INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Slezská univerzita v Opavě
Obchodně podnikatelská fakulta v Karviné

STATISTICKÉ ZPRACOVÁNÍ DAT

Pro kombinovanou formu studia

Jaroslav Ramík, Radmila Stoklasová

Karviná 2013

Projekt OP VK č. CZ.1.07/2.2.00/28.0017
„Inovace studijních programů na Slezské univerzitě,
Obchodně podnikatelské fakultě v Karviné“

Název:	**Statistické zpracování dat**
Autor:	**prof. RNDr. Jaroslav Ramík, CSc., Mgr. Radmila Stoklasová, Ph.D.**
Vydavatel:	Slezská univerzita v Opavě Obchodně podnikatelská fakulta v Karviné
Určeno:	studentům Obchodně podnikatelské fakulty v Karviné
Počet stran:	162
AA – VA:	7,07 – 7,35
Náklad:	100
Tiskárna:	Z + M Partner, spol. s r.o. Ostrava
Číslo publikace:	595-200-2013
ISBN:	**978-80-7248-842-1**

Tato publikace neprošla jazykovou úpravou.

evropský sociální fond v ČR · EVROPSKÁ UNIE · MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY · OP Vzdělávání pro konkurenceschopnost · Slezská univerzita v Opavě

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Slezská univerzita v Opavě
Obchodně podnikatelská fakulta v Karviné

STATISTICKÉ ZPRACOVÁNÍ DAT

Pro kombinovanou formu studia

Jaroslav Ramík a Radmila Stoklasová

Karviná 2013

Projekt OP VK č. CZ.1.07/2.2.00/28.0017
„Inovace studijních programů na Slezské univerzitě,
Obchodně podnikatelské fakultě v Karviné“

Obor: Statistika.

Anotace: Publikace představuje studijní oporu vysokoškolského kurzu Statistické metody pro ekonomy pro navazující studium na vysoké škole ekonomického zaměření. Obsahově pokrývá základní témata: analýza rozptylu – 1 faktor, analýza rozptylu – 2 faktory, jednorozměrná a vícerozměrná analýzy rozptylu a časové řady.

Klíčová slova: Analýza rozptylu, jednoduchá regresní analýza, vícerozměrná regresní analýza, analýza časových řad, Box – Jenkinsova metodologie.

Autor: **prof. RNDr. Jaroslav Ramík, CSc.**
Mgr. Radmila Stoklasová, Ph.D.

Recenzenti: doc. Ing. Jana Hančlová, CSc.
Ing. Filip Tošenovský, Ph.D.

ISBN **978-80-7248-842-1**

OBSAH

ÚVOD

Tento text představuje studijní oporu pro studium všech akreditovaných studijních programů v navazujícím magisterském studiu na Slezské univerzitě, Obchodně podnikatelské fakultě v Karviné. Předmět *Statistické zpracování dat* navazuje na předmět Statistika (dříve Kvantitativní metody B) obsahující základní bakalářský kurz statistiky na SU OPF, nebo na obdobný ekvivalentní předmět základů statistiky v bakalářském stupni studia na jiné VŠ ekonomického zaměření v ČR. Tento text je inovací předchozí studijní opory s názvem *Statistika* pro navazující magisterské studium, specializované pro studenty distanční a kombinované formy studia. Inovací studijních oborů na SU OPF v rámci projektu OPVK vznikl také předmět Statistické zpracování dat. V tomto předmětu je kladen důraz především na uplatnění statistických metod při zpracování ekonomických dat v aplikovaných ekonomických disciplínách, jako jsou zejména marketing a management.

Samotný učební text, nebo jak se říká v moderní terminologii: *studijní opora* - umožňující studentovi plnohodnotné a zároveň samostatné studium – je rozčleněn do 12 tematických kapitol. Jednotlivé kapitoly odpovídají obvyklým výukovým týdnům jednoho semestru a jsou přibližně stejně obsahově rozsáhlé a obtížné. Takový rozsah učiva odpovídá klasické dvouhodinové přednášce v prezenčním studiu na vysoké škole ekonomického zaměření. V prezenčním studiu je ovšem na rozdíl od kombinované formy studia přednáška doplněna seminářem, kde se probraná látka aplikuje na konkrétní číselné příklady, které se řeší až k požadovanému výsledku pomocí počítače.

Vysokoškolské studium v případě předmětu Statistické zpracování dat vyžaduje enormní úsilí studenta zaměřené na pravidelnost a vytrvalost ve studiu i samostudiu, schopnost koncentrace na předmět, aktivní přístup spočívající v samostatném řešení příkladů. V tom všem by tato studijní opora měla studentům kombinované formy studia pomoci nahradit kvalitní prezenční výuku i úlohu učebnic a skript. Studijní opora je k tomu účelu vybavena určitými nástroji, o jejichž funkcích byste měli být informováni a mohli je tudíž účelně využívat ve svůj prospěch. Pro lepší zvládnutí látky jsou vám v elektronické verzi kurzu Statistické zpracování dat k dispozici ještě doplňkové materiály v elektronické podobě. Dalšími podpůrnými zdroji ke studiu mohou být klasické učebnice a skripta a další doporučená literatura.

Předpokladem pro úspěšné zvládnutí tohoto předmětu Statistické zpracování dat je zvládnutí bakalářského předmětu Statistika na SU OPF nebo odpovídajícího základního bakalářského kurzu Pravděpodobnosti – Statistiky, a to podle typu bakalářského studia na některé VŠ v ČR. Ne všechno, co jste se v základních kurzech statistiky naučili, zde využijete, řada věcí tam prezentovaných měla jiný účel. Rozhodně se vám však vyplatí nabytá schopnost přesného a logického uvažování, nezbytností je též zvládnutí matematické symboliky a základních partií teorie pravděpodobnosti a základů inferenční statistiky.

Nyní něco k obsahu předmětu Statistické zpracování dat. Přesnější název předmětu by zněl: Vybrané statistické metody zpracování dat pro ekonomy, nebo ještě jinak: Vybrané statistické metody zpracování dat a jejich použití v marketingu a managementu. To jsou totiž významné oblasti uplatnění statistických metod, s nimiž se absolventi Obchodně podnikatelské fakulty SU často v praxi setkávají. Obsahem kapitol 1 a 2 je analýza rozptylu - ANOVA, kapitoly 3 až 6 jsou věnovány regresní analýze - jednoduché i vícerozměrné, zbývající kapitoly 7 až 12 se věnují analýze ekonomických časových řad, ty jsou v ekonomických disciplínách mimořádně významné. Během studia budete využívat k řešení úloh známého programu Excel, s nímž jste pracovali již v předmětu Statistika.

Odměna, která vás na konci studia našeho předmětu očekává, stojí za to: je to pocit, že jste překonali něco významného, že jste se přenesli přes překážku, za níž se nachází svět profesionálů, kteří rozumějí odborným metodám a postupům, jež jsou obyčejným smrtelníkům nepřístupné. Získaný nadhled vám umožní snadněji pochopit a osvojit si praktické zásady analýzy informací, jimiž jsme všichni dnes zahlceni a v nichž je nám určeno žít.

1 ANALÝZA ROZPTYLU (ANOVA) – JEDEN FAKTOR

RYCHLÝ NÁHLED KAPITOLY

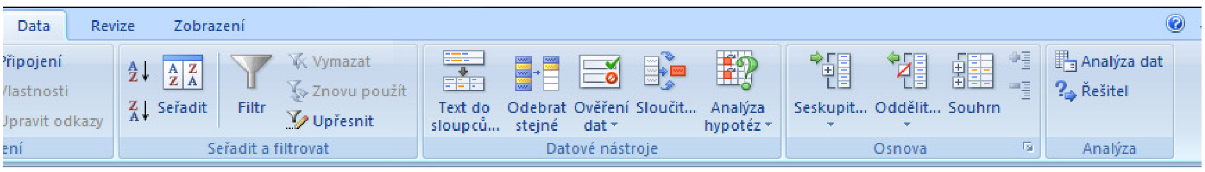
Jednofaktorová metoda ANOVA, kterou prokazujeme závislost hodnot znaků *Y* na faktoru *X*, pro něž jsou k dispozici příslušná data, spočívá v tom, že celkovou variabilitu měřenou součtem čtverců odchylek od celkového průměru rozdělíme na variabilitu uvnitř jednotlivých výběrů a na variabilitu mezi jednotlivými výběry. Cílem, k němuž směřujeme, je buď přijmout nulovou hypotézu o vzájemné nezávislosti *Y* na *X*, nebo ji zamítnout (na zvolené hladině významnosti). Jedná se tedy o běžný statistický postup nazývaný testování statistických hypotéz, známý ze základního kurzu statistiky. V případě přijetí nulové hypotézy vyvozujeme nezávislost hodnot *Y* na *X*, v opačném případě konstatujeme, že *Y* na *X* závisí. V této kapitole se naučíte, jak tento test statistické hypotézy konkrétně provést: jak vypočítat hodnotu testového kritéria a příslušnou kritickou hodnotu a jak vyvodit z těchto hodnot příslušný závěr týkající se eventuální závislosti nebo nezávislosti hodnot znaku *Y* na faktoru *X*.

Analýza rozptylu umožňuje ověřit významnost rozdílu mezi výběrovými průměry většího počtu náhodných výběrů, umožňuje posoudit vliv různých faktorů na hospodářský proces charakterizovaný kvantitativním statistickým znakem. Taktéž dovoluje hodnotit účinky různých přijatých hospodářských opatření. Základní myšlenka analýzy rozptylu spočívá v rozkladu celkového rozptylu na dílčí rozptyly příslušející jednotlivým vlivům, podle nichž jsou data roztříděna. Kromě dílčích rozptylů je jednou složkou celkového rozptylu tzv. reziduální rozptyl, způsobený nepostiženými vlivy. Podle počtu analyzovaných faktorů rozlišujeme *jednofaktorovou*, *dvoufaktorovou* a *vícefaktorovou* analýzu rozptylu. Všeobecně používané označení ANOVA je akronymem anglických slov „ANalysis Of VAriance" (doslovný překlad: analýza rozptylu).

Formálně vzato je ANOVA, ať jednofaktorová nebo vícefaktorová, testem statistické hypotézy, s nímž jste se seznámili v základním kurzu statistiky. Klasická ANOVA vychází, jak uvidíte, z předpokladu normality rozdělení hodnot daného faktoru. Pokud je takový předpoklad neudržitelný, lze použít jiného typu ANOVA, konkrétně Kruskal-Wallisovu verzi ANOVA. Jednofaktorovou ANOVA se zabývá tato kapitola, vícefaktorová a Kruskal-Wallisova ANOVA je obsahem kapitoly následující.

V tomto studijním textu předpokládáme, že čtenář má k dispozici verzi Excel 2007, eventuálně vyšší, pro osobní počítač typu PC s operačním systémem typu Windows. Pro zjednodušení práce je vhodné mít aktivovaný doplňky „Analýza dat" a „Řešitel" ve složce „Data" (viz Obr.1.1)

Obrázek 1.1

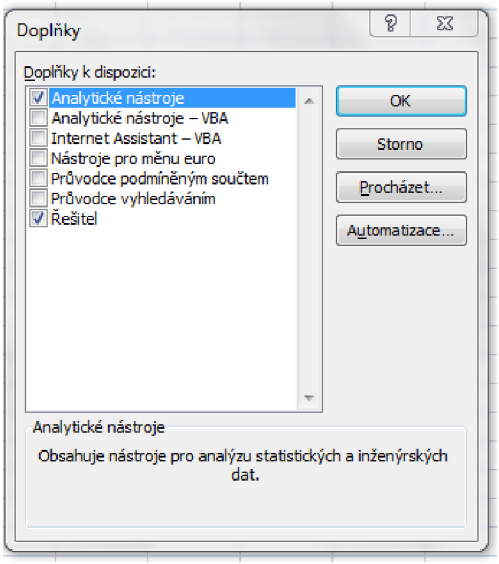


Zdroj: Vlastní zpracování.

V případě, že tyto doplňky nejsou ve složce „Data", lehce je nainstalujete tímto postupem:

„Tlačítko Office" → „Možnosti aplikace Excel" → „Doplňky" → „Přejít…" a v dialogovém okně zaškrtnout položky „Analytické nástroje" a „Řešitel" (viz Obr. 1.2).

Obrázek 1.2



Zdroj: Vlastní zpracování.

Kromě doplňků „Analýza dat“ a „Řešitel“ tabulkový procesor MS Excel disponuje širokým spektrem statistických funkcí. Všechny funkce procesoru MS Excel použité v následujícím textu budou značeny ve tvaru: =FUNKCE(proměnná1;…; proměnná N) se znaménkem „=“ na začátku; použití analytického nástroje bude značeno podobným způsobem.

1.1 NEZÁVISLÝ A ZÁVISLÝ FAKTOR

Často se vyskytuje situace, kdy máme k nezávislých náhodných výběrů které obecně nemusí pocházet z jednoho základního souboru, nebo jinak řečeno, nemusí být stejného typu, s rozsahy, tj. počty prvků $n_1, n_2, \ldots, n_k$. Číslo k může být libovolné podle konkrétní situace, např. 2, 3, 4, ... Tyto rozsahy výběrů rovněž nemusí být stejné, v každém z nich budiž znám průměr $\bar{x}_i$, a také rozptyl s_i^2, i = 1,2,...,k. V praktických situacích obvykle tyto výběry vzniknou tak, že základní soubor rozdělíme podle určitého statistického znaku X do k skupin, např. věkových, v každé z nich pak máme n_i prvků, i = 1,2,...,k. Znak X pak označujeme jako *nezávislý faktor*, jehož hodnoty předem stanovíme, stanovíme např. věkové skupiny takto: do 18 let, 19 až 29 let, 30 až 59 let, 60 a více let, v tomto příkladu je k = 4. Hovoříme proto často o *faktoru kontrolovaném*. Další příklady faktorů: velikost rodiny, měsíční příjem rodiny, velikost podniku, typ ekonomické činnosti, apod. Hodnotami faktoru X jsou obvykle *kvalitativní* (nečíselné) *veličiny*, označujeme je symbolicky $x_1, x_2, \ldots, x_k$. Tyto hodnoty mohou, ale nemusejí být nutně vzájemně uspořádány.

Faktor X, jež nabývá k kvalitativních hodnot, může, ale nemusí ovlivňovat hodnoty statistického znaku Y, o kterém předpokládáme, že má na rozdíl od X *kvantitativní* (tedy číselnou) povahu.

Cílem ANOVA je právě prokázat, že hodnoty kvalitativního znaku X ovlivňují hodnoty kvantitativního znaku Y- *závislého faktoru*. Hodnoty znaku Y, které přísluší hodnotě x_i faktoru X, označujeme $y_{i1}, y_{i2}, \ldots, y_{in_i}$. Pro analýzu rozptylu je výhodné uspořádat výchozí údaje do přehledné tabulky, viz Tab. 1.1.

Princip metody ANOVA, kterou prokazujeme závislost Y na X, spočívá v tom, že celkovou variabilitu měřenou součtem čtverců odchylek od celkového průměru rozdělíme na variabilitu uvnitř jednotlivých výběrů a na variabilitu mezi jednotlivými výběry. V následujícím odstavci tento postup upřesníme.

Číslo výběru	**Zjištěné hodnoty sledovaného znaku**	**Počet prvků**	**Průměr**	**Rozptyl**
1	$y_{11}, y_{12}, \dots, y_{1j}, \dots, y_{1n_1}$	n_1	$\overline{y}_1$	s_1^2
2	$y_{21}, y_{22}, \dots, y_{2j}, \dots, y_{2n_2}$	n_1	$\overline{y}_2$	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{in_i}$	n_i	$\overline{y}_i$	s_i^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$y_{k1}, y_{k2}, \dots, y_{kj}, \dots, y_{kn_k}$	n_k	$\overline{y}_k$	s_k^2
Celkem		n	$\overline{y}$	s^2

Tab. 1.1. *Schéma výchozí tabulky analýzy rozptylu pro jeden faktor*

Zdroj: Vlastní zpracování.

1.2 PŘEDPOKLADY ANALÝZY ROZPTYLU S JEDNÍM FAKTOREM

Předpokládáme, že faktor X má k úrovní (hodnot x_i), s účinkem na znak Y, který lze vyjádřit vztahem:

$$\mu_i = \mu + \alpha_i, \; i = 1,2,\dots,k,$$

kde μ_i je průměr znaku Y v i-té skupině (příslušné k hodnotě faktoru x_i),

μ je celkový průměr znaku Y,

α_i je efekt hodnoty faktoru x_i na znak Y.

Formulujeme nyní *nulovou hypotézu* H_0, že všechny výběry pocházejí ze stejné základní populace (základního souboru), jinak řečeno, že hodnoty faktoru X nemají na hodnoty znaku Y žádný efekt (vliv).

Budeme dále předpokládat, že hodnoty α_i pocházejí z *normálně rozdělené* populace s *nulovou střední hodnotou* a *konstantním rozptylem* σ^2.

Formulujeme nulovou hypotézu:

$$H_0: E(\alpha_1) = E(\alpha_2) = \dots = E(\alpha_k) = 0,$$

proti *alternativní hypotéze*, že H_0 neplatí, tudíž alespoň pro dvě položky, např. *i* a *j*, platí:

$$H_1: E(\alpha_i) \neq E(\alpha_j).$$

Symbolem $E(\alpha_i)$ označujeme *střední hodnotu* náhodné veličiny α_i. Předpoklad konstantního rozptylu pro všechny veličiny α_i je podstatný, je ho možno ověřit statistickým testem, a to buď tzv. *Bartlettovým testem*, s nímž se seznámíte později. Normalitu rozdělení veličin α_i lze taktéž ověřit příslušným testem, např. *Chi-kvadrát testem dobré shody*, známým ze základního kurzu statistiky, viz Ramík (2003). V praxi obvykle předpokládáme (na podkladě věcné znalosti problému), že zmíněné dva předpoklady jsou automaticky splněny a při aplikaci ANOVA je již obvykle neověřujeme.

Cílem, k němuž směřujeme, je buď *přijmout* nulovou hypotézu H_0, nebo H_0 *zamítnout* (na zvolené *hladině významnosti*). Jedná se tedy o běžný statistický postup nazývaný *testování statistických hypotéz*, známý ze základního kurzu statistiky, viz Ramík (2003). V případě přijetí nulové hypotézy vyvozujeme nezávislost hodnot faktoru Y na faktoru X, jinak řečeno: faktor Y na faktoru X nezávisí. V opačném případě (při zamítnutí H_0), konstatujeme, že faktor Y na faktoru X závisí, neboli faktor X ovlivňuje Y.

1.3 POSTUP PŘI ANALÝZE ROZPTYLU S JEDNÍM FAKTOREM

Celkovou variabilitu znaku Y změříme *výběrovým rozptylem*

$$s^2 = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y}\right)^2}{n-1}. \tag{1.1}$$

V souvislosti s analýzou rozptylu se budeme zabývat pouze čitatelem výše uvedeného zlomku, totiž součtem čtverců odchylek zjištěných hodnot y_{ij} od celkového průměru $\bar{y}$, přičemž průměr vypočítáme podle známého vztahu: sečteme všechny hodnoty a výsledek podělíme jejich počtem, tedy

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}.$$

Tento *celkový součet čtverců* budeme označovat symbolem S_y, tj.

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2. \tag{1.2}$$

Celkovému součtu čtverců přísluší počet stupňů volnosti $df_y = n - 1$.
Variabilitu mezi skupinami budeme měřit *meziskupinovým součtem čtverců* $S_{y,m}$, který definujeme následovně

$$S_{y,m} = \sum_{i=1}^{k} n_i\left(\bar{y}_i - \bar{y}\right)^2. \tag{1.3}$$

Meziskupinovému součtu čtverců přísluší počet stupňů volnosti $df_m = k - 1$.

Variabilitu uvnitř skupin označujeme jako vnitroskupinovou, nebo také reziduální a používáme přitom označení $S_{y,v}$, přičemž definujeme *vnitroskupinový* (*reziduální*) *součet čtverců* takto

$$S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2. \tag{1.4}$$

Vnitroskupinovému součtu čtverců přísluší počet stupňů volnosti $df_v = n - k$.

Aritmetickými úpravami výše uvedených vzorců lze snadno dokázat základní vztah analýzy rozptylu, totiž, že *celkový součet čtverců je roven sumě meziskupinového a vnitroskupinového součtu čtverců,* symbolicky:

$$S_y = S_{y,m} + S_{y,v}. \tag{1.5}$$

Pro ověření nulové hypotézy H_0 použijeme statistiku:

$$F = \frac{\dfrac{S_{y,m}}{k-1}}{\dfrac{S_{y,v}}{n-k}} = \frac{\dfrac{S_{y,m}}{df_m}}{\dfrac{S_{y,v}}{df_v}}\ , \tag{1.6}$$

která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F(k-1, n-k)$. Kritické hodnoty Fisherova rozdělení $F_\alpha(df_1, df_2)$ jsou tabelovány pro různé hodnoty hladiny významnosti α a různé hodnoty parametrů (stupňů volnosti: *d*egree of *f*reedom) df_1 a df_2 . Někdy se namísto kritických hodnot tabelují kvantily Fisherova rozdělení $F^k_{1-\alpha}(df_1, df_2)$. Vztah mezi kritickými hodnotami a kvantily je jednoduchý:

$F_\alpha(df_1, df_2) = F^k_{1-\alpha}(df_1, df_2)$.

Např. 5-ti procentní kritická hodnota je rovna 95-ti procentnímu kvantilu při stejných hodnotách parametrů df_1 a df_2.

Pro výpočet kritických hodnot lze využít Excelu. Postupuje se přitom takto: v hlavním menu postupně vybíráte: Vložit → Funkce → Statistické → FINV(α;df_1 ; df_2).

Postup testování hypotézy H_0 charakterizujeme následujícími 3 kroky:

Krok 1. Zvolte hladinu významnosti α, která představuje chybu 1. druhu, tj. pravděpodobnost zamítnuti správné hypotézy. Praktické hodnoty hladiny významnosti α jsou: 0,1 , 0,05 , 0,01 , nebo-li v procentech: 10%, 5%, 1%.

Krok 2. Vypočtěte hodnotu statistiky F podle vzorce (1.6), přičemž pro hodnoty meziskupinového součtu čtverců $S_{y,m}$ a pro výpočet vnitroskupinového součtu čtverců $S_{y,v}$ použijte vzorce (1.3) a (1.4). Výpočetně výhodnější, např. pro výpočet na kalkulačce, jsou následující vzorce:

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2, \tag{1.7}$$

$$S_{y,m} = \sum_{i=1}^{k} n_i \bar{y}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2. \tag{1.8}$$

K výpočtu $S_{y,v}$ lze využít základního vztahu (1.5) a právě uvedených vztahů (1.7) a (1.8):

$S_{y,v} = S_y - S_{y,m}$.

Krok 3. Porovnejte hodnotu statistiky F vypočtené v Kroku 2 s kritickou hodnotou $F_\alpha(k-1, n-k)$. Výsledek tohoto porovnání může být dvojí:

I. Platí $F \le F_\alpha(k-1, n-k)$.

Potom se nulová hypotéza H_0 *přijímá* (nezamítá) a tudíž se konstatuje, že hodnoty faktoru X *nemají* na hodnoty znaku Y *statisticky významný vliv* (na zvolené hladině významnosti). Jinak řečeno, faktor X je *neúčinný*.

II. Platí $F > F_{\alpha}(k-1, n-k)$.

Potom se nulová hypotéza H_0 *zamítá,* přijímá se hypotézu alternativní H_1, a tudíž se konstatuje, že hodnoty faktoru *X mají* na hodnoty znaku *Y* statisticky významný vliv. Jinak řečeno, faktor *X* je *účinný.*

Podaří-li se výše uvedeným testem prokázat, že hodnoty faktoru *X* mají na hodnoty znaku *Y* statisticky významný vliv, mohou nás zajímat další informace o tom, které skupiny se významně odlišují od průměru, eventuálně jak skupinové průměry seřadit, případně zařadit do společných celků. V krajním případě by se totiž mohlo stát, že významnost rozdílnosti *k* skupin způsobuje jediná skupina a ostatní skupiny se navzájem neliší. Touto problematikou se zabývají metody tzv. *simultánního testování*, z nichž nejznámější je metoda Shaffeho. Vy se touto problematikou zde nezabývat nebudete, zájemce odkazujeme na literaturu, viz např. Anděl (2007).

Metoda analýzy rozptylu je založena na předpokladech shody rozptylů v jednotlivých *k* skupinách. Pokud jsou předpoklady splněny, pak popsaná metoda ANOVA poskytuje nejlepší výsledky – je nejúčinnější. Není-li tento předpoklad splněn, pak použití výše uvedeného testu může poskytnout nesprávný výsledek. V takovém případě lze použít jiné metody, např. Kruskal-Wallisova ANOVA, která používá Chi-kvadrát test, s níž se seznámíte v příští kapitole.

V Excelu jsou k dispozici funkce, které umožňují řešit jednofaktorové i vícefaktorové úlohy ANOVA. Naleznete je v hlavním menu: Nástroje → Analýza dat → ANOVA: jeden faktor... V tomto textu se s nimi naučíte pracovat.

1.4 MÍRA TĚSNOSTI ZÁVISLOSTI

Variabilita podmíněných (skupinových) průměrů $\overline{y}_i$ kolem celkového průměru $\overline{y}$ je způsobena závislostí znaku *Y* na znaku *X*. Tuto variabilitu jsme vyjádřili meziskupinovým součtem čtverců $S_{y,m}$. Variabilita znaku *Y* uvnitř jednotlivých skupin – vyjádřena vnitroskupinovým (reziduálním) součtem čtverců $S_{y,v}$, je způsobena jinými (neuvažovanými) činiteli. Čím větší je $S_{y,m}$, tím větší je těsnost závislosti znaků *X* a *Y*. Protože však jsou jednotlivé součty čtverců vzájemně vázány vztahem (1.5), lze míru těsnosti závislosti vyjádřit jako podíl meziskupinového a celkového součtu čtverců. Zavádíme proto jako míru těsnosti závislosti znaku *Y* na znaku *X poměr determinace* P^2 takto:

$$P^2 = \frac{S_{y,m}}{S_y}. \tag{1.9}$$

Odmocninu z poměru determinace *P* nazýváme *poměr korelace*.

Poměr determinace nabývá hodnot z intervalu [0,1]. Čím těsnější je závislost *Y* na *X*, tím více se hodnota poměru determinace blíží k 1, tím více se také vnitroskupinový součet čtverců blíží k celkovému součtu čtverců, přičemž meziskupinový součet čtverců se blíží k nule. Naopak, čím více se poměr determinace blíží k 0, tím menší část z celkového součtu čtverců tvoří meziskupinový součet čtverců (na úkor vnitroskupinového), a tím menší je těsnost závislosti znaku *Y* na *X*. Způsob výpočtu determinačního a korelačního poměru si procvičíte na numerických příkladech. V Excelu bohužel funkce pro výpočet poměru determinace nebo korelace chybí, musí se proto k výpočtu použít vzorce (1.9).

Uvědomte si však, že poměr determinace P^2 je náhodná veličina (jakožto podíl dvou veličin – součtu čtverců, které jsou samy náhodnými veličinami), proto může být výsledkem kladné číslo i v případě, že výsledkem ANOVA je fakt, že zkoumaný faktor není statistický významný, neboli sledovaná veličina na faktoru nezávisí. V takovém případě by logicky mělo platit, že poměr determinace P^2 je nulový, tj. $P^2 = 0$. Tento zdánlivý rozpor vysvětlujeme statistickým přístupem: testem statistické hypotézy. V tomto případě je nulová hypotéza H_0: $P^2 = 0$. Jako testové kritérium se použije statistika F ze vzorce (1.6).

Pokud platí $F \leq F_\alpha(k-1, n-k)$, potom se nulová hypotéza H_0 přijímá (a tudíž konstatujeme, že hodnoty faktoru X nemají na hodnoty znaku Y statisticky významný vliv na zvolené hladině významnosti) a poměr determinace (samozřejmě i poměr korelace) je roven nule, jinak řečeno, je statisticky nevýznamný.

V opačném případě se nulová hypotéza zamítá a poměr determinace je statisticky významný. Hodnota poměru determinace i poměru korelace je nenulová. V tom případě má smysl hovořit o síle závislosti veličiny Y na faktoru X.

ŘEŠENÝ PŘÍKLAD 1.1

Na testovacím okruhu byla testována průměrná spotřeba tří automobilů téže třídy různých výrobců Škoda, Renault a Fiat. Řidič absolvoval s každým automobilem 5 testovacích jízd. Tabulka ukazuje spotřebu benzínu na 100 kilometrů v jednotlivých jízdách.

Automobil	Spotřeba				
Škoda	7,4	7,8	6,8	7,6	8,1
Renault	6,7	7,2	8,3	7,1	7,5
Fiat	6,8	6,9	7,3	7,9	7,6

Na hladině významnosti $\alpha = 0{,}05$ zjistěte, zda má typ automobilu vliv na spotřebu benzínu. V kladném případě vypočtěte determinační a korelační poměr.

Řešení:

Chceme zjistit závislost znaku Y (průměrná spotřeba) na jediném znaku X (výrobce automobilu). Provedeme proto jednofaktorovou analýzu rozptylu.

Faktor X má tři hodnoty: x_1 = Škoda, x_2 = Renault, x_3 = Fiat, tzn. $k = 3$, s počty hodnot $n_1 = n_2 = n_3 = 5$ v každé z nich budeme testovat nulovou hypotézu

H_0: $E(\alpha_1) = E(\alpha_2) = E(\alpha_2) = 0$, tj. průměrná spotřeba je u všech vozidel stejná.

Alternativní hypotéza H_1 je negací nulové hypotézy.

Nejprve vypočítáme podmíněné průměry $\bar{y}_1, \bar{y}_2, \bar{y}_3$

$$\bar{y}_1 = \frac{\sum_{j=1}^{5} y_{1j}}{5} = \frac{7{,}4 + 7{,}8 + \ldots + 8{,}1}{5} = 7{,}54$$

$$\bar{y}_2 = \frac{\sum_{j=1}^{5} y_{2j}}{5} = \frac{6{,}7 + 7{,}2 + \ldots + 7{,}5}{5} = 7{,}36$$

$$\bar{y}_3 = \frac{\sum_{j=1}^{5} y_{3j}}{5} = \frac{6{,}8 + 6{,}9 + \ldots + 7{,}6}{5} = 7{,}3$$

a celkový průměr znaku Y

$$\bar{y} = \frac{\sum y_{ij}}{n} = \frac{7{,}4 + 7{,}8 + \ldots + 7{,}6}{15} = 7{,}4\,.$$

Dále vypočítáme pomocí vztahů (1.2), (1.3), popř. (1.7), (1.8) součty S_y a S_{ym}.

$$S_y = \sum_{i=1}^{3}\sum_{j=1}^{5}(y_{ij} - \bar{y})^2 = (7{,}4 - 7{,}4)^2 + (7{,}8 - 7{,}4)^2 + \ldots + (8{,}1 - 7{,}4)^2 +$$

$$+ (6{,}7 - 7{,}4)^2 + (7{,}2 - 7{,}4)^2 + \ldots + (7{,}5 - 7{,}4)^2 +$$

$$+ (6{,}8 - 7{,}4)^2 + \ldots + (7{,}6 - 7{,}4)^2 = 3{,}4$$

$$S_{ym} = \sum_{i=1}^{3} n_i (y_{ij} - \bar{y})^2 = 5(\bar{y}_1 - \bar{y})^2 + 5(\bar{y}_2 - \bar{y})^2 + 5(\bar{y}_3 - \bar{y})^2 =$$

$$= 5(7{,}54 - 7{,}4)^2 + 5(7{,}36 - 7{,}4)^2 + 5(7{,}3 - 7{,}4)^2 = 0{,}16\,.$$

Součet S_{ym} má k - 1 stupňů volnosti, v našem případě $df_m = 3 - 1 = 2$.
Pomocí součtů S_y a S_{ym} dopočítáme součet S_{yv}, neboť $S_y = S_{yv} + S_{ym}$.
Proto

$$S_{yv} = S_y - S_{ym} = 3{,}4 - 0{,}16 = 3{,}24.$$

Součet S_{yv} má $n - k$ stupňů volnosti, proto $df_v = 15 - 3 = 12$.

Testové kritérium F vypočítáme podle vztahu (1.6):

$$F = \frac{\dfrac{S_{ym}}{k-1}}{\dfrac{S_{yv}}{n-k}} = \frac{\dfrac{0{,}16}{2}}{\dfrac{3{,}24}{12}} = 0{,}296\,.$$

Pro stanovení kritického oboru C najdeme v tabulkách kritických hodnot $F_\alpha(k - 1, n - k)$ kritickou hodnotu $F_{0,05}(2, 12) = 3{,}89$ (ověřte v Excelu pomocí funkce FINV). Kritický obor je proto interval od 3,89 do nekonečna, tj.

$$C = (3{,}89, +\infty)\,.$$

Zřejmě platí $0{,}296 < 3{,}89$, tzn. $F \notin C$, proto nulovou hypotézu H_0 přijímáme.

Znamená to, že faktor X-výrobce automobilu je neúčinný nebo-li, že průměrná spotřeba benzínu není statisticky významně ovlivněna výrobcem automobilu. Poměr determinace i korelace je tedy 0.

ŘEŠENÝ PŘÍKLAD 1.2

Rozhodněte, zda velikost výnosů petržele (faktor *Y*) závisí na použitém druhu hnojiva (faktor *X*). Pokud závisí, pak pomocí determinačního poměru zjistěte těsnost této závislosti. Data jsou uvedena v následující tabulce, použijte hladinu významnosti 0,05.

Hnojivo	Výnosy (1kg/10 m^2)					
A	40	42	45	40	44	47
B	76	75	82	68		
C	60	58	62	64	70	

Řešení:

U tohoto příkladu si ukážeme řešení s pomocí Excelu. Nejprve však příklad vyřešíme klasickým postupem.

K výpočtu hodnot součtů čtverců S_{ym} a S_y, potřebujeme znát celkový průměr $\overline{y}$ a podmíněné průměry $\overline{y}_1, \overline{y}_2, \overline{y}_3$.

$$\overline{y}_1 = \frac{\sum_{j=1}^{6} y_{1j}}{n_1} = \frac{40+42+\ldots+47}{6} = 43,$$

$$\overline{y}_2 = 75{,}25; \overline{y}_3 = 62{,}8,$$

$$\overline{y} = \frac{\sum_{i=1}^{3} n_i \overline{y}_i}{n} = \frac{43\cdot 6 + 75{,}25\cdot 4 + 62{,}8\cdot 5}{15} = 58{,}2.$$

Nyní již můžeme vypočítat součty S_{ym} a S_y, podle vztahů (1.2), (1.3)

$$S_y = \sum_{ij} (y_{ij} - \overline{y})^2 = (40-58{,}2)^2 + \ldots + (47-58{,}2)^2 +$$
$$+ (76-58{,}2)^2 + \ldots + (68-58{,}2)^2 +$$
$$+ (60-58{,}2)^2 + \ldots + (70-58{,}2)^2 = 2878{,}4.$$

$$S_{ym} = \sum_{i=1}^{3} n_i (\overline{y}_i - \overline{y})^2 = 6(43-58{,}2)^2 + 4(75{,}25-58{,}2)^2 + 5(62{,}8-58{,}2)^2 = 2654{,}85.$$

Hodnota testového kritéria je $F = \dfrac{\dfrac{S_{ym}}{k-1}}{\dfrac{S_{yv}}{n-k}} = \dfrac{\dfrac{2654{,}85}{2}}{\dfrac{2878{,}4-2654{,}85}{12}} = 71{,}26$.

Kritická hodnota je $F_{0,05}(2,\ 12) = 3{,}89$ a je mnohem menší než hodnota testového kritéria *F*. Proto nulovou hypotézu zamítáme a konstatujeme, faktor hnojiva významně ovlivňuje hodnoty výnosů petržele.

Hodnotu determinačního poměru P^2 zjistíme dosazením hodnot S_{ym} a S_y do vztahu (1.9).

$$P^2 = \frac{2654{,}85}{2878{,}4} = 0{,}92.$$

Hodnoty determinačního poměru blízké 1 svědčí o vysoké závislosti faktoru *Y* na faktoru *X*. Hodnota 0,92 proto znamená, že závislost výnosů petržele na použitém druhu hnojiva je vysoká.

Řešení pomocí Excelu:

Nejprve je zapotřebí připravit v Excelu data. Jednotlivé hodnoty y_{ij} pro faktoru Y pro hodnotu x_i faktoru X uspořádáme do řádků, podobně jako v tabulce v zadání. V prvním sloupci umístíme kvůli lepší orientaci název hodnoty faktoru (popisky) x_i, v tomto případě název hnojiva: A, B, C. Data ve worksheetu Excelu vypadají tedy například takto:

	A	B	C	D	E	F	G	H
1	A	40	42	45	40	44	47	
2	B	76	75	82	68			
3	C	60	58	62	64	70		
4								

Data je možné uspořádat také do sloupců, přitom do prvního řádku umístíme názvy hodnot faktoru X (popisky). To je výhodné zejména u velkého množství dat, tj. pro velkou hodnotu počtu dat n.

Dále otevřeme v hlavním menu postupně položky:

Data → Analýza dat... → ANOVA: jeden faktor

Pokud se tam položka Analýza dat nevyskytuje je ji zapotřebí doinstalovat (viz začátek této kapitoly).

Zvolíte-li pak první položku ANOVA: jeden faktor, otevře se zadávací okno, kde postupně zadáte:

Vstupní oblast: $A$1:$G$3

Sdružit: zakliknete tlačítko *Řádky* (je možné uspořádat data do sloupců, pak ovšem zakliknete tlačítko *Sloupce*

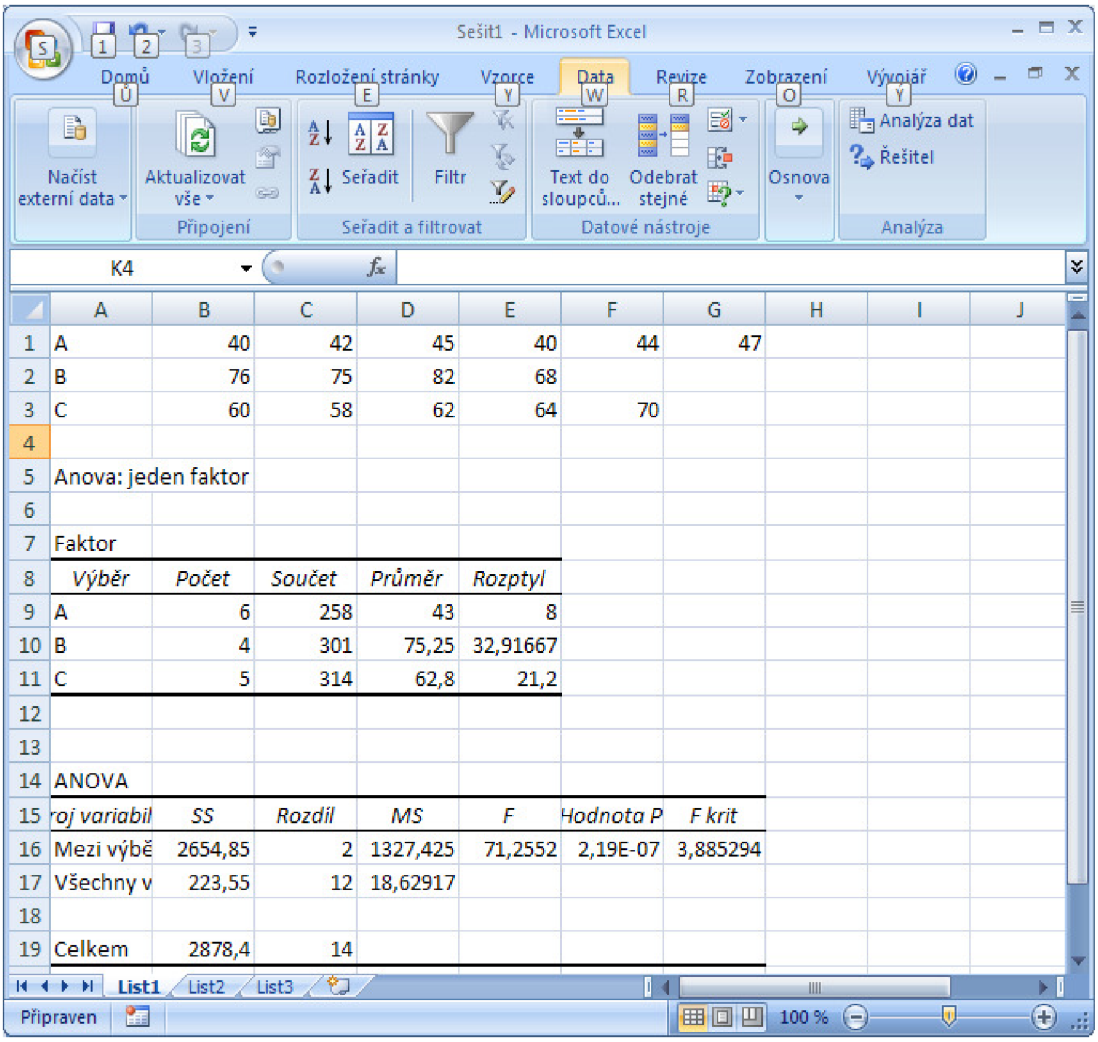
Popisky v prvním sloupci – zakliknete

Alfa: 0,05 (hladina významnosti je předvolena, lze ji však změnit)

Výstupní oblast: $A$5 (levý horní roh výstupní oblasti)

Potvrdíte OK

Obdržíte následující výstup, kterého “levý horní roh” začíná v buňce A5 nadpisem Anova jeden faktor:



	A	B	C	D	E	F	G
1	A	40	42	45	40	44	47
2	B	76	75	82	68		
3	C	60	58	62	64	70	
4							
5	Anova: jeden faktor						
6							
7	Faktor						
8	*Výběr*	*Počet*	*Součet*	*Průměr*	*Rozptyl*		
9	A	6	258	43	8		
10	B	4	301	75,25	32,91667		
11	C	5	314	62,8	21,2		
12							
13							
14	ANOVA						
15	*roj variabil*	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
16	Mezi výbě	2654,85	2	1327,425	71,2552	2,19E-07	3,885294
17	Všechny v	223,55	12	18,62917			
18							
19	Celkem	2878,4	14				

V první tabulce s názvem Faktor jsou uvedeny základní statistické údaje o datech: Počet, Součet, Průměr a Rozptyl.

Ve druhé tabulce nazvané ANOVA jsou uvedeny výpočty metodou ANOVA, jednotlivé položky mají následující význam:

Mezi výběry = meziskupinový
Všechny výběry = vnitroskupinový
Celkem = celkový
SS = Součet čtverců (Sum of Squares)
Rozdíl = stupeň volnosti (DF – Degree of Freedom)
MS = Průměr čtverců (Mean Square)
F = testové kritérium = 71,25
Hodnota P = Signifikance (p-hodnota) = 0,000000219 $< 0{,}05 = \alpha$
F krit = kritická hodnota rozdělení F = 3,89

Hodnoty získané řešením v Excelu jsou stejné jako při použití „ručního" výpočtu, proto i závěry jsou stejné. V Excelu máme navíc vypočtenu p-hodnotu testu (tzv. signifikanci), která, pokud je menší než zvolená hladina významnosti α, znamená, že nulovou hypotézu zamítáme. V opačném případě nulovou hypotézu nezamítáme (přijímáme).

1.5 Samostatné úkoly

1.1 Pan Novák může jet do zaměstnání čtyřmi různými trasami. Čtyřikrát projel jednotlivé trasy a zaznamenal si dobu, po kterou jel do zaměstnání. Na hladině významnosti $\alpha = 0{,}01$ zjistěte, zda záleží na tom, kterou trasou pojede.

Cesta 1	Cesta 2	Cesta 3	Cesta 4
22	27	26	28
26	29	33	30
25	26	25	32
30	28	30	26

1.2 Učitel fyziky zkoumal, jaký vliv má druh zkušebního testu na jeho úspěšnost. Vytvořil tři typy stejně obtížných testů a náhodně je rozdal mezi studenty ve třídě. Tabulka uvádí bodové zisky studentů v jednotlivých testech. Na hladině významnosti $\alpha = 0{,}05$ zjistěte, zda má typ testu vliv na úspěšnost studentů.

Typ testu		
T1	**T2**	**T3**
75	72	64
90	78	78
70	94	70
90	78	90
85		50

1.3 Ve vepříně zjišťovali, jestli váhové přírůstky vepřů závisí na použitém druhu krmiva, či nikoli. Na hladině významnosti $\alpha = 0{,}05$ rozhodněte, zda jsou váhové přírůstky pro různá krmiva různé, eventuálně zjistěte, který druh krmiva dává nejmenší váhové přírůstky.

Krmivo		
A	**B**	**C**
21,5	19,9	23,7
22,8	24,3	22,5
26,3	20,1	20,6
24,2	20,9	21,4
25,6	21,1	
28,1		

1.4 Výroba součástek může v podniku probíhat na jednom ze čtyř rozdílných strojů. I když každý stroj provádí stejné operace, má každý svá specifika. Na hladině významnosti $\alpha = 0{,}01$ testujte hypotézu o tom, že počet vyrobených součástek není ovlivněn volbou stroje.

Stroj			
A	**B**	**C**	**D**
93	108	123	133
98	153	143	163
80	123	150	168
88	158	165	145
60	143	140	130

1.5 Školský úřad Karviná chtěl srovnat úroveň znalostí maturantů gymnázií okresu Karviná. Za tímto účelem byl vytvořen test zahrnující otázky ze všech oblastí učiva a zadán náhodně vybraným studentů jednotlivých škol. Bodové výsledky studentů jsou uvedeny v následující tabulce.

Gymnázium Karviná	Gymnázium Český Těšín	Gymnázium Bohumín	Gymnázium Orlová	Gymnázium Havířov
79	62	74	73	86
86	54	81	67	52
49	88	64	59	61
72			76	

a. Na hladině významnosti $\alpha = 0{,}05$ zjistěte, je-li průměrná úroveň maturantů jednotlivých škol stejná.
b. Jak ovlivní výsledek průzkumu změna hladiny významnosti na 0,01?

1.6 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

1.1 F = 1,0 F krit = 5,95 p-hodnota = 0,43 – H_0 přijímáme (je jedno, kterou cestu zvolí).

1.2 F = 1,43 F krit = 3,98 p-hodnota = 0,28 – H_0 přijímáme (typ testu nemá vliv na úspěch).

1.3 F = 4,7 F krit = 3,89 p-hodnota = 0,03 – H_0 zamítáme (krmivo má vliv, nejvíce A).

1.4 F =15,02 F krit = 5,29 p-hodnota = 0,000 – H_0 zamítáme (typ stroje má vliv).

1.5 a) F = 0,12 F krit = 3,26 p-hodnota = 0,97 – H_0 přijímáme (škola nemá vliv).
b) F = 0,12 F krit = 5,41 p-hodnota = 0,97 – H_0 přijímáme (škola nemá vliv).

2 ANALÝZA ROZPTYLU (ANOVA) – DVA A VÍCE FAKTORŮ

RYCHLÝ NÁHLED KAPITOLY

Jednofaktorová metoda ANOVA, kterou prokazujeme závislost znaků (faktorů) *Y* na *X*, pro něž jsou k dispozici příslušná data, spočívá v tom, že celkovou variabilitu měřenou součtem čtverců odchylek od celkového průměru rozdělíme na variabilitu uvnitř jednotlivých výběrů a na variabilitu mezi jednotlivými výběry. Cílem, k němuž směřujeme nyní, je situace, kdy budeme uvažovat, že se kromě třídění do skupin vyskytují další faktory, říkáme jim bloky, podle nichž výsledky (tj. hodnoty znaku *Y*) rovněž třídíme.

2.1 ANALÝZA ROZPTYLU SE DVĚMA FAKTORY

U analýzy rozptylu s jedním faktorem jste uvažovali výsledky tříděné podle jistého kvalitativního znaku *X* do několika (konkrétně do *k*) skupin o rozsazích $n_1, n_2, ..., n_k$. Proto v tomto případě hovoříme také o ANOVA při jednoduchém třídění, neboli třídění podle jednoho faktoru. V této kapitole budeme uvažovat situaci, kdy se kromě třídění do skupin, vyskytují další faktory, říkáme jim bloky, podle nichž výsledky (tj. hodnoty znaku *Y*) rovněž třídíme. Přehledná situace vzniká, když kromě prvního faktoru uvažujeme ještě faktor druhý, říkáme pak, že je třídíme do bloků a v takovém případě se jedná o dvoufaktorovou ANOVA. Formálně vzato je ANOVA, ať jednofaktorová, dvoufaktorová nebo vícefaktorová, parametrickým testem statistické hypotézy, s nímž jste se seznámili v základním kurzu statistiky. Tato tzv. klasická ANOVA vychází z předpokladu normality rozdělení hodnot uvažovaných faktorů. Pokud je takový předpoklad neudržitelný, lze použít jiného typu ANOVA, tedy neparametrického testu statistické hypotézy (tento pojem si připomeňte ze základního kurzu statistiky). Konkrétně se v této kapitole seznámíte s Kruskal-Wallisovu verzi ANOVA, která využívá Chi-kvadrát test statistické hypotézy.

U analýzy rozptylu s jedním faktorem jsme uvažovali výsledky tříděné podle jistého kvalitativního znaku *X* do několika (konkrétně do *k*) *skupin* o rozsazích $n_1, n_2, ..., n_k$. V tomto odstavci budeme uvažovat situaci, kdy se kromě třídění do skupin, vyskytuje další faktor, podle něhož výsledky (tj. hodnoty znaku *Y*) rovněž třídíme, říkáme, že je třídíme do *bloků*. Začneme výklad příkladem známým již z předchozí kapitoly.

Příklad 1. Testovacími jízdami na zkušebním okruhu se zjišťuje průměrná spotřeba paliva automobilu Octavia při použití benzínu od různých výrobců (např. Aral, Shell, Benzina, Slovnaft). Všechny testy provede jeden řidič, když s každým druhem benzínu uskuteční několik testovacích jízd, a to tak, že pro každou značku benzínu uskuteční jiný počet jízd. Zjištěné výsledky testů, tj. změřené průměrné spotřeby na 100 km, podrobíme jednofaktorové analýze rozptylu, která nám umožní zjistit, zda značka (tj. výrobce) použitého benzínu má vliv na průměrnou spotřebu automobilu.

Příklad 2. Nyní budeme uvažovat podobnou situaci, kdy výsledky testů byly získány různými řidiči (např. A, B, C, D, E, F), a to tak, že každý řidič uskutečnil jednu testovací jízdu s každou značkou benzínu. Výsledky testů proto budeme členit nejen podle značky benzínu - do skupin (1. faktor), ale také podle testovacích řidičů - do bloků (2. faktor). Podle předpokladů je nyní počet výsledků ve všech skupinách stejný a je roven počtu řidičů (každý řidič jel s jednou značkou benzínu jedenkrát). Zjištěné výsledky podrobíme dvoufaktorové analýze rozptylu, která umožní jednak zjistit, zda značka (tj. výrobce) použitého benzínu má vliv na průměrnou spotřebu automobilu, jednak zjistit, zda různí řidiči mají vliv na tuto spotřebu.

Příklad 3. Nyní budeme uvažovat stejnou situaci jako v příkladu 2, přitom výsledky testů byly získány různými řidiči (např. A, B, C, D, E, F), a to tak, že každý řidič uskutečnil tři testovací jízdy s každou značkou benzínu. Zjištěné výsledky podrobíme dvoufaktorové analýze rozptylu s opakováním, která umožní jednak zjistit, zda značka (tj. výrobce) použitého benzínu má vliv na průměrnou spotřebu automobilu, jednak zjistit, zda různí řidiči mají vliv na tuto spotřebu.

Na konci této kapitoly budou všechny tři příklady podrobně analyzovány na konkrétních číselných datech. Nyní budeme postupovat ve výkladu s obecnými daty, nejprve pro případ popsaný v příkladu 2. Taková data, podobně jako u jednofaktorové analýzy rozptylu, uspořádáme do přehledné tabulky Tab. 2.1.

Číslo skupiny	Hodnoty sledovaného znaku Číslo bloku						
	1	**2**	**...**	***j***	**...**	***r***	**Průměr skupiny**
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1r}	$\overline{y}_{1\bullet}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2r}	$\overline{y}_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ir}	$\overline{y}_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kr}	$\overline{y}_{k\bullet}$
Průměr bloku	$\overline{y}_{\bullet 1}$	$\overline{y}_{\bullet 2}$	...	$\overline{y}_{\bullet j}$	...	$\overline{y}_{\bullet r}$	$\overline{y}$

Tab. 2.1. *Schéma výchozí tabulky analýzy rozptylu pro dva faktory*

V Tab. 2.1. značíme symbolem $\overline{y}_{i\bullet}$ průměr v i-té skupině, symbolem $\overline{y}_{\bullet j}$ označujeme průměr hodnot v j-tém bloku, symbolem $\overline{y}$ značíme celkový průměr.

Celkový součet čtverců (celkovou variabilitu) označujeme stejně, jako v (1.2), tedy:

$$S_y = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(y_{ij} - \overline{y}\right)^2 . \tag{2.1}$$

Variabilitu mezi skupinami budeme měřit *meziskupinovým součtem čtverců* $S_{y,m}$, který definujeme následovně:

$$S_{y,m} = r\sum_{i=1}^{k}\left(\overline{y}_{i\bullet} - \overline{y}\right)^2 . \tag{2.2}$$

Meziskupinovému součtu čtverců přísluší počet stupňů volnosti $df_m = k - 1$.

Variabilitu mezi bloky budeme měřit *meziblokovým součtem čtverců* $S_{y,b}$, který definujeme následovně:

$$S_{y,b} = k\sum_{j=1}^{r}\left(\bar{y}_{\bullet j} - \bar{y}\right)^2 . \tag{2.3}$$

Meziskupinovému součtu čtverců přísluší počet stupňů volnosti $df_b = r - 1$.

Variabilitu uvnitř skupin označujeme jako vnitroskupinovou, nebo také reziduální a používáme přitom označení $S_{y,v}$, přičemž definujeme *vnitroskupinový* (*reziduální*) *součet čtverců* takto

$$S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}\right)^2 . \tag{2.4}$$

Vnitroskupinovému součtu čtverců přísluší počet stupňů volnosti $df_v = (k - 1)(r - 1)$.

Aritmetickými úpravami výše uvedených vzorců lze dokázat totiž, že *celkový součet čtverců je roven sumě meziskupinového, vnitroskupinového a blokového součtu čtverců,* symbolicky

$$S_y = S_{y,m} + S_{y,v} + S_{y,b} . \tag{2.5}$$

Tento vztah se nazývá *základní vztah dvoufaktorové analýzy rozptylu.*

2.2 Předpoklady ANOVA se 2 faktory

Předpokládáme, že faktor X_1 má k úrovní, faktor X_2 má r úrovní s efektem na znak Y, který lze vyjádřit vztahem

$$\mu_{ij} = \mu + \alpha_i + \beta_j ,\quad i = 1,2,...,k,\quad j = 1,2,...,r, \tag{2.6}$$

kde μ_{ij} je průměr znaku Y v i-té skupině a j-tém bloku,

μ je celkový průměr znaku Y,

α_i je efekt hodnoty faktoru X_1 na znak Y,

β_j je efekt hodnoty faktoru X_2 na znak Y.

V modelu (2.6) nejprve předpokládáme, že efekty obou faktorů na znak Y jsou *aditivní a vzájemně nezávislé*, tj. bez vzájemných interakcí. Tento předpoklad nám umožní oddělit od sebe hypotézy o efektech jednotlivých faktorů.

Formulujeme nejprve nulovou hypotézu, že všechny *skupiny* pocházejí ze stejné základní populace (základního souboru), jinak řečeno, že hodnoty faktoru X_1 nemají na hodnoty znaku *Y* žádný efekt (vliv). Budeme tedy v nulové hypotéze předpokládat, že α_i pocházejí z *normálně rozdělené* populace s *nulovou střední hodnotou* a *konstantním rozptylem* σ^2, tedy formulujeme nulovou hypotézu

$$H_0 : E(\alpha_1) = E(\alpha_2) = ... = E(\alpha_k) = 0 ,$$

proti *alternativní hypotéze*, že H_0 neplatí, tudíž alespoň pro dvě hodnoty, např. *i* a *j*, platí:

$$H_1 : E(\alpha_i) \neq E(\alpha_j).$$

Cílem, k němuž směřujeme, je *přijmout* nulovou hypotézu H_0, eventuálně H_0 *zamítnout* (na zvolené *hladině významnosti*). Pro ověření nulové hypotézy H_0 použijeme statistiku:

$$F_1 = \frac{\dfrac{S_{y,m}}{k-1}}{\dfrac{S_{y,v}}{(k-1)(r-1)}}, \qquad (2.7)$$

která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F(k-1,(k-1)(r-1))$. Kritické hodnoty lze nalézt v tabulkách, nebo lze využít funkce z Excelu: FINV(α;*k* – 1;(*k* – 1)(*r* – 1)).

Dále formulujeme nulovou hypotézu, že všechny *bloky* pocházejí ze stejné základní populace (základního souboru), jinak řečeno, že hodnoty faktoru X_2 nemají na hodnoty znaku *Y* žádný efekt. Budeme tedy v nulové hypotéze předpokládat, že β_j pocházejí z *normálně rozdělené* populace s *nulovou střední hodnotou* a *konstantním rozptylem* σ^2, tedy formulujeme nulovou hypotézu

$$H_0' : E(\beta_1) = \ldots = E(\beta_r) = 0 ,$$

proti *alternativní hypotéze*, že H_0' neplatí, tudíž alespoň pro dvě hodnoty, např. $i' \neq i''$, platí

$$H_1' : E(\beta_{i'}) \neq E(\beta_{i''}) .$$

Pro ověření nulové hypotézy H_0' použijeme statistiku:

$$F_2 = \frac{\dfrac{S_{y,b}}{r-1}}{\dfrac{S_{y,v}}{(k-1)(r-1)}}, \qquad (2.8)$$

která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F(r-1,(k-1)(r-1))$.

Zásadní rozdíl mezi dvoufaktorovou a jednofaktorovou analýzou rozptylu spočívá v tom, že u jednofaktorové ANOVA *neuvažujeme* působení dalšího faktoru, zatímco u dvoufaktorové ANOVA tak činíme. Tento rozdíl je vyjádřen ve výpočtu testového kritéria

(2.7) a (2.8), kde se ve jmenovateli zlomku vyskytuje člen $(k-1)(r-1)$. Kdybychom na stejnou situaci aplikovali pouze jednofaktorovou ANOVA, pak by ve výpočtu hodnoty testového kritéria podle vztahu (1.6) byl na stejném místě člen $(n-k)$ nebo člen $(n-r)$, podle toho, zda bychom brali v úvahu skupiny nebo bloky. Tento rozdíl může zapříčinit rozdílné výsledky získané jednofaktorovou nebo dvoufaktorovou ANOVA!

ŘEŠENÝ PŘÍKLAD 2.1

Testovacími jízdami na zkušebním okruhu se zjišťuje průměrná spotřeba benzínu Natural 95 automobilu Octavia při použití benzínu od různých výrobců (Aral, Shell, Benzina, Slovnaft). Bylo vybráno 6 řidičů A, B, C, D, E, F, z nichž každý absolvoval s každým typem benzínu jednu zkušební jízdu. Na hladině významnosti 0,05 testujte, je-li průměrná spotřeba paliva závislá na typu použitého benzínu a na tom, který řidič s vozem jel.

	Řidiči					
Značka benzínu	**A**	**B**	**C**	**D**	**E**	**F**
Aral	7,5	6,9	7,9	7,3	6,9	7,8
Shell	7,6	7,2	7,5	8,0	7,3	8,2
Benzina	7,2	8,1	7,8	7,6	7,8	6,9
Slovnaft	7,0	7,3	7,2	7,5	8,2	7,7

Řešení:

Máte za úkol prozkoumat závislost průměrné spotřeby (znak Y) na typu použitého benzínu (znak X_1) a na řidiči (znak X_2), který s vozem jel.

Znak X_1 má $k = 4$ skupiny, znak X_2 má $r = 6$ bloků.

Pro faktor X_1 formulujeme nulovou hypotézu:

$$\mathrm{H}_0\colon E(\alpha_1)=E(\alpha_2)=E(\alpha_3)=E(\alpha_4), \qquad (2.9)$$

proti

H_1: neplatí (2.9), tj. průměrná spotřeba závisí na použitém druhu benzínu.

Pro faktor X_2 formulujeme nulovou hypotézu

$$\mathrm{H}_0'\colon E(\beta_1)=E(\beta_2)=\ldots=E(\beta_6), \qquad (2.10)$$

proti alternativní hypotéze H_1': neplatí (2.10), tj. průměrná spotřeba benzínu závisí na řidiči, který s vozem jel.

Pro ověření těchto hypotéz, tj. pro výpočet testových kritérií, musíme znát hodnotu součtů $S_{y,m}$, $S_{y,v}$ a S_y.

Nejdříve vypočítáme podmíněné průměry $\bar{y}_{i.}$, $i = 1, 2, 3, 4$, $\bar{y}_{.j}$, $j = 1, 2, \ldots, 6$ a také celkový průměr $\bar{y}$.

$$\bar{y}_{1.} = \frac{7{,}5+6{,}9+\ldots+7{,}8}{6} = 7{,}38,$$

další průměry $\bar{y}_{2.}, \bar{y}_{3.}, \bar{y}_{4.}$ vypočítáme analogicky, viz Tab. 2.2.

$$\bar{y}_{.1} = \frac{7{,}5+7{,}6+7{,}2+7}{4} = 7{,}33,$$

další průměry $\bar{y}_{.2}, \ldots, \bar{y}_{.6}$ vypočítáme analogicky. Celkový průměr je

$$\overline{y} = \frac{7{,}5 + 6{,}9 + \ldots + 7{,}7}{24} = 7{,}50\,.$$

Hodnoty všech průměrů jsou uvedeny v tabulce. Nyní lze přistoupit k výpočtu jednotlivých součtů:

$$S_{y,m} = r\sum_{i=1}^{4}(\overline{y}_{i.} - \overline{y})^2 = 6\cdot\left[(7{,}38 - 7{,}5)^2 + \ldots + (7{,}48 - 7{,}5)^2\right] = 0{,}21\,.$$

$$S_{y,b} = k\sum_{j=}^{6}\left(\overline{y}_{.j} - \overline{y}\right)^2 = 4\cdot\left[(7{,}33 - 7{,}5)^2 + \ldots + (7{,}38 - 7{,}5)^2\right] = 0{,}35\,.$$

Potřebujeme znát i hodnotu součtu $S_{y,v}$, z praktického hlediska je však výhodnější vypočítat hodnotu součtu S_y. Součet $S_{y,v}$ pak snadno dopočítáme, neboť $S_y = S_{y,m}+S_{y,v}+S_{y,b}$.

$$S_y = \sum_{i=1}^{4}\sum_{j=1}^{6}\left(y_{i,j} - \overline{y}\right)^2 = (7{,}5 - 7{,}5)^2 + (6{,}9 - 7{,}5)^2 + \ldots + (7{,}8 - 7{,}5)^2 +$$

$$+ (7{,}6 - 7{,}5)^2 + \ldots + (8{,}2 - 7{,}5)^2 + \ldots + (7{,}7 - 7{,}5)^2 = 3{,}79.$$

Potom vypočítáme

$$S_{y,v} = S_y - S_{y,m} - S_{y,b} = 3{,}79 - 0{,}21 - 0{,}36 = 3{,}22.$$

Pro ověření hypotézy H_0 určíme testové kritérium F_1

$$F_1 = \frac{\dfrac{S_{y,m}}{k-1}}{\dfrac{S_{y,v}}{(k-1)(r-1)}} = \frac{\dfrac{0{,}21}{3}}{\dfrac{3{,}22}{3\cdot 5}} = 0{,}32\,.$$

V tabulce kritických hodnot F-rozdělení nebo pomocí Excelu najdeme

$F_{0,05}(3,15)$ = FINV(0,05; 3,15) = 3,29.

Protože 0,32 < 3,29, přijímáme H_0, což znamená, že použitá značka benzínu nemá na průměrnou spotřebu vliv.

Pro ověření hypotézy H_0' určíme testové kritérium F_2

$$F_2 = \frac{\dfrac{S_{y,b}}{r-1}}{\dfrac{S_{y,v}}{(k-1)(r-1)}} = \frac{\dfrac{0{,}36}{5}}{\dfrac{3{,}22}{3\cdot 5}} = 0{,}33\,.$$

$F_{0,05}(5,15)$ = FINV(0,05; 5,15) = 2,9.

Protože 0,33 < 2,9, přijímáme i hypotézu H_0', tzn., že ani volba řidiče nemá na průměrnou spotřebu statisticky významný vliv.

Na rozdíl od jednofaktorové ANOVA jsme zde v obou situacích uvažovali současné působení *dvou faktorů*!

	Řidiči						
Zn. benzínu	**A**	**B**	**C**	**D**	**E**	**F**	**Průměry**
Aral	7,5	6,9	7,9	7,3	6,9	7,8	7,38
Shell	7,6	7,2	7,5	8,0	7,3	8,2	7,63
Benzina	7,2	8,1	7,8	7,6	7,8	6,9	7,57
Slovnaft	7,0	7,3	7,2	7,5	8,2	7,7	7,48
Průměry	7,33	7,38	7,6	7,6	7,55	7,65	7,50

Tab. 2.2. *Průměry*

Nakonec si ještě ukážeme řešení pomocí Excelu. Využijeme přitom funkci menu:
Nástroje → Analýza dat... → ANOVA: dva faktory bez opakování

Nejprve je zapotřebí připravit v Excelu data. Jednotlivé hodnoty y_{ij} pro faktoru Y pro hodnoty faktorů X_1 = benzín a X_2 = řidič uspořádáme do řádků a sloupců, podobně jako v tabulce v zadání. Data ve worksheetu Excelu vypadají tedy například takto:

	A	B	C	D	E	F	G	I
1	benzin/řidič	A	B	C	D	E	F	
2	Aral	7,5	6,9	7,9	7,3	6,9	7,8	
3	Shell	7,6	7,2	7,5	8	7,3	8,2	
4	Benzina	7,2	8,1	7,8	7,6	7,8	6,9	
5	Slovnaft	7	7,3	7,2	7,5	8,2	7,7	
6								

Dále otevřeme v hlavním menu postupně položky:
Data → Analýza dat... → ANOVA :dva faktory bez opakování

Po volbě třetí položky ANOVA: dva faktory bez opakování, se otevře zadávací okno kde postupně zadáte:

Vstupní oblast*:* \$A\$1:\$G\$5
Popisky v prvním sloupci – zakliknete
Alfa*:* 0,05 (hladina významnosti je předvolena, lze ji však změnit)
Výstupní oblast*:* \$L\$1 (levý horní roh výstupní oblasti)
Potvrdíte OK

Obdržíte následující výstup, kterého "levý horní roh" začíná v buňce L1 nadpisem ANOVA: dva faktory bez opakování:

Anova: dva faktory bez opakování

Faktor	*Počet*	*Součet*	*Průměr*	*Rozptyl*
Aral	6	44,3	7,383333	0,185667
Shell	6	45,8	7,633333	0,154667
Benzina	6	45,4	7,566667	0,194667
Slovnaft	6	44,9	7,483333	0,181667
A	4	29,3	7,325	0,075833
B	4	29,5	7,375	0,2625
C	4	30,4	7,6	0,1
D	4	30,4	7,6	0,086667
E	4	30,2	7,55	0,323333
F	4	30,6	7,65	0,296667

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Řádky	0,21	3	0,07	0,325581	0,806868	3,287383
Sloupce	0,358333	5	0,071667	0,333333	0,884913	2,901295
Chyba	3,225	15	0,215			
Celkem	3,793333	23				

V první tabulce jsou uvedeny základní statistické údaje o datech: Faktor, Počet, Součet, Průměr a Rozptyl.

Ve druhé tabulce nazvané ANOVA jsou uvedeny výpočty metodou ANOVA: dva faktory bez opakování, jednotlivé položky mají následující význam:

Řádky = meziskupinový
Sloupce = vnitroskupinový
Chyba = meziblokový
Celkem = celkový
SS = Součet čtverců (Sum of Squares)
Rozdíl = stupeň volnosti (DF – Degree of Freedom)
MS = Průměr čtverců (Mean Square)
F = testové kritérium
Hodnota P = Signifikance (p-hodnota)
F krit = kritická hodnota rozdělení F

Hodnoty získané řešením v Excelu jsou stejné jako při použití „ručního" výpočtu, proto i závěry jsou stejné. V Excelu máme navíc vypočtenu p-hodnotu testu (tzv. signifikanci), která, pokud je menší než zvolená hladina významnosti α, znamená, že nulovou hypotézu zamítáme. V opačném případě nulovou hypotézu přijímáme.

V předchozích úvahách jsme měli situaci právě jednoho výskytu všech kombinací hodnot skupin a bloku obou uvažovaných faktorů. Například každý řidič absolvoval jedinou jízdu s každým typem benzínu. Dále budeme uvažovat situaci vícenásobného opakování všech kombinací hodnot skupin a bloku obou uvažovaných faktorů. Například každý řidič absolvuje několik jízd (například 3 jízdy – viz následující příklad 2.2) s každým typem benzínu, přitom samozřejmě mohou být dosažené hodnoty průměrné spotřeby různé. Zda se tyto výsledky odlišují výrazně či nikoliv, se opět zjišťuje statistickým testem. Podrobnou analýzu situace, která je analogická analýze případu bez opakování, již zde uvádět nebudeme. Omezíme se pouze na řešení příkladu s využitím Excelu, konkrétně položky ANOVA: dva faktory s opakováním.

ŘEŠENÝ PŘÍKLAD 2.2

Podobně jako v příkladu 2.1 se zjišťuje průměrná spotřeba benzínu Natural 95 automobilu Octavia při použití benzínu od různých výrobců (Aral, Shell, Benzina, Slovnaft). Bylo vybráno 6 řidičů A, B, C, D, E, F, z nichž každý absolvoval s každým typem benzínu tři zkušební jízdy. Na hladině významnosti 0,05 testujte, je-li průměrná spotřeba paliva závislá na typu použitého benzínu a na řidiči. Údaje jsou uvedeny v následující tabulce.

benzin/řidič	Aral	Shell	Benzina	Slovnaft
A	7,5	7,6	7,2	7
	7,7	7,4	7,6	7,4
	8	7,3	8,1	7,7
B	6,9	7,2	8,1	7,3
	6,7	7,4	8,5	7,6
	6,6	7,6	8,8	7,8
C	7,9	7,5	7,8	7,2
	8	7,8	7,7	7,1
	8,3	8,1	7,6	7
D	7,3	8	7,6	7,5
	7,2	8	7,8	7,7
	7,1	7,9	8	7,8
E	6,9	7,3	7,8	8,2
	6,8	7,2	8	8,1
	6,7	7	8,1	8
F	7,8	8,2	6,9	7,7
	7,7	8,4	7,5	7,7
	7,5	8,5	7,9	7,7

Tab. 2.3. *Řidiči verus Benzíny s opakováním*

Řešení:

Data ve worksheetu Excelu vypadají přesně tak jako v Tab. 2.3, jsou umístěny např. v poli A1 až E19. Dále otevřeme v hlavním menu postupně položky:

Data → Analýza dat... → ANOVA: dva faktory s opakováním

Po volbě druhé položky ANOVA: dva faktory s opakováním, se otevře zadávací okno, kde postupně zadáte:

Vstupní oblast*:* \$A\$1:\$E\$19
Řádků na výběr: 3 (tj. počet opakování)
Alfa*:* 0,05 (hladina významnosti je předvolena, lze ji však změnit)
Výstupní oblast*:* např. \$L\$1 (levý horní roh výstupní oblasti)
Potvrdíte OK.

Obdržíte následující výstup, kterého "levý horní roh" začíná v buňce L1 nadpisem ANOVA: dva faktory s opakováním. V první tabulce jsou uvedeny základní statistické údaje o datech: Faktor, Počet, Součet, Průměr a Rozptyl.

Anova: dva faktory s opakováním

Faktor	Aral	Shell	Benzina	Slovnaft	Celkem
A					
Počet	3	3	3	3	12
Součet	23,2	22,3	22,9	22,1	90,5
Průměr	7,73	7,43	7,63	7,37	7,54
Rozptyl	0,06	0,02	0,20	0,12	0,10
B					
Počet	3	3	3	3	12
Součet	20,2	22,2	25,4	22,7	90,5
Průměr	6,73	7,40	8,47	7,57	7,54
Rozptyl	0,02	0,04	0,12	0,06	0,46
C					
Počet	3	3	3	3	12
Součet	24,2	23,4	23,1	21,3	92
Průměr	8,07	7,80	7,70	7,10	7,67
Rozptyl	0,04	0,09	0,01	0,01	0,16
D					
Počet	3	3	3	3	12
Součet	21,6	23,9	23,4	23	91,9
Průměr	7,200	7,967	7,800	7,667	7,658
Rozptyl	0,010	0,003	0,040	0,023	0,103
E					
Počet	3	3	3	3	12
Součet	20,4	21,5	23,9	24,3	90,1
Průměr	6,80	7,17	7,97	8,10	7,51
Rozptyl	0,01	0,02	0,02	0,01	0,33
F					
Počet	3	3	3	3	12
Součet	23	25,1	22,3	23,1	93,5
Průměr	7,67	8,37	7,43	7,70	7,79
Rozptyl	0,02	0,02	0,25	0,00	0,19
Celkem					
Počet	18	18	18	18	
Součet	132,6	138,4	141	136,5	
Průměr	7,37	7,69	7,83	7,58	
Rozptyl	0,28	0,20	0,19	0,13	

Ve druhé tabulce nazvané ANOVA jsou uvedeny výpočty metodou ANOVA: dva faktory s opakováním.

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Výběr	0,69	5	0,14	2,64	0,03	2,41
Sloupce	2,08	3	0,69	13,23	0,00	2,80
Interakce	10,23	15	0,68	12,99	0,00	1,88
Dohromady	2,52	48	0,05			
Celkem	15,53	71				

Jednotlivé položky mají následující význam:

Výběr = meziskupinový

Sloupce = vnitroskupinový

Interakce = meziblokový

Celkem = celkový

SS = Součet čtverců (Sum of Squares)

Rozdíl = stupeň volnosti (DF – Degree of Freedom)
MS = Průměr čtverců (Mean Square)
F = testové kritérium
Hodnota P = Signifikance (p-hodnota)
F krit = kritická hodnota rozdělení *F*

Hodnoty získané řešením v Excelu jsou analogické jako v příkladu 2.1, tedy v případě ANOVA bez opakování. Navíc je tu p-hodnota uvedená v řádku Interakce, která se týká testu vzájemné závislosti faktorů. Nulová hypotéza předpokládá, že faktoru jsou vzájemně nezávislé. Pokud je tato hodnota menší než zvolená hladina významnosti α, znamená to, že nulovou hypotézu zamítáme. V opačném případě nulovou hypotézu přijímáme.

V této kapitole jsme uvažovali situaci, kdy se kromě třídění do skupin vyskytují další faktory, říkáme jim bloky, podle nichž výsledky (tj. hodnoty znaku *Y*) rovněž třídíme. Přehledná situace vzniká, když kromě prvního faktoru uvažujeme ještě faktor druhý, říkáme pak, že je třídíme do bloků a v takovém případě se jedná o dvoufaktorovou ANOVA. Formálně vzato je ANOVA, ať jednofaktorová, dvoufaktorová nebo vícefaktorová, parametrickým testem statistické hypotézy, s nímž jste se seznámili v základním kurzu statistiky. Nejprve jsme měli situaci právě jednoho výskytu všech kombinací hodnot skupin a bloku obou uvažovaných faktorů. Například každý řidič absolvoval jedinou jízdu s každým typem benzínu. Poté jsme uvažovali situaci vícenásobného opakování všech kombinací hodnot skupin a bloku obou uvažovaných faktorů. Například každý řidič absolvuje několik jízd s každým typem benzínu, přitom samozřejmě mohou být dosažené hodnoty průměrné spotřeby různé. Zda se tyto výsledky odlišují výrazně či nikoliv, se opět zjistilo statistickým testem. K řešení příkladů jsme použili Excel, konkrétně položku Analýza dat.

2.3 SAMOSTATNÉ ÚKOLY

Řešte v Excelu.

2.1 Ve čtyřech městech okresu Karviná jsme v jednotlivých dnech sledovali průměrnou spotřebu pitné vody (v m^3) na jednoho obyvatele. Zjistěte, zda je průměrná spotřeba vody závislá na dni v týdnu, a je-li spotřeba v různých městech různá. Uvažujte hladinu významnosti 0,01. Zjištěné údaje jsou uvedeny v tabulce.

	Karviná	Orlová	Bohumín	Český Těšín
Po	0,64	0,75	0,54	0,76
Út	0,78	0,63	0,61	0,83
St	0,93	0,82	0,7	0,91
Čt	0,66	0,62	0,56	0,62
Pá	0,99	1,3	0,79	0,99
So	1,22	1,65	1,3	0,98
Ne	1,05	1,3	1,24	1,1

2.2 Výroba součástek může v podniku probíhat na jednom ze čtyř rozdílných strojů. I když každý stroj provádí stejné operace, má svá specifika. U každého stroje pracuje jeden dělník. Na hladině významnosti $\alpha = 0{,}01$ testujte hypotézu o tom, že počet vyrobených součástek není ovlivněn volbou stroje ani dělníkem, který na něm pracuje.

Dělník	**Stroj**			
	A	**B**	**C**	**D**
1	93	108	123	133
2	98	153	143	163
3	80	123	150	168
4	88	158	165	145
5	60	143	140	130

2.4 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

2.1 DNY: $F = 12{,}95$ F krit $= 4{,}01$ p-hodnota $= 0{,}000$ – H_0 zamítáme (průměrná spotřeba pitné vody závisí na dnu v týdnu)
MĚSTO: $F = 2{,}07$ F krit $= 5{,}1$ p-hodnota $= 0{,}14$ – H_0 přijímáme (nebyla prokázána závislost průměrné spotřeby pitné vody na městě).

2.2 DĚLNÍK: $F = 2{,}45$ F krit $= 5{,}41$ p-hodnota $= 0{,}1$ – H_0 přijímáme (nebyla prokázána závislost počtu vyrobených součástek na dělníkovi, který na stroji pracuje).
STROJ: $F = 20{,}47$ F krit $= 5{,}95$ p-hodnota $= 0{,}000$ – H_0 zamítáme (počet vyrobených součástek závisí na stroji).

3 REGRESNÍ ANALÝZA – JEDNOROZMĚRNÁ LINEÁRNÍ REGRESE

RYCHLÝ NÁHLED KAPITOLY

Analýzu rozptylu z první kapitoly je možné chápat jako analýzu závislosti kvantitativního znaku (proměnné) na kvalitativním znaku - faktoru (proměnné). Naproti tomu závislostí kvantitativního znaku na kvantitativním znaku (nebo více kvantitativních znacích) se zabývá *regresní analýza*. V případě závislosti dvou znaků mluvíme o *jednorozměrné regresi* (případně *jednoduché regresi*), u znaku závislém na více kvantitativních veličinách hovoříme o *vícerozměrné regresi* (*vícenásobné regresi*). V této kapitole budeme vyšetřovat nejprve nejjednodušší *lineární* závislost dvou znaků, v další kapitole se budeme zabývat i nelineárními závislostmi dvou znaků důležitých z hlediska ekonomických aplikací. V následujících kapitolách pak budeme vyšetřovat závislosti více než dvou statistických znaků. Nejprve si ozřejmíte základní rozdíl mezi ANOVA a regresní analýzou, formulujete model jednoduché lineární regresní analýzy, definujete a ozřejmíte pojem regresní přímky a regresních koeficientů. Poté bude vysvětlena metoda nejmenších čtverců k nalezení „nejlepších" hodnot regresních koeficientů v regresním modelu. Míra přiléhavosti dat k regresní křivce bude stanovena pomocí koeficientu determinace a jeho odmocniny – koeficientu korelace. Nakonec se seznámíte s tzv. klasickým jednoduchým regresním modelem, který stanovuje 3 základní podmínky, kterým by měl vyhovovat regresní model vzhledem k existujícím datům. Vše bude demonstrováno na příkladech, které budou řešeny mimo jiné pomocí funkcí Excelu.

3.1 REGRESNÍ ANALÝZA

Analýzu rozptylu z první kapitoly je možné chápat jako analýzu závislosti kvantitativního znaku (proměnné) na kvalitativním znaku - faktoru (proměnné). Naproti tomu závislostí kvantitativního znaku na kvantitativním znaku (nebo více kvantitativních znacích) se zabývá *regresní analýza*. V případě závislosti dvou znaků mluvíme o *jednorozměrné regresi* (případně *jednoduché regresi*), u znaku závislém na více kvantitativních veličinách hovoříme o *vícerozměrné regresi* (*vícenásobné regresi*). V této kapitole budeme vyšetřovat nejprve nejjednodušší *lineární* závislost dvou znaků, v další kapitole se budeme zabývat i nelineárními závislostmi dvou znaků důležitých z hlediska ekonomických aplikací. V následujících kapitolách pak budeme vyšetřovat závislosti více než dvou statistických znaků.

V regresní analýze studujeme vztah mezi jedinou proměnnou (hodnotami statistického znaku) nazývanou *závisle proměnnou* (někdy *vysvětlovanou proměnnou*), označujeme ji Y, a obecně několika proměnnými (hodnotami statistických znaků), které nazýváme *nezávisle proměnné* (někdy *vysvětlující proměnné*), a označujeme je symboly X_1, X_2,.... Pokud se zabýváme jedinou nezávisle proměnnou X, hovoříme o *jednoduché regresi*, pokud je nezávisle proměnných více než jedna, mluvíme o *vícrozněrnéné (vícenásobné) regresi* (někdy též mnohonásobné regresi). V této a následující kapitole se věnujeme jednoduché regresi.

Závisí-li veličina Y na veličině X, pak to matematicky vyjadřujeme zápisem

$$Y = f(X) \, , \tag{3.1}$$

což je *funkční vztah*, známý mimo jiné z fyziky (například Newtonův gravitační zákon: Y je přitažlivá síla, X je vzdálenost hmotných bodů).

V našem případě jsou Y a X *statistické znaky* (náhodné veličiny), pak hovoříme o *statistické závislosti*, funkční vztah (3.1) přejde v *regresní vztah* (*regresní model*)

$$y = f(x) + \varepsilon, \tag{3.2}$$

kde y, resp. x, představují hodnoty znaku Y, resp. X, ε je *náhodná složka,* funkci f nazýváme *regresní funkce*.

Jestliže je regresní funkce f lineární, což značí, že má tvar regresní přímky

$$f(x) = \beta_0 + \beta_1 x, \tag{3.3}$$

potom hovoříme o *jednoduché lineární regresi*, nemá-li regresní funkce lineární tvar, hovoříme o *jednoduché nelineární regresi.* Ve vzorci (3.3) jsou β_0, β_1 *parametry* regresní funkce neboli *regresní koeficienty*.

Mezi nejpoužívanější nelineární regresní funkce patří:

regresní parabola:	$f(x) = \beta_0 + \beta_1 x^2$,	(3.4)
regresní hyperbola:	$f(x) = \beta_0 + \beta_1 \frac{1}{x}$,	(3.5)
regresní logaritmická funkce:	$f(x) = \beta_0 + \beta_1 \log x$.	(3.6)
regresní mocninná funkce:	$f(x) = \beta_0 x^{\beta_1}$,	(3.7)
regresní exponenciální funkce:	$f(x) = \beta_0 \beta_1^x$.	(3.8)

Výše uvedené nelineární regresní funkce lze převést na lineární vhodnou transformaci, jak uvidíme v následující kapitole.

Kromě výše uvedených příkladů nelineárních regresních funkcí existuje celá řada dalších významných nelineárních funkcí, např. Törnquistovy funkce, které nelze na lineární funkci jednoduše převést. Budeme se jimi zabývat v následující kapitole.

3.2 JEDNODUCHÁ REGRESNÍ ANALÝZA

Představte si výběr párových hodnot (y_1, x_1), (y_2, x_2), (y_3, x_3),..., (y_n, x_n), získaných (např. změřených) na statistických jednotkách základního souboru. Zde jsou y_i hodnotami závisle proměnné Y a x_i jsou hodnotami nezávisle proměnné X. Zmíněné párové hodnoty můžeme získat zejména dvojím způsobem:

(A) Hodnoty nezávisle proměnné x_i jsme předem pevně zvolili a k nim jsme „změřili“ příslušné hodnoty y_i. V této situaci jsou hodnoty znaku X pevné (nenáhodné), zatímco hodnoty znaku Y považujeme za náhodné veličiny.

(B) Párové hodnoty (y_i, x_i) „změříme“ na n náhodně zvolených jednotkách základního souboru. V této situaci jak hodnoty znaku X, tak hodnoty znaku Y považujeme za náhodné veličiny.

Výše uvedený datový soubor párových hodnot můžeme geometricky znázornit v rovině *bodovým grafem*, kde na vodorovnou osu „x“ nanášíme hodnoty nezávisle proměnné a na svislou osu „y“ příslušné hodnoty závisle proměnné. Výsledkem je geometrické znázornění n bodů v rovině, z jejichž vzájemné polohy můžeme soudit na regresní závislost znaku Y na X. Úkolem jednoduché lineární regrese je „proložit“ danými body přímku (tj. nalézt lineární regresní funkci), která nejlépe charakterizuje polohu daných n bodů. Z předchozího odstavce víme, že tato regresní funkce má tvar $f(x) = \beta_0 + \beta_1 x$, kde β_0, β_1 jsou zatím neznámé hodnoty parametrů regresní přímky. Regresní model (3.2) má nyní tvar

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, ..., n. \tag{3.9}$$

Odhady b_0, b_1 těchto neznámých parametrů – *regresní koeficienty* získáme *metodou nejmenších čtverců*. Této metodě, která patří mezi nejdůležitější metody používané ve statistice, bude věnován následující odstavec.

3.3 METODA NEJMENŠÍCH ČTVERCŮ

Uvažujte data ve formě párových hodnot – bodů: (y_1, x_1), (y_2, x_2), (y_3, x_3),..., (y_n, x_n). Úkolem jednoduché regrese je najít regresní funkci, která „nejlépe charakterizuje polohu“ daných n bodů. Nejprve budeme uvažovat obecný tvar regresní funkce $f(x; \beta_0, \beta_1)$ se dvěma parametry β_0, β_1 (nemusí to být nutně regresní přímka). Speciálními případy této regresní funkce je lineární funkce (3.3) a také nelineární funkce (3.4) – (3.8). Postup metody nejmenších čtverců bude vždy stejný, tj. nezávislý na konkrétním tvaru regresní funkce. Odhady b_0, b_1 neznámých parametrů β_0, β_1 získáme tak, že nalezneme hodnoty b_0, b_1, pro něž nabývá své minimální hodnoty *reziduální součet čtverců* odchylek hodnot závisle proměnné y_i od teoretické hodnoty $Y_i = f(x_i; b_0, b_1)$, tj.

$$S_R = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i, b_0, b_1))^2 . \tag{3.10}$$

Jak je známo z matematické analýzy, své minimum funkce S_R (zde je to funkce proměnných b_0, b_1) vždy nabývá pro ty hodnoty b_0, b_1, pro něž se anulují její parciální derivace:

$$\frac{\partial S_R}{\partial b_0} = 0, \qquad \frac{\partial S_R}{\partial b_1} = 0. \tag{3.11}$$

Vztahy (3.11) představují soustavu 2 rovnic o 2 neznámých b_0, b_1, která se nazývá *soustava normálních rovnic.* Jejím řešením získáme hledané odhady regresních parametrů zvolené regresní funkce.

Vyřešíme nyní soustavu (3.11) pro speciální případ, který nás zejména zajímá, totiž pro lineární regresní funkci $f(x;\beta_0,\beta_1) = \beta_0 + \beta_1 x$. Dosadíme-li tuto funkci do vztahu (3.10), vypočteme příslušné parciální derivace, které položíme rovny 0, získáme konkrétní soustavu normálních rovnic

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i , \tag{3.12}$$
$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 .$$

Z těchto rovnic již snadno (v konkrétním případě pro dané hodnoty y_i , x_i známou „dosazovací metodou“) vypočteme hledané odhady b_0, b_1 takto:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad , \quad b_0 = \bar{y} - b_1 \bar{x} . \tag{3.13}$$

Z analytické geometrie si připomeňte, že regresní koeficient b_0 představuje průsečík regresní přímky s osou „*y*“, tedy hodnotu Y_0 pro $x = 0$, tento regresní koeficient se někdy nazývá *úrovňová konstanta*. Regresní koeficient b_1 vyjadřuje směrnici přímky, tedy sklon přímky k ose „*x*“, tj. změnu funkční hodnoty *Y* při změně nezávisle proměnné *x* o jednotku.

Pro jiné než lineární tvary regresní funkce je postup metody nejmenších čtverců obdobný. Výsledkem je rovněž soustava 2 normálních rovnic, tyto rovnice však již nemusí být lineární a proto soustavu již obvykle nelze snadno vyřešit. K řešení pak používáme *iterační numerické metody*, které zde nejsou předmětem našeho zájmu. V části Řešené příklady uvedeme způsob nalezení odhadů regresních koeficientů metodou linearizace exponenciální a mocninné regresní funkce pomocí logaritmické transformace.

Na tomto místě bychom chtěli zvýraznit jeden důležitý fakt, který budeme v následujícím výkladu neustále využívat. Data pro regresní analýzu jsou výsledkem náhodného výběru, ať již jsme použili při jejich získání postup (A), nebo (B). Proto také výsledek jednoduché lineární regresní analýzy – odhady neznámých parametrů β_0, β_1, tj. regresní koeficienty b_0, b_1, budou náhodné veličiny. Při každém dalším náhodném výběru dat bude výsledek, tj. odhad b_0, b_1, obecně jiný! Má proto význam hovořit dále o statistických charakteristikách těchto odhadnutých parametrů, jako např. střední hodnota, rozptyl, apod.

3.4 MÍRA VARIABILITY, KOEFICIENT DETERMINACE

Metoda nejmenších čtverců nás nyní přivedla k postupu, který jsme již použili v předchozí kapitole při analýze rozptylu. V ANOVA se jednalo o rozklad celkové variability znaku *Y*, vyjádřené jako celkový součet čtverců, na meziskupinový a vnitroskupinový (reziduální) součet čtverců. V analýze rozptylu jsme pracovali se znakem *X*, který měl kvalitativní povahu, a proto nebylo možné vyjádřit závislost regresním modelem. V regresní analýze má znak *X* – nezávisle proměnná – kvantitativní povahu, a proto je regresní model závislosti Y na X možný. Použijeme analogii s ANOVA v tom, že znak X zde bude nabývat hodnot x_1, x_2,...,x_n a i-tá skupina bude nyní charakterizována teoretickou hodnotou $Y_i = f(x_i; b_0, b_1)$, namísto skupinového průměru $\overline{y}_i$ v ANOVA. Potom celkovou variabilitu vysvětlované proměnné charakterizuje *celkový součet čtverců:*

$$S_y = \sum_{i=1}^{n} (y_i - \overline{y})^2 . \tag{3.14}$$

Část celkové variability vysvětlenou regresním modelem charakterizuje *teoretický součet čtverců:*

$$S_T = \sum_{i=1}^{n} (Y_i - \overline{y})^2 , \tag{3.15}$$

nevysvětlenou část celkové variability představuje *reziduální součet čtverců* (3.10):

$$S_R = \sum_{i=1}^{n} (y_i - Y_i)^2 , \tag{3.16}$$

kde $e_i = y_i - Y_i$ nazýváme *reziduum*.

Lze dokázat, že mezi jednotlivými součty čtverců platí základní vztah:

$$S_y = S_T + S_R . \tag{3.17}$$

Obdobně jako v analýze rozptylu jsme zavedli k vyjádření těsnosti vztahu Y a X poměr determinace, nyní zavedeme analogický pojem charakterizující přiléhavost dat k regresnímu modelu. Tímto pojmem je *koeficient determinace*, který definujeme vztahem

$$R^2 = 1 - \frac{S_R}{S_y} . \tag{3.18}$$

Ze vztahu (3.17) vyplývá, že koeficient determinace nabývá hodnoty z intervalu [0,1] a určuje tu část celkové variability pozorovaných hodnot S_y, kterou lze vysvětlit daným regresním modelem. Jinak řečeno, po vynásobení koeficientu determinace hodnotou 100 obdržíme, kolik procent celkové variability je vysvětlitelných regresním modelem. Koeficient determinace je proto důležitou charakteristikou vhodnosti zvoleného regresního modelu.

Vztah (3.18) vzniká podílem náhodných veličin, a proto jakožto náhodná veličina je odhadem koeficientu determinace R^2. Pro malé rozsahy výběru n je odhad (3.18) *vychýlený*, viz Ramík (2003), tj. nadhodnocuje přiléhavost k regresnímu modelu. Proto se používá *nevychýlený odhad* koeficientu determinace R^2_{adj} (z angl. *adj*usted), který *nazýváme korigovaný (upravený) koeficient determinace*:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-2} . \tag{3.19}$$

Pro velké hodnoty n je však zlomek ve vzorci (3.19) blízký k jedné a korigovaný koeficient se blíží k „nekorigovanému“.

3.5 KLASICKÝ LINEÁRNÍ MODEL

Klasickým jednoduchým lineárním regresním modelem se nazývá regresní model (3.9):

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1,2,...,n$,

splňující následující podmínky:

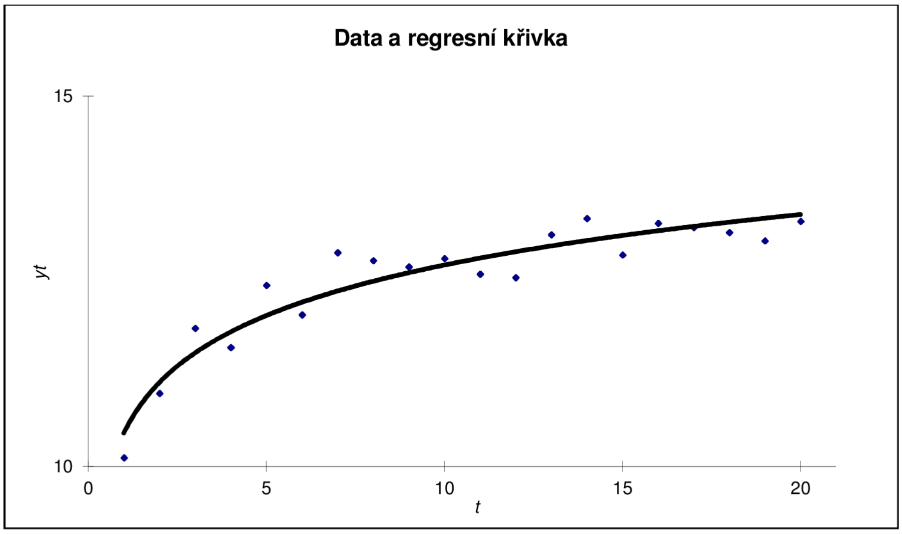
(1) Hodnoty vysvětlující proměnné x_i se volí předem, viz (A) odstavec 3.2, nejsou to tedy náhodné veličiny.

(2) Náhodné složky ε_i v modelu (3.9) mají *normální rozdělení* pravděpodobnosti se střední hodnotou 0 a (neznámým) rozptylem σ^2.
Konstantnost rozptylu nazýváme *homoskedasticita*.

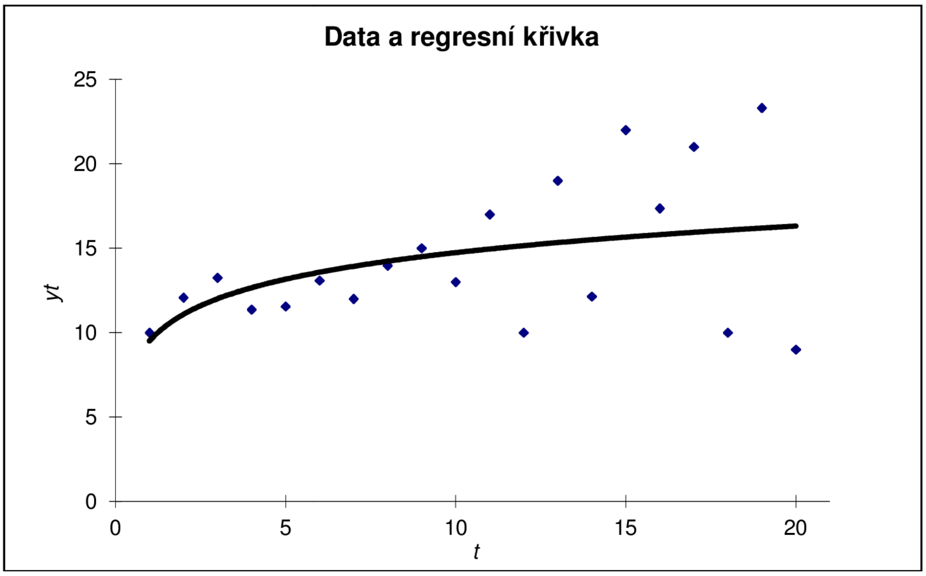
(3) Náhodné složky jsou *nekorelované*, tj.
$Cov(\varepsilon_i , \varepsilon_j) = 0$ pro každé $i \neq j$, $i,j = 1,2,...,n$. (*Cov* značí kovarianci, viz Ramík (2003))

Podmínky (1) až (3) požadujeme tehdy, chceme-li zajistit splnění některých dalších vlastností: např. zjistit intervaly spolehlivosti koeficientů regresní funkce, interval spolehlivosti hodnoty regresní funkce, eventuálně chceme-li provádět testy hypotéz o některých prvcích regresního modelu. Těmito tématy se budeme zabývat v následujících odstavcích. Pokud totiž tyto podmínky splněny nejsou, nelze zajistit „spolehlivé předpovědi“.

V praxi jsou podmínky klasického modelu často splněny, nejsme-li si však jejich platností jisti, můžeme provést testy hypotéz jak o normalitě rozdělení náhodné složky (např. test dobré shody, viz např. Ramík (2003)), tak i testy o nekorelovanosti náhodných složek (např. *t*-test). Další testy uvedeme později v souvislosti s časovými řadami. Na Obr. 3.1 je znázorněna situace, kdy podmínky klasického lineárního modelu jsou splněny, na Obr. 3.2 je zachycena situace, kdy není splněna ani podmínka normality náhodných složek (na obrázku jsou všechny ε_i prakticky stejné), ani podmínka nekorelovanosti (hodnoty y_i se nacházejí vedle sebe po jedné straně grafu regresní funkce).



Obr. 3.1. *Podmínky klasického modelu jsou splněny*



Obr. 3.2 *Podmínky klasického modelu nejsou splněny*

ŘEŠENÝ PŘÍKLAD 3.1

Společnost na výrobu bytového textilu zkoumala, jak souvisí zisk z prodeje s výdaji na reklamu. Tab. 3.1 uvádí údaje obdržené v deseti náhodně vybraných firmách.

a. Načrtněte bodový graf a určete typ regresní funkce popisující danou závislost.
b. Stanovte koeficienty regresní funkce z a.
c. Vypočítejte koeficient determinace a zhodnoťte těsnost závislosti vyjádřenou regresním modelem z bodu **b**.

Pozorování	Výdaje na reklamu (tis. Kč)	Zisk z prodeje (10 tis. Kč)
1	6	5
2	8	8
3	9	9
4	9	12
5	12	21
6	15	25
7	16	32
8	20	36
9	22	51
10	23	59

Tab. 3.1. *Výdaje na reklamu*

Řešení („ruční“ výpočet):

a. Zkoumá se závislost zisku z prodeje na výdajích na reklamu, proto sestrojíte bodový graf tak, že na osu x nanesete výdaje, na osu y zisk.

Z grafu vidíte, že jde o přímou závislost, kterou je možné popsat regresní přímkou
$Y = \beta_0 + \beta_1 x$.

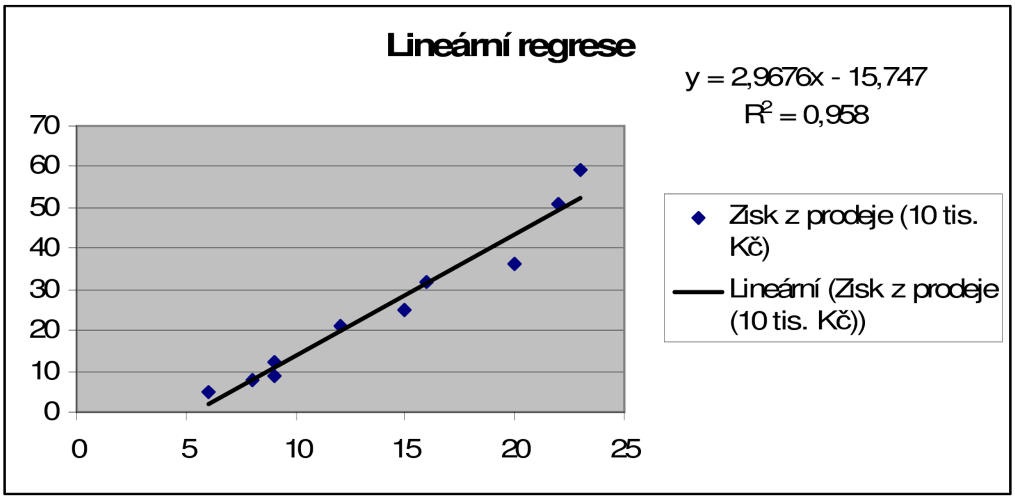
b. Máte za úkol stanovit hodnoty koeficientů b_0, b_1, neboli na základě dat z tabulky odhadnout hodnoty parametrů β_1, β_2. Využijeme výsledků metody nejmenších čtverců, nebudete však dosazovat přímo do soustavy rovnic (3.12), ale použijete vztahy pro b_0, b_1, tj. (3.13), které je možné z dané soustavy vyjádřit, a to v numericky výhodném a snadno zapamatovatelném tvaru:

$$b_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{462{,}1 - 14 \cdot 25{,}8}{230 - 14^2} = \frac{100{,}9}{34} = 2{,}97$$

$$b_0 = \overline{y} - b_1 \overline{x} = 25{,}8 - 2{,}97 \cdot 14 = -15{,}75\,.$$

Výpočty potřebných hodnot pomocí kalkulačky jsou uvedeny v následující tabulce.

i	x_i	y_i	x_i^2	$x_i y_i$	Y_i	$(Y_i - \overline{y})^2$	$(y_i - \overline{y})^2$
1	6	5	36	0	2,04	565,21	432,64
2	8	8	64	64	7,98	318,22	316,84
3	9	9	81	81	10,95	221,15	282,24
4	9	12	81	108	10,95	221,15	190,44
5	12	21	144	252	19,86	35,62	23,04
6	15	25	225	375	28,77	8,61	0,64
7	16	32	256	512	31,74	34,84	38,44
8	20	36	400	720	43,62	315,88	104,04
9	22	51	484	1122	49,56	562,08	635,04
10	23	59	529	1357	52,53	711,60	1102,24
Součet	140	258	2300	4621	258	2994,3	3125,6
Průměr	14	25,8	230	462,1			



Obr. 3.3. *Graf regresní přímky*

Hledaná regresní přímka má tvar:

$$Y = -15{,}75 + 2{,}97x\,.$$

c. K tomu, abychom vypočítali determinační koeficient, musíme znát hodnotu součtu S_T a součtu S_y. Tyto součty vypočítáme podle vztahů (3.14), (3.15). Pro výpočet teoretického součtu musíme pro každé x_i, $i = 1,\dots,10$, znát teoretickou hodnotu Y_i, $i = 1,\dots,10$

$$Y_1 = -15{,}75 + 2{,}97 \cdot x_1 = -15{,}78 + 2{,}97 \cdot 6 = 2{,}04.$$

Tato hodnota udává, jaký by měl být zisk při výdajích $x = 6$. Protože však jde o stochastickou závislost mezi společenskými veličinami, může se tato hodnota lišit od skutečně zjištěné hodnoty $y = 5$. Všechny teoretické hodnoty Y_i i hodnoty součtů S_y a S_T jsou uvedeny v tabulce. Koeficient determinace vypočítáme dosazením součtů S_y, S_T do vztahu (3.18).

$$R^2 = \frac{S_T}{S_y} = \frac{2994{,}3}{3125{,}6} = 0{,}958.$$

Tato hodnota znamená, že pomocí regresní přímky $Y = -15{,}78 + 2{,}97x$ je vysvětleno 95,8% chování proměnné Y.

Nakonec ještě ukážeme řešení pomocí Excelu. Využijeme přitom graf funkce s funkcí Přidat spojnici trendu. V dalším řešeném příkladu si pak ukážeme ještě další možnost řešení úlohy jednoduché (i vícenásobné) regrese s využitím menu:

Data → Analýza dat... → Regrese.

Data jsou uspořádána ve worksheetu ve 2 sloupcích:

	A	B
1	**Výdaje na reklamu (tis. Kč)**	**Zisk z prodeje (10 tis. Kč)**
2	6	5
3	8	8
4	9	9
5	9	12
6	12	21
7	15	25
8	16	32
9	20	36
10	22	51
11	23	59
12		

Po volbě položky Vložit graf → XY bodový..., se otevře zadávací okno, kde zadáte:

Oblast dat: \$A\$1:\$B\$11

Sloupce: √ (zakliknout)

Potvrdíte OK

Obdržíte bodový graf, viz Obr. 3.3. (ještě bez regresní přímky). Poklepem pravým tlačítkem myši na některý z bodů grafu obdržíte nabídku menu, kde zvolíte: Přidat spojnici trendu

Typ trendu regrese: zvolíte Lineární

Dále otevřete záložku *Možnosti*, kde zakliknete:

Zobrazit rovnici regrese (rovnice regresní přímky) a

Zobrazit hodnotu spolehlivosti R (hodnotu koeficientu determinace R^2).

Potvrdíte OK.

Obdržíte výsledek téměř takový, jaký je na Obr. 3.3. K původním bodům se zobrazí regresní přímka, dále rovnice regresní přímky a hodnotu koeficientu determinace R^2.

ŘEŠENÝ PŘÍKLAD 3.2

Společnost Air - Ostrava, zajišťující lety na trase Ostrava - Praha, sleduje při plánování letů také na hmotnost užitečného zatížení letadla, jehož významnou část tvoří pasažéři a jejich zavazadla. Zjistilo se, že hmotnost zavazadel cestujících souvisí s dobou, na kterou odcestovali.

a. Najděte rovnici regresní přímky popisující danou závislost.

b. S jakou hmotností zavazadel lze počítat, bude-li na palubě 15 cestujících vracejících se za 2 dny, 7 cestujících vracejících se za 5 dnů, 5 cestujících vracejících se za 6 dnů a 1 cestující vracející se za 14 dní.

Výsledky průzkumu jsou zaznamenány v tabulce.

Pozorování	Dny	Hmotnost
1	13	46
2	12	43
3	9	29
4	16	52
5	10	31
6	5	18
7	2	11
8	3	12
9	8	25
10	2	10
11	14	48
12	19	60
13	3	15
14	5	20
15	2	12

Řešení:

Prezentujeme zde pouze „ruční“ výpočet řešení (s kalkulačkou), řešení pomocí Excelu s využitím funkce Přidat spojnici trendu v bodovém grafu ponecháváme na čtenáři.

a. K výpočtu regresních koeficientů b_0, b_1 použijeme opět vztahů (3.13):

$$b_1 = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{324{,}4 - 8{,}2 \cdot 28{,}8}{96{,}73 - 8{,}2^2} = 2{,}99\,, \qquad b_0 = \overline{y} - b_1\overline{x} = 28{,}8 - 2{,}99 \cdot 8{,}2 = 4{,}27\,.$$

Regresní přímka má tedy tvar

$$Y = 4{,}27 + 2{,}99x\,.$$

i	x_i	y_i	$x_i y_i$	x_i^2
1	13	46	598	169
2	12	43	516	144
3	9	29	261	81
4	16	52	832	256
5	10	31	310	100
6	5	18	90	25
7	2	11	22	4

8	3	12	36	9
9	8	25	200	64
10	2	10	20	4
11	14	48	672	196
12	19	60	1140	361
13	3	15	45	9
14	5	20	100	25
15	2	12	24	4
Součet	123	432	4866	1451
Průměr	8,2	28,8	324,4	96,73

b. Vypočítáme hodnotu Y pro $x = 2$: $Y(2) = 4{,}27 + 2{,}99 \cdot 2 = 10{,}25$,

$x = 5$: $Y(5) = 4{,}27 + 2{,}99 \cdot 5 = 19{,}22$,

$x = 6$: $Y(6) = 4{,}27 + 2{,}99 \cdot 6 = 22{,}21$,

$x = 14$: $Y(14) = 4{,}27 + 2{,}99 \cdot 14 = 46{,}13$.

Potom hmotnost zavazadel m, se kterou lze počítat, snadno zjistíte, uvážíte-li počty příslušných cestujících:

$$m = 15 \cdot Y(2) + 7 \cdot Y(5) + 5 \cdot Y(6) + 1 \cdot Y(14) = 153{,}75 + 134{,}54 + 111{,}05 + 46{,}13 = 445{,}47.$$

3.6 Samostatné úkoly

3.1 Personální ředitel firmy shromáždil údaje o věku (X) a době pracovní neschopnosti (Y) dvaceti náhodně vybraných stálých zaměstnanců. Zjištěné údaje jsou zaznamenány v tabulce.

X	*Y*	*X*	*Y*
20	4	58	20
35	14	46	13
35	15	43	16
34	10	33	10
32	10	29	10
28	9	36	11
25	12	48	14
46	15	55	15
38	15	36	14
50	16	19	6

Načrtněte bodový graf a najděte rovnici regresní funkce vyjadřující danou závislost. Zhodnoťte výstižnost (přiléhavost) regresní funkce vzhledem k datům.

3.2 Bylo sledováno, jak souvisí množství vadných výrobků (v % z vyrobených výrobků) s výkonem soustružníka (v % z předepsané normy). Bylo vybráno deset pracovníků, naměřené údaje jsou uvedeny v tabulce.

Výkon	56	68	72	85	92	102	107	111	123	142
Vadné výrobky	5,2	3,9	3,5	2,4	2,04	2	2,2	2,24	2,4	2,51

Stanovte regresní model a určete přiléhavost regresní přímky k datům.

3.3 Tabulka zachycuje stáří (v letech) osmi vybraných strojů v potravinářském závodě a týdenní náklady (v Kč) na provoz těchto strojů.

Stáří stroje	1	2	3	4	5	6	7	8
Náklady	44	52	61	80	94	108	111	116

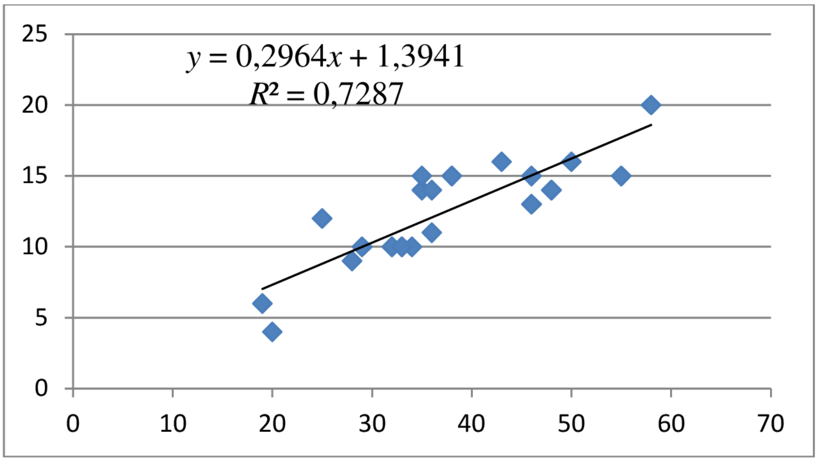
a. Odhadněte parametry lineární regresní funkce, která by měla vystihovat průběh závislosti nákladů na stáří.

b. Určete koeficient determinace R^2 a interpretujte jej.

c. Jaké týdenní náklady můžeme očekávat u stroje starého 4 roky?

3.7 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

3.1



3.2 $Y = -0{,}0285x + 5{,}56; \quad R^2 = 0{,}53$.

3.3 a) $Y = 32{,}14 + 11{,}36x$

b) $R^2 = 0{,}97$ tzn. modelem je vysvětleno 97% celkové variability.

c) $Y(4) = 32{,}14 + 11{,}36.4 = 77{,}58\ Kč$.

4 REGRESNÍ ANALÝZA – JEDNOROZMĚRNÁ: INTERVALY SPOLEHLIVOSTI, TESTY HYPOTÉZ, NELINEÁRNÍ REGRESE

RYCHLÝ NÁHLED KAPITOLY

Tato kapitola vám rozšíří znalosti v jednorozměrné regresní analýze. Za předpokladů jednorozměrného klasického regresního modelu se budete zabývat stanovením intervalů spolehlivosti a dále testy hypotéz regresních koeficientů a testem nulovosti koeficientu determinace. Další odstavce se zabývají jednorozměrnou nelineární regresí. Nejprve budou vyšetřovány regresní funkce, které lze s pomocí vhodné transformace převést na funkce lineární dále parabolická regresní funkce a nakonec nelineární regresní funkce tzv. Tornquiustova typu. Pro výpočet parametrů těchto funkcí se používá metoda vybraných bodů s pomocí Excelu. V této kapitole si rozšíříte znalosti v jednorozměrné regresní analýze. V návaznosti na jednorozměrný klasický regresní model se budete zabývat intervaly spolehlivosti a testy hypotéz regresních koeficientů a také testem koeficientu determinace. Poté se začnete zabývat jednorozměrnou nelineární regresí. Nejprve budou vyšetřovány ty regresní funkce, které lze s pomocí vhodné transformace převést na funkce lineární. Následuje parabolická regresní funkce a nakonec nelineární funkce tzv. Törnquiustova typu. Pro výpočet parametrů těchto funkcí, jež mají uplatnění především v marketingu, poznáte novu metodu tzv. metodu vybraných bodů, která zde nahradí známou metodu nejmenších čtverců s využitím Excelu.

4.1 INTERVALY SPOLEHLIVOSTI

Jsou-li splněny předpoklady klasického lineárního modelu (3.9), tj. modelu

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,2,\dots,n,$$

potom pro rozdělení odhadů regresních koeficientů b_0, b_1 jakožto náhodných veličin platí toto: Regresní koeficient b_j má normální rozdělení pravděpodobnosti se střední hodnotou β_j a rozptylem $\sigma^2 h_j$, kde $j = 0$ nebo 1, čísla h_j jsou definována následujícími vztahy:

$$h_0 = \frac{\sum x_i^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}, \tag{4.1}$$

$$h_1 = \frac{n}{n\sum x_i^2 - \left(\sum x_i\right)^2}. \tag{4.2}$$

V klasickém lineárním modelu předpokládáme, že náhodné složky mají konstantní rozptyl σ^2, jeho hodnotu však neznáme. Neznámý rozptyl σ^2 můžeme nahradit jeho bodovým odhadem

$$s_R^2 = \frac{S_R}{n-2}, \tag{4.3}$$

který nazýváme *reziduální rozptyl*. Jak je vidět, v reziduálním rozptylu vystupuje v čitateli reziduální součet čtverců (3.16) dělený číslem $n - 2$, což je *počet stupňů volnosti*, tj. rozsah dat n mínus počet regresních parametrů v modelu: 2. Odmocninu reziduálního rozptylu s_R nazýváme *směrodatná chyba*.

Oboustranný interval spolehlivosti pro regresní koeficient b_j, při zadaném koeficientu spolehlivosti $(1 - \alpha)$, je následující interval:

$$[b_j - t_{1-\alpha/2}(n-2)\ s_R\sqrt{h_j}\,,\ b_j + t_{1-\alpha/2}(n-2)\ s_R\sqrt{h_j}\ ],\ j = 0 \text{ nebo } 1. \tag{4.4}$$

Připomínáme, že zde $t_{1-\alpha/2}(n-2)$ je příslušný *kvantil* Studentova *t*-rozdělení, podrobnosti, viz Ramík (2003), h_j jsou dány vztahy (4.1), (4.2).

Bodový odhad regresních koeficientů b_j neříká nic o eventuální variabilitě tohoto koeficientu. Tuto informaci doplňuje směrodatná chyba (4.3) a zejména interval spolehlivosti (4.4), který informuje, v jakém rozmezí se regresní koeficient může pohybovat v rámci zadané spolehlivosti.

Odhadnutý lineární regresní model (3.1), který má tvar

$$y = b_0 + b_1x + e\,, \tag{4.5}$$

resp. regresní funkce

$$Y = b_0 + b_1x\,, \tag{4.6}$$

má praktický význam zejména při odhadu chování modelu v případě, že nezávisle proměnná nabývá nějakou v datech se nevyskytující hodnotu, označme ji např. x_0. Model (4.5), resp. regresní funkce (4.6), pak slouží k *předpovědi (predikci, prognóze, extrapolaci) hodnoty závisle proměnné*. Bodový odhad předpovědi získáme dosazením x_0 do (4.5), resp. (4.6), neboť predikovaná hodnota chyby (rezidua) e je 0, tedy

$$Y_0 = b_0 + b_1x_0\,. \tag{4.7}$$

Informaci o tom, v jakém rozmezí se predikovaná hodnota závisle proměnné y může pohybovat, poskytne oboustranný interval spolehlivosti:

$$[Y_0 - t_{1-\alpha/2}(n-2)\ s_R\sqrt{H}\,,\ Y_0 + t_{1-\alpha/2}(n-2)\ s_R\sqrt{H}\ ], \tag{4.8}$$

kde $H = 1 + \dfrac{1}{n}\left[1 + \dfrac{\left(nx_0 - \sum x_i\right)^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\right]$. Ostatní symboly v (4.8) mají stejný význam, jako v intervalu (4.4).

4.2 Testy hypotéz

Metodou nejmenších čtverců lze zjistit, zda regresní koeficienty b_j jsou nenulová čísla, musíme mít však stále na paměti, že se jedná o realizace náhodných veličin, a tudíž má smysl testovat, zda naše původní parametry β_j jsou přesto nulové. Za předpokladů klasického lineárního modelu je možno testovat nulovou hypotézu:

$$H_0:\ \beta_j = 0,\ j = 0 \text{ nebo } 1 \tag{4.9}$$

proti oboustranné alternativní hypotéze

$$H_1:\ \beta_j \neq 0,\ j = 0 \text{ nebo } 1. \tag{4.10}$$

Při tomto testu použijeme testové kritérium

$$T = \frac{b_j}{\sqrt{\frac{S_R}{n-2} h_j}}, \tag{4.11}$$

které má při platnosti H_0 *t*-rozdělení s $n - 2$ stupni volnosti, S_R je reziduální součet čtverců, h_j je dáno vztahy (4.1), (4.2), přičemž $j = 0$ nebo 1.

Na hladině významnosti α (viz Ramík (2003)) je kritický obor vymezen nerovností

$$|T| > t_{1-\alpha/2}(n-2),$$

kde $t_{1-\alpha/2}(n-2)$ je příslušný kvantil Studentova *t*-rozdělení, který lze nalézt v tabulkách, nebo v Excelu pomocí funkce TINV.

Přijmete-li např. na dané hladině významnosti α nulovou hypotézu H_0: $\beta_1 = 0$, pak to znamená, že *y nezávisí* na *x*, jinak řečeno, pro libovolnou hodnotu nezávisle proměnné *x* nabývá závisle proměnná *y* neustále stejné hodnoty β_0.

Vypočítaná hodnota koeficientu determinace je prakticky vždy kladná. Musíme však mít stále na paměti, že u hodnot vstupujících do výpočtu koeficientu determinace se jedná o realizace náhodných veličin, a tudíž má smysl testovat, zda teoretický koeficient determinace $\mathbf{R}^2$ není přesto nulový. Za předpokladů klasického lineárního modelu je možno testovat nulovou hypotézu:

H_0: $\mathbf{R}^2 = 0$,

proti oboustranné alternativní hypotéze

H_1: $\mathbf{R}^2 \neq 0$.

Při tomto testu použijeme testové kritérium

$$T = \sqrt{\frac{R^2(n-2)}{1-R^2}}, \tag{4.11*}$$

které má při platnosti H_0 *t*-rozdělení s $n - 2$ stupni volnosti, R^2 je vypočítaný koeficient determinace.

Na hladině významnosti α (viz Ramík (2003)) je kritický obor vymezen nerovností

$$T > t_{1-\alpha}(n-2),$$

kde $t_{1-\alpha}(n-2)$ je příslušný kvantil Studentova *t*-rozdělení, který lze nalézt v tabulkách, nebo v Excelu pomocí funkce TINV.

4.3 Nelineární regresní analýza

V tomto odstavci si povšimneme jednoduchého regresního modelu s nelineární regresní funkcí, který se však dá pouhou substitucí na lineární model převést. Konkrétně se jedná o dvě regresní funkce zmíněné již v kapitole 3:

regresní mocninná funkce: $$f(x) = \beta_0 x^{\beta_1}, \tag{4.12}$$

regresní exponenciální funkce: $$f(x) = \beta_0 \beta_1^{\,x}. \tag{4.13}$$

Regresní model s regresní funkcí (4.12) má tvar:

$$y = \beta_0 x^{\beta_1} + \varepsilon\,, \tag{4.14}$$

avšak namísto něj uvažujeme model, jež vznikne logaritmováním (4.12), kde položíme $y = f(x)$, tj. $\ln y = \ln\beta_0 + \beta_1 \ln x + \varepsilon'$, přitom *ln* označuje přirozený logaritmus o základu *e* = 2,718... Jestliže nyní položíte substituce

$$y' = \ln y\,,\ x' = \ln x\,, \tag{4.15}$$

$$\beta_0' = \ln\beta_0\,,\ \beta_1' = \beta_1\,, \tag{4.16}$$

pro transformaci (4.15) původních dat y_i, x_i, obdržíte „čárkovaný“ jednoduchý lineární regresní model

$$y' = \beta_0' + \beta_1' x' + \varepsilon'\,, \tag{4.17}$$

jehož parametry β_0', β_1' (regresní koeficienty) lze odhadnout metodou nejmenších čtverců aplikovanou na lineární model (4.17), a obdržíte tak jejich odhady b_0', b_1'. S použitím vztahů (4.15) a (4.16) dostanete nazpět odhady b_0, b_1 původního nelineárního regresního modelu (4.12):

$$b_0 = e^{b_0'}, b_1 = b_1'\,.$$

Analogickým postupem lze linearizovat jednoduchý nelineární regresní model s exponenciální regresní funkcí (4.13), která je v ekonomii známa jako *Cobb-Douglasova jednofaktorová produkční funkce:*

$$y = \beta_0 \beta_1^x + \varepsilon\,, \tag{4.18}$$

který substitucemi

$$y' = \ln y\,,\ x' = x\,, \tag{4.19}$$

$$\beta_0' = \ln\beta_0\,,\ \beta_1' = \ln\beta_1\,, \tag{4.20}$$

lze rovněž transformovat na „čárkovaný“ lineární model (4.17), jehož parametry β_0', β_1' odhadneme metodou nejmenších čtverců, a obdržíme tak jejich odhady b_0', b_1'. S použitím vztahů (4.20) vypočteme nazpět odhady b_0, b_1 původního nelineárního regresního modelu (4.18):

$$b_0 = e^{b_0'}, b_1 = e^{b_1'}\,. \tag{4.21}$$

Je však třeba upozornit, že na intervalové odhady, resp. testy hypotéz, regresních koeficientů b_0', b_1' lze použít postup z počátku této kapitoly pouze tehdy, když transformovaná, tj. „čárkovaná“ data y_i', x_i', splňují podmínky klasického regresního modelu z kapitoly 3. Meze intervalových odhadů, tedy krajní body intervalů spolehlivosti pak vypočítáme s použitím zpětných transformací (4.21).

Dalšími užitečnými nelineárními regresními funkcemi s uplatněním především v marketingu a výzkumu trhu (logistické funkce, Gompertzovy funkce, aj.) se budete zabývat v kapitole věnované analýze časových řad. Tam se budete zabývat i problémem výběru vhodného typu regresní funkce. V následujících odstavcích se ještě věnujeme známé parabolické regresní funkci a dále Törnquistovým funkcím, které nelze převést jednoduše na lineární tvar, jak tomu bylo v tomto odstavci.

4.4 PARABOLICKÁ REGRESE

V kapitole 3.1. jsme označili parabolickou regresní funkci (3.4) za regresní funkci, kterou lze substitucí $x' = x^2$ převést na lineární tvar. V tomto případě se však jednalo pouze o speciální tvar paraboly (s vrcholem na ose y) se dvěma parametry. Obecný tvar paraboly však má parametry tři a vypadá takto:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 . \qquad (4.22)$$

Jednoduchý regresní model s parabolickou regresní funkcí pak má tvar

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon . \qquad (4.23)$$

Máme-li tedy k dispozici data, tj, dvojice hodnot (y_1, x_1), (y_2, x_2), (y_3, x_3),..., (y_n, x_n), pak lze odhady b_0, b_1, b_2 regresních parametrů $\beta_0, \beta_1, \beta_2$ získat metodou nejmenších čtverců, přičemž je zapotřebí řešit soustavu 3 normálních rovnic o 3 neznámých:

$$\begin{aligned} \sum y_i &= nb_0 + b_1 \sum x_i + b_2 \sum x_i^2 , \\ \sum y_i x_i &= b_0 \sum x_i + b_1 \sum x_i^2 + b_2 \sum x_i^3 , \\ \sum y_i x_i^2 &= b_0 \sum x_i^2 + b_1 \sum x_i^3 + b_2 \sum x_i^4 . \end{aligned} \qquad (4.24)$$

Uvědomte si, že neznámé jsou v této soustavě rovnic b_0, b_1, b_2, zatímco y_i, x_i jsou známé hodnoty, které se dosadí do sum $\sum$ v soustavě (4.24). Tuto soustavu 3 lineárních rovnic o 3 neznámých je snadné vyřešit např. známou Gaussovou eliminační metodou. Číselný příklad uvedeme v následující části věnované řešeným příkladům.

4.5 TÖRNQUISTOVY FUNKCE

Zejména v marketingu se využívají Törnquistovy regresní funkce (též Törnquistovy křivky), což jsou regresní funkce s více parametry, které podle použití rozdělujeme na tři typy:

Törnquistovy křivky I. typu vyjadřují závislosti *poptávky* po spotřebním zboží $f(x)$ na *výši příjmů x* ekonomických subjektů (např. rodin). Tyto křivky mají tvar:

$$f(x) = \frac{\beta_0 x}{x + \beta_1} . \qquad (4.25)$$

Křivky tohoto typu se používají například při plánování a prognózování ve spotřebním průmyslu. Regresní funkce (4.25) slouží k modelování poptávky po *zboží nezbytného charakteru* (mléko, pečivo, obuv, apod.).

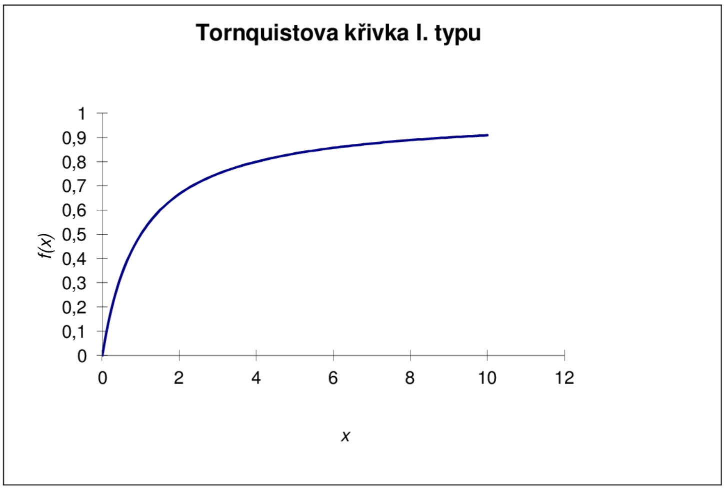
Při modelování poptávky po zboží *relativně nezbytného charakteru* (elektrospotřebiče, maso a uzeniny, apod.) se používají *Törnquistovy křivky II. typu*, které mají tvar:

$$f(x) = \frac{\beta_0 (x - \beta_1)}{x + \beta_2} . \qquad (4.26)$$

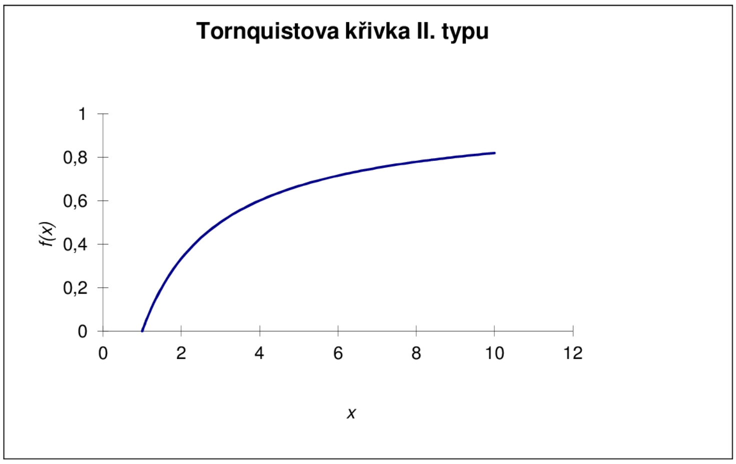
Törnquistovy křivky III. typu se používají při modelování poptávky po zboží zbytného charakteru (auta, šperky, umělecká díla, apod.). Tyto regresní funkce se třemi parametry mají tvar:

$$f(x) = \frac{\beta_0 x (x - \beta_1)}{x + \beta_2} . \qquad (4.27)$$

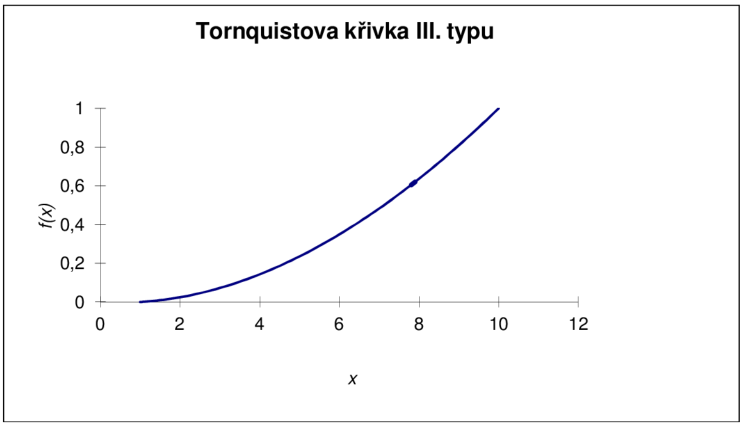
Odhady regresních parametrů funkcí (4.25) - (4.27) lze získat opět metodou nejmenších čtverců, avšak s použitím PC a Excelu, neboť soustava 3 normálních rovnic o 3 neznámých je nelineární, a proto se k řešení používají iterační numerické metody. Pro ruční výpočet můžeme alternativně využít i *metodu vybraných bodů*.



Obr. 4.1. *Törnquistova křivka I. typu,* $\beta_0 = \beta_1 = \beta_2 = 1$



Obr. 4.2. *Törnquistova křivka II. typu,* $\beta_0 = \beta_1 = \beta_2 = 1$



Obr. 4.3. *Törnquistova křivka III. typu,* $\beta_0 = \beta_1 = 1, \beta_2 = 80$

4.6 Metoda vybraných bodů

Ukážeme si zde jinou metodu výpočtu neznámých parametrů, která sice nevede z teoretického pohledu k nejlepším odhadům, avšak její výhoda spočívá ve výpočetní nenáročnosti umožňující „ruční" výpočet. Tato metoda se nazývá *metoda vybraných bodů* a spočívá v tom, že z daných údajů (Y_i, x_i) vybereme 3 charakteristické hodnoty - body, kterými necháme Törnquistovu křivku procházet, jinými slovy, položíme empirické hodnoty rovny hodnotám teoretickým. Jestliže charakteristické hodnoty poptávky Y_1, Y_2, Y_3 odpovídají hodnotám výše příjmů x_1, x_2, x_3, pak ze vztahu (4.26) obdržíte soustavu 3 rovnic o 3 neznámých b_0, b_1, b_2:

$$Y_1 = \frac{b_0(x_1 - b_1)}{x_1 + b_2}, \qquad Y_2 = \frac{b_0(x_2 - b_1)}{x_2 + b_2}, \qquad Y_3 = \frac{b_0(x_3 - b_1)}{x_3 + b_2}, \tag{4.28}$$

jejichž řešením např. postupným dosazováním získáme odhady neznámých parametrů b_0, b_1, b_2.

Řešený příklad 4.1

Data v tabulce představují ceny brožovaných knih a k nim příslušné počty jejich stran.

a. Určete lineární regresní model popisující závislost ceny knih na počtu stran.
b. Určete interval, ve kterém bude s pravděpodobností 95% ležet regresní koeficient b_1.
c. Na hladině významnosti 5% testujte, zda je regresní koeficient b_1 statisticky významný.
d. Vypočtěte koeficient determinace a na hladině významnosti 5% testujte, zda je statisticky významný.
e. V jakém rozmezí se bude pohybovat cena knihy s 250 stranami? Uvažujte hladinu významnosti 0,01.

Měření č.	1	2	3	4	5	6	7
Počet stran	20	35	48	50	130	200	86
Cena knihy	40	50	70	106	118	179	100

Řešení:

a. Koeficienty regresní přímky $Y = b_0 + b_1 x$ určíte pomocí vztahů (3.13):

$$b_1 = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{10135{,}71 - 81{,}29 \cdot 94{,}71}{10103{,}57 - 81{,}29^2} = \frac{2436{,}73}{3495{,}51} = 0{,}70$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 94{,}71 - 0{,}7 \cdot 81{,}29 = 37{,}81.$$

Hledaná regresní přímka má tvar $Y = 37{,}81 + 0{,}7x$.

b. Úkolem je najít 95% oboustranný interval spolehlivosti pro koeficient b_1. Obecný tvar tohoto intervalu je následující (viz (4.4)):

$$[b_1 - t_{1-\alpha/2}(n-2)\, s_R \sqrt{h_1},\ b_1 + t_{1-\alpha/2}(n-2)\, s_R \sqrt{h_1}],$$

kde s_R je odmocnina z reziduálního rozptylu $s_R^2 = \dfrac{S_R}{n-2}$, h_1 je definováno vztahem (4.2).

i	x_i	y_i	x_i^2	x_iy_i	Y_i	$(y_i - Y_i)^2$	$(y_i - \bar{y})^2$
1	20	40	400	800	51,81	139,48	2993,18
2	35	50	1225	1750	62,31	151,54	1998,98
3	48	70	2304	3360	71,41	1,99	610,58
4	50	106	2500	5300	72,81	1101,58	127,46
5	130	118	16900	15340	128,81	116,86	542,42
6	200	179	40000	35800	177,81	1,42	7104,80
7	86	100	7396	8600	98,01	3,96	27,98
Součet	569	663	70725	70950		1516,83	13405,43
Průměr	81,29	94,71	10103,57	10135,7			

Nejprve se vypočítá reziduální součet čtverců S_R (v tabulce výpočtů je to hodnota v předposledním sloupci dole):

$$S_R = \sum_{i=1}^{7}(y_i - Y_i)^2 = 1516{,}83\,.$$

Teoretické hodnoty Y_i obdržíme postupným dosazováním hodnot x_i do rovnice regresní přímky. Hodnoty Y_i, jednotliví sčítanci i součet S_R jsou uvedeni v tabulce. Nyní můžeme vypočítat hodnotu reziduálního rozptylu s_R^2.

$$s_R^2 = \frac{1516{,}83}{7-2} = 303{,}37.$$

Potom

$$s_R = \sqrt{s_R^2} = \sqrt{303{,}37} = 17{,}42\,.$$

Dále stanovíme hodnotu h_1.

$$h_1 = \frac{n}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{7}{7 \cdot 70725 - 569^2} = \frac{7}{171314} = 0{,}00004.$$

V tabulkách Studentova rozdělení nalezneme $(1 - \alpha/2) = 97{,}5\%$ kvantil t-rozdělení o $n - 2 = 7 - 2 = 5$ stupních volnosti, tj. $t_{0,975}(5) = 2{,}57$.

Dosazením výše vypočítaných hodnot do vztahu pro interval spolehlivosti určíme jeho pravou a levou stranu:

$$L = 0{,}7 - 2{,}57 \cdot 17{,}42 \cdot \sqrt{0{,}00004} = 0{,}42\,.$$

$$P = 0{,}7 + 2{,}57 \cdot 17{,}42 \cdot \sqrt{0{,}00004} = 0{,}98\,.$$

Regresní koeficient b_1 bude s 95%-ní pravděpodobností ležet v intervalu [0,42; 0,98].

c. Ačkoliv je hodnota koeficientu b_1= 0,7, nesmíte zapomínat na to, že pracujete s náhodným výběrem a že teoretická hodnota parametru β_1 přesto může být nulová. Bude se proto testovat nulová hypotéza

H_0: $\beta_1 = 0$

proti oboustranné alternativní hypotéze

H_1: $\beta_1 \neq 0$.

K ověření nulové hypotézy vypočítáme hodnotu testového kritéria (4.11)

$$T = \frac{b_1}{\sqrt{\frac{S_R}{n-2}h_1}}. = \frac{0{,}7}{\sqrt{\frac{1516{,}8}{7-2} \cdot 0{,}00004}} = \frac{0{,}7}{0{,}11} = 6{,}35\,.$$

V tabulkách *t*-rozdělení nalezneme $t_{0,975}(5) = 2,57$. Protože 6,35 > 2,57, zamítáme nulovou hypotézu ve prospěch hypotézy alternativní, což znamená, že na zvolené hladině významnosti je parametr β_1 nenulový a tedy statisticky významný.

d. Koeficient determinace R^2 vypočítáme podle vztahu

$$R^2 = 1 - \frac{S_R}{S_y} = 1 - \frac{1516,83}{13405,43} = 0,89 .$$

Testové kritérium stanovíte podle vztahu (4.11*)

$$T = \sqrt{\frac{R^2(n-2)}{1-R^2}} = \sqrt{\frac{0,89.5}{1-0,89}} = 6,35 .$$

Protože 6,35 > 2,57, zamítá se nulová hypotéza ve prospěch hypotézy alternativní, což znamená, že na zvolené hladině významnosti je koeficient determinace $\mathbf{R}^2$ nenulový a tedy statisticky významný.

e. Máte stanovit 99% interval spolehlivosti pro predikovanou hodnotu *Y*, je-li $x_0 = 250$. Podle (4.8) je tvar tohoto intervalu

$$[Y_0 - t_{1-\alpha/2}(n-2)\ s_R\sqrt{H}\ ,\ Y_0 + t_{1-\alpha/2}(n-2)\ s_R\sqrt{H}\ ],$$

kde

$Y_0 = b_0 + b_1 x = 37,81 + 0,7 \cdot 250 = 212,81,$

$t_{1-\alpha/2}(n-2) = 4,032$,

$s_R = 17,42$,

$$H = 1 + \frac{1}{n}\left[1 + \frac{\left(nx_0 - \sum x_i\right)^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\right] = 1 + \frac{1}{7}\left[1 + \frac{(7\cdot 250 - 569)^2}{7\cdot 70725 - 569^2}\right] = 1 + \frac{1}{7}\left(1 + \frac{1394761}{171314}\right) =$$

$$= 1 + \frac{1}{7}\cdot 9,14 = 2,31 .$$

Meze hledaného intervalu jsou:

$$L = 212,81 - 4,032\cdot 17,42\cdot\sqrt{2,31} = 106,06.$$

$$P = 212,81 + 4,032\cdot 17,42\cdot\sqrt{2,31} = 319,56.$$

Cena knihy se bude s 99%-ní pravděpodobností pohybovat v intervalu [106,06;319,56].

Nakonec si ukážeme řešení pomocí Excelu. Na tomto místě to bude další možnost řešení úlohy jednoduché (i vícenásobné) regrese s využitím menu:

Data → Analýza dat... → Regrese.

Data jsou uspořádána ve worksheetu ve 2 sloupcích:

	A	B	C
1	Počet stran	Cena knihy	
2	20	40	
3	35	50	
4	48	70	
5	50	106	
6	130	118	
7	200	179	
8	86	100	
9			

Otevře se okno regrese, které vyplníte takto:

Po potvrzení OK obdržíte:

VÝSLEDEK

Regresní statistika	
Násobné R	0,942
Hodnota spolehlivosti R	0,887
Nastavená hodnota spolehlivosti R	0,864
Chyba stř. hodnoty	17,416
Pozorování	7

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*ýznamnost F*
Regrese	1	11888,84	11888,84	39,19608	0,001525
Rezidua	5	1516,586	303,3172		
Celkem	6	13405,43			

	Koeficienty	*ba stř. hod*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 99,0%*	*Horní 99,0%*
Hranice	38,059	11,19022	3,401	0,019	9,294	66,825	-7,061	83,180
Počet stran	0,697	0,111327	6,261	0,002	0,411	0,983	0,248	1,146

V první části výstupu jsou popisky s nepřesnými překlady do češtiny, uvádíme proto jejich správné významy:

Násobné R	= R - koeficient korelace
Hodnota spolehlivosti R	= R^2 - koeficient determinace
Nastavená hodnota spolehlivosti R	= R^2_{adj} - upravený koeficient determinace
Chyba stř. hodnoty	= s^2 - směrodatná chyba (odhad směrodatné odchylky náhod. složky)

V této části výstupu je důležitá druhá hodnota – koeficient determinace R^2 = 0,887, který odpovídá ručně získanému výsledku z části d.

Druhá tabulka ve výstupu – ANOVA není v pravém slova smyslu metoda ANOVA, jak jsme se jí zabývali v kapitolách 1 a 2, jde tu o analogii využívající podobnosti vztahů (1.5) a (3.17). Analogicky jako v metodě ANOVA je zde výsledek F-testu statistické významnosti celého regresního modelu: Významnost *F* = 0,001525. Tato hodnota je menší než 0,05 a proto je celý regresní model statisticky významný.

Ve třetí – poslední tabulce jsou uvedeny relevantní informace k vypočítanému regresnímu modelu. Nejprve jsou uvedeny odhady regresních koeficientů:

Hranice = úrovňová konstanta = b_0

Počet stran = sklon regresní přímky = koeficient u nezávisle proměnné „počet stran" = b_1

Ve sloupci Hodnota P jsou uvedeny p-hodnoty (signifikance) testů nulovosti příslušných regresních koeficientů:

Pro regresní koeficient b_0 je tato hodnota 0,019 < 0,05 - b_0 je statisticky významný tj. $\beta_0 \neq 0$.

Pro regresní koeficient $b_1$je tato hodnota 0,002 < 0,05 – b_1 je statisticky významný tj. $\beta_1 \neq 0$.

Intervaly spolehlivosti regresních koeficientů jsou uvedeny ve sloupcích:
Dolní 95%, Horní 95%, resp. *Dolní 99,0%, Horní 99,0%.*
Konkrétně, 95%-ní interval spolehlivosti koeficientu β_1 je [0,411 ; 0,983], což je stejný výsledek, jaký jsme obdrželi předtím ručním výpočtem.

ŘEŠENÝ PŘÍKLAD 4.2

Při sledování závislosti vlastních nákladů na skladování zahrnující i ztráty způsobené zastavením výroby z nedostatku součástek (Y) na velikosti dodávek (X) v 18 obuvnických závodech jsme obdrželi následující údaje - viz. tabulka.

a. Nalezněte regresní funkci popisující závislost Y na X a určete její rovnici.
b. Stanovte optimální velikost dodávky.

Podnik	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Dodávka	28	32	35	40	42	45	49	51	53	56	57	60	61	64	69	72	75	77
Náklady	62	59	58	53	50	46	44	42	40	41	38	35	36	36	38	40	42	46

Řešení:

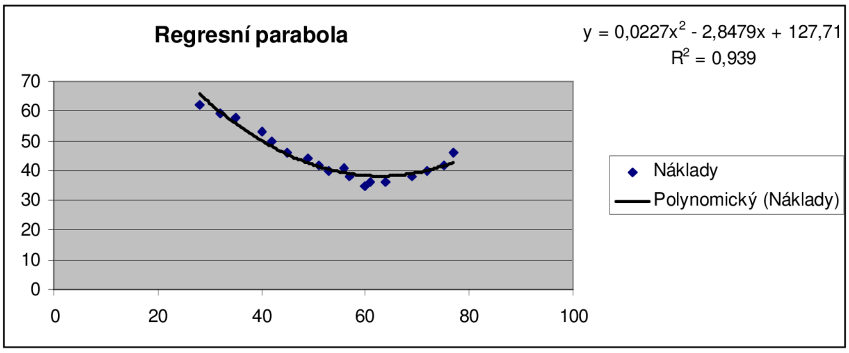
Jak z průběhu bodového diagramu, tak i rozboru empirických údajů plyne, že závislost mezi velikostí dodávek a náklady na skladování dobře vystihuje parabolická regresní funkce

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Náklady na skladování mají zpočátku klesající tendenci- malá dodávka způsobuje vysoké náklady na převzetí připadající na jednu součástku a způsobuje výpadky ve výrobě. Tuto tendenci později vystřídá vzestup – příliš velká dodávka zvyšuje stav zásob, prodlužuje skladovací dobu a vyvolává nutnost úvěrového krytí – viz Obr. 4.4.

Odhady hodnot parametrů parabolické regrese obdržíme řešením soustavy normálních rovnic

$$\sum_i y_i = nb_0 + b_1 \sum_i x_i + b_2 \sum_i x_i^2$$
$$\sum_i y_i x_i = b_0 \sum_i x_i + b_1 \sum_i x_i^2 + b_2 \sum_i x_i^3$$
$$\sum_i y_i x_i^2 = b_0 \sum_i x_i^2 + b_1 \sum_i x_i^3 + b_2 \sum_i x_i^4 .$$



Obr. 4.4. *Parabolická regrese*

Dosazením hodnot ze součtového řádku tabulky do těchto rovnic dostaneme:

$$806 = 18b_0 + 966b_1 + 55534b_2$$

$$41618 = 966b_0 + 55534b_1 + 3372084b_2$$

$$2330182 = 55534b_0 + 3372084b_1 + 213664858b_2 .$$

Řešením této soustavy rovnic (např. Cramerovým pravidlem) získáme regresní koeficienty

$b_0 = 127,71$; $b_1 = -2,8479$; $b_2 = 0,0227$.

Hledaná parabola má tvar

$$Y = 127{,}71 - 2{,}8479x + 0{,}0227x^2 .$$

b. Optimální velikost objednávky zjistíme jako minimum funkce

$$Y = 127{,}71 - 2{,}8479x + 0{,}0227x^2$$

tak, že položíme její první derivaci rovnu nule, tj.

$Y' = -2{,}8479 + 0{,}0454x = 0$, tudíž $x = 62{,}7$.

Optimální velikost dodávky je 62 nebo 63 kusů.

Nakonec provedeme výpočet pomocí Excelu s využitím funkce Přidat spojnici trendu v bodovém grafu.

Po zobrazení dat pomocí grafu XY bodový poklepete pravým tlačítkem myši,
zvolíte položku Typ trendu a rergrese: Polynomický (stupeň 2),
Dále otevřete záložku Možnosti, kde zakliknete:
Zobrazit rovnici regrese (rovnice regresní přímky) a současně zakliknete
Zobrazit hodnotu spolehlivosti *R* (hodnotu koeficientu determinace R^2).
Potvrdíte OK.

i	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
1	28	62	784	21952	614656	1736	48608
2	32	59	1024	32768	1048576	1888	60416
3	35	58	1225	42875	1500625	2030	71050
4	40	53	1600	64000	2560000	2120	84800
5	42	50	1764	74088	3111696	2100	88200
6	45	46	2025	91125	4100625	2070	93150
7	49	44	2401	117649	5764801	2156	105644
8	51	42	2601	132651	6765201	2142	109242
9	53	40	2809	148877	7890481	2120	112360

10	56	41	3136	175616	9834496	2296	128576
11	57	38	3249	185193	10556001	2166	123462
12	60	35	3600	216000	12960000	2100	126000
13	61	36	3721	226981	13845841	2196	133956
14	64	36	4096	262144	16777216	2304	147456
15	69	38	4761	328509	22667121	2622	180918
16	72	40	5184	373248	26873856	2880	207360
17	75	42	5625	421875	31640625	3150	236250
18	77	46	5929	456533	35153041	3542	272734
Součet	966	806	55534	3372084	213664858	41618	2330182

Obdržíte výsledek téměř takový, jaký je na následujícím obrázku. K původním bodům se zobrazí regresní parabola, dále rovnice regresní paraboly a hodnotu koeficientu determinace R^2. Výsledek je stejný, jako při ručním výpočtu, viz výše.

ŘEŠENÝ PŘÍKLAD 4.3

V jisté firmě zkoumali, jak závisí vlastní náklady na jednotku produkce (Y) na objemu produkce (X). Následující tabulka uvádí zjištěné údaje v různých obdobích.

a. Najděte regresní hyperbolický model popisující danou závislost.

b. Pomocí koeficientu determinace zhodnoťte přiléhavost regresní funkce k datům.

Období	1	2	3	4	5	6	7	8	9	10	11	12	13
Objem produkce	0,5	0,7	0,9	1,4	1,9	3,2	4,2	4,8	6,9	7,9	8,8	9,2	10,1
Náklady / jednotka	456	297	206	165	118	79	57	54	40	35	30	23	14

Řešení:

a. Dosadíte potřebné údaje do normálních rovnic, které získáte z hyperbolické regresní funkce (3.5) tak, že k nalezení minima součtu čtverců odchylek:

$F(b_0, b_1) = \sum \left(y_i - (b_0 + b_1 \frac{1}{x_i}) \right)^2$ se anulují parciální derivace, tj. $\frac{\partial F}{\partial b_0} = 0$ a $\frac{\partial F}{\partial b_1} = 0$. Tím

obdržíte následující normální rovnice:

$$\sum y_i = n \cdot b_0 + b_1 \sum \frac{1}{x_i}$$

$$\sum \frac{y_i}{x_i} = b_0 \sum \frac{1}{x_i} + b_1 \sum \frac{1}{x_i^2}$$

a obdržíme soustavu 2 rovnic o 2 neznámých

$$1574 = 13 \cdot b_0 + b_1 \cdot 7{,}13$$

$$1812{,}19 = b_0 \cdot 7{,}13 + b_1 \cdot 8{,}33.$$

Řešením této soustavy získáte odhady regresních parametrů:

$b_0 = 3{,}32$; $b_1 = 214{,}71$.

Hledaná regresní hyperbola má tvar: $Y = 3{,}32 + \frac{214{,}71}{x}$.

b. Nejdříve vypočítáte teoretické hodnoty Y_i postupným dosazením hodnot x_i do rovnice regresní hyperboly

$$Y_1 = 3{,}32 + \frac{214{,}71}{x_1} = 3{,}32 + \frac{214{,}71}{0{,}5} = 432{,}74\,.$$

Všechny hodnoty Y_i jsou uvedeny v tabulce, viz níže.

Dále vypočítáte součty S_T, S_y

$$S_T = \sum_{i=1}^{13}(Y_i - \bar{y})^2 = (432{,}74 - 121{,}08)^2 + (310{,}05 - 121{,}08)^2 + \ldots + (24{,}58 - 121{,}08)^2 =$$
$$= 203722{,}02.$$

$$S_y = \sum_{i=1}^{13}(y_i - \bar{y})^2 = (456 - 121{,}08)^2 + (297 - 121{,}08)^2 + \ldots + (14 - 121{,}08)^2 = 2060{,}97\,.$$

i	x_i	y_i	$1/x_i$	$1/x_i^2$	y_i/x_i	Y_i	$(Y_i - \bar{y})^2$	$(y_i - \bar{y})^2$
1	0,5	456	2,00	4,00	912,00	432,74	97131,96	112171,41
2	0,7	297	1,43	2,04	424,29	310,05	35709,66	30947,85
3	0,9	206	1,11	1,23	228,89	241,89	14595,06	7211,41
4	1,4	165	0,71	0,51	117,86	156,68	1267,36	1928,97
5	1,9	118	0,53	0,28	62,11	116,33	22,56	9,49
6	3,2	79	0,31	0,10	24,69	70,42	2566,44	1770,73
7	4,2	57	0,24	0,06	13,57	54,44	4440,89	4106,25
8	4,8	54	0,21	0,04	11,25	48,05	5333,38	4499,73
9	6,9	40	0,14	0,02	5,80	34,44	7506,49	6573,97
10	7,9	35	0,13	0,02	4,43	30,50	8204,74	7409,77
11	8,8	30	0,11	0,01	3,41	27,72	8716,09	8295,57
12	9,2	23	0,11	0,01	2,50	26,66	8915,14	9619,69
13	10,1	14	0,10	0,01	1,39	24,58	9312,25	11466,13
Součet	60,5	1574	7,13	8,33	1812,19		203722,02	206010,97
Průměr	4,65	121,08	0,55	0,64	139,40			

Hodnoty jednotlivých sčítanců i součtů S_T, S_y jsou uvedeny v tabulce.

Koeficient determinace R^2 vypočítáte podle vztahu (3.18).

$$R^2 = \frac{S_T}{S_y} = \frac{203722{,}02}{206011{,}97} = 0{,}99\,.$$

Hodnota koeficientu determinace 0,99 je vysoká, což znamená, že daným regresním modelem s vysvětlující proměnnou „objem produkce“ je vysvětleno 99% variability znaku Y. Pouze 1% chování proměnné Y je ovlivněno jinými faktory.

ŘEŠENÝ PŘÍKLAD 4.4

Data v tabulce ukazují poptávku po určitém druhu zboží (v tis. ks) při různých cenách (v Kč). Popište závislost poptávky na ceně mocninnou regresní funkcí.

Pozorování	**1**	**2**	**3**	**4**	**5**	**6**
Cena	8,5	40	92	180	200	250
Poptávka	200	140	80	45	42	18

Řešení:

Úkolem je nalézt odhady parametrů β_1, β_0 regresní funkce $Y = \beta_0 x^{\beta_1}$.
Použijete linearizující transformace, a to tak, že obě strany rovnice zlogaritmujete a použijete vhodnou substituci (viz odstavec 4.3), čímž získáte rovnici

$$Y' = \beta_0' + \beta_1' x',$$

kde $Y' = \ln Y, x' = \ln x, \beta_0' = \ln \beta_0, \beta_1' = \beta_1$, což je rovnice regresní přímky.

Regresní koeficienty b_0', b_1' určíme pomocí známých vztahů takto:

$$b_1' = \frac{\overline{x'y'} - \overline{x'} \cdot \overline{y'}}{\overline{x'^2} - \overline{x'}^2} = \frac{17{,}49 - 4{,}39 \cdot 4{,}18}{20{,}7 - 4{,}39 \cdot 4{,}39} = \frac{-0{,}86}{1{,}43} = -0{,}6$$

$$b_0' = \overline{y'} - b_1' \overline{x'} = 4{,}18 - (-0{,}6 \cdot 4{,}39) = 6{,}8.$$

i	***x***	***y***	***x′***	***y′***	***x′y′***	***x′*²**
1	8,5	200	2,14	5,30	11,34	4,58
2	40	140	3,69	4,94	18,23	13,61
3	92	80	4,52	4,38	19,81	20,45
4	180	45	5,19	3,81	19,77	26,97
5	200	42	5,30	3,74	19,80	28,07
6	250	18	5,52	2,89	15,96	30,49
Průměr			4,39	4,18	17,49	20,70

Odhady b_0, b_1 původního modelu snadno vypočítáte zpětnou transformací

$$b_1' = b_1, b_0 = e^{b_0'}.$$

Proto bude

$$b_1 = -0{,}6; b_0 = 897{,}85.$$

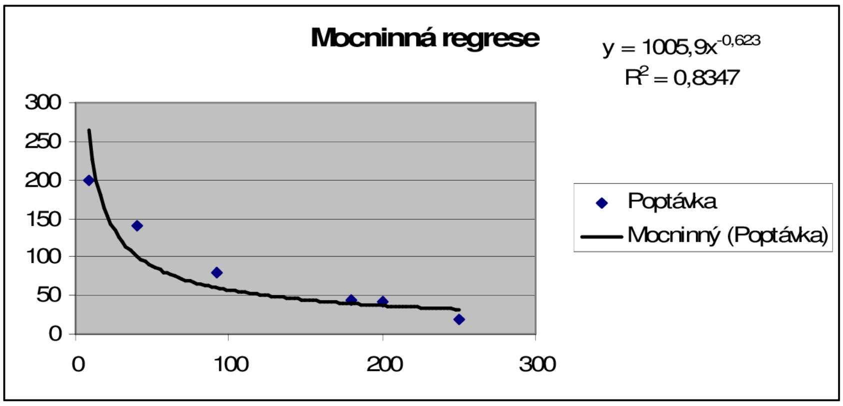
Hledaná mocninná regresní funkce má tvar

$$Y = 897{,}85 \cdot x^{-0{,}6}.$$

Nakonec provedeme výpočet pomocí Excelu s využitím funkce Přidat spojnici trendu v bodovém grafu.
Po zobrazení dat pomocí grafu XY bodový poklepete pravým tlačítkem myši, zvolíte položku

Typ trendu a rergrese: Mocninný,
Dále otevřete záložku Možnosti, kde zakliknete:
Zobrazit rovnici regrese (rovnice regresní přímky) a současně zakliknete
Zobrazit hodnotu spolehlivosti R (hodnotu koeficientu determinace R^2).
Potvrdíte OK.

Obdržíte výsledek, jaký je na následujícím obrázku. K původním bodům se zobrazí regresní mocninná funkce, dále její rovnice a hodnotu koeficientu determinace R^2. Výsledek je poněkud odlišný od výsledku, který jsme získali při ručním výpočtu, viz výše. Tato odlišnost je způsobena tím, že Excel počítá koeficienty přímo metodou nejmenších čtverců bez použití linearizace s logaritmickou transformací. Metoda použita Excelem je přesnější než metoda linearizace a proto bychom ji dali při aplikaci přednost. Metoda linearizace je zase výpočetně jednodušší, je ji možno provést ručně, v době počítačů však tato výhoda ztrácí na významu.



Obr. 4.5. *Mocninná regrese*

ŘEŠENÝ PŘÍKLAD 4.5

Tabulka uvádí stáří pletacích strojů (X) v letech a náklady na jejich údržbu (Y) v tis. Kč. Popište závislost Y na X exponenciální regresní funkcí.

Měření	1	2	3	4	5	6	7	8	9	10	11	12
Stáří	14	0,8	3	7,5	8,4	14,8	4,5	15,6	17,3	11,5	13,2	1,5
Náklady	47,5	8	10	17	22	76,4	12,5	76	94,5	25	30,6	12

Řešení:

Úkolem je nalézt odhady regresních parametrů exponenciální regresní funkce

$$y = \beta_0 \beta_1^x .$$

Pomocí logaritmické transformace převedeme tuto funkci na funkci lineární:

$$\ln y = \ln \beta_0 + x \ln \beta_1 .$$

Použitím substituce

$$y' = \ln Y, x' = x, \beta_0' = \ln \beta_0, \beta_1' = \ln \beta_1$$

obdržíte regresní přímku $y' = \beta_0' + \beta_1' x'$.

Odhady parametrů β_0', β_1' této přímky určíme použitím známých vztahů

$$b_1' = \frac{\overline{x'y'} - \overline{x'} \cdot \overline{y'}}{\overline{x'^2} - \overline{x'}^2} = \frac{34{,}8 - 9{,}34 \cdot 3{,}25}{118{,}59 - 9{,}34^2} = \frac{4{,}45}{31{,}35} = 0{,}14$$

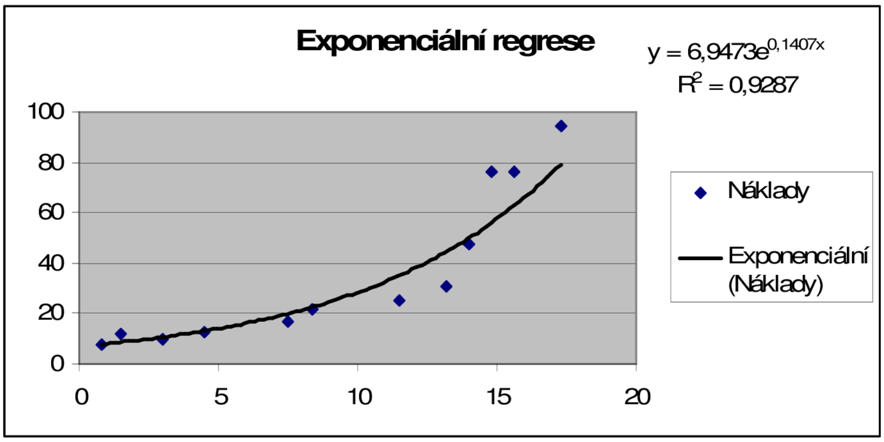
$$b_0' = \overline{y'} - b_1' \overline{x'} = 3{,}25 - (0{,}14 \cdot 9{,}34) = 1{,}94.$$

Regresní koeficienty původní funkce snadno vypočítáme zpětnou transformací:

$b_0 = e^{b_0'} = 6{,}96; b_1 = e^{b_1'} = 1{,}15$. Hledaná exponenciální regresní funkce má tvar:

$$y = 6{,}96 \cdot 1{,}15^x = 6{,}96 \cdot e^{0{,}14x}.$$

i	$x_i = x_i'$	y_i	y_i'	$x_i' y_i'$	x'^2
1	14	47,5	3,86	54,04	196,00
2	0,8	8	2,08	1,66	0,64
3	3	10	2,30	6,90	9,00
4	7,5	17	2,83	21,23	56,25
5	8,4	22	3,09	25,96	70,56
6	14,8	76,4	4,34	64,23	219,04
7	4,5	12,5	2,53	11,39	20,25
8	15,6	76	4,33	67,55	243,36
9	17,3	94,5	4,55	78,72	299,29
10	11,5	25	3,22	37,03	132,25
11	13,2	30,6	3,42	45,14	174,24
12	1,5	12	2,48	3,72	2,25
Průměr	**9,34**		**3,25**	**34,80**	**118,59**



Obr. 4.6. *Exponenciální regrese*

Nakonec provedeme výpočet pomocí Excelu s využitím funkce Přidat spojnici trendu v bodovém grafu. Po zobrazení dat pomocí grafu XY bodový poklepete pravým tlačítkem myši, zvolíte položku

Typ trendu a rergrese: Exponenciální,

Dále otevřete záložku Možnosti, kde zakliknete:

Zobrazit rovnici regrese (rovnice regresní přímky) a současně zakliknete

Zobrazit hodnotu spolehlivosti R (hodnotu koeficientu determinace R^2).

Potvrdíte OK.

Obdržíte výsledek, jaký je na následujícím obrázku. K původním bodům se zobrazí regresní exponenciální funkce, dále její rovnice a hodnotu koeficientu determinace R^2. Výsledek je prakticky stejný jako výsledek, který jsme získali při ručním výpočtu, viz výše.

4.7 SAMOSTATNÉ ÚKOLY

4.1 Tabulka zachycuje stáří (v letech) osmi vybraných strojů v potravinářském závodě a týdenní náklady (v Kč) na provoz těchto strojů.

Stáří stroje	1	2	3	4	5	6	7	8
Náklady	44	52	61	80	94	108	111	116

a. Odhadněte parametry regresní funkce $f(x)=\beta_0+\beta_1 \ln x$, která by měla vystihovat průběh závislosti nákladů na stáří.

b. Jaké týdenní náklady můžeme očekávat u stroje starého 4 roky?

c. Určete koeficient determinace a interpretujte jej.

4.2 V tenisovém zápase má významný vliv na vítězství hráče úspěšnost jeho prvního podání. Data v tabulce představují počet úspěšných prvních podání (X) a počet vyhraných bodů při úspěšném prvním podání (Y) deseti vybraných hráčů z předních míst žebříčku ATP.

X	31	42	39	41	50	38	33	49	37	46
Y	22	31	29	26	33	26	23	30	29	31

Zvolte nejprve lineární a potom parabolický typ regresní funkce popisující závislost Y na X.

a. Určete regresní parametry obou zvolených regresních funkcí.

b. Stanovte 95% interval spolehlivosti pro regresní koeficient b_1 u lineární regrese.

c. Zhodnoťte výstižnost obou zvolených regresních funkcí. Která z nich lépe vystihuje data?

4.8 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

4.1 a) $Y = 32{,}29 + 38{,}44 \cdot \ln x$

b) $Y(4) = 32{,}29 + 38{,}44 \cdot \ln 4 = 85{,}58 Kč$

c) $R^2 = 0{,}92$

4.2	**lineární regresní funkce**	**kvadratická regresní funkce**
a)	$Y = 7{,}95 + 0{,}49x$	$Y = -25{,}94 + 2{,}19x - 0{,}02x^2$
b)	$b1 \in \langle 0{,}26;\ 0{,}73 \rangle$	
c)	$R^2 = 0{,}75$	$R^2 = 0{,}79$

Model lépe vystihuje kvadratická regresní funkce.

5 REGRESNÍ ANALÝZA - VÍCEROZMĚRNÁ

RYCHLÝ NÁHLED KAPITOLY

V této kapitole navážete na jednoduchou regresi vyšetřovanou v předchozí kapitole. Nyní budeme předpokládat, že vysvětlovaná proměnná závisí na několika (více než jedné) vysvětlujících proměnných. Vícenásobný lineární regresní model je zobecněním jednoduchého lineárního regresního modelu. Lineární regresní model bude rozšířen na vícenásobný regresní model lineární v parametrech, který předpokládá lineární vztah pouze v regresních koeficientech, nikoliv nutně v nezávisle proměnných. Odhady regresních koeficientů se stanoví opět metodou nejmenších čtverců, přitom lze využít maticové symboliky, která usnadňuje práci s vektory a maticemi. Podobně jako v případě jednoduché regrese budou formulovány předpoklady klasického regresního modelu, přičemž obdržíte analogické výsledky pro intervaly spolehlivosti regresních koeficientů a odpovídající testy hypotéz jako v případě jednoduché regrese. Nejprve budeme předpokládat, že vysvětlovaná proměnná Y závisí na několika (konkrétně k) vysvětlujících proměnných X_1, X_2,..., X_k. Zkonstruujeme vícenásobný lineární regresní model, který předpokládá lineární vztah pouze v regresních koeficientech, (nikoliv nutně v nezávisle proměnných – uvědomte si ten rozdíl!). Odhady regresních koeficientů se stanoví opět metodou nejmenších čtverců, přitom lze využít maticové symboliky, která usnadňuje práci s vektory a maticemi. Podobně jako v případě jednoduché regrese budou dále formulovány předpoklady klasického regresního modelu, přičemž obdržíte analogické výsledky pro intervaly spolehlivosti regresních koeficientů a odpovídající testy hypotéz jako v případě jednoduché regrese.

5.1 VÍCEROZMĚRNÁ REGRESNÍ ANALÝZA

Na rozdíl od předchozích dvou kapitol, kde jsme předpokládali, že vysvětlovaná proměnná Y závisí na jediné vysvětlující proměnné X, budeme nyní předpokládat, že vysvětlujících proměnných je několik (tj. alespoň 2), řekněme k, kde $k \geq 2$, přitom k je celé číslo. Vysvětlující statistické znaky (proměnné) označíme X_1, X_2,...,X_k, i-tému pozorování (i-té realizaci) hodnot vysvětlujících znaků $x_{i1}, x_{i2}, \ldots, x_{ik}$ odpovídá hodnota vysvětlovaného znaku y_i. Vícenásobný lineární regresní model je zobecněním jednoduchého lineárního regresního modelu (4.9) a má následující tvar:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1,2,\ldots,n. \tag{5.1}$$

Jak jste viděli v předchozí kapitole při aplikaci metody linearizace, bylo pro použití metody nejmenších čtverců podstatné, že regresní funkce byla lineární v parametrech β_i, nikoliv v proměnné x. Tohoto důležitého faktu využijeme nyní a formulujeme poněkud obecnější model, než (5.1), totiž vícenásobný regresní model *lineární v parametrech*. Ten vypadá takto

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}, x_{i2}, \ldots, x_{ik}) + \beta_2 f_2(x_{i1}, x_{i2}, \ldots, x_{ik}) + \ldots + \beta_k f_k(x_{i1}, x_{i2}, \ldots, x_{ik}) + \varepsilon_i,$$
$$i = 1,2,\ldots,n. \tag{5.2}$$

kde $f_j(x_1, x_2, \ldots, x_k)$, $j = 1,2,\ldots,k$, jsou funkce proměnných $x_1, x_2, \ldots, x_k$, nezávislé na parametrech β_i.

5.2 METODA NEJMENŠÍCH ČTVERCŮ

Odhady regresních koeficientů $b_0, b_1, ..., b_k$ lze stanovit metodou nejmenších čtverců, která spočívá v minimalizaci součtu kvadrátů (tj. druhých mocnin) odchylek skutečných hodnot dat y_i od teoretických hodnot $Y_i = b_0 + b_1 f_1(x_{i1}, x_{i2}, ..., x_{ik}) + ... + b_k f_k(x_{i1}, x_{i2}, ..., x_{ik})$. Podobně, jako u jednoduchého modelu, vypočteme odhady ze soustavy normálních rovnic:

$$\frac{\partial S_R}{\partial b_0} = 0, \; \frac{\partial S_R}{\partial b_1} = 0, \; ..., \; \frac{\partial S_R}{\partial b_k} = 0 \; . \tag{5.3}$$

V (5.3) se jedná o parciální derivace funkce S_R podle proměnných b_i. Označení

$$F_{ij} = f_i(x_{j1}, x_{j2}, ..., x_{jk}), \; i = 1,2,...,k, \; j = 1,2,...,n, \tag{5.4}$$

umožní využít maticovou symboliku. Soustavu rovnic (5.2) lze maticově zapsat takto:

$$\mathbf{y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.5}$$

kde matice:

$$\mathbf{F} = \begin{bmatrix} 1 & F_{11} & \cdots & F_{k1} \\ 1 & F_{12} & \cdots & F_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & F_{1n} & \cdots & F_{kn} \end{bmatrix}$$ se nazývá *matice regresorů*,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$ je vektor pozorování vysvětlované proměnné *Y*,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$, resp. $$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$, je vektor regresních koeficientů, resp. vektor jejich odhadů.

Dále

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$, je vektor náhodných složek.

Při výpočtu vektoru odhadů **b** regresních koeficientů metodou nejmenších čtverců obdržíte soustavu normálních lineárních rovnic, které lze maticově vyjádřit. Pozor, používáte přitom pravidla pro sečítání a násobení matic - pravidlo „řádek krát sloupec“. Toho lze dosáhnout tak, že regresní rovnici $\mathbf{y} = \mathbf{F}.\mathbf{b}$, vynásobíte zleva transponovanou maticí $\mathbf{F}^T$**,** takže obdržíte

$$\mathbf{F}^T\mathbf{y} = \mathbf{F}^T\mathbf{F}.\mathbf{b} \; , \tag{5.6}$$

a za předpokladu, že matice $\mathbf{F}^T\mathbf{F}$ je regulární, a tedy existuje k ní matice inverzní $(\mathbf{F}^T\mathbf{F})^{-1}$, lze nalézt řešení soustavy, tj. vektor odhadů regresních koeficientů modelu (5.5), a to po vynásobení (5.6) zleva maticí $(\mathbf{F}^T\mathbf{F})^{-1}$, ve tvaru:

$$\mathbf{b} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}. \tag{5.7}$$

Ve speciálním případě jednoduché lineární regrese je k = 1, pak matice regresorů a další prvky z (5.6) mají tvar:

$$\mathbf{F} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}, \ \mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}, \ \mathbf{F}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix},$$

a soustava normálních rovnic (5.6) je následující:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}, \tag{5.8}$$

což je tvar ekvivalentní rovnicím (3.12) , (3.13).

5.3 NÁHODNÝ VEKTOR A JEHO CHARAKTERISTIKY

Nyní ještě rozšíříme pojmy střední hodnoty a rozptylu používané doposud pro náhodnou veličinu (skalár), a to pro náhodný vektor:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \tag{5.9}$$

kde složky X_i jsou náhodné veličiny. *Střední hodnota* $\boldsymbol{E}(\mathbf{X})$ *vektorové náhodné veličiny* $\mathbf{X}$ je vektor středních hodnot jednotlivých složek, tj.:

$$\boldsymbol{E}(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}. \tag{5.10}$$

Rozptyl (*variance*) $\boldsymbol{Var}(\mathbf{X})$ *vektorové náhodné veličiny* $\mathbf{X}$ je matice:

$$\boldsymbol{Var}(\mathbf{X}) = \boldsymbol{E}((\mathbf{X}- \boldsymbol{E}(\mathbf{X}))^{\mathrm{T}}(\mathbf{X} - \boldsymbol{E}(\mathbf{X}))), \tag{5.11}$$

kde čárkou ´ označujeme transponovanou matici (vektor). Jistě jste si všimli, že rozptyl náhodného vektoru (5.11) je čtvercová matice typu ($n\times n$).

5.4 KLASICKÝ LINEÁRNÍ MODEL

O *klasickém (vícerozměrném) lineárním regresním modelu* hovoříme tehdy, když matice regresorů má nejjednodušší tvar, tj. když je matice tvořena danými hodnotami pozorování vysvětlujících proměnných:

$$F_{ij} = x_{ij}, \ i = 1,2,...,k, \ j = 1,2,...,n. \tag{5.12}$$

V tom případě má matice regresorů tvar:

$$\mathbf{F} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}. \tag{5.13}$$

U klasického lineárního modelu požadujeme splnění podmínek 1. až 3. z minulé kapitoly, přitom u těchto podmínek nebylo důležité, zda jde o jednoduchý nebo vícerozměrný regresní model:

1. Hodnoty vysvětlujících proměnných X_1, X_2,...,X_k, tvořící matici regresorů **F** podle (5.13) se volí předem, nejsou to tedy náhodné veličiny.

2. Reziduum ε v modelu (3.5) má *normální rozdělení* pravděpodobnosti s nulovou střední hodnotou a (neznámým) rozptylem σ^2, tj.:

$$\boldsymbol{E}(\boldsymbol{\varepsilon}) = \mathbf{0}\,, \tag{5.14}$$

$$\boldsymbol{Var}(\boldsymbol{\varepsilon}) = \sigma^2\,\mathbf{I}\,, \tag{5.15}$$

kde symbol **I** označuje jednotkovou matici.

Vztah (5.15) zahrnuje zároveň podmínku 3. z klasického lineárního modelu, viz kapitola 3.5, neboť na diagonále matice $\boldsymbol{Var}(\boldsymbol{\varepsilon})$ jsou rozptyly σ^2 jednotlivých složek náhodného vektoru $\boldsymbol{\varepsilon}$ a mimo diagonálu vystupují nulové kovariance těchto složek. V tom případě hovoříme o *homoskedasticitě*. V opačném případě hovoříme o přítomnosti *heteroskedasticity.*

3. Vysvětlující proměnné X_1, X_2,...,X_k, nejsou kolineární, tj. sloupcové vektory matice regresorů (5.13) jsou nekorelované. V opačném případě hovoříme o přítomnosti multikolinearity.

5.5 MÍRY VARIABILITY A KOEFICIENT DETERMINACE

Podobně jako u jednoduché regrese, zajímáme se nyní o *celkovou variabilitu* vysvětlované proměnné, kterou charakterizuje celkový součet čtverců:

$$S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2\,. \tag{5.16}$$

Část celkové variability vysvětlenou regresním modelem charakterizuje *teoretický součet čtverců*:

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{y})^2\,, \tag{5.17}$$

kde $Y_i = b_0 + b_1 f_1(x_{i1}, x_{i2},..., x_{ik}) + ... + b_k f_k(x_{i1}, x_{i2},..., x_{ik})$, b_i jsou odhady regresních parametrů získané MNČ. Nevysvětlenou část celkové variability představuje reziduální součet čtverců:

$$S_R = \sum_{i=1}^{n} (y_i - Y_i)^2\,, \tag{5.18}$$

kde $e_i = y_i - Y_i$ je *reziduum*, tj. odhad náhodné složky ε_i.

Mezi jednotlivými součty čtverců platí základní vztah:

$$S_y = S_T + S_R . \tag{5.19}$$

Obdobně, jako v případě jednoduché regrese, zavedeme analogický pojem, charakterizující přiléhavost dat k regresnímu modelu, *koeficient determinace*, který definujeme vztahem:

$$R^2 = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y} . \tag{5.20}$$

Koeficient determinace nabývá hodnoty z intervalu [0,1] a určuje tu část celkové variability pozorovaných hodnot y_i, kterou lze vysvětlit daným regresním modelem. Jinak řečeno, po vynásobení koeficientu determinace stem obdržíme, kolik procent celkové variability je vysvětlitelných regresním modelem.

Nevychýlený odhad koeficientu determinace R^2_{adj}, který nazýváme *korigovaný (upravený) koeficient determinace*, definujeme takto:

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p} , \tag{5.21}$$

kde $p = k+1$ označuje počet parametrů v regresním modelu (5.2).

5.6 INTERVALY SPOLEHLIVOSTI A TESTY HYPOTÉZ

Tento odstavec je přirozeným rozšířením kapitoly 4 pro jednoduchý klasický lineární model, tj. model (3.9) se dvěma parametry β_0, β_1. Nyní máme analogický model, avšak s k+1 parametry $\beta_0, \beta_1, ..., \beta_k$.

Jsou-li splněny předpoklady klasického lineárního modelu (5.5), tj. modelu:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i , \; i = 1,2,...,n, \tag{5.22}$$

potom pro rozdělení odhadů regresních koeficientů $b_0, b_1, \ldots, b_k$, jakožto náhodných veličin, platí toto: Regresní koeficient b_j má normální rozdělení pravděpodobnosti se střední hodnotou β_j a rozptylem $\sigma^2 h_{jj}$, kde $j = 0,1,...,k$, čísla h_{jj} jsou diagonálními prvky matice:

$$\mathbf{H} = (\mathbf{F}^T\mathbf{F})^{-1}, \tag{5.23}$$

kde matice **F** je definována vztahem (5.13).

V klasickém lineárním modelu předpokládáme, že reziduální složky mají konstantní rozptyl σ^2, jeho hodnotu však zpravidla neznáme. Neznámý rozptyl σ^2 můžeme nahradit jeho bodovým odhadem:

$$s_R^2 = \frac{S_R}{n-p} , \tag{5.24}$$

který nazýváme v souladu s (5.22) *reziduální rozptyl.* V reziduálním rozptylu vystupuje v čitateli reziduální součet čtverců (5.18) dělený číslem $n - p$, což je *počet stupňů volnosti*, tj. rozsah dat n mínus počet regresních koeficientů v modelu: $p = k + 1$. Odmocninu reziduálního rozptylu s_R nazýváme *směrodatná chyba.*

Oboustranný interval spolehlivosti pro regresní koeficient b_j, při zadaném koeficientu spolehlivosti $(1 - \alpha)$, je následující interval:

$$\left[b_j - t_{1-\alpha/2}(n-p)\sqrt{\frac{S_R h_{jj}}{n-p}} ,\; b_j + t_{1-\alpha/2}(n-p)\sqrt{\frac{S_R h_{jj}}{n-p}}\right], \; j = 0,1,...,k. \tag{5.25}$$

Zde $t_{1-\alpha/2}(n-p)$ je příslušný kvantil Studentova t-rozdělení, h_{jj} diagonální prvky matice (5.23). Interval (4.23) je speciálním případem intervalu (5.25) v případě $k = 1$.

Bodový odhad regresních koeficientů b_j, vypočtený metodou nejmenších čtverců, doplňuje interval spolehlivosti (5.25), který informuje, v jakém rozmezí se regresní koeficient může pohybovat v rámci zadané spolehlivosti v případě jiného náhodného výběru dat (ze stejného základního souboru). Odhadnutý lineární regresní model (3.9), který má tvar:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + e, \tag{5.26}$$

kde e je reziduum, tj. odhad náhodné složky ε, resp. regresní funkce:

$$Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k, \tag{5.27}$$

má praktický význam zejména při odhadu chování modelu pro nezávisle proměnné nevyskytující se v datech, např. hodnoty $x_{01}, x_{02}, \ldots, x_{0k}$. Model (5.26), resp. regresní funkce (5.27), pak slouží k predikci hodnoty závisle proměnné. Bodový odhad předpovědi získáme dosazením $\mathbf{x_0} = (x_{01}, x_{02}, \ldots, x_{0k})'$ do (5.27):

$$Y_0 = b_0 + b_1x_{01} + b_2x_{02} + \ldots + b_kx_{0k}. \tag{5.28}$$

Informaci o tom, v jakém rozmezí se predikovaná hodnota vysvětlované proměnné může pohybovat, poskytuje oboustranný interval spolehlivosti:

$$[Y_0 - t_{1-\alpha/2}(n-p)\, s_R\sqrt{1+\mathbf{x}_0^{\mathrm{T}}\mathbf{H}\mathbf{x}_0}\,,\ Y_0 + t_{1-\alpha/2}(n-p)\ s_R\sqrt{1+\mathbf{x}_0^{\mathrm{T}}\mathbf{H}\mathbf{x}_0}\ ], \tag{5.29}$$

kde $\mathbf{H} = (\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}$ a matice $\mathbf{F}$ je definována vztahem (5.13). Ostatní symboly v (5.29) mají stejný význam, jako v intervalu spolehlivosti (5.25).

5.7 Individuální *t*-testy o hodnotách regresních koeficientů

Tento odstavec je rozšířením kapitoly 4.2 pro vícerozměrný lineární regresní model. Zjistíme-li metodou nejmenších čtverců, že regresní koeficienty b_j jsou nějaká nenulová čísla, musíme mít stále na paměti, že se jedná o realizace náhodných veličin, a tudíž má smysl testovat, zda naše původní parametry β_j nemohou být přesto nulové. Za předpokladů klasického lineárního modelu je možno pro $j = 0,1,\ldots,k$ testovat nulovou hypotézu:

$$\mathrm{H_0}: \beta_j = 0, \tag{5.30}$$

proti oboustranné alternativní hypotéze:

$$\mathrm{H_1}: \beta_j \neq 0. \tag{5.31}$$

Při tomto testu použijeme testové kritérium:

$$t = \frac{b_j}{\sqrt{\dfrac{S_R}{n-p}h_{jj}}}, \tag{5.32}$$

které má při platnosti H_0 t-rozdělení s $n-p$ stupni volnosti, S_R je reziduální součet čtverců, h_{jj} jsou diagonální prvky matice $\mathbf{H}$ z (5.23), přičemž $j = 0,1,\ldots,k$, $p = k + 1$.
Na hladině významnosti α je kritický obor vymezen nerovností:

$$|t| > t_{1-\alpha/2}(n-p),$$

kde $t_{1-\alpha/2}(n-p)$ je příslušný kvantil Studentova t-rozdělení, viz funkci v Excelu TINV,

Nemůžeme-li např. na dané hladině významnosti α zamítnout nulovou hypotézu $\mathrm{H_0}$: $\beta_j = 0$, pak to znamená, že y nezávisí na x_j, jinak řečeno, pro libovolnou hodnotu vysvětlující proměnné x_j nabývá vysvětlovaná proměnná y stále stejné hodnoty.

5.8 F-TEST HYPOTÉZY O HODNOTÁCH REGRESNÍCH KOEFICIENTŮ

V minulém odstavci jste individuálními *t*-testy zjišťovali vliv jednotlivých vysvětlujících proměnných na vysvětlovanou proměnnou. V tomto odstavci se budeme zabývat testem, který najednou odhalí, zda vůbec existuje nějaká vysvětlující proměnná, která má na vysvětlovanou proměnnou nějaký vliv. Testuje se nulová hypotéza:

$$H_0: \ \beta_1 = \beta_2 = ... = \beta_k = 0, \tag{5.33}$$

proti alternativní hypotéze, že pro alespoň jeden regresní koeficient platí $\beta_j \neq 0$.

Testové kritérium:

$$T = \frac{\dfrac{S_T}{p-1}}{\dfrac{S_R}{n-p}} \tag{5.34}$$

má Fisherovo rozdělení F s (p – 1) a (n – p) stupni volnosti. Na hladině významnosti α je kritický obor vymezen nerovností:

$$T > F_{1-\alpha}(p-1, n-p), \tag{5.35}$$

kde $F_{1-\alpha}(p-1, n-p)$ je příslušný kvantil rozdělení. Pokud hodnota testového kritéria padne do kritického oboru, tedy pokud platí (5.35), potom H_0 zamítáme, což znamená, že některá z vysvětlujících proměnných má statisticky významný efekt na vysvětlovanou proměnnou y. Pokud však nulovou hypotézu nelze na dané hladině významnosti zamítnout, pak vysvětlující proměnné x_i nemají statisticky významný efekt na y.

ŘEŠENÝ PŘÍKLAD 5.1

Při zjišťování vlivů na pracovní neschopnost zaměstnanců 10 podniků byly získány následující údaje:

Průměrný věk (roky)	Podíl žen v počtu pracovníků (%)	Pracovní neschopnost (%)
37	55	4,4
33	32	0,7
46	59	7,6
34	36	1,8
25	18	0,1
32	47	3,4
38	22	1,6
40	36	3,5
32	29	3,3
41	38	4,7

a. Odhadněte parametry lineární regresní funkce popisující závislost pracovní neschopnosti na průměrném věku zaměstnanců a na podílu žen mezi zaměstnanci.

b. Pomocí koeficientu determinace charakterizujte přiléhavost daného regresního modelu k datům.

c. Jak se změní pracovní neschopnost zaměstnanců, zvýší-li se jejich průměrný věk o 2 roky při stejném podílu žen?

d. Určete 95% intervaly spolehlivosti pro regresní koeficienty b_0, b_1, b_2.
Na hladině významnosti $\alpha = 0{,}01$ testujte hypotézu $\beta_1 = \beta_2 = 0$.

Řešení:

a. Naším úkolem je nalézt regresní koeficienty b_0, b_1, b_2 regresní funkce

$Y = b_0 + b_1X_1 + \; b_2X_2,$

kde X_1 je průměrný věk zaměstnanců,

X_2 je podíl žen v počtu zaměstnanců.

Regresní koeficienty b_0, b_1, b_2 vypočítáme pomocí metody nejmenších čtverců. Využijeme přitom nejprve maticové symboliky, kterou jsme použili v textu.

$$\mathbf{F} = \begin{bmatrix} 1 & 37 & 55 \\ 1 & 33 & 32 \\ 1 & 46 & 59 \\ 1 & 34 & 36 \\ 1 & 25 & 18 \\ 1 & 32 & 47 \\ 1 & 38 & 22 \\ 1 & 40 & 36 \\ 1 & 32 & 29 \\ 1 & 41 & 38 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 4{,}4 \\ 07 \\ 7{,}6 \\ 1{,}8 \\ 0{,}1 \\ 3{,}4 \\ 1{,}6 \\ 3{,}5 \\ 3{,}3 \\ 4{,}7 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

Vektor **b** vypočítáme pomocí vztahu (5.7). Matice $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ a $\mathbf{F}^{\mathrm{T}}\mathbf{y}$ mají obecně tvar:

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix}, \; \mathbf{F}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{bmatrix}.$$

Hodnoty potřebné k výpočtu těchto matic jsou uvedeny v následující tabulce:

Pozorování	X_1	X_2	Y	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
1	37	55	4,4	1369	3025	2035	162,8	242,0
2	33	32	0,7	1089	1024	1056	23,1	22,4
3	46	59	7,6	2116	3481	2714	349,6	448,4
4	34	36	1,8	1156	1296	1224	61,2	64,8
5	25	18	0,1	625	324	450	2,5	1,8
6	32	47	3,4	1024	2209	1504	108,8	159,8
7	38	22	1,6	1444	484	836	60,8	35,2
8	40	36	3,5	1600	1296	1440	140,0	126,0
9	32	29	3,3	1024	841	928	105,6	95,7
10	41	38	4,7	1681	1444	1558	192,7	178,6
Σ	358	372	31,1	13128	15424	13745	1207,1	1374,7

Potom

$$\mathbf{F}^T\mathbf{F} = \begin{bmatrix} 10 & 358 & 372 \\ 358 & 13128 & 13745 \\ 372 & 13745 & 15424 \end{bmatrix} \quad \mathbf{F}^T\mathbf{y} = \begin{bmatrix} 31,1 \\ 1207,1 \\ 1374,7 \end{bmatrix}.$$

K matici $\mathbf{F}^T\mathbf{F}$ musíme vypočítat matici inverzní:

$$(\mathbf{F}^T\mathbf{F})^{-1} = \begin{bmatrix} 4,355 & -0,131 & -0,012 \\ -0,131 & 0,005 & -0,001 \\ -0,012 & -0,001 & 0,001 \end{bmatrix}.$$

Vektor **b** je výsledkem součinu matic $(\mathbf{F}^T\mathbf{F})^{-1}$ a $\mathbf{F}^T\mathbf{y}$:

$$(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y} = \begin{bmatrix} -6,59 \\ 0,18 \\ 0,09 \end{bmatrix}.$$

Hledaná regresní funkce má tvar: $Y = -6,59 + 0,18x_1 + 0,09x_2$.

b. K tomu, abychom vypočítali determinační koeficient, musíme znát hodnotu teoretického součtu čtverců S_T a celkového součtu čtverců S_y. Tyto součty vypočítáme podle vztahů (5.17), (5.16). Pro výpočet teoretického součtu musíme pro každé x_{1i}, x_{2i}, $i = 1,\ldots, 10$, znát teoretickou hodnotu Y_i, $i = 1,\ldots,10$, např. Y_1 vypočítáme takto:

$$Y_1 = -6,59 + 0,18x_{11} + 0,09x_{22} = -6,59 + 0,18\cdot 37 + 0,09\cdot 55 = 5,02$$

	X_1	X_2	y	Y	$(y-\bar{y})^2$	$(Y-\bar{y})^2$
1	37	55	4,4	5,02	1,664	3,648
2	33	32	0,7	2,23	5,808	0,774
3	46	59	7,6	7,00	20,160	15,132
4	34	36	1,8	2,77	1,716	0,116
5	25	18	0,1	–0,47	9,060	12,816
6	32	47	3,4	3,40	0,084	0,084
7	38	22	1,6	2,23	2,280	0,774
8	40	36	3,5	3,85	0,152	0,548
9	32	29	3,3	1,78	0,036	1,769
10	41	38	4,7	4,21	2,528	1,210
Součet	358	372	31,1	32,02	43,489	36,872

Tato hodnota udává, jaká by měla být teoreticky pracovní neschopnost při průměrném věku zaměstnanců téměř 37 let a podílu žen v počtu pracovníků 55%. Protože však jde o stochastickou závislost, liší se tato hodnota od skutečně zjištěné hodnoty $y = 4,4$. Všechny teoretické hodnoty Y_i jsou uvedeny v následující tabulce. Jednotliví sčítanci i hodnoty součtů S_y a S_T jsou rovněž uvedeni v tabulce.

Koeficient determinace vypočítáme dosazením do vztahu (3.20):

$$R^2 = \frac{S_T}{S_y} = \frac{36,87}{43,49} = 0,848.$$

Tato hodnota znamená, že pomocí regresní funkce $Y = -6,59 + 0,18x_1 + 0,09x_2$ je vysvětleno 84,8% celkové variability proměnné Y.

c. Velikost změny znaku Y je při změně znaku X_1 o jednotku rovna b_1. Má-li se tedy zvýšit průměrný věk o 2 roky při nezměněné zaměstnanosti žen X_2, zvýší se pracovní neschopnost o $2b_1$, tj. o 0,36%.

d. Obecný tvar těchto intervalů je následující (viz (3.25)):

$$[b_i - t_{1-\alpha/2}(n-p)\sqrt{\frac{S_R h_{ii}}{n-p}},\ b_i + t_{1-\alpha/2}(n-p)\sqrt{\frac{S_R h_{ii}}{n-p}}],$$

kde S_R je reziduální součet čtverců,
$t_{1-\alpha/2}(n-\mathrm{p})$ je kvantil t-rozdělení o $n-p$ stupních volnosti,
p je počet parametrů regresní funkce,
h_{ii} prvek matice $\mathbf{H} = (\mathbf{F}'\mathbf{F})^{-1}$.

Hodnotu S_R vypočítáme ze vztahu:

$$S_R = S_y - S_T = 43,49 - 36,87 = 6,62.$$

V tabulce t-rozdělení nalezneme $(1-\alpha/2) = 97,5\%$ kvantil t-rozdělení o $n - p = 10 - 3 = 7$ stupních volnosti:

$$t_{0,975}(7) = 2,365,$$

$$h_{00} = 4,355;\ h_{11} = 0,0051;\ h_{22} = 0,001,\ \mathbf{H} = \{h_{ij}\},\ i,j = 0,1,2.$$

Dosazením výše vypočítaných hodnot do vztahu pro interval spolehlivosti určíme jeho pravou a levou krajní hodnotu L a P:

Pro b_0, tj. $i = 0$:

$$L = 6,59 - 2,365\sqrt{\frac{6,62 \cdot 4,355}{7}} = 1,79,$$

$$P = 6,59 + 2,365\sqrt{\frac{6,62 \cdot 4,355}{7}} = 11,39.$$

95% interval spolehlivosti pro regresní koeficient b_0 je [1,79;11,39]. Pro b_1, tj. $i = 1$:

$$L = 0,18 - 2,365\sqrt{\frac{6,62 \cdot 0,0051}{7}} = 0,016,$$

$$P = 0,18 + 2,365\sqrt{\frac{6,62 \cdot 0,0051}{7}} = 0,344.$$

Pak 95% interval spolehlivosti pro regresní koeficient b_1 je [0,016; 0,344].

Pro b_2, tj. $i = 2$:

$$L = 0,09 - 2,365\sqrt{\frac{6,62 \cdot 0,001}{7}} = 0,017,$$

$$P = 0,09 + 2,365\sqrt{\frac{6,62 \cdot 0,001}{7}} = 0,163.$$

Potom 95% interval spolehlivosti pro regresní koeficient b_2 je [0,017; 0,163].

e. Pro ověření hypotézy použijeme F-test. Budeme testovat nulovou hypotézu:

H_0: $\beta_1 = \beta_2 = 0$

proti alternativní hypotéze

H_1: alespoň jedno β_i je různé od nuly.

K ověření nulové hypotézy použijeme testové kritérium (3.34):

$$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}} = \frac{\frac{36{,}87}{2}}{\frac{6{,}62}{7}} = 19{,}49.$$

V tabulce F-rozdělení najdeme (1-α)% kvantil F-rozdělení o $p-1$ a $n-p$ stupních volnosti:

$F_{1-0{,}01}(2{,}7) = 9{,}55$.

Protože je 19,49 > 9,55, zamítáme nulovou hypotézu ve prospěch alternativní hypotézy, což znamená, že regresní parametry jsou vesměs nenulové, a tudíž existuje statisticky významná závislost Y na X_1 a nebo X_2.

Řešení v Excelu.

Regresní statistika	
Násobné R	0,912
Hodnota spolehlivosti R (koeficient determinace)	0,831
Nastavená hodnota spolehlivosti R	0,783
Chyba stř. hodnoty	1,024
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	36,155	18,078	17,255	0,002
Rezidua	7	7,334	1,048		
Celkem	9	43,489			

e) Protože hodnota Významnost F je menší než hladina významnosti 0,01; nulovou hypotézu zamítáme, tzn. že regresní parametry jsou vesměs nenulové.

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	-6,595	2,136	-3,087	0,018	-11,645	-1,544
průměrný věk X1	0,178	0,073	2,441	0,045	0,006	0,351
podíl žen (%) X2	0,089	0,032	2,758	0,028	0,013	0,166

ŘEŠENÝ PŘÍKLAD 5.2

Následující tabulka obsahuje údaje o tržbách, velikosti výdajů na reklamu a o počtu obchodních zástupců pro 11 firem zabývajících se nákupem a prodejem:

Reklamní výdaje (tis. Kč)	Obchodní zástupci	Objem prodeje (mil. Kč)
180	35	260
230	38	310
260	33	280
240	40	300
280	38	340
300	32	380
340	42	410
320	49	440
360	53	400
380	55	430
260	33	310

a. Popište závislost objemu produkce na reklamních výdajích a na počtu obchodních zástupců dvourozměrný lineárním regresním modelem.
b. F-testem posuďte významnost tohoto regresního modelu. Uvažujte hladinu významnosti $\alpha = 0{,}01$.
c. Na hladině významnosti α = 0,01 testujte individuální významnost regresního parametru β_1.
d. Jaký objem produkce lze očekávat, vydá-li firma na reklamu 450 tis. Kč a současně bude mít 50 obchodních zástupců? Určete bodový odhad objemu produkce.

Řešení:

Řešení v Excelu.

Regresní statistika		
Násobné R	0,916	
Hodnota spolehlivosti R	0,839	koeficient determinace
Nastavená hodnota spolehlivosti R	0,799	
Chyba stř. hodnoty	28,434	
Pozorování	11	

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	33822,799	16911,399	20,917	0,001
Rezidua	8	6468,110	808,514		
Celkem	10	40290,909			

b) Hodnota Významnost F je menší než 0,01;
model je zvolen správně,
zamítáme nulovou hypotézu o nulovosti obou koeficientů

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*
Hranice	63,830	47,652	1,340	0,217
reklamní výdaje (tis.Kč)	0,849	0,224	3,789	0,005
obchodní zástupci	1,076	1,656	0,650	0,534

a) $Y = 63{,}83 + 0{,}85.x1 + 1{,}08.x2$

c) Koeficient $b1 = 0{,}849$ je statisticky významný na hladině významnosti 0,01; protože Hodnota P je mešní než 0,01.

d) 500,33 mil.Kč

5.9 SAMOSTATNÉ ÚKOLY

5.1 Firma sledovala, jak jsou její tržby ovlivněny výdaji na reklamu v různých sdělovacích prostředcích. Výsledky průzkumu jsou uvedeny v následující tabulce.

Rádio, TV (tis. Kč)	Noviny, časopisy (tis. Kč)	Tržby (tis. Kč)
0	16	254
22	29	765
28	30	864
33	35	1001
39	27	911
41	36	1121
49	0	856
55	12	932
60	23	1152
63	34	1403
68	54	1702

a. Určete jednoduchý lineární regresní model popisující závislost obratu na velikosti prostředků vydaných na reklamu v novinách a časopisech.
b. Určete dvourozměrný lineární regresní model popisující závislost obratu na velikosti prostředků vydaných na reklamu v novinách a časopisech a na velikosti prostředků vydaných na reklamu v rozhlase a v televizi.
c. Pomocí F–testu rozhodněte, je-li vhodné k popisu závislosti používat zvolený vícenásobný lineární model. Uvažujte hladinu významnosti $\alpha = 0{,}05$.
d. Přispělo významně zavedení další vysvětlující proměnné k zlepšení výstižnosti modelu?

e. Jaký obrat je možné očekávat, vydá-li se na reklamu v tisku 32 tis. Kč a na reklamu v rozhlase a televizi 47 tis. Kč? Proveďte bodový odhad.

5.2 Mezinárodní organizace WHO zjistila údaje o dětské úmrtnosti (v promile) - DÚ, gramotnosti žen (v procentech) - GŽ a HDP na hlavu (v dolarech) - HDP u 64 rozvojových zemí:

DÚ	GŽ	HDP	DÚ	GŽ	HDP
128	37	1870	142	50	8640
204	22	130	104	62	350
202	16	310	287	31	230
197	65	570	41	66	1620
96	76	2050	312	11	190
209	26	200	77	88	2090
170	45	670	142	22	900
240	29	300	262	22	230
241	11	120	215	12	140
55	55	290	246	9	330
75	87	1180	191	31	1010
129	55	900	182	19	300
24	93	1730	37	88	1730
165	31	1150	103	35	780
94	77	1160	67	85	1300
96	80	1270	143	78	930
148	30	580	83	85	690
98	69	660	223	33	200
161	43	420	240	19	450
118	47	1080	312	21	280
269	17	290	12	79	4430
189	35	270	52	83	270
126	58	560	79	43	1340
12	81	4240	61	88	670
167	29	240	168	28	410
135	65	430	28	95	4370
107	87	3020	121	41	1310
72	63	1420	115	62	1470
128	49	420	186	45	300
27	63	19830	47	85	3630
152	84	420	178	45	220
224	23	530	142	67	560

a. Určete lineární regresní model popisující závislost dětské úmrtnosti na gramotnosti žen a HDP v rozvojových zemích.

b. Pomocí F–testu rozhodněte, je-li vhodné k popisu závislosti používat zvolený vícenásobný lineární model. Uvažujte hladinu významnosti $\alpha = 0{,}05$.

c. Jsou regresní koeficienty modelu statisticky významné? Stanovte jejich intervaly spolehlivosti pro hladinu významnosti $\alpha = 0{,}10$.

d. Pomocí koeficientu determinace určete přiléhavost dat k modelu. Jak se změní dětská úmrtnost při zvýšení HDP o 1000 USD při stejném stupni negramotnosti žen? Naopak: jak se změní dětská úmrtnost při zvýšení gramotnosti žen o 1 procento při stejné úrovni HDP?

5.10 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

5.1 a) jednoduchý lineární regresní model

Regresní statistika	
Násobné R	0,658
Hodnota spolehlivosti R	0,433
Nastavená hodnota spolehlivosti R	0,370
Chyba stř. hodnoty	292,354
Pozorování	11

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	587103,478	587103,478	6,869	0,028
Rezidua	9	769235,250	85470,583		
Celkem	10	1356338,727			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*
Hranice	538,482	195,714	2,751	0,022
Noviny,časopisy (tis.Kč)	17,019	6,494	2,621	0,028

$Y = 539{,}5 + 17{,}2.x$

b) dvourozměrný lineární regresní model

Regresní statistika	
Násobné R	0,992
Hodnota spolehlivosti R	0,985
Nastavená hodnota spolehlivosti R	0,981
Chyba stř. hodnoty	50,634
Pozorování	11

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	2	1335828,082	667914,041	260,514	0,000
Rezidua	8	20510,645	2563,831		
Celkem	10	1356338,727			

	Koeficienty	Chyba stř. hodnoty	t Stat	Hodnota P
Hranice	87,214	42,969	2,030	0,077
Rádio,TV (tis.Kč)	13,905	0,814	17,089	0,000
Noviny,časopisy (tis.Kč)	12,275	1,158	10,596	0,000

$$Y = 87{,}21 + 13{,}9.x1 + 12{,}27.x2$$

c) Ano, hodnota Významnost F je menší než 0,05; proto vícenásobný lineární model je vhodný.

d) Ano, koeficient determinace se z hodnoty 0,43 zvýšil na hodnotu 0,98.

e) 1 133,15 tis.Kč = 1 133 150 Kč

5.2 a)

Regresní statistika	
Násobné R	0,841
Hodnota spolehlivosti R	0,708
Nastavená hodnota spolehlivosti R	0,698
Chyba stř. hodnoty	41,748
Pozorování	64

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	2	257362,373	128681,187	73,833	0,000
Rezidua	61	106315,627	1742,879		
Celkem	63	363678,000			

	Koeficienty	Chyba stř. hodnoty	t Stat	Hodnota P	Dolní 90,0%	Horní 90,0%
Hranice	263,642	11,593	22,741	0,000	244,278	283,005
GŽ	-2,232	0,210	-10,629	0,000	-2,582	-1,881
HDP	-0,006	0,002	-2,819	0,006	-0,009	-0,002

$$Y = 263{,}64 - 2{,}23.x1 - 0{,}006.x2$$

b) Ano, hodnota Významnost F je menší než 0,05; proto vícenásobný lineární model je vhodný.

c) Oba regresní koeficienty jsou statisticky významné, protože Hodnota P je menší než 0,1. Intervaly spolehlivosti: $b1 \in (-2{,}5;\ -1{,}8)$; $b2 \in (-0{,}009;\ -0{,}002)$

d) Koeficient determinace je roven 0,71; tzn., že 71% celkové variability je vysvětleno modelem.

e) Při zvýšení HDP o 1000 USD při stejném stupni negramotnosti žen klesne dětská úmrtnost o 5,6 promile. Při zvýšení gramotnosti žen o 1% při stejné úrovni HDP klesne dětská úmrtnost o 0,22 promile.

6 REGRESNÍ ANALÝZA – VÍCEROZMĚRNÁ: MULTIKOLINEARITA, HETEROSKEDASTICITA, AUTOKORELACE

RYCHLÝ NÁHLED KAPITOLY

V této kapitole se naučíte identifikovat, analyzovat a odstraňovat problémy, které způsobuje nesplnění hlavních předpokladů klasického vícerozměrného lineárního regresního modelu formulované v kapitole 5.4: multikolinearita, heteroskedasticita a autokorelace. Multikolinearitou tedy rozumíme vzájemnou statistickou závislost, tj. korelaci, mezi vysvětlujícími proměnnými ve vícenásobném lineárním regresním modelu. Další důležitou vlastností klasického lineárního regresního modelu je homoskedasticita. Jde o vlastnost (5.15), která spočívá v tom, že rozptyl poruchy ε_i v populačním lineárním regresním modelu je konstantní. Autokorelace je korelace mezi pozorováními uspořádanými v čase (data jsou časové řady) nebo v prostoru (data jsou průřezová, tj. v jednom časovém okamžiku/intervalu). Říkáme, že v regresním modelu není přítomná autokorelace, jestliže náhodné veličiny jsou vzájemně nekorelované.

6.1 CO JE MULTIKOLINEARITA?

Multikolinearitou tedy rozumíme vzájemnou statistickou závislost, tj. korelaci, mezi vysvětlujícími proměnnými ve vícenásobném lineárním regresním modelu:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \,. \qquad (6.1)$$

Informaci o této vzájemné závislosti poskytuje matice výběrových korelačních koeficientů:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}. \qquad (6.2)$$

Zřejmě je matice (6.2) symetrická, tj. $r_{ij} = r_{ji}$ pro všechna *i,j*. Pokud jsou všechny dvojice vysvětlujících proměnných vzájemně nekorelované, potom platí, že $r_{ij} = r_{ji} = 0$, tj. $\mathbf{R} = \mathbf{I}$, čili $\mathbf{R}$ je jednotkovou maticí.

Uvědomte si, že na diagonále matice **R** musejí být všechny prvky rovny 1, neboť korelace vektoru dat se sebou samým je vždy rovna 1! Jsou-li však alespoň některé nediagonální prvky matice **R** nenulové, hovoříme o *multikolinearitě*. Matice **R** pak není jednotkovou maticí a její determinant je menší než 1. Je-li multikolinearita vysoká, hovoříme o *škodlivé multikolinearitě*, pak se determinant matice **R** blíží k nule. V tom případě dává metoda nejmenších čtverců odhady regresních koeficientů s širokými intervaly spolehlivosti, takže výsledky jsou prakticky neupotřebitelné.

Na to, kdy je multikolinearita „škodlivá“, existují různé názory, opírající se víceméně o zkušenost. Někteří autoři považují za škodlivou multikolinearitu, když alespoň jeden nediagonální prvek matice **R** je větší než 0,8.

Zjistí-li se škodlivá multikolinearita, je možno postupovat v zásadě dvojím způsobem. Buď vysvětlující proměnnou, která je zdrojem multikolinearity, vypustíme z modelu, nebo doplníme data, eventuálně získáme nový vzorek dat. Škodlivá multikolinearita je totiž často důsledkem „špatného" vzorku dat. Projevuje se obvykle vysokým koeficientem determinace (blízkým k 1) a zároveň jsou individuální koeficienty statisticky nevýznamné (t-test), model jako celek je naopak statisticky významný (F-test), viz kap. 5.7 a 5.8.
Celou záležitost ilustrujeme na příkladu.

ŘEŠENÝ PŘÍKLAD 6.1

V následující tabulce jsou uvedeny měsíční výdaje, měsíční příjmy a majetek (v Kč) u 10 českých rodin. Proveďte regresní analýzu měsíčních výdajů rodin v závislosti na měsíčních příjmech a majetku. Vysvětlete dosažené výsledky pomocí jednorozměrné regrese.

Y vydaje	X1 prijmy	X2 majetek
8400	9600	100000
7800	12000	120000
10800	14400	150000
11400	16800	170000
13200	19200	200000
13800	21600	225000
14400	24000	246000
16800	26400	264000
18600	28800	292000
18000	31200	322000

Řešení:

Data z Ta. 6.1 uložíme v excelovské tabulce. Známým postupem v menu:

Data → Analýza dat... → Regrese

získáme po vyplnění příslušných políček tento výsledek:

VÝSLEDEK

Regresní statistika	
Násobné R	0,981
Hodnota spolehlivosti R	0,962
Nastavená hodnota spol	0,951
Chyba stř. hodnoty	832,660
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*ýznamnost F*
Regrese	2	1,23E+08	61581370	88,82062	1,06E-05
Rezidua	7	4853260	693322,9		
Celkem	9	1,28E+08			

	Koeficienty	*ba stř. hod*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	2943,676	832,579	3,536	0,010	974,940	4912,413
X1 prijmy	0,569	0,847	0,672	0,523	-1,433	2,571
X2 majetek	-0,006	0,083	-0,071	0,946	-0,203	0,191

V tomto výstupu se vyskytují zdánlivě paradoxní výsledky. Z Tabulky ANOVA vyplývá, že regresní model

$$y = 2943{,}676 + 0{,}569x_1 - 0{,}006x_2 + \varepsilon$$

je jako celek statisticky významný (F-test), zatímco individuální regresní koeficienty u proměnných „příjmy" resp. „majetek" jsou statisticky nevýznamné, neboť obě odpovídající p-hodnoty (signifikance) jsou větší než 0,05 (0,672 resp. 0,946). Koeficient determinace $R^2 = 0{,}962$ je vysoký – blízký k 1, což svědčí o vysoké přiléhavosti dat k modelu. Navíc je u regresního koeficientu u proměnné x_2 záporné znaménko, což je evidentně v rozporu s intuicí, která říká: čím je větší majetek, tím je vyšší spotřeba rodiny. Tento zdánlivý rozpor je způsoben kolinearitou regresorů, o čemž svědčí jejich korelační matice

$$\mathbf{R} = \begin{bmatrix} 1{,}000 & 0{,}999 \\ 0{,}999 & 1{,}000 \end{bmatrix},$$

kterou lze snadno zjistit tak, že vypočítáte $r_{12} = r_{21} = 0{,}999012$ pomocí excelovské funkce =CORREL(B4:B13;C4:C13), za předpokladu, že data pro x_1 jsou uložena v oblasti B4:B13, data pro x_2 jsou uložena v oblasti C4:C13. Vysvětlující proměnné x_1 a x_2 jsou kolineární, neboť koeficient korelace $r_{12} = r_{21} = 0{,}999012$ je blízký k 1.

Vypustíme-li nyní jednu z vysvětlujících proměnných, např. x_2 – majetek, a provedeme-li (jednoduchou) regresi x_1 na y, obdržíme s analogickým využitím Excelu tento výsledek:

VÝSLEDEK

Regresní statistika	
Násobné R	0,981
Hodnota spolehlivosti R	0,962
Nastavená hodnota spole	0,957
Chyba stř. hodnoty	779,160
Pozorování	10

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	1	1,23E+08	1,23E+08	202,8679	5,75275E-07
Rezidua	8	4856727	607090,9		
Celkem	9	1,28E+08			

	Koeficienty	ba stř. hodı	t stat	Hodnota P	Dolní 95%	Horní 95%
Hranice	2934,545	769,658	3,813	0,005	1159,710	4709,381
X1 prijmy	0,509	0,036	14,243	0,000	0,427	0,592

Vidíte, že v novém regresním modelu je regresní koeficient statisticky významný, neboť odpovídající p-hodnota (signifikance) je menší než 0,05 (0,000...), což je ve shodě s tabulkou ANOVA.

Podobně, vypustíme-li nyní vysvětlující proměnnou x_1 – příjem, a provedeme-li (jednoduchou) regresi x_2 na y, obdržíme s analogickým využitím Excelu výsledek z následujícího výstupu. Opět vidíte, že v novém regresním modelu je regresní koeficient statisticky významný, neboť odpovídající p-hodnota (signifikance) je menší než 0,05 (0,000...), což je ve shodě s tabulkou ANOVA. Navíc je znaménko u regresního koeficientu 0,050 kladné, což je v souhlasu s intuicí, že totiž velikost spotřeby je přímo úměrná velikosti majetku.

VÝSLEDEK

Regresní statistika	
Násobné R	0,979614
Hodnota spolehlivosti R	0,959644
Nastavená hodnota spole	0,954599
Chyba stř. hodnoty	803,6024
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	1,23E+08	1,23E+08	190,2357	7,37266E-07
Rezidua	8	5166214	645776,8		
Celkem	9	1,28E+08			

	Koeficienty	*ba stř. hodı*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	2880,627	798,404	3,608	0,007	1039,503	4721,750
X2 majetek	0,050	0,004	13,793	0,000	0,042	0,058

6.2 CO JE HETEROSKEDASTICITA?

Další důležitou vlastností klasického lineárního regresního modelu je *homoskedasticita*. Jde o vlastnost (5.15), která spočívá v tom, že rozptyl poruchy ε_i v populačním lineárním regresním modelu je konstantní, tj. v modelu

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i,\ i = 1,2,\ldots,n, \qquad (5.1)$$

platí podmínka

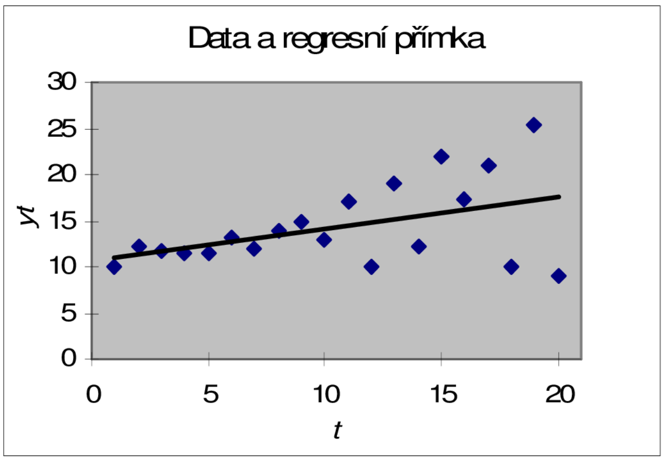
$$\boldsymbol{Var}(\boldsymbol{\varepsilon}) = \sigma^2\, \mathbf{I}\,, \qquad (5.15)$$

kde symbol **I** označuje jednotkovou matici.
Podmínku (5.15) je možné ekvivalentně vyjádřit také takto

$$\boldsymbol{E}(\varepsilon_i^2) = \sigma^2\,, \quad i = 1,2,\ldots,n, \qquad (6.3)$$

kde ***E*** je známý operátor střední hodnoty.

Pokud podmínka (5.15) není splněna, potom hovoříme o *heteroskedasticitě*. Příkladem heteroskedasticity v případě jednorozměrného lineárního regresního modelu je na Obr. 6.1. Je zřejmé, že rozptyl hodnoty y se zvětšuje s rostoucí hodnotou x.



Obr. 6.2. *Případ heteroskedasticity v regresním modelu*

Heteroskedasticita může být způsobena různými příčinami. Častou příčinou heteroskedasticity je fakt, že při postupném sběru dat se technika sběru postupně zlepšuje a chyba se proto zmenšuje. Naopak se chyba zvětšuje s přítomnosti odlehlých hodnot. Dalším zdrojem heteroskedasticity je nesprávná specifikace modelu, např. tím, že jsou opominuty důležité vysvětlující proměnné regresního modelu. Přítomnost heteroskedasticity v regresním modelu je silně nežádoucí, a to zejména z těchto důvodů:

- Přítomnost heteroskedasticity způsobuje neplatnost odhadů rozptylů regresních koeficientů a tudíž také odhadů jejich intervalů spolehlivosti a testů hypotéz o jejich statistické významnosti atd., viz kap. 5.6.
- Prognózy s využitím regresního modelu obsahujícího heteroskedasticitu jsou často nespolehlivé a dokonce nerealistické.

6.2.1 JAK ZJIŠŤOVAT HETEROSKEDASTICITU?

Jak poznáme, že v regresním modelu, který jsme sestavili na základě nějakých dat, je přítomna heteroskedasticita? Podobně jako v případě multikolinearity neexistují přesná pravidla, jak detekovat přítomnost heteroskedasticitu, pouze pár heuristických zásad.

Velmi často poznáme přítomnost heteroskedasticity z věcné povahy problému. Například je známo, že s rostoucím věkem zaměstnanců se zvětšuje rozptyl jejich platů. Ať je typ závislosti platu na věku lineární nebo ne, bude v modelu přítomna heteroskedasticita.

Pokud však nemáme podobné předběžné empirické informace o povaze problému, předpokládáme, že heteroskedasticita není přítomna, že tudíž je rozptyl náhodné složky modelu konstantní. Takové tvrzení pak můžeme podrobit zkoumání např. grafické analýze nebo statistickému testu reziduí e_i. S oběma postupy se zde seznámíte.

Grafická analýza

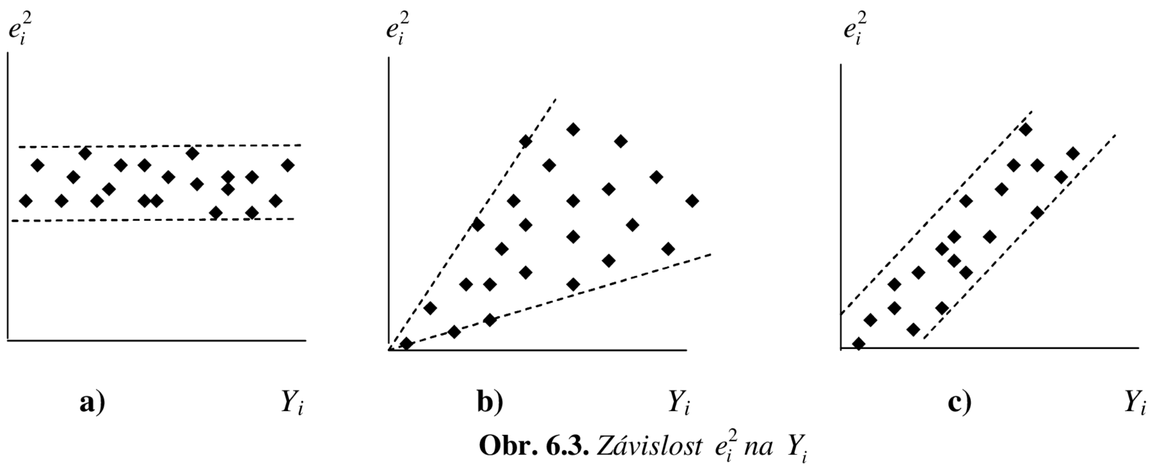
Zobrazíme si závislost kvadrátu reziduí e_i^2 na teoretické hodnotě Y_i. Na Obr. 6.3 jsou zobrazeny 4 důležité případy tvaru, které mohou nastat, kde

$$Y_i = b_0 + b_1 f_1(x_{i1}, x_{i2}, ..., x_{ik}) + ... + b_k f_k(x_{i1}, x_{i2}, ..., x_{ik}), \qquad (6.4)$$

přitom b_i jsou odhady regresních parametrů získané MNČ,

$$e_i = y_i - Y_i \qquad (6.5)$$

je reziduum, tj. odhad náhodné složky ε_i.



Obr. 6.3. *Závislost e_i^2 na Y_i*

Na Obr. 6.3 a) hodnota e_i^2 v zásadě nezávisí na Y_i, což naznačuje, že náhodná složka je konstatntní a tudíž heteroskedasticita není přítomna. Na druhou stranu Obr. 6.3 b) a c) hodnota e_i^2 v zřejmě závisí na Y_i, což naznačuje přítomnost heteroskedasticity. Konkrétní tvar závislosti vám dobře potvrdí zobrazení bodového diagramu závislosti y_i na vybrané datové hodnoty j-té vysvětlující proměnné x_{ji}.

Testy heteroskedasticity

Detekce heteroskedasticity s pomocí statistického testu hypotézy je obvykle založena na nulové hypotéze, že rozptyly náhodné složky ε_i^2 jsou konstantní, přičemž se analyzují jejich odhady, tj. rezidua e_i^2. V literatuře můžete nalézt podrobné testy heteroskedasticity s názvy jako Parkův test, Glejserův test, Goldfeld-Quandtův test aj., viz např. Gujarati (2003). Tyto statistické testy lze provádět pomocí specializovaných statistíckých programů, např. SPSS, v Excelu specializované funkce na tyto testy bohužel chybí. My si zde proto ukážeme tzv. Bartletův test heteroskedasticity, který představuje zjednodušený Goldfeld-Quandtův test a lze k jeho provedení využít funkce Excelu.

Bartletův test

Test vychází z rozdělení dat podle velikosti (některé) vysvětlující proměnné – označíme ji X - do dvou částí: $x_i \leq \hat{x}$ a $x_i > \hat{x}$, přitom jsou data uspořádána podle X, $\hat{x}$ je medián z x_i .

- Testuje se hypotéza o rovnosti rozptylů reziduí v obou částech (v Excelu: Analýza dat, Dvouvýběrový F-test pro rozptyl,…)
- Pokud se hypotéza o rovnosti rozptylu reziduí v obou částech zamítá, potom se hypotéza o konstantnosti rozptylu náhodné složky neboli hypotéza o přítomnosti heteroskedasticity, přijímá (a obráceně).

Použití Bartletova testu si ukážeme na příkladu. Ještě předtím se budeme zabývat otázkou, jak odstranit zjištěnou heteraskedasticitu, tj. jak modifikovat původní model tak, aby heteraskedasticitu neobsahoval.

6.2.2 JAK ODSTRAŇOVAT HETEROSKEDASTICITU?

Nejznámější metodou k odstranění heteroskedasticity je *metoda vážených nejmenších čtverců* MVNČ. V MVNČ předpokládáme určitý typ nekonstantního chování rozptylu náhodné složky.

Předpoklad 1: Rozptyl náhodné složky je přímo úměrný kvadrátu vysvětlující proměnné x, tj.

$$\boldsymbol{E}(\varepsilon_i^2) = \sigma^2 x_i^2 \text{ , } i = 1,2,...,n. \tag{6.6}$$

Transformovaný regresní model získáme tak, že regresní rovnici

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{, } i = 1,2,...,n, \tag{6.7}$$

vydělíme hodnotou x_i , čímž obdržíme

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i} = \beta_0 \frac{1}{x_i} + \beta_1 + \delta_i \text{, } i = 1,2,...,n, \tag{6.8}$$

kde pro novou náhodnou chybu δ_i platí po dosazení z (6.6)

$$E(\delta_i^2) = E(\frac{\varepsilon_i^2}{x_i^2}) = \sigma^2,\ i = 1,2,...,n. \qquad (6.9)$$

Provedením transformace $y_i' = \frac{y_i}{x_i},\ x_i' = \frac{1}{x_i},\ i = 1,2,...n.$ (6.10)

obdržíme z (6.8) nový regresní model

$$y_i' = \beta_1 + \beta_0 x_i' + \delta_i,\ i = 1,2,...,n. \qquad (6.11)$$

což je nový lineární regresní model podle (6.9) však bez heteroskedasticity.

Uvažovali jsme jednoduchý regresní model, avšak rozšíření výše uvedeného postupu na vícerozměrný regresní model je snadné. Předpoklad 1 modifikujeme tak, že rozptyl náhodné složky je přímo úměrný kvadrátu vysvětlující proměnné x_j, tj.

$$\boldsymbol{E}(\varepsilon_i^2) = \sigma^2 x_{ij}^2, \qquad i = 1,2,...,n. \qquad (6.6)$$

Namísto modelu (6.7) uvažujeme model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \varepsilon_i,\ i = 1,2,...,n. \qquad (6.7^*)$$

Pro nový vícerozměrný regresní model použijeme namísto transformace (6.10) nová transformovaná data

$$y_i' = \frac{y_i}{x_{ij}},\ x_{ij}' = \frac{1}{x_{ij}}, x_{ik}' = \frac{x_{ik}}{x_{ij}}, k \neq j,\ i = 1,2,...,n. \qquad (6.10^*)$$

Předpoklad 2: Rozptyl náhodné složky je přímo úměrný vysvětlující proměnné x, tj.

$$\boldsymbol{E}(\varepsilon_i^2) = \sigma^2 x_i, \qquad i = 1,2,...,n. \qquad (6.12)$$

Transformovaný regresní model získáme tak, že regresní rovnici

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,2,...,n, \qquad (6.13)$$

vydělíme hodnotou $\sqrt{x_i}$, čímž obdržíme

$$\frac{y_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1\sqrt{x_i} + \frac{\varepsilon_i}{\sqrt{x_i}} = \beta_0\frac{1}{\sqrt{x_i}} + \beta_1\sqrt{x_i} + \vartheta_i,\ i = 1,2,...,n, \qquad (6.14)$$

kde pro novou náhodnou chybu δ_i platí po dosazení z (6.12)

$$E(\vartheta_i^2) = E(\frac{\varepsilon_i^2}{x_i}) = \sigma^2,\ i = 1,2,...,n. \qquad (6.15)$$

Provedením transformace $y_i' = \frac{y_i}{\sqrt{x_i}},\ x_i' = \frac{1}{\sqrt{x_i}}, x_i'' = \sqrt{x_i},\ i = 1,2,...,n.$ (6.16)

obdržíme z (6.16) nový regresní model

$$y_i' = \beta_0 x_i' + \beta_1 x_i'' + \vartheta_i,\ i = 1,2,...,n, \qquad (6.17)$$

což je nový lineární regresní model bez úrovňové konstanty podle (6.15) však bez heteroskedasticity. Rozšíření na vícerozměrný regresní model je možné udělat analogicky jako v případě Předpokladu 1. Odstranění heteroskedasticity si prakticky vyzkoušíte v následujícím řešeném příkladu.

ŘEŠENÝ PŘÍKLAD 6.2

V následující tabulce jsou uvedeny příjmy a spotřební výdaje 30 rodin v tis. Kč/rok. Vytvořte lineární regresní model závislosti výdajů na příjmech, graficky a statistickým testem zjistěte přítomnost heteroskedasticity. Z původního modelu pak heteroskedasticitu odstraňte pomocí MVNČ. Použijte přitom Excel.

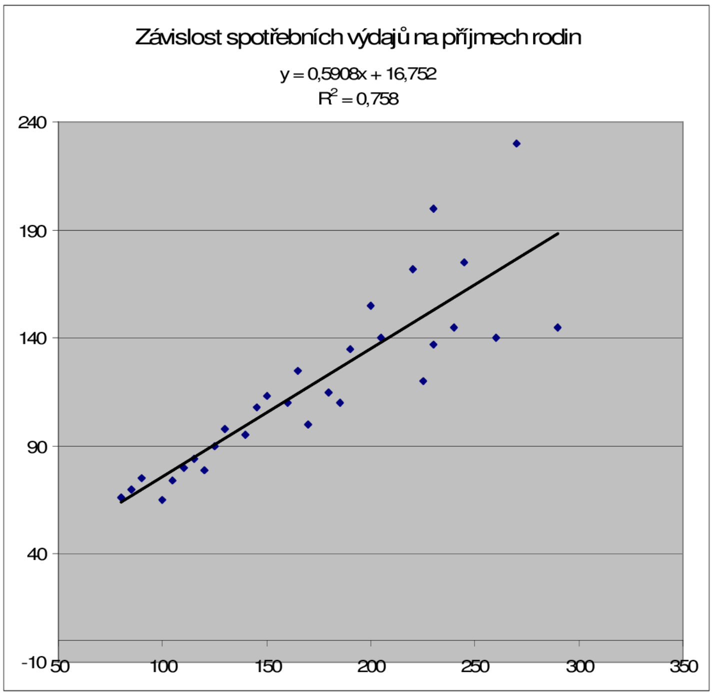
č.rodiny	Výdaje	Příjmy	č.rodiny	Výdaje	Příjmy
1	66	80	16	115	180
2	65	100	17	120	225
3	70	85	18	100	170
4	80	110	19	145	240
5	79	120	20	110	185
6	84	115	21	172	220
7	98	130	22	200	230
8	95	140	23	175	245
9	90	125	24	140	260
10	75	90	25	135	190
11	74	105	26	140	205
12	110	160	27	155	200
13	113	150	28	230	270
14	125	165	29	137	230
15	108	145	30	145	290

Řešení:

V Excelu vytvoříme z daných údajů graf: XY bodový a pomocí pravého tlačítka iniciujeme nabídku s volbou Přidat spojnici trendu... V podnabídce Možnosti zaklikneme 2 položky: Zvolit rovnici regrese a Zvolit koeficient spolehlivosti (tj. koeficient determinace). Obdržíme výsledek, z něhož vyplývá lineární regresní model:

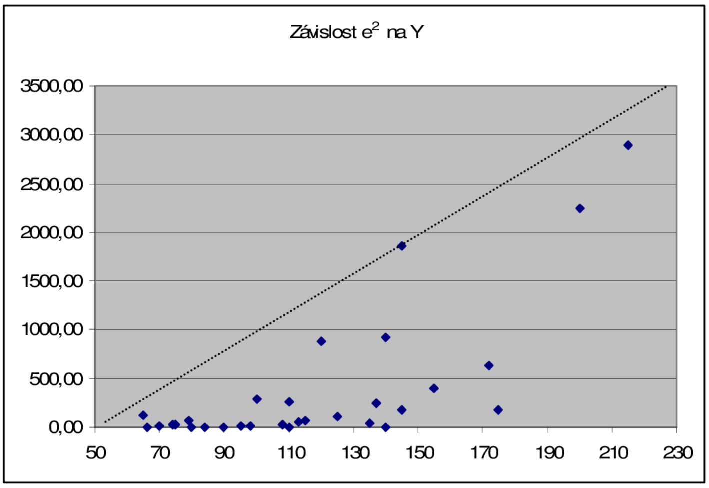
$$y = 9{,}29 + 0{,}64.x + \varepsilon,$$

viz následující graf.



Dále vedle sloupce y_i vytvoříme pomocí vzorce regresní rovnice sloupec teoretických hodnot Y_i. Další sloupec vytvoříme jako rozdíl sloupců y_i a Y_i, což bude sloupec reziduí. Poslední sloupec bude druhá mocnina reziduí. Společně pak vytvoříme XY bodový graf mezi Y_i a e_i^2.

Výsledkem je následující graf, který napovídá přítomnost heteroskedasticity, neboť body v grafu netvoří pás rovnoběžný s vodorovnou osou, jako na Obr. 6.3 a), ale spíše kužel, jako na Obr. 6.3 b).



Obr. 6.4. *Příklad: Kužel závislosti* e_i^2 *na* Y_i

K exaktnímu prokázání heteroskedasticity použijeme Bartletův test. Podle rostoucích hodnot *X* – Příjmů seřadíme hodnoty reziduí a z nich vytvoříme dva stejně velké soubory e1 a e2:

Příjmy	e1	Příjmy	e2
80	1,99	170	-8,09
85	-10,83	180	-29,68
90	3,03	185	-17,19
100	-1,74	190	-13,54
105	-8,65	200	-16,05
110	-0,69	205	25,28
115	4,45	220	47,37
120	-4,46	225	13,51
125	-0,60	230	-30,36
130	5,08	230	6,00
140	-4,78	240	2,14
145	-1,28	245	20,09
150	7,63	260	53,74
160	10,77	270	-15,63
165	5,58	290	-43,08

Budeme testovat, zda rozptyly obou souborů jsou stejné pomocí F-testu z Excelu:
V menu: Data → Analýza dat → Dvouvýběrový F-test pro rozptyl zadáme umístění oblastí sloupců e1 a e2, eventuální popisky a oblast výstupu. Obdržíme výstup:

Dvouvýběrový F-test pro rozptyl

	Soubor 1	*Soubor 2*
Stř. hodnota	0,366225	-0,366225
Rozptyl	35,88461	792,7791
Pozorování	15	15
Rozdíl	14	14
F	0,045264	
P(F<=f) (1)	3,89E-07	
F krit (1)	0,402621	

V tomto výstupu je důležitá P-hodnota: P(F<=f) (1) = 3,89 E-07 = 0,000000389 < 0,05. Na hladině $\alpha = 0{,}05$ proto nulovou hypotézu H_o: „Rozptyly obou uvažovaných souborů jsou stejné“ zamítáme. Uvažované soubory mají různý rozptyl, což znamená, že rozptyl náhodné složky regresního modelu není konstantní, neboli že heteroskedasticita je v modelu přítomna.

Nakonec ukážeme, jak přítomnou heteroskedasticitu odstranit. V Obr. 6.4 se body grafu nacházejí v „lineárním kuželu“, proto zvolíme pro transformaci Předpoklad 2.

Transformace podle (6.16): $y_i{'} = \dfrac{y_i}{\sqrt{x_i}}$, $x_i{'} = \dfrac{1}{\sqrt{x_i}}$, $x_i{''} = \sqrt{x_i}$, $i = 1,2,...,30$.

obdržíme nový regresní model

$$y_i{'} = 16{,}75x_i{'} + 0{,}59_1 x_i{''} + \vartheta_i,\ i = 1,2,...,\ 30,$$

který je bez heteroskedasticity.

č.rodiny	y´	x´	x´´	č.rodiny	y´	x´	x´´
1	7,379	0,112	8,944	16	8,572	0,075	13,416
2	6,500	0,100	10,000	17	8,000	0,067	15,000
3	7,593	0,108	9,220	18	7,670	0,077	13,038
4	7,628	0,095	10,488	19	9,360	0,065	15,492
5	7,212	0,091	10,954	20	8,087	0,074	13,601
6	7,833	0,093	10,724	21	11,596	0,067	14,832
7	8,595	0,088	11,402	22	13,188	0,066	15,166
8	8,029	0,085	11,832	23	11,180	0,064	15,652
9	8,050	0,089	11,180	24	8,682	0,062	16,125
10	7,906	0,105	9,487	25	9,794	0,073	13,784
11	7,222	0,098	10,247	26	9,778	0,070	14,318
12	8,696	0,079	12,649	27	10,960	0,071	14,142
13	9,226	0,082	12,247	28	13,997	0,061	16,432
14	9,731	0,078	12,845	29	9,034	0,066	15,166
15	8,969	0,083	12,042	30	8,515	0,059	17,029

6.3 CO JE AUTOKORELACE?

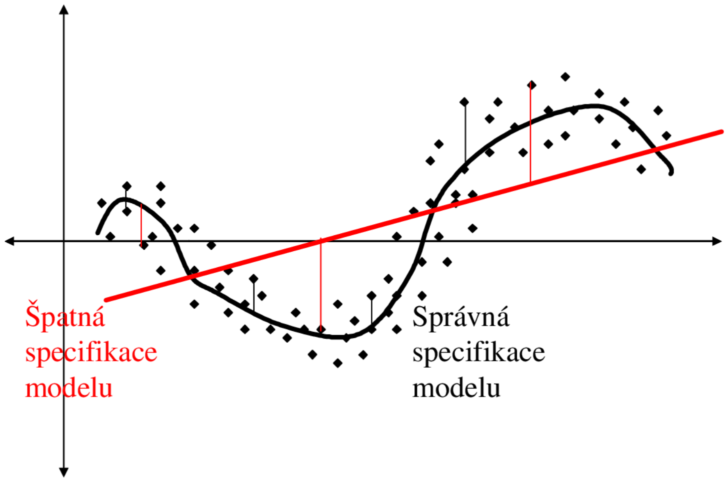
Autokorelace je korelace mezi pozorováními uspořádanými v čase (data jsou časové řady) nebo v prostoru (data jsou průřezová, tj. v jednom časovém okamžiku/intervalu). Říkáme, že v regresním modelu není přítomná autokorelace, jestliže náhodné veličiny jsou vzájemně nekorelované, symbolicky to lze vyjádřit takto

$$\boldsymbol{E}(\varepsilon_i.\varepsilon_j) = 0\ , \qquad i \neq j,\ i,j = 1,2,...,n. \qquad (6.18)$$

Jestliže naopak existuje dvojice indexů $i \neq j$, přičemž platí $\boldsymbol{E}(\varepsilon_i.\varepsilon_j) \neq 0$, řekneme, že v regresním modelu *je přítomna autokorelace*.

Autokorelace se nejčastěji vyskytuje v regresních modelech založených na datech ve formě časových řad. Potom indexy *i*, (resp. *j*) představují časové okamžiky *t*. Časovým řadám a jejich analýze se budou věnovat následující kapitoly 8 až 12, kde bude podrobněji pojednáno také o autokorelaci.

Následující obr. 6.5 dává příklad dvou regresních modelů dat, z nichž jeden je správně specifikován (nelineární regresní model – černá křivka), druhý je nesprávně specifikován (lineární regresní křivka – červená přímka). Nesprávná specifikace modelu způsobuje, že rezidua jsou vzájemně korelována, což se projevuje tak, že datové body leží vždy ve větší oblasti podél vodorovné osy na jedné straně regresní křivky, zatímco v případě nekorelovaných reziduí leží datové body rovnoměrně po obou stranách regresní křivky v celé oblasti vodorovné osy (tj. nezávisle proměnné).



Obr. 6.5. *Autokorelace: špatná a správná specifikace modelu*

6.4 Samostatné úkoly

6.1 V následující tabulce jsou uvedeny hodnoty obratu, výdajů na vědu a výzkum (VaV) a zisku za 18 průmyslových odvětví v USA v roce 1990. Vytvořte lineární regresní model závislosti zisku na obratu a výdajích na VaV. Zjistěte, zda je v modelu přítomna multikolinearita a heteroskedasticita. Použijte postupy, které jste se naučili v této kapitole.

Obrat	VaV	Zisk
6375,3	62,5	185,1
11626,4	92,9	1569,5
14655,1	178,3	276,8
21869,2	258,4	2828,1
26408,3	494,7	225,9
32405,6	1083,0	3751,9
35107,7	1620,6	2884,1
40295,4	421,7	4645,7
70761,6	509,2	5036,4
80552,8	6620,1	13869,9
95294,0	3918,6	4487,8
101314,1	1595,3	10278,9
116141,3	6107,5	8787,3
122315,7	4454,1	16438,8
141649,9	3163,8	9761,4
175025,8	13210,7	19774,5
230614,5	1703,8	22626,6
293543,0	9528,2	18415,4

6.5 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

6.1 $Y = 791{,}54 + 0{,}069.x1 + 0{,}369.x2$

$x1\ldots$obrat; $x2\ldots$výdaje na VaV; koeficient $b2 = 0{,}369$ není statisticky významný

Korelační koeficient = 0,692 je statisticky významný na hladině významnosti 0,01. V modelu je přítomna multikolinearita.

Závislost zisku na obratu:

$Y = 862{,}85 + 0{,}08.x1$

Koeficient 0,08 je statisticky významný.

Dvouvýběrový F-test pro rozptyl

	Soubor 1	*Soubor 2*
Stř. hodnota	-809,8808	809,8807591
Rozptyl	1219536	20761396,39
Pozorování	9	9
Rozdíl	8	8
F	0,058741	
P(F<=f) (1)	0,000289	
F krit (1)	0,290858	

Nulovou hypotézu: rozptyly obou souborů jsou stejné, <u>zamítáme</u>, rozptyl náhodné složky není konstantní, neboli heteroskedasticita je v modelu přítomna.

Závislost zisku na obratu:

$Y = 3817{,}11 + 1{,}4.x2$

Koeficient 1,4 je statisticky významný.

Dvouvýběrový F-test pro rozptyl

	Soubor 1	*Soubor 2*
Stř. hodnota	-1348,771	1348,770762
Rozptyl	7292620	43919891,06
Pozorování	9	9
Rozdíl	8	8
F	0,166044	
P(F<=f) (1)	0,010033	
F krit (1)	0,290858	

Nulovou hypotézu: rozptyly obou souborů jsou stejné, <u>nezamítáme</u>, rozptyl náhodné složky je konstantní, neboli heteroskedasticita není v modelu přítomna. (hladina významnosti 0,01)

7 FIKTIVNÍ PROMĚNNÉ

RYCHLÝ NÁHLED KAPITOLY

Jedním ze způsobů jak kvantifikovat přítomnost nebo nepřítomnost nějaké vlastnosti je konstrukce nových proměnných, které toto vyjadřují pomocí čísel 1 nebo 0, přitom 1 indikuje přítomnost a 0 nepřítomnost vlastnosti. Tyto proměnné, které se nazývají fiktivní proměnné (anglicky dummy variables), mají časté použití při analýze ekonomických problémů závislosti mezi ekonomickými veličinami. V této kapitole se naučíte použít fiktivní vysvětlující proměnné ve vícerozměrných regresních modelech. Fiktivní proměnné lze zavést do regresního modelu stejně snadno, jako jakékoliv jiné kvantitativní proměnné, které jsme např. použili v předchozí kapitole o vícenásobné regresní analýze. V této kapitole si ukážeme, jak vytvořit regresní model, který obsahuje výlučně fiktivní vysvětlující proměnné, jak tento model souvisí s dříve použitou metodou ANOVA, a také jak využít model, který je kombinací fiktivních a obvyklých kvantitativních vysvětlujících proměnných. Dále ukážeme, jak lze fiktivní proměnné výhodně použít pro sezónní data k identifikaci jednotlivých sezón a analýze problému závislostí mezi ekonomickými veličinami.

7.1 CO JSOU FIKTIVNÍ PROMĚNNÉ?

V regresní analýze je závisle proměnná (kritérium) ovlivňována často nejen kvantitativními proměnnými (regresory) jako například příjem rodin, ceny výrobků, náklady podniků atd., ale také proměnnými, které mají kvalitativní povahu, jako jsou pohlaví zákazníků, jejich národnost, vzdělání, region apod. Tyto proměnné obvykle představují přítomnost nebo naopak nepřítomnost nějaké „kvality“ nebo vlastnosti, jako jsou v případě pohlaví zákazníků muž nebo žena, v případě národnosti Čech nebo Slovák, v případě vzdělání základní (Z), středoškolské (S) nebo vysokoškolské (V) apod. Jedním ze způsobů jak kvantifikovat takové vlastnosti je konstrukce nových proměnných, které vyjadřují přítomnost nebo nepřítomnost příslušné vlastnosti pomocí čísel 1 nebo 0, přitom 1 indikuje přítomnost a 0 nepřítomnost vlastnosti. Tyto proměnné se nazývají *fiktivní proměnné* (anglicky „dummy variables“).

Fiktivní proměnné lze zavést do regresního modelu stejně snadno, jako jakékoliv jiné kvantitativní proměnné, které jsme např. použili v předchozí kapitole. Můžeme však vytvořit regresní model, který obsahuje výlučně fiktivní vysvětlující proměnné. S nimi jste se vlastně již setkali v 1. a 2. kapitole v rámci ANOVA.

7.2 FIKTIVNÍ PROMĚNNÉ A ANOVA

Nejprve budeme vyšetřovat situaci pouze s kvalitativní vysvětlující proměnnou, která nabývá K hodnot (kategorií) $Z_0, Z_1, Z_2, ..., Z_{K-1}$. Tuto kvalitativní proměnnou nahradíme $K-1$ fiktivními vysvětlujícími proměnnými: $d_1, ..., d_{K-1}$, definovanými takto:

$d_1 = 1$ pokud kvalitativní proměnná nabývá hodnoty Z_1,
$\quad = 0$ jinak.
$d_2 = 1$ pokud kvalitativní proměnná nabývá hodnoty Z_2,
$\quad = 0$ jinak.
....
$d_{K-1} = 1$ pokud kvalitativní proměnná nabývá hodnoty Z_{K-1},
$\quad = 0$ jinak.

Budeme nadále uvažovat regresní model:

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \ldots + \beta_{K\text{-}1} d_{K\text{-}1,} + \varepsilon \qquad (7.1)$$

s vysvětlovanou proměnnou y představující vysvětlovanou (závisle) proměnnou a $K - 1$ fiktivními vysvětlujícími proměnnými: $d_1, \ldots, d_{K\text{-}1}$, ε je náhodná složka.

Mohli jste si všimnout, že k vyjádření K kategorií jsme použili $K - 1$ fiktivních proměnných. Možná vás napadla otázka: Proč jsme nepoužili též proměnnou d_0 pro kategorii Z_0 podobně jako pro ostatní kategorie? Pokud kvalitativní proměnná nabývá hodnoty Z_0, potom podle definice všechny zavedené fiktivní proměnné: $d_1, \ldots, d_{K\text{-}1}$, nabývají hodnoty 0, a tudíž je v regresním modelu (7.1) situace popsána rovnicí

$$y = \beta_0 + \varepsilon. \qquad (7.2)$$

Průměrná hodnota vysvětlované proměnné y je vyjádřena regresní úrovňovou konstantou β_0.

Poznámka. Kdybychom však postupovali tak, že bychom použili proměnnou d_0 pro kategorii Z_0, podobně jako pro ostatní kategorie, pak by došlo k situaci perfektní kolinearity mezi vysvětlujícími fiktivními proměnnými, což je nežádoucí situace popsaná v předchozí kapitole. V tomto případě bychom mohli použít regresní analýzu s modelem bez úrovňové konstanty, tj. regresní model

$$y = \beta_0 d_0 + \beta_1 d_1 + \beta_2 d_2 + \ldots + \beta_{K\text{-}1} d_{K\text{-}1,} + \varepsilon. \qquad (7.3)$$

K tomuto regresnímu modelu se vrátíme ještě v subkapitole 7.4, která se bude zabývat použitím fiktivních proměnných u sezónních dat.

Ke stanovení odhadů regresních koeficientů modelu (7.1), eventuálně model (7.3), použijeme metodu nejmenších čtverců, tedy vícerozměrnou regresní analýzu z kapitoly 5. Stejně tak můžete k řešení výchozí situace použít jednofaktorovou ANOVA, přitom nezávislým faktorem bude uvažovaná kvalitativní proměnná s K kategoriemi. Vztah mezi metodou ANOVA a metodou regresní analýzy vysvětlíme na konkrétním příkladu, který budeme řešit s pomocí Excelu.

ŘEŠENÝ PŘÍKLAD 7.1

Analyzujte závislost výdajů rodin na letní dovolenou na vzdělání rodičů (nejvyšší vzdělání alespoň jednoho z rodičů je základní - ZŠ, středoškolské - SŠ, vysokoškolské - VŠ). Použijte metodu ANOVA a poté vícerozměrnou regresní analýzu. Srovnejte oba výsledky. Data za 15 rodin jsou uvedena v následující tabulce.

Přitom Y_i představují výdaje rodiny na letní dovolenou,

d_{1i} = 1 jestliže rodiče mají vzdělání SŠ, 0 jinak,

d_{2i} = 1 jestliže rodiče mají vzdělání VŠ, 0 jinak.

y_i - výdaje na dov.	d_{1i} - SŠ	d_{2i} - VŠ
39	1	0
33	1	0
31	1	0
31	1	0
36	1	0
60	0	1
72	0	1
64	0	1
79	0	1
62	0	1
18	0	0
19	0	0
17	0	0
15	0	0
20	0	0

Řešení:

Budeme uvažovat regresní model

$$y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \varepsilon_i$$

s vysvětlovanou proměnnou y_i představující výdaje rodiny na letní dovolenou a dvěma fiktivními proměnnými: d_{1i} - rodiče mají vzdělání SŠ, d_{2i}- rodiče mají vzdělání VŠ. Všimněte si, že k vyjádření $K = 3$ kategorií (ZŠ, SŠ, VŠ) jsme použili 2 fiktivní proměnné. K výpočtu regresních koeficientů použijeme Excelu. V hlavním menu otevřeme postupně položky:

Data → Analýza dat... → Regrese

Data uložíme ve worksheetu v poli s adresou a1:c16 (viz níže), zadávací okno vyplníme takto:

Zadáme OK.

Ve výstupu dostaneme jak výsledek metody ANOVA, tak i výsledek regresní analýzy. V první tabulce výstupu: *Regresní statistika* nás zajímá druhá hodnota – koeficient determinace (Hodnota spolehlivosti R), tj. $R^2 = 0{,}953$.

Ve druhé tabulce ANOVA (viz níže) jednotlivé položky mají následující význam:

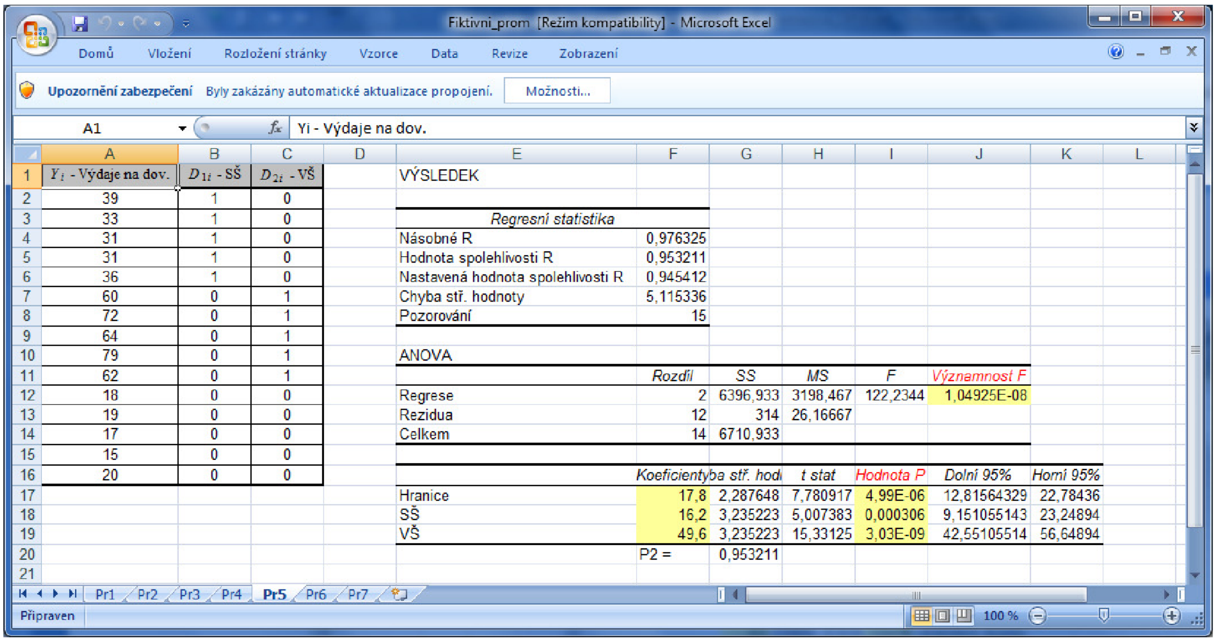
Regrese = meziskupinový součet čtverců
Rezidua = vnitroskupinový součet čtverců
Celkem = celkový součet čtverců
SS = Součet čtverců (Sum of Squares)
Rozdíl = stupeň volnosti (DF – Degree of Freedom)
MS = Průměr čtverců (Mean Square)
F = testové kritérium = 122,234
Významnost F = Signifikance (p-hodnota) = 0,00000001049 $< 0{,}05 = \alpha$
Faktor vzdělání vyjádřený kategoriemi ZŠ, SŠ, VŠ je tudíž statisticky významný.

Poměr determinace $P^2 = 0{,}953$ se vypočte jako podíl meziskupinového a celkového součtu čtverců (vypočtěte, konfrontujte s kapitolou 1!). Vidíte, že $R^2 = P^2$, tedy v případě fiktivních proměnných je koeficient determinace definovaný v regresní analýze totožný s poměrem determinace z ANOVA. Odtud plyne, že také koeficient korelace je totožný s poměrem korelace z ANOVA. Přiléhavost dat k regresní rovině je tedy totéž, co těsnost závislosti, s níž jste se setkali v metodě ANOVA.

Poslední tabulka výstupu přináší hodnoty odhadů regresních koeficientů:

Hranice = b_0 = 17,8 ; SŠ = b_1 = 16,2; VŠ = b_2 = 49,6.

Ve sloupci *Hodnota P* (signifikance) jsou všechna čísla mnohem menší, než je běžná hladina významnosti α =0,05, tedy hodnoty regresních koeficientů jsou statisticky významné. Přitom hodnota b_0 = 17,8 představuje průměrné výdaje rodiny, kde rodiče mají pouze základní vzdělání, b_1 = 16,2 představuje průměrný nárůst výdajů na letní dovolenou při nárůstu fiktivní proměnné d_{1i} z 0 na 1, tedy průměrné výdaje rodiny, kde rodiče mají středoškolské vzdělání, jsou $b_0 + b_1$ = 34,0 tis. Kč. Hodnota regresního koeficientu b_2 = 49,6 představuje průměrný nárůst výdajů na letní dovolenou při nárůstu fiktivní proměnné d_{2i} z 0 na 1, tedy průměrné výdaje rodiny, kde rodiče mají vysokoškolské vzdělání, jsou $b_0 + b_2$ = 67,4 tis. Kč. Poslední dva sloupce tabulky udávají dolní a horní hranici 95%-ního intervalu spolehlivosti pro příslušný regresní koeficient.



Y_i - Výdaje na dov.	D_{1i} - SŠ	D_{2i} - VŠ
39	1	0
33	1	0
31	1	0
31	1	0
36	1	0
60	0	1
72	0	1
64	0	1
79	0	1
62	0	1
18	0	0
19	0	0
17	0	0
15	0	0
20	0	0

VÝSLEDEK

Regresní statistika	
Násobné R	0,976325
Hodnota spolehlivosti R	0,953211
Nastavená hodnota spolehlivosti R	0,945412
Chyba stř. hodnoty	5,115336
Pozorování	15

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	2	6396,933	3198,467	122,2344	1,04925E-08
Rezidua	12	314	26,16667		
Celkem	14	6710,933			

	Koeficienty	Chyba stř. hodi	t stat	Hodnota P	Dolní 95%	Horní 95%
Hranice	17,8	2,287648	7,780917	4,99E-06	12,81564329	22,78436
SŠ	16,2	3,235223	5,007383	0,000306	9,151055143	23,24894
VŠ	49,6	3,235223	15,33125	3,03E-09	42,55105514	56,64894
	P2 =	0,953211				

7.3 Společné fiktivní a kvantitativní proměnné

Nyní budeme vyšetřovat situaci, která se v ekonomické oblasti často vyskytuje: současně se vyskytují jak s fiktivní vysvětlující proměnné (nabývají pouze hodnot 0 a 1), tak kvantitativní vysvětlující proměnné (mohou nabývat libovolných číselných hodnot). Pro jednoduchost se v následujícím textu omezíme na přítomnost pouze jediné kvantitativní proměnné, případ přítomnosti více kvantitativních proměnných je analogický. Uvažujme tedy regresní model

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + ... + \beta_{K-1} d_{K-1,} + b_K x_{K,} + \varepsilon \qquad (7.4)$$

s vysvětlovanou proměnnou y, $K - 1$ fiktivními vysvětlujícími proměnnými: $d_1,...,d_{K-1}$ a kvantitativní vysvětlující proměnnou x_K, přitom ε je náhodná složka. Pro ilustraci budeme uvažovat následující příklad, který je rozšířením příkladu z předchozí subkapitoly.

ŘEŠENÝ PŘÍKLAD 7.2

Analyzujte závislost výdajů rodin na letní dovolenou na vzdělání rodičů (nejvyšší vzdělání alespoň jednoho z rodičů je základní - ZŠ, středoškolské - SŠ, vysokoškolské - VŠ), jakož i na příjmech rodin. Data za 15 rodin jsou uvedena v následující tabulce. Přitom y_i představují výdaje rodin na letní dovolenou, x_i představují roční příjem rodin (obojí v Kč), dále uvažujeme dvě fiktivní proměnné:

d_{1i} = 1 jestliže rodiče mají vzdělání SŠ, 0 jinak,
d_{2i} = 1 jestliže rodiče mají vzdělání VŠ, 0 jinak.

y_i - výdaje na dov.	d_{1i} - SŠ	d_{2i} - VŠ	x_i - příjmy
39	1	0	400
33	1	0	310
31	1	0	180
31	1	0	190
36	1	0	470
60	0	1	270
72	0	1	260
64	0	1	170
79	0	1	430
62	0	1	490
18	0	0	150
19	0	0	250
17	0	0	290
15	0	0	200
20	0	0	410

Řešení:

Budeme uvažovat regresní model

$$y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 x_i + \varepsilon_i$$

s vysvětlovanou proměnnou y_i představující výdaje rodiny na letní dovolenou, dvěma fiktivními proměnnými: d_{1i} - rodiče mají vzdělání SŠ, d_{2i}- rodiče mají vzdělání VŠ a jednou kvantitativní vysvětlující proměnnou x_i. V hlavním menu otevřeme postupně položky:

Data → Analýza dat... → Regrese

Data uložíme ve worksheetu v poli s adresou a1:d16 (viz níže), zadávací okno vyplníme takto:

Zadáme OK.

V první tabulce výstupu: Regresní statistika nás zajímá druhá hodnota – koeficient determinace (Hodnota spolehlivosti R), tj. $R^2 = 0{,}96$. Přiléhavost dat k regresní nadrovině je 96%.

Druhá tabulka ANOVA (viz níže) však již nemá význam klasické metody ANOVA jako v příkladu 7.1. Dále se jí nebudeme zabývat.

Poslední tabulka výstupu přináší hodnoty odhadů regresních koeficientů:

Hranice = $b_0 = -12{,}86$, d1i-SŠ = $b_1 = -0{,}68$, d2i-VŠ = $b_2 = 4{,}472$, xi-příjmy = $b_3 = 0{,}172$, viz následující tabulka:

VÝSLEDEK

Regresní statistika	
Násobné R	0,9982662
Hodnota spolehlivosti R	0,99653541
Nastavená hodnota spolehlivosti R	0,99559053
Chyba stř. hodnoty	1,4538542
Pozorování	15

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	3	6687,682721	2229,2	1054,6605	8,27968E-14
Rezidua	11	23,25061224	2,1137		
Celkem	14	6710,933333			

	Koeficienty	Chyba stř. hodnoty	t stat	Hodnota P	Dolní 95%	Horní 95%
Hranice	-12,860	2,694	-4,774	0,001	-18,789	-6,931
d1i - SŠ	-0,680	1,708	-0,398	0,698	-4,439	3,079
d2i - VŠ	4,472	3,956	1,130	0,282	-4,236	13,179
xi - příjmy	0,172	0,015	11,728	0,000	0,140	0,205

Ve sloupci *Hodnota P* (signifikance) jsou u koeficientů b_1 (0,698) a b_2 (0,282) hodnoty signifikance větší, než běžná hladina významnosti $\alpha = 0{,}05$, tedy hodnoty těchto regresních koeficientů **nejsou** statisticky významné (nulovou hypotézu, že $b_i = 0$ přijímáme) a nemají proto na vysvětlovanou proměnnou (tj. výdaje na dovolenou) vliv. Ukazuje se, že mnohem významnější vliv má vysvětlující proměnná x, u které je regresní koeficient b_3 (0,172) statisticky významný, stejně jako koeficient b_0 (–12,86). Přitom hodnota $b_0 = -12{,}86$ představuje teoretickou hodnotu výdajů na dovolenou v (nerealistickém) případě nulových příjmů rodiny. Ve srovnání s příkladem 7.1, kde jsme použili stejná data ochuzená o údaje o příjmech rodin, jsme obdrželi velmi odlišný výsledek. Tento výsledek však není s předchozím v rozporu, jak by se mohlo zdát. K původním datům jsme totiž přidali data nová, která jsme použili pro naši analýzu, a přitom se ukázalo, že pak výdaje na dovolenou nezávisí na vzdělání rodin, nýbrž závisí na příjmech rodin, které jsou pro danou závislost mnohem významnější.

7.4 FIKTIVNÍ PROMĚNNÉ V SEZÓNNÍCH MODELECH

Mnoho ekonomických údajů se nachází ve formě tzv. časových řad, kdy údaje jsou uspořádány podle sledované rostoucí časové posloupnosti. Může se např. jednat o denní (měsíční, roční) hodnoty tržeb v jistém supermarketu. V těchto časových řadách se často projevují tzv. sezónní vlivy, které se pravidelně opakují v jistých obdobích, např. v letní (zimní) sezónu, o Vánocích, Velikonocích apod. Během uvažovaných sezón se chování hodnot sledovaného znaku odlišuje od běžného chování ve zbytku sledovaného období, zaznamenáváme v chování tzv. sezónní složku. Podrobněji se budeme analýzou časových řad zabývat v následujících kapitolách tohoto textu. Na tomto místě uvedeme příklad situace, kdy je možné k analýze časové řady se sezónními složkami výhodně využít fiktivních proměnných, které identifikují tyto sezónní složky a umožňují tak odstranění sezónnosti. To přispívá k lepšímu pochopení chování sledovaného ukazatele. V ekonomické praxi jsou důležité časové řady, jako např. míra nezaměstnanosti, spotřebitelský cenový index (CPI), index průmyslové výroby aj., publikovány ve formě časových řad „očištěných“ od sezónnosti.

Pro ilustraci techniky fiktivních proměnných budeme uvažovat příklad měsíčních tržeb prodeje ledniček v jistém supermarketu. Sezónami zde budou 4 čtvrtletí roku, proto budeme uvažovat regresní model se 4 fiktivními proměnnými bez úrovňové konstanty. Tento postup jsme zvolili proto, abychom ukázali druhou možnost volby počtu fiktivních vysvětlujících proměnných, viz subkapitolu 7.2 (a Poznámku tam uvedenou), kde jsme použili metodu, která by vedla v tomto případě ke 3 fiktivním proměnným, avšak s úrovňovou konstantou. Zde tedy uvádíme druhý způsob řešení se 4 fiktivními proměnnými, avšak bez úrovňové konstanty. Budeme uvažovat následující regresní model

$$y = \beta_0 d_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_{3,} + \varepsilon, \qquad (7.5)$$

kde y je vysvětlovaná proměnná – měsíční prodej ledniček, d_i, i = 0,1,2,3, jsou fiktivní proměnné pro první, druhé, třetí a čtvrté čtvrtletí roku. Data prodejů za 32 po sobě jdoucích měsíců (v tis. Kč) jsou uvedena v následující tabulce.

y - tržby	d0	d1	d2	d3	y - tržby	d0	d1	d2	d3
1317	1	0	0	0	943	1	0	0	0
1615	0	1	0	0	1175	0	1	0	0
1662	0	0	1	0	1269	0	0	1	0
1295	0	0	0	1	973	0	0	0	1
1271	1	0	0	0	1102	1	0	0	0
1555	0	1	0	0	1344	0	1	0	0
1639	0	0	1	0	1641	0	0	1	0
1238	0	0	0	1	1225	0	0	0	1
1277	1	0	0	0	1429	1	0	0	0
1258	0	1	0	0	1699	0	1	0	0
1417	0	0	1	0	1749	0	0	1	0
1185	0	0	0	1	1117	0	0	0	1
1196	1	0	0	0	1242	1	0	0	0
1410	0	1	0	0	1684	0	1	0	0
1417	0	0	1	0	1764	0	0	1	0
919	0	0	0	1	1328	0	0	0	1

Uvedeme řešení pomocí Excelu. V hlavním menu otevřeme postupně položky:
Data → Analýza dat... → Regrese
Data uložíme ve worksheetu v poli s adresou a1:e33 (viz níže), zadávací okno vyplníme takto:

Všimněte si, že jsme zaklikli položku Konstanta je nula. Tato volba umožňuje řešení regresního modelu **bez úrovňové konstanty**. Po zadání OK obdržíme řešení uvedené v následující tabulce. Odhady regresních koeficientů, které jsou všechny statisticky významné (viz sloupec Hodnota P, jde o velmi malá čísla), má regresní nadrovina rovnici

$$y = 1222{,}13d_0 + 1467{,}50d_1 + 1569{,}75d_2 + 1160{,}00d_3. \qquad (7.6)$$

Regresní koeficienty v (7.6) představují průměrné hodnoty prodejů v jednotlivých čtvrtletích. Zvolíme-li jako výchozí úroveň hodnoty prodejů v 1. čtvrtletí, potom *sezónní faktor* S_i *i*-tého čtvrtletí obdržíme jako rozdíl mezi b_i a b_0, tj. $S_i = b_i - b_0$. V našem příkladu je $S_0 = 0$, $S_1 = 245{,}38$, $S_2 = 247{,}63$, $S_3 = -62{,}13$.

Další použití fiktivních proměnných si ukážeme v kapitole o sezónních modelech časových řad.

VÝSLEDEK

Regresní statistika	
Násobné R	0,729244
Hodnota spolehlivosti R	0,531797
Nastavená hodnota spolehlivosti R	0,445918
Chyba stř. hodnoty	169,6785
Pozorování	32

ANOVA

	Rozdíl	*MS*	*F*	*ýznamnost F*
Regrese	4	228909	7,950768	0,000227
Rezidua	28	28790,8		
Celkem	32			

	Koeficienty	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	0	#N/A	#N/A	#N/A	#N/A
d0	1222,13	20,372	2,5E-18	1099,24	1345,01
d1	1467,50	24,46224	1,94E-20	1344,615	1590,385
d2	1569,75	26,16668	3,18E-21	1446,865	1692,635
d3	1160,00	19,33642	9,81E-18	1037,115	1282,885

Můžeme tedy shrnout: Jedním ze způsobů, jak kvantifikovat přítomnost nebo nepřítomnost nějaké vlastnosti, je konstrukce nových proměnných, které toto vyjadřují pomocí čísel 1 nebo 0, přitom 1 indikuje přítomnost a 0 nepřítomnost vlastnosti. Tyto proměnné se nazývají fiktivní proměnné (anglicky dummy variables). Fiktivní proměnné lze zavést do regresního modelu stejně snadno, jako jakékoliv jiné kvantitativní proměnné, které jsme použili např. v předchozí kapitole. V této kapitole jsme ukázali, jak vytvořit regresní model, který obsahuje výlučně fiktivní vysvětlující proměnné a také model, který je kombinací fiktivních a obvyklých vysvětlujících proměnných. Na konec jsme ukázali, jak lze fiktivní proměnné výhodně použít pro sezónní data k identifikaci jednotlivých sezón.

7.5 SAMOSTATNÉ ÚKOLY

7.1 V následující tabulce jsou uvedeny roční úspory a příjmy rodin za období od roku 1979 do roku 1996. Protože po roce 1989 došlo ke změně ekonomického chování rodin, použijte fiktivní proměnnou k vysvětlení změny závislosti úspor na příjmech, viz sloupec „období“.

rok	úspory	příjmy	období
1979	57,7	831,8	0
1980	66,3	894	0
1981	61,4	981,6	0
1982	89	1101,7	0
1983	96,7	1210	0
1984	104,6	1313,4	0
1985	95,8	1451,4	0
1986	90,7	1607,5	0
1987	110,2	1812,4	0
1988	118,1	2034	0
1989	136,9	2258,5	0
1990	159,4	2520,9	1
1991	153,9	2670,8	1
1992	130,6	2836,6	1
1993	164,1	3008,7	1
1994	125,4	3325,3	1
1995	121,7	3531	1
1996	104,2	3780	1

7.6 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

7.1 $y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 x_i + \varepsilon_i$; y_i…úspory; x_i…příjmy; d_{1i} … fiktivní proměnná
$Y_i = 71{,}86 + 0{,}015\, x_i + 17{,}84\, d_{1i}$

Hodnota P (signifikance) je u koeficientu b_1 (0,236) a u koeficientu b_2 (0,465) tedy nulovou hypotézu o nulovosti koeficientů nelze zamítnout a tyto proměnné nemají vliv na vysvětlovanou proměnnou.
Vynecháme-li fiktivní proměnnou, dostáváme: $Y_i = 61{,}87 + 0{,}023\, x_i$ a hodnota P u koeficientu b_1 je 0,0009 a nulovou hypotézu o nulovosti koeficientu zamítáme. To znamená, že úspory nezávisí na období, ale na příjmech rodin.

8 ZÁKLADY ANALÝZY ČASOVÝCH ŘAD

RYCHLÝ NÁHLED KAPITOLY

Důležitým nástrojem ke zkoumání dynamiky ekonomických procesů je analýza časových řad. Časovou řadou přitom rozumíme věcně a prostorově srovnatelná pozorování uspořádaná v čase směrem od minulosti přes přítomnost k budoucnosti. Obsahem této kapitoly je objasnit typizaci ekonomických časových řad, vysvětlit elementární charakteristiky časových řad, uvést základní modely časových řad a popsat jejich složky. Analýza časových řad je vedena snahou po vysvětlení minulosti a předvídání budoucnosti, v ekonomické oblasti se jedná o vývojové trendy ukazatelů hospodářské činnosti. Analýza časových řad jako soubor metod a postupů nabízí širokou škálu nástrojů a technik. Ke klasickým analytickým postupům založeným na regresi z předchozích kapitol a syntetickým přístupům založeným na technikách vyrovnání časových řad, přistupuje moderní, výpočetně náročnější harmonická analýza a Box - Jenkinsova metodologie využívající současného mohutného rozvoje výpočetní techniky.

8.1 TYPY EKONOMICKÝCH ČASOVÝCH ŘAD

Důležitým nástrojem ke zkoumání dynamiky ekonomických procesů je analýza časových řad. *Časovou řadou* přitom rozumíme věcně a prostorově srovnatelná pozorování uspořádaná v čase směrem od minulosti přes přítomnost k budoucnosti. Časové řady členíme následujícím způsobem:

- podle charakteru časové řady na *intervalové časové řady* a *okamžikové časové řady,*
- podle periodicity, s jakou jsou sledovány, na *krátkodobé* časové řady (méně než roční periodicita), *střednědobé* časové řady (roční periodicita) a *dlouhodobé* časové řady (delší, než roční periodicita),
- podle druhu sledovaných ukazatelů (údajů) na časové řady *absolutních* ukazatelů a časové řady *odvozených* ukazatelů.

Intervalovou časovou řadou se rozumí časová řada intervalového ukazatele y_t, tj. ukazatele, jehož velikost (hodnota) závisí na délce intervalu, za který je sledován. Pro ukazatele tohoto typu je možné tvořit součty, z jejich povahy však vyplývá, že se vztahují ke stejně dlouhým časovým intervalům, jinak by byly hodnoty vzájemně nesrovnatelné. Není např. správné srovnávat výrobu za leden a únor, neboť únor je z hlediska počtu pracovních dní kratší. Abychom zajistili srovnatelnost, přepočítáváme všechna sledovaná období na stejný časový interval. Tato operace se nazývá očišťování časových řad od kalendářních variací. Údaje očištěné časové řady $y_t^{(0)}$ dostaneme z hodnoty očišťovaného ukazatele y_t takto:

$$y_t^{(0)} = y_t \frac{\bar{k}_t}{k_t}, \tag{8.1}$$

kde $\bar{k}_t$ je průměrný počet dnů v příslušném dílčím období, k_t je skutečný počet dnů v příslušném dílčím období *t*.

Okamžikovou časovou řadou rozumíme časovou řadu ukazatelů, které se vztahují k určitému okamžiku, např. počátku nebo konci určitého časového intervalu (období). Protože součet za několik za sebou jdoucích okamžikových hodnot obvykle nemá reálný smysl, shrnují se řady tohoto typu pomocí chronologického průměru.

Pro dané *ekvidistantní* (stejně vzdálené) časové okamžiky $t_1, t_2,...,t_n$, ke kterým přísluší hodnoty okamžikových ukazatelů $y_1,y_2,...,y_n$ je *prostý chronologický průměr* definován jako aritmetický průměr z aritmetických průměrů vždy dvou po sobě jdoucích hodnot, tedy:

$$\bar{y}_{ch} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + ... + \frac{y_{n-1} + y_n}{2}}{n-1}, \tag{8.2}$$

Není-li délka mezi jednotlivými časovými okamžiky stejná, definujeme *vážený chronologický průměr*, kde vahami jsou délky jednotlivých časových intervalů $d_k = t_{k+1} - t_k$, $k = 1,2,...,n-1$:

$$\bar{y}_{ch} = \frac{\frac{y_1 + y_2}{2} d_1 + \frac{y_2 + y_3}{2} d_2 + ... + \frac{y_{n-1} + y_n}{2} d_{n-1}}{d_1 + d_2 + ... + d_{n-1}}. \tag{8.3}$$

Časový rozdíl mezi časovými okamžiky, tedy délka časového intervalu v okamžikové časové řadě, se nazývá *periodicita* časové řady. Je-li periodicita ekonomických časových řad kratší než jeden rok, hovoříme o *krátkodobých časových řadách*. Nejčastější periodicitou je *měsíční* periodicita. Je-li periodicita *roční*, hovoříme často o *střednědobých časových řadách*, při delší periodicitě, např. pětileté, hovoříme o *dlouhodobých časových řadách.*

Časovou řadou absolutních hodnot se obvykle rozumí časová řada přímo zjištěných údajů (v naturálních jednotkách) očištěná od kalendářních variací. Odvozené údaje a z nich vytvořené časové řady získáme obvykle matematickými operacemi z absolutních údajů. Většinu důležitých ekonomických časových řad tvoří časové řady ukazatelů vyjádřených v peněžní formě. Vzhledem ke změnám cenové hladiny, které jsou v tržní ekonomice přirozené, však v delší časové řadě často dostáváme posloupnost údajů, které nejsou vždy zcela souměřitelné. Proto důležitým problémem v analýze časových řad je srovnatelnost údajů, konkrétně cenová srovnatelnost. Při sestavování delší časové řady je možno v zásadě postupovat dvojím způsobem: použít *běžné ceny* a vyjádřit v nich absolutní objem určitého ukazatele, resp. tempa růstu, nebo vycházet ze *stálých cen*, tj. cen fixovaných k určitému datu. Používání stálých cen v ekonomice vede ke zmírnění negativních tendencí v účinnosti základních fondů vyplývajících z vlivu technického rozvoje na výrobu, dále vede ke zreálnění výsledků hospodářského vývoje vzhledem k mezinárodnímu srovnání.

Vývoj základních ekonomických ukazatelů v České republice je možné sledovat jednak za jednotlivé roky ve *statistických ročenkách*, jednak podle jednotlivých měsíců ve *statistických přehledech* a *bulletinech* vydávaných Českým statistickým úřadem. Pro potřeby vrcholového řízení ve firmách a podnicích slouží především údaje o vývoji základních ukazatelů podle měsíců, neboť jde o informace s určitým vztahem k okamžité odezvě v chování ekonomických subjektů, ať už výrobců, nebo spotřebitelů. Jsou to zejména informace o inflaci (index spotřebitelských cen a indexy životních nákladů), dále informace o peněžních příjmech a výdajích obyvatelstva, o celkovém prodeji v maloobchodě, průmyslové, zemědělské a stavební výrobě a též údaje o nezaměstnanosti.

Bohatým zdrojem informaci a dat jsou webové stránky Českého statistického úřadu (ČSÚ), www.czso.cz případně Statistického úřadu Evropské komise EUROSTAT: http://epp.eurostat.ec.europa.eu .

8.2 ELEMENTÁRNÍ CHARAKTERISTIKY ČASOVÝCH ŘAD

Mezi elementární metody analýzy časových řad patří *vizuální analýza* chování ukazatele využívající grafů spolu s určováním elementárních statistických charakteristik, ke kterým patří *absolutní diference* různého řádu a *koeficient růstu* časové řady.

Označíme-li y_t hodnoty určitého ukazatele v čase $t = 1,2,...,n$ (např. v jednotlivých měsících), potom *absolutní diferencí prvního řádu* rozumíme rozdíl:

$$\Delta^{(1)} y_t = y_t - y_{t-1},\ t = 2,3,...,n. \tag{8.4}$$

Obdobně lze definovat *absolutní diference vyšších řádů* - druhého, třetího, atd.:

$$\Delta^{(2)} y_t = \Delta^{(1)} y_t - \Delta^{(1)} y_{t-1} = y_t - 2y_{t-1} + y_{t-2},\ t = 3,4,...,n,$$

$$\Delta^{(3)} y_t = \Delta^{(2)} y_t - \Delta^{(2)} y_{t-1} = y_t - 3y_{t-1} + 3y_{t-2} - y_{t-3},\ t = 4,5,...,n, \text{ atd.}$$

Další používanou elementární charakteristikou je *koeficient růstu*, který udává, o kolik procent vzrostla hodnota časové řady v daném časovém okamžiku oproti období v předchozím časovém okamžiku:

$$k_t = \frac{y_t}{y_{t-1}},\ t = 2,3,...,n. \tag{8.5}$$

Při hodnocení vývoje za celou analyzovanou řadu zjišťujeme souhrnné charakteristiky – *průměrný absolutní přírůstek*:

$$\overline{\Delta} = \frac{1}{n-1}\sum_{t=2}^{n} \Delta^{(1)} y_t = \frac{y_n - y_1}{n-1}, \tag{8.6}$$

a průměrný koeficient růstu:

$$\overline{k} = \sqrt[n-1]{k_2 k_3 \ldots k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}. \tag{8.7}$$

Jak průměrný absolutní přírůstek, tak průměrný koeficient růstu závisí pouze na první a poslední hodnotě časové řady. Průměrný absolutní přírůstek ukazuje, o kolik by se měl ukazatel pravidelně měnit (v absolutních jednotkách), aby se hodnota ukazatele změnila z původní první hodnoty y_1 na poslední hodnotu y_n. Naproti tomu průměrný koeficient růstu poskytuje informaci, o kolik procent by se měla hodnota ukazatele měnit, tj. jaká by měla být rychlost růstu (poklesu), aby se hodnota ukazatele změnila z původní první hodnoty y_1 na poslední hodnotu y_n.

8.3 MODELY EKONOMICKÝCH ČASOVÝCH ŘAD

Modelový přístup k analýze časových řad bude vycházet z předpokladu, že *jediným faktorem dynamiky ukazatele v časové řadě je čas*. Ostatní faktory působící na hodnotu ukazatele budeme většinou zanedbávat. Model časové řady tohoto typu můžeme zapsat ve formě:

$$y_t = f(t, \varepsilon_t), \tag{8.8}$$

kde y_t je hodnota analyzovaného ukazatele v čase t, f je určitá funkce (typ závislosti), t je časová proměnná, ε_t je hodnota náhodné složky. Modely časových řad založené na výše uvedeném principu se nazývají *jednorozměrné modely*.

Každá časová řada může obsahovat 4 složky, které vyjadřují různé druhy pohybu analyzovaného ukazatele:

- trendovou složku (trend) T_t,
- sezónní složku S_t,
- cyklickou složku C_t,
- náhodnou složku ε_t.

Trendová, sezónní a cyklická složka tvoří společně *systematickou (deterministickou) složku* kterou značíme Y_t , tj. $Y_t = T_t + S_t + C_t$. Zpravidla se uvažuje, že složky Y_t jsou v aditivním vztahu, takže model časové řady můžeme zapsat ve tvaru:

$$y_t = T_t + S_t + C_t + \varepsilon_t . \qquad (8.9)$$

V tom případě mluvíme o *aditivním modelu* časové řady. V ekonomických časových řadách se nejčastěji setkáváme se dvěma speciálními případy modelu (8.9). U střednědobých modelů (s roční periodicitou) se obvykle předpokládá $S_t = C_t = 0$, pak model časové řady (8.9) má tvar:

$$y_t = T_t + \varepsilon_t . \qquad (8.10)$$

U krátkodobých modelů časových řad (s čtvrtletní nebo měsíční periodicitou) se předpokládá, že $C_t = 0$, a tedy model (8.9) má tvar:

$$y_t = T_t + S_t + \varepsilon_t , \qquad (8.11)$$

mluvíme pak o *časové řadě se sezónní složkou.*

Vedle aditivního modelu (8.9) je *multiplikativní model* založen na předpokladu, že vzájemný vztah jednotlivých složek obsažených v modelu je dán vzájemným násobením:

$$y_t = T_t \cdot S_t \cdot C_t \cdot \varepsilon_t . \qquad (8.12)$$

Popis a kvantifikace jednotlivých složek modelu časové řady patří k hlavním úkolům analýzy časových řad.

ŘEŠENÝ PŘÍKLAD 8.1

V tabulce jsou uvedeny průměrné měsíční mzdy zaměstnanců ve státní správě v letech 1989-1997. Pro tuto časovou řadu vypočítejte:

a. absolutní přírůstky a průměrný absolutní přírůstek,
b. koeficienty růstu a průměrný koeficient růstu.

Roky	1989	1990	1991	1992	1993	1994	1995	1996	1997
Mzda	2980	3110	4500	5650	7460	8930	10670	12820	13250

Řešení:

a. Absolutní přírůstky vypočítáme podle vztahu (8.4):

$$\Delta^{(1)} y_2 = y_2 - y_1 = 3110 - 2980 = 130, \text{ atd.}$$

Výsledek říká, že průměrná měsíční mzda stoupla v letech 1989-1990 o 130 Kč.
Všechny absolutní přírůstky jsou uvedeny v následující tabulce.
Průměrný absolutní přírůstek je podle (8.6):

$$\overline{\Delta} = \frac{y_n - y_1}{n-1} = \frac{13250 - 2980}{8} = 1283{,}75.$$

b. Koeficienty růstu vypočítáme podle vztahu (4.5). Např.:

$$k_2 = \frac{y_2}{y_1} = \frac{3110}{2980} = 1{,}0436.$$

Průměrná měsíční mzda vzrostla v letech 1989-1990 o 4,36%.

Hodnoty ostatních koeficientů růstu jsou uvedeny v následující tabulce.
Průměrný koeficient růstu vypočítáme podle (8.7):

$$\bar{k} = \sqrt[n-1]{\frac{y_n}{y_1}} = \sqrt[8]{\frac{13250}{2980}} = 1{,}205.$$

Výsledek ukazuje, že mzdy rostly ročně v průměru o 20,5%.

Roky	1989	1990	1991	1992	1993	1994	1995	1996	1997
Mzda	2980	3110	4500	5650	7460	8930	10670	12820	13250
$\Delta^{(1)} y$	.	130	1390	1150	1810	1470	1740	2150	430
k	.	1,04	1,45	1,26	1,32	1,20	1,19	1,20	1,03

8.4 SAMOSTATNÉ ÚKOLY

8.1 V tabulce jsou uvedeny počty prodaných automobilů v autocentru A+A v letech 1990 až 1997. Pro tuto časovou řadu vypočítejte:

a) absolutní přírůstky a průměrný absolutní přírůstek
b) koeficienty růstu a průměrný koeficient růstu.

Rok	1990	1991	1992	1993	1994	1995	1996	1997
Počet	120	159	167	175	197	172	199	240

8.5 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

8.1

Rok	**Počet**	**Abs.přírůstky**	**Koeficienty růstu**
1990	120	xxx	xxx
1991	159	39	1,325
1992	167	8	1,050
1993	175	8	1,048
1994	197	22	1,126
1995	172	-25	0,873
1996	199	27	1,157
1997	240	41	1,206

Průměrný absolutní přírůstek je podle (8.6): $\bar{\Delta} = 17{,}14$.
Průměrný koeficient růstu vypočítáme podle (8.7): $\bar{k} = 1{,}104$.
Počet prodaných automobilů rostl ročně v průměru o 10,4%.

9 ANALÝZA TRENDU ČASOVÝCH ŘAD

RYCHLÝ NÁHLED KAPITOLY

V této kapitole se budete zabývat trendovou složkou časové řady, která představuje nejdůležitější komponentu analyzované časové řady. Proto popis trendu je jedním z nejdůležitějších úkolů analýzy časových řad. Vycházíme přitom z předpokladu, že jediným faktorem vývoje dynamiky analyzovaného ukazatele je čas. Trendová složka totiž poskytuje rozhodující informaci pro prognózování hodnot časové řady do budoucna. K určení trendové složky používáme dva obecné přístupy: analytický a syntetický. Analytický přístup stanovení trendu vychází z předem známých typů trendových funkcí vyznačujících se přítomností parametrů, které je třeba stanovit co nejlépe s ohledem na skutečné hodnoty ukazatele časové řady. Z velkého množství používaných trendových funkcí se zaměříme na několik z nich, které mají význam především v ekonomických aplikacích. Jsou to: lineární trend, parabolický trend, exponenciální trend, logistický trend a Gompertzův trend. Syntetický přístup stanovení trendu spočívá ve vyrovnání odchylek daného ukazatele v časové řadě tak, že získané vyrovnané hodnoty vyjadřují trendový faktor obsažený pouze v časové řadě, nikoliv faktor vložený z vnějšku. Nemusíte proto znát předem typ trendové funkce, což je přednost syntetického přístupu oproti přístupu analytickému. Jeho nevýhodou je naopak obtížnější využití pro prognózování hodnot časové řady. Z existujících metod syntetického přístupu uvedeme metody klouzavého průměru a exponenciální vyrovnání.

9.1 TRENDOVÁ SLOŽKA ČASOVÝCH ŘAD

Jak již bylo v průvodci studiem řečeno, v této kapitole vycházíme z předpokladu, že jediným faktorem vývoje dynamiky analyzovaného ukazatele je čas t. Jednoduchý způsob volby časové proměnné spočívá v jejím zavedení tak, že časová řada začíná v okamžiku 1, ke kterému se vztahuje první člen analyzované časové řady y_1. Další časové okamžiky označujeme po řadě přirozenými čísly 2,3,...,n. Symbol n označuje poslední uvažovaný časový okamžik a zároveň i počet uvažovaných časových okamžiků.

Jiný jednoduchý a výhodný způsob označení časové proměnné spočívá v zavedení nové časové proměnné t' následujícím způsobem:

$$t' = (t - \bar{t}), \tag{9.1}$$

je-li počet členů časové řady n lichý, pak $\bar{t} = \frac{n+1}{2}$, nebo

$$t' = 2(t - \bar{t}), \tag{9.2}$$

je-li počet členů n sudý. Nová časová proměnná splňuje důležitý požadavek:

$$\sum_{t=1}^{n} t' = 0\,. \tag{9.3}$$

Trendová složka představuje nejdůležitější komponentu analyzované časové řady, a proto popis trendu je jedním z nejdůležitějších úkolů analýzy časových řad. Trendová složka totiž poskytuje rozhodující informaci pro prognózování hodnot časové řady do budoucna. K určení trendové složky používáme dva obecné přístupy: *analytický* a *syntetický*.

Analytický přístup stanovení trendu vychází z předem známých typů trendových funkcí vyznačujících se přítomností parametrů, které je třeba stanovit co nejlépe s ohledem na skutečné hodnoty ukazatele časové řady. Z velkého množství používaných trendových funkcí se zaměříme na několik z nich, které mají význam především v ekonomických aplikacích.

Jsou to: lineární trend, parabolický trend, exponenciální trend, logistický trend a Gompertzův trend. Výhodou těchto trendových funkcí je to, že je lze snadno použít pro účely prognózování. Nevýhodou je fakt, že typ trendové funkce musíme stanovit předem na základě externích, mnohdy subjektivních předpokladů a informací. Nejužívanější metodou odhadu neznámých parametrů trendové funkce je *metoda nejmenších čtverců (MNČ)*, s níž jsme se setkali již v kapitole 3. Zde tuto metodu aplikujeme na speciální typ jednoduché regrese pro data ve formě ekonomické časové řady, tedy případ, kdy nezávisle proměnnou je čas a závisle proměnnou tvoří sledovaný ekonomický ukazatel. Kromě metody nejmenších čtverců pro nelineární trendové funkce uvedeme alternativní *metodu vybraných bodů (MVB).*

Syntetický přístup stanovení trendu spočívá ve vyrovnání odchylek daného ukazatele v časové řadě (tzv. vyrovnání) tak, že získané vyrovnané hodnoty vyjadřují trendový faktor obsažený pouze v časové řadě, nikoliv faktor vložený z vnějšku. Nemusíme proto znát předem typ trendové funkce, což je přednost syntetického přístupu oproti přístupu analytickému. Jeho nevýhodou je naopak obtížnější využití pro prognózování hodnot časové řady. Z existujících metod syntetického přístupu uvedeme *metody klouzavého průměru* a *exponenciální vyrovnání*.

9.2 LINEÁRNÍ TREND

Nejčastěji používanou trendovou funkcí je lineární trendová funkce:

$$T_t = \beta_0 + \beta_1 t\,, \tag{9.4}$$

kde β_0, β_1 jsou neznámé parametry a t = 1,2,...,n je časová proměnná. Odhady neznámých parametrů, které označujeme b_0, b_1, získáme metodou nejmenších čtverců, která dává nejlepší nestranné odhady. V souladu s postupem z kapitoly 3 je zapotřebí vyřešit 2 normální rovnice (3.12), kde x_i nahradíme t :

$$\sum y_t = b_0 n + b_1 \sum t\,, \tag{9.5}$$

$$\sum t y_t = b_0 \sum t + b_1 \sum t^2\,. \tag{9.6}$$

Použijeme-li nyní časové transformace (9.1), (9.2) a s využitím vztahu (9.3) dostaneme jednoduché řešení normálních rovnic (9.5), (9.6):

$$b_0 = \frac{\sum y_t}{n}, \quad b_1 = \frac{\sum t' y_t}{\sum (t')^2}. \tag{9.7}$$

Parametr b_0 interpretujeme jako aritmetický průměr hodnot časové řady, parametr b_1 udává, jaký přírůstek hodnoty T_t odpovídá jednotkovému přírůstku proměnné t.

9.3 PARABOLICKÝ TREND

Rozšířením lineárního trendu o kvadratický člen dostaneme *parabolickou trendovou funkci*:

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2\,, \tag{9.8}$$

kde $\beta_0, \beta_1, \beta_2$ jsou neznámé parametry a t = 1,2,...,n je časová proměnná. Odhady neznámých parametrů, které označujeme b_0, b_1, b_2, získáme metodou nejmenších čtverců řešením soustavy 3 lineárních rovnic o 3 neznámých:

$$\sum y_t = b_0 n + b_1 \sum t' + b_2 \sum (t')^2 \,,$$
$$\sum t' y_t = b_0 \sum t' + b_1 \sum (t')^2 + b_2 \sum (t')^3 \,, \qquad (9.9)$$
$$\sum (t')^2 y_t = b_0 \sum (t')^2 + b_1 \sum (t')^3 + b_2 \sum (t')^4 \,.$$

Z podmínky (9.3) dostaneme z rovnice (9.9) ihned řešení:

$$b_1 = \frac{\sum t' y_t}{\sum (t')^2} \,. \qquad (9.10)$$

Dosazením (9.10) do zbývajících dvou normálních rovnic obdržíme ještě řešení b_0, b_2:

$$b_0 = \frac{\sum y_t \sum (t')^4 - \sum (t')^2 \sum y_t (t')^2}{n \sum (t')^4 - \left(\sum (t')^2\right)^2} \,, \qquad (9.11)$$

$$b_2 = \frac{n \sum y_t (t')^2 - \sum y_t \sum (t')^2}{n \sum (t')^4 - \left(\sum (t')^2\right)^2} \,. \qquad (9.12)$$

9.4 MOCNINNÝ TREND

Mocninná trendová funkce má tvar:

$$T_t = \beta_0 t^{\beta_1} \,, \qquad (9.13)$$

avšak namísto něj uvažujeme model, jenž vznikne logaritmováním obou stran (9.13):

$$\ln T_t = \ln \beta_0 + \beta_1 \ln t \,,$$

kde ln je přirozený logaritmus o základu e = 2,718... Použijeme analogický postup jako v případě jednoduché lineární regrese v kapitole 2.2.6. Jestliže nyní použijeme substituce

$$T_t' = \ln T_t \,, \; t'' = \ln t \,, \qquad (9.14)$$

$$\beta_0' = \ln \beta_0 \,, \; \beta_1' = \beta_1 \,, \qquad (9.15)$$

obdržíme „čárkovaný“ lineární trend:

$$T_t' = \beta_0' + \beta_1' t'' \,, \qquad (9.16)$$

jehož parametry β_0', β_1' (regresní koeficienty) odhadneme metodou nejmenších čtverců a obdržíme tak jejich odhady b_0', b_1'. Ze vztahů (9.15) vypočteme zpětně odhady b_0, b_1:

$$b_0 = e^{b_0'} \,, b_1 = b_1' \,.$$

9.5 EXPONENCIÁLNÍ TREND

Exponenciální trendová funkce má tvar:

$$T_t = \beta_0 \beta_1^t \,, \qquad (9.17)$$

který substitucemi:

$$T_t' = \ln T_t \,, \; t'' = t \,, \qquad (9.18)$$

$$\beta_0' = \ln \beta_0 \,, \; \beta_1' = \ln \beta_1 \,, \qquad (9.19)$$

lze rovněž transformovat na „čárkovaný“ lineární trend, jehož parametry β_0', β_1' odhadneme metodou nejmenších čtverců, a obdržíme tak odhady b_0', b_1'. Ze vztahů (9.19) vypočteme odhady b_0, b_1 původního nelineárního regresního modelu (9.17):

$$b_0 = e^{b_0'} \,, b_1 = e^{b_1'} \,.$$

Použití exponenciálního trendu bude demonstrováno na příkladu v závěru této kapitoly.

9.6 LOGISTICKÝ TREND

Logistická trendová funkce patří k nelineárním trendům, které se vyznačují horní asymptotou, tj. hranicí, k níž se hodnoty ukazatele přibližují pro neomezeně rostoucí hodnoty času, a jedním inflexním bodem, v němž graf logistické funkce přechází z konvexního do konkávního tvaru. Pro tvar podobný písmenu *S* se takovým křivkám říká *S-křivky*. V ekonomické oblasti, speciálně v marketingu, se tato funkce používá při modelování poptávky po zboží dlouhodobé spotřeby, ale také při modelování vývoje výroby a prodeje některých druhů výrobků.

Na rozdíl od předchozích trendových funkcí, které byly definovány jednoznačně, logistická funkce bývá vyjadřována v několika různých variantách, uvedeme zde nejpoužívanější tvar:

$$T_t = \frac{\kappa}{1+\beta_0\beta_1^t}, \qquad (9.20)$$

kde β_0, β_1, κ jsou neznámé parametry a t = 1,2,...,n je časová proměnná, přitom se kvůli zachování tvaru *S*-křivky předpokládá, že 0 < κ; $0 < \beta_0, 0 < \beta_1 < 1$. Odhady neznámých parametrů, označujeme je b_0, b_1, *k,* lze opět získat metodou nejmenších čtverců, která dává nejlepší výsledky, i když vede na řešení soustavy nelineárních rovnic vyžadující použití složitějších výpočetních metod - *iteračních metod*. Proto zde ukážeme jinou metodu výpočtu neznámých parametrů, která sice nevede z teoretického pohledu k nejlepším odhadům, avšak její výhoda spočívá ve výpočetní nenáročnosti umožňující „ruční" výpočet. Tato metoda se nazývá *metoda vybraných bodů* a spočívá v tom, že z daných údajů časové řady vybereme 3 charakteristické hodnoty - body, kterými necháme logistickou trendovou křivku procházet, jinými slovy, položíme empirické hodnoty rovny hodnotám teoretickým. Jestliže charakteristické hodnoty $T_{t_1}, T_{t_2}, T_{t_3}$ odpovídají časovým okamžikům t_1, t_2, t_3, kde $t_1 < t_2 < t_3$, pak ze vztahu (4.33) obdržíme soustavu 3 rovnic o 3 neznámých β_0, β_1, κ:

$$T_{t_1} = \frac{\kappa}{1+\beta_0\beta_1^{t_1}},\ T_{t_2} = \frac{\kappa}{1+\beta_0\beta_1^{t_2}},\ T_{t_3} = \frac{\kappa}{1+\beta_0\beta_1^{t_3}}, \qquad (9.21)$$

jejichž řešením získáme odhady neznámých parametrů b_0, b_1, *k*. Výpočty v metodě vybraných bodů můžeme usnadnit, když charakteristické body zvolíme ekvidistantně:

$$t_1 = 0,\ t_2 = \Delta,\ t_3 = 2\Delta,$$

kde Δ je určitý časový interval. Za tohoto předpokladu je řešení soustavy následující:

$$b_0 = \frac{k - T_{t_1}}{T_{t_1}},$$

$$b_1 = \left(\frac{T_{t_1}\left(k - T_{t_2}\right)}{T_{t_2}\left(k - T_{t_1}\right)}\right)^{\frac{1}{t_2}}, \qquad (9.22)$$

$$k = \frac{2T_{t_1}T_{t_2}T_{t_3} - T_{t_2}^2\left(T_{t_1} + T_{t_3}\right)}{T_{t_1}T_{t_3} - T_{t_2}^2}. \qquad (9.23)$$

Z výše uvedeného vztahu (9.23) lze přímo vypočíst parametr *k*, jeho dosazením do vztahu (9.22) vypočítáme parametry b_0, b_1. Jak se snadno zjistí, hodnota asymptoty logistické křivky je $\frac{\kappa}{1+\beta_0}$, což představuje horní mez, k níž se limitně přibližuje hodnota trendové funkce při velkých hodnotách času *t*.

9.7 Gompertzův trend

Ve srovnání s předchozí logistickou trendovou funkcí je *Gompertzův trend* jiným typem *S*-křivky:

$$T_t = \kappa\beta_0^{\beta_1^t}, \tag{9.24}$$

kde opět β_0, β_1, κ jsou neznámé parametry a t = 1,2,...,n je časová proměnná, přitom se kvůli zachování tvaru S-křivky předpokládá, že 0 < κ; $0 < \beta_0, 0 < \beta_1 < 1$. Odhady b_0, b_1, těchto *k* parametrů získáme opět metodou nelineární regrese (metodou nejmenších čtverců), eventuálně metodou vybraných bodů, jako v předchozím odstavci. Asymptota Gompertzovy křivky je rovnoběžná s osou t ve vzdálenosti *k*, přičemž inflexní bod křivky není na rozdíl od logistického trendu (9.20) umístěn uprostřed mezi časovou osou a asymptotou.

9.8 Volba vhodného modelu trendu

Závažným problémem analýzy časových řad je problém stanovení konkrétního typu trendové funkce. Základem pro rozhodnutí o vhodném typu funkce by měla být věcně-ekonomická kritéria, tedy trendová funkce by měla být volena na základě věcné analýzy zkoumaného ekonomického jevu. Během věcného rozboru lze obvykle posoudit, zda jde o funkci rostoucí (nebo klesající), s trendem růstu nade všechny meze, či k určité konečné hodnotě (asymptotě).

Grafické znázornění časové řady umožní v hrubých rysech odhalit základní tendence ve vývoji analyzovaného ukazatele. Nebezpečí volby na základě vizuálního výběru spočívá však v jeho subjektivitě. Různí analytici mohou danou situaci posoudit různě a zvolit rozdílné typy trendové funkce. Nebezpečí tu plyne i z toho, že tvar grafu je do značné míry závislý na volbě použitého měřítka.

Přiléhavost dat k trendové (regresní) křivce jsme v kapitole 3 měřili koeficientem determinace R^2, viz (3.18):

$$R^2 = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y}. \tag{9.25}$$

Tento koeficient můžeme k porovnání vhodnosti různých modelů trendu použít i nyní. V zásadě lze přijmout hodnocení, v němž nejvhodnější model trendu dává nejvyšší hodnotu koeficientu determinace R^2. Vzhledem k tomu, že hodnota S_y je dána, závisí velikost R^2 na velikosti reziduálního součtu čtverců S_R; čím je jeho hodnota menší, tím je hodnota R^2 větší (blíže k jedné). Taková metoda hodnocení trendu časové řady však upřednostňuje modely s větším počtem parametrů. Protože se zejména u ekonomických časových řad snažíme o nalezení jednoduchého tvaru trendu, je lepší k hodnocení vhodnosti modelu použít *reziduální rozptyl*:

$$s_R^2 = \frac{S_R}{n-p}, \tag{9.26}$$

kde

$$S_R = \sum_{i=1}^{n} \left(y_i - Y_i\right)^2$$

je reziduální součet čtverců, n je počet datových bodů a p je počet parametrů v modelu. Z tvaru (9.26) je zřejmé, že hodnota reziduálního rozptylu roste s rostoucím počtem parametrů, což odpovídá výše uvedenému požadavku po co nejmenším počtu parametru v trendové funkci. Vhodný model trendu bude tedy „kompromisem“ mezi velikostmi hodnot R^2 a p.

Volbu vhodné trendové funkce lze podpořit také *testy hypotéz*. Z celé řady různých testů uvedeme známý F-test, který slouží pro rozhodování, zda má smysl dávat přednost složitějšímu modelu (s větším počtem parametrů) před jednodušším modelem (s menším počtem parametrů). Testujeme nulovou hypotézu, že totiž pokud jde o přiléhavost dat ke zvoleným trendovým funkcím, není mezi modely statisticky významný rozdíl. Tento test je založen na statistice:

$$F = \frac{\dfrac{S_T^{(2)} - S_T^{(1)}}{p_1 - p_2}}{\dfrac{S_R^{(1)}}{n - p_1}}, \qquad (9.27)$$

kde hodnoty $S_T^{(1)}, S_R^{(1)}, p_1$ přísluší ke složitějšímu modelu, hodnoty $S_T^{(2)}, p_2$ přísluší k jednoduššímu modelu, tj. $p_1 > p_2$, $S_T^{(2)} > S_T^{(1)}$. Statistika (9.27) má přibližně Fisherovo rozdělení F s $p_1 - p_2$ a $n - p_1$ stupni volnosti. V případě, že vypočítaná hodnota statistiky padne do kritického oboru, lze na zvolené hladině významnosti α usuzovat, že model s větším počtem parametrů přináší výrazné zlepšení oproti jednoduššímu modelu.

9.9 Klouzavé průměry

V posledních dvou odstavcích této kapitoly věnované trendovým funkcím se budeme zabývat dvěma metodami syntetického způsobu stanovení trendu časové řady. Jak již bylo řečeno, *syntetický přístup stanovení trendu* spočívá ve vyhlazení a vyrovnání odchylek daného ukazatele v časové řadě takovým způsobem, že získané vyrovnané hodnoty vyjadřují trendový faktor obsažený pouze v časové řadě, nikoliv však vnější faktory. Nemusíme proto znát předem typ trendové funkce, což je přednost syntetického přístupu oproti přístupu analytickému, kde jsme typ trendové funkce museli stanovit předem. Jeho nevýhodou je naopak obtížnější využití pro prognózování hodnot časové řady. Z existujících metod syntetického přístupu uvedeme metody *klouzavého průměru* a *exponenciální vyrovnání*.

Podstata vyrovnání časové řady pomocí klouzavých průměrů spočívá v tom, že posloupnost hodnot časové řady nahradíme novou řadou průměrů vypočítaných s kratších úseků časové řady, přičemž tyto kratší úseky postupně posouváme (kloužeme) směrem od začátku ke konci časové řady, a současně vypočítáváme dílčí průměry - *klouzavé průměry*. Vzniká důležitý problém, který je nutno předem řešit: jaký má být počet členů klouzavé části průměru. Klouzavou částí průměru budeme tedy rozumět časový interval určité délky, který se posunuje po časové ose vždy o jednotku. Volba rozsahu klouzavé části závisí na věcném (ekonomickém) charakteru časové řady a nelze ji obvykle stanovit na podkladě exaktních statistických metod. V praxi jsou u ekonomických neperiodických časových řad voleny většinou klouzavé části menší liché délky, např. 3, 5 nebo 7 časových jednotek, což souvisí se snadnější interpretovatelností výsledků, neboť pak můžeme hodnotu klouzavého průměru přiřadit prostřednímu časovému okamžiku klouzavé části. U periodických časových řad se volí délka klouzavých části totožná s délkou periody (sezóny, cyklu).

Uvažujme časovou řadu $y_1, y_2, y_3, ... y_n$. *Prosté klouzavé průměry* získáme tak, že úseky časové řady o délce $m = 2p+1$, přičemž $m < n$, $p \geq 1$, celé číslo, vyrovnáme lineárním trendem s využitím metody nejmenších čtverců. Výsledkem je vzorec pro hodnoty vyrovnané časové řady ve formě aritmetického průměru:

$$\bar{y}_t = \frac{1}{2p+1}\sum_{t=-p}^{p} y_{t+i} = \frac{y_{t-p} + y_{t-p+1} + ... + y_{t+p-1} + y_{t+p}}{2p+1}, \tag{9.28}$$

kde $t = p+1, p+2, ..., n-p$. Přitom p hodnot na začátku a p hodnot na konci časové řady zůstává nevyrovnáno.

Kromě prostých klouzavých průměrů se někdy používají složitější *vážené klouzavé průměry*, případně *centrované klouzavé průměry*. Ty získáme tak, že namísto lineárního trendu v každém úseku použijeme polynomický trend vyššího řádu, tj. kvadratickou parabolu, kubickou parabolu apod. Metodou nejmenších čtverců obdržíme poměrně složité vzorce pro výpočet vyrovnaných hodnot. Vzhledem k poměrně řídkému použití těchto složitějších klouzavých průměrů se jimi zde nebudeme dále zabývat. Zájemce odkazujeme na literaturu, např. Seger (1998).

9.10 EXPONENCIÁLNÍ VYROVNÁNÍ

Další metodou vyhlazování časové řady, tedy syntetického stanovení trendu, je *exponenciální vyrovnání*. Při něm se nová vyrovnaná hodnota stanoví na základě exponenciálně váženého průměru současné hodnoty a všech předchozích hodnot časové řady. Přitom se používá systém koeficientů - *vah*, kdy novější hodnota má vždy větší váhu (tj. důležitost), než hodnota starší.

Nechť y_t značí pozorovanou hodnotu v časovém okamžiku t, w je váha přiřazená současné hodnotě, $0 < w < 1$, $\hat{y}_t$ je vyrovnaná hodnota v čase t. Metoda exponenciálního vyrovnání začíná tím, že první vyrovnanou hodnotu časové řady $\hat{y}_1$ (v čase 1) položíme rovnu pozorované hodnotě y_1, tedy:

$$\hat{y}_1 = y_1 .$$

Následující vyrovnané hodnoty definujeme rekurentním vztahem:

$$\hat{y}_t = wy_t + (1-w)\hat{y}_{t-1}, \; t = 2,3,...n, \tag{9.29}$$

který umožňuje postupně vypočítat všechny vyrovnané hodnoty dané časové řady. Ze vztahu (9.28) lze snadno odvodit vztah:

$$\hat{y}_t = wy_t + w(1-w)y_{t-1} + w(1-w)^2 y_{t-2} + ... + w(1-w)^{t-2} y_2 + (1-w)^{t-1} y_1 .$$

Z posledního vztahu je vidět, že vyrovnaná hodnota časové řady v čase t závisí na všech předchozích nevyrovnaných hodnotách s tím, že do celkového součtu vstupují starší hodnoty s menší vahou

$$w_{t-i} = w(1-w)^i , \tag{9.30}$$

kde $i = 0,1,...,t-2$. Vzhledem k tomu, že platí $0 < w < 1$, je zřejmé, že se hodnota w_{t-i} exponenciálně zmenšuje s rostoucím i, tj. rostoucím „stářím“ dat. Váhu w nazýváme *koeficient exponenciálního zapomínání*. Ze vztahu (9.30) vyplývá, že čím vyšší je koeficient zapomínání, tím menší je hodnota $(1-w)$, a tedy také $(1-w)^i$, což znamená, že váha - význam starších dat klesá, starší data se rychleji zapomínají. Je-li např. $w = 0{,}9$, tedy koeficient zapomínání je 90%, potom za jednotku času se vliv hodnoty y_{t-i} zmenší na

$(1-w)y_{t-i} = 0{,}1y_{t-i}$, což znamená, že se „zapomene“ 90% hodnoty. V praxi se používají obvykle váhy z intervalu 0,7 až 1,0. Pro výpočet exponenciálně vyrovnaných hodnot časové řady je ovšem výhodnější rekurentní vztah (9.29).

Kromě výše uvedené metody se v praxi využívají i složitější postupy exponenciálního vyrovnání, které se zařazují do skupiny metod, kterým se říká *adaptivní metody*. Zájemce odkazujeme např. na práce Seger (1998), Cipra (1986).

ŘEŠENÝ PŘÍKLAD 9.1

V následující tabulce jsou uvedeny počty prodaných automobilů v autocentru A+A v letech 2000 až 2007. Pro tuto časovou řadu vypočítejte:

Rok	2000	2001	2002	2003	2004	2005	2006	2007
Počet	120	159	167	175	197	172	199	240

a. Trend v prodeji automobilů popište lineární trendovou funkcí.
b. Jaký počet prodaných automobilů lze očekávat v roce 2008 s 95% pravděpodobností? (Stanovte bodový odhad a 95%-ní interval spolehlivosti prognózy.)
c. Stanovte koeficient determinace a na jeho základě určete přiléhavost dat k trendové funkci.

Řešení:

a. Podle vztahu (9.2) zavedeme novou časovou proměnnou t' (viz následující tabulka).

Rok	t'	y_t	t'^2	$y_t t'$	$\hat{T}$	$(y-\hat{T})^2$	$(y-\bar{y})^2$
2000	-7	120	49	–840	133,818	190,937	3436,891
2001	-5	159	25	–795	146,620	153,264	385,141
2002	-3	167	9	–501	159,422	57,426	135,141
2003	-1	175	1	–175	172,224	7,706	13,141
2004	1	197	1	197	185,026	143,377	337,641
2005	3	172	9	516	197,828	667,086	43,891
2006	5	199	25	995	210,630	135,257	415,141
2007	7	240	49	1680	223,432	274,499	3766,891
Součet	0	1429	168	1077		1629,552	8533,875

Odhady b_0, b_1 parametrů β_0, β_1 trendové funkce:

$$T_t = \beta_0 + \beta_1 t', t' = -7, -5, -3, \ldots$$

vypočítáme podle vztahů:

$$b_0 = \frac{\sum y_t}{n} = \frac{1429}{8} = 178{,}625 \ , \qquad b_1 = \frac{\sum t' y_t}{\sum t'^2} = \frac{1077}{168} = 6{,}410.$$

Odhadnutá trendová funkce má tvar:

$$\hat{T} = 178{,}625 + 6{,}41\, t' ,\ t' = -7, -5, -3, \ldots$$

b. Očekávaný prodej v roce 2008 vypočítáme dosazením t', které odpovídá roku 2008, do rovnice trendu:

$$\hat{T} = 178{,}625 + 6{,}401 \cdot 9 \cong 236{,}32.$$

Intervalovou předpověď obdržíme dosazením potřebných hodnot do vztahu (4.8). Ve speciálním případě časové řady, kdy $t_i = x_i$, obdržíme po úpravách následující vztah pro interval spolehlivosti predikce na i časových okamžiků dopředu:

$$[\, y(n+i) - t_{1-\alpha/2}(n-2)\, s_R \sqrt{Q_n(i)}\,,\ y(n+i) + t_{1-\alpha/2}(n-2)\, s_R \sqrt{Q_n(i)}\,],$$

kde $y(n+i) = \hat{T} = 236{,}32$,

$t_{1-\alpha/2}(n-2) = 2{,}45$,

$$s_R = \sqrt{\frac{S_R}{n-p}}\,,$$

$$Q_n(i) = \sqrt{(1-R^2)\frac{n(n^2-1)+12i^2}{(n^2-1)(n-2)}}\,,\ i = 1.$$

Z tabulky obdržíte $S_R = 1629{,}552$. Potom směrodatná chyba odhadu s_R je

$$s_R = \sqrt{\frac{1629{,}552}{8-2}} = 16{,}48\,.$$

K výpočtu $Q_n(i)$ je zapotřebí znát hodnotu koeficientu determinace R^2 , tj.

$$R^2 = 1 - \frac{S_R}{S_y} = 1 - \frac{1629{,}552}{8533{,}875} = 0{,}809\,.$$

Výpočet součtu S_y je uveden v tabulce. Potom

$$Q_n(i) = \sqrt{(1-0{,}809)\frac{8(64-1)+12}{(64-1)(8-2)}} = \sqrt{0{,}191\cdot\frac{516}{378}} = 0{,}51\,.$$

Dosazením výše vypočítaných hodnot do obecného vztahu obdržíte levou (L) a pravou (P) mez intervalové předpovědi.

$$L = 236{,}315 - 2{,}447\cdot 16{,}48\cdot\sqrt{0{,}51} = 207{,}52.$$

$$P = 236{,}315 + 2{,}447\cdot 16{,}48\cdot\sqrt{0{,}51} = 265{,}11.$$

Bodový odhad prodeje v roce 2008 je 236 automobilů. S 95% pravděpodobností by se mělo v roce 2008 prodat mezi 208 a 265 automobily.

c. Koeficient determinace byl vypočten v **b**: $R^2 = 0{,}809$. Tato hodnota říká, že přiléhavost dat k trendové funkci je „vysoká“.

ŘEŠENÝ PŘÍKLAD 9.2

V tabulce jsou uvedeny údaje o počtu vyrobených myček nádobí v letech 2004-2012.

a. Trend ve výrobě tohoto výrobku popište exponenciální trendovou funkcí.

b. Vypočítejte bodovou prognózu výroby na rok 2013, dále zjistěte koeficient determinace a na jeho základě zhodnoťte „přiléhavost“ dat k trendové funkci.

Rok	2004	2005	2006	2007	2008	2009	2010	2011	2012
Myčky nádobí (tis. ks)	8	9	17	20	38	40	70	101	180

Řešení:

Nejprve vypočítáte odhady b_0, b_1 parametrů exponenciální trendové funkce

$$T_t = \beta_0 \beta_1^t .$$

Logaritmováním této rovnice obdržíte vztah

$$\ln T_t = \ln \beta_0 + t \ln \beta_1 .$$

Zavedením substituce

$$T_t' = \ln T_t ,\ t' = t ,$$

$$\beta_0' = \ln \beta_0 ,\ \beta_1' = \ln \beta_1$$

se původní rovnice exponenciálního trendu transformuje na rovnici lineárního trendu.
Zavedete novou časovou proměnnou t'' viz (9.1) a vypočítáte koeficienty b_0', b_1'

$$b_0' = \frac{\sum y_t'}{n} = \frac{31{,}4886}{9} = 3{,}4987 , \qquad b_1' = \frac{\sum t'' y_t'}{\sum t''^2} = \frac{23{,}2315}{60} = 0{,}3872 .$$

Potom

$$b_0 = e^{b_0'} = e^{3{,}4987} = 33{,}07 , \qquad b_1 = e^{b_1'} = e^{0{,}3872} = 1{,}47 .$$

Rok	t''	y	$y' = \ln y$	t''^2	$t'' y'$	T	$(y - T)^2$	$(y - \bar{y})^2$
2004	-4	8	2,0794	16	-8,3178	7,0285	0,8425	2085,7489
2005	-3	9	2,1972	9	-6,5917	10,3519	1,9904	1995,4089
2006	-2	17	2,8332	4	-5,6664	15,2466	2,8771	1344,6889
2007	-1	20	2,9957	1	-2,9957	22,4558	6,2330	1133,6689
2008	0	38	3,6376	0	0	33,0737	24,3049	245,5489
2009	1	40	3,6889	1	3,6889	48,7122	74,1821	186,8689
2010	2	70	4,2485	4	8,4970	71,7452	2,1345	266,6689
2011	3	101	4,6151	9	13,8453	105,6690	16,3831	2240,1289
2012	4	180	5,1930	16	20,7718	155,6333	654,3364	15959,2689
Součet	0	490	31,4886	60	23,2315		783,2839	25458,0001

Hledaná trendová funkce má tvar

$$\hat{T}_{t''} = 33{,}07 \cdot 1{,}47^{t''} , t'' = -4, -3, -2, \ldots .$$

K bodovému odhadu využijeme nalezenou trendovou funkci, kam dosadíme t“ = 5, což je hodnota, která odpovídá netransformované časové hodnotě t = 2013.

Koeficient determinace vyžaduje znát hodnotu celkového součtu S_y a reziduálního součtu S_R (viz poslední dva sloupce v tabulce).
Pro výpočet reziduálního součtu čtverců je dále třeba znát odhady teoretické hodnoty $\hat{T}_{t''}$, které obdržíme postupným dosazováním za t'' do rovnice trendu, tedy např. pro $t'' = -4$:

$$\hat{T} = 33{,}07 \cdot 1{,}47^{-4} = 7{,}08.$$

Všechny hodnoty $\hat{T}$ i součtů S_y, S_R najdete v tabulce. Pro koeficient determinace platí:

$$R^2 = 1 - \frac{S_R}{S_y} = 1 - \frac{783{,}2839}{25458{,}0001} = 0{,}969.$$

Hodnota 0,969 říká, že přiléhavost dat k trendové křivce je vysoká.

ŘEŠENÝ PŘÍKLAD 9.3

V tabulce jsou uvedeny údaje o počtu výrobků určitého typu (v tis. ks) v letech 1999 - 2009. Nalezněte logistickou trendovou funkci, která charakterizuje trend dané časové řady. Prognózujte výrobu pomocí bodového odhadu na rok 2012.

Čas	1999	2000	2001	2002	2003	2004	2006	2006	2007	2008	2009
Zjištěné hodnoty	5	6	9	16	22	25	32	34	41	44	45

Řešení:

Hledáme odhady parametrů trendové funkce ve tvaru (9.20)

$$T_t = \frac{\kappa}{1+\beta_0\beta_1^t}.$$

Tyto odhady stanovíte metodou vybraných bodů. Abyste mohli k výpočtu použít vztahy (9.21), (9.22), (9.23), zvolíte opět novou časovou proměnnou t', viz následující tabulka. Ze všech údajů v časové řadě vyberete tři časové okamžiky, např. na počátku, uprostřed a na konci časové osy: $t_1' = 0, t_2' = 5, t_3' = 10$. V těchto okamžicích (jsou vyznačeny tučně) položíte empirické hodnoty rovny hodnotám teoretickým, tedy $T_{t_1'} = 5, T_{t_2'} = 25, T_{t_3'} = 45$.

t	1999	2000	2001	2002	2003	2004	2006	2006	2007	2008	2009
t'	**0**	1	2	3	4	**5**	6	7	8	9	**10**
Zjištěné hodnoty	**5**	6	9	16	22	**25**	32	34	41	44	**45**

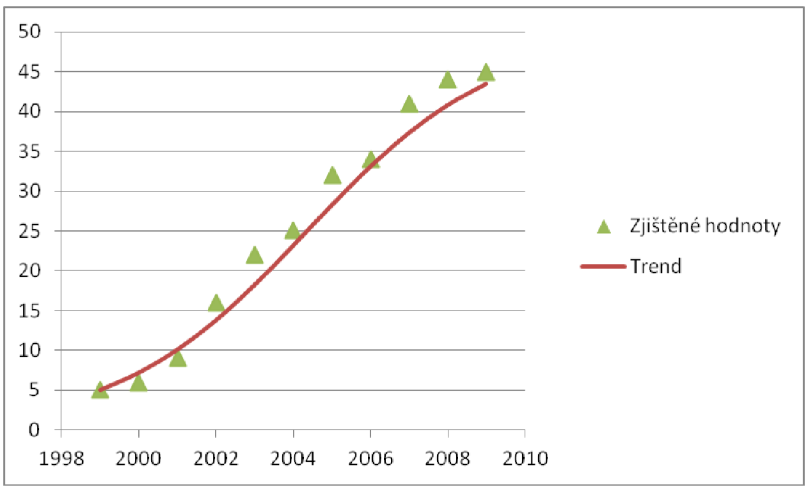
Potom ze vztahů (9.22), (9.23) postupně vypočítáte:

$$k = \frac{2T_{t_1'}T_{t_2'}T_{t_3'} - T_{t_2'}^2(T_{t_1'} + T_{t_3'})}{T_{t_1'}T_{t_3'} - T_{t_2'}^2} = \frac{2\cdot 5\cdot 25\cdot 45 - 25^2(5+45)}{5\cdot 45 - 25^2} = 50,$$

$$b_0 = \frac{k - T_{t_1'}}{T_{t_1'}} = \frac{50-5}{5} = 9, \qquad b_1 = \left(\frac{T_{t_1'}(k - T_{t_2'})}{T_{t_2'}(k - T_{t_1'})}\right)^{\frac{1}{t_2'}} = \left(\frac{5(50-25)}{25(50-5)}\right)^{\frac{1}{5}} = 0{,}644.$$

Odhadovaný logistický trend má tvar

$$\hat{T}_{t'} = \frac{50}{1+9\cdot 0{,}644^{t'}}.$$



Obr. 9.1. *Logistický trend*

Rok 2012 odpovídá v transformované časové ose hodnotě $t' = 13$. Dosazením do rovnice zjištěné trendové funkce obdržíte

$$\hat{T}_{2008} = \frac{50}{1 + 9 \cdot 0{,}644^{13}} = 48{,}57 \cong 49,$$

tj. prognózovaná výroba daného výrobku v roce 2012 je 49 tis. ks.

ŘEŠENÝ PŘÍKLAD 9.4

V následující tabulce jsou uvedeny údaje o spotřebě pitné vody v jednotlivých dnech tří po sobě jdoucích týdnů.

a. Stanovte odpovídající interval klouzavého průměru a vyrovnejte tuto řadu prostými klouzavými průměry.

b. Vyrovnejte časovou řadu pomocí metody exponenciálního vyrovnání, použijte koeficient zapomínání $w = 0{,}7$.

Po	0,64	0,75	0,54
Út	0,78	0,63	0,61
St	0,93	0,82	0,7
Čt	0,66	0,63	0,56
Pá	0,99	1,3	0,79
So	1,22	0,65	1,3
Ne	1,05	1,3	1,24

Řešení:

a. Z charakteru dat vyplývá, že pro analyzovanou časovou řadu budou vhodné klouzavé průměry o délce $m = 7$ pozorování, tj. v rámci týdne. Použijete proto prosté7-členné klouzavé průměry, které vypočítáte podle vztahu (9.28):

$$\bar{y}_1 = \frac{y_1 + y_2 + \ldots + y_7}{7} = \frac{0{,}64 + 0{,}78 + 0{,}93 + 0{,}66 + 0{,}99 + 1{,}22 + 1{,}05}{7} = 0{,}896.$$

Tuto hodnotu přiřadíte prostřednímu časovému okamžiku klouzavé části, tj. ke čtvrté hodnotě dané časové řady.

Druhý klouzavý průměr vypočítáte analogicky posunutím o jeden den a přiřadíte jej k páté hodnotě původní časové řady:

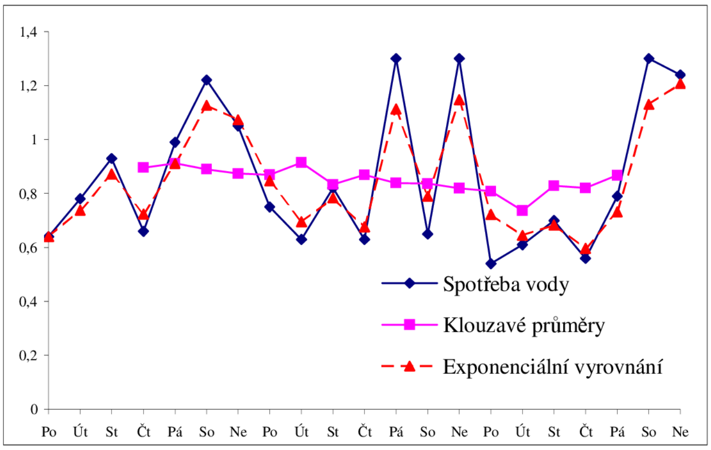
$$\bar{y}_2 = \frac{y_2 + y_3 + \ldots + y_8}{7} = \frac{0{,}78 + 0{,}93 + 0{,}66 + 0{,}99 + 1{,}22 + 1{,}05 + 0{,}75}{7} = 0{,}911.$$

Ostatní klouzavé průměry vypočítáte obdobně postupným klouzáním směrem ke konci časové řady. Empirické hodnoty jakož i klouzavé průměry ukazuje Obr. 9.2 .

b. Exponenciální vyrovnání se provede podle (9.29):

$$\hat{y}_1 = y_1,$$

$$\hat{y}_t = w y_t + (1 - w)\hat{y}_{t-1}, \; t = 2,3,\ldots n, \qquad \text{kde } w = 0{,}7.$$



Obr. 9.2. *Klouzavé průměry a exponenciální vyrovnání*

Proto:

$$\hat{y}_1 = 0{,}64,$$

$$\hat{y}_2 = 0{,}7y_2 + (1 - 0{,}7)\cdot \hat{y}_1 = 0{,}7\cdot 0{,}78 + 0{,}3\cdot 0{,}64 = 0{,}738.$$

Další hodnoty $\hat{y}_t$ vypočítáme rekurentně, viz následující tabulka.

Den	Spotřeba vody (m^3/os.)	Klouzavé průměry	Exponenciální vyrovnání
Po	0,64		0,640
Út	0,78		0,738
St	0,93		0,872
Čt	0,66	0,896	0,724
Pá	0,99	0,911	0,910
So	1,22	0,890	1,127
Ne	1,05	0,874	1,073
Po	0,75	0,870	0,847
Út	0,63	0,914	0,695
St	0,82	0,833	0,783
Čt	0,63	0,869	0,676
Pá	1,30	0,839	1,113
So	0,65	0,836	0,789
Ne	1,30	0,819	1,147
Po	0,54	0,809	0,722
Út	0,61	0,736	0,644
St	0,70	0,829	0,683
Čt	0,56	0,820	0,597
Pá	0,79	0,867	0,732
So	1,30		1,130
Ne	1,24		1,207

Je zřejmé, že koeficient zapomínání w = 0,7 ještě nevyhlazuje původní data dostatečně, k většímu vyhlazení by byla zapotřebí menší hodnota koeficientu zapomínání.

Zopakujme si získané poznatky této kapitoly: Jediným faktorem vývoje dynamiky analyzovaného ukazatele byl zde čas. Trendová složka představuje nejdůležitější komponentu analyzované časové řady, a proto popis trendu je jedním z nejdůležitějších úkolů analýzy časových řad. K určení trendové složky jsme použili dva obecné přístupy: analytický a syntetický. Analytický přístup stanovení trendu vychází z předem známých typů trendových funkcí vyznačujících se přítomností parametrů, které je třeba stanovit co nejlépe s ohledem na skutečné hodnoty ukazatele časové řady. Z velkého množství používaných trendových funkcí jsme se zaměřili na několik, které mají význam především v ekonomických aplikacích. Byly to: lineární trend, parabolický trend, exponenciální trend, logistický trend a Gompertzův trend. Syntetický přístup stanovení trendu spočívá ve vyrovnání odchylek daného ukazatele v časové řadě tak, že získané vyrovnané hodnoty vyjadřují trendový faktor obsažený pouze v časové řadě, nikoliv faktor vložený z vnějšku. Nemuseli jste proto znát předem typ trendové funkce, což je přednost syntetického přístupu oproti přístupu analytickému. Jeho nevýhodou je naopak obtížnější využití pro prognózování hodnot časové řady. Z existujících metod syntetického přístupu jsme uvedli metodu klouzavého průměru a jednoduché exponenciální vyrovnání.

9.11 SAMOSTATNÉ ÚKOLY

9.1 V tabulce jsou uvedeny údaje o počtu vyrobených kuchyňských robotů v letech 2001 až 2011.

Rok	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Kuchyňské roboty (tis. ks)	5	4	8	16	35	32	40	56	100	120	195

a. Trend ve výrobě tohoto výrobku popište exponenciální trendovou funkcí.
b. Jaké množství vyrobených kuchyňských robotů lze očekávat v roce 2014?
c. Znaménkovým testem ověřte na hladině významnosti $\alpha = 0,05$ náhodnost reziduí.

9.2 Následující časová řada představuje počet vyrobených pneumatik Barum v letech 2001 až 2012.

Rok	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Pneumatiky (mil.ks)	0,8	1,6	1,5	2,4	5	3,88	4,47	3,88	6,89	7,69	5,83	8,25

a. Nalezněte lineární trend časové řady.
b. Jaké množství vyrobených pneumatik lze očekávat v roce 2013? Stanovte bodový i intervalový odhad na hladině významnosti $\alpha = 0,05$.

9.12 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

9.1 **a)** $\hat{T} = 29{,}55 \cdot 1{,}47^t, t = -5, -4, -3, \ldots$

b) v roce 2014, tzn. $t = 8$; $\hat{T} = 644{,}31$.

c) $S = 5$; testové kritérium $U = 0$; obor přijetí $A = (-1{,}96;\ 1{,}96)$; přijímáme nulovou hypotézu o náhodném uspořádání reziduí

9.2 **a)** $\hat{T} = 4{,}35 + 0{,}32.t,\ \ t = -11, -9, -7, \ldots$

b) v roce 2013, tzn. $t = 13$; $\hat{T} = 8{,}5$ mil.ks ; 95%-ní intervalový odhad (5,97; 11,05)

10 ANALÝZA SEZÓNNÍ SLOŽKY A NÁHODNÉ SLOŽKY

RYCHLÝ NÁHLED KAPITOLY

Při analýze ekonomických časových řad se setkáváme téměř vždy s existencí sezónních vlivů, reprezentovaných v modelu časové řady sezónní složkou. Sezónními vlivy rozumíme soubor příčin, které se pravidelně opakují v důsledku koloběhu přírody. Důsledkem působení sezónních vlivů na analyzovanou časovou řadu jsou pravidelné výkyvy nahoru a dolů vůči určitému normálnímu vývoji. Pokud se u časových řad vyskytují podobné vlivy v delším časovém horizontu, hovoříme o cyklické složce časové řady, v kratším časovém horizontu, hovoříme o sezónní složce časové řady. Souhrnně se sezónní a cyklické složky označují jako periodické složky časové řady. Úkolem modelování periodické složky časové řady je nalézt její vhodné vyjádření, které by umožnilo periodickou (nejčastěji sezónní) složku nejen vhodně identifikovat, ale i následně použít k predikci chování časové řady v budoucnu. V této kapitole se budete zabývat časovými řadami, jejichž hodnoty se periodicky opakují: jedná se o sezónní časové řady. Nejprve si objasníte význam sezónní složky časové řady, poté se věnujete metodě harmonické analýzy, která k modelování časové řady využívá známé matematické periodické funkce sinus a kosinus. Poté se naučíte aplikovat jednoduché metody konstantní sezónnosti se schodovitým a lineárním trendem a rovněž metodu proporcionální sezónnosti. V závěru se budete věnovat analýze náhodné složky. Zmíněné metody si ozřejmíte na konkrétních příkladech řešených s využitím Excelu.

10.1 PERIODICKÁ SLOŽKA ČASOVÝCH ŘAD

Při analýze časových řad s periodicitou kratší než jeden rok se setkáváme téměř vždy s existencí sezónních vlivů, reprezentovaných v modelu časové řady sezónní složkou. *Sezónními vlivy* rozumíme soubor příčin, které se pravidelně opakují v důsledku koloběhu přírody. Důsledkem působení sezónních vlivů na analyzovanou časovou řadu jsou pravidelné výkyvy nahoru a dolů vůči určitému normálnímu vývoji. Pokud se u časových řad vyskytují podobné vlivy v delším časovém horizontu, hovoříme o *cyklické složce* časové řady. Souhrnně se sezónní a cyklické složky označují jako *periodické složky* časové řady, takže model časové řady (8.9) můžeme vyjádřit ve tvaru

$$y_t = T_t + P_t + \varepsilon_t, \qquad t = 1,2,...n, \tag{10.1}$$

kde P_t je periodická složka $P_t = S_t + C_t$, S_t - sezónní složka, C_t - cyklická složka. Úkolem popisu periodické složky časové řady je nalézt její vhodné modelové vyjádření, které by umožnilo periodickou (nejčastěji sezónní) složku nejen vhodně identifikovat, ale i následně použít k predikci chování časové řady v budoucnu.

10.2 HARMONICKÁ ANALÝZA

Pro vyjádření periodické složky časových řad se využívají různé modely. Mezi nejčastěji používané modely sezónní složky patří *harmonická analýza*. Základní ideou tohoto přístupu je vyjádřit periodickou složku jako součet určitého počtu „vln“ známých periodických goniometrických křivek - funkcí sinus a kosinus. Jde tedy o následující model periodické složky časové řady:

$$P_t = \sum_{j=1}^{H} (\alpha_j \sin \omega_j t + \beta_j \cos \omega_j t), \; t = 1,2,...n, \tag{10.2}$$

Výraz (10.2) se nazývá *trigonometrický polynom*, přičemž jeho část v závorce za znakem sumace je *harmonický člen*. Přitom je přirozené číslo $H \leq n/2$, dále $\omega_j = \frac{2\pi j}{n}$ se nazývá *frekvence*, čas $t = 1,2,...,n$. Harmonický člen v (10.2) je možné vyjádřit také takto:

$$\alpha_j \sin \omega_j t + \beta_j \cos \omega_j t = A_j \sin(\omega_j t + \varphi_j), \tag{10.3}$$

přičemž platí:

$$\alpha_j^2 + \beta_j^2 = A_j^2, \quad \frac{\beta_j}{\alpha_j} = tg\,\varphi_j,$$

kde A_j se nazývá *amplituda*, φ_j je tzv. *fázový posuv*.

Předpokládejme, že máme dánu časovou řadu y_t, $t = 1,2,...n$, kde trend T_t je již stanoven například některým z postupů uvedených v předchozí kapitole. V takovém případě lze identifikovanou trendovou složku od hodnot časové řady odečíst, čímž dostaneme nový model časové řady, avšak s konstantním trendem μ. Model (10.1) pak má tvar:

$$y_t = \mu + P_t + \varepsilon_t, \qquad t = 1,2,...n, \tag{10.4}$$

kde periodická složka P_t je dána vztahem (10.2). Metodou nejmenších čtverců se vypočítají bodové odhady m, a_j, b_j regresních koeficientů μ, α_j, β_j takto:

$$m = \frac{1}{n}\sum_{t=1}^{n} y_t = \bar{y}, \tag{10.5}$$

$$a_j = \frac{2}{n}\sum_{t=1}^{n} y_t \sin \omega_j t, \tag{10.6}$$

$$b_j = \frac{2}{n}\sum_{t=1}^{n} y_t \cos \omega_j t, \; j = 1,2,...,H. \tag{10.7}$$

Přiléhavost modelu (10.2) k datům je dána koeficientem determinace:

$$R^2 = \frac{\mathrm{var}(P_t)}{\mathrm{var}(y_t)}, \tag{10.8}$$

kde

$$\mathrm{var}(P_t) = \frac{1}{2}\sum_{j=1}^{H}(a_j^2 + b_j^2), \tag{10.9}$$

$$\mathrm{var}(y_t) = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2. \tag{10.10}$$

Vyhledávání periodicity v časových řadách lze realizovat následujícími postupy:

- *Subjektivní odhady* zahrnují vizuální odhady z grafů, eventuálně z klouzavých průměrů. Tyto metody přes svoji jednoduchost a subjektivnost neztratily dosud význam, zvláště v situacích, kdy potřebujeme rychlé předběžné informace o chování časové řady.

- *Objektivní odhady* zahrnují některé složitější a také výpočetně náročnější metody a postupy jako autokorelační funkce, spektrální analýza a periodogram. Tyto nástroje slouží k hlubšímu a obecnějšímu zkoumání chování časových řad a jsou obvyklou součástí pokročilých statistických softwarových paketů, jako SPSS, STATISTICA, UNISTAT apod. Zde se budeme podrobněji věnovat periodogramu, který je relativně nejjednodušší.

Periodogram představuje souhrn všech hodnot rozptylů jednotlivých harmonik periodické složky vyjádřených pomocí amplitud A_j, respektive s využitím vztahu $A_j^2 = a_j^2 + b_j^2$ také pomocí odhadů regresních koeficientů. Podle velikosti jednotlivých rozptylů (pro jednotlivá $j = 1,2,\dots,H$) a jejich příspěvků k vysvětlení celkového rozptylu rozhodneme, kterou frekvenci (a tedy periodu neboli „délku vlny") vybereme. Schéma periodogramu je uvedeno v následující tabulce:

j	**Frekvence $\omega_j = 2\pi j/n$**	$I(\omega_j) = 1/2(a_j^2 + b_j^2)$
1	ω_1	$1/2(a_1^2 + b_1^2)$
2	ω_2	$1/2(a_2^2 + b_2^2)$
...	...	...
H	ω_H	$1/2(a_H^2 + b_H^2)$
$\sum$	---	$var(y_j)$

Periodogram se často zobrazuje také graficky v podobě sloupcového grafu.
Pokud některá hodnota $I(\omega_j)$ z periodogramu, eventuálně několik takových hodnot, značně převyšuje zbývající hodnoty, je intuitivně zřejmé, že odpovídající frekvence identifikují významné harmoniky periodické složky dané časové řady. Exaktní metodou pro rozhodnutí o existenci statisticky významné periodické složky v časové řadě poskytuje *Fisherův test*. V něm se testuje nulová hypotéza, že y_t, $t = 1,2,\dots n$, je časová řada nezávislých náhodných veličin, majících normální rozdělení se střední hodnotou 0, proti alternativní hypotéze, že existuje periodická složka ve tvaru (10.2).

V testu se postupuje tak, že se nejprve seřadí sestupně hodnoty $I(\omega_j)$, tak, že největší z nich označíme I_1, nejmenší I_H , pak se tyto hodnoty „normují", tj. položí se:

$$Y_j = \frac{I_j}{\sum_{i=1}^{H} I_i}. \tag{10.11}$$

Testová statistika W má tvar:

$$W = \max_{j=1,\dots,H} Y_j . \tag{10.12}$$

Nulová hypotéza se zamítá, když hodnota testové statistiky $\boldsymbol{W}$ překročí kritickou hodnotu $g_f(\alpha)$, která je tabelována pro různé hodnoty hladiny významnosti α.

10.3 MODEL KONSTANTNÍ SEZÓNNOSTI SE SCHODOVITÝM TRENDEM

Při popisu trendové složky i periodické složky v předchozí subkapitole jsme používali posloupnost časové proměnné $t = 1,2,\dots n$, nyní budeme toto označení používat pro označení časových intervalů (např. roků), které se člení na dalších r dílčích časových období, které nazýváme *sezóny* (např. měsíce nebo čtvrtletí) a označujeme $j = 1,2,\dots,r$ (např. v případě, že sezóny jsou měsíce je $r = 12$, v případě že sezóny představují kvartály, platí $r = 4$). Model (10.1) lze s použitím uvedené symboliky zapsat ve tvaru:

$$y_{tj} = T_{tj} + P_{tj} + \varepsilon_{tj},\ t = 1,2,\dots,n,\ j = 1,2,\dots,r. \tag{10.13}$$

U modelu *konstantní sezónnosti* se vychází z předpokladu, že:

$$P_{tj} = \gamma_j \text{ pro sezónu } j \text{ v letech } t = 1,2,\dots,n, \tag{10.14}$$

kde γ_j jsou neznámé sezónní parametry, o nichž dále předpokládáme, že splňují rovnost:

$$\sum_{j=1}^{r} \gamma_j = 0. \tag{10.15}$$

Předpoklady (10.14) a (10.15) vycházejí z představy, že v důsledku pravidelného (ročního) koloběhu sezónních vlivů se v *j*-té sezóně opakují sezónní výkyvy γ_j, které se mezi léty neliší - podmínka (10.14). Dále se tyto vlivy během roku (*r* sezón) vykompenzují, takže jejich roční součet je nulový - podmínka (10.15).

Nejprve budeme předpokládat, že trendová složka T_{tj} nabývá ve všech sezónách hodnotu roku t hodnotu α_t, takže posloupnost těchto hodnot v letech $t = 1,2,...,n$ představuje *schodovitý trend*. Model (10.13) pak bude mít tvar:

$$y_{tj} = \alpha_t + \gamma_j + \varepsilon_{tj},\ t = 1,2,...,n,\ j = 1,2,...,r. \tag{10.16}$$

Odhady a_t, c_j $n + r$ parametrů tohoto modelu získáme metodou nejmenších čtverců:

$$a_t = \frac{1}{r}\sum_{j=1}^{r} y_{tj} = \bar{y}_t,\quad c_j = \frac{1}{n}\sum_{t=1}^{n} y_{tj} - \frac{1}{rn}\sum_{t=1}^{n}\sum_{j=1}^{r} y_{tj}. \tag{10.17}$$

Všimněte si v prvním vzorci, že odhadem výšky schodu v roce t je průměr hodnot v roce t. Z druhého vzorce pak vyplývá, že hodnota sezónního vlivu c_j, tzv. *j*-tého sezónního koeficientu, je představována průměrnou hodnotou vypočítanou z *j*-tých sezón ve všech letech po odečtení celkového průměru ze všech hodnot v celé časové řadě. Například sezónní koeficient c_1 se vypočítá jako průměr ze všech lednových hodnot v časové řadě měsíčních údajů po odečtení celkového průměru ze všech hodnot v celé časové řadě. V tomto příkladu je měsíc leden uvažován jako první sezóna z 12 měsíčních sezón.
Konkrétní použití modelu si ukážeme na řešeném příkladu 10.1 v závěru této kapitoly.

10.4 MODEL KONSTANTNÍ SEZÓNNOSTI S LINEÁRNÍM TRENDEM

Při popisu trendové složky v předchozím odstavci jsme používali posloupnost časové proměnné $t = 1,2,...n$, o trendové funkci jsme předpokládali, že je konstantní během všech sezón daného roku t, tj. $T_{tj} = \alpha_t$ pro $j = 1,2,...,r$. Přitom hodnota α_t mohla být v každém roce jiná a tvořila výšku „schodu“ v roce t. Model časové řady bude opět aditivní, tedy

$$y_{tj} = T_t + \gamma_j + \varepsilon_{tj},\ t = 1,2,...,n,\ j = 1,2,...,r, \tag{10.18}$$

kde stejně jako v modelu (10.13) jsou γ_j neznámé sezónní parametry, o nichž dále předpokládáme, že splňují podmínku $\sum_{j=1}^{r} \gamma_j = 0$.

Nyní budeme předpokládat, že trendová složka T_{tj} má lineární tvar, potom model (10.18) bude mít tvar:

$$y_{tj} = \alpha + \beta(t - \bar{t}) + \gamma_j + \varepsilon_{tj},\ t = 1,2,...,n,\ j = 1,2,...,r. \tag{10.19}$$

Odhady a, b, c_j z $(r + 2)$ parametrů tohoto modelu získáme metodou nejmenších čtverců, řešení má komplikovaný tvar, který zde neuvádíme, zájemce odkazujeme na Segera (1998).

10.5 MODEL PROPORCIONÁLNÍ SEZÓNNOSTI

Nyní budeme používat t = 1,2,...,n, k označení časových intervalů (např. roků), které se člení na dalších r dílčích časových období, které nazýváme *sezóny* (např. měsíce nebo čtvrtletí) a označujeme j = 1,2,...,r (např. v případě, že sezóny jsou měsíce je r = 12, v případě že sezóny představují kvartály, platí r = 4). Regresní model lze s použitím uvedené symboliky zapsat ve tvaru:

$$y_{tj} = T_{tj} + P_{tj} + \varepsilon_{tj},\ t = 1,2,...,n,\ j = 1,2,...,r. \qquad (10.20)$$

U modelu *proporcionální sezónnosti* se vychází z předpokladu, že periodická složka je proporcionální (tj. přímo úměrná) velikosti trendové složky:

$$P_{tj} = C_j T_{tj} \text{ pro sezónu } j \text{ v letech } t = 1,2,...,n, \qquad (10.21)$$

tedy po dosazení (10.21) do (10.20) obdržíte

$$y_{tj} = (1 + C_j)T_{tj} + \varepsilon_{tj}. \qquad (10.22)$$

Aplikací MNČ obdržíme c_j odhad koeficientů C_j takto

$$1 + c_j = \frac{\sum_{i=1}^{n} y_{ij}\bar{y}_i}{\sum_{i=1}^{n} \bar{y}_i^2},\ j = 1,2,...,r. \qquad (10.23)$$

Dosazením do (10.22) obdržíte konečnou podobu modelu proporcionální sezónnosti

$$y_{tj} = \frac{\sum_{i=1}^{n} y_{ij}\bar{y}_i}{\sum_{i=1}^{n} \bar{y}_i^2} T_{tj} + \varepsilon_{tj},\ t = 1,2,...,n,\ j = 1,2,...,r. \qquad (10.24)$$

Přitom $\bar{y}_i = \frac{1}{r}\sum_{j=1}^{r} y_{ij}$ je aritmetický průměr y_{ij} přes j. V konkrétním případě můžeme uvažovat, že trendová složka má lineární tvar, tedy například

$$T_{tj} = \alpha + \beta(t - \bar{t}). \qquad (10.25)$$

Vzorec (10.24) lze snadno realizovat v Excelu.

10.6 ANALÝZA NÁHODNÉ SLOŽKY

Náhodnou složku ε_t lze v modelu (10.20) vyjádřit v tvaru:

$$\varepsilon_t = y_t - Y_t,\quad t = 1,2,...n, \qquad (10.26)$$

kde $Y_t = T_t + P_t$. Jedná se zde o vyjádření blíže nespecifikovaných náhodných vlivů. Zdrojem této složky jsou obvykle nepodchycené drobné vzájemně nezávislé náhodné vlivy. Chceme-li zajistit spolehlivé předpovědi na základě modelu časové řady, potom je třeba mít zajištěny některé předpoklady o náhodné složce. Konkrétně je výhodné, když jsou splněny předpoklady klasického lineárního regresního modelu, které jsme uvedli v kapitole 3.5. Byly to předpoklady 1. až 3., které pro přehlednost zopakujeme, avšak při současném označení, kdy nezávisle proměnná x je nyní čas t. Jedná se tedy o tyto předpoklady:

1. Hodnoty vysvětlující proměnné t se volí předem, obvykle t = 1,2,...,n.
2. Náhodné složky ε_t v modelu (10.20) mají *normální rozdělení* pravděpodobnosti se střední hodnotou 0 a (neznámým) rozptylem σ^2.
   Konstantnost rozptylu nazýváme *homoskedasticita*.

3. Náhodné složky jsou *nekorelované*, tj.
   $Cov(\varepsilon_t, \varepsilon_{t'}) = 0$ pro každé $t \neq t'$, $t, t' = 1,2,...,n$.
   (*Cov* značí kovarianci)

Jak již bylo řečeno v kapitole 3.5, v praxi jsou podmínky klasického modelu často splněny. Nejsme-li si však jejich platností jisti, můžeme provést testy hypotéz jak o normalitě rozdělení náhodné složky (např. Chi-kvadrát test dobré shody), tak i testy homoskedasticity (Bartleyův test). Při ověřování těchto předpokladů zjišťujeme, zda jsou všechny systematické složky z časové řady eliminovány. Jakákoliv nenáhodnost u reziduí naznačuje nevhodnost zvoleného modelu časové řady.

Jednoduchým nástrojem, kterým lze ověřit náhodnost reziduí, je *znaménkový test*. Při tomto testu vyčíslíme počet případů, kdy rozdíl sousedních reziduí $e_t - e_{t-1}$ je kladný, jejich počet označíme *S*. Přitom je:

$$e_t = y_t - Y_t, \qquad (10.27)$$

kde $Y_t = T_t + P_t$ je odhad teoretické hodnoty časové řady, T_t je odhad trendu (s regresními koeficienty získanými např. metodou nejmenších čtverců), P_t je odhad periodické složky, např. (10.23), kde parametry α_j, β_j jsou rovněž odhadnuty metodou nejmenších čtverců. Náhodné složky ε_t, které jsou dány (10.26), jsou tedy náhodné veličiny, zatímco rezidua e_t, (10.27), jsou realizacemi - odhady těchto náhodných veličin. Je-li posloupnost reziduí e_t náhodně uspořádána, potom pro střední hodnotu *S* platí: $E(S) = \frac{n-1}{2}$. Testujeme proto nulovou hypotézu: $H_0 : E(S) = \frac{n-1}{2}$, proti alternativní hypotéze $H_1 : E(S) \neq \frac{n-1}{2}$. Použijeme testové kritérium:

$$U = \frac{\sqrt{12}\left(S - \frac{1}{2}(n-1)\right)}{\sqrt{n+1}}, \qquad (10.28)$$

které má již pro $n \geq 13$ přibližně normované normální rozdělení. Pro stanovení kritických hodnot tedy použijeme *kvantily* normovaného normálního rozdělení $u_{1-\alpha/2}$.

Vlastnost časových řad, která často způsobuje porušení předpokladů 1. až 3. je *autoregrese* náhodných složek, viz též kapitola 6.5, která znamená, že mezi náhodnými složkami platí následující vztah:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \qquad (10.29)$$

kde $0 < \rho < 1$ je *autokorelační koeficient* a u_t splňuje předpoklady 1. až 3. Nulovou hypotézu: $H_0 : \rho = 0$ (což je totéž, jako $\varepsilon_t = u_t$) testujeme proti alternativní hypotéze $H_1 : \rho \neq 0$ pomocí testového kritéria:

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \qquad (10.30)$$

Funkce *D,* nazývaná *Durbin-Watsonova statistika*, bývá tabelována pro různé hladiny významnosti α, viz např. Gujarati (2003). Test založený na této statistice nazýváme *Durbin-Watsonův test autokorelace*.

V této kapitole jste se zabývali časovými řadami, jejichž hodnoty se periodicky opakují, tzv. sezónními časovými řadami. Nejprve jste si objasnili význam sezónní složky časové řady, poté jste se věnovali metodě harmonické analýzy, která k modelování časové řady využívá známé matematické periodické funkce sinus a kosinus. Poté jste se naučili aplikovat jednoduché metody konstantní sezónnosti se schodovitým a lineárním trendem a rovněž metodu proporcionální sezónnosti. Zmíněné metody si ozřejmíme na konkrétních příkladech řešených s využitím Excelu.

ŘEŠENÝ PŘÍKLAD 10.1

Časová řada y_t udává počet měsíčně ubytovaných v Penzionu Madonna za období let 2006 a 2007 - celkem 24 hodnot. Harmonickou analýzou modelujte sezónní složku této časové řady. Zvolte $j = 2$.

t	1	2	3	4	5	6	7	8	9	10	11	12
y_t	332	223	267	319	455	507	492	500	350	253	178	401
t	13	14	15	16	17	18	19	20	21	22	23	24
y_t	301	213	247	433	399	466	505	455	314	222	184	335

Řešení:

Pomocí vztahů (10.1) - (10.7) stanovíme regresní koeficienty m, a_2, b_2. Všechny potřebné výpočty jsou uvedeny v následující tabulce. V této tabulce jsou také uvedeny teoretické hodnoty Y_t a rezidua e_t.

t	y_t	$\sin\frac{2\pi jt}{n}$	$\cos\frac{2\pi jt}{n}$	$y_t \sin\frac{2\pi jt}{n}$	$y_t \cos\frac{2\pi jt}{n}$	Y_t	e_t
1	332	0,50	0,87	166,00	287,52	236,32	95,68
2	223	0,87	0,50	193,12	111,50	269,28	-46,28
3	267	1,00	0,00	267,00	0,00	323,33	-56,33
4	319	0,87	-0,50	276,26	-159,50	383,97	-64,97
5	455	0,50	-0,87	227,50	-394,04	434,97	20,03
6	507	0,00	-1,00	0,00	-507,00	462,65	44,35
7	492	-0,50	-0,87	-246,00	-426,08	459,60	32,40
8	500	-0,87	-0,50	-433,01	-250,00	426,64	73,36
9	350	-1,00	0,00	-350,00	0,00	372,59	-22,59
10	253	-0,87	0,50	-219,10	126,50	311,94	-58,94
11	178	-0,50	0,87	-89,00	154,15	260,95	-82,95
12	401	0,00	1,00	0,00	401,00	233,27	167,73
13	301	0,50	0,87	150,50	260,67	236,32	64,68
14	213	0,87	0,50	184,46	106,50	269,28	-56,28
15	247	1,00	0,00	247,00	0,00	323,33	-76,33
16	433	0,87	-0,50	374,99	-216,50	383,97	49,03
17	399	0,50	-0,87	199,50	-345,54	434,97	-35,97
18	466	0,00	-1,00	0,00	-466,00	462,65	3,35
19	505	-0,50	-0,87	-252,50	-437,34	459,60	45,40
20	455	-0,87	-0,50	-394,04	-227,50	426,64	28,36
21	314	-1,00	0,00	-314,00	0,00	372,59	-58,59
22	222	-0,87	0,50	-192,26	111,00	311,94	-89,94
23	184	-0,50	0,87	-92,00	159,35	260,95	-76,95
24	335	0,00	1,00	0,00	335,00	233,27	101,73
Součty	8351	0,00	0,00	-295,57	-1376,32	8351	0,00

Potom platí:

$$m = \frac{1}{n}\sum_{t=1}^{n} y_t = \frac{1}{24}\sum_{t=1}^{24} y_t = \frac{8351}{24} = 347{,}95\,.$$

$$a_2 = \frac{2}{n}\sum_{t=1}^{n} y_t \sin\omega_j t = \frac{2}{24}\sum_{t=1}^{24} y_t \sin\frac{2\pi jt}{n} = \frac{1}{12}\sum_{t=1}^{24} y_t \sin\frac{2\pi 2t}{24} = \frac{-295{,}57}{12} = -24{,}63\,.$$

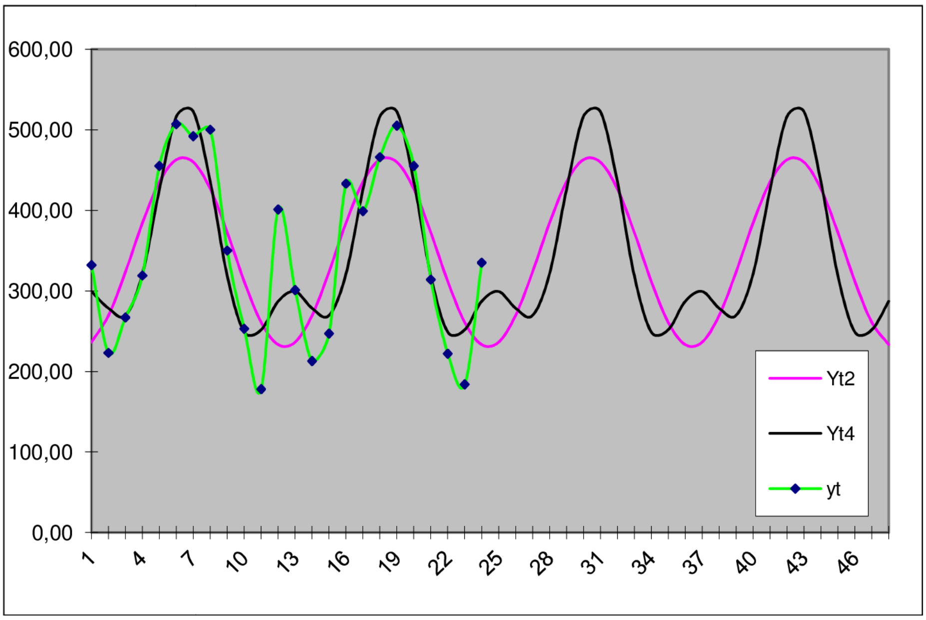
$$b_2 = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\omega_j t = \frac{2}{24}\sum_{t=1}^{24} y_t \cos\frac{2\pi jt}{n} = \frac{1}{12}\sum_{t=1}^{24} y_t \cos\frac{2\pi 2t}{24} = \frac{-1376{,}32}{12} = -114{,}69\,.$$

Teoretické hodnoty obdržíme dosazením m, a_2, b_2 do modelu (10.2), např.:

$$Y_1 = m + a_2 \sin\varpi_j t + b_2 \cos\varpi_j t = 347{,}95 - 24{,}63 \sin\frac{2\pi 2t}{24} - 114{,}69 \cos\frac{2\pi 2t}{24} =$$

$$= 347{,}95 - 24{,}63 \cdot 0{,}5 - 114{,}69 \cdot 0{,}87 = 236{,}32\,.$$

Na Obr. 10.1. je znázorněna harmonická analýza pro $j = 2$, $j = 4$. Regresní parametry m, a_4, b_4 vypočítáte analogicky.



Obr. 10.1. *Harmonická analýza*

ŘEŠENÝ PŘÍKLAD 10.1

Data v tabulce představují objem přepravy po vodních tocích ČR v jednotlivých čtvrtletích pěti po sobě jdoucích let.

t/j	Čtvrtletí					
Roky	**1**	**2**	**3**	**4**	**Součet**	**Průměr**
1	120	138	132	114	504	126,00
2	118	138	150	119	525	131,25
3	149	161	155	145	610	152,50
4	150	173	181	159	663	165,75
5	178	195	198	183	754	188,50
Součet	715	805	816	720	3056	
Průměr	143	161	163,2	144		152,80

a. Nalezněte pro tuto časovou řadu model konstantní sezónnosti se schodovitým trendem.
b. Na hladině významnosti $\alpha = 0{,}05$ ověřte náhodnost reziduí.

Řešení:

a. Úkolem je nalézt odhady parametrů α_t, γ_j modelu

$$y_{tj} = \alpha_t + \gamma_j + \varepsilon_{tj},\ t = 1,2,...,n,\ j = 1,2,...,r,$$

kde α_t je trendová složka
γ_j je sezónní složka.

Odhady a_t, c_j $n + r$ parametrů tohoto modelu vypočítáme ze vztahů (10.28):

$$a_t = \frac{1}{r}\sum_{j=1}^{r} y_{tj} = \bar{y}_t,\quad c_j = \frac{1}{n}\sum_{t=1}^{n} y_{tj} - \frac{1}{rn}\sum_{t=1}^{n}\sum_{j=1}^{r} y_{tj}.$$

Všechny potřebné součty a průměry jsou uvedeny v tabulce, jejich dosazením do daných vztahů obdržíte :

trendová složka: $a_1 = 126$, $a_2 = 131{,}25$, $a_3 = 152{,}5$, $a_4 = 165{,}75$, $a_5 = 188{,}5$,
sezónní složka: $c_1 = 143 - 152{,}8 = -9{,}8$, $c_2 = 161 - 152{,}8 = 8{,}2$,
$c_3 = 163{,}2 - 152{,}8 = 10{,}4$, $c_4 = 144 - 152{,}8 = -8{,}8$.

Výsledky ukazují, že působení sezónních vlivů klesl v prvním čtvrtletí objem přepravy o 9,8 tun a ve čtvrtém čtvrtletí o 8,8 tuny. Tento pokles je vykompenzován růstem přepravy ve zbylých dvou čtvrtletích o 8,2 a 10,4 tun, tj. ve čtvrtletích pro říční přepravu klimaticky příznivějších. Z vývoje ročních průměrů a_t je zřejmé, že se průměrný roční objem přepravy neustále zvyšoval.

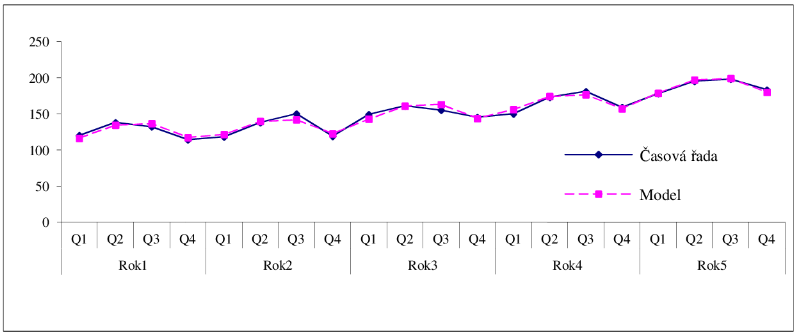
b. Nejdříve vypočítáte odhady teoretických hodnot $\hat{Y}$ dané časové řady tak, že odhadnete trendovou i sezónní složku. Např.:

$$\hat{Y}_{1,1} = a_1 + c_1 = 126 + (-9{,}8) = 116{,}2$$

$$\hat{Y}_{1,2} = a_1 + c_2 = 126 + 8{,}2 = 134{,}2$$

Všechny hodnoty $\hat{Y}_{t,j}$ jsou uvedeny v následující tabulce.

t/j	1	2	3	4
1	116,20	134,2	136,4	117,2
2	121,45	139,5	141,7	122,5
3	142,70	160,7	162,9	143,7
4	155,95	174,0	176,2	157,0
5	178,70	196,7	198,9	179,7



Obr. 10.2. *Model konstantní sezónnosti se schodovitým trendem.*

Dále podle (10.28) vypočítáme hodnoty reziduí. Např.:

$$e_{1,1} = y_{1,1} - \hat{Y}_{1,1} = 120 - 116{,}2 = 3{,}8,\; e_{1,2} = y_{1,2} - \hat{Y}_{1,2} = 138 - 134{,}2 = 3{,}8\,.$$

Hodnoty všech reziduí jsou uvedeny v následující tabulce:

t/j	1	2	3	4
1	3,80	3,80	–4,40	–3,20
2	–3,45	–1,45	8,35	–3,45
3	6,30	0,30	–7,90	1,30
4	–5,95	–0,95	4,85	2,05
5	–0,70	–1,70	–0,90	3,30

K testu náhodnosti reziduí použijeme znaménkový test. Je proto třeba určit počet případů S, kdy je rozdíl sousedních reziduí $e_t - e_{t-1}$ kladný. Např.:

$e_{1,2} - e_{1,1} = 3{,}8 - 3{,}8 = 0,$
$e_{1,3} - e_{1,2} = -4{,}4 - 3{,}8 = -8{,}2.$

V následující tabulce jsou případy, kdy $e_t - e_{t-1} > 0$, označeny „+“, ostatní „–“.

t/j	1	2	3	4
1	.	–	–	+
2	–	+	+	–
3	+	–	–	+
4	–	+	+	–
5	–	–	+	+

Z tabulky vidíme, že $S = 9$. Hodnotu testového kritéria vypočítáme podle (10.28):

$$U = \frac{\sqrt{12}\left(S - \frac{1}{2}(n-1)\right)}{\sqrt{n+1}} = \frac{\sqrt{12}\left(9 - \frac{1}{2}(20-1)\right)}{\sqrt{20+1}} = -0{,}378\,.$$

V tabulce normovaného normálního rozdělení nalezneme $u_{1-\alpha/2}$, tj.: $u_{0{,}975} = 1{,}96$.

Protože hodnota testového kritéria –0,378 leží v oboru přijetí A = (–1,96;1,96), lze na zvolené hladině významnosti přijmout nulovou hypotézu, tj. hypotézu o náhodném uspořádání reziduí.

10.7 SAMOSTATNÉ ÚKOLY

10.1 V následující tabulce jsou uvedeny měsíční tržby jedné obchodní organizace za posledních 60 měsíců od ledna 2007 až do prosince 2011.

a. Nalezněte model konstantní sezónnosti se schodovým trendem.

b. Pro rok 2012 uvažujte s růstem 5% (tj. výška schodu). Prognózujte tržby na rok 2012.

1	2	3	4	5	6	7	8	9	10	11	12
6489	5971	6272	6944	7217	7448	7259	7602	7651	8064	7952	8498
13	14	15	16	17	18	19	20	21	22	23	24
6930	6391	6979	7315	7798	7861	7994	7798	8022	8155	8694	8764
25	26	27	28	29	30	31	32	33	34	35	36
7560	7182	7077	7847	8603	8659	8827	8855	8337	8379	8834	9709
37	38	39	40	41	42	43	44	45	46	47	48
7833	7406	7791	8190	8869	8988	8736	9254	9240	9380	9422	9954
49	50	51	52	53	54	55	56	57	58	59	60
8442	7987	8673	8925	9534	9534	9331	9877	9695	9730	10192	10661

10.2 Použijte data z řešeného příkladu 10.1. Nalezněte pro tuto časovou řadu model konstantní sezónnosti s lineárním trendem.

10.8 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

10.1 **a)** a1 = 7280,6; a2 = 7630,6; a3 = 8322,4; a4 = 8755,3; a5 = 9381,7; a6 = 9850,8
c1 = –823,3; c2 = –1286,7; c3 = –915,7; c4 = –429,9; c5 = 130,1; c6 = 223,9;
c7 = 155,3; c8 = 403,1; c9 = 314,9; c10 = 467,5; c11 = 744,7; c12 = 1243,1

b)

leden 2012	9027,51
únor 2012	8564,11
březen 2012	8935,11
duben 2012	9420,91
květen 2012	9980,91
červen 2012	10074,7
červenec 2012	10006,1
srpen 2012	10253,9
září 2012	10165,7
říjen 2012	10318,3
listopad 2012	10595,5
prosinec 2012	11093,9

10.2 $Y_t = 6782{,}2 + 49{,}536.t + c_j$

c1 = –569,8; c2 = –1082,7; c3 = –761,3; c4 = –325; c5 = 185,4; c6 = 229,7;
c7 = 111,6; c8 = 309,8; c9 = 172,1; c10 = 275,2; c11 = 502,8; c12 = 951,7

leden 2012	9234,1
únor 2012	8770,7
březen 2012	9141,7
duben 2012	9627,5
květen 2012	10187,5
červen 2012	10281,3
červenec 2012	10212,7
srpen 2012	10460,5
září 2012	10372,3
říjen 2012	10524,9
listopad 2012	10802,1
prosinec 2012	11300,5

11 STOCHASTICKÉ PROCESY

RYCHLÝ NÁHLED KAPITOLY

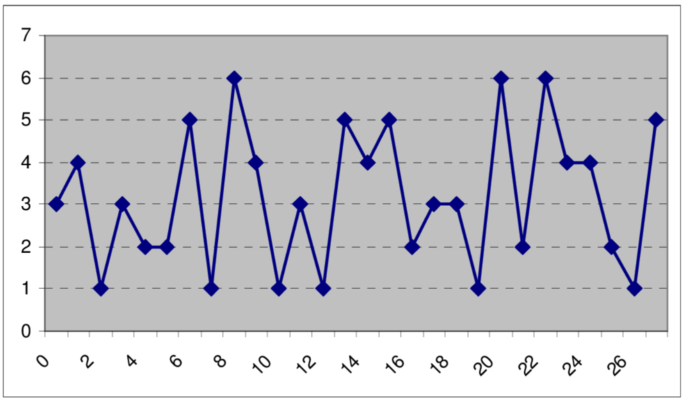
V této kapitole se nejprve zaměříte na obecnější pojetí časové řady s pomocí pojmu stochastický (náhodný) proces. V klasickém pojetí je časová řada posloupnost číselných veličin v čase, v rozšířeném pojetí je to posloupnost náhodných veličin v čase. Speciální důležitou třídu tvoří tzv. stacionární procesy, zbývající procesy jsou nestacionární. Výjimečné místo mezi stacionárními procesy zaujímá bílý šum, mezi nestacionárními je to náhodná procházka. Těmi se budete v této kapitole podrobněji zabývat. S rozšířením pojmu časové řady vzniká i rozšíření klasického deterministického trendu na tzv. stochastický trend. V závěru kapitoly se budete zabývat problémem, jak rozhodnout, zda daná časová řada je nebo není stacionární. K řešení ilustračních příkladů budete používat nejen Excel, nýbrž také speciální statistický SW: SPSS (Statistical Package for Social Sciences).

11.1 STOCHASTICKÝ (NÁHODNÝ) PROCES

V kapitole 8 jste se o časové řadě dozvěděli, že časovou řadou rozumíme věcně a prostorově srovnatelná pozorování uspořádaná v čase směrem od minulosti přes přítomnost k budoucnosti. Tato definice časové řady jako posloupnost čísel $\{y_t\}$, kde t představuje časový index, nám doposud vystačila, avšak pro další rozvoj metod analýzy časových řad se ukázalo vhodné definici časové řady rozšířit. V tomto rozšíření je každá hodnota časové řady pojímána jako realizace nějaké náhodné veličiny. Jinak řečeno, v konkrétním čase t je hodnotou časové řady náhodná veličina Y_t, přičemž se realizovala konkrétní hodnota časové řady y_t. Kdyby však bylo možné čas vrátit zpět do t, pak by se mohla realizovat jiná hodnota, řekněme z_t. V čase t by bylo možné realizovat různé hodnoty s různou pravděpodobností, neboli s jistým rozdělením pravděpodobnosti. Náhodná veličina Y_t přestavuje jednak množinu hodnot, které se mohou nabývat, jednak rozdělení pravděpodobnosti, s níž se hodnoty mohou nabývat. Stochastický proces je pak posloupnost náhodných veličin $\{ Y_t \}$, kde t představuje časový index. Ačkoliv časový index může být intervalem reálných čísel (tzv. *spojitý čas*), v ekonomických aplikacích vystačíme zpravidla s tzv. *diskrétním časem*, konkrétně t =1,2,3,…

Stejně tak lze uvažovat, že dvě nebo více sousedních hodnot časové řady v čase spolu buď nesouvisejí – nejsou vzájemně korelovány, nebo naopak souvisejí – jsou vzájemně korelovány. Budeme-li nadále hovořit o časové řadě, budeme tím mít na mysli stochastický proces ve výše zmíněném smyslu.

Příklad 1. Uvažujte jednoduchý stochastický proces „hod kostkou v čase“: pro t =1,2,3,… je Y_t náhodná veličina nabývající hodnoty 1,2,3,4,5,6 se stejnou pravděpodobností $p = {}^1/_6$. Výsledky hodů kostkou v jednotlivých časech jsou vzájemně nezávislé – nekorelované. Grafické zobrazení takové „časové řady“, tj. konkrétních realizací takových náhodných veličin, je na Obr. 11.1.



Obr. 11.1 *Stochastický proces „Hod kostkou v čase"*

V každém časovém intervalu se realizuje diskrétní náhodná veličina Y_t, která má střední hodnotu $\mu = {}^{21}/_{6} = 3{,}5$ a rozptyl $\sigma^2 = {}^{18,5}/_{6} = 2{,}9167$.

11.2 STACIONÁRNÍ A NESTACIONÁRNÍ PROCES

Typem stochastického procesu, který si získal velkou pozornost analytiků časových řad je tzv. stacionární stochastický proces. Stručně a poněkud nepřesně řečeno, stochastický proces (časová řada) se nazývá *stacionární*, jestliže střední hodnota a rozptyl náhodných veličin jsou konstantní v čase a také hodnota kovariance mezi dvěma veličinami vzdálenými v čase závisí pouze na jejich vzdálenosti v čase a nikoliv na konkrétním časovém okamžiku, v němž se kovariance zjišťuje. V odborné literatuře se takovýto typ stacionarity nazývá též *slabá stacionarita*, zde vystačíme s jednoduchým *stacionarita*.

K přesnější definici stacionarity použijeme známou symboliku ze statistiky. Uvažujte stochastický proces – časovou řadu $\{Y_t\}$, nechť je střední hodnota konstantní, tj.

$$E(Y_t) = \mu \text{ pro všechna } t, \tag{11.1}$$

rozptyl je konstantní, tj.

$$Var(Y_t) = E(Y_t - \mu)^2 = \sigma^2 \text{ pro všechna } t, \tag{11.2}$$

kovariance nezávisí na čase t, tj.

$$Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)] = \gamma_k \quad \text{pro všechna } t, \tag{11.3}$$

kde *Var* značí rozptyl (z angl. „Variance"), *Cov* označuje kovarianci („Covariance", *E* je operátor střední hodnoty („*E*xpected value"). Potom se stochastický proces $\{Y_t\}$ nazývá *stacionární*.

V (11.3) se jedná o kovarianci mezi dvěma hodnotami Y, které jsou od sebe vzdáleny o k časových jednotek – tzv. *posuv* (angl. *lag*). Jestliže posuv $k = 0$, obdržíme γ_0, která se v tom případě rovná rozptylu, tj. $\gamma_0 = \sigma^2$. Jestliže je posuv $k = 1$, obdržíme kovarianci γ_1, která je v tom případě kovariancí sousedních hodnot. Důležitou roli hraje tzv. *autokorelační funkce* (ACF) stochastického procesu, která je definována jako normovaná kovarianční funkce, tedy

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad \text{pro } k = 1,2,\ldots \tag{11.4}$$

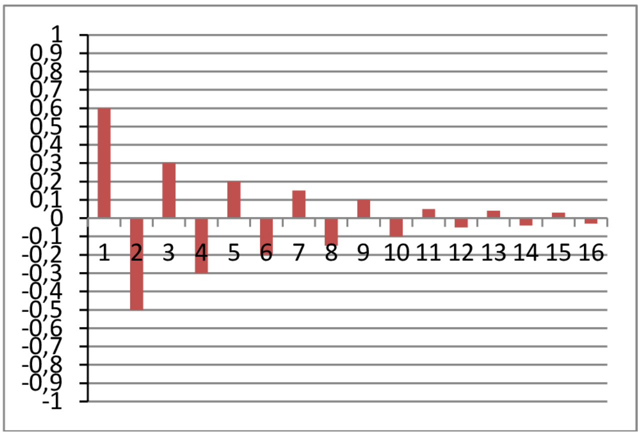
V případě stacionárního procesu má ACF následující vlastnosti:

(a) $\rho_0 = 1$,

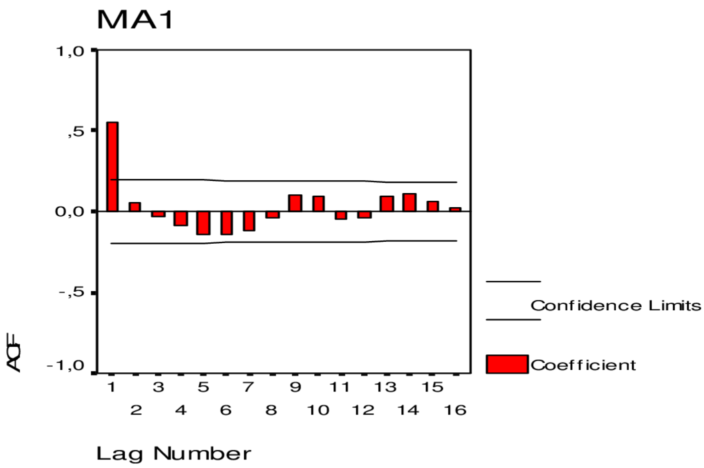
(b) $-1 \le \rho_k \le 1$ pro $k = 1,2,\ldots$

(c) $\rho_k = \rho_{-k}$ pro $k = 1,2,\ldots$, tj. autokorelační funkce je symetrická kolem $k = 0$.

Grafickým znázorněním autokorelační funkce je *korelogram*, jehož příklady jsou na Obr. 11.2. a 11.3. Vzhledem k uvedeným vlastnostem stačí, aby korelogram zobrazoval hodnoty korelace pro posuvy (angl. „lag“) $k > 0$.



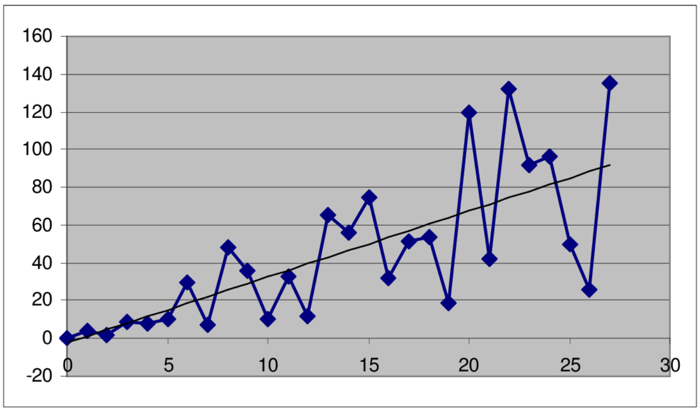
Obr. 31.2 *Korelogram – ACF*



Obr. 11.3 *Korelogram – ACF (výstup z programu SPSS)*

Proč jsou stacionární stochastické procesy důležité? Je to proto, že když je časová řada nestacionární, můžeme její chování analyzovat pouze v časovém intervalu, kde máme k dispozici data. Pro analýzu mimo tento časový interval, např. pro prognózování, nemá daná časová řada praktický význam. Jak poznáme, že časová řada je stacionární? K tomuto problému se vrátíme na konci této kapitoly, kdy se seznámíte se statistickým testem, kterým lze stacionaritu zjišťovat (tj. testovat). Na tomto místě uvedeme příklad **nestacionární** časové řady.

Příklad 2. Uvažujte stochastický proces „hod kostkou závislý na čase“: pro t =1,2,3,… je Y_t náhodná veličina nabývající hodnoty $1{\cdot}t$, $2{\cdot}t$, $3{\cdot}t$, $4{\cdot}t$, $5{\cdot}t$, $6{\cdot}t$ se stejnou pravděpodobností $p = {}^1/_6$, přitom t je časový okamžik. Výsledky hodů kostkou v jednotlivých časech jsou vzájemně nezávislé – nekorelované. Grafické zobrazení takové „časové řady“, tj. konkrétních realizací takových náhodných veličin, je na Obr. 11.4.

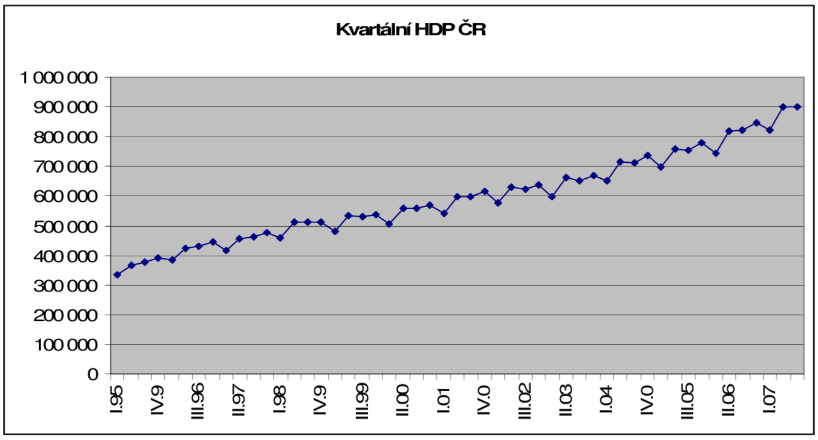


Obr. 11.4 *Stochastický proces „Hod kostkou závislý na čase"*

Příklad 3. Uvažujte stochastický proces „Čtvrtletní HDP České republiky". Data v následující tabulce byla převzata z českého statistického úřadu, viz www.czso.cz . Je zřejmé, že se jedná o stochastický proces – časovou řadu v širším pojetí, kdy konečná hodnota za každé čtvrtletí je realizací náhodné veličiny (s neznámým rozdělením pravděpodobnosti), která může nabývat různých hodnot z číselného intervalu v závislosti na konkrétních ekonomických podmínkách závislých na mnoha faktorech. Grafické znázornění je na obr. 11.5.

Q	I.95	II.95	III.95	IV.95	I.96	II.96	III.96	IV.96	I.97	II.97	III.97	IV.97
HDP_Q	332995	366618	376688	390221	382859	423953	432152	444324	415593	455790	461902	477809
Q	I.98	II.98	III.98	IV.98	I.99	II.99	III.99	IV.99	I.00	II.00	III.00	IV.00
HDP_Q	457925	512225	512408	513925	481895	532968	529465	536469	504479	558691	557780	568219
Q	I.01	II.01	III.01	IV.01	I.02	II.02	III.02	IV.02	I.03	II.03	III.03	IV.03
HDP_Q	540124	598842	599262	613986	576665	630141	621004	636622	598385	660401	650791	667533
Q	I.04	II.04	III.04	IV.04	I.05	II.05	III.05	IV.05	I.06	II.06	III.06	IV.06
HDP_Q	650448	715163	712103	737048	696387	759740	753836	777759	745496	817867	821754	846459
Q	I.07	II.07	III.07									
HDP_Q	820689	900606	900022									

Tab. 11.1 *Čtvrtletní HDP ČR 1995 - 2007*



Obr. 11.5 *Stochastický proces „Čtvrtletní HDP České republiky"*

11.3 BÍLÝ ŠUM A NÁHODNÁ PROCHÁZKA

V této části se budeme věnovat dvěma speciálním typům stochastického procesu. První z nich je nejjednodušší stochastický proces nazývaný *bílý šum*. Setkali jste se s ním již v kapitolách o regresní analýze v souvislosti s klasickým jednoduchým lineárním regresním modelem, kde jsme jej nazývali *náhodná porucha*. O té jsme předpokládali, že splňuje jisté podmínky:

- má nulovou střední hodnotu,
- má konstantní rozptyl σ^2,
- veličiny poruchy jsou vzájemně nekorelované, tj. $Cov(Y_t, Y_{t+k}) = 0$ pro vš. t a $k \neq 0$.

Poslední podmínka vlastně říká, že autokorelační funkce je identicky nulová (pro $k > 0$). Tyto podmínky zároveň zaručují, že posloupnost poruch tvoří speciální stacionární stochastický proces, konkrétně podmínky (11.1) – (11.3), který nazýváme *bílý šum* (angl. white noise). Můžeme tedy říci, že daná data vyhovují jistému klasickému regresnímu modelu, jestliže se od něj odlišují o bílý šum. Ještě jinak, můžeme říci, že jistý regresní model je vhodným modelem pro daná data, jestliže je jeho odchylka od dat bílým šumem, v opačném případě není vhodným modelem. Totéž platí o modelech časových řad, které jsou speciálními jednoduchými regresními modely, kde nezávislou proměnnou je čas. Bílý šum je tedy výjimečným stacionárním stochastickým procesem (časovou řadou v širším pojetí), který hraje významnou roli při analýze a modelování časových řad.

Na druhou stranu podobnou roli u nestacionárních stochastických procesů sehrává tzv. *náhodná procházka* (angl. random walk). Její název je odvozen od představy, která spočívá v tom, že další krok vzniká přičtením náhodně zvolené veličiny (bílého šumu) ke kroku předchozímu. Říká se, že ceny akcií nebo kurzy měn se řídí náhodnou procházkou – jsou nestacionární. Rozlišujeme dva typu náhodné procházky: (1) náhodná procházka bez posuvu a (2) náhodná procházka s posuvem.

Uvažujme nyní bílý šum u_t s nulovou střední hodnotou a rozptylem σ^2. Časová řada $\{Y_t\}$ daná předpisem

$$Y_T = Y_{T-1} + U_T, \qquad (11.5)$$

je podle naší definice náhodná procházka. Podle modelu (11.5) se hodnota časové řady v čase t rovná hodnotě v čase (t-1) plus náhodná chyba. Podle příznivců hypotézy eficientního kapitálového trhu jsou ceny akcií v zásadě náhodné, chovají se podle modelu (11.5), a není proto důvodů pro spekulaci o jejich cenách: kdybychom totiž uměli predikovat zítřejší cenu akcie na základě ceny dnešní, byli bychom všichni milionáři. Rovnici (11.5) můžeme rozepsat takto:

$$Y_1 = Y_0 + u_1,$$
$$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2,$$
$$Y_3 = Y_2 + u_3 = Y_0 + u_1 + u_2 + u_3$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Y_t = Y_{t-1} + u_t = Y_0 + \Sigma u_i.$$

Aplikujete-li operátor střední hodnoty E, obdržíte díky vlastnostem střední hodnoty

$$E(Y_t) = E(Y_0 + \Sigma u_i) = Y_0. \text{ (Víte proč?)} \qquad (11.6)$$

Podobně lze odvodit, že v tomto případě (nyní díky vlastnostem rozptylu Var) platí

$$Var(Y_t) = t\sigma^2. \qquad (11.7)$$

Jak vyplývá z předchozích vztahů, střední hodnota časové řady je rovna Y_0, což je konstanta (obvykle je rovna nule, tj. $Y_0 = 0$), avšak rozptyl se přímo úměrně zvětšuje s rostoucím t, není tedy konstantní, a proto časová řada **není** stacionární. Můžete dále vztah (11.5) zapsat také takto

$$Y_t - Y_{t-1} = \Delta Y_t = u_t, \qquad (11.8)$$

kde ΔY_t značí první diferenci časové řady (viz kap. 8). Vztah (11.8) nám tedy dává důležitou informaci: přestože časová řada – náhodná procházka $\{Y_t\}$ je nestacionární časová řada, její první diference ΔY_t je stacionární, neboť je to vlastně bílý šum u_t.

Modifikujme nyní vztah (11.5) následovně:

$$Y_t = \delta + Y_{t-1} + u_t, \qquad (11.9)$$

kde δ je tzv. *parametr posuvu* neboli *drift*. Podobně jako v případě, kdy bylo $\delta = 0$, aplikujeme operátor střední hodnoty E a obdržíme (díky vlastnostem střední hodnoty)

$$E(Y_t) = Y_0 + t\delta \qquad (11.10)$$

a analogicky lze odvodit, že stejně jako v předchozím případě platí

$$Var(Y_t) = t\sigma^2.$$

V tomto případě, na rozdíl od případu bez driftu (tj. $\delta = 0$), je nejen rozptyl, ale též střední hodnota rostoucí nebo klesající v závislosti na tom, zda je δ kladné nebo záporné číslo. Náhodná procházka s posuvem je proto rovněž nestacionární časová řada. V závěru této subkapitoly uvedeme příklad náhodné procházky bez driftu a s driftem.

ŘEŠENÝ PŘÍKLAD 11.1

Pomocí funkce NÁHČÍSLO() v Excelu simulujte bílý šum u_t pro t = 1,2,...,30 s rozptylem $\sigma^2 = 1$. Přitom funkce NÁHČÍSLO() generuje spojitou náhodnou veličinu se stejnoměrným rozdělením a s hodnotami v intervalu [0,1]. S generovanými hodnotami bílého šumu pak vytvořte náhodnou procházku

a. bez driftu,

b. s driftem $\delta = 2$.

Simulace časových řad zobrazte graficky.

Řešení:

Nejprve si uvědomte, že funkce NÁHČÍSLO() v Excelu simuluje náhodnou veličinu X se stejnoměrným rozdělením pravděpodobnosti a s hodnotami v intervalu [0,1], přitom střední hodnota $E(X) = 0{,}5$ a rozptyl $Var(X) = {}^1/_{12}$. Abyste obdrželi požadovaný bílý šum se střední hodnotou $E(X) = 0$ a rozptylem $Var(X) = 1$, hodnoty generované funkcí NÁHČÍSLO() transformovat na hodnoty bílého šumu, tj. musíte od každé simulované hodnoty nejprve odečíst střední hodnotu 0,5 a poté výsledek násobit $\sqrt{12} = 3{,}464$ (aby byl výsledný rozptyl roven požadované hodnotě 1). Pro spojitou náhodnou veličinu X se stejnoměrným rozdělením pravděpodobnosti, která nabývá hodnoty v intervalu $[a,b]$ totiž platí, viz [Statika A]:

$$E(X) = (b-a)/2 \text{ a } Var(X) = (b-a)^2/12.$$

V následující tabulce z Excelu je ve. sloupci A hodnota časového indexu t, ve sloupci B jsou generovány hodnoty funkce NÁHČÍSLO(), ve sloupci C je hodnota z 2. sloupce transformována na hodnotu bílého šumu u_t pomocí vzorce

=odmocnina(12)*(B3-0,5), atd.

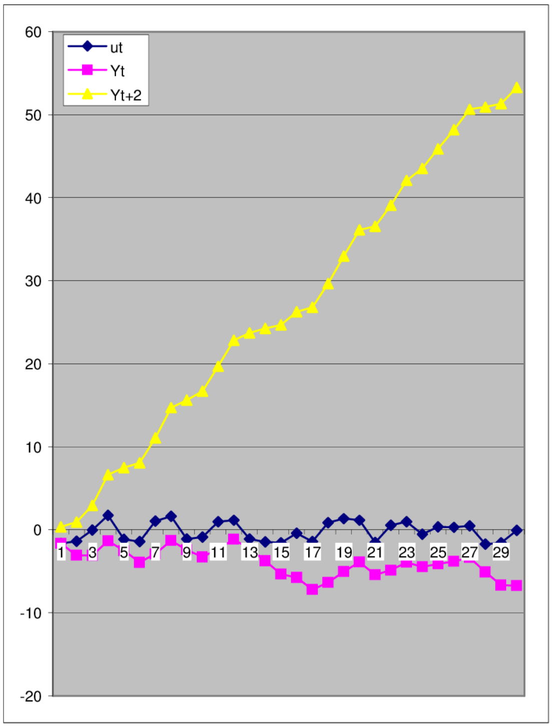
Ve sloupci D se uloží kumulované hodnoty bílého šumu u_t a vypočítá se hodnota náhodné procházky Y_t pomocí vzorce

=D2+C3, atd., přitom Y_0=0 a dále

v je buňce C2 uložena hodnota u_1. V dalších buňkách C4, C5,... jsou následně uloženy postupně kumulované součty hodnot bílého šumu. V buňce D2, D3,... jsou uloženy aktuální hodnoty bílého šumu.

Ve sloupci E se kumulací hodnot bílého šumu u_t a driftu $\delta = 2$ vypočítá hodnota náhodné procházky Y_{t+1} a to tak, že v buňce E3 je uložen vzorec
=2+E2+C3, atd.,
přitom je v buňce E2 uložena hodnota u_1+2.
Průběhy časových řad jsou zobrazeny spojnicovými grafy.

	A	B	C	D	E
1	t	NÁHČÍSLO	u_t	Y_t	Y_t+2
2	1	0,02388619	-1,649307	-1,649307	0,350693
3	2	0,09182492	-1,41396	-3,063267	0,936733
4	3	0,49199476	-0,027731	-3,090998	2,909002
5	4	0,99972311	1,731092	-1,359906	6,640094
6	5	0,16421835	-1,163182	-2,523088	7,476912
7	6	0,08807177	-1,426961	-3,950049	8,049951
8	7	0,79755234	1,030752	-2,919297	11,0807
9	8	0,97280984	1,637861	-1,281436	14,71856
10	9	0,1767674	-1,119711	-2,401147	15,59885
11	10	0,24392359	-0,887075	-3,288221	16,71178
12	11	0,78491179	0,986963	-2,301258	19,69874
13	12	0,83005406	1,143341	-1,157917	22,84208
14	13	0,17227059	-1,135288	-2,293205	23,70679
15	14	0,08082969	-1,452049	-3,745254	24,25475
16	15	0,03972789	-1,594429	-5,339683	24,66032
17	16	0,37897088	-0,419257	-5,75894	26,24106
18	17	0,08436329	-1,439808	-7,198748	26,80125
19	18	0,74316992	0,842365	-6,356383	29,64362
20	19	0,8830687	1,326989	-5,029394	32,97061
21	20	0,82902386	1,139772	-3,889622	36,11038
22	21	0,05462249	-1,542833	-5,432455	36,56755
23	22	0,65885022	0,550273	-4,882181	39,11782
24	23	0,77936355	0,967744	-3,914438	42,08556
25	24	0,3407385	-0,551698	-4,466136	43,53386
26	25	0,60415156	0,360792	-4,105344	45,89466
27	26	0,58379199	0,290264	-3,81508	48,18492
28	27	0,63250072	0,458996	-3,356084	50,64392
29	28	0,00089919	-1,728936	-5,08502	50,91498
30	29	0,04228401	-1,585575	-6,670595	51,32941
31	30	0,48263301	-0,060161	-6,730756	53,26924



Tab.11.2 *Hodnoty náhodné procházky*

Obr. 11.6 *Náhodná procházka bez driftu a s driftem* $\delta = 2$

11.4 DETERMINISTICKÝ A STOCHASTICKÝ TREND

Rozdíl mezi stacionárním a nestacionárním procesem – časovou řadou spočívá zejména v tom, že příslušný trend časové řady (viz kap. 9) je buď *deterministický* nebo *stochastický*. Zhruba řečeno, je-li trend časové řady plně předvídatelný a neměnný, říkáme, že je deterministický, není-li předvídatelný, řekneme, že je stochastický. Pro účely formálního zpřesnění této definice uvažujte následující model časové řady

$$Y_t = \beta_1 + \beta_2\, t + \beta_3 Y_{t-1} + u_t, \qquad (11.11)$$

kde u_t je bílý šum, t je časový index. Rozebereme následující možnosti:

Náhodná procházka: Ve vztahu (11.11) uvažujte $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 1$, potom obdržíte

$$Y_t = Y_{t-1} + u_t,$$

což je náhodná procházka bez driftu (11.5), která je nestacionární časovou řadou, jejíž první diference je podle (11.8) stacionární, neboť první diference je bílý šum. Proto se taková časová řada nazývá *diferenčně stacionární*.

Náhodná procházka s driftem: Ve vztahu (11.11) uvažujte $\beta_1 \neq 0$, $\beta_2 = 0$, $\beta_3 = 1$, potom obdržíte

$$Y_t = \beta_1 + Y_{t-1} + u_t,$$

což je náhodná procházka s driftem (11.9), která je nestacionární časovou řadou. Pro první diferenci je pak

$$\Delta Y_t = \beta_1 + u_t, \tag{11.12}$$

což představuje *stochastický trend*, který je buď pozitivní ($\beta_1 > 0$), nebo je negativní ($\beta_1 < 0$). První diference je podle (11.12) též stacionární, i když to obecně není bílý šum.

Deterministický trend: Ve vztahu (11.11) uvažujte $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 = 0$, potom obdržíte

$$Y_t = \beta_1 + \beta_2\, t + u_t, \tag{11.13}$$

což je nestacionární časová řada, kterou nazýváme *trendově stacionární*. Časová řada (11.13) je nestacionární proto, že její střední hodnota není konstantní v čase, neboť z (11.13) vyplývá

$$E(Y_t) = \beta_1 + \beta_2\, t, \tag{11.14}$$

a dále po dosazení z (11.13) a (11.14) obdržíte

$$Var(Y_t) = E(Y_t - E(Y_t))^2 = E(u_t)^2 = \sigma^2.$$

Náhodná procházka s driftem a deterministickým trendem: Ve vztahu (11.11) uvažujte $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 = 1$, potom obdržíte

$$Y_t = \beta_1 + \beta_2\, t + Y_{t-1} + u_t, \tag{11.15}$$

což po diferenciaci dává

$$\Delta Y_t = \beta_1 + \beta_2\, t + u_t, \tag{11.16}$$

a to je nestacionární časová řada.

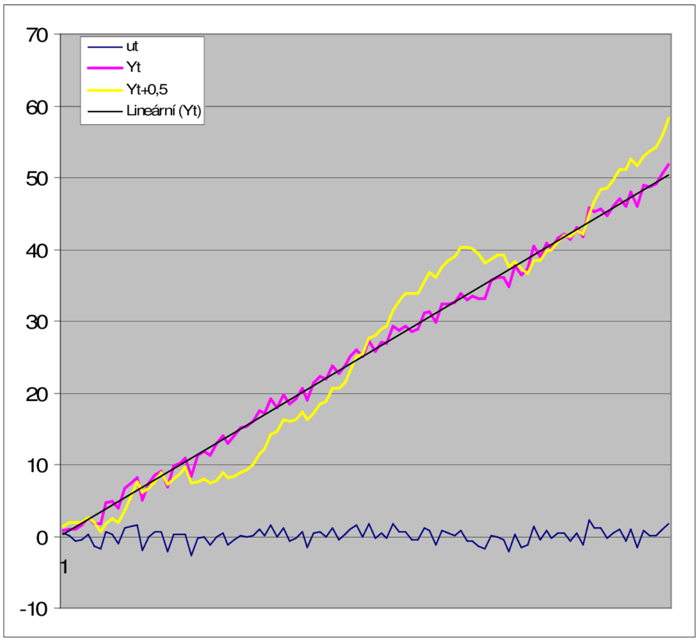
Příklad 4. Uvažujte stochastický proces

$$Y_t = 0{,}5t + u_t,$$

kde u_t je bílý šum, který při simulaci pro t = 1,2,...,100 generuje časovou řadu s deterministickým trendem, viz Obr. 11.7 (fialová křivka), a stochastický proces

$$Y_t = 0{,}5 + Y_{t-1} + u_t,$$

který při simulaci generuje časovou řadu stochastickým trendem, viz Obr. 11.7 (žlutá křivka).



Obr. 11.7. *Časová řada s deterministickým a stochastickým trendem*

Jak je zřejmé, v případě deterministického trendu tvoří odchylky od lineárního trendu $Y_t = 0{,}5t$ pouze bílý šum, zatímco v případě stochastického trendu ovlivňuje bílý šum dlouhodobý průběh časové řady.

11.5 JAK POZNÁME, ŽE ČŘ JE STACIONÁRNÍ?

Možná si kladete otázku, jak poznat stacionární, resp. nestacionární časovou řadu (ČŘ)? Přirozený návod nám dávají vlastnosti (11.1) – (11.3). Pokud alespoň jedna z těchto 3 podmínek není u časové řady splněna, potom je daná časová řada nestacionární. Zejména splnění prvních dvou podmínek: konstantní střední hodnoty ČŘ v čase a konstantní rozptyl ČŘ v čase, lze často odhadnout jednoduše grafickou metodou, tj. posouzením spojnicového grafu, viz Příklady 1 a 2. V některých případech však může být toto posouzení obtížné, zejména nemáme-li k dispozici dostatečně dlouhý časový úsek ČŘ a tehdy může pomoci třetí podmínka: analýza autokorelační funkce, resp. korelogramu. Všimněme si nejprve, jak vypadá korelogram nejjednodušších ČŘ: bílého šumu a náhodné procházky.

Z definice bílého šumu v subkapitole 11.3 přímo plyne, že korelační funkce ρ_k je nulová pro $k \neq 0$. Z toho plyne, že v korelogramu, tj. sloupcovém grafu, který zobrazuje hodnoty autokorelací pro posuvy $k = 1,2,...$ pomocí výšky sloupců, mají všechny sloupce nulovou výšku. V konkrétní ČŘ je třeba tento fakt ověřit statistickým testem hodnoty *výběrového autokorelačního koeficientu*, který z dané časové řady vypočteme podle vztahů

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \text{ pro } k = 1,2,\ldots, \tag{11.17}$$

kde

$$\hat{\gamma}_k = \frac{\sum (Y_t - \overline{Y})(Y_{t+k} - \overline{Y})}{n-k}, \text{ resp.} \tag{11.18}$$

$$\hat{\gamma}_0 = \frac{\sum (Y_t - \overline{Y})^2}{n-1}. \tag{11.19}$$

je *výběrový autokovarianční koeficient* řádu k, resp. *výběrový rozptyl* ČŘ. Hodnoty výběrové autokorelační funkce budou zcela jistě nenulové, statistickým testem zjistíme, zda jsou tyto hodnoty statisticky významné, tj. zda nulovou hypotézu: $\rho_k = 0$ zamítneme, či nikoliv. Pokud nulovou hypotézu nezamítnete pro všechna $k = 1, 2,...$, pak je původní ČŘ bílý šum. V opačném případě ČŘ není bílý šum (může však být přesto stacionární ČŘ, jak uvidíme v následující kapitole).

ŘEŠENÝ PŘÍKLAD 11.2

V následující tabulce je uvedena časová řada Y_t pro $t = 1,2,...,30$. Sestrojte korelogram a na jeho základě rozhodněte, zda se jedná o bílý šum. Použijte program SPSS.

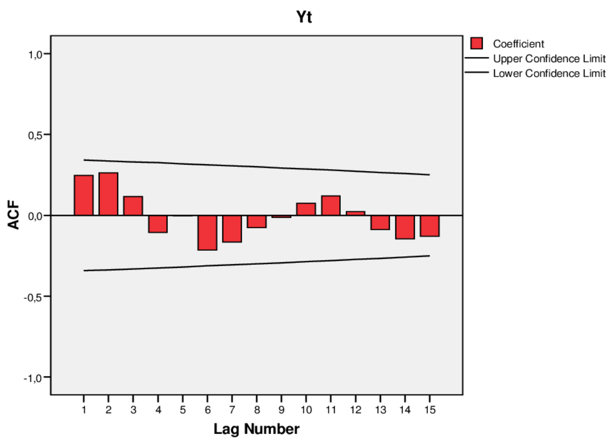
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y_t	-0,437	-0,360	1,014	0,210	-1,723	1,509	1,152	0,858	-1,157	1,462	1,701	1,557	0,089	-0,752	1,506
t	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Y_t	-0,145	-1,604	1,576	-1,511	-0,362	-0,255	1,239	0,469	0,376	0,534	-1,240	0,364	0,599	1,289	-0,723

Tab. 11.3. *Hodnoty časové řady*

Řešení:

Data z tabulky Tab. 11.3. zapíšeme v SPSS (v.15) do jediného sloupce editoru Data View. V záhlaví ji nazveme Yt. V hlavním menu vybereme postupně položky:

Analyze → Time Series→Autocorelations...
Proměnnou Yt vložíme do Variables, v Options... nastavíme
Maximum number of Lags: 15, Continue a potvrdíme OK.
Obdržíme následující výstup, který je požadovaným korelogramem.



Lag	Coeff
1	0,247
2	0,263
3	0,117
4	-0,105
5	-0,004
6	-0,216
7	-0,166
8	-0,076
9	-0,012
10	0,073
11	0,12
12	0,023
13	-0,088
14	-0,145
15	-0,13

Poloha čar, které stanovují horní a dolní konfidenční meze (Upper/Lower Confidence Limit) vymezuje pás, v němž všechny hodnoty autokorelační funkce ACF jsou statisticky nevýznamné (tedy s 95%-ní pravděpodobností nulové).

Nyní se podíváme na korelogram náhodné procházky (NP), tedy nestacionární ČŘ. Poměrně snadno se dá dokázat, že pro autokorelační funkci platí vztah (Arlt, 1999)

$$\rho_k = \sqrt{1 - \frac{k}{t}}\,. \qquad (11.20)$$

Již ze vztahu (11.20) je zřejmé, že autokorelační funkce závisí nejen na posuvu k, nýbrž také na konkrétním časovém indexu t. NP tedy nesplňuje také 3. podmínku stacionarity, tj. (11.3). Navíc je vidět, že pro $t \to +\infty$ konverguje ρ_k k hodnotě 1, tj. $\rho_k \to 1$. Z (11.20) je vidět, že u NP je pro k = 1,2,... hodnota autokorelační funkce blízká 1 a s rostoucím k tato hodnota pomalu klesá.

ŘEŠENÝ PŘÍKLAD 11.3

V Tab. 11.3. jsou hodnoty časové řady, o níž jste v Řešeném příkladu 11-2 rozhodli, že se jedná o bílý šum. Vytvořte z této ČŘ náhodnou procházku a s pomocí programu SPSS zobrazte její korelogram.

Řešení:

Z bílého šumu vytvoříme náhodnou procházku postupnou kumulací jeho hodnot. Data si takto připravíme v Excelu do následující tabulky

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y_t	-0,437	-0,797	0,217	0,426	-1,296	0,213	1,365	2,224	1,067	2,529	4,230	5,786	5,875	5,123	6,629
t	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Y_t	6,483	4,880	6,456	4,945	4,583	4,328	5,567	6,036	6,412	6,946	5,706	6,070	6,669	7,958	7,235

Tab. 11.4. *Příprava dat v Excelu*

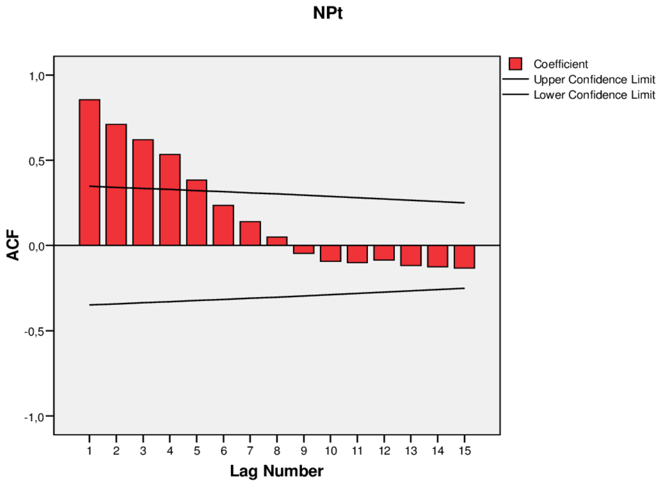
Data z tabulky zapíšeme v SPSS (v.15) do jediného sloupce editoru Data View. V záhlaví ji nazveme NPt. V hlavním menu vybereme postupně položky:

Analyze → Time Series→Autocorelations...

Proměnnou NPt vložíme do Variables, v Options... nastavíme

Maximum number of Lags: 15, Continue a potvrdíme OK.

Obdržíme následující výstup, který je požadovaným korelogmem.



Z korelogramu je vidět, že dochází k pozvolnému poklesu hodnot autokorelační funkce ACF, což potvrzuje fakt, že se jedná o časovou řadu, která je náhodnou procházkou.

Shrňme tedy poznatky z této kapitoly: Nejprve jsme se zaměřili na obecnější pojetí časové řady s pomocí pojmu stochastický (náhodný) proces. V klasickém pojetí je časová řada posloupnost číselných veličin v čase, v rozšířeném pojetí je to posloupnost náhodných veličin v čase. Speciální důležitou třídu tvoří stacionární procesy, zbývající procesy jsou nestacionární. Výjimečné místo mezi stacionárními procesy zaujímá bílý šum, mezi nestacionárními je to náhodná procházka. Oběma jsme se v této kapitole podrobněji zabývali. S rozšířením pojmu časové řady vzniká také rozšíření klasického deterministického trendu na stochastický trend. V závěru kapitoly jsme se zabývali problémem, jak rozhodnout, zda daná časová řada je nebo není stacionární. K řešení ilustračních příkladů jsme využili nejen Excel, ale také speciální statistický SW: SPSS.

11.6 Samostatné úkoly

11.1 V následující tabulce je uvedena časová řada U_t pro $t = 1,2,...,30$.

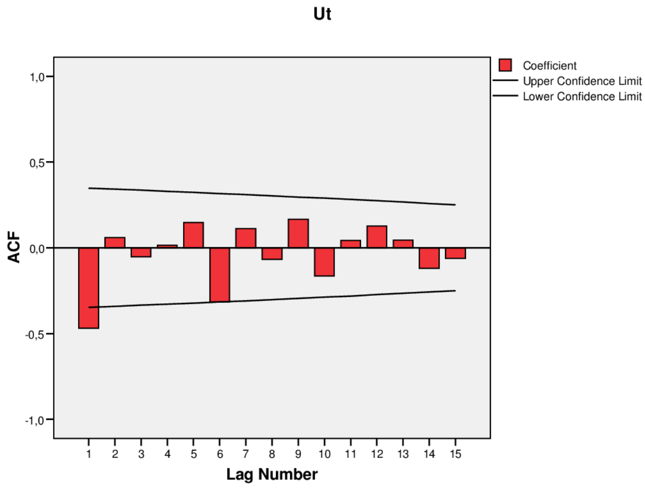
a. Sestrojte a zobrazte korelogram ČŘ a na jeho základě potvrďte, že se jedná o bílý šum.

b. Vytvořte z této ČŘ kumulací hodnot náhodnou procházku a s pomocí programu SPSS zobrazte její korelogram.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
U_t	1,657	0,835	-0,516	-1,656	-1,702	-1,260	-0,037	0,423	-1,634	-1,508	0,511	0,547	0,353	-0,331	-0,716
t	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
U_t	1,416	0,545	-0,997	0,990	0,931	0,188	-0,641	-0,295	-1,483	-0,380	-0,279	1,301	-0,065	0,813	-0,845

11.7 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

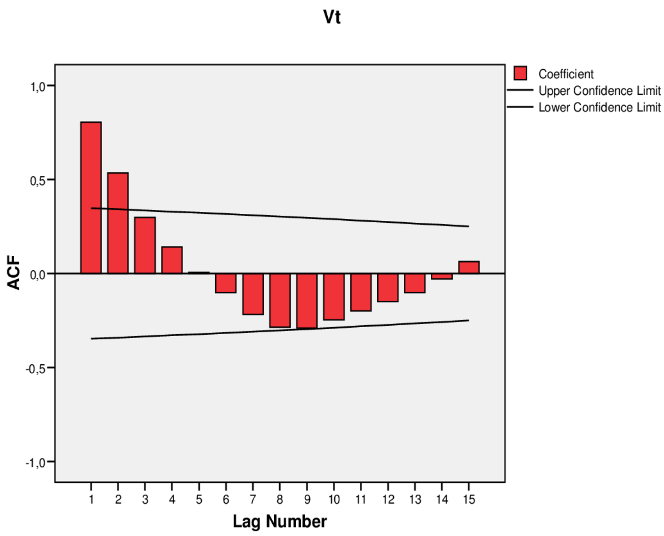
11.1 a. Jak je vidět z korelogramu, první hodnota autokorelační funkce vybočuje z 95%-ního intervalu, v němž jsou hodnoty statisticky nevýznamné. Zadaná ČŘ je proto bílým šumem „s nižší spolehlivostí“.



b. Z bílého šumu se vytvoří NP postupnou kumulací jeho hodnot v Excelu do následující tabulky:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
V_t	1,657	2,492	1,976	0,321	-1,381	-2,641	-2,678	-2,256	-3,889	-5,397	-4,886	-4,339	-3,986	-4,317	-5,033
t	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
V_t	-3,377	-2,831	-3,828	-2,838	-1,908	-1,719	-2,361	-2,656	-4,140	-4,520	-4,798	-3,498	-3,562	-2,750	-3,594

Korelogram potvrzuje, i když ne příliš přesvědčivě, že ČŘ V_t z kumulovaných hodnot U_t má charakter náhodné procházky.



12 MODELY TYPU ARIMA A PROGNÓZOVÁNÍ ČASOVÝCH ŘAD

RYCHLÝ NÁHLED KAPITOLY

Tato kapitola bezprostředně navazuje na kapitolu předchozí a to zejména v tom, že o vyšetřovaných časových řadách (v širším smyslu) budeme předpokládat, že jsou buď stacionární, nebo je lze na stacionární ČŘ převést (několikerým) diferencováním. Nejprve se budete zabývat časovými řadami typu ARIMA. Box-Jenkinsova metodologie, která se modely analýzy časových řad typu ARIMA zabývá, klade důraz nikoliv na konstrukci jedno-rovnicového nebo vícerovnicového modelu, jak je tomu např. v regresní analýze, nýbrž na analýzu vlastních stochastických vlastností ekonomických ČŘ. Postupně se seznámíte s vlastnostmi autoregresivních procesů AR, procesů pohyblivých průměrů MA, integračních procesů I, jakož i procesů vzniklých jejich kombinací: ARIMA. Dále lze tyto procesy rozšířit též na sezónní procesy. Úkolem pak je pro konkrétní proces – časovou řadu nalézt vhodný konkrétní proces (model) typu ARIMA a nalezený model použít pro účely prognózy (predikce, extrapolace) hodnot dané časové řady. Celý postup tvorby prognózy ČŘ autoři metody ARIMA formulovali ve 4 krocích, které nazýváme Box-Jenkinsova metodologie prognózování ČŘ. Jednotlivé kroky jsou (1) Identifikace modelu, (2) Odhad modelu, (3) Verifikace modelu a (4) Prognóza pomocí modelu. Jednotlivé kroky Box-Jenkinsovy metodologie budou ilustrovány na příkladu časové řady čtvrtletního HDP České republiky s pomocí statistického programu SPSS. K řešení ilustračních příkladů budete používat nejen Excel, nýbrž také speciální statistický SW: SPSS (Statistical Package for Social Sciences).

12.1 ÚVOD

Tuto kapitolu lze obtížně studovat izolovaně, neboť bezprostředně navazuje na kapitolu předchozí, se kterou je úzce propojena jak tematicky, tak použitými příklady. Ke zvládnutí látky kapitoly je podstatnou měrou využit statistický program SPSS, který je dostupný ve všech PC učebnách na SU OPF. SPSS je typem SW, který je do značné míry intuitivní a uživatelsky přátelský. Řešení příkladů uvedených v této kapitole, jež SPSS verzi 11.5 využívají, jsou podrobně komentována postupným procházením vložených menu, např. takto:

Analyze → Time Series → Autocorrelations…,

což znamená, že nejprve zvolíte položku hlavního menu Analyze, potom Time Series, konečně Autocorrelations… Jednotlivé prvky, které jsou součástmi SPSS, jsou odlišeny jiným fontem písma (Arial), zatímco ostatní text je psán fontem Times New Roman. Přesto před studiem této kapitoly naléhavě doporučujeme seznámit se podrobněji s hlavními funkcemi a způsobem ovládání programu SPSS, ať již starší verze 11.5, nebo nejnovější verze 20, která je již také na některých počítačových učebnách k dispozici. K tomu účelu může velmi dobře posloužit položka Tutorial, kterou naleznete pod položkou hlavního menu Help.
Na tomto místě zmíníme pouze jedinou informaci avšak prvořadé důležitosti: Přenos číselných dat mezi worksheetem v Excelu a Data View v SPSS funguje naprosto bezproblémově, a to na obě strany tak, jak jste zvyklí z MS Office: pomocí kombinace kláves Ctrl+C (kopírovat do schránky), Ctrl+V (vložit ze schránky).

Prognózování (předvídání, předpovídání) je důležitou součástí ekonomických (ekonometrických) analýz, dá se říci, že z určitého pohledu nejdůležitější. Jak prognózovat budoucí hodnoty ekonomických veličin, jako jsou HDP, inflace, kurzy měn, ceny akcií, míra nezaměstnanosti a dalších? Jednu klasikou metodu již znáte: lineární, (resp. nelineární) regresní analýza, s níž jste se seznámili již v kapitolách 3 a 4. V této kapitole se dozvíte o nové metodě, která se stala v posledních letech velmi populární: tzv. modely autoregresivních a integrovaných procesů a klouzavých průměrů - ARIMA (z angl. Auto Regresive Integrated Moving Average), která je známa také pod názvem Box-Jenkinsova metodologie (podle autorů metody G.P.E. Boxe a G.M. Jenkinse ze 70. let 20. století).

Téma ekonomického prognózování je velmi široké a existuje k němu množství specializovaných knih a dalších publikací. My zde chceme podat pouze stručný vhled do problematiky. Naštěstí k problematice prognózování ekonomických ČŘ existuje nejen vhodná literatura, její přehled lze nalézt např. u Arlta (1999), u Gujaratho (2003) aj., ale též příslušný specializovaný SW v podobě programových balíků jakými jsou SPSS (v současnosti je k dispozici na všech PC učebnách SU OPF), STATISTICA, SAS a další. V této kapitole budeme využívat konkrétně program SPSS, který obsahuje modul Time Series, umožňující modelování pomocí metody ARIMA.

Tato kapitola bezprostředně navazuje na kapitolu předchozí a to zejména v tom, že o níže vyšetřovaných časových řadách (v širším smyslu) budeme předpokládat, že jsou buď stacionární, nebo je lze na stacionární ČŘ převést (několikerým) diferencováním. Přitom zde využijeme pojmy zavedené v předchozí kapitole a přidáme ještě pojmy další, které se nám budou pro prognózování ČŘ hodit.

Jak jsme již dříve zmínili, k analýze ČŘ existuje řada různých metod a přístupů. Kromě již zmíněné (1) jednoduché regresní analýzy a (2) metody ARIMA, které jsou předmětem tohoto textu, je zapotřebí ještě jmenovat (3) metody *exponenciálního vyrovnání* (Holtova-Wintersova metoda a jejich varianty), (4) metody *simultánních rovnic* a (5) *vektorové autoregresivní metody* VAR, (6) metody ARCH a GARCH a další. S nimi se zájemci mohou blíže seznámit např. v Seger (1998).

12.2 MODELOVÁNÍ ČASOVÝCH ŘAD POMOCÍ ARIMA

Podle svých autorů známa jako Box-Jenkinsova metodologie, avšak technicky nazývaná ARIMA metodologie klade důraz nikoliv na konstrukci jednorovnicového nebo vícerovnicového modelu, jak je tomu např. v regresní analýze, nýbrž na analýzu vlastních stochastických vlastností ekonomických ČŘ podle filosofie „ať data hovoří sama za sebe". V regresních modelech je závisle proměnná Y vysvětlována několika vysvětlujícími proměnnými – regresory, zatímco v ARIMA metodách je závisle proměnná Y v čase t vysvětlována hodnotami téže Y v minulých časových okamžicích a zároveň chybovými členy v současných anebo minulých okamžicích. Na rozdíl od regresních modelů a modelů simultánních rovnic, které jsou založeny na ekonomické teorii, nejsou modely ARIMA na teorii přímo závislé. Teoretické závislosti jsou u nich vyjádřeny zprostředkovaně skrze sledované hodnoty v minulých časových okamžicích.

Příklad 1. Uvažujte stochastický proces Y_t „Čtvrtletní HDP České republiky" z kapitoly 11, viz www.czso.cz. Z ČŘ vypočítáme novou ČŘ indexu růstu I_t podle vztahu

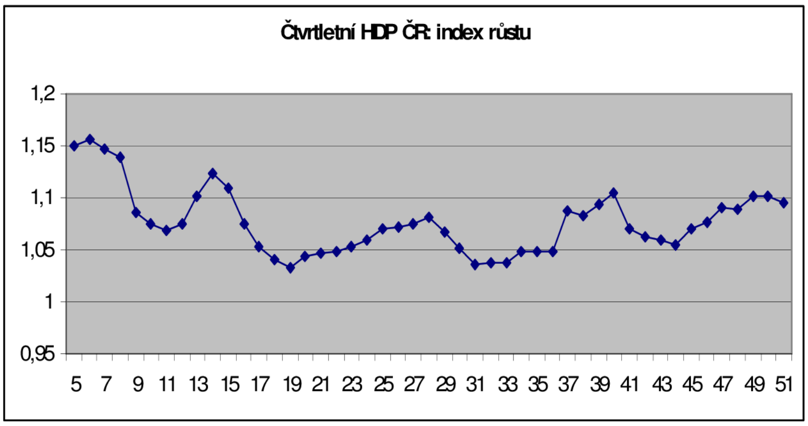
$$I_t = \frac{Y_t}{Y_{t-4}} \quad \text{pro } t = 5,6,...,30. \qquad (12.1)$$

Hodnoty indexu růstu čtvrtletního HDP jsou uvedeny v níže uvedené tabulce.

Je zřejmé, že se jedná o stochastický proces – časovou řadu v širším pojetí, kdy konečná hodnota za každé čtvrtletí je realizací náhodné veličiny (s neznámým rozdělením pravděpodobnosti), která může nabývat různých hodnot z číselného intervalu v závislosti na konkrétních ekonomických podmínkách závislých na mnoha faktorech. Grafické znázornění je na Obr. 12.1.

Q	I.95	II.95	III.95	IV.95	I.96	II.96	III.96	IV.96	I.97	II.97	III.97	IV.97
I_HDP_Q					1,150	1,156	1,147	1,139	1,085	1,075	1,069	1,075
Q	I.98	II.98	III.98	IV.98	I.99	II.99	III.99	IV.99	I.00	II.00	III.00	IV.00
I_HDP_Q	1,102	1,124	1,109	1,076	1,052	1,040	1,033	1,044	1,047	1,048	1,053	1,059
Q	I.01	II.01	III.01	IV.01	I.02	II.02	III.02	IV.02	I.03	II.03	III.03	IV.03
I_HDP_Q	1,071	1,072	1,074	1,081	1,068	1,052	1,036	1,037	1,038	1,048	1,048	1,049
Q	I.04	II.04	III.04	IV.04	I.05	II.05	III.05	IV.05	I.06	II.06	III.06	IV.06
I_HDP_Q	1,087	1,083	1,094	1,104	1,071	1,062	1,059	1,055	1,071	1,077	1,090	1,088
Q	I.07	II.07	III.07									
I_HDP_Q	1,101	1,101	1,095									

Tab. 12.1 *Index růstu I_t čtvrtletního HDP ČR*



Obr. 12.1 *Časová řada indexu růstu I_t čtvrtletního HDP ČR*

12.3 AUTOREGRESIVNÍ PROCES (AR)

Uvažujte ČŘ indexu růstu I_t z Tab. 12.1., který budeme pro konzistenci s kap. 11 označovat Y_t. Budeme předpokládat, že Y_t se chová podle vztahu

$$(Y_t - \mu) = \varphi_1(Y_{t-1} - \mu) + u_t, \qquad (12.2)$$

kde μ je střední hodnota Y_t a u_t je bílý šum, φ_1 je konstanta. V tom případě říkáme, že ČŘ Y_t je *autoregresivní proces 1. řádu*, neboli AR(1). Podle modelu (12.2) je prognóza $Y-\mu$ v čase t je přímo úměrná $Y-\mu$ v čase $(t-1)$ prostřednictvím koeficientu úměry φ_1 plus/mínus náhodná chyba (bílý šum). Pokud pro konstantu v modelu (12.2) platí $-1 < \varphi_1 < 1$, pak se dá ukázat, že proces AR(1) je stacionární. Dále si všimněte, že speciálně při $\varphi_1 = 0$ je z (12.2) proces AR(1) bílý šum a při $\varphi_1 = 1$ je z (12.2) proces AR(1) náhodná procházka. Také pro $\varphi_1 \geq 1$ nebo $\varphi_1 < -1$ je proces AR(1) nestacionární (Arlt, 1999).

Podobně *autoregresivní proces 2. řádu*, neboli AR(2) má tvar

$$(Y_t - \mu) = \varphi_1(Y_{t-1} - \mu) + \varphi_2(Y_{t-2} - \mu) + u_t. \qquad (12.3)$$

Analogicky *autoregresivní proces p-tého řádu*, neboli AR(p) má tvar

$$(Y_t - \mu) = \varphi_1(Y_{t-1} - \mu) + \varphi_2(Y_{t-2} - \mu) + \ldots + \varphi_p(Y_{t-p} - \mu) + u_t. \qquad (12.4)$$

Otázka stacionarity procesů AR(p) pro $p > 1$ je složitější problém, kterým se zde zabývat nebudeme. Eventuální zájemce odkazujeme na literaturu, např. knihu Arlt (1999).

Všimněte si, že kromě hodnot Y v různých časových okamžicích se ve výše uvedených modelech nevyskytují jiné regresory. V tomto smyslu říkáme, že „data hovoří sama za sebe“.

12.4 PROCES KLOUZAVÝCH PRŮMĚRŮ (MA)

Výše uvedený AR proces není jediný, kterým lze generovat hodnoty Y. Nyní budeme předpokládat, že Y_t se chová podle vztahu

$$(Y_t - \mu) = u_t - \theta_1 u_{t-1}, \qquad (12.5)$$

kde μ je střední hodnota Y_t a u_t je bílý šum. V tom případě říkáme, že ČŘ Y_t je *proces klouzavých průměrů 1. řádu*, neboli MA(1). Podle modelu (12.5) je prognóza $Y-\mu$ v čase t je přímo úměrná náhodné chybě v čase $(t-1)$ prostřednictvím koeficientu úměry $-\theta_1$ plus/mínus náhodná chyba (bílý šum).

Podobně *proces klouzavých průměrů 2. řádu*, neboli MA(2) má tvar

$$(Y_t - \mu) = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2}, \qquad (12.6)$$

Analogicky *proces klouzavých průměrů q-tého řádu*, neboli MA(q) má tvar

$$(Y_t - \mu) = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \ldots - \theta_q u_{t-q}. \qquad (12.7)$$

Jednoduše řečeno, proces klouzavých průměrů je lineární kombinací minulých náhodných chyb bílého šumu. Na rozdíl od AR procesů jsou procesy MA(q) pro všechna $q \geq 1$ stacionární nezávisle na hodnotách koeficientů θ_i.

12.5 AUTOREGRESIVNÍ PROCES KLOUZAVÝCH PRŮMĚRŮ (ARMA)

Časová řada, která má charakteristiky jak AR tak MA procesů, je ARMA proces. Konkrétně ARMA proces 1. řádu, tj. ARMA(1,1) má tvar

$$Y_t = \delta + \varphi_1 Y_{t-1} + u_t - \theta_1 u_{t-1}, \qquad (12.8)$$

kde δ je konstantní člen. Analogicky můžete uvažovat procesy ARMA(p,q), které mají p autoregresivních a q klouzavých členů. Vzhledem ke stacionaritě procesu MA(q) je podmínka stacionarity procesu ARMA(p,q) totožná s podmínkou stacionarity procesu AR(p). Jinak řečeno, proces ARMA(p,q) je stacionární, právě když je stacionární proces AR(p).

12.6 AUTOREGRESIVNÍ A INTEGROVANÝ PROCES KLOUZAVÝCH PRŮMĚRŮ (ARIMA)

Časové procesy, které jste doposud poznali, byly vesměs za určitých podmínek stacionární. Dobře však víte, že mnohé ekonomické časové řady jsou nestacionární. Říkáme, že časová řada Y_t, tj. *stochastický proces Y_t je integrovaný 1. řádu*, neboli je to I(1) proces, jestliže 1. diference této časové řady je stacionární. Jinak řečeno, ČŘ Y_t je integrovaná 1. řádu, jestliže $\Delta Y_t = Y_t - Y_{t-1}$ je stacionární ČŘ. Analogicky lze zavést pojem integrované časové řady d-tého řádu, jestliže d-tá diference této ČŘ je stacionární, neboli $\Delta^d Y_t = \Delta^{d-1} Y_t - \Delta^{d-1} Y_{t-1}$ je stacionární, přitom $\Delta^1 = \Delta$. Stacionární proces se této symbolice označuje jako I(0) proces.

Proto když nejprve proces d-krát diferencujeme a poté obdržíme ARMA(p,q) proces, nazývá se původní proces ARIMA(p,d,q). V tomto symbolickém vyjádření znamenají např. ARIMA(p,0,q) a ARMA(p,q) stejný proces, stejně tak ARIMA(0,0,q) = MA(q), ARIMA(p,0,0) = AR(p), ARMA(p,0) = AR(p), apod.

12.7 SEZÓNNÍ PROCESY ARIMA

Uvažujte nyní stochastický proces Y_t o němž je známo, že je sezónní se sezónní složkou a r sezónami, viz kapitolu 9, které budeme modelovat jako AR proces. Konkrétně si můžete vzít naši časovou řadu čtvrtletního HDP ČR z Příkladu 1, kde počet sezón (tj. čtvrtletí) je $r = 4$. Budeme proto předpokládat, že Y_t se chová podle vztahu

$$Y_t = \Phi_1 Y_{t-r} + u_t, \qquad (12.9)$$

kde u_t je bílý šum, Φ_1 je konstanta. V tom případě říkáme, že ČŘ Y_t je *sezónní autoregresivní proces 1. řádu*, neboli SAR(1) s r sezónami. Podobně můžeme uvažovat další sezónní ARIMA procesy, říkáme jim SARIMA procesy. Podrobněji se touto problematikou zabývat nebudeme, viz např. Arlt (1999).

12.8 BOX-JENKINSOVA METODOLOGIE PROGNÓZOVÁNÍ ČŘ

Představte si, že máte analyzovat nějakou časovou řadu, jako třeba čtvrtletní HDP ČR z příkladu 3, Tab. 11.2 v kapitole 11. Jak zjistíte, o který typ procesu se jedná? Jde o realizaci AR procesu, nebo snad MA procesu, či jejich kombinaci ARMA? Může byt konkrétní časová řada realizací více různých typů procesu, např. jak AR(1), tak současně MA(1)? V této souvislosti hledáme *model časové řady* a hned je třeba říci, že konkrétní časová řada může mít **několik** „správných“ modelů. Zda je model „správný“ ověříme postupem zvaným *verifikace modelu*, viz Krok 3 v následujícím postupu. Předtím však musíte v Kroku 1 identifikovat, o jaký typ procesu se ve vašem případě jedná, zda je to proces AR, MA, ARI apod., a také stanovit řád příslušného procesu, např. AR(2), kde řád $p = 2$, nebo ARI(2,1), kde se jedná o integrovaný proces I(1) řádu d=1 v kombinaci autoregresivním procesem AR(2) řádu $p = 2$,. Abyste mohli „správně“ prognózovat hodnoty časové řady v budoucnosti (Krok 4), musíte mít k dispozici „správný“ model ČŘ, který následně použijete pro výpočet prognózy. Celý postup tvorby prognózy ČŘ autoři metody ARIMA formulovali ve 4 krocích, které nazýváme Box-Jenkinsova metodologie prognózování ČŘ. Jednotlivé kroky si teď přiblížíme i s ohledem na použití statistického programu SPSS.

Krok1. **Identifikace modelu:** Stanovení typu modelu (AR, MA, I, ARMA, ARIMA, sezónnosti apod.) a řádů, tj. čísel p, d, q v modelu ARIMA(p,d,q). V sezónních modelech SARIMA se stanoví ještě další parametry *sp*, *sd*, *sq* (viz Analyze → Time Series → ARIMA v SPSS). Využívá se přitom tvarů ACF a PACF (viz další subkapitola).

Krok 2. **Odhad modelu:** Odhad parametrů modelu - výpočet koeficientů modelu φ_i a θ_j (v SPSS je použita metoda maximální věrohodnosti, což je obdoba metody nejmenších čtverců – MNČ). Diferencování modelu (d-krát) vede ke stacionarizaci ČŘ.

Krok 3. **Verifikace modelu:** Výpočet Rezidua – rozdílu mezi modelovými hodnotami a příslušnými hodnotami z dat. (V SPSS jsou to hodnoty proměnné ERR). Model je správný, pokud reziduum je bílý šum, jinak je třeba přejít na Krok 1 – k nové identifikaci a přehodnocení modelu. Tento krok do značné míry závisí na zkušenostech analytika – nejde o přísně exaktní postup (např. hodnocení tvarů ACF a PACF, resp. statistické významnosti hodnot ACF a PACF na hladině spolehlivosti 95%).

Krok 4. **Prognózování:** Výpočet modelem prognózovaných hodnot v zadaném časovém horizontu prognózy a intervalů spolehlivosti prognózy. (V SPSS jsou to hodnoty proměnné FIT, 95%UCL a 95%LCL).

Aplikaci jednotlivých kroků s využitím SPSS si ukážeme na konkrétním příkladu v závěru této kapitoly. Ještě předtím se seznámíte s dalšími nástroji a metodami, které se využívají v prvním kroku při identifikaci modelu ČŘ.

12.9 PROGNÓZOVÁNÍ POMOCÍ ARIMA MODELŮ

Významným nástrojem ke stanovení typu modelu (AR, MA, I, ARMA, ARIMA) je autokorelační funkce ρ_k, k = 1,2,..., (ACF) a korelogram, resp. výběrová autokorelační funkce $\hat{\rho}_k$, k = 1,2,..., a výběrový korelogram, s kterými jste se seznámili v subkapitolách 11.2. a 11.5. Korelace mezi 2 náhodnými veličinami je často způsobena tím, že obě tyto veličiny jsou korelovány s veličinou třetí. velká část korelace mezi veličinami Y_t a Y_{t-k} může být zapříčiněna jejích korelací s mezilehlými veličinami Y_{t-1}, Y_{t-2},,,,,Y_{t-k+1}. Pojem parciální autokorelace zachycuje korelaci mezi veličinami Y_t a Y_{t-k} očištěnou o vliv veličin mezi nimi. *Parciální autokorelační koeficient* ρ_{kk}, k = 0,1,2,..., (2 indexy kk) je analogií k pojmu parciální regresní koeficient. Uvažujte k-násobnou lineární regresi Y_t s regresory Y_{t-1}, Y_{t-2},,,,,Y_{t-k}:

$$Y_t = \rho_{k1}\, Y_{t-1} + \rho_{k2}\, Y_{t-2} + ... + \rho_{kk}\, Y_{t-k} + e_t. \qquad (12.10)$$

Regresní koeficient ρ_{kk} je ve (12.10) právě parciální autokorelační koeficient. Vztahu (12.10) se také využívá k výpočtu *výběrového parciálního autokorelačního koeficientu* $\hat{\rho}_{kk}$, viz Arlt (1999).

Důležitou roli hraje tzv. *parciální autokorelační funkce* (PACF) stochastického procesu ρ_{kk} pro k =0,1,2,… PACF má následující vlastnosti:

$$\rho_{00} = 1,$$
$$-1 \le \rho_{kk} \le 1 \text{ pro } k = 1,2,\ldots$$
$$\rho_{kk} = \rho_{-k,-k} \text{ pro } k = 1,2,\ldots, \text{ tj. PACF je symetrická kolem } k = 0.$$

Grafickým znázorněním PACF je *parciální korelogram*. Vzhledem k uvedeným vlastnostem stačí, aby parciální korelogram zobrazoval hodnoty pro posuvy $k > 0$. Výpočet PACF bývá dnes samozřejmou součástí statistických softwarových programů, např. SPSS.

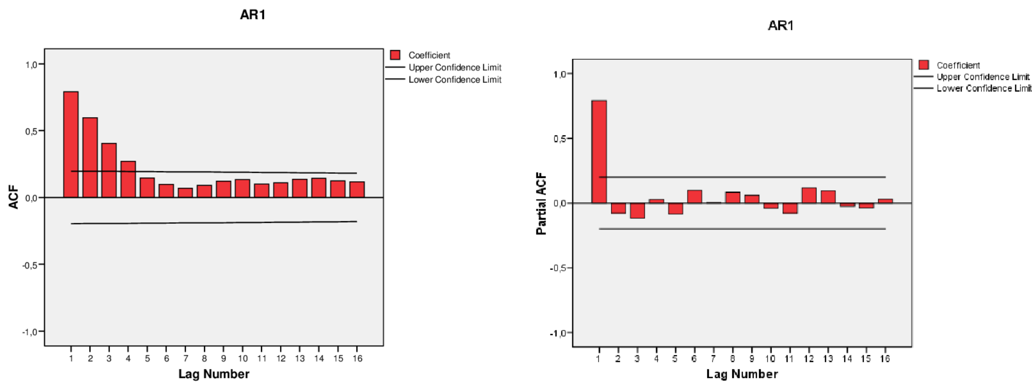
12.10 IDENTIFIKACE PROCESŮ ARIMA POMOCÍ ACF A PACF

Při identifikaci typu procesu ARIMA a jeho řádů využíváme charakteristických tvarů ACF a PACF. Různé typy procesů ARIMA mají charakteristické tvary korelogramů a parciálních korelogramů. V SPSS využíváme nabídku: Analyze → Time Series → Autocorrelations…

Jednotlivé typy procesů mají následující charakteristiky:

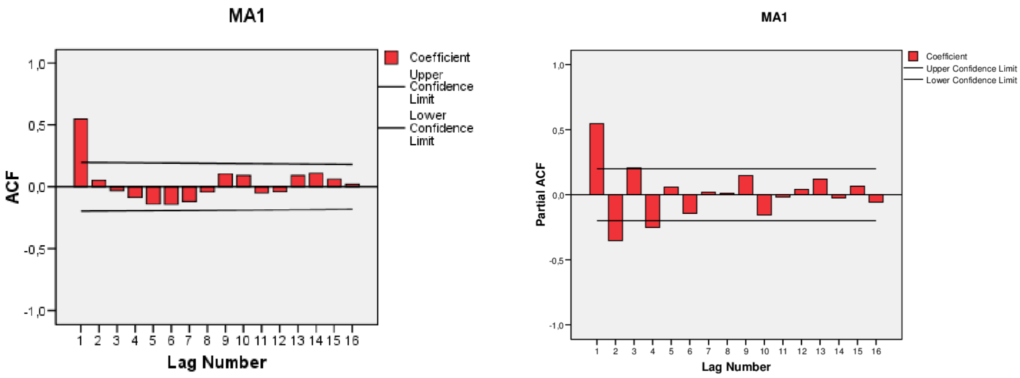
a. Proces AR(*p*): Prvních p hodnot PACF je „velkých“, další = 0 a „rychlý“ pokles (v absolutních hodnotách) ACF.

Příklad korelogramů AR(1):



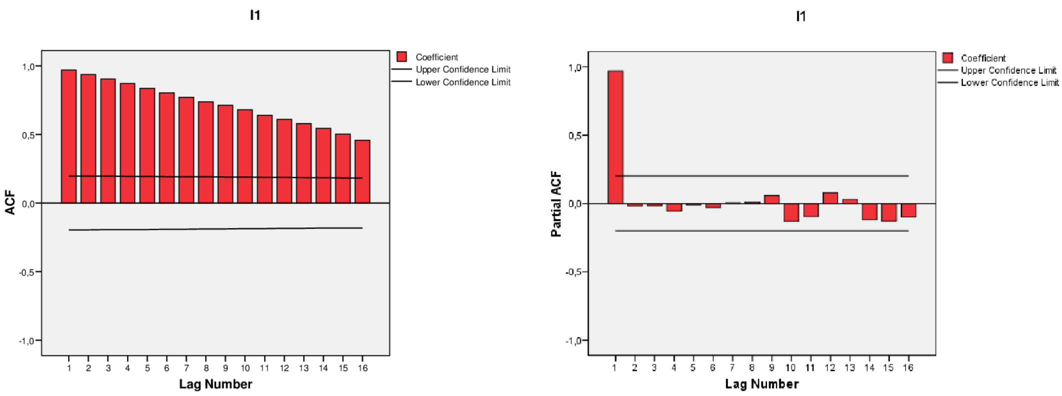
b. Proces MA(*q*): Prvních *q* hodnot ACF je „velkých“, další = 0 a „rychlý“ pokles (v absolutních hodnotách) PACF.

Příklad korelogramů MA(1):



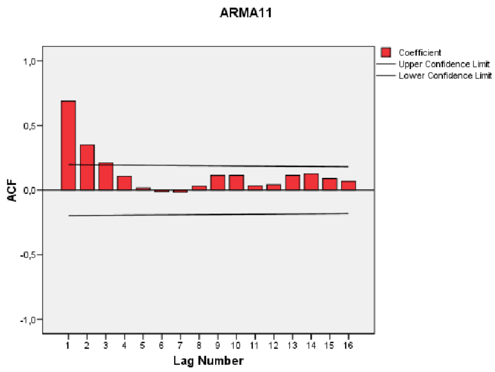
c. Proces I(*d*): „Pomalý“ pokles ACF, prvních *d* hodnot PACF je „velkých“, další = 0.

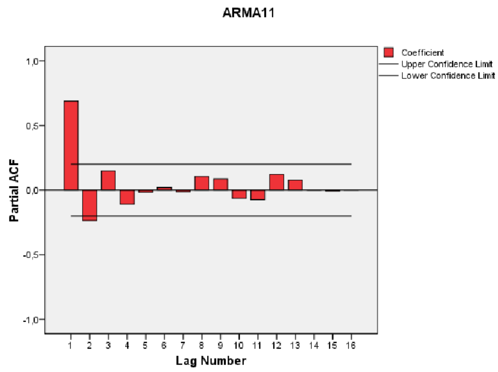
Příklad korelogramů I(1): „Náhodná procházka“



d. Proces ARMA(*p*,*q*): Prvních *q* hodnot ACF je „velkých“, další = 0 a prvních *p* hodnot PACF je „velkých“, další = 0.

Příklad korelogramů ARMA(1,1):



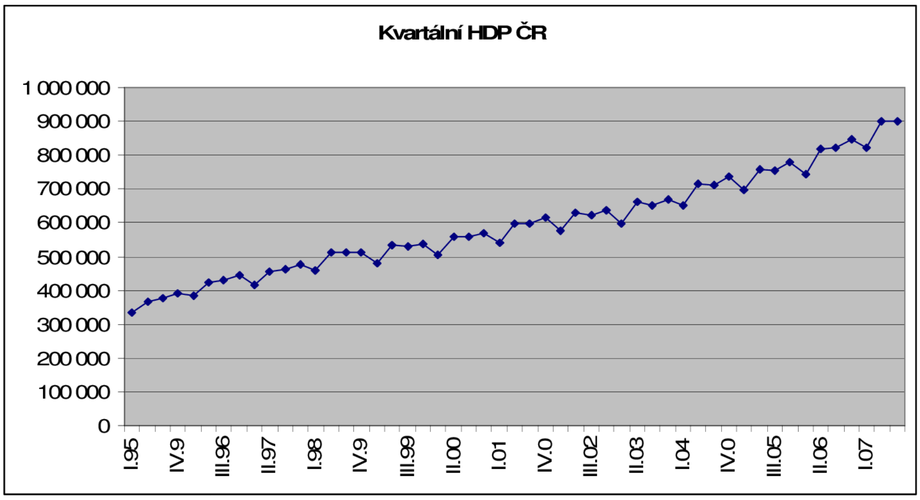


ŘEŠENÝ PŘÍKLAD 12.1

Uvažujte ještě naposledy časovou řadu „Čtvrtletní HDP České republiky“ z Příkladu 3 v kap. 11, viz www.czso.cz . Hodnoty časové řady jsou uvedeny v následující tabulce Tab. 12.2 a zobrazeny v grafu na Obr. 12.2:

Q	I.95	II.95	III.95	IV.95	I.96	II.96	III.96	IV.96	I.97	II.97	III.97	IV.97
HDP_Q	332995	366618	376688	390221	382859	423953	432152	444324	415593	455790	461902	477809
Q	I.98	II.98	III.98	IV.98	I.99	II.99	III.99	IV.99	I.00	II.00	III.00	IV.00
HDP_Q	457925	512225	512408	513925	481895	532968	529465	536469	504479	558691	557780	568219
Q	I.01	II.01	III.01	IV.01	I.02	II.02	III.02	IV.02	I.03	II.03	III.03	IV.03
HDP_Q	540124	598842	599262	613986	576665	630141	621004	636622	598385	660401	650791	667533
Q	I.04	II.04	III.04	IV.04	I.05	II.05	III.05	IV.05	I.06	II.06	III.06	IV.06
HDP_Q	650448	715163	712103	737048	696387	759740	753836	777759	745496	817867	821754	846459
Q	I.07	II.07	III.07									
HDP_Q	820689	900606	900022									

Tab. 12.2 *HDP ČR v letech 1995 – 2007*



Obr. 12.2 *HDP ČR ČR v letech 1995 – 2007: grafické znázornění*

Nalezněte vhodný ARIMA model této časové řady a pomocí něj prognózujte čtvrtletní hodnoty HDP až do konce roku 2009.

Řešení:

K řešení využijeme Box-Jenkinsovu metodologii prognózování ČŘ formulovanou ve 4 krocích popsaných v subkapitole 12.3. Použijeme k tomu statistický program SPSS. Zde v Data View do proměnné nazvané HDP_Q uložíme 51 hodnot kvartálního HDP. V menu:

Data → Define Data →Years, Quaters

vytvoříme hodnoty časové osy (počínaje rok – čtvrtletí):

Year: 1995

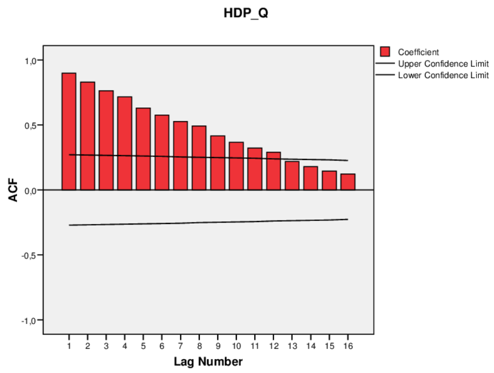
Quarter: 1

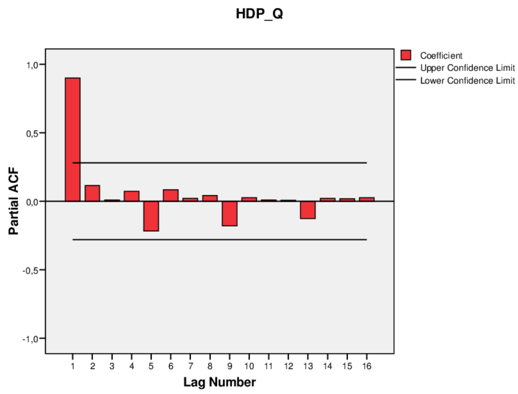
Krok 1: Identifikace modelu procesu ARIMA.

Z prostého pohledu na spojnicový graf na Obr. 12.2. lze usoudit, že se jedná o nestacionární časovou řadu, zároveň vykazuje sezónní složku se 4 sezónami. Tento předpoklad potvrdíme analýzou korelogramů ACF a PACF. V menu:

Analyze → Time Series → Autocorrelations…

vložíme proměnnou HDP_Q a ve výstupu Output obdržíme korelogramy:

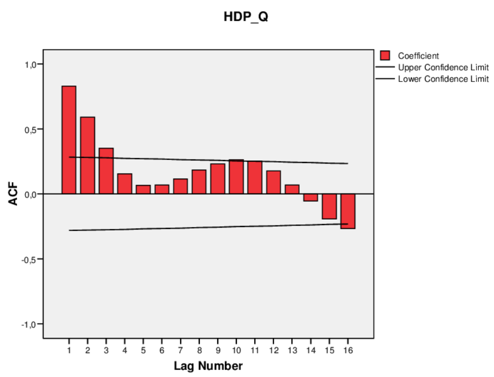


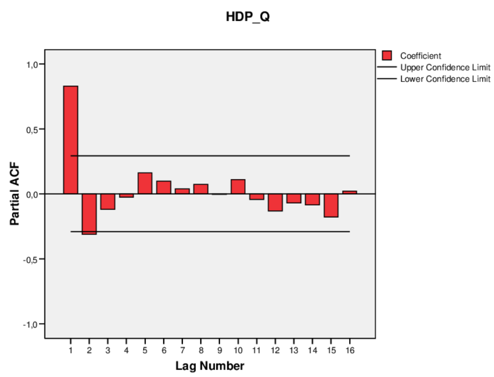


V korelogramu hodnoty ACF pomalu klesají, v PACF je „velká" první hodnota, ostatní jsou statisticky nevýznamné, tedy nulové, neboť se nacházejí v pásu 95% spolehlivosti. Z toho vyvozujeme, že se jedná o nestacionaritu 1. řádu, tj. typu I(1). Stacionarizujeme proto ČŘ jedním diferencováním a zobrazíme korelogramy této ČŘ: V menu:

Analyze → Time Series → Autocorrelations…,

dále zvolíme Seasonally difference: 1 , obdržíme korelogramy:





Z tvarů kolerogramů je zřejmé, že stacionarizovaný proces je typu AR(1), srovnejte s příkladem v subkapitole 12.4.1. Celkovým výsledkem kroku identifikace je, že naše časová řada je ARI(1,1) proces se sezónní složkou (se 4 sezónami) řádu 1.

Krok 2: Odhad parametrů modelu - výpočet koeficientů modelu provedeme v programu SPSS v menu:

Analyze → Time Series → ARIMA…,

vložíme proměnnou Dependent: HDP_Q, Independent: QUARTER_, dále zvolíme:

Autoregressive: p = 1, Seasonal sp = 1

Difference: d = 1, Seasonal sd = 1

Moving Average: q = 0, Seasonal sq = 0

Potvrdíme OK a v Output obdržíme vypočítané parametry:

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	0,168	0,136	1,235	0,223
Seasonal Lags	Seasonal AR1	0,947	0,044	21,726	0,000
Regression Coefficients	QUARTER, period 4	12614,393	6130,4	2,058	0,045
Constant		14435,108	13570,9	1,064	0,293

Melard's algorithm was used for estimation.

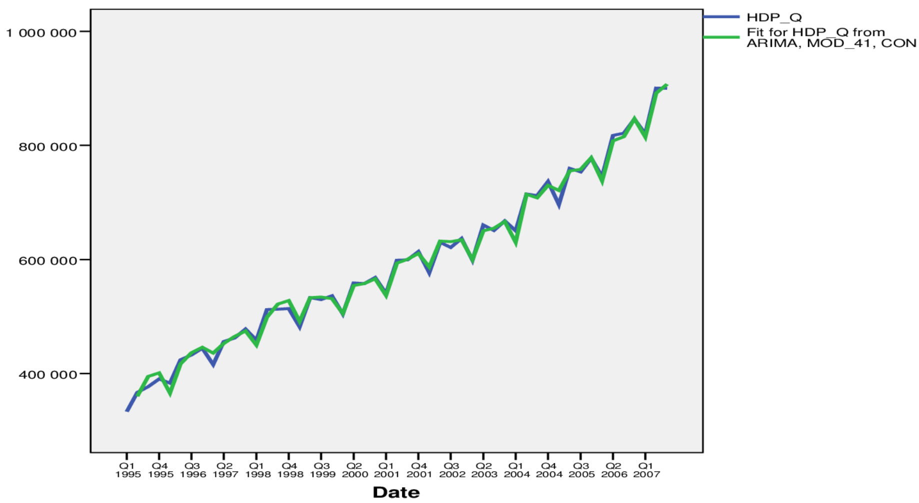
Dále zobrazíme v jednom grafu hodnoty ČŘ (modrá křivka) a hodnoty modelu ČŘ (zelená křivka). Použijeme k tomu menu:

Analyze → Time Series → Sequence Charts…,

vložíme proměnné: Variables: HDP_Q, Fit for HDP_Q,

Time Axis Labels: QUARTER_,

potvrdíme OK. Ve výstupu Output obdržíme graf:

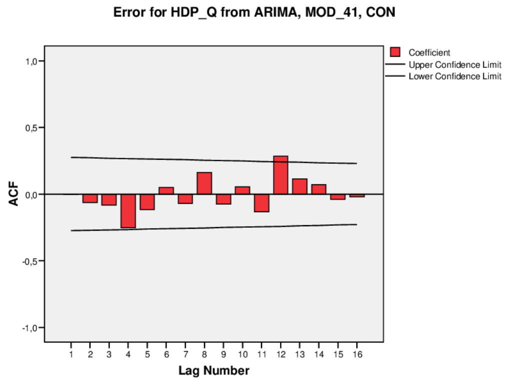


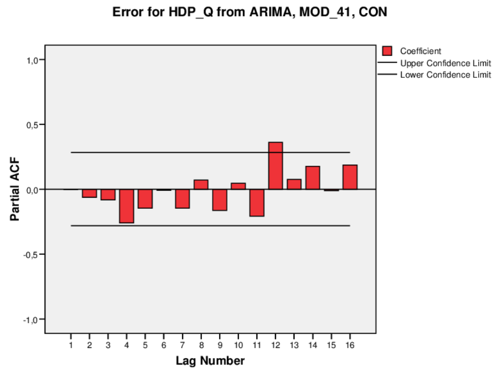
V Data View se vytvořilo 5 nových proměnných: FIT, ERR, LCL, UCL, SEP.

Krok 3: Verifikace modelu – spočívá v ověření předpokladu, že reziduum tj. odchylka modelu od dat, je bílý šum. V SPSS jsou to hodnoty proměnné ERR. Model je správný, pokud reziduum je bílý šum. V menu:

Analyze → Time Series → Autocorrelations…

vložíme proměnnou ERR , potvrdíme OK a ve výstupu Output obdržíme korelogramy:





Výše uvedené korelogramy potvrzují, že ACF i PACF jsou nulové, proto Reziduum ERR je bílý šum (i když u obou korelogramů v jednom případě hodnota mírně přesahuje pás 95% spolehlivosti).

Krok 4: Prognózu vytvoříme v prognózovaném časovém období 2007_IV až 2009_IV pomocí verifikovaného modelu. V menu:

Analyze → Time Series → ARIMA…,

ponecháme proměnnou Dependent: HDP_Q, Independent: QUARTER_, a také hodnoty řádů: p = 1, d = 1, q = 0, sp = 1, sd = 1, sq = 0. Navíc klikneme na tlačítko SAVE a ve vloženém okně klikneme Predict trough a vyplníme konec prognózovaného intervalu:

Year: 2009
Quarter: 4

Potvrdíme CONTINUE a OK.

V Data View se vytvořilo 5 nových proměnných: FIT, ERR, LCL, UCL, SEP s hodnotami také v časovém intervalu predikce (kromě proměnné ERR). Obdržíte mimo jiné hodnoty predikce čtvrtletního HDP ČR:

Q	HDP_Q
IV.07	923892
I.08	898085
II.08	975193
III.08	976058
IV.08	1000094
I.09	974413
II.09	1048887
III.09	1051130
IV.09	1075325

Nakonec zobrazíme ČR od roku 2005 do konce roku 2009 včetně 95% intervalu spolehlivosti prognózy. V SPSS nejprve vybereme zobrazovaný časový interval. V menu:

Data → Select Cases...

klikneme tlačítka Based on time or case range a Range, pak vybereme zobrazovaný interval (od - rok, čtvrtletí do - rok čtvrtletí):

Year: 2005 Year: 2009
Quarter: 1 Quarter: 4

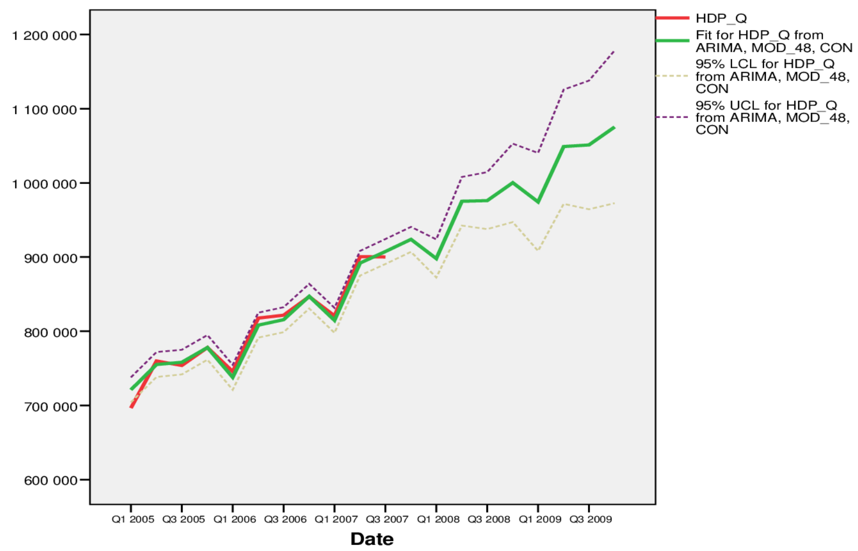
Poté v menu:

Analyze → Time Series → Sequence Charts…,

vložíme proměnné

Variables: HDP_Q, Fit for HDP_Q,
Time Axis Labels: QUARTER_,

potvrdíme OK. Ve výstupu Output obdržíme po mírné editaci graf:



Všechny zadané úkoly jsou tímto vyřešeny.

Můžeme tedy shrnout: Tato závěrečná kapitola 12 bezprostředně navázala na předchozí kapitolu 11. Nejprve jste se seznámili s časovými řadami (procesy) typu ARIMA. Box-Jenkinsova metodologie, která se touto problematikou zabývá, klade důraz nikoliv na konstrukci jednorovnicového nebo vícerovnicového modelu, jak tomu bylo např. v regresní analýze, nýbrž na analýzu vlastních stochastických vlastností ekonomických ČŘ. Postupně jste se seznámili s vlastnostmi autoregresivních procesů AR, procesů pohyblivých průměrů MA, integračních procesů I, jakož i procesů vzniklých jejich kombinací ARIMA. Dále byly tyto procesy rozšířeny též na sezónní procesy. Úkolem pak bylo pro konkrétní proces – časovou řadu nalézt vhodný konkrétní proces (model) typu ARIMA a nalezený model použít pro účely prognózy (predikce, extrapolace) hodnot dané časové řady. Celý postup tvorby prognózy ČŘ autoři metody ARIMA formulovali ve 4 krocích, které nazýváme Box-Jenkinsova metodologie prognózování ČŘ. Jednotlivé kroky jsou (1) Identifikace modelu, (2) Odhad modelu, (3) Verifikace modelu a (4) Prognóza pomocí modelu. Jednotlivé kroky Box-Jenkinsovy metodologie byly ilustrovány na příkladu konkrétní časové řady čtvrtletního HDP České republiky s pomocí statistického programu SPSS.

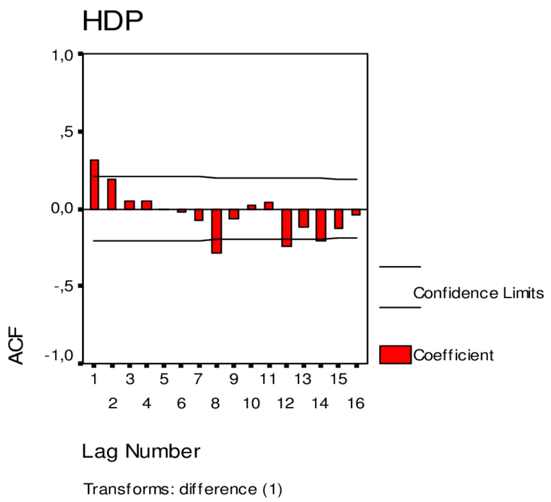
12.11 SAMOSTATNÉ ÚKOLY

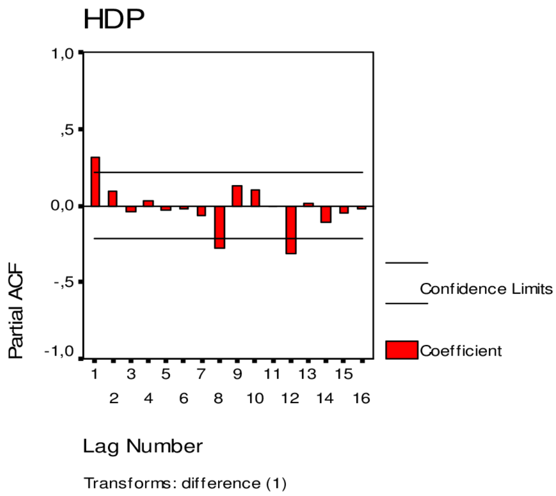
12.1 Uvažujte časovou řadu „Čtvrtletní HDP USA“ v letech 1970 - 1991. Hodnoty časové řady jsou uvedeny v následující tabulce. Najděte vhodný ARIMA model této časové řady a pomocí něj prognózujte čtvrtletní hodnoty HDP až do konce roku 1994 (chybějící hodnoty v tabulce). Použijte přitom 4 kroky Box-Jenkinsovy metodologie.

Q.Rok	HDP_Q	Q.Rok	HDP_Q	Q.Rok	HDP_Q	Q.Rok	HDP_Q	Q.Rok	HDP_Q
I.70	2872,8	I.75	3154	I.80	3830,8	I.85	4221,8	I.90	4880,8
II.70	2860,3	II.75	3190,4	II.80	3732,6	II.85	4254,8	II.90	4900,3
III.70	2896,6	III.75	3249,9	III.80	3733,5	III.85	4309	III.90	4903,3
IV.70	2873,7	IV.75	3292,5	IV.80	3808,5	IV.85	4333,5	IV.90	4855,1
I.71	2942,9	I.76	3356,7	I.81	3860,5	I.86	4390,5	I.91	4824
II.71	2947,4	II.76	3369,2	II.81	3844,4	II.86	4387,7	II.91	4840,7
III.71	2966	III.76	3381	III.81	3864,5	III.86	4412,6	III.91	4862,7
IV.71	2980,8	IV.76	3416,3	IV.81	3803,1	IV.86	4427,1	IV.91	4868
I.72	3037,3	I.77	3466,4	I.82	3756,1	I.87	4460	I.92	
II.72	3089,7	II.77	3525	II.82	3771,1	II.87	4515,3	II.92	
III.72	3125,8	III.77	3574,4	III.82	3754,4	III.87	4559,3	III.92	
IV.72	3175,5	IV.77	3567,2	IV.82	3759,6	IV.87	4625,5	IV.92	
I.73	3253,3	I.78	3591,8	I.83	3783,5	I.88	4655,3	I.93	
II.73	3267,6	II.78	3707	II.83	3886,5	II.88	4704,8	II.93	
III.73	3264,3	III.78	3735,6	III.83	3944,4	III.88	4734,5	III.93	
IV.73	3289,1	IV.78	3779,6	IV.83	4012,1	IV.88	4779,7	IV.93	
I.74	3259,4	I.79	3780,8	I.84	4089,5	I.89	4809,8	I.94	
II.74	3267,6	II.79	3784,3	II.84	4144	II.89	4832,4	II.94	
III.74	3239,1	III.79	3807,5	III.84	4166,4	III.89	4845,6	III.94	
IV.74	3226,4	IV.79	3814,6	IV.84	4194,2	IV.89	4859,7	IV.94	

12.12 ŘEŠENÍ ÚKOLŮ, VÝSLEDKY

12.1 a) Identifikace modelu





Vybíráme model ARIMA (1; 1; 0).

b) Odhad parametrů modelu

FINAL PARAMETERS:

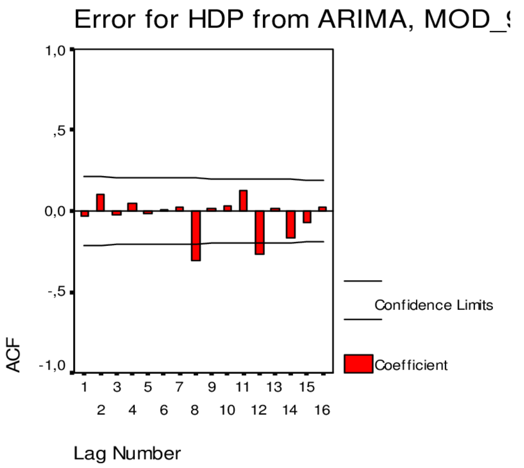
Number of residuals	87
Standard error	34,260641
Log likelihood	-429,95729
AIC	863,91457
SBC	868,84639

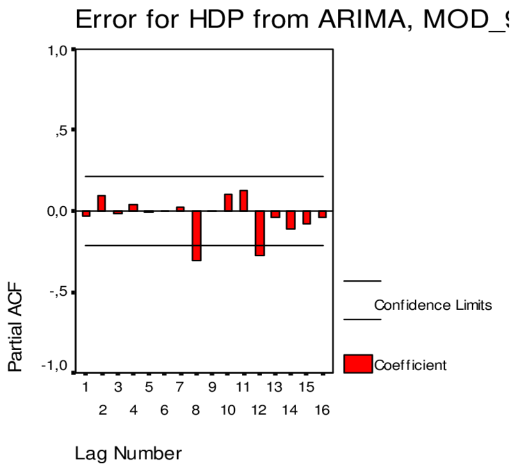
Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	**,317328**	,1028097	3,0865560	**,00273359**
CONSTANT	22,652801	5,3519931	4,2325916	,00005813

Koeficient AR1 = 0,32 je statisticky významný na hladině významnosti 0,05 (protože hodnota 0,0027 je menší než 0,05).

c) Verifikace modelu

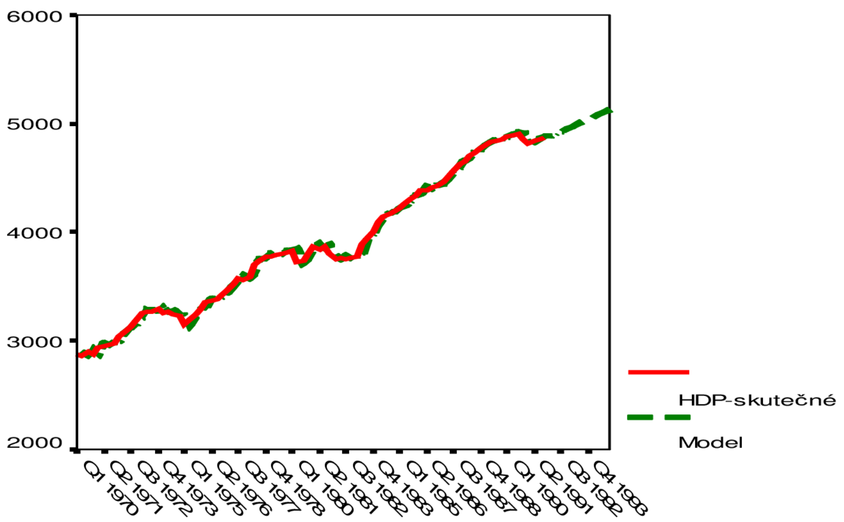




Korelogramy potvrzují, že ACF i PACF jsou nulové, proto Reziduum ERR je bílý šum (i když u obou korelogramů ve dvou případech hodnota mírně přesahuje pás 95% spolehlivosti).

d) Prognóza do 4.čtvrtletí 1994

	Bodový odhad	Intervalový odhad (95%)	
Q1 1992	4885,1	4816,6	4953,6
Q2 1992	4906	4792,1	5019,9
Q3 1992	4928,2	4777,3	5079
Q4 1992	4950,6	4768,4	5132,8
Q1 1993	4973,2	4763,4	5182,9
Q2 1993	4995,8	4761,2	5230,5
Q3 1993	5018,5	4760,8	5276,2
Q4 1993	5041,1	4761,9	5320,3
Q1 1994	5063,8	4764,3	5363,3
Q2 1994	5086,5	4767,5	5405,4
Q3 1994	5109,1	4771,5	5446,7
Q4 1994	5131,8	4776,2	5487,3



ZÁVĚR

Tento text představuje studijní oporu pro studium všech akreditovaných studijních programů v navazujícím magisterském studiu na Slezské univerzitě, Obchodně podnikatelské fakultě v Karviné. Předmět Statistické zpracování dat navazuje na předmět Statistika (dříve Kvantitativní metody B) obsahující základní bakalářský kurz statistiky na SU OPF, nebo na obdobný ekvivalentní předmět základů statistiky v bakalářském stupni studia na jiné VŠ ekonomického zaměření v ČR. Tento text je inovací předchozí studijní opory s názvem Statistika pro navazující magisterské studium, specializované pro studenty distanční a kombinované formy studia. Inovací studijních oborů na SU OPF v rámci projektu OPVK vznikl také předmět Statistické zpracování dat. V tomto předmětu je kladen důraz především na uplatnění statistických metod při zpracování ekonomických dat v aplikovaných ekonomických disciplínách, jako jsou zejména marketing a management.

Samotný učební text, nebo jak se říká v moderní terminologii: studijní opora - umožňující studentovi plnohodnotné a zároveň samostatné studium – je rozčleněn do 12 tematických kapitol. Jednotlivé kapitoly odpovídají obvyklým výukovým týdnům jednoho semestru a jsou přibližně stejně obsahově rozsáhlé a obtížné. Takový rozsah učiva odpovídá klasické dvouhodinové přednášce v prezenčním studiu na vysoké škole ekonomického zaměření. V prezenčním studiu je ovšem na rozdíl od kombinované formy studia přednáška doplněna seminářem, kde se probraná látka aplikuje na konkrétní číselné příklady, které se řeší až k požadovanému výsledku pomocí počítače.

Vysokoškolské studium v případě předmětu Statistické zpracování dat vyžaduje enormní úsilí studenta zaměřené na pravidelnost a vytrvalost ve studiu i samostudiu, schopnost koncentrace na předmět, aktivní přístup spočívající na samostatném řešení příkladů. V tom všem by tato studijní opora měla studentům kombinované formy studia pomoci nahradit kvalitní prezenční výuku i úlohu učebnic a skript. Pro lepší zvládnutí látky jsou vám v elektronické verzi kurzu Statistické zpracování dat k dispozici ještě doplňkové materiály v elektronické podobě. Dalšími podpůrnými zdroji ke studiu mohou být klasické učebnice a skripta a další doporučená literatura.

SEZNAM DOPORUČENÉ LITERATURY

ANDĚL,Jiří, 2007. *Statistické metody.* 4. upr. vyd. Praha: Marfyzpress, 299 s. ISBN 80-7378-003-8.

ARLT, Josef, 1999. *Moderní metody modelování ekonomických časových řad*. 1.vyd. Praha: Grada Publishing, 307 s. ISBN 80-716-9539-4.

CIPRA,Tomáš, 1986. *Analýza časových řad s aplikacemi v ekonomii.* 1.vyd. Praha: Státní nakladatelství technické literatury, 246 s.

GUJARATI, Damodar N, c2003. *Basic econometrics*. 4th ed. Boston: McGraw-Hill, xxix, 1002 s. ISBN 978-0-07-233542-2.

HÁTLE, Jaroslav a LIKEŠ, Jiří, 1974. *Základy počtu pravděpodobnosti a matematické statistiky*. 2. vyd. Praha: SNTL. 463 s.

HINDLS, Richard, SEGER, Jan a HRONOVÁ, Stanislava, 2002. *Statistika pro ekonomy*. 1. vyd. Praha: Professional Publishing, 415 s. ISBN 80-864-1926-6.

KAŇKA, Miloš, 1998. *Vybrané partie z matematiky pro ekonomy*. 1.vyd. Praha: VŠE, 231 s. ISBN 80-707-9537-9.

MAREK, Luboš a kol., 2007. *Statistika pro ekonomy: aplikace*. 2. vyd. Praha: Professional Publishing. 485 s. ISBN 978-80-86946-40-5.

RAMÍK, Jaroslav a Šárka ČEMERKOVÁ, 2000. *Statistika A*. Vyd. 3., rozš. a upr. V Opavě: Slezská univerzita, Obchodně podnikatelská fakulta v Karviné, 162 s. ISBN 80-7248-097-9.

RAMÍK, Jaroslav a Šárka ČEMERKOVÁ, 2000. *Statistika B*. Vyd. 2., rozš. a upr. V Opavě: Slezská univerzita, Obchodně podnikatelská fakulta v Karviné, 143 s. ISBN 80-724-8099-5.

RAMÍK, Jaroslav a Šárka ČEMERKOVÁ, 2003. *Kvantitativní metody B: statistika*. Vyd. 1. Karviná: Slezská univerzita v Opavě, Obchodně podnikatelská fakulta v Karviné, 206 s. ISBN 80-724-8198-3.

SEGER, Jan, HRONOVÁ, Stanislava a HINDLS, Richard, 1998. *Statistika v hospodářství*. 1.vyd. Praha: ETC Publishing, 636 s. ISBN 80-860-0656-5.