \beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=0}^{k} \beta_i g_i(x)$ kde g_i jsou funkce nezávisle proměnných x = (x_1, x_2, …, x_n). Jako ukázky lineárních modelů jsou uvedeny následující funkce *[3]:*

- regresní přímka $g(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x$ (2.0)
- regresní parabola $g(x, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x + \beta_2 x^2$ (2.1)
- regresní hyperbola $g(x, \beta_0, \beta_1) = \beta_0 + \beta_1 \frac{1}{x}$ (2.2)
- regresní logaritmická funkce $g(x, \beta_0, \beta_1) = \beta_0 + \beta_1 log x$ (2.3)
- regresní rovina $g(x_1, x_2, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (2.4)

2.3 Modely nelineární vzhledem k parametrům

Tyto modely se dají vhodnou transformací upravit na lineární tvar. Odhady parametrů lineárních modelů a modelů, které lze transformovat na lineární tvar, se provádějí nejčastěji metodou nejmenších čtverců, která je popsána v následující kapitole. Jako ukázky nelineárních modelů jsou uvedeny následující funkce *[4]:*

- regresní mocninná funkce $g(x, \beta_0, \beta_1) = \beta_0 * x^{\beta_1}$ (2.5)
- regresní exponenciální funkce $g(x, \beta_0, \beta_1) = \beta_0 * {\beta_1}^x$ (2.6)

2.4 Modely, které se nedají jednoduše transformovat na lineární tvar

U těchto modelů není vhodné použít pro odhady parametrů metodu nejmenších čtverců. Proto se odhady u těchto modelů provádějí jinými metodami např. metodou částečných součtů, metodou dílčích průměrů nebo metodou vybraných bodů. Jako ukázku můžeme použít například funkci *[4]*

$$g(x, \beta_0, \beta_1, \beta_2) = \beta_0 * {\beta_1}^x * \beta_2 \quad (2.7)$$

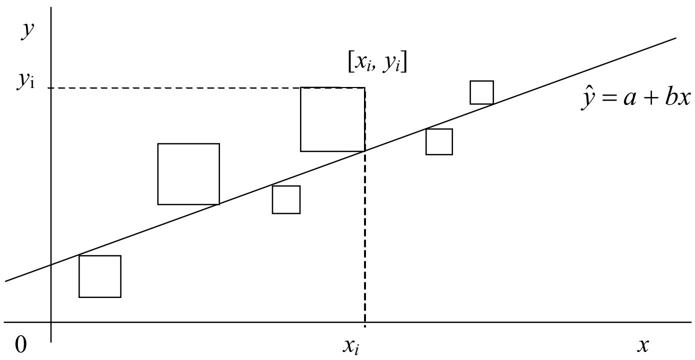
3 METODA NEJMENŠÍCH ČTVERCŮ

Tato metoda slouží k získání odhadu neznámých parametrů pro průběh lineární regrese, proto je důležité vysvětlit základní principy této metody.

Předpokládejme, že máme konkrétní dvojice naměřených hodnot (x_1,y_1), (x_2,y_2), …, (x_n,y_n). Hledáme takovou funkci (odhad) $\hat{y} = a+bx$, aby nám co nejvíce „přiléhala“ k bodům (x_1,y_1), (x_2,y_2), …, (x_n,y_n), kde „přiléhání“ měříme součtem rozdílů $\hat{y}i$ - y_i (tzv. reziduí). Z toho nám vyplývá, že se snažíme nalézt takové odhady, u kterých budou rozdíly mezi skutečnými hodnotami a hodnotami odhadnutými co nejmenší. Chceme – li se vyhnout tomu, že při značných odchylkách mezi $\hat{y}i$ a y_i se kladné a záporné rozdíly navzájem odečtou, vezmeme jako míru přiléhání ne prostý součet reziduí, ale součet jejich čtverců $\Sigma^{n}_{i=1}(y_i - \hat{y}i)^2$. Čím menší tento součet bude, tím lépe nám bude funkce $\hat{y}$ přiléhat k naměřeným bodům. Snažíme se proto nalézt takové odhady a, b, aby platilo: *[4]*

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = min \tag{3.0}$$

Na následujícím obrázku je metoda znázorněna graficky. Tento obrázek popisuje myšlenku metody, kde dvojice (x_i,y_i) je i-tá hodnota skutečně naměřených bodů a body regresní přímky $\hat{y} = a+bx$ jsou odhadem skutečných hodnot. Čím menší jsou čtverce, tím lepší je odhad a proložení funkce.



Obrázek 1: Metoda nejmenších čtverců

Zdroj:*[3]*

Nyní si postupně ukážeme teoretický postup jak zjistit bodové odhady a, b lineární regrese. V jednoduchém modelu lineární regrese budeme minimum funkce, ve které Y je náhodnou veličinou a A, B jsou odhady konstanty α a směrnice β dané přímky. Vztahy, které jsou zde použity, jsou převzaty ze zdroje *[4]*.

$$S(A,B) = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2 \tag{3.1}$$

Hledáme parciální derivace prvního řádu a vypočítáme extrém funkce dvou proměnných. Nutnou podmínkou pro zjištění extrému je nulovost parciálních derivací:

$$\frac{\partial S}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - Bx_i) = 0;\ \frac{\partial S}{\partial B} = -2\sum_{i=1}^{n}(Y_i - A - Bx_i)\cdot x_i = 0,$$

Po úpravě dostaneme soustavu normálních rovnic:

$$nA + B\,\Sigma_{i=1}^{n}\, x_i = \Sigma_{i=1}^{n}\, Y_i, \tag{3.2}$$

$$A\,\Sigma_{i=1}^{n}\, x_i + B\,\Sigma_{i=1}^{n}\, x_i^{\,2} = \Sigma_{i=1}^{n}\, x_i\, Y_i\,.$$

Toho pro odhady A, B plyne:

$$B = \frac{n\sum_{i=1}^{n}x_iY_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}Y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2} \quad ; \quad A = \frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iY_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}. \tag{3.3}$$

Odhad parametru α se často vyjadřuje ve tvaru, který vychází z první rovnice soustavy,

$$A = \frac{\sum_{i=1}^{n}Y_i - B\sum_{i=1}^{n}x_i}{n} = \overline{Y} - B\overline{x}$$

K tomu abychom si ověřili, že funkce S (3.1) nabývá ve stacionárním bodě minima, musíme zjistit parciální derivace druhého řádu:

$$\frac{\partial^2 S}{\partial A^2} = 2n\,;\ \frac{\partial^2 S}{\partial A \partial B} = 2\sum_{i=1}^{n}x_i\ ;\ \frac{\partial^2 S}{\partial B^2} = 2\sum_{i=1}^{n}x_i^2$$

Podle věty z diferenciálního počtu víme, že pokud výraz

$$\frac{\partial^2 S}{\partial A^2} * \frac{\partial^2 S}{\partial B^2} - \left(\frac{\partial^2 S}{\partial A \partial B}\right) = 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2$$

$$= 4n\left[\sum_{i=1}^{n} x_i^2 - 2\tfrac{1}{2}\left(\sum_{i=1}^{n} x_i^2\right)^2 + \tfrac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)^2\right] = 4n\sum_{i=1}^{n}(x_i - \bar{x})^2$$

nabývá kladné hodnoty, pak i parciální derivace $\frac{\partial^2 S}{\partial A^2} = 2n$ a $\frac{\partial^2 S}{\partial B^2} = 2\sum_{i=1}^{n} x_i^2$ nabývají kladných hodnot a funkce S má ve stacionárním bodě ostré minimum.

Regresní přímka, kterou jsme takto získali metodou nejmenších čtverců, má tedy tvar

$\hat{y} = A + Bx$. Uvedenou rovnici můžeme upravit na tvar $\hat{Y} = \overline{Y} + B(x - \overline{x})$

Dále je zapotřebí dokázat, že odhady A, B parametrů α a β jsou nevychýlené. Musí platit, že EB = β a EA = α.

Nejprve musíme upravit tvar odhadu parametru β (vzorec 3.3):

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} \overline{x}\, Y_i}{\sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + nx^{-2}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} * Y_i$$

Nyní ověříme EB = β

$$\mathrm{EB} = \mathrm{E}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} * Y_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} * \mathrm{E}Y_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} * (\alpha + \beta x_i)$$

$$= \alpha\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \beta\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

neboť $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$; $\sum_{i=1}^{n}(x_i - \overline{x})\, x_i = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

Dále ověříme EA = α

$$EA = E\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Y_i} - \mathrm{B}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}}{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} EY_i - \frac{1}{n}\sum_{i=1}^{n} x_i EB = \frac{1}{n}\left[\sum_{i=1}^{n}(\alpha + \beta x_i) - \sum_{i=1}^{n} \beta x_i\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \alpha = \alpha$$

Tím jsme dokázaly, že odhady A, B parametrů α, β jsou skutečně nevychýlené. Funkce $\hat{Y} = A + Bx$ je nevychýleným odhadem regresní přímky y = α + βx. *[4]*

4 KVALITA REGRESNÍ FUNKCE A INTENZITA ZÁVISLOSTI

Jedním z úkolů regresní a korelační analýzy je posouzení kvality regresní funkce a zjištění intenzity (síly, těsnosti) závislosti. Čím více jsou empirické hodnoty dané proměnné soustředěné kolem odhadnuté regresní funkce, tím silnější máme posuzovaný vztah, získáváme tím lepší regresní funkci. Naopak čím více jsou empirické hodnoty vzdáleny hodnotám vyrovnaným Y, tím je vztah slabší. Míra intenzity závislosti úzce souvisí s hodnocením odhadnuté regresní funkce a tedy s kvalitou regresního odhadu. Ke zjištění kvality regresní funkce se využívají zejména charakteristiky jako index determinace, index korelace, rozptyl reziduí.

4.1 Rozptyly empirických, vyrovnaných a skutečných hodnot

Můžeme zkonstruovat tři rozptyly se zcela odlišnou vypovídající schopností *[3]*:

- Rozptyl empirických hodnot $$s_y^2 = \frac{1}{n}\sum(y_i - \overline{y})^2 \quad (4.0)$$
- Rozptyl vyrovnaných hodnot $$s_y^2 = \frac{1}{n}\sum(Y_i - \overline{y})^2 \quad (4.1)$$
- Rozptyl skutečných hodnot (tzv. reziduální rozptyl)
$$s_{(y-Y)}^2 = \frac{1}{n}\sum(y_i - Y_i - \overline{y} - \overline{y})^2 = \frac{1}{n}\sum(y_i - Y_i)^2 \quad (4.2)$$

Při použití metody nejmenších čtverců mezi uvedenými rozptyly platí vztah:

$$S_y^2 = {S_Y}^2 + {S^2}_{(y-Y)} \quad (4.3)$$

Rozptyl empirických hodnot můžeme rozložit na rozptyl vyrovnaných hodnot a rozptyl reziduálních hodnot. Všechny empirické hodnoty by byly zároveň hodnotami vyrovnanými (ležely by na regresní čáře), pokud by existovala funkční závislost mezi závisle proměnnou y a vysvětlující proměnnou x. Potom by se rozptyl empirických hodnot rovnal rozptylu vyrovnaných hodnot a reziduální rozptyl by byl nulový. Platilo by $S_y^2 = {S_y}^2$. Pokud by mezi oběma proměnnýma existovala úplná nezávislost, pak by všechny vyrovnané hodnoty byly stejné a jejich rozptyl by byl nulový. Optimální regresní funkce je ta, která má menší rozptyl *[3]*.

4.2 Index determinace

Z uvedeného nám vyplývá, že závislost proměnné y a proměnné x pravděpodobně bude tím silnější, čím větší bude podíl rozptylu vyrovnaných hodnot na celkovém rozptylu. A naopak čím bude podíl tohoto rozptylu menší, tím slabší bude intenzita závislosti. Sílu závislosti můžeme tedy měřit poměrem rozptylu vyrovnaných a empirických hodnot.

$$I_{yx}^2 = \frac{s_y^2}{s_y^2} \tag{4.3}$$

Tento poměr nazýváme index determinace *[3]*. U funkční závislosti získáme hodnotu 1, hodnotu nula získáme v případě nezávislosti. Čím více se blíží jedné, tím je závislost silnější a dobře vystihuje regresní funkci. Naopak čím více se blíží nule, tím je závislost slabší a regresní funkce je méně výstižná. Pokud hodnotíme intenzitu závislosti pomocí indexu determinace, musíme brát v úvahu, že jeho velikost ovlivňuje skutečnost, zda se nám podařilo nalézt vhodný typ regresní funkce pro popis dané závislosti. Z toho nám vyplývá, že pokud vyjde nízká hodnota indexu determinace, tak to nemusí znamenat nízký stupeň závislosti, ale může nám poukázat na špatnou volbu regresní funkce.

4.3 Index Korelace

V praxi se k měření těsnosti závislosti obyčejně nepoužívá pouze index determinace. Používá se i jeho odmocnina, která se nazývá index korelace.

$$I_{yx} = \sqrt{\frac{s_y^2}{s_y^2}} \tag{4.4}$$

Index korelace nám prakticky udává stejné údaje o těsnosti závislosti jako index determinace, pouze s tím rozdílem, že tyto informace mají menší vypovídací hodnotu *[3]*.

5 KORELAČNÍ ANALÝZA

Korelační analýza slouží k posouzení kvality regresní funkce a zjištění těsnosti (velikosti, síly) oboustranné závislosti. Předpokladem je, že všechny proměnné, jejichž závislost zkoumáme, jsou náhodnými veličinami a jejich sdružené rozdělení je vícerozměrné normální rozdělení tj.:

- U dvou proměnných uvažujeme, že pozorované dvojice pochází z dvourozměrného rozdělení.
- U m-proměnných uvažujeme, že pozorované m-tice pocházejí z m-rozměrného normálního rozdělení.

Pro naše účely si vystačíme pouze s dvourozměrným rozdělením

5.1 Teoretické základy – jednorozměrný náhodný výběr

Předpokládejme, že máme dvě náhodné veličiny X a Y s konečnými nenulovými rozptyly DX, DY. Pokud jsou tyto náhodné veličiny závislé, je zapotřebí tuto závislost kvantitativně vyjádřit. Abychom vyjádřili míru závislosti, u lineárního typu závislosti použijeme korelační koeficient. Korelační koeficient dvou náhodných veličin je definován následujícím způsobem*[4]*:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{DX} * \sqrt{DY}} \tag{5.0}$$

Pokud nám vyjde, že $\rho_{x,y}=0$, tvrdíme, že náhodné veličiny X a Y jsou nekorelované. V opačném případě říkáme, že mezi náhodnými veličinami existuje korelační vztah. Dále je dobré se zmínit, že když jsou dvě náhodné veličiny nekorelovaná, tak to neznamená, že musí být nezávislé.

Uvažujeme, že $X=(X_1, X_2, \ldots, X_m)$ a $Y = (Y_1, Y_2, \ldots, Y_m)$ jsou m-rozměrné náhodné vektory s vektory středních hodnot $EX = (EX_1, EX_2, \ldots, EX_m)$ a $EY = (EY_1, EY_2, \ldots, EY_m)$ a s konečnými nenulovými rozptyly jednotlivých složek. Vztah mezi těmito veličinami nám pak vyjadřuje kovariance náhodných veličin X a Y, a je dána výrazem $cov(X,Y) = E[(X - EX)(Y - EY)] = EXY - EXEY$. Budeme se zabývat kovariancemi takových náhodných veličin, kdy jedna je složkou náhodného vektoru X a druhá Y. Kovariance takovýchto dvojic můžeme zapsat do kovariační matice náhodných vektorů X a Y *[4]*.

$$cov(X,Y) = \begin{pmatrix} cov(X_1,Y_1) & cov(X_1,Y_2) & \cdots & cov(X_1,Y_n) \\ cov(X_2,Y_1) & cov(X_2,Y_2) & & cov(X_2,Y_n) \\ \cdots & & \cdots & \cdots \\ cov(X_m,Y_1) & cov(X_m,Y_2) & \cdots & cov(X_m,Y_n) \end{pmatrix}$$

Nyní si můžeme uvést i korelační matici vektorů X a Y

$$\rho = \begin{pmatrix} \rho_{X_1,Y_1} & \rho_{X_1,Y_2} & \cdots & \rho_{X_1,Y_n} \\ \rho_{X_2,Y_1} & \rho_{X_2,Y_2} & & \rho_{X_2,Y_n} \\ \cdots & & \cdots & \cdots \\ \rho_{X_m,Y_1} & \rho_{X_m,Y_2} & \cdots & \rho_{X_m,Y_n} \end{pmatrix}$$

5.2 Dvourozměrný náhodný výběr

5.2.1 Výběrová kovariance

Nechť $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ je náhodný výběr dvojrozměrného základního souboru (X, Y). Výběrový průměr definujeme jako dvojici $(\overline{X}, \overline{Y})$, kde $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ a $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Podobně můžeme vyjádřit i výběrový rozptyl (S_X^2, S_Y^2), kde $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ a $S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$. Výběrovou kovariancí nazýváme statistiku:

$$\overline{COV}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{5.1}$$

Kde $\overline{X}$ resp. $\overline{Y}$ je aritmetický průměr z náhodných veličin X_i resp. Y_i *[3].*

5.2.2 Výběrový koeficient korelace a korelační matice

Definice *[4]*: Nechť je dán dvojrozměrný náhodný výběr $[(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)]$. Výběrový koeficient korelace $R_{x,y}$ náhodných veličin X a Y definujeme jako podíl:

$$R_{X,Y} = \frac{\overline{COV}(X,Y)}{S_X * S_Y} \tag{5.2}$$

Na rozdíl od korelačního koeficientu náhodných veličin X a Y, výběrový koeficient korelace obsahuje v čitateli výběrovou kovarianci a ve jmenovateli součin výběrových směrodatných odchylek.

Pro praktické výpočty je vhodnější tvar tohoto vztahu *[3]:*

$$R_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} = \frac{\sum_{i=1}^{n}X_iY_i-n\overline{X}\overline{Y}}{\sqrt{\sum_{i=1}^{n}X_i^2-n\overline{X}^2}\sqrt{\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2}} \tag{5.3}$$

Vlastnosti výběrového koeficientu korelace *[4]*

- $R_{X,Y} \in \langle -1; 1\rangle$
- $R_{X,Y} = R_{Y,X}$
- $R_{aX+bcX+d} = R_{X,Y}$ pro ac > 0
- $R_{aX+bcX+d} = -\,R_{X,Y}$ pro ac < 0

Výběrový koeficient korelace je odhadem korelačního koeficientu dvou náhodných veličin.

Výběrovou korelační maticí vektoru X = (X_1, X_2, …, X_n) vyjadřujeme vzájemné vztahy náhodných veličin z náhodného výběru m-rozměrného rozdělení pravděpodobností.

Definice *[3]*: Nechť X = (X_1, X_2, …, X_n) je náhodný výběr z m – rozměrného základního souboru X = (X_1, X_2, …, X_n). Výběrovou korelační maticí budeme nazývat matici

$$R = \begin{pmatrix} 1 & R_{X_1,X_2} & \dots & R_{X_1,X_n} \\ R_{X_2,X_1} & 1 & \dots & R_{X_2,X_n} \\ \dots & & \dots & \dots \\ R_{X_m,X_1} & R_{m,X_2} & \cdots & 1 \end{pmatrix}$$

5.2.3 Test významnosti pro koeficient korelace

Často se setkáváme s tím, že potřebujeme zjistit, jestli jsou náhodné veličiny X a Y korelované nebo ne. Zjistíme to tím, budeme-li testovat hypotézu H_0: ρ = 0 proti alternativní hypotéze H_1: ρ ≠ 0.

Nechť (X_1, Y_1), (X_2, Y_2), ….,(X_n, Y_n) je dvojrozměrný náhodný výběr ze základního souboru (X, Y). Základní soubor má dvojrozměrné normální rozdělení pravděpodobností N (μ_1, μ_2, σ_1, σ_2). R je výběrový koeficient korelace.

Testujeme hypotézu, že koeficient korelace základního souboru je roven nule, tedy *[4]*

H_0: ρ = 0 proti H_1: ρ ≠ 0

Testovací kritérium má tvar: $$T = \frac{R}{\sqrt{1-R^2}} * \sqrt{n-2} \tag{5.4}$$

Podle této hypotézy má náhodná veličina T Studentovo rozdělení pravděpodobnosti s n-2 stupni volnosti. Kritická oblast je definována vztahem:

$$W = \{T; |T| > t_{\alpha,n-2}\}, \text{ kde } t_{\alpha,n-2} = F_{t,n-2}^{-1}\left(\frac{2-\alpha}{2}\right) \tag{5.5}$$

Hlavním předpokladem tohoto testu je, že náhodný výběr pochází ze základního souboru s normálním rozdělením pravděpodobností. Vychází se z kritických hodnot Studentova rozdělení pravděpodobnosti. Pro kritickou hodnotu r_α za předpokladu hypotézy H_0: $\rho = 0$, platí: $P(|R| \geq r_\alpha) = \alpha$. Upraveno podle *[4]*.

6 REGRESNÍ A KORELAČNÍ ANALÝZA V PODNIKÁNÍ

Metody regresní a korelační analýzy se dají využívat nejen pro řešení statistických úloh, ale dají se použít i pro usnadnění výpočtů všude kde je potřeba zjistit závislost mezi ukazateli. Nejinak tomu je i v podnikání. Nemusíme se bavit pouze o zjišťování závislosti mezi ukazateli, regresní analýza se dá také využít k odhadu budoucích hodnot ukazatelů, nebo pro zjištění jakým trendem se nám bude daný ukazatel vyvíjet. Zjišťováním trendu se myslí sestavování a analýza časových řad. Právě v analýze časových řad využíváme regresní analýzy k sestavení časové řady a k následnému odhadu budoucích hodnot.

Jednou z nejdůležitějších složek, které se vyskytují v podnikání takovém nebo v řízení podniku jsou náklady, které daný podnik vydá na získání výnosů. K tomu aby se náklady daly účinně analyzovat a řídit, je potřeba vzniklé náklady roztřídit dle různých hledisek. Náklady se třídí podle:

- Druhů – spotřeba surovin a materiálu, spotřeba energie, mzdové a osobní náklady, odpisy dlouhodobého hmotného i nehmotného majetku, finanční náklady, a jiné,
- Účelu vynaložení – náklady technologické a náklady na obsluhu a řízení, náklady jednicové a režijní, náklady přímé a nepřímé,
- Závislosti nákladů na změnách výroby – variabilní a fixní náklady.

Z hlediska řízení nákladů patří členění nákladů na variabilní a fixní mezi nejdůležitější. Vztah mezi celkovými náklady a objemem výroby se dá vyjádřit matematickými funkcemi, které nazýváme nákladové funkce. Nejjednodušší je lineární funkce tvaru: $N = F + n*Q$, kde N jsou celkové náklady v Kč, Q je objem výroby v kusech, n jsou variabilní náklady na 1 jednotku a F jsou fixní náklady. *[7]*

K zjišťování fixních a variabilních nákladů a pro stanovení nákladových funkcí se v praxi používá řada metod (klasifikační analýza, metoda dvou období) a jednou z nich je i využití regresní a korelační analýzy. Metoda regresní a korelační analýzy je nejspolehlivější metodou pro stanovení nákladových funkcí. Umožňuje stanovit i nelineární nákladové funkce, které jsou vhodné pro případný nadproporcionální nebo podproporcionální vývoj nákladů a to v případech, kdy průběh nákladů nemůžeme spolehlivě vyjádřit lineární funkcí. Tato metoda nám umožňuje stanovit i spolehlivost zjištěných funkcí pomocí měr korelace a provádět předběžné odhady chyb zjišťovaných hodnot pomocí tzv. **mezí spolehlivosti**. Dále nám umožňuje sestavit i grafický obrázek (tzv. bodový diagram) vývoje nákladů. Jak již jste si mohli povšimnout lineární nákladová funkce, má stejný tvar jako lineární regresní model, proto se dá pro ruční výpočet použít metoda nejmenších čtverců, o které jsem se již zmínil v kapitole 3.*[6]*

Regresní a korelační analýzu lze v podniku využít také při sestavování rozpočtů, plánování výnosů (zejména analýzou časových řad faktorů ovlivňující tržby např. HDP, příjmy domácností, změny produkce dominantních odběratelů apod.), plánování rozvahy (pomocí regresního modelu, časových řad minulých tržeb a časových řad vybrané položky rozvahy, která vystupuje jako závisle proměnná). Dále je regresní analýza jednou z metod pro výpočet normy počtu pracovníků k zajištění funkce organizační jednotky. V tomto případě regresní analýza vychází ze sledování závislosti řady faktorů charakterizujících technicko-organizační podmínky práce (resp. Pracnost dané činnosti a počtu pracovníků zajišťujících tuto činnost).*[2]*

7 PŘÍKLAD Z EKONOMICKÉ PRAXE

Když jsme si vysvětlili problematiku regresní a korelační analýzy, rád bych přešel k aplikaci této teorie na praktické příklady z ekonomické praxe. Příklady jsou zpracované v tabulkovém programu MS Excel s využitím doplňků pro analýzu dat, které obsahují metody korelace a regrese.

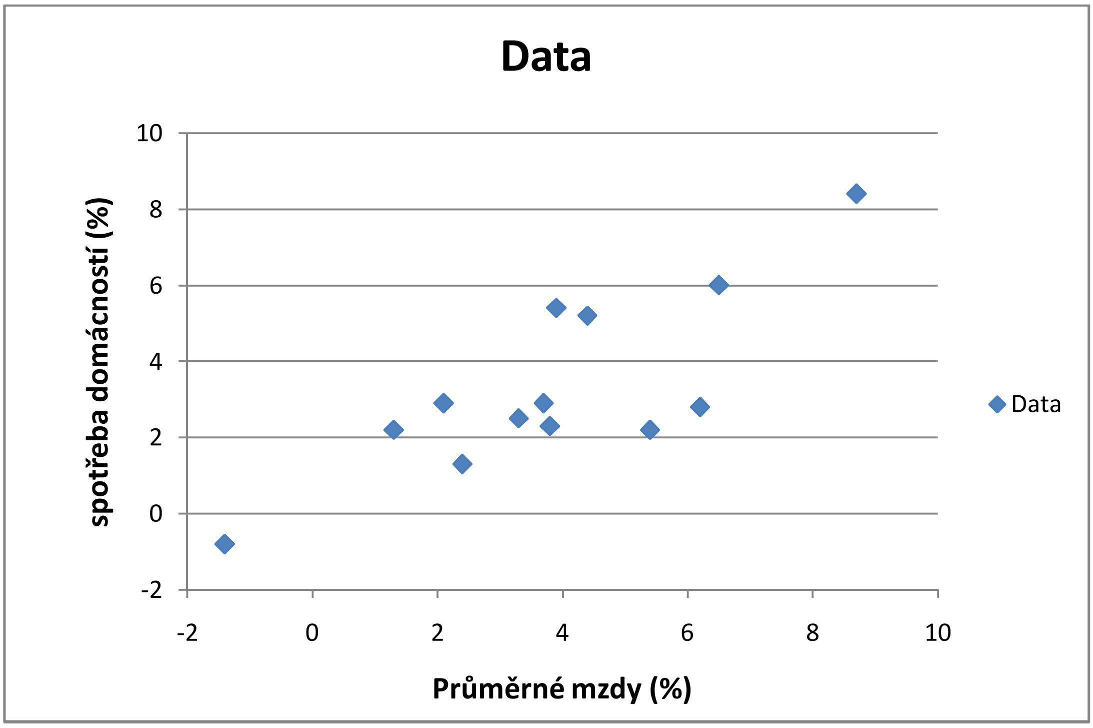
7.1 Zadání

Jako první byl vybrán příklad, kdy jsem zjišťoval závislost míry spotřeby domácností na měsíční hrubé reálné mzdě. Mým cílem je vytvořit regresní přímku a zjistit pomocí regresní a korelační analýzy, zda existuje závislost mezi danými ukazateli. Dále pak zjistit parametry α a β regresní přímky a zjistit těsnost dané závislosti. Data jsou pořízena z ekonomického serveru Měšec.cz [8], [9]. Změny hodnot u obou skupin dat jsou vyjádřeny v procentech a vyznačují změnu ukazatele v daném období. Jako nezávislá data x použijeme hrubé mzdy a na nich závislými daty y použijeme spotřebu domácností. Vstupní data jsou zobrazena v následující tabulce (Tabulka 1) a graficky znázorněna v grafu (Obrázek 2).

Tabulka 1: Vstupní data

rok	průměrné mzdy (%) x	spotřeba domácností (%) y
1996	8,7	8,4
1997	1,3	2,2
1998	-1,4	-0,8
1999	6,2	2,8
2000	2,4	1,3
2001	3,8	2,3
2002	5,4	2,2
2003	6,5	6
2004	3,7	2,9
2005	3,3	2,5
2006	3,9	5,4
2007	4,4	5,2
2008	2,1	2,9

Zdroj: Vlastní zpracování



Obrázek 2: Grafické znázornění vstupních dat

Zdroj: Vlastní zpracování

7.2 Regresní a korelační analýza

7.2.1 Regresní analýza

K posouzení závislosti budeme využívat hlavně ukazatelů regresní analýzy. Tyto ukazatele nalezneme v regresní statistice, kterou nám vytvořil tabulkový program MS Excel. Z výsledků, které nám zobrazil MS Excel, si vybereme tzv. p – hodnotu, u které platí, že pokud hladina významnosti je větší než p – hodnota, tak existuje mezi danými veličinami závislost. Z regresní statistiky dále použijeme hodnotu spolehlivosti R (index determinace), násobné R (korelační koeficient). U indexu korelace a indexu determinace víme, že čím je hodnota větší, tím lépe. Dále pak ověříme výsledek ještě dalším výpočtem a to výpočtem testu významnosti pro koeficient korelace.

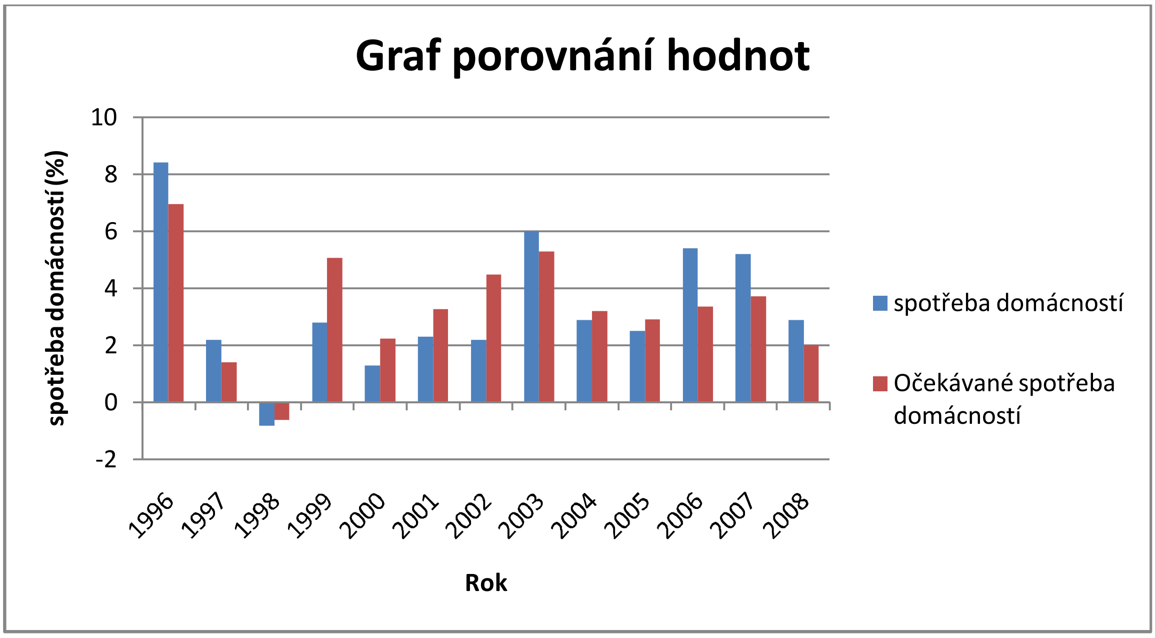
Tabulka níže (Tabulka 2) obsahuje jednotlivé nezávislé hodnoty x a závislé hodnoty y, tyto data jsou vstupními daty. Hodnoty v poli očekávaná spotřeba domácností byly vytvořeny regresním modelem. Tyto hodnoty nám představují body, kterými nám bude procházet regresní přímka. Jinými slovy to je již zmiňovaný odhad regresního modelu. Již z těchto dat dokážeme alespoň částečně odhadnout, zda regresní přímka bude dobře prokládat data.

Tabulka 2: Vstupní data a očekávané hodnoty regresní přímky

průměrné mzdy (%) x	spotřeba domácností (%) y	očekávaná spotřeba domácností (%) Yi
8,7	8,4	6,95
1,3	2,2	1,41
-1,4	-0,8	-0,61
6,2	2,8	5,08
2,4	1,3	2,23
3,8	2,3	3,28
5,4	2,2	4,48
6,5	6	5,30
3,7	2,9	3,20
3,3	2,5	2,90
3,9	5,4	3,35
4,4	5,2	3,73
2,1	2,9	2,01

Zdroj: Vlastní zpracování

Jen pro přehled ještě dodávám grafické znázornění porovnání hodnot skutečné a očekávané spotřeby domácností (Obrázek 3). Porovnáním skutečných hodnot spotřeby domácností s očekávanými hodnotami spotřeby domácností zjistíme, že skutečná spotřeba domácností, byla u většiny případů vyšší než očekávaná spotřeba. Z tohoto grafu lze také vyčíst, že v období mezi lety 1999 – 2002, byla skutečná spotřeba o dost menší, než by se dalo očekávat.

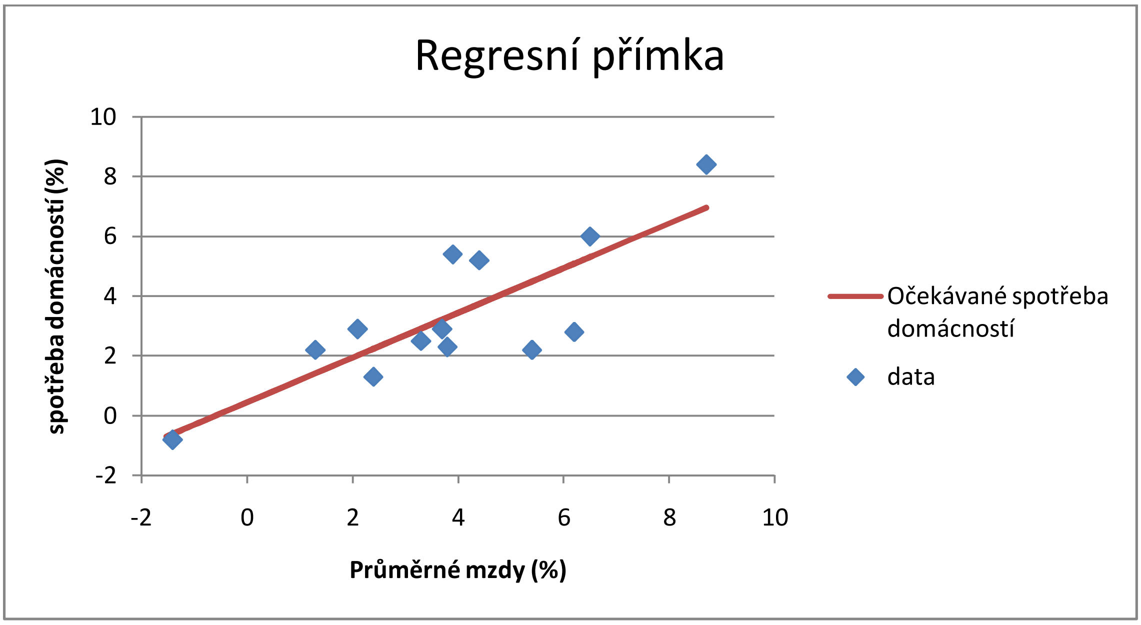


Obrázek 3: Grafické porovnání hodnot

Zdroj: Vlastní zpracování

Z tabulkového programu MS Excelu jsme obdrželi následující výsledky, které nám pomohou s vyhodnocením našeho úkolu:

Graf regresní funkce (Obrázek 4):



Obrázek 4: Regresní přímka

Zdroj: vlastní zpracování

rovnici regresní přímky: $\hat{y} = 0{,}434 + 0{,}748x$

z které nám vyplývá, že parametr α = 0,434 a parametr β = 0,748

T – testem pro beta koeficient byla zjištěna p – hodnota:

P – hodnota = $0{,}77 * 10^{-3}$

Nulovou hypotézu H_0: β = 0 zamítáme na všech hladinách významnosti, protože α > p – hodnota. Vzhledem k tomu, že výsledná p – hodnota je velmi malá (p- hodnota = $0{,}77 * 10^{-3}$), můžeme tvrdit, že existuje významná lineární závislost mezi průměrnou hrubou mzdou a mírou spotřeby domácností.

Regresní statistika	
Násobné R	**0,810888599**
Hodnota spolehlivosti R	**0,65754032**
Nastavená hodnota spolehlivosti R	0,626407622
Chyba stř. hodnoty	1,441539371
Pozorování	13

Obrázek 5: Výsledky regresní statistiky

Zdroj: vlastní zpracování

Z výsledků regresní statistiky (Obrázek 5) vidíme, že hodnota spolehlivosti R (index determinace) je 0,65754032, tedy 65,75% z celkové variability se dá vysvětlit jednoduchým lineárním modelem.

7.2.2 Korelační analýza

Dalším výsledkem, který budeme brát v úvahu je násobné R (koeficient korelace), hodnota je 0,810888599. Díky této hodnotě můžeme vytvořit korelační analýzu, která by nám měla potvrdit, že mezi veličinami X a Y existuje závislost. Velikostí tohoto koeficientu zjistíme sílu (těsnost) závislosti, čím vice se koeficient blíží 1 nebo -1 tím je těsnost závislosti větší. K určení závislosti použijeme test významnosti pro koeficient korelace:

$$T = \frac{R}{\sqrt{1-R^2}} * \sqrt{n-2} = \frac{0{,}810888599}{\sqrt{1-(0{,}810888599)^2}} * \sqrt{13-2} = 4{,}595712094$$

Hranice kritické oblasti má hodnotu: $t_{0,05;11}$= 2,2010

|4,5957| > 2,2010, z toho vyplývá, že testovací kritérium má realizaci v kritické oblasti, tedy zamítáme H_0: ρ=0. Korelace je tím pádem nenulová, tedy existuje lineární závislost mezi průměrnou hrubou mzdou a mírou spotřeby domácností.

Hodnota koeficientu korelace je 0,810888599, z toho vyplývá, že se zde nachází výrazná těsnost závislosti.

8 POUŽITÍ KORELACE A REGRESE NA PODNIKOVÝCH DATECH KONKRÉTNÍHO PODNIKU

V této části si uvedeme použití regresní a korelační analýzy na příkladech z konkrétního podniku, z kterého jsem získal potřebná data k řešení různých příkladů. Tyto data jsou získána z výroční zprávy podniku za rok 2010 a za rok 2012.

8.1 Informace o firmě

Nejprve si uvedeme nějaké základní informace o firmě, z které jsem získal potřebná data.

Název společnosti: Dalkia Česká republika, a.s.

Sídlo společnosti: 28. Října 3337/7, Moravská Ostrava 702 00, Ostrava, Česká republika

Právní forma: akciová společnost

Datum založení: 24. Dubna 1992

Základní kapitál: 3 146 446 440 Kč

Firma Dalkia Česká republika, a.s. je členem skupiny Dalkia v České republice, která je považována za největšího nezávislého výrobce tepla a elektrické energie na českém trhu. Tato firma je členem stejnojmenné francouzské nadnárodní skupiny, která v Evropě zaujímá přední pozici v oblasti energetiky. Jejím majoritním vlastníkem je skupina folia Environnement, světová jednička v oblasti environmentálních služeb.

Skupina Dalkia působí na českém trhu již více než dvacet let a zásobuje teplem více než 260 tisíc domácností. Hlavními zákazníky Dalkie jsou města a jejich obyvatelé, průmyslové podniky, zdravotnická zařízení, školy, veřejné organizace, obchodní a administrativní centra, atd. Je také prvním provozovatelem sítě chladu v ČR a kromě toho zajišťuje i výrobu a dodávky stlačeného vzduchu a průmyslových utilit. *[11]*

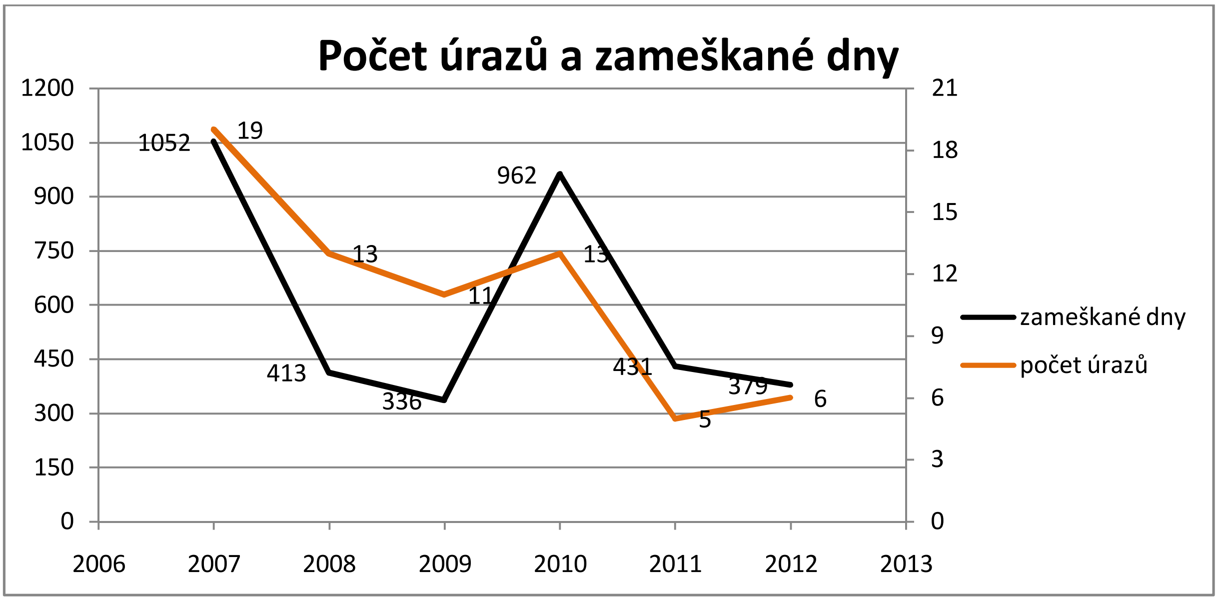
8.2 Příklad 1

V tomto příkladu budeme zjišťovat pomocí korelační analýzy vztah mezi počtem úrazů a zameškaných dnů v období mezi roky 2007 až 2012.

Tabulka 3: Vstupní data

rok	počet úrazů	zameškané dny
2007	19	1052
2008	13	413
2009	11	336
2010	13	962
2011	5	431
2012	6	379

Zdroj: Vlastní zpracování



Obrázek 6: Počet úrazů a zameškané dny

Zdroj:[11]

Pomocí tabulkového programu MS Excel zjistíme hodnotu korelačního koeficientu, s jehož pomocí zjistíme vztah mezi počtem úrazů a zameškanými dny. Hodnota korelačního koeficientu je 0,73461465. K určení vztahu mezi těmito veličinami použijeme test významnosti pro koeficient korelace:

$$T = \frac{R}{\sqrt{1-R^2}} * \sqrt{n-2} = \frac{0{,}73461465}{\sqrt{1-(0{,}73461465)^2}} * \sqrt{6-2} = 2{,}165457233$$

Hranice kritické oblasti má hodnotu: $t_{0,05,4} = 2{,}7764$

|2.1655| < 2.7764, z toho vyplývá, že testovací kritérium nemá realizaci v kritické oblasti, tedy H_0: $\rho=0$ nezamítáme. Z toho vyplývá, že korelace může být nulová a tím pádem neexistuje žádná významná závislost mezi počtem úrazů a zameškanými dny.

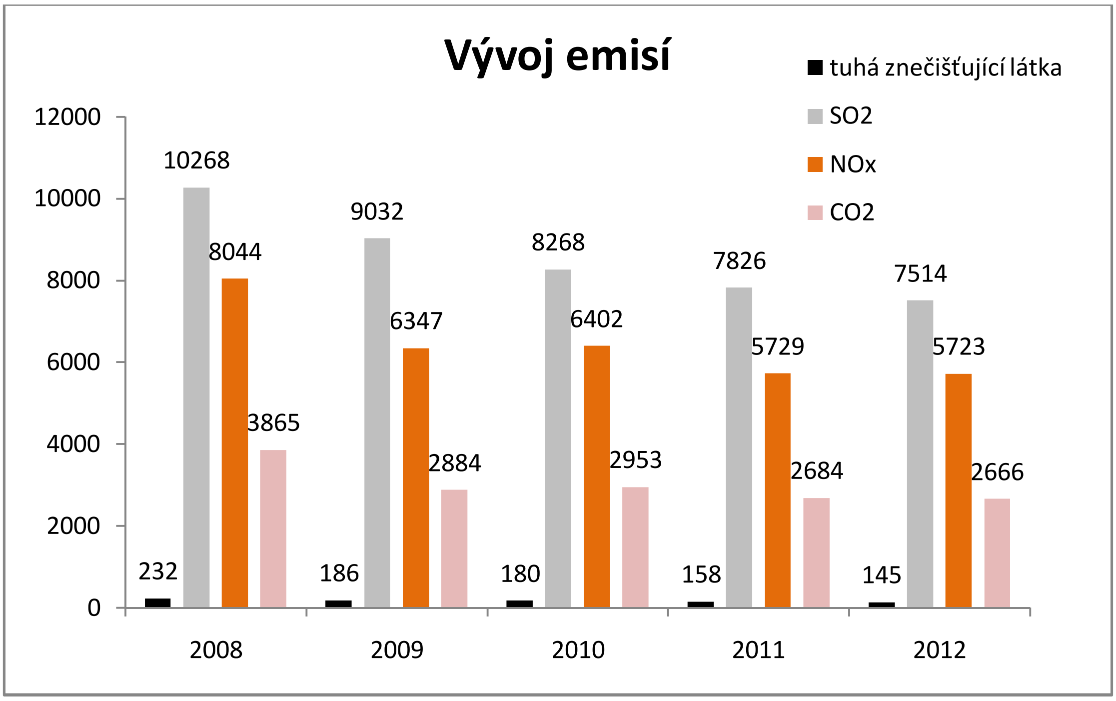
8.3 Příklad 2

Ve druhém příkladu provedeme lineární regresi množství emisí v čase, a to z hlediska 4 různých složek (tuhá znečišťující látka, SO_2, NO_x, CO_2). Regresi provedeme za období 2008 až 2012.

Tabulka 4: Vstupní data

rok	tuhá znečišťující látka	SO_2	NO_x	CO_2
2008	232	10268	8044	3865
2009	186	9032	6347	2884
2010	180	8268	6402	2953
2011	158	7826	5729	2684
2012	145	7514	5723	2666

Zdroj: Vlastní zpracování



Obrázek 7: Vývoj emisí

Zdroj:[11]

Pomocí tabulkového programu MS Excel si pro každou složku vytvoříme lineární regresi a získáme rovnice regresních přímek.

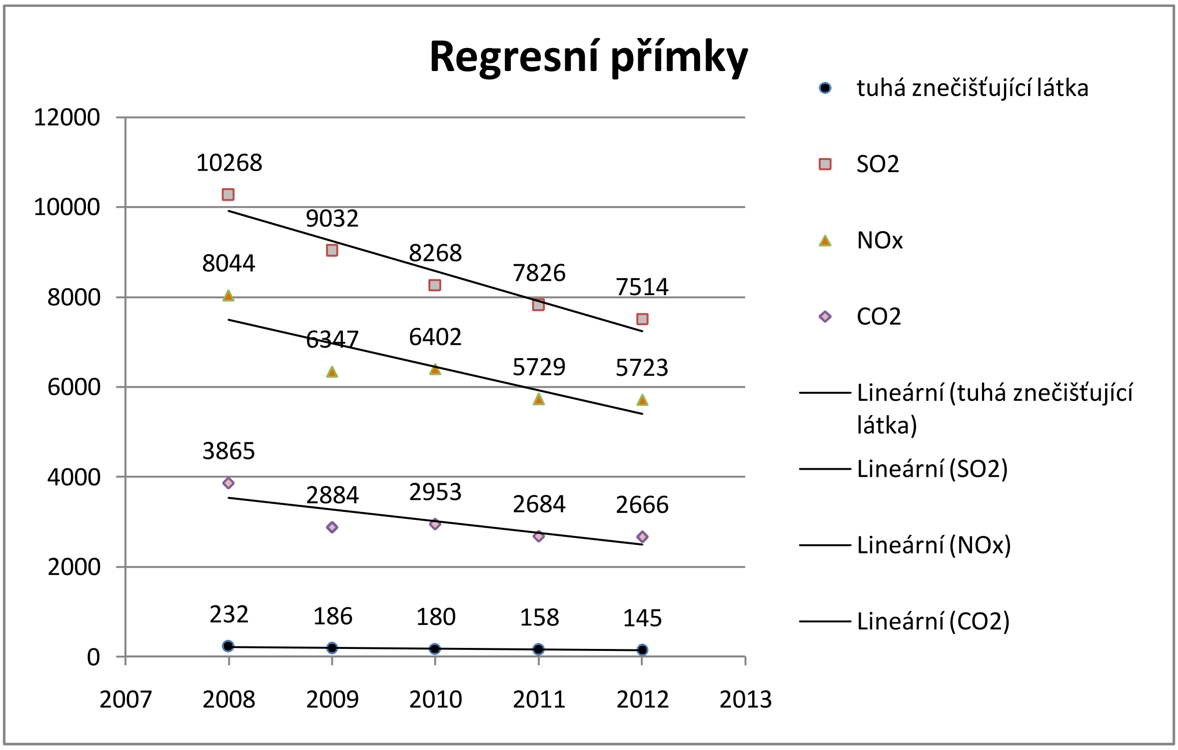
Tuhá znečišťující látka: $\hat{y} = 40782{,}2 - 20{,}2x$

SO_2: $\hat{y} = 1358096 - 671{,}4x$

NO_x: $\hat{y} = 1063709 - 526x$

CO_2: $\hat{y} = 525208{,}4 - 259{,}8x$

Z těchto rovnic vyplývá, že množství tuhých znečišťujících látek klesalo v průměru o 20,2 tun každý rok od roku 2008. Množství SO_2 v průměru klesalo o 671,4 tun každý rok od roku 2008. Množství NO_x v průměru klesalo o 526 tun každý rok od roku 2008. A množství CO2 v průměru klesalo o 259,8 tisíc tun každý rok od roku 2008.



Obrázek 9: Regresní přímky jednotlivých složek vývoje emisí

Zdroj: Vlastní zpracování

T - testy pro beta koeficienty byly zjištěny p – hodnoty pro každou složku:

Tuhá znečišťující látka: p – hodnota = 0,010377

SO_2: p – hodnota = 0,008327

NO_x: p – hodnota = 0,051179

CO_2: p – hodnota = 0,080455

Porovnáním P – hodnoty s hladinou významnosti $\alpha = 0{,}05$ zjistíme, zda změny vývoje emisí mají dlouhodobý trend růstu či poklesu. U tuhých znečišťujících látek je výsledná p – hodnota menší než hladina významnosti $\alpha >$ p – hodnota. Tedy H_0: $\beta=0$ zamítáme na hladině významnosti $\alpha = 0{,}05$. Z toho vyplývá, že existuje lineární závislost a můžeme tedy konstatovat, že vývoj tuhých znečišťujících látek má dlouhodobý trend poklesu. U SO_2 je výsledná p – hodnota také menší než hladina významnosti, tedy výsledek je stejný jako u tuhých znečišťujících látek. U NO_x je výsledná p – hodnota je o něco málo větší než hladina významnosti $\alpha=0{,}05$. Z toho vyplývá, že H_0: $\beta=0$ nezamítáme a můžeme tvrdit, že neexistuje významná lineární závislost a můžeme konstatovat, že vývoj NO_x nemá žádný dlouhodobý trend růstu či poklesu. U CO_2 je výsledná hodnota vyšší než hladina významnosti, tedy výsledek je stejný jako u NO_x.

Z výsledků regresních statistik pro každou složku vývoje emisí, získáme index determinace. Pomocí tohoto indexu zjistíme, kolik procent z celkové variability je možné vysvětlit jednoduchým lineárním modelem.

Tuhá znečišťující látka: index determinace = 0,91719115, tedy 91,72% z celkové variability se dá vyjádřit jednoduchým lineárním modelem.

SO_2: index determinace = 0,92831914, tedy 92,83 % z celkové variability.

NO_x: index determinace = 0,76809340, tedy 76,81 % z celkové variability.

CO_2: index determinace = 0,69243667, tedy 69,24 % z celkové variability.

U tuhých znečišťujících látek a u SO_2 nám vyšly indexy determinace přes 90 %. Takto vysoké hodnoty značí, že nám regresní přímka velice dobře prokládá data.

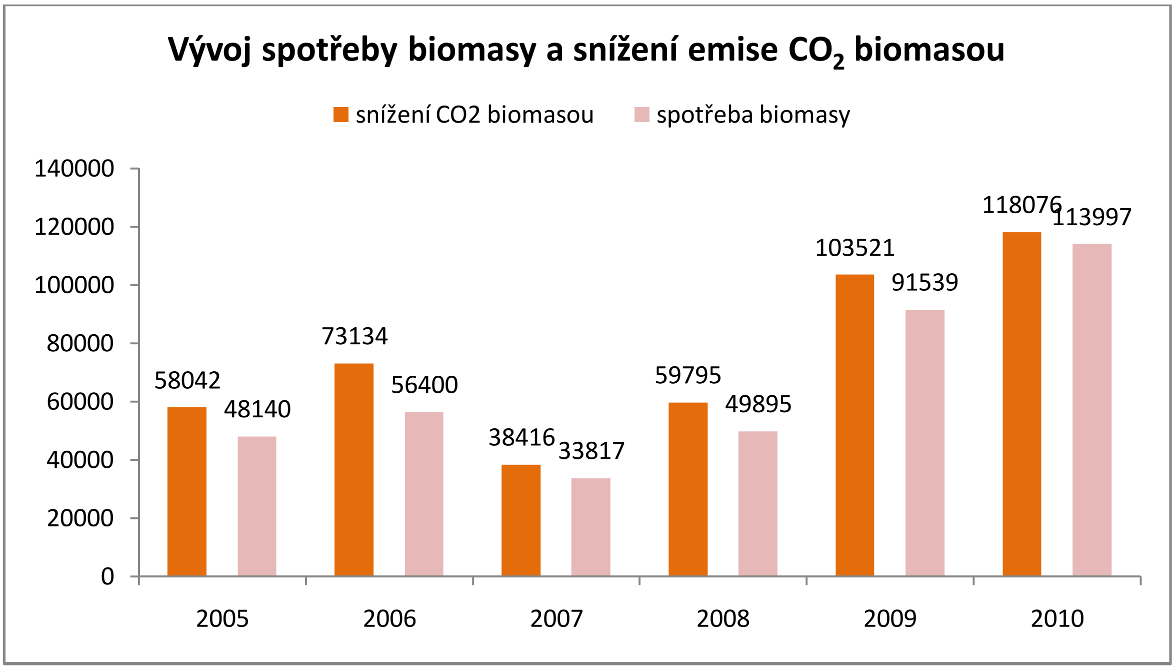
8.4 Příklad 3

V tomto příkladu využijeme korelační a regresní analýzy, pro zjištění závislosti mezi spotřebou biomasy a snížení emise CO_2 biomasou. Závislost budeme zjišťovat v rozmezí let 2005 až 2010.

Tabulka 5: Vstupní data

rok	spotřeba biomasy	snížení CO_2 biomasou
2005	48140	58042
2006	56400	73134
2007	33817	38416
2008	49895	59795
2009	91539	103521
2010	113997	118076

Zdroj Vlastní zpracování



Obrázek 10: Vývoj spotřeby biomasy a snížení emise CO_2 biomasou

Zdroj:[10]

8.4.1 Regresní analýza

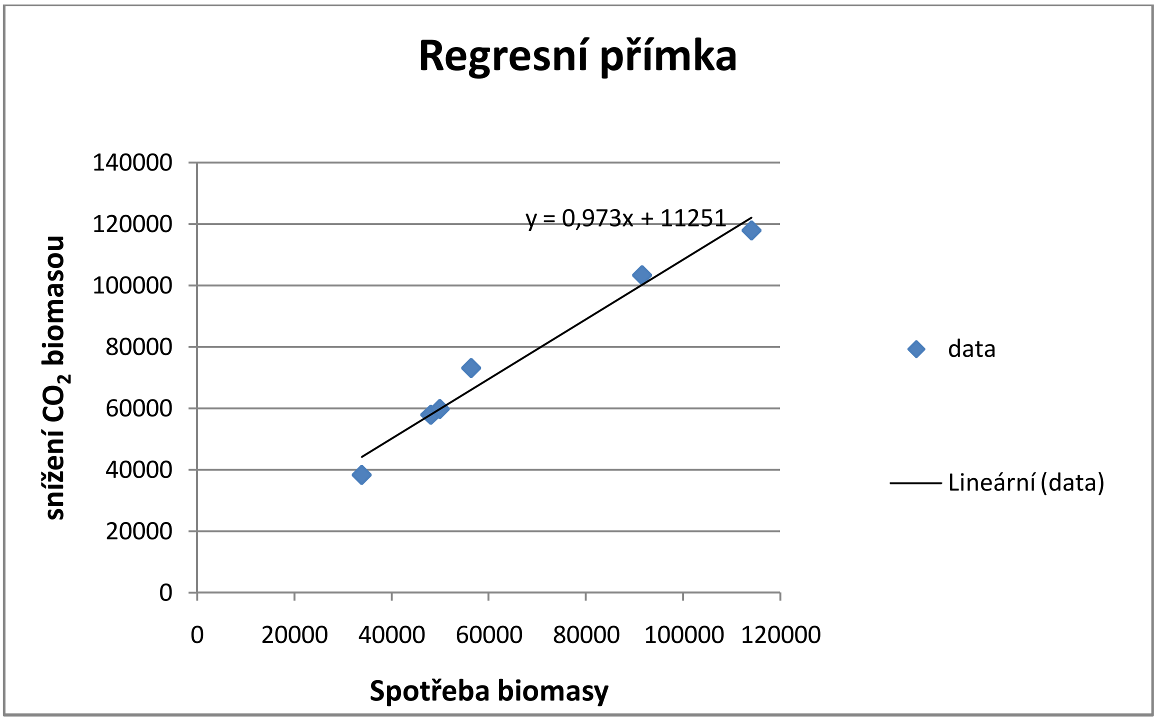
Pomocí tabulkového programu MS Excel vytvoříme regresní analýzu. Z této analýzy získáme výsledky, které nám pomohou s řešením příkladu. Získaly jsme hodnoty parametrů regresní přímky:

α =11251

β = 0,973

Znalostí těchto parametrů, nyní můžeme sestavit rovnici regresní přímky:

$$\hat{y} = 11251 + 0{,}973x$$



Obrázek 11: Regresní přímka

Zdroj: *Vlastní zpracování*

T – testem pro koeficient beta jsme získali p – hodnota:

P – hodnota: $0{,}219 * 10^{-3}$

Nulovou hypotézu H_0: $\beta = 0$ zamítáme na všech hladinách významnosti, protože $\alpha > p$ – hodnota. Vzhledem k tomu, že výsledná p – hodnota je velmi malá (p- hodnota = $0{,}219 * 10^{-3}$), můžeme tvrdit, že existuje významná lineární závislost mezi spotřebou biomasy a snížením CO^2 biomasou.

Regresní statistika	
Násobné R	**0,987893**
Hodnota spolehlivosti R	**0,975933**
Nastavená hodnota spolehlivosti R	0,969917
Chyba stř. hodnoty	5220,389
Pozorování	6

Obrázek 12: Regresní statistika

Zdroj: Vlastní zpracování

Z výsledků regresní statistiky vidíme, že hodnota spolehlivosti R (index determinace) je 0,975933, tedy 97,59 % z celkové variability se dá vyjádřit jednoduchým lineárním modelem.

8.4.2 Korelační analýza

Dalším výsledkem regresní statistiky, který budeme brát v úvahu je násobné R (koeficient korelace), hodnota je 0,987893. K určení závislosti použijeme test významnosti pro koeficient korelace:

$$T = \frac{R}{\sqrt{1-R^2}} * \sqrt{n-2} = \frac{0{,}987893}{\sqrt{1-(0{,}987893)^2}} * \sqrt{6-2} = 12{,}73598$$

Hranice kritické oblasti má hodnotu: $t_{0,05;4} = 2{,}7764$

|12,73598| > 2,7764, z toho vyplývá, že testovací kritérium má realizaci v kritické oblasti, tedy zamítáme H_0: $\rho=0$. Korelace je tím pádem nenulová, tedy existuje lineární závislost mezi spotřebou biomasy a snížením CO_2 biomasou.

Hodnota koeficientu korelace je 0,987893, z toho vyplývá, že se zde nachází velmi výrazná těsnost závislosti.

8.5 Příklad 4

V tomto příkladu použijeme metod korelace a regrese pro zjištění vztahu mezi výnosy a zisky za účetní období dceřiných společností za rok 2012.

Tabulka 6: Vstupní data

Dceřinná společnost	výnosy	zisk
Dalkia Česká republika, a.s.	8 883 102	1 990 721
OLTERM & TD Olomouc, a.s.	402 768	20 922
AmpluServis, a.s.	356 543	3 023
Dalkia Kolín, a.s.	418 886	88 991
Dalkia Mariánské Lázně, s.r.o.	156 216	13 102
Dalkia Industry CZ, a.s.	1 856 444	435 105
Dalkia Commodities CZ, s.r.o.	2 507 990	79 477
Dalkia Powerline Sp. z.o.o.	566 814	16 869

Zdroj: Vlastní zpracování

8.5.1 Regresní analýza

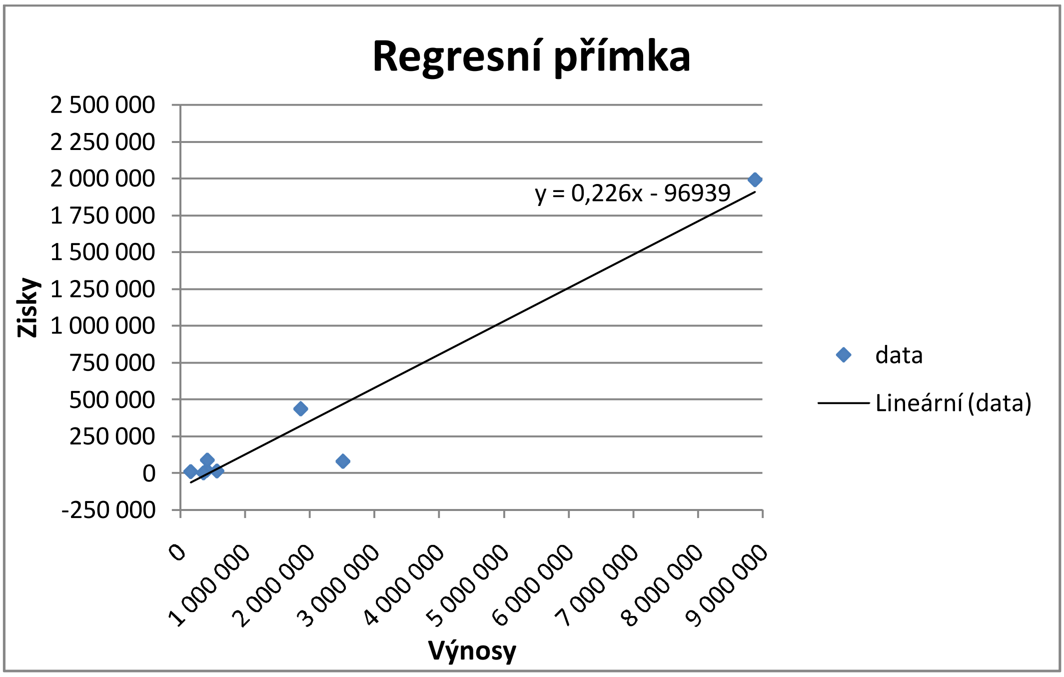
Pomocí tabulkového programu MS Excel vytvoříme regresní analýzu. Z této analýzy získáme výsledky, které nám pomohou s řešením příkladu. Získaly jsme hodnoty parametrů regresní přímky:

α =-96939

β = 0,226

Znalostí těchto parametrů, nyní můžeme sestavit rovnici regresní přímky:

$$\hat{y} = -96939 + 0{,}226x$$



Obrázek 13: Regresní přímka

Zdroj: Vlastní zpracování

T – testem pro koeficient beta jsme získali p – hodnota:

P – hodnota: $0{,}583 * 10^{-4}$

Nulovou hypotézu H_0: $\beta = 0$ zamítáme na všech hladinách významnosti, protože $\alpha > p$ – hodnota. Vzhledem k tomu, že výsledná p – hodnota je velmi malá (p- hodnota = $0{,}583 * 10^{-4}$), můžeme tvrdit, že existuje významná lineární závislost mezi výnosy a zisky dceřiných společností.

Regresní statistika	
Násobné R	**0,971221261**
Hodnota spolehlivosti R	**0,943270737**
Nastavená hodnota spolehlivosti R	0,93381586
Chyba stř. hodnoty	176391,8813
Pozorování	8

Obrázek 14: Regresní statistika

Zdroj: Vlastní zpracování

Z výsledků regresní statistiky vidíme, že hodnota spolehlivosti R (index determinace) je 0,943270737, tedy 94,32 % z celkové variability se dá vyjádřit jednoduchým lineárním modelem

8.5.2 Korelační analýza

Dalším výsledkem regresní statistiky, který budeme brát v úvahu je násobné R (koeficient korelace), hodnota je 0,971221261. K určení závislosti použijeme test významnosti pro koeficient korelace:

$$T = \frac{R}{\sqrt{1-R^2}} * \sqrt{n-2} = \frac{0{,}971221261}{\sqrt{1-(0{,}971221261)^2}} * \sqrt{6-2} = 9{,}988269$$

Hranice kritické oblasti má hodnotu: $t_{0,05;6} = 2{,}4469$

$|9{,}988269| > 2{,}4469$, z toho vyplývá, že testovací kritérium má realizaci v kritické oblasti, tedy zamítáme H_0: $\rho=0$. Korelace je tím pádem nenulová, tedy existuje lineární závislost mezi výnosy a zisky dceřiných společností.

Hodnota koeficientu korelace je 0,971221261, z toho vyplývá, že se zde nachází velmi výrazná těsnost závislosti.

ZÁVĚR

Cílem této práce bylo teoreticky popsat metody a pojmy regresní a korelační analýzy a následné řešení problémů s reálnými ekonomickými daty. V teoretické části jsem kromě základních pojmů a metod regresní analýzy charakterizoval také jednotlivé ukazatele kvality regresní funkce a intenzity závislosti, kterých se následně využívalo při aplikaci na daných příkladech. Teoretická část obsahuje také kapitolu o využití regresní a korelační analýzy v ekonomické praxi. Tato kapitola nám nastínila možné použití metod korelace a regrese v řízení podniku. V praktických ukázkách metod regresní a korelační analýzy byla vytvořena analýza v tabulkovém programu MS Excel. Z výsledků této analýzy jsem se zaměřil hlavně na parametry α a β, které nám pomohly sestavit regresní funkci. Dále jsem se zaměřil na p – hodnotu, index determinace a korelační koeficient. Tyto ukazatele mi pomohli ke zjištění závislosti a také k intenzitě závislosti. Následně jsem pomocí očekávaných hodnot yi, které jsou také jedním z výsledků analýzy, vytvořil příslušnou regresní funkci prokládající vstupní data.

U příkladu v sedmé kapitole byla analyzována data se závislostí mezi mírou spotřeby domácností a průměrnou hrubou mzdou. Regresní analýzou nám vyšla regresní přímka ve tvaru $\hat{y} = 0{,}434 + 0{,}748x$. T – testem pro beta koeficient byla zjištěna p - hodnota = 0,77 * 10^{-3}, bylo zjištěno, že existuje významná závislost mezi ukazateli míry spotřeby domácností a průměrnou hrubou mzdou. Pomocí indexu determinace bylo zjištěno, že regresní funkce se dá vyjádřit z 65,75 % jednoduchým lineárním modelem.

V osmé kapitole jsem se zaměřil na využití korelační a regresní analýzy na příkladech s reálnými podnikovými daty z konkrétního podniku. Prvním příkladem byla korelace počtu úrazů a zameškaných dnů. Korelační analýzou byl zjištěn koeficient korelace R = 0,734614656. Testem významnosti pro koeficient korelace bylo zjištěno, že neexistuje žádná výrazná závislost mezi počtem úrazů a zameškaných dnů. Druhým příkladem byla lineární regrese množství emisí v čase. U tohoto příkladu se počítalo se 4 různými složkami emise. Regresní analýzou nám vyšli regresní přímky ve tvarech:

Tuhá znečišťující látka: $\hat{y} = 40782{,}2 - 20{,}2x$

SO_2: $\hat{y} = 1358096 - 671{,}4x$

NO_x: $\hat{y} = 1063709 - 526x$

CO_2: $\hat{y} = 525208{,}4 - 259{,}8x$

Následně T – testem pro beta koeficient pro každou složku emise byly zjištěny p – hodnoty. Porovnáním těchto hodnot s hladinou významnosti $\alpha = 0{,}05$ bylo zjištěno, že u emise tuhých znečišťujících látek a u SO_2 byla p – hodnota menší než hladina významnosti. Z toho vyplývá, že existuje významná závislost a můžeme konstatovat, že tyto složky mají dlouhodobý trend poklesu. Naopak u NO_x a CO_2 byla p – hodnota větší než hladina významnosti. Z toho vyplývá, že neexistuje významná závislost a můžeme konstatovat, že tyto složky nemají žádný dlouhodobý trend růstu či poklesu. U třetího příkladu se zjišťovala závislost pomocí korelační a regresní analýzy mezi spotřebou biomasy a snížením CO_2 biomasou. Regresní analýzou nám vyšla regresní přímka ve tvaru $\hat{y} = 11251 + 0{,}973x$. T – testem pro beta koeficient byla zjištěna p - hodnota = $0{,}213 * 10^{-3}$, bylo zjištěno, že existuje významná závislost mezi spotřebou biomasy a snížením CO_2 biomasou. Pomocí indexu determinace bylo zjištěno, že regresní funkce se dá vyjádřit z 97,59 % jednoduchým lineárním modelem. Korelační analýzou byla zjištěna hodnota koeficientu korelace. Hodnota koeficientu korelace je 0,987893, z toho vyplývá, že se zde nachází velmi výrazná těsnost závislosti. U posledního příkladu se zjišťoval vztah mezi výnosy a zisky dceřiných společností za rok 2012. Regresní analýzou nám vyšla regresní přímka ve tvaru $\hat{y} = -96939 + 0{,}226x$. T – testem pro beta koeficient byla zjištěna p - hodnota = $0{,}213 * 10^{-3}$, bylo zjištěno, že existuje významná závislost mezi Výnosy a zisky dceřiných společností. Pomocí indexu determinace bylo zjištěno, že regresní funkce se dá vyjádřit z 94,32 % jednoduchým lineárním modelem. Korelační analýzou byla zjištěna hodnota koeficientu korelace. Hodnota koeficientu korelace je 0,971221261, z toho vyplývá, že se zde nachází velmi výrazná těsnost závislosti.

POUŽITÁ LITERATURA

[1] BUDÍKOVÁ, Marie, Maria KRÁLOVÁ a Bohumil MAROŠ. Průvodce základními statistickými metodami. 1. vyd. Praha: Grada, 2010, 272 s. ISBN 978-80-247-3243-5.

[2] BUCHTA, Miroslav. Organizace provozu podniku: distanční opora na CD. 1. vyd. Pardubice, 201, 1 CD-ROM. ISBN 978-80-7395-417-8.

[3] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6.

[4] KUBANOVÁ, J. Matematická statistika. Statistické metody pro ekonomickou a technickou praxi. 3. vyd. Bratislava: Statis, 2008. ISBN 978-80-85659-47-4.

[5] MAREK, L. *Statistika pro ekonomy – aplikace.* vydání 1. Praha: Professional Publishing, 2005. 423 s. ISBN 80-86419-68-1.

[6] SYNEK, Miloslav. Manažerská ekonomika. 4.aktualiz. a rouš. vyd. Praha: Grada, 2007, 452 s. ISBN 978-80-247-1992-4.

[7] SYNEK, M., KISLINGEROVÁ, E a kol. Podniková ekonomika 5. vyd. Praha: C.H. Beck, 2010. ISBN 978-80-7400-336-3.

[8] Měšec.cz: Vývoj spotřeby v ČR [online]. 2013 [cit. 2013-08-13]. Dostupné z: https://www.mesec.cz/dane/ekonomika/pruvodce/spotreba/

[9] Měšec.cz: Vývoj reálných mezd v ČR [online]. 2013 [cit. 2013-08-13]. Dostupné z: https://www.mesec.cz/dane/ekonomika/pruvodce/mzdy/

[10] Výroční zpráva a.s. Dalkia Česká republika. Ostrava (CZ): Dalkia Česká republika, 2010.

[11] Výroční zpráva a.s. Dalkia Česká republika. Ostrava (CZ): Dalkia Česká republika, 2012.