Series on Advanced Economic Issues
Faculty of Economics, VŠB-TU Ostrava

Dušan Marček

PRAVDEPODOBNOSTNÉ MODELOVANIE A SOFT COMPUTING V EKONOMIKE

Ostrava, 2013

Dušan Marček
Department of Applied Informatics
Faculty of Economics
VŠB-Technical University Ostrava
Sokolská 33
701 21 Ostrava, CZ
dusan.marcek@vsb.cz

Recenze

Valerie Novitzká, Technická univerzita v Košiciach
Milan Terek, Ekonomická univerzita v Bratislave

This work was supported by European Social Fund within the project CZ.1.07/2.3.00/20.0296.

The text should be cited as follows: Marček, D. (2013). Pravdepodobnostné modelovanie a soft computing v ekonomike, SAEI, vol. 18. Ostrava: VSB-TU Ostrava.

Printed in Tiskárna Grafico, s.r.o.
Cover design by MD communications, s.r.o.

ISBN 978-80-248-2955-5

Predhovor

V ekonomickom modelovaní snáď najviac je využívaná teória pravdepodobnosti. Spomenieme napr. klasické modely založené na regresnej, faktorovej, kanonickej a komponentnej analýze, modely logistickej regresie a novšie modely ako sú napr. modely založené na Kalmanovej filtrácii, Box-Jenkinsovej metodológií, teórie kointegračnej analýzy a modely autoregresné s podmienenou heteroskedasicitou. V poslednom období zaznamenali náhly rozmach v ekonomickom modelovaní soft výpočtové metódy (anglický názov *soft computing* – SC). Slovné spojenie soft computing je v domácej odbornej terminológii zaužívané. SC je veľmi blízky známemu informatickému odboru umelej inteligencii vo zmysle inteligentného riešenia problémov strojom pomocou učenia sa z dát. Dá sa povedať že SC je už etablovaný ako vedná disciplína, ktorá sa stala neoddeliteľným komplementom aj spomenutej teórie pravdepodobnosti a mnohých prípadoch ju úspešne dopĺňa a nahradzuje. Kniha sa zaoberá technikami štatistického a strojového učenia predikčných modelov s aplikáciami v ekonomike. Uvádzajú sa najnovšie poznatky v danej oblasti s reálnymi aplikáciami. Cieľom publikácie je priblížiť a zoznámiť študentov a odbornú verejnosť s týmito novými inovačnými smermi a rozširovať tieto metódy do praxe

Publikácia je rozdelená na dve časti. Prvá časť – kapitoly 1 až 5 – sa venuje nevyhnutnému teoretickému základu pravdepodobnostných výpočtových metód rozvíjaných v teórii štatistiky a ekonometrie, ako sú základy tvorby a diagnostikovania modelov založených na regresnej analýze s doplnením o problematiku modelovania založenej na Kalmanovej filtrácii, Box-Jenkinsovej metodológii a modelovania autoregresných procesov s podmienenou heteroskedasicitou pre analýzu vzťahov medzi ekonomickými veličinami. Predpokladá sa, že čitatelia majú základné znalosti z teórie pravdepodobnosti a opisnej štatistiky. Text knihy bol písaný tak, aby pokiaľ možno obsahoval len nevyhnutné teoretické definície a ich dôkazy. Poskytujú sa však odkazy, kde k nim je možné nájsť viacej technických a teoretických detailov.

Druhá časť – kapitoly 6 až 11 sa zaoberá možnosťami konštrukcie modelov ekonomických procesov metódami SC a umelej inteligencie. Sústreďuje sa na modelovanie a predikciu ekonomických procesov prostriedkami umelých neurónových sietí a ich učeniu, a špeciálne strojovému učeniu metódou podporných vektorov (Support Vector Machine – SVM). V tejto časti sa porovnávajú ich výsledky na reálnych aplikáciách s výsledkami získanými metódami klasických pravdepodobnostných výpočtov a ekonometrických modelov z prvej časti knihy.

Text je spravidla tvorený tak, že určitá ucelená problematika obsahuje riešené príklady, na ktorom sa ilustruje teória. správnosti pochopenia danej problematiky.

Kapitola prvá obsahuje úvodný pohľad na analýzu a modelovanie ekonomických časových radov založenom na regresnej analýze. Zaoberá sa overovaním správnosti modelu a vyhodnocovaním aproximačnej a predikčnej schopnosti modelov.

Kapitola druhá je venovaná Box-Jenkinsovej metodológii analýzy dát – triede ARIMA modelov. Prezentuje metódy transformácie dát na stacionárne časové rady, teoretický základ ARMA procesov a vývojové stupne ARMA modelovania ekonomických veličín a procesov zaťažených sezónnosťou.

Kapitola tretia sa zaoberá triedou ARCH modelov ako alternatívy k modelom, v ktorých rozptyl je premenlivý v čase. Sú v nej uvedené teoretické základy modelov jednak klasických a taktiež novších modelov, ktoré boli uvedené pre modelovanie procesov na finančných trhoch s vysokofrekvenčnými dátami. Vyčerpávajúcim spôsobom sa poskytuje postup vývoja týchto modelov počnúc metódami kvantifikácie a testovania správnosti špecifikácie až do ich prognostických aplikácií.

Kapitola štvrtá je venovaná základom kointegračnej analýzy ekonomických procesov, dynamickému modelovaniu v ekonomike a modelom opravy chybou. Analyzuje sa koncept rovnováhy, integrovaných premenných, koncept kointegrácie a ich význam pri konštrukcii modelov.

Kapitola piata sa zaberá modelovaním štrukturálnych stavových veličín Kalmanovou filtráciou s inicializáciou počiatočných hodnôt pre Kalmanove rekurzie, postupom vývoja modelov štrukturálnych stavových veličín s aplikačnými príkladmi.

V druhej časti knihy kapitola šiesta je úvodným pohľadom na SC koncept a podstatu SC výpočtov prostriedkami umelých neurónových sietí a strojového učenia.

V kapitole siedmej sa opisuje teória spracovania signálov na báze perceptróna a viacvrstvovej neurónovej siete s dopredným šírením signálov vrátane učiacich algoritmov pre modelovanie ekonomických procesov.

Kapitola ôsma sa zaoberá konštrukciou a teóriou rekurentných sietí a ich učiacimi algoritmami a ich využitie v praxi.

Deviata kapitola je venovaná RBF (Radial Basic Function) sieťam. V nej sa analyzujú špecifiká učenia RBF sietí. Poskytujú sa reálne aplikácie s porovnaním modelovania pomocou RBF sietí oproti modelom založeným na najnovších štatistických (ekonometrických) metódach.

Desiata kapitola sa zaoberá neurónovými sieťami s nesupervizovaným učení a typickými úlohami pre ich aplikácie.

Jedenásta kapitola sa podrobne zaoberá teóriou strojového učenia. Na jej metódach sa vyvíjajú prognostické modely a zhodnocujú sa jej prednosti a obmedzenia.

Kniha bola napísaná za podpory štatistického, matematického a ekonometrického softvéru, pomocou ktorého boli tlačené aj niektoré obrázky a za podpory doktorandov a diplomantov autora.

Dušan Marček, Ostrava, jún 2013

Obsah

Predhovor ..... **V**

Obsah ..... **IX**

Podrobný obsah ..... **XIII**

Kapitola 1 Modelovanie závislostí s použitím regresnej analýzy **1**

1.1 Regresný model s jednou vysvetľujúcou premennou a viacnásobný lineárny regresný model **1**

1.2 Odhad parametrov regresného modelu **4**

1.3 Testy a intervaly spoľahlivosti **6**

1.4 Testy predpokladov aplikácie odhadovej metódy najmenších štvorcov **10**

1.5 Zovšeobecnená metóda najmenších štvorcov **13**

1.6 Konštrukcia prognóz a predikčné intervaly **15**

1.7 Hodnotenie presnosti predpovedí **22**

Kapitola 2 Box-Jenkinsova analýza dát, trieda ARIMA procesov a ich modelovanie **25**

2.1 ARIMA modely **26**

2.2 Autoregresné procesy **31**

2.3 Procesy kĺzavých priemerov **35**

2.4 Zmiešané ARMA(p, q) procesy **38**

2.5 Invertibilita procesov **40**

2.6 Nestacionárne procesy **42**

2.7 Sezónne procesy **44**

2.8 Modelovanie časových radov **46**

2.9 Súhrnné príklady **61**

Kapitola 3 Modelovanie ARCH/GARCH procesov ... 65

3.1 Jednopremenné ARCH modely ... 67

3.2 Odhad parametrov ... 73

3.3 Testovanie heteroskedasticity ... 74

3.4 Diagnostická kontrola ... 75

3.5 Aplikačné príklady ... 77

Kapitola 4 Dynamické modelovanie a kointegračná analýza ... 87

4.1 Dynamické modelovanie vzťahov ekonomických veličín a kointegrácia ... 88

4.2 Testovanie integrovanosti veličín ... 92

4.3 Testovanie kointegrácie ... 94

4.4 Odhad parametrov Engleovou-Grangerovou dvojstupňovou metódou ... 96

4.5 Kointegrácia v systéme viac ako dvoch premenných ... 96

4.6 Využitie kointegračných modelov v prognózovaní ... 97

4.7 Aplikačný príklad ... 98

Kapitola 5 Stavová forma štrukturálnych modelov a Kalmanove rekurzívne procedúry ... 103

5.1 Všeobecná reprezentácia Kalmanovho filtra ... 103

5.2 Kalmanove rekurzie ... 108

5.3 Počiatočné hodnoty pre Kalmanovu filtráciu ... 109

5.4 Odhad rozptylov štrukturálneho modelu a predikcia ... 111

5.5 Vývoj modelovania časových radov štrukturálnymi modelmi – aplikačné príklady na sezónne procesy ... 112

Kapitola 6 Úvod do umelých neurónových sietí a strojového učenia ... 125

6.1 Funkčný model a model spracovania informácií neurónovej bunky ... 126

6.2 Matematický model umelého neurónu a neurónovej siete ... 127

6.3 Učiace schémy ... 134

6.4 Model strojového učenia ... 137

6.5 Zhrnutie ... 138

Kapitola 7 UNS na báze perceptrónu ... 141

7.1 Perceptrón ... 141

7.2 Viacperceptrónový systém ... 142

7.3 Učenie perceptrónu – *delta pravidlo* ... 144

7.4 Viacvrstvové siete – učiaci algoritmus ... 146

Kapitola 8 Rekurentné neurónové siete ... 159

8.1 Bloková reprezentácia neurónových sietí ... 161

8.2 Učenie rekurentnej siete ... 164

8.3 Učenie rekurentnej siete v reálnom čase ... 167

8.4 Stavová reprezentácia a učenie rekurentných sietí Kalmanovými rekurziami ... 172

8.5 Boltzmannov stroj ... 172

8.6 Niektoré ďalšie architektúry rekurentných neurónových sieti ... 174

8.7 Zhrnutie ... 176

Kapitola 9 RBF siete ... 177

9.1 Architektúra RBF sietí ... 177

9.2 Adaptácia parametrov RBF siete ... 180

9.3 Problém voľby počtu a odhadu parametrov RB funkcií a dvojfázová učiaca metóda ... 187

9.4 Soft a granulárna RBF sieť ... 190

Kapitola 10 Nesupervizované učenie ... 199

10.1 Hebbovo učenie ... 200

10.2 Neurónová priamoväzbová sieť na extrakciu hlavných komponentov ... 202

10.3 Kompetitívne (konkurenčné) učenie ... 203

10.4 Samoorganizujúce mapy – SOM ... 208

10.5 Učiaci algoritmus SOM siete (Kohonenov algoritmus) ... 211

10.6 Kvantovanie vektorov učením ... 213
10.7 Adaptívna rezonančná teória – ART siete ... 216
10.8 Siete s hybridnými učiacimi schémami – sieť typu *counterpropagation* ... 219
10.9 Zhrnutie ... 222
Kapitola 11 SVM (SupportVectorMachine) ... 225
11.1 SVM ako lineárny klasifikátor ... 226
11.2 SVM ako nelineárny klasifikátor a aproximátor ... 233
11.3 SV regresia ... 237
11.4 Aplikačné príklady ... 241
Kapitola 12 Záver ... 253
Prílohy ... 255
Literatúra ... 259
Zoznam tabuliek ... 269
Zoznam obrázkov ... 271
Register ... 277
Summary ... 281

Podrobný obsah

Predhovor ..... **V**

Obsah ..... **IX**

Podrobný obsah ..... **XIII**

Kapitola 1 Modelovanie závislostí s použitím regresnej analýzy **1**
1.1 Regresný model s jednou vysvetľujúcou premennou a viacnásobný lineárny regresný model 1
1.2 Odhad parametrov regresného modelu 4
1.3 Testy a intervaly spoľahlivosti 6
1.4 Testy predpokladov aplikácie odhadovej metódy najmenších štvorcov 10
1.5 Zovšeobecnená metóda najmenších štvorcov 13
1.6 Konštrukcia prognóz a predikčné intervaly 15
1.7 Hodnotenie presnosti predpovedí 22

Kapitola 2 Box-Jenkinsova analýza dát, trieda ARIMA procesov a ich modelovanie **25**
2.1 ARIMA modely 26
2.2 Autoregresné procesy 31
2.3 Procesy kĺzavých priemerov 35
2.4 Zmiešané ARMA(p, q) procesy 38
2.5 Invertibilita procesov 40
2.6 Nestacionárne procesy 42
2.7 Sezónne procesy 44
2.8 Modelovanie časových radov 46
2.8.1 Identifikácia 47
2.8.2 Odhad parametrov 50
2.8.3 Diagnostická kontrola 54
2.8.4 Konštrukcia predpovedí 56
2.9 Súhrnné príklady 61

Kapitola 3 Modelovanie ARCH/GARCH procesov **65**
3.1 Jednopremenné ARCH modely 67
3.1.1 ARCH(p) proces 69
3.1.2 GARCH(p, q) proces 70
3.1.3 ARCH-GARCH regresné modely 71
3.1.4 Ďalšie typy ARCH/GARCH modelov 72
3.2 Odhad parametrov 73
3.3 Testovanie heteroskedasticity 74

3.4 Diagnostická kontrola ... 75
3.5 Aplikačné príklady ... 77

Kapitola 4 Dynamické modelovanie a kointegračná analýza ... 87
4.1 Dynamické modelovanie vzťahov ekonomických veličín a kointegrácia ... 88
4.2 Testovanie integrovanosti veličín ... 92
4.3 Testovanie kointegrácie ... 94
4.4 Odhad parametrov Engleovou-Grangerovou dvojstupňovou metódou ... 96
4.5 Kointegrácia v systéme viac ako dvoch premenných ... 96
4.6 Využitie kointegračných modelov v prognózovaní ... 97
4.7 Aplikačný príklad ... 98

Kapitola 5 Stavová forma štrukturálnych modelov a Kalmanove rekurzívne procedúry ... 103
5.1 Všeobecná reprezentácia Kalmanovho filtra ... 103
5.2 Kalmanove rekurzie ... 108
5.3 Počiatočné hodnoty pre Kalmanovu filtráciu ... 109
5.4 Odhad rozptylov štrukturálneho modelu a predikcia ... 111
5.5 Vývoj modelovania časových radov štrukturálnymi modelmi – aplikačné príklady na sezónne procesy ... 112

Kapitola 6 Úvod do umelých neurónových sietí a strojového učenia ... 125
6.1 Funkčný model a model spracovania informácií neurónovej bunky ... 126
6.2 Matematický model umelého neurónu a neurónovej siete ... 127
6.2.1 Matematický model jednoduchého neurónu ... 129
6.2.2 Viacvrstvové neurónové siete ... 132
6.3 Učiace schémy ... 134
6.4 Model strojového učenia ... 137
6.5 Zhrnutie ... 138

Kapitola 7 UNS na báze perceptrónu ... 141
7.1 Perceptrón ... 141
7.2 Viacperceptrónový systém ... 142
7.3 Učenie perceptrónu – *delta pravidlo* ... 144
7.4 Viacvrstvové siete – učiaci algoritmus ... 146
7.4.1 Adaptácia váh siete s lineárnou aktivačnou funkciou neurónu ... 146
7.4.2 Adaptácia váh siete s nelineárnou aktivačnou funkciou neurónu ... 147
7.4.3 Adaptácia váh v sieti s jednou skrytou vrstvou, *Back-Propagation* algoritmus ... 148
7.4.4 *Back-Propagation* algoritmus vo viacvrstvových sieťach ... 150
7.4.5 Prognózovanie ekonomických časových radov – aplikačný príklad ... 153

Kapitola 8 Rekurentné neurónové siete 159
8.1 Bloková reprezentácia neurónových sietí 161
8.2 Učenie rekurentnej siete 164
8.3 Učenie rekurentnej siete v reálnom čase 167
8.4 Stavová reprezentácia a učenie rekurentných sietí Kalmanovými rekurziami 172
8.5 Boltzmannov stroj 172
8.6 Niektoré ďalšie architektúry rekurentných neurónových sieti 174
8.7 Zhrnutie 176

Kapitola 9 RBF siete 177
9.1 Architektúra RBF sietí 177
9.2 Adaptácia parametrov RBF siete 180
9.3 Problém voľby počtu a odhadu parametrov RB funkcií a dvojfázová učiaca metóda 187
9.4 Soft a granulárna RBF sieť 190

Kapitola 10 Nesupervizované učenie 199
10.1 Hebbovo učenie 200
10.2 Neurónová priamoväzbová sieť na extrakciu hlavných komponentov 202
10.3 Kompetitívne (konkurenčné) učenie 203
10.3.1 Učenie kompetitívnej siete 204
10.3.2 Modifikácia učenia kompetitívnej siete 207
10.4 Samoorganizujúce mapy – SOM 208
10.5 Učiaci algoritmus SOM siete (Kohonenov algoritmus) 211
10.6 Kvantovanie vektorov učením 213
10.7 Adaptívna rezonančná teória – ART siete 216
10.8 Siete s hybridnými učiacimi schémami – sieť typu *counter-propagation* 219
10.9 Zhrnutie 222

Kapitola 11 SVM (SupportVectorMachine) 225
11.1 SVM ako lineárny klasifikátor 226
11.2 SVM ako nelineárny klasifikátor a aproximátor 233
11.3 SV regresia 237
11.4 Aplikačné príklady 241

Kapitola 12 Záver 253

Prílohy 255

Literatúra 259

Zoznam tabuliek 269

Zoznam obrázkov 271

Register 277

Summary 281

Kapitola 1

Modelovanie závislostí s použitím regresnej analýzy

V konkrétnom hospodárskom procese je sledovaná viac alebo menej zložitá sústava ukazovateľov. Medzi jednotlivými ukazovateľmi existujú vzájomné známe alebo menej preskúmané vzťahy, ktoré majú svoj vecný ekonomický alebo technicko-ekonomický obsah. Ak existujú vzťahy medzi ekonomickými veličinami, prejavia sa v časových radoch ukazovateľov, ktoré ich vyjadrujú. Ak existuje hypotéza alebo teória o týchto vzťahoch, môžeme tieto vzťahy opísať pomocou modelu, ktorý zjednodušene zobrazuje skutočnú realitu skúmaného hospodárskeho procesu. Zmyslom tohto a ďalších častí je uviesť základné techniky, postupy a metódy, ktoré sa používajú pri modelovaní vzťahov medzi ekonomickými veličinami zobrazenými regresnými modelmi s následnou kvantifikáciou a verifikáciou modelov. Ide tu prakticky o ukázanie matematickej formulácie modelu, jeho konfrontovanie s realitou, resp. ekonomickou teóriou.

1.1 Regresný model s jednou vysvetľujúcou premennou a viacnásobný lineárny regresný model

Najjednoduchším regresným modelom je model s jednou nezávislou veličinou, pomocou ktorého sa určuje vzťah medzi závislou veličinou y_t a nejakou nezávislou veličinou x_t. Ak budeme predpokladať, že hodnoty závislej a nezávislej veličiny sú časovými radmi, t. j. časovo usporiadaná postupnosť pozorovaní (realizácií) veličín a že závislosť medzi hodnotami týchto dvoch veličín je lineárna, vzťah medzi pozorovanými hodnotami časových radov $\{y_t\}$ a $\{x_t\}$ môžeme vyjadriť modelom

$$y_t = b_0 + b_1 x_t + u_t, \text{ pre } t = 1, 2, ..., N, \tag{1.1}$$

kde u_t je náhodný člen modelu, o ktorom predpokladáme, že má normálne rozdelenie s nulovou strednou hodnotou a konštantným rozptylom. Symbolmi b_0 a b_1 sú označené neznáme parametre modelu. Model (1.1) je jednoduchý lineárny regresný model časových radov $\{y_t\}$ a $\{x_t\}$.

Jednoduchý lineárny regresný model (1.1) pre jednu nezávislú veličinu sa dá jednoducho rozšíriť pre viac nezávislých veličín (viacnásobný regresný model) v tvare

$$y_t = b_0 + b_1 x_{t1} + b_2 x_{t2} + \ \ldots \ + b_k x_{tk} + u_t. \tag{1.2}$$

Pomocou tohto modelu explicitne sa vyjadruje lineárny vzťah jednej endogénnej (vysvetľovanej) veličiny s viacerými vysvetľujúcimi veličinami a s jednou náhodnou zložkou modelu, kde b_0 je neznámy absolútny člen regresie, b_i pre i = 1,2, ..., k sú neznáme skutočné parametre modelu, x_{ti} pre i = 1,2, ..., k sú pozorované hodnoty vysvetľujúcich (nezávislých) veličín, u_t je náhodná zložka modelu.

V modeli (1.2) sú zmeny závislej veličiny vysvetľované jednak zmenami vysvetľujúcich veličín a taktiež zmenami náhodnej zložky. Model obsahuje k + 1 parametrov b_0, b_1, ..., b_k.

Rozdiel medzi počtom pozorovaní N a počtom parametrov modelu sa nazýva počet stupňov voľnosti modelu, pričom musí platiť $N > k + 1$. Príkladom lineárneho modelu môže byť závislosť maloobchodného obratu potrieb určitých výrobkov (Y_t) od disponibilných peňažných príjmov obyvateľstva (X_t) a cien týchto výrobkov (P_t) v tvare

$$Y_t = b_0 + b_1 X_t + b_2 P_t + u_t. \tag{1.3}$$

Ide o model viacnásobnej regresie, pomocou ktorého zmeny závislej veličiny Y_t sú vysvetľované pomocou dvoch nezávislých veličín – X_t, P_t. Vo všeobecnom zápise jednorovnicového modelu (1.2) sú vysvetľujúce veličiny označené symbolmi x_{t1}, x_{t2}, ..., x_{tk}. V zápisoch konkrétnych modelov sú tieto označenia premenných zamenené za konkrétne označenia veličín tak, ako je zaužívané ich označenie v ekonomickej teórii a praxi. Analogicky to platí i pre závislú veličinu y_t. V našom príklade je označenie y_t zamenené za Y_t, x_{t1} za X_t, x_{t2} za P_t. Hodnoty vysvetľujúcich a závislej veličiny pre t = 1, 2 ,..., N môžeme získať napr. pozorovaním v minulých obdobiach a sú známe. Vysvetľujúce veličiny modelu (1.2) nemusia byť všetky z toho istého obdobia. Môžu byť časovo posunuté (oneskorené alebo predbiehajúce). Je možné, že lepšie než príjem X_t v modeli (1.3) vysvetlí maloobchodný obrat Y_t príjem z predchádzajúceho obdobia X_{t-1}.

Vplyvy nepodstatných faktorov, ktoré neboli explicitne zahrnuté do modelu vo forme ďalších premenných alebo vplyvy nesprávnej (nepresnej) špecifikácie modelu sú zohľadnené prostredníctvom náhodnej zložky modelu u_t. Hodnoty náhodnej zložky modelu nie je možné získať pozorovaním ako hodnoty ostatných veličín modelu.

Zápis modelu (1.2) možno vyjadriť aj v tvare

$$y_t = \sum_{i=0}^{k} b_i x_{ti} + u_t, \tag{1.4}$$

pričom v tomto zápise x_{t0} má v každom časovom okamihu t = 1,2,..., n hodnotu 1, t. j.

$$x_{t0} = 1. \tag{1.5}$$

Zápisy modelov (1.2) a (1.3) je možné, s prihliadnutím na (1.5), jednoduchšie vyjadriť v maticovom zápise ako

$$\mathbf{y} = \mathbf{Xb} + \mathbf{u} \tag{1.6}$$

alebo s rozpísaním prvkov matíc a vektorov v (1.6), t. j. ako

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{X} & \cdot & \mathbf{b} & + & \mathbf{u} \end{array}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & .. & x_{1k} \\ 1 & x_{21} & x_{22} & .. & x_{2k} \\ .. & .. & .. & .. & .. \\ 1 & x_{n1} & x_{n2} & .. & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ .. \\ b_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ .. \\ u_n \end{pmatrix}, \tag{1.7}$$

kde **y** je stĺpcový vektor pozorovaných hodnôt vysvetľovanej veličiny rozmeru N, **X** je matica pozorovaných hodnôt všetkých vysvetľujúcich veličín rozmeru $N \cdot (k+1)$. Prvý stĺpec matice obsahuje samé jednotky, s ohľadom na (1.5), **b** je stĺpcový vektor skutočných parametrov modelu rozmeru k + 1, **u** je stĺpcový vektor hodnôt náhodnej zložky rozmeru N.

Zo vzťahu (1.6) je vidieť, že parameter b_0 je pokladaný za koeficient pre premennú x_{0t}, o ktorej sa predpokladá že má vo všetkých pozorovaniach hodnoty rovné 1.

Parametre modelu b_0, b_1, ..., b_k v rovnici (1.2) sú neznámymi parametrami. Ich hodnoty sa odhadujú štatisticky na základe niektorej odhadovej metódy z výberu dát premenných veličín.

Pokiaľ ide o ekonomickú interpretáciu parametrov **b**, možno povedať, že až na úrovňovú konštantu, je ich možné priamo interpretovať ako koeficienty absolútnej elasticity. Ak sa napr. v modeli (1.3) nezmení úroveň cien výrobkov (P_t), vykazuje zvýšenie príjmov obyvateľstva napr. o 1 mld. Sk tendenciu zvýšiť maloobchodný obrat daných výrobkov (y_t) o b_1. Ak sa na druhej strane pri nezmenených príjmoch (X_t) zvýši cenová hladina o 1 jednotku, zníži sa maloobchodný obrat o b_2 jednotiek. Znamienko pri parametri b_2 bude záporné.

1.2 Odhad parametrov regresného modelu

Odhad parametrov regresného modelu na základe empirických pozorovaní spadá do problematiky štatistiky a ekonometrie. Najčastejšie používaným odhadovým kritériom je minimum súčtu štvorcov odchýlok, určených ako rozdiel medzi pozorovanými hodnotami vysvetľovanej veličiny a jej vypočítanými hodnotami.

Označme vektor vypočítaných hodnôt parametrov rovnice (1.6) ako $\hat{\mathbf{b}}$, ďalej vektor vypočítaných (teoretických) hodnôt vysvetľovanej premennej ako $\hat{\mathbf{y}}$. Potom môžeme teoretické hodnoty vysvetľovanej premennej určiť ako

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} \tag{1.8}$$

a pozorované hodnoty vysvetľovanej premennej ako

$$\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + \mathbf{e}, \tag{1.9}$$

kde **e** je vektor hodnôt reziduálnych odchýlok, vypočítaných z tohto vzťahu ako

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{y} - \hat{\mathbf{y}}. \tag{1.10}$$

Odhad parametrov modelu (1.2) metódou najmenších štvorcov (v literatúre je táto metóda označovaná anglickými iniciálami OLS – Ordinary Least Squares) vychádza z nasledovných predpokladov o náhodnej zložke:

1. Náhodná zložka modelu má normálne rozdelenie a má v každom pozorovaní nulovú strednú hodnotu, takže platí

   $$E(\mathbf{u}) = \mathbf{0} \text{ alebo} \qquad E(u_t) = 0 \text{ pre } t = 1, 2, ..., N, \tag{1.11}$$

   kde **0** je stĺpcový vektor dimenzie N so všetkými nulovými prvkami. Keďže predpokladáme, že náhodná zložka má v každom pozorovaní nulovú strednú hodnotu, potom aj očakávané hodnoty vysvetľovanej premennej, v súlade s výrazmi (1.6) a (1.8), sú $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$.
2. Rozptyl náhodnej zložky modelu je konštantný, je rovnaký v každom pozorovaní, t. j.

   $$E\left(u_t^2\right) = \sigma^2, \text{ pre } t = 1, 2, ..., N, \tag{1.12}$$

   lebo rozptyl náhodnej veličiny sa rovná priemeru jej štvorca zmenšeného o štvorec priemeru, t. j. $E(u_t^2) = E(u_t^2) - \left[E(u_t)\right]^2$. Predpoklad (1.12) je známy ako predpoklad homoskedasticity. Ak predpoklad neplatí hovoríme, že náhodná zložka je heteroskedastická alebo aj model je heteroskedastický. Rozptyl náhodnej zložky, ktorý je konštantný v každom pozorovaní nepoznáme, a preto musí sa taktiež odhadnúť. Niekedy sa hovorí v tejto súvislosti, že lineárny regresný model (1.2) má k+1 neznámych parametrov a neznámy rozptyl náhodnej zložky. Takže model má potom celkom k+2 neznámych parametrov.
3. Náhodné zložky sú štatisticky vzájomne nezávisle. Hodnoty náhodnej zložky z nerovnakých období sú ortogonálne (nie sú vzájomne korelované), t. j. majú nulové kovariancie, čo môžeme zapísať

$$E\left(u_i u_j\right) = 0, \text{ pre } j \neq i. \tag{1.13}$$

Ak by predpoklad (1.13) neplatil, hodnoty náhodnej zložky v jednotlivých pozorovaniach boli by vzájomne korelované (autokorelované), teda vykazovali by určitú pravidelnosť a teda aj určité systematické správanie. Vtedy hovoríme o sériovej korelácii alebo autokorelácii. Predpoklad konštantnosti rozptylu a predpoklad o nekorelovanosti hodnôt náhodnej zložky sa vyjadriť variačno-kovariančnou maticou Σ v tvare

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma^2 \end{pmatrix}. \tag{1.14}$$

Variačno-kovariančná matica náhodnej zložky modelu Σ spĺňa predpoklad (1.12) o konštantnosti rozptylov vo všetkých jej pozorovaniach rovných σ^2 a predpoklad (1.13) o vzájomnej nekorelovanosti hodnôt náhodnej zložky modelu, lebo má všetky nediagonálne prvky nulové.

Ak sú splnené uvedené predpoklady (1.11) až (1.13) je možné**,** na základe výberu hodnôt pozorovaní vysvetľovanej a všetkých vysvetľujúcich veličín odhadnúť metódou najmenších štvorcov parametre b_0, b_1, ..., b_k lineárneho regresného modelu (1.2).

Odhadový výraz OLS metódy minimalizuje súčet štvorcov hodnôt reziduálnych odchýlok

$$\ell\left(\hat{b}_0, \hat{b}_1, \ ..., \hat{b}_k\right) = \mathbf{e}^2 = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^2 = \sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} \left[y_t - \left(\hat{b}_0 + \sum_{i=1}^{k} \hat{b}_i x_{ti} \right) \right]^2 . \tag{1.15}$$

Parciálnymi deriváciami výrazu (1.15) podľa parametrov, t. j. $\partial \ell / \partial \hat{b}_i$, i = 0, 1, ..., k sa získa k + 1 normálnych rovníc, z ktorých možno určiť hodnoty parametrov $\hat{b}_0$, $\hat{b}_1$, ..., $\hat{b}_k$. Odhadový výraz OLS metódy pre parametre regresnej rovnice (1.2) je (Garaj a Šujan, 1980)

$$\hat{\mathbf{b}} = \left(\mathrm{X'X}\right)^{-1} \mathbf{X'y}. \tag{1.16}$$

Vektor $\hat{\mathbf{b}}$ je stĺpcový vektor, ktorého prvky sú hľadané odhady parametrov $\hat{b}_0$, $\hat{b}_1$, ..., $\hat{b}_k$. Prvky matice **X** a vektor **y** sú určené výrazom (1.7). Matica **X'** je transponovaná matica **X**.

Bodové odhady parametrov $\hat{\mathbf{b}}$ regresného modelu (1.2) podľa výrazu (1.16) sú náhodnými veličinami, lebo **y** je náhodnou veličinou. Pokiaľ ide o jeho štatistické vlastností (nevychýlenosť, výdatnosť), dá sa dokázať (Garaj a Šujan, 1980), že estimátor (1.16) je nevychýleným estimátorom, t. j.

$$E\left(\hat{\mathbf{b}}\right) = \mathbf{b}, \text{ resp. } E\left(\hat{b}_j\right) = b_j \text{ pre } j = 0,1, ..., k. \tag{1.17}$$

Ak sú splnené predpoklady (1.11) až (1.13) o náhodnej zložke regresného modelu, odhadová funkcia parametrov $\hat{\mathbf{b}}$ získaná OLS metódou je výdatná s rozptylmi parametrov

$$\sigma^2_{\hat{b}j} = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj} \text{ pre } j = 0, 1, ..., k, \tag{1.18}$$

kde index *jj* pri matici $(\mathbf{X}'\mathbf{X})^{-1}$ vyjadruje príslušný prvok tejto matice.

Odhadová funkcia (1.18) pre rozptyl parametrov modelu má najmenší výberový rozptyl zo všetkých lineárnych nevychýlených estimátorov vektora **b**. Odhad vektora parametrov **b** (odhadov najmenších štvorcov koeficientov regresnej rovnice) modelu (1.6) je najlepším lineárnym nevychýleným odhadom (NLNO). Smerodajné odchýlky vektora parametrov sú interpretované ako chyby odhadu parametrov. Vypočítajú sa ako

$$\sigma_{\hat{b}j} = \sigma \sqrt{\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}} \text{ pre } j = 0, 1, ..., k. \tag{1.19}$$

Vo výraze (1.18) pre odhad rozptylu $\sigma^2_{\hat{b}j}$ vektora parametrov $\hat{\mathbf{b}}$ resp. výraz pre odhad ich smerodajných odchýlok (1.19), vystupuje rozptyl náhodnej zložky modelu σ^2, ktorý je obvykle neznámy. Rozptyl σ^2 náhodnej zložky modelu môžeme odhadnúť. Označme s^2 odhad rozptylu náhodnej zložky σ^2. Potom odhadový výraz pre rozptyl náhodnej zložky s^2 modelu je

$$s^2 = \frac{\sum_{t=1}^{N} e_t^2}{N-(k+1)} = \frac{\mathbf{e}'\mathbf{e}}{N-(k+1)} \tag{1.20}$$

a odhadový výraz pre smerodajnú odchýlku odhadov parametrov modelu je

$$s_{\hat{b}j} = s \sqrt{\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}} \text{ pre } j = 0, 1, ..., k. \tag{1.21}$$

Odhad rozptylu náhodnej zložky podľa (1.20) je nevychýleným odhadom. Odhad smerodajnej odchýlky (odhad výberovej chyby) náhodnej zložky $s = \sqrt{s^2}$ je charakteristikou presnosti modelu.

1.3 Testy a intervaly spoľahlivosti

Dôležitými informáciami pri posudzovaní kvality odhadnutých parametrov regresného modelu je test na štatistickú verifikáciu modelu ako celku a testy významnosti jednotlivých parametrov modelu.

Miera, na základe ktorej sa posudzuje presnosť či vhodnosť špecifikovaného modelu k dátam je koeficient determinácie, označovaný ako R^2. Koeficient determinácie je voľne interpretovaný ako miera vysvetlenia premenlivosti dát vysvetľovanej premennej regresným modelom. Je to miera celkovej variability vysvetľovanej veličiny vzhľadom na jej variabilitu vysvetľovanú pomocou

modelu. Čím lepšie sú regresným modelom aproximované dáta vysvetľovanej premennej, tým viac je R^2 bližšie k hodnote 1. Ak R^2 je napr. 0,7, môžeme povedať, že 70 % rozptylu y_t je vysvetlené modelom.

Mierou variability veličiny y_t je jej rozptyl. Ten je úmerný sume štvorcov odchýlok pozorovaní premennej y_t od jej aritmetického priemeru. Zdrojom celkovej variability sú pozorovania y_t, t. j. dáta. Označíme ju symbolom *C*. Modelom vysvetlenú variabilitu premennej y_t označíme písmenom *V*. Zdrojom vysvetlenej variability je regresný model. Vysvetlená variabilita a teda aj rozptyl je úmerný sume rozdielu štvorcov modelom vypočítaných hodnôt $\hat{y}_t$ od aritmetického priemeru radu $\{y_t\}$. Zdrojom nevysvetleného zvyšku variability sú rezíduá e_t. Rozklad celkovej variability *C* a modelom vysvetlenej variability *V* je prehľadne uvedený v tabuľke 1–1.

Tabuľka 1–1 Rozklad rozptylu pre testovanie významnosti regresného modelu

Suma štvorcov odchýlok	Definícia	Označenie	Zdroj variability	Počet stupňov voľnosti
Celková	$\Sigma(y_t-\overline{y})^2$	*C*	$\{y_t\}$	*N*–1
Vysvetlená	$\Sigma(\hat{y}_t-\overline{y})^2$	*V*	parametre modelu	*K*
Nevysvetlená	$\Sigma(y_t-\hat{y}_t)^2$	N	$\{e_t\}$	*N*–(*k*+1)

Hodnota koeficientu determinácie je určená vzťahom

$$R^2=\frac{V}{C}=\frac{C-\mathrm{N}}{C}=1-\frac{\mathrm{N}}{C}= \\ =\frac{\sum_{t=1}^{N}\left(\hat{y}_t-\overline{y}\right)^2}{\sum_{t=1}^{N}\left(y_t-\overline{y}\right)^2}=1-\frac{\sum_{t=1}^{N}\left(y_t-\hat{y}_t\right)^2}{\sum_{t=1}^{N}\left(y_t-\overline{y}\right)^2}=1-\frac{\sum_{t=1}^{N}e_t^2}{\sum_{t=1}^{N}\left(y_t-\overline{y}\right)^2}. \tag{1.22}$$

Z tabuľky 1–1 je vidieť, že modelom vysvetlená variabilita je monotónne neklesajúca funkcia počtu vysvetľujúcich veličín, a teda aj funkcia počtu parametrov. S rastúcim počtom vysvetľujúcich premenných, nebude sa hodnota koeficienta determinácie znižovať. V snahe dosiahnuť čo najvyššiu vysvetľujúcu schopnosť modelu, niekto by mohol do modelu zaraďovať ekonomickou teóriou alebo inou hypotézou nepodložené *vysvetľujúce* veličiny. Z tohto dôvodu sa zaviedol korigovaný koeficient determinácie, čím sa znevýhodňuje model za také vysvetľujúce veličiny, ktoré význačne alebo vôbec neprispievajú k vysvetleniu variability endogénnej veličiny. Takto korigovaný koeficient determinácie je označovaný ako R^2_{adj}. Jeho hodnotu určíme tak, že vo výraze pre koeficient determinácie

$$R^2 = 1 - \frac{\mathrm{N}}{C} = 1 - \left(1 - R^2\right)$$

podelíme jednotlivé súčty štvorcov N a C príslušnými stupňami voľnosti, t. j.

$$R^2_{adj} = 1 - \frac{N-1}{N-\left(k+1\right)}\left(1 - R^2\right).$$

Hodnota korigovaného koeficienta determinácie spravidla so zvyšujúcim počtom vysvetľovaných veličín spočiatku rastie, prechádza cez maximum, a potom klesá. Mnoho analytikov za najlepší model považuje model s maximálnou hodnotou korigovaného koeficienta determinácie.

Test štatistickej významnosti vzťahu medzi premennými. Vzťahmi (1.22) sa určuje výpočet koeficienta determinácie. Štatistická významnosť vzťahu medzi premennými sa testuje nulovou hypotézou, ktorá predpokladá, že všetky jeho parametre sú rovné nule, t. j.

H_0: $b_1 = b_2 = ... = b_k = 0$

oproti alternatíve (1.23)

H_1: aspoň jeden $b_j \neq 0$ pre $j = 1, 2, ..., k$.

Štatistika, ktorou sa overuje platnosť H_0 je $\mathrm{F_R}$, ktorá je určená vzťahom

$$\mathrm{F_R} = \frac{V/k}{N/\left[N-\left(k+1\right)\right]} = \frac{\left(\sum_{t=1}^{N}\left(\hat{y}_t - \bar{y}\right)^2\right)/k}{\left(\sum_{t=1}^{N} e_t^2\right)/\left[N-\left(k+1\right)\right]}. \qquad (1.24)$$

Veličina $\mathrm{F_R}$ má F- rozdelenie so stupňami voľnosti k a $[N– (k + 1)]$. Stupne voľnosti sú určené počtami nezávislých sčítancov v príslušných sumách štvorcov odchýlok v tabuľke 1–1. Zamietnutie nulovej hypotézy H_0 (1.23) a prijatie platnosti alternatívnej hypotézy H_1, t. j. štatistickej významnosti vzťahu medzi premennými, hodnota veličiny $\mathrm{F_R}$ vypočítaná vzťahom (1.24) musí mať väčšiu hodnotu, ako je jej teoretická kritická hodnota na hladine významnosti α a pri daných stupňoch voľnosti, t. j. musí platiť

$$\mathrm{F_R} > F_{\alpha,k,\left[N-(k+1)\right]}, \qquad (1.25)$$

kde $F_{\alpha,k,\left[N-(k+1)\right]}$ je tabuľovaná teoretická kritická hodnota na hladine významnosti α pri stupňoch voľnosti k a $[N – (k + 1)]$.

Aj keď bola potvrdená štatistickej významnosť vzťahu medzi premennými modelu, neznamená to ešte, že model je vhodný pre použitie na konštrukciu prognóz. Predikčnú schopnosť modelu lepšie vystihuje variačný koeficient (*VK*), ktorý sa vypočíta ako pomer modelom nevysvetlenej variability vysvetľovanej premennej k hodnote jej aritmetického priemeru, t. j

$$VK = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2}}{\overline{y}} 100.$$

Podľa Montgomery a kol. (1990) ak hodnota *VK* je menšia ako 20, ide o vhodný model pre prognostickú aplikáciu.

Významnosť jednotlivých parametrov modelu a ich intervaly spoľahlivosti. Významnosť jednotlivých parametrov modelu môže byť testovaná pomocou *t* testu. Následkom predpokladu o normálnom rozdelení náhodnej zložky u_t pomer medzi odhadnutou hodnotou parametra $\hat{b}_j$, j = 0, 1, 2, ..., k a smerodajnou odchýlkou jeho rozdelenia $s_{\hat{b}j} = s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}$ určuje veličinu, ktorá má Studentovo *t*-rozdelenie so stupňami voľnosti $v = N - \left(k+1\right)$. Nulová hypotéza testu významnosti parametrov je formulovaná ako

H_0: $b_j = 0$,

oproti alternatíve

H_1: $b_j \neq 0$.

Ak absolútna hodnota pomeru $t_j = \dfrac{\hat{b}_j}{s_{\hat{b}j}}$ je väčšia, ako je tabuľovaná kritická hodnota $t_{\alpha,(N-k=1)}$ Studentovho rozdelenia, t. j. ak

$$\left|t_j\right| > t_{\alpha,(N-k=1)}, \tag{1.26}$$

kde $t_{\alpha,(N-k=1)}$ je tabuľovaná kritická hodnota Studentovho rozdelenia, potom parameter b_j je štatisticky významný. Ak sa pri testovaní významnosti individuálnych parametrov modelu ukáže, že aspoň jeden z jeho parametrov je štatisticky nevýznamný, treba túto nevýznamnosť zohľadniť aj pri hodnotení modelu ako celku. Parameter b_0 v regresnom modeli pre ekonomickú aplikáciu nemá ekonomickú interpretáciu. Vyjadruje len počiatočnú úroveň vysvetľovanej premennej. Štatistická verifikácia tohto parametra modelu nemá praktický význam a nemá preto vplyv pre celkové hodnotenie významnosti modelu.

Zo vzťahu (1.26) pre parametre b_j regresného modelu môžu byť určené obojstranné intervaly spoľahlivosti. (1–α) 100% interval spoľahlivosti parametra b_j je

$$b_j = \hat{b}_j \pm t_{0.05[N-(k+1)]} \mathrm{S}_{bj}. \tag{1.27}$$

Skutočná hodnota parametra b_j sa bude nachádzať v intervale určeného vzťahom (1.27) s pravdepodobnosťou (1–α).

1.4 Testy predpokladov aplikácie odhadovej metódy najmenších štvorcov

Predpoklady o náhodnej zložke u_t modelu (1.2), ktoré sme definovali vzťahmi (1.11) až (1.13) sú len teoretickými predpokladmi, lebo hodnoty náhodnej zložky nemôžeme merať. Testovanie teoretických predpokladov je založené na skúmaní rezíduí e_t, ktoré po kvantifikácii modelu môžeme určiť ako $e_t = y_t - \hat{y}_t$.

Test náhodnej zložky na normálne rozdelenie. Regresný model predpokladá, že jeho náhodná zložka má normálne rozdelenie s nulovou strednou hodnotou. Posúdenie, či náhodná zložka má normálne rozdelenie, možno vykonať na základe grafického priebehu rezíduí e_t v závislosti od normovaných rezíduí. Normované rezíduá získame tak, že rezíduá e_t podelíme ich smerodajnou odchýlkou. Ak náhodná zložka má normálne rozdelené, priebeh v závislosti rezíduí od normovaných rezíduí musí byť približne priamočiary. Ilustráciu tohto grafického priebehu (tzv. QQ plot) je vidno na obrázku 1–1. v príklade 1.1.

Testovanie na normálne rozdelenie náhodnej zložky je formálne podložené na Jarque-Beraovej testovacej štatistike χ^2, ktorá má χ^2 rozdelenie s 2 stupňami voľnosti. Hodnota Jarque-Beraovej štatistiky je

$$\chi^2 = \frac{N-k}{6}\left[S^2 + \frac{1}{4}(K-3)\right], \tag{1.28}$$

kde N je počet pozorovaní, k je počet vysvetľujúcich premenných, S je šikmosť (štvrtý centrálny moment normovaných rezíduí), K je špicatosť. Nulová hypotéza predpokladá, že náhodné zložky sú normálne rozdelené, alternatívna hypotéza zamieta nulovú hypotézu. Ak vypočítaná veličina χ^2 podľa vzťahu (1.28) je väčšia ako tabuľovaná $\chi^2_{\alpha,2}$ hodnota na hladine významnosti α, nulovú hypotézu zamietneme, náhodné zložky nemajú normálne rozdelenie.

Testovanie autokorelácie. Predpoklad (1.13) o náhodnej zložke modelu vyjadruje, že náhodná zložka regresného modelu je náhodnou veličinou len vtedy, ak jej hodnoty z rôznych pozorovaní nie sú vzájomne závislé. Ak by boli hodnoty náhodnej zložky vzájomne závislé, malo by to za následok, podhodnotenia smerodajných odchýlok parametrov modelu. Viedlo by to k zväčšovaniu hodnôt testovacích veličín na významnosť parametrov a v konečnom dôsledku k mylným optimistickým záverom o dobrej modelovej aproximácii dát časového radu vysvetľovanej premennej. Predovšetkým modely založené na časových radoch často porušujú predpoklad o nekorelovanosti náhodných zložiek modelu.

Ak nie je splnený predpoklad, že kovariancie náhodnej zložky modelu a zvyčajne aj kovariancie rezíduí nie sú nulové, variačno-kovariančná matica Σ podľa (1.14) náhodnej zložky modelu nie je diagonálna (všetky nediagonálne prvky tejto matice nie sú nulové). Hodnoty náhodnej zložky modelu môžu byť v tomto prípade rôznym spôsobom korelované. Dôvodom k tomu môže byť napr. nezahrnutie niektorej premennej pri špecifikácii modelu, časovo oneskorený

efekt prechodových vplyvov, voľba nesprávneho funkčného tvaru modelu, chyby pozorovaní a pod.

V probléme autokorelácie rezíduí sa obvykle riešia dve úlohy. Prvá úloha je spojená so zisťovaním autokorelácie hodnôt náhodnej zložky modelu. V prvej úlohe sa testuje významnosti autokorelácie. Druhá úloha je spojená s odstraňovaním autokorelácie, predovšetkým príčin jej vzniku. Najprv sa budeme zaoberať exaktným zisťovaním (testovaním) výskytu autokorelácie.

Aby sme mohli zistiť výskyt autokorelácie rezíduí, musíme prijať určité predpoklady o charaktere, či závislostí hodnôt náhodnej zložky modelu. Najčastejšie sa predpokladá autoregresný typ závislosti hodnôt náhodnej zložky modelu v tvare

$$u_t = \rho\, u_{t-1} + \varepsilon_t, \text{ pre } t = 2, 3, ..., N \tag{1.29}$$

v ktorom ρ je koeficient autokorelácie náhodných zložiek ε_t je náhodná zložka tejto schémy (modelu). O náhodnej zložke ε_t schémy (1.29) sa predpokladá, že má normálne rozdelenie s nulovou strednou hodnotou, konštantným rozptylom a jednotlivé hodnoty ε_t nie sú korelované.

Pri výskyte autokorelácie rezíduí variačno-kovariančná matica Σ (1.14) náhodnej zložky modelu, ako sme uviedli, nebude mať nediagonálne prvky nulové. Uvedieme výsledne vzťahy pre výpočet rozptylu a kovariancií variačno-kovariančnej matice náhodnej zložky modelu Σ. Postup ich odvodenia je uvedený v Marček, D. a Marček, M. (2001).

V prípade autokorelácie rezíduí náhodnej zložky modelu podľa schémy (1.29) rozptyl náhodnej zložky je daný výrazom

$$\sigma^2 = \frac{1}{1-\rho^2}\sigma_\varepsilon^2, \tag{1.30}$$

kde σ_ε^2 je rozptyl náhodnej zložky autokorelačnej schémy (1.29).

Nediagonálne prvky variačno-kovariančnej matice v prípade autokorelácie náhodnej zložky budú vyjadrovať závislosti medzi hodnotami náhodnej zložky. Táto závislosť je úmerná kovarianciam náhodnej zložky pre rôzne posuny $j = 1, 2, ..., n - 1$. Hodnoty prvkov sú dané výrazom

$$\text{cov}(u_t, u_{t+j}) = E(u_t, u_{t+j}) = \frac{\rho^j}{1-\rho^2}\sigma_\varepsilon^2, \text{ pre } j \neq 0. \tag{1.31}$$

Napokon kovariančnú maticu získame tak, že jej rozptyly, určené vzťahom (1.30), budú na hlavnej diagonále a jej nediagonálne prvky budú kovariancie hodnôt náhodnej zložky modelu určovanými vzťahom (1.31), t. j.

$$\mathrm{E}(\mathbf{u'u}) = \Sigma = \frac{\sigma_\varepsilon^2}{1-\rho^2} \cdot \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ \rho & 1 & \rho^3 & \rho^4 & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & 1 \end{pmatrix}. \qquad (1.32)$$

Skutočné hodnoty náhodnej zložky u_t modelu nepoznáme, preto ich odhadneme rezíduálnymi odchýlkami e_t. Analogicky koeficient autokorelácie ρ nahradíme jeho odhadom $\hat{\rho}$, ktorý vypočítame OLS odhadovou metódou aplikovanú na schému (1.29) ako

$$\hat{\rho} = \frac{\sum_{t=2}^{N} e_t e_{t-1}}{\sum_{t=1}^{N} e_t^2}. \qquad (1.33)$$

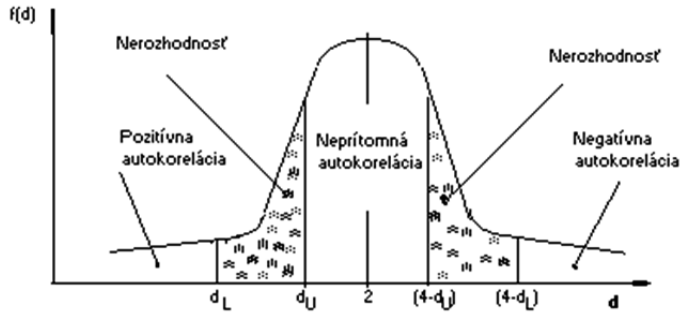
Autoregresná schéma (1.29) sa nazýva schémou autokorelácie rezíduí prvého rádu a je v praxi najčastejšia. Koeficient autokorelácie ρ môže nadobúdať hodnoty v intervale [1, –1], pričom, ak jeho hodnota je menšia ako nula, hovoríme o negatívnej autokorelácii, ak $\rho > 0$, hovoríme o pozitívnej autokorelácii. Ďalej sa budeme zaoberať len zisťovaním autokorelácie prvého rádu.

Najčastejšie používaným testom zisťovania autokorelácie prvého rádu je štatistika *d* Durbin-Watsonovho testu (Durbin and Watson, 1950), ktorá je definovaná ako

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \qquad (1.34)$$

Durbin-Watsonov test umožňuje štatisticky určiť, či existuje alebo neexistuje (je prítomná alebo nie je prítomná) autokorelácia rezíduí. Kritické hodnoty štatistiky boli tabuľované. Na príslušnej hladine významnosti α a sú označované ako d_L a d_U. Symbolmi d_L a d_U sú vyjadrené dolné (Lower) a horné (Upper) hranice hodnôt štatistiky *d* určenej vzťahom (1.34), v rámci ktorých nie je výsledok testu jednoznačný, ako to ukazuje obrázok 1–1. Z obrázku 1–1 vyplýva:

- ak je $d < d_L$ alebo $d > (4 - d_L)$, odmieta sa nulová hypotéza H_0: $\rho = 0$, čo indikuje prítomnosť autokorelácie.
- ak $d_U < d < (4 - d_U)$, nemôžeme odmietnuť nulovú hypotézu H_0: $\rho = 0$, čo indikuje neprítomnosť autokorelácie
- ak *d* nachádza sa medzi d_L a d_U, nie je možné jednoznačne indikovať prítomnosť alebo neprítomnosť autokorelácie. Obvykle je potrebné test overovať s väčším počtom pozorovaní.



Obrázok 1–1 Test autokorelácie: Durbin-Watsonova štatistika *d*

Postup odstraňovania autokorelácie závisí od konkrétnych príčin jej vzniku. Autokoreláciu je možné zmierniť alebo úplne odstrániť zmenou špecifikácie modelu (špecifikovaním ďalšej premennej) alebo zmenou analytického tvaru modelu. Ide o postupy, ktoré predovšetkým znamenajú prehodnotenie východiskovej hypotézy o modeli a vedú k novej konštrukcii modelu. Ak tieto pokusy v konečnom dôsledku neodstraňujú autokoreláciu, treba tento fakt brať ako skutočnosť a pri kvantifikácii modelu ho zohľadniť napr. použitím inej odhadovej metódy, než je OLS metóda. V ďalšej časti ukážeme metódy odhadu parametrov modelu pri výskyte autokorelácie rezíduí. Ide o metódu, ktorá je označovaná ako zovšeobecnená metóda najmenších štvorcov.

1.5 Zovšeobecnená metóda najmenších štvorcov

Zovšeobecnená metóda najmenších štvorcov (v literatúre označovaná anglickými iniciálkami GLS – Generalized Least Squares) je metódou na získanie najlepšieho lineárneho nevychýleného odhadu parametrov modelu **y** = **Xb** + **u** v prípade pozitívnej autokorelácie rezíduí v heteroskedastických modeloch. Ide o prípady, kde náhodná zložka modelu z rozličných dôvodov nespĺňa predpoklad o vzájomnej nekorelovanosti, resp. o konštantnom rozptyle v každých ich pozorovaniach. Predpokladajme, že koeficient autokorelácie ρ sme odhadli a je teda známy. Vieme, že regresný model (1.2) v tvare

$$y_t = b_0 + b_1 x_{t1} + b_2 x_{t2} + \ldots + b_k x_{tk} + u_t$$

platí pre všetky pozorovania *t*. Napíšme jeho tvar pre časový oneskorený posun o jedno obdobie, t. j. pre periódu *t* – 1

$$y_{t-1} = b_0 + b_1 x_{t-1,1} + b_2 x_{t-1,2} + \ldots + b_k x_{t-1,k} + u_{t-1}, \; t = 2, 3, ..., N.$$

Ak vynásobíme obidve strany poslednej rovnice koeficientom autokorelácie ρ a potom ju odčítame od predchádzajúcej rovnice, získame model v transformovanom tvare

$$_T y_t = {}_T b_0 + b_1 \; {}_T x_{t1} + b_2 \; {}_T x_{t2} + \; \ldots \; + b_k \; {}_T x_{tk} + {}_T u_t, \qquad (1.35)$$

v ktorom ${}_T y_t = y_t - \rho y_{t-1}$, ${}_T x_{tj} = x_{tj} - \rho x_{t-1,j}$ pre $j = 1, 2, ..., k$ sú zovšeobecnené diferencie, ${}_T b_0 = b_0 - \rho b_0 = b_0(1-\rho)$ a ${}_T u_t = u_t - \rho u_{t-1}$. Na odhad parametrov pre takto transformované dáta modelu (1.35) môžeme použiť OLS metódu, lebo v súlade s predpokladom (1.29) transformovaná náhodná zložka modelu (1.35) ${}_T u_t = u_t - \rho u_{t-1} = \varepsilon_t$ je náhodnou veličinou.

Model v transformovanom tvare (1.35) môžeme napísať v maticovom tvare ako

$${}_T\mathrm{y} = {}_T\mathrm{X}\,\mathrm{b} + {}_T\mathrm{u}$$

v ktorom vektory ${}_T\mathbf{y}$ a ${}_T\mathbf{u}$, matica ${}_T\mathbf{X}$ sú určené vzťahmi v (1.35) a vektor parametrov má prvky $\mathbf{b}` = \left[b_0(1-\rho) \cdot b_1 \cdot b_2 \cdot \ldots b_k \right]$. Estimátor pre odhad týchto parametrov je

$$\hat{\mathbf{b}} = ({}_T\mathbf{X}'\,{}_T\mathbf{X})^{-1}\,{}_T\mathbf{X}'\,{}_T\mathbf{y}. \qquad (1.36)$$

V transformovanom modeli (1.35) skutočný koeficient autokorelácie je nahradený jeho odhadom. Odhad parametrov GLS metódou síce znižuje autokoreláciu rezíduí, ale nemusí ju odstrániť úplne. Preto GLS metóda sa používa viacnásobne (cyklicky). V praxi sa najčastejšie používa pre odhad parametrov modelu s významnou autokoreláciou Cochrane-Orcuttova metóda, ktorá vedie ku konzistentným odhadom koeficienta autokorelácie. Algoritmus odhadu podľa tejto metódy je iteratívny, je vhodný pre počítač. Uvedieme ho v nasledovných výpočtových blokoch (bodoch).

Označme počítadlo iterácií a súčasne aj poradie iterácií identifikátorom *i*. Všetky premenné v tomto algoritme s indexom *i* budú sa vzťahovať na *i*-tú iteráciu, *i*-tý iteračný cyklus. Algoritmus môže mať napr. nasledovnú podobu:

1. Počiatočný (inicializačný) blok
   Počítač iterácií nastav na 1 $i = 1$
   Inicializuj $\hat{\rho}_0 = 0$
   Inicializuj $\delta = 0{,}001$
2. Výpočtový blok
   Prvotný odhad parametrov $\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$
   Vypočítaj **e** výraz (1.10)
   Vypočítaj $\hat{\rho}_i$ formula (1.33)
3. Rozhodovací blok
   Je $| \hat{\rho}_i - \hat{\rho}_{i-1} | < \delta$ áno: koniec výpočtu
   nie: skok na blok 4.
4. Transformácia dát (transformácia vektora **y** na ${}_T\mathbf{y}$ a matice **X** na ${}_T\mathbf{X}$) pomocou výrazov
   ${}_T y_t = y_t - \hat{\rho}_i y_{t-1}$
   ${}_T x_{tj} = x_{tj} - \hat{\rho}_i x_{t-1,j}$, $j = 1, 2, ..., k$, $t = 2, 3, ..., N$

5. Odhad parametrov z transformovaných dát
   Odhadni parametre $\hat{\mathbf{b}}_i = ({}_T\mathbf{X}'\,{}_T\mathbf{X})^{-1}\,{}_T\mathbf{X}'\,{}_T\mathbf{y}$
   Inkrementácia počítadla $i = i + 1$
   Vypočítaj **e** výraz (1.10)
   Vypočítaj $\hat{\rho}_i$ formula (1.33)
   Skok na blok 3.

Algoritmus ukončí výpočet, ak absolútna hodnota z rozdielu dvoch po sebe vypočítaných hodnôt autokorelačných koeficientov je menšia, ako je zvolená hodnota premennej δ (napr. 0,001).

Súhrnne možno k GLS metóde uviesť, že odhad parametrov modelu, za predpokladu porušenia homoskedasticity a za predpokladu autoregresného vzťahu náhodných zložiek modelu určuje sa podľa vzťahu (1.36). Pozornému čitateľovi zrejme neuniklo, že v transformačnom predpise (1.35) ide o transformáciu pozorovaných dát. Konkrétne touto transformáciou sa upravujú pozorované hodnoty vysvetľovanej a vysvetľujúcich premenných tak, že od pozorovaní v období t odpočítava sa $\hat{\rho}$ násobok hodnoty pozorovaní v období $t - 1$, pre $t = 2, 3, ..., N$. Potom odhad parametrov pri výskyte autokorelácie pomocou GLS metódy je identický s OLS metódou, ak sa použije OLS metóda na takto transformované pozorované dáta. Určitou nevýhodou GLS metódy, v prípade autokorelácie rezíduí, je strata prvého pozorovania pri transformácii dát. Vo väčšine prípadov pri dostatočnom počte pozorovaní to nie je vážnym nedostatkom. Vo viacrovnicových ekonometrických modeloch, môže byť táto skutočnosť dôvodom pre jej problematické uplatnenie pre odhad parametrov. Stratu informácie možno eliminovať tak, ak v modeli (1.35) doplníme *stratené* prvé pozorovania hodnotami ${}_T y_1^*$, ${}_T x_{ij}^*$ dané

$${}_T y_1^* = \sqrt{1-\hat{\rho}^2}\, y_1, \quad {}_T x_1^* = \sqrt{1-\hat{\rho}^2}\, x_{1j} \text{ pre } j = 1, 2, ..., k,$$

kde $\hat{\rho}$ je odhad autokorelačného koeficienta a hviezdičkou sme označili skutočnosť, že ide o hodnoty dodatočne doplnené.

1.6 Konštrukcia prognóz a predikčné intervaly

Regresné modely časových radov sa často používajú na určenie bodových predpovedí, t. j. na odhad hodnôt vysvetľovanej veličiny v období $p = N + 1, N + 2, ...$ a na určenie predikčných intervalov bodových predpovedí. Bodové predpovede, resp. ich predikčné intervaly je možné určiť, ak je model kvantifikovaný, ak možno predpokladať stabilitu parametrov modelu a ak sú k dispozícii očakávané hodnoty vysvetľujúcich premenných modelu. Za týchto predpokladov možno vypočítať bodové prognózy vysvetľovanej veličiny ako

$$\hat{y}_p = \mathbf{x}'_p \hat{\mathbf{b}}. \tag{1.37}$$

pre $p = N + 1, N + 2, ...$

kde $\mathbf{x}'_p$ je vektor $k + 1$ vysvetľujúcich premenných v obdobiach p, $\hat{\mathbf{b}}$ je vektor odhadnutých parametrov modelu. Vidíme, určovanie prognóz, t. j. prognózovanie vysvetľovanej veličiny znamená určovanie jej hodnôt na obdobie nasledujúce po období kvantifikácie.

Ak vo vzťahu (1.37) dosadíme za $\hat{\mathbf{b}}$ odhadový výraz, možno vypočítať bodové prognózy vysvetľovanej veličiny ako

$$\hat{y}_p = \mathbf{x}'_p \hat{\mathbf{b}} = \mathbf{x}'_p \, (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \tag{1.38}$$

kde $\mathbf{X}$ je matica pozorovaných hodnôt všetkých vysvetľujúcich premenných v obdobiach $t = 1, 2, ..., N$ rozmeru $N \cdot (k + 1)$, ktorej tvar a prvky je vidieť zo vzťahu (1.7). Symbol $\mathbf{y}$ vyjadruje stĺpcový vektor pozorovaných hodnôt vysvetľovanej veličiny rozmeru $N \cdot 1$. Skutočnú hodnotu vysvetľovanej premennej v období $p = N + 1, N + 2, ...$ označíme y_p. Potom chybu prognózy e_p určíme ako rozdiel skutočnej hodnoty vysvetľovanej premennej a vypočítanej prognózy, t. j.

$$e_p = y_p - \hat{y}_p. \tag{1.39}$$

Rozptyl chyby prognózy e_p označíme σ_p^2, ktorý je daný vzťahom (Montgomery a kol., 1990)

$$\sigma_p^2 = \mathbf{x}'_p \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p + \sigma^2, \tag{1.40}$$

kde σ^2 je skutočný rozptyl náhodnej zložky modelu. Skutočný rozptyl náhodnej zložky modelu nevieme vyčísliť, nahradíme ho odhadom s^2, daným výrazom (1.20), t. j.

$$s^2 = \frac{\sum_{t=1}^{N} e_t^2}{N - (k+1)}.$$

Po dosadení odhadu σ^2 ako s^2 do (1.40) označme odhad rozptylu chyby bodovej prognózy σ_p^2 ako s_p^2, t. j.

$$s_p^2 = \mathbf{x}'_p s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p + s^2$$

a po úprave, odhad rozptylu chyby bodovej prognózy je daný výrazom

$$s_p^2 = s^2 \left[1 + \mathbf{x}'_p \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p\right]. \tag{1.41}$$

Smerodajná odchýlka chyby predpovedi je daná odmocninou z jej rozptylu, t. j.

$$s_p = s\sqrt{1 + \mathbf{x}'_p \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p}. \tag{1.42}$$

Ak je známy odhad smerodajnej odchýlky chyby prognózy, za predpokladu normálneho rozdelenia chýb prognóz, môžeme vypočítať 100 $(1-\alpha)$ percentný predikčný interval bodovej prognózy $\hat{y}_p$ ako

$$\hat{y}_{p\,\min}^{\max} = \hat{y}_p \pm t_{\alpha,N-(k+1)} s_p = \hat{y}_p \pm t_{\alpha,N-(k+1)} s \sqrt{1 + \mathbf{x}'_p \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p}\,. \qquad (1.43)$$

Príklad 1.1

Na ilustráciu postupov a výrazov odhadu parametrov regresných modelov s aplikáciou na časové rady a na overenie ich vhodnosti uvažujme úlohu modelovania vývoja ročnej spotreby elektrickej (technologickej) energie y_t hutnej prevádzky na výrobu hliníkových polotovarov v závislosti od celkového ročného objemu výroby x_{t1} a indexu energetickej nákladovosti sortimentnej štruktúry výroby x_{t2} v jednotlivých rokoch. Jednotlivé skutočnosti (pozorovania) za obdobie rokov t = 1, 2, ..., 18 boli zistené z plánovacej dokumentácie a sú uvedené v tabuľke 1–2. Hutná prevádzka, na základe uzavretých kontraktov, plánuje vyrobiť v najbližších troch rokoch objemy hliníkových polotovarov s príslušnými indexovými energetickými nákladmi, ktoré sú uvedené v tabuľke 1–2 ako

Tabuľka 1–2 Data a pomocné výpočty príkladu 1.1

t	y_t	x_{t1}	x_{t2}	$\hat{y}_t$	$e_t = y_t - \hat{y}_t$	e_t^2	$(e_t - e_{t-1})^2$
1	9,99	421,11	24,75	9,877	0,112	0,0126	
2	10,65	466,13	24,78	10,810	–0,160	0,0256	0,0743
3	11,79	600,12	24,09	13,571	–1,781	3,1734	2,6280
4	11,99	526,63	24,87	12,064	–0,074	0,0055	2,9133
5	12,65	636,24	24,73	14,331	–1,681	2,8267	2,5815
6	13,18	513,6	24,92	11,795	1,384	1,9161	9,3976
7	14,07	588,46	25,23	13,351	0,718	0,5162	0,4432
8	14,16	587,14	26,66	13,351	0,808	0,6543	0,0081
9	13,88	648,97	28,96	14,674	–0,794	0,6312	2,5709
10	17,23	739,4	24,36	16,460	0,769	0,5926	2,4472
11	16,53	730,25	27,04	16,321	0,208	0,0436	0,3147
12	16,17	780,83	29,39	17,412	–1,242	1,5441	2,1067
13	17,92	789,37	27,24	17,548	0,371	0,1376	2,6039
14	18,77	780,14	26,81	17,349	1,420	2,0170	1,1008
15	18,71	813,77	27,15	18,052	0,657	0,4323	0,5816
16	17,01	803,37	24,25	17,782	–0,772	0,5967	2,0451
17	18,14	818,15	22,92	18,063	0,076	0,0058	0,7208
18	18,16	819,43	27,74	18,180	–0,020	0,0004	0,0094
19	18,93	820,14	28,74	18,214			
20	19,01	799,37	26,93	17,750			
21	18,54	826,21	27,53	18,317			
Suma						15,132	32,548

budúce hodnoty nezávisle premenných x_{1t} a x_{2t}, pre t = 19, 20, 21. Našou úlohou je opísať proces vývoja spotreby elektrickej energie, v závislosti od objemu výroby a indexnej nákladovosti a na základe neho určiť spotrebu elektrickej energie pre plánovanú výrobu v budúcich troch rokoch.

Regresný vzťah spotreby elektrickej energie v závislosti na ročnom objeme výroby a indexe energetickej nákladovosti sortimentnej štruktúry vyjadríme v súlade s tvarom modelu (1.2) lineárnym vzťahom, t. j.

$$y_t = b_0 + b_1 x_{t1} + b_2 x_{t2} + u_t, \qquad \text{(a)}$$

v ktorom parametre b_0, b_1, b_2 odhadneme pomocou výrazu (1.16)

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \qquad \text{(b)}$$

Matica **X** v odhadovom výraze (b) je tvorená hodnotami prvkov

$$\mathbf{X} = \begin{pmatrix} 1 & 421{,}11 & 24{,}75 \\ 1 & 466{,}13 & 24{,}76 \\ 1 & 600{,}12 & 24{,}09 \\ 1 & 526{,}63 & 24{,}87 \\ 1 & 636{,}24 & 24{,}73 \\ 1 & 513{,}60 & 24{,}92 \\ 1 & 588{,}46 & 25{,}23 \\ 1 & 587{,}14 & 26{,}66 \\ 1 & 648{,}97 & 28{,}96 \\ 1 & 739{,}40 & 24{,}36 \\ 1 & 730{,}25 & 27{,}04 \\ 1 & 780{,}83 & 29{,}39 \\ 1 & 789{,}37 & 27{,}24 \\ 1 & 780{,}14 & 26{,}81 \\ 1 & 813{,}77 & 27{,}15 \\ 1 & 803{,}37 & 24{,}25 \\ 1 & 818{,}15 & 22{,}92 \\ 1 & 819{,}43 & 27{,}74 \end{pmatrix}.$$

Matica $(\mathbf{X}`\mathbf{X})^{-1}$ má hodnoty prvkov

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 12{,}322161 & -0{,}000185 & -0{,}469134 \\ -0{,}000185 & 3{,}747E-06 & -8{,}99E-05 \\ -0{,}469134 & -8{,}99E-05 & 0{,}0204518 \end{pmatrix}.$$

Vypočítané hodnoty parametrov regresného modelu (a) výrazom (b) potom sú

$$\hat{\mathbf{b}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = \begin{pmatrix} 0,693 \\ 0,0207 \\ 0,0188 \end{pmatrix}.$$

Smerodajné odchýlky odhadov (estimátorov) parametrov $\hat{b}_1$ a $\hat{b}_2$ modelu (a) vypočítame na základe výrazov (1.20) a (1.21), t. j.

$$s_{\hat{b}_1} = \left[s^2 \left(\mathbf{X}'\mathbf{X} \right)^{-1}_{22} \right]^{1/2} = \left[\frac{\sum_{t=1}^{18} e_t^2}{18-(2+1)} 3,75\ 10^{-6} \right]^{1/2} = 0,002,$$

kde súčet štvorcov rezíduí je vypočítaný v tabuľke 1–2, hodnota výrazu $(\mathbf{X}'\mathbf{X})^{-1}_{22}$ je prevzatá z matice $(\mathbf{X}'\mathbf{X})^{-1}$ ako hodnota druhého diagonálneho prvku. Analogickým spôsobom vypočítame

$$s_{\hat{b}_2} = 0,144 .$$

Posúdenie významnosti parametrov $\hat{b}_1$ a $\hat{b}_2$ urobíme na základe kritéria (1.26) na hladine významnosti $\alpha = 0,05$

$$\left| t_j = \frac{\hat{b}_j}{s_{\hat{b}j}} \right| > t_{\alpha,(N-k-1)}.$$

Pre parameter $\hat{b}_1$, ($j = 1$) dostaneme

$\left| t_1 = \frac{0,0207}{0,002} \right| = 10,649 > t_{0.05,(18-2-1)} = 2,1315,$ z čoho usúdime, že príslušný regresor $\hat{b}_1$ je štatisticky významný. Analogicky pre parameter $\hat{b}_2$, ($j = 2$) dostaneme

$\left| t_2 = \frac{0,0188}{0,144} \right| = 0,131 < t_{0.05,(18-2-1)} = 2,1315,$ z čoho usúdime, že príslušný regresor $\hat{b}_2$ je štatisticky nevýznamný.

Ako mieru presnosti vysvetlenia závislej veličiny y_t modelom použijeme koeficient determinácie R^2, ktorý vypočítame podľa vzťahu (1.22)

$$R^2 = 1 - \frac{\sum_{t=1}^{18} e_t^2}{\sum_{t=1}^{18} (y_t - \overline{y})^2} = 1 - \frac{15,1327}{144,0624} = 0,8949,$$

v ktorom $\overline{y} = 15,056$ je aritmetický priemer časového radu y_t pre t = 1, 2 , ..., 18. Vidíme, že modelom je na 89 % vysvetlená variabilita y_t. Pomocou koeficienta determinácie overíme významnosť zhody dát vysvetľovanej premennej, ktoré boli odhadnuté (vypočítané) modelom s pozorovanými dátami. Štatistika, pomocou ktorej sa otestujeme platnosť H_0: $b_1 = b_2 = ... = b_k = 0$ v súlade s (1.24) je

$$F_R = \frac{V / k}{\mathrm{N}/\left[N-(k+1)\right]} = \frac{R^2 / k}{\left(1-R^2\right)/\left[N-(k+1)\right]} = 63,899.$$

Vidíme, že v súlade s (1.25) pre hladinu významnosti $\alpha = 0,05$, $k = 2$, $N = 18$ platí

$$F_R > F_{0.05,2,15} = 3,682,$$

čo znamená, že model je významný ako celok.

Na testovanie autokorelácie náhodnej zložky modelu vypočítame štatistiku d Durbin-Watsonovho testu podľa (1.34)

$$d = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2} = \frac{32,548}{15,132} = 2,151.$$

Kritické hodnoty pre $\alpha = 0,05$, $k = 2$, $N = 18$ sú: $d_U = 1,53$, $d_L = 1,05$. V našom príklade platí: $d_U < d < (4 - d_U)$, čo indikuje neprítomnosť autokorelácie náhodnej zložky modelu.

Nakoniec posúdime, či rozdelenie náhodných chýb máú normálne rozdelenie. Posúdenie vykonáme na základe grafického priebehu rezíduí e_t v závislosti od normovaných rezíduí. Normované rezíduá získame tak, že rezíduá e_t podelíme ich smerodajnými odchýlkami. Smerodajné odchýlky chýb odhadu pre pozorovania t = 1, 2, ..., N vypočítame analógiou výrazu (1.22), t. j.

$$s_{e_t} = s\sqrt{1 + \mathbf{x}_t'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_t}, \qquad \text{(c)}$$

kde s je odhad smerodajnej odchýlky náhodnej zložky modelu. Určíme ho pomocou výrazu $s = \left(\frac{\sum_{t=1}^{n} e_t^2}{N-(k+1)}\right)^{\frac{1}{2}} = \left(\frac{15,132}{18-(2+1)}\right)^{\frac{1}{2}} = 1,0044$. Vo výraze (c) $\mathbf{x}_t'$ je riadkový vektor $k + 1$ vysvetľujúcich premenných v pozorovaní t. Napr. pre pozorovanie $t = 2$, z tabuľky 1–2 priamo vyčítame jeho prvky $\mathbf{x}_2' = [1\ 466,13\ 24,78]$. Prvky matice $\mathbf{(X`X)}^{-1}$ sme určili vo výraze (b). Potom hodnota smerodajnej odchýlky chyby odhadu pre pozorovanie t = 2, vypočítaná

výrazom (c) je $s_{e_2} = 1,0984$ a normovaná hodnota rezíduá $\left({}_n e_2\right)$ v tomto pozorovaní je

$${}_n e_2 = e_2 / s_{e_2} = -0,163 / 1,0984 = -0,14594.$$

Analogicky sme vypočítali všetky normované rezíduá ${}_N e_t$, pre $t = 1, 2, ..., 18$ a spolu s rezíduami e_t graficky zobrazili v obrázku 1–2.

Na obrázku 1–2 je vidno, že priebeh rezíduí v závislosti od normovaných rezíduí je takmer priama čiara. Na základe toho môžeme konštatovať, že náhodná zložka u_t modelu (a) má normálne rozdelenie. Súhrnne môžeme získaný model prezentovať v nasledovnom tvare, v ktorom sú uvedené výsledky testov, resp. jeho dôležité charakteristiky presnosti

$$\hat{y}_t = \underset{(0,002)}{0,693} + \underset{(0,1436)}{0,0207 x_{t1}} + 0,0188 x_{t2}$$

$$R^2 = 0,8949 \qquad F_R = 63,899 \qquad d = 2,150$$

kde hodnoty v zátvorkách sú vypočítané smerodajné odchýlky príslušných parametrov.

Po získaní celkove dobrých výsledkov testov modelu, ako aj na základe jeho pomerne vysokej presnosti môžeme za predpokladu jeho nemennosti v období predpovedí prikročiť ku konštrukcii predpovedí vysvetľovanej veličiny. Na tento účel vypočítame najskôr variačný koeficient *VK*.

$$VK = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2}}{\bar{y}} 100 = \frac{\sqrt{\frac{1}{18}15,132}}{15,056} 100 = 6,09\ \%.$$

Vidíme, že variačný koeficient je nízky (pod 20 %), čo nás oprávňuje k aplikácii navrhnutého modelu na konštrukciu krátkodobých prognóz.

Bodové predpovede vysvetľovanej veličiny určíme výrazom (1.37)

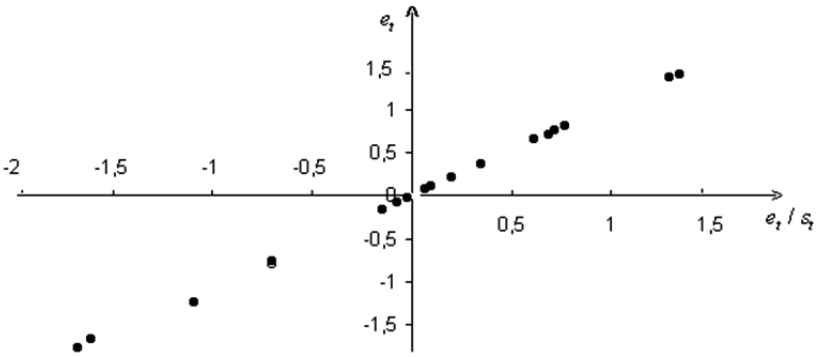
$$\hat{y}_p = \mathbf{x}'_p \hat{\mathbf{b}}, \text{ pre } p = N + 1, N + 2, \ldots .$$

Napr. bodovú predpoveď pre obdobie $p = N + 1 = 18 + 1 = 19$ vypočítame

$$\hat{y}_{19} = \mathbf{x}'_{19} \hat{\mathbf{b}} = [1 \quad 820,14 \quad 28,74] \begin{pmatrix} 0,693 \\ 0,0207 \\ 0,0188 \end{pmatrix} = 18,214.$$

Podobne sa vypočítajú bodové predpovede pre $p = 20$ a $p = 21$. Ich hodnoty sú uvedené v tabuľke 1–1.

Na určenie predikčných intervalov bodových predpovedí musíme poznať smerodajné odchýlky s_p chýb jednotlivých predpovedí. Tieto určíme pomocou



Obrázok 1–2 Priebeh rezíduí e_t verzus normované rezíduá e_t / s_t pri normálnom rozdelení (príklad 1.1)

výrazu (1.42). Napr. smerodajná odchýlka bodovej predpovede $\hat{y}_p$, pre $p = 19$ je

$$s_{19} = s\sqrt{1 + \mathbf{x}'_{19}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_{19}} = 1{,}0044\sqrt{1 + \mathbf{x}'_{19}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_{19}} = 1{,}114.$$

Analogickým postupom určíme aj smerodajné odchýlky bodových predpovedí pre $p = 20$ a $p = 21$ ako

$$s_{20} = 1{,}0044\sqrt{1 + \mathbf{x}'_{20}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{20}} = 1{,}0612,$$

$$s_{21} = 1{,}0044\sqrt{1 + \mathbf{x}'_{21}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{21}} = 1{,}08.$$

Ak máme určené odhady smerodajných odchýlok chýb prognóz, za predpokladu normálneho rozloženia chýb prognóz, môžeme vypočítať $100(1-\alpha)$ percentný predikčný interval bodovej prognózy $\hat{y}_p$ pomocou výrazu (1.43) ako

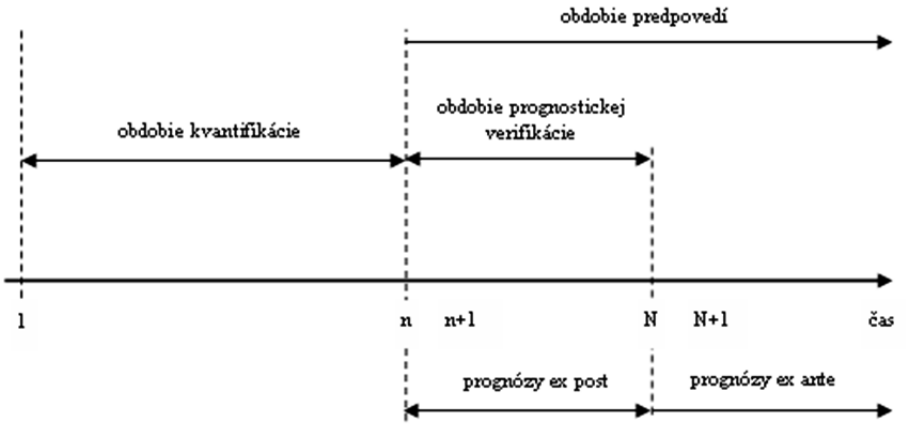
$$y_{p\,\min}^{\max} = \hat{y}_p \pm t_{\alpha,N-(k+1)}s_p.$$

Pre zvolenú hladinu významnosti $\alpha = 0.05$ a počtu stupňov voľnosti $v = N - (k + 1) = 18 - (2 + 1) = 15$ je kritická hodnota Studentovho t rozdelenia rovná 2,1315. Takže pre predikčné intervaly bodových prognóz v období $p = 19,20,21$ platí

$$y_{p\,\min}^{\max} = \hat{y}_p \pm 2{,}1315\, s_p = \begin{matrix} \min & \max \\ 15{,}840 & 20{,}588 \\ 15{,}488 & 20{,}012 \\ 16{,}015 & 20{,}619. \end{matrix}$$

1.7 Hodnotenie presnosti predpovedí

Problém hodnotenia chýb predpovedí spočíva v tom, že spravidla skutočnú hodnotu prognózovanej veličiny y_p v období koncového pozorovania N nepo-



Obrázok 1–3 Časová os prognózovania

známe. Budeme predpokladať a zabezpečovať, aby bola minimálna. Aby sme mohli ohodnotiť veľkosť chyby prognózy a hodnotiť prognostickú schopnosť jednotlivých modelov navzájom, prakticky postupujeme tak, že do odhadu parametrov nezahrňujeme všetky pozorovania, ktoré máme k dispozícii, ale len časť, obyčajne vekovo staršie pozorovania a ostatnú časť, obvykle niekoľko posledných pozorovaní (vekovo mladšie pozorovania), ponecháme ako rezervu na hodnotenie prognóz. Uvedenú situáciu rozdelenia *histórie* pozorovaní znázorňuje obrázok 1–3.

Označme N celkový počet pozorovaní, ktoré máme k dispozícii pre kvantifikáciu modelu. Na obrázku 1–3 je vidno, nie všetky dáta, ktoré máme k dispozícii za N pozorovaní použijeme na kvantifikáciu, ale len n pozorovaní. Pozorovania $n + 1, n + 2, ..., N$ ponechávame na prognostickú verifikáciu modelu. Prognózam, ktoré konštruujeme pre obdobie $n + 1, n + 2, ..., N$, hovoríme prognózy ex post (alebo pseudoprognózy) a prognózam, ktoré konštruujeme pre obdobia $N + 1, N + 2, ...$, hovoríme prognózy ex ante. Pre ekonomické rozhodovania majú zmysel prognózy ex ante.

Pri prognózach ex post sú známe skutočnosti endogénnych premenných, poznáme i pozorovania nezávislých premenných. Je možné pre ne vyčísliť chyby prognóz podľa (1.5) pre $p = n + 1, n + 2, ..., N$.

Pre posúdenie prognostickej vhodnosti modelu používa sa celý rad charakteristík, ktoré sú založené na vyčíslení chýb prognóz ex post. Uvedieme niektoré z nich. Označme počet prognózovaných období M, z ktorých budeme vyčíslovať chyby prognóz ex post. Na obrázku 1–3 je zrejmé, že $M = N – n$.

Základnou charakteristikou, ktorou sa posudzuje presnosť modelu je priemerná štvorcová chyba prognóz, označovaná *MSE* (Mean Square Error). Vypočíta sa podľa vzorca