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

Dušan Marček

SUPERVIZOVANÉ A NESUPERVIZOVANÉ UČENÍ Z DAT: STATISTICKÝ A SOFT PŘÍSTUP

Ostrava, 2016

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

Reviews
Petr Dostál, Brno University of Technology
Peter Fabián, University of Žilina

This work was partially supported by project reg. No. OP VK CZ.1.07/2.3.00/20.0296 *Researcher team for modelling of economic and financial processes* at VSB − Technical University of Ostrava.

The text should be cited as follows: Marček, D. (2016). *Supervizované a nesupervizované učení z dat: statistický a soft přístup*, SAEI, vol. 45. Ostrava: VSB-TU Ostrava.

Printed in Tribun EU, s.r.o.
Cover design by BELISA Advertising, s.r.o.

ISBN 978-80-248-3884-7

Předmluva

V poslední dekádě nastala exploze výpočtových a informačních technologií. Paralelně s tím a jako důsledek tohoto fenoménu došlo k velkému nárůstu dat ve všech oblastech společenského života. S velkým nárůstem dat se data stávají čím dále tím více složitější, strukturovanější, v důsledku čehož, i jejich informační interpretace a porozumění je čím dále tím víc komplikovanější a pracnější. Aby bylo možné data správně informačně interpretovat a využít, je naléhavou výzvou ve všech vědních oblastech dostat se pod jejich vnější projev k vnitřním strukturám a objevovat v nich skryté vzory a souvislosti, a tak usuzovat na jejich vývoj, tento vývoj modelovat, kvantifikovat, predikovat a konfrontovat s realitou. Jednoduše řečeno, porozumět datům. Jako reakce na porozumění těmto datům vznikly ve statistice a informatice nové informatické obory, jako jsou dolování a strojové učení z dat.

Procesy a metody, na základě kterých se proniká k vnitřním strukturám dat, kterými se analyzují a klasifikují data, kterými se v nich vyhledávají souvislosti, modelují struktury, závislosti, trendy, vývojové tendence a konfrontují se s realitou, budeme nazývat učením z dat.

V teorii statistických metod pro analýzu dat je učení z dat klasifikované do dvou základních skupin. Prvá skupina je nazývána učením s učitelem nebo kontrolované učení (*supervised learning*). Je to učení, ve kterém je znám nějaký vzor, reference, šablona, předloha nebo jednoduše a obrazně řečeno učitel, kteří na proces učení dohlížejí tak, aby výsledek učení se co nejvíc shodoval s předlohami. Druhá skupina je nazvána učením bez učitele nebo nekontrolované učení (*unsupervised learning*). V nekontrolovaném učení, na rozdíl od kontrolovaného učení, nejsou známy vzory či předlohy nebo osvědčené postupy. Pokud není k dispozici žádná reference na dohlížení správného postupu analýzy a modelování, musí být do procesu učení zakomponovaný samoorganizující mechanizmus. Samoorganizující mechanizmus musí umožnit na základě lokálních informací usuzování o vlastnostech dat, jejich vzájemných závislostech a určení strategie nebo postupu dosažení požadovaných informací.

V současné době už existuje početná literatura, která se zabývá na všeobecné konceptuální úrovni metodami učení z dat. Tato kniha je zaměřena na celkový výklad učení z dat, tj. teorie, použité metody, vývoj modelu až do aplikačních příkladů na reálných datech a též na zpřístupnění vyvinutých učicich metod autora úspěšně prezentovaných na mezinárodních fórech nebo publikovaných

v prestižních mezinárodních časopisech. Kniha se zabývá v teoretické a aplikační rovině vybranými modely strojového učení s učitelem a bez učitele z dat metodami statistické analýzy a UI. Poskytuje se jejich charakteristika a zhodnocují se z hlediska praktických aplikací.

Text knihy je po věcné stránce organizovaný do dvou částí. Prvá část – kapitoly 1 až 5 se věnují problematice kontrolovaného učení. Jejím cílem je v teoretické a aplikační rovině obeznámit se s vybranými metodami a modely učení s učitelem založené na klasické regresní analýze, ARMA modelech a modelech přenosových funkcích, modelech logistické regrese a exponenciálního vyrovnávaní, dále modely založené na strojovém učení SVM a na nejnovějších modelech dopředních neuronových sítí typu perceptron a RBF sítí. Druhá část – kapitoly 6 až 11 v teoretické a aplikační rovině se zabývá vybranými statistickými metodami nekontrolovaného učení, jako jsou diskriminační a faktorová analýza, metoda hlavních komponent, a skupinou umělých neuronových sítí s nekontrolovaným učením založeným Hebbovem a na konkurenčním (kompetitivním) učení. V kapitole 11 jsou uvedeny a diskutovány algoritmy pro vyhledávání shluků a jejich popisných charakteristik z vícerozměrných dat, které byly vyvinuty v rámci informatických věd.

Kniha je určena studentům studujícím informační technologie, aplikovanou informatiku a výpočetní techniku a taktéž pro studenty všech studijních programů s předměty matematická statistika, ekonometrie, prognostika v doméně analýzy, modelování a kvantitativní predikce hospodářských procesů, finančních trhů nebo společenských jevů. Poskytuje učební základ studentům doktorandského studia informačních technologií a informatiky pro odkrývání závislosti dat v databázích o ekonomických procesech, ve finančních trzích, marketingových procesech, v manažerských teoriích a v praxi.

Obrovský dík patří studentům a doktorandům, kterým byl autor vedoucím diplomových prací nebo školitelem a kteří všichni úspěšně obhájili diplomové nebo dizertační práce. Byli to studenti T. Laštík, L. Falát, J. Kolčák, P. Kuběj, J. Bábel, A. Kotillová, Z. Mečiarová, L. Zajíc a další, kteří se v rámci výuky zčásti podíleli na zpracování programů a výpočtů, a také Milanovi Marčekovi za cenné připomínky a souhlas k převzetí textu z jeho publikace.

Autor vyslovuje poděkování recenzentům: Prof. Petru Dostálovi, Ph.D., a doc. Petrovi Fabiánovi. Autor rovněž děkuje Ekonomické fakultě Vysoké školy báňské – Technické univerzity Ostrava za vytvoření podmínek k vydání knihy.

Obsah

Kapitola 1

Úvod do modelování a predikce časových řad

Předmětem publikace je popis základních statistických postupů a technik, které se používají pro analýzu a modelování časových řad ekonomických veličin a na predikci hodnot časových řad. Základní schéma, pomocí kterého budeme popisovat pozorování časové řady y_t v čase $t = 1, 2, ..., N$, je

$$y_t = D_t + S_t + \varepsilon_t \text{ , pro } t = 1, 2, ..., N, \qquad (1.1)$$

v níž D_t vyjadřuje deterministickou část schématu, S_t je sezónní nebo oscilační komponent schématu a ε_t je chybový, poruchový nebo náhodný komponent schématu. Naším úkolem bude funkčně specifikovat, tj. modelovat deterministickou část D_t a sezónní komponent S_t schématu (1.1) tak, aby statistická míra závislosti dat y_t na D_t a S_t byla co nejvyšší. Při specifikaci deterministické a sezónní části schématu (1.1) budeme v podstatě používat dvě kvantitativní metody či techniky. První technika se zakládá na popisu vztahů mezi ekonomickými veličinami pomocí regresní analýzy. Jde o skupinu metod a modelů, pomocí kterých je vývoj vysvětlováním (endogenní, závislé) proměnné popisován pomocí kauzálních, resp. symptomatických proměnných. Předpokládá se, že pro časovou závislost mezi endogenní proměnou a kauzálními (exogenními, nezávislými) proměnnými existuje ověřená ekonomická teorie. V případě, že odpovídající ekonomická teorie není k dispozici nebo mohou vzniknout pochybnosti o její vhodnosti, pak regresní vztah lze postavit na kauzálních i symptomatických proměnných, příp. jen na symptomatologických proměnných. Jednou z nejčastěji používaných symptomatických proměnných je samotná časová proměnná, a to ve formě vhodných funkcí časových period $t = 1, 2, ..., N$.

Druhá technika vychází z předpokladu, že na hodnoty pozorování časové řady veličiny je možné se dívat jako na realizaci náhodného procesu. Proměnné a mechanismus jejich závislostí, pomocí kterých jsou generovány hodnoty časové řady, se nehledají mimo časovou řadu, tj. v exogenních proměnných, ale metodami statistické analýzy dat se vyhledávají takové časové posuny a závislosti

v hodnotách časové řady a v náhodném členu schématu (1.1), které adekvátně popisují pozorování časové řady. Dá se to zjednodušeně říci i tak, že zde jde o vyhledávání vnitřní struktury dat a závislostí, které generují pozorovaná data s přijatelnou přesností. Proces generování dat je popisován třídou lineárních modelů nazývanou ARMA modely.

Mezi oběma technikami není přesná rozlišovací hranice, uvedené techniky se vzájemně doplňují. Např. v některých postupech k určení náhodné složky ARMA modelu se používá technika regresní analýzy. Na specifikaci nezávislých proměnných a jejich časových posunů v modelech založených na regresní analýze se používají ARMA modely. Můžeme proto hovořit o kombinovaných technikách. Jak uvidíme např. v kapitole ARCH modely nebo v podkapitole 2.4 – Modely přenosových funkcí, na popis deterministické části, resp. i sezónního komponentu jsou aplikovány techniky regresní analýzy a na popis náhodného člena je aplikována ARMA metodologie.

Uvedená klasifikace statistických technik na modelování časových řad má dopad na posouzení přesnosti či adekvátnosti toho kterého modelu časové řady. Základní mírou přesnosti modelů založených na regresní analýze je koeficient determinace. Jím je vyjádřena míra závislosti endogenní proměnné na exogenních proměnných. Je to míra, která procentně vyjadřuje stupeň vysvětlení proměnlivosti endogenní proměnné regresním modelem.

V modelovacím přístupu, ve kterém se vyhledává mechanismus, kterým jsou nejlépe reprodukovány pozorovaná data, kromě uvedené základní míry těsnosti závislosti přistupují další kritéria, založené na analýze reziduí e_t. Rezidua jsou vytvořena posloupností, v níž jednotlivé prvky jsou určovány jako rozdíl hodnot pozorování od modelu generovaných hodnot dat $\hat{y}_t$, tj.

$$e_t = y_t - \hat{y}_t = y_t - \left(D_t + S_t\right). \tag{1.2}$$

Rezidua by měla být časově stabilní, náhodné proměnné s přibližně normálním rozdělením s nulovou střední hodnotou a konstantním rozptylem.

Všimněme si, že časová řada vyjádřená schématem (1.1) je modelem pro data y_t pro $t = 1, 2, ..., N$. Tato data nebo každé pozorování vyjádřené schématem (1.1), vzhledem k její náhodné komponentě, lze považovat za realizace určité náhodné veličiny Y_t a potom časovou řadu y_t pro $t = 1, 2, ..., N$ lze považovat za realizaci posloupnosti náhodných proměnných veličin Y_t pro $t = 1, 2, ..., N$. Na specifikované schéma (1.1) se proto budeme dívat jako na model stochastického procesu.

Výrazem (1.1) je vyjádřena struktura modelu časové řady, která říká jen tolik, že časová řada je určitým předpokládaným souhrnem deterministické složky, charakterizující dlouhodobé pravidelnosti, sezónní a náhodné složky. Při analýze časových řad vzniká otázka, jak vyjádřit deterministickou a sezónní složku. Budeme předpokládat, že deterministickou i sezónní můžeme vyjádřit třídou

lineárních funkcí. Např. pokud budeme abstrahovat od sezónní složky modelu, deterministickou složku vyjádříme např. jako funkci času *t* ve tvaru

$$y_t = a_0 + a_1 t + a_2 t^2 + \varepsilon_t \text{ , } t = 1, 2, ..., N, \tag{1.3}$$

kde a_0, a_1, a_2 jsou parametry modelu. Tyto parametry musíme statisticky odhadnout, tj. na základě *N* pozorovaní o hodnotách y_t a časové proměnné *t* pomocí určité statistické metody tak, aby byl minimalizován zpravidla výraz $\sum_{t=1}^{N}(y_t - \hat{y}_t)^2$. Odhadnuté parametry budeme označovat se stříškou jako $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$.

Proces statistického odhadu parametrů modelu se nazývá kvantifikace modelu. Pokud se odhadnuté parametry dosadí do modelu (1.3), potom platí pro každé pozorování

$$y_t = \hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2 + e_t, \tag{1.4}$$

kde e_t je chyba vztahu, reziduum určené výrazem (1.2).

V dalších kapitolách se budeme zabývat různými metodami odhadů deterministické a oscilační složky modelu (1.1), resp. jejich odstraňováním tak, abychom získali časovou řadu nazývanou stacionární časová řada. Přesnější k této problematice pojednává podkapitola 2.4.

1.1 Koncept predikce časových řad

Našim primárním cílem při modelování časových řad je specifikace modelu, pomocí kterého je možné v určitém smyslu zaručit minimální chybu předpovědí. Chybu předpovědi budeme definovat jako následnou chybu předpovědi e_p v období předpovědí $p = N + 1, N + 2, ...$ jako

$$e_p = y_p - \hat{y}_p, \tag{1.5}$$

kde $\hat{y}_p$ je modelem určená předpověď pro období *p*, y_p je skutečná hodnota veličiny. V dalším textu se omezíme na skupinu tzv. lineárních předpovědí, podle níž hledáme takovou lineární kombinaci známých pozorování časové řady y_t pro $t = 1, 2,..., N$, která nejlépe aproximuje budoucí skutečnou hodnotu y_p. Na formální vyjádření, v tomto smyslu nejlepší lineární předpovědi, předpokládejme, že stojíme před problémem vyhledání předpovědi o jedno období dopředu s minimální chybou předpovědi, podle níž je vyhledávána předpověď o jedno období dopředu lineární kombinací všech pozorování, tj.

$$\hat{y}_N(1) = \hat{y}_{N+1} = \sum_{i=1}^{N} b_i(1)\, y_{N+1-i} \tag{1.6}$$

při podmínce minimalizace výrazu

$$E\left[e_N^2(1)\right] = E\left(y_{N+1} - \hat{y}_{N+1}\right)^2 = E\left(y_{N+1} - \sum_{i=1}^{N} b_i(1)\, y_{N+1-i}\right)^2. \qquad (1.7)$$

Lineární prediktor (1.6) je velmi jednoduchý. Jak uvidíme v dalších kapitolách v konkrétních modelech, lze na základě něj snadno vytvořit rekurzivní algoritmus k určení předpovědi pro libovolný horizont τ, tj. předpověď o τ období dopředu. Algoritmus je v literatuře znám pod označením Durbin–Levinsonovi rekurze (Brockwell a Davis, 1987).

1.2 Hodnocení přesnosti předpovědí

Problém hodnocení chyb předpovědí spočívá v tom, že zpravidla skutečnou hodnotu prognózované veličiny y_p v období koncového pozorování *N* neznáme. Víme o ní jen to, že je minimální v souladu s výrazy (1.6) a (1.7). Abychom mohli ohodnotit velikost chyby prognózy a hodnotit prognostickou schopnost jednotlivých modelů navzájem, prakticky postupujeme tak, že do odhadu parametrů nezahrnuje všechna pozorování, která máme k dispozici, ale jen část, obvykle věkově starší pozorování, a ostatní část, obvykle několik posledních pozorování (věkově mladší pozorování), ponecháme jako rezervu na hodnocení prognóz. Uvedenou situaci rozdělení *historie* pozorování znázorňuje obrázek 1–1.

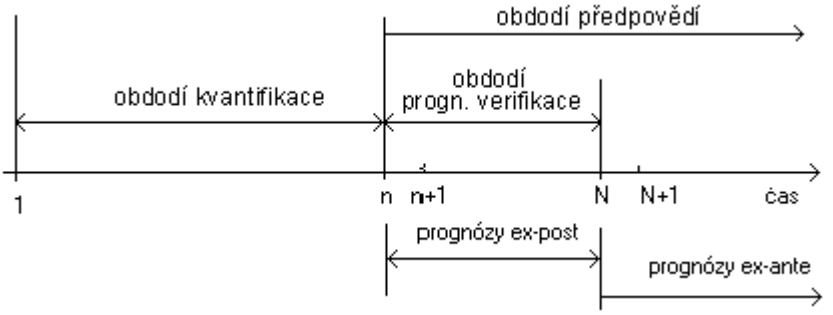
Označme *N* celkový počet pozorovaní, která máme k dispozici pro kvantifikaci modelu. Z obrázku 1–1 je vidět, že ne všechna data, která máme k dispozici za *N* pozorování, použijeme na kvalifikaci, ale jen *n* pozorovaní. Pozorování *n* + 1, *n* + 2, ..., *N* ponecháváme na prognostickou verifikaci modelu. Prognózám, které konstruujeme pro období *n* + 1, *n* + 2, ..., *N*, říkáme prognózy ex post (nebo pseudoprognózy) a prognózám, které konstruujeme pro období *N* + 1, *N* + 2, ..., říkáme prognózy ex-ante. Pro ekonomické rozhodování mají smysl prognózy ex-ante.

Při prognózách ex-post jsou známé skutečnosti endogenních proměnných, známe i pozorování nezávislých proměnných. Je možné pro ně vyčíslit chyby prognóz podle (1.5) pro *p* = *n* + 1, *n* + 2, ..., *N*.

Pro posouzení prognostické vhodnosti modelu se používá celá řada charakteristik, které jsou založeny na vyčíslení chyb prognóz ex-post. Uvedeme některé z nich. Označme počet prognózovaných období *M*, z nichž budeme vyčíslovat chyby prognóz ex-post. Z obrázku 1–1 je zřejmé *M* = *N* – *n*.

Základní charakteristikou, kterou se posuzuje přesnost modelu, je průměrná čtvercová chyba prognóz, označovaná MSE (Mean Square Error). Vypočítá se podle vzorce

$$MSE = \tfrac{1}{M} \sum_{p=n+1}^{N} \left(y_p - \hat{y}_p\right)^2. \qquad (1.4)$$



Obrázek 1–1 Časová osa prognózování

Odmocnina z MSE je ukazatelem variability posloupnosti chyb prognóz e_p.

Za vhodnou míru variability relativních chyb prognóz se považuje Theilův koeficient nesouladu T^2 (Kozák a Seger, 1971)

$$T^2 = \frac{\sum_{p=n+1}^{N} \left(y_p - \hat{y}_p \right)^2}{\sum_{p=n+1}^{N} y_p^2}. \tag{1.7}$$

Obě míry interpretujeme tak, že čím jsou vyšší, tím větší je nepřesnost posuzovaných prognóz.

Kromě uvedených dvou souhrnných měr přesnosti prognóz v praxi se často používají další míry, ze kterých orientačně uvedeme:

- průměrná absolutní relativní chyba (Mean Absolute Percentage Error)

$$MAPE = \tfrac{1}{M} \sum_{p=n+1}^{N} \frac{\left| y_p - \hat{y}_p \right|}{y_p}, \tag{1.8}$$

- průměrná relativní chyba (Mean Percentage Error)

$$MPE = \tfrac{1}{M} \sum_{p=n+1}^{N} \frac{y_p - \overline{y}_p}{y_p}. \tag{1.9}$$

Při posuzování vhodnosti modelu časové řady na prognózování pomocí charakteristik (1.6) až (1.9) lze dospět ke konkrétním závěrům, pokud tyto charakteristiky jsou vypočítány i pro některé alternativní modely (např. modely trendu, naivní modely apod.). Pomocí těchto charakteristik se usuzuje pak na vhodnost konkrétního modelu v porovnání s jinými modely. RMSE a MAPE jsou obecně široce akceptované ve statistických vývojových nástrojích. RMSE je standardní metrikou, která je vyzžívána ve výzkumné komunitě (akademiky), a MAPE metrika je spíše preferována v průmyslové sféře (praktiky). Za dobré predikční modely se považují ty, jejichž MAPE hodnoty jsou menší než 5 %.

Kapitola 2

Regresní analýza a její modely

Výchozím bodem pro studium modelů založených na regresní analýze je klasický regresní model, pomocí kterého se odhalují a modelují vztahy mezi proměnnými. Z teorie tohoto základního regresního modelu se odvíjí další novější modely, známé jako modely exponenciálního vyrovnávání, SV (Support Vector) regrese nebo model ARMA (AutoRegressive Moving Averege) a modely logistické regrese. Základnímu (klasickému) regresnímu modelu jsme se podrobně věnovali jinde (Marček, 2013). V této části opět uvedeme nezbytný teoretický základ klasického regresního modelu. Důvodem je, abychom tento přístup uměli porovnat s uvedenými novějšími modely a tak pochopili jejich vzájemné rozdíly a současně i demonstrovali, pro jaké úkoly jsou tyto modelu vhodné.

2.1 Model klasické regresní analýzy

Regresní analýza v nejobecnějším pojetí se zabývá kvantifikací (odhadem) vztahů mezi skupinami veličin. Vztahy mezi těmito veličinami mohou být složité. Ve většině případů při zjišťování těchto vztahů se vychází z relevantní teorie nebo hypotéz, na základě kterých se formuluje jednoduchý regresní model, který má tvar

$$y_t = b_0 + b_1 x_{1t} + ... + b_k x_{kt} + u_t, \tag{2.1}$$

v němž veličina y_t je eličina vysvětlovaá (závislá, endogenní) pomocí p vysvětlujících (nezávislých, exogenních) veličin $x_{1t},...,x_{p-kt}$ a $(b_0, b_1, ..., b_k)$ je vektor parametrů, jehož prvky jsou odhadované z časových řad veličin y_t a $x_{1t},...,x_{kt}$, u_t je náhodná složka modelu s vlastnostmi bílého šumu. Po odhadu parametrů modelu se model testuje na kontrolu hodnot jeho parametrů, aby hodnoty parametrů neodporovaly ekonomické teorii, tj. aby měly správně znaménka a hodnoty, aby byly stabilní v čase a aby model poskytoval přijatelné předpovědi mimo časové období, ve kterém byly parametry odhadnuty. Správnost použití modelu se testuje pomocí odhadnutých hodnot reziduí, které by měly vykazovat vlastnost bílého šumu, tj. odhadnuté reziduá nesmí vykazovat autokorelační strukturu a proměnlivý rozptyl v čase. V případě, že model

nevyhovuje nějakému testovacímu kritériu, musí být modifikován a znovu odhadnuty jeho parametry. Uvedený postup modelování ekonomických veličin je v literatuře znám jako metoda zprůměrování ekonomické regrese AER (Average Economic Regression) (Kennedy, 1992). Výhodou metody AER je, že pro odhad parametrů modelu postačuje omezený relativně malý počet pozorování časových řad veličin. Jednotlivými předtím zmíněnými kroky konstruování regresního modelu, tj. odhadem parametrů a jejich testováním, testováním modelu jako celku a testováním správnosti použití předběžné odhadové metody se zabýváme v následujícím textu.

2.1.1 Odhad parametrů regresního modelu

Nejčastěji používaným odhadovaným kritériem je minimum součtu čtverců odchylek, určených jako rozdíl mezi pozorovanými hodnotami vysvětlované veličiny a jejími vypočítanými hodnotami.

Označme vektor vypočítaných hodnot parametrů rovnice (2.1) jako $\hat{\mathbf{b}}$, dále vektor vypočítaných (teoretických) hodnot vysvětlované proměnné jako $\hat{\mathbf{y}}$. Pak můžeme teoretické hodnoty vysvětlované proměnné určit jako

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} \tag{2.2}$$

a pozorované hodnoty vysvětlované proměnné jako

$$\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + \mathbf{e}, \tag{2.3}$$

kde $\mathbf{e}$ je vektor hodnot reziduálních odchylek, vypočtených z tohoto vztahu jako

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \mathbf{y} - \hat{\mathbf{y}}. \tag{2.4}$$

Odhad parametrů modelu (2.1) metodou nejmenších čtverců (v literatuře je tato metoda označována anglickými iniciálami OLS – Ordinary Least Squares) vychází z následujících předpokladů o náhodné složce:

1. Náhodná složka modelu má normální rozdělení a má v každém pozorování nulovou střední hodnotu, takže platí

   $$E(\mathbf{u}) = \mathbf{0}, \tag{2.5}$$

   kde $\mathbf{0}$ je sloupcový vektor dimenze N se všemi nulovými prvky.

2. Rozptyl náhodné složky modelu je konstantní, je stejný v každém pozorování, tj.

   $$E = \left(u_t^2\right) = \sigma^2, \text{ pro } t = 1, 2, ..., N. \tag{2.6}$$

   Pokud předpoklad neplatí, říkáme, že náhodná složka je heteroskedastická nebo i model je heteroskedastický.

3. Hodnoty náhodné složky jsou statisticky vzájemně nezávislé. Hodnoty náhodné složky z rozdílných období jsou ortogonální (nejsou vzájemně korelovány), tj. mají nulové kovariance, což můžeme zapsat

$$E\left(u_i u_j\right) = 0, \text{ pro } j \# i. \tag{2.7}$$

Předpoklad konstantnosti rozptylu a předpoklad o nekorelovanosti hodnot náhodné složky se vyjádří variačně-kovarianční maticí Σ ve tvaru

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix}. \tag{2.8}$$

Odhadovaný výraz OLS metody pro parametry regresní rovnice (2.1) je

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.9}$$

Vektor $\hat{\mathbf{b}}$ je sloupcový vektor, jehož prvky jsou hledané odhady parametrů $\hat{b}_0,\ \hat{b}_1, \ldots,\ \hat{b}_k$.

Směrodatné odchylky vektoru parametrů jsou interpretovány jako chyby odhadu parametrů. Vypočítávají se jako

$$\sigma_{\hat{b}j} = \sigma \sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}} \text{ pro } j = 0, 1, \ldots, k. \tag{2.10}$$

Ve výrazu pro odhad směrodatných odchylek (2.10) vystupuje rozptyl náhodné složky modelu σ^2, který je obvykle neznámý. Rozptyl σ^2 náhodné složky modelu můžeme odhadnout. Označme s^2 odhad rozptylu náhodné složky σ^2. Potom odhadovaný výraz pro rozptyl náhodné složky s^2 modelu je

$$s^2 = \frac{\sum_{t=1}^{N} e_t^2}{N-(k+1)} = \frac{\mathbf{e}'\mathbf{e}}{N-(k+1)} \tag{2.11}$$

a odhadovaný výraz pro směrodatnou odchylku parametrů modelu je

$$s_{\hat{b}j} = s \sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}} \text{ pro } j = 0, 1, \ldots, k. \tag{2.12}$$

Odhad směrodatné odchylky (odhad výběrové chyby) náhodné složky $s = \sqrt{s^2}$ je charakteristikou přesností modelu.

2.1.2 Testování parametrů a intervaly spolehlivosti

Významnost jednotlivých parametrů modelu může být testována pomocí *t* testu. Následkem předpokladu o normálním rozdělení náhodné složky poměr mezi odhadovanou hodnotou parametru $\hat{b}_j$, j = 0, 1, 2, ..., k a jeho směrodatnou odchylkou $s_{\hat{b}j} = s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}$ určuje veličinu, která má Studentovo *t*-rozdělení se stupni volnosti $v = N-(k+1)$. Nulová hypotéza testu významnosti parametrů je formulovaná jako

$$H_0: b_j = 0,$$

oproti alternativě

$$H_1: b_j \neq 0.$$

Když poměr

$$\left| t_j = \frac{\hat{b}_j}{s_{\hat{b}j}} \right| > t_{\alpha,(N-k=1)}, \tag{2.13}$$

kde $t_{\alpha,(N-k=1)}$ je tabelovaná kritická hodnota Studentova rozdělení, pak parametr b_j je statisticky významný. Pokud se při testování významnosti jednotlivých parametrů modelu ukáže, že alespoň jeden z jeho parametrů je statisticky nevýznamný, je třeba tuto nevýznamnost zohlednit i při hodnocení modelu jako celku. Parametr b_0 v regresním modelu pro ekonomickou aplikaci nemá ekonomickou interpretaci. Vyjadřuje jen počáteční úroveň vysvětlované proměnné. Statistická verifikace tohoto parametru modelu nemá praktický význam, a nemá tedy vliv na celkové hodnocení významnosti modelu.

Ze vztahu (2.13) pro parametr b_j regresního modelu mohou být určeny oboustranné intervaly spolehlivosti. (1 – α) 100% interval spolehlivosti parametru b_j je

$$b_j = \hat{b}_j \pm t_{0.05[N-(k+1)]} s_{bj}. \tag{2.14}$$

Skutečná hodnota parametru b_j se bude nacházet v intervalu určeného vztahem (2.14) s pravděpodobností (1 – α).

2.1.3 Testy předpokladů aplikace odhadované metody nejmenších čtverců

Jedna z měr, kterou se posuzuje přesnost či vhodnost specifikovaného modelu k datům je koeficient determinace, označovaný jako R^2. Koeficient determinace je volně interpretován jako míra vysvětlení proměnlivosti dat vysvětlované proměnné regresním modelem. Je to míra celkové variability vysvětlované

veličiny vzhledem k její variabilitě vysvětlované pomocí modelu. Koeficient determinacee se vypočte vztahem

$$R^2 = 1 - \frac{\sum_{t=1}^{N} e^2}{\sum_{t=1}^{N} (y_t - \bar{y})^2}. \tag{2.15}$$

Na bázi koeficientu determinace je založen i test významnosti jako celku. Statistika, kterou se ověřuje významnost regresního modelu jako celku, je označována jako $F_{R,}$ která je určena vztahem

$$F_R = \frac{\sum_{t=1}^{N} (y_t - \bar{y})^2 / k}{(\sum_{t=1}^{N} e^2)/[N-(k+1)]}. \tag{2.16}$$

Statistika F_R má *F*- rozdělení se stupni volnosti *k* a [*N* – (*k* + 1)]. Aby byl koeficient korelace statisticky významný a tím potvrdil statistickou významnost modelu jako celku, hodnota veličiny F_R vypočtená vztahem (2.16) musí mít větší hodnotu, než je její teoretická kritická hodnota na hladině významnosti α a při daných stupních volnosti. Pro významnost modelu jako celku musí platit

$$F_R > F_{\alpha,k,[N-(k+1)]}, \tag{2.17}$$

kde $F_{\alpha,k,[N-(k+1)]}$ je tabelovaná teoretická kritická hodnota na hladině významnosti α při stupních volnosti *k* a [*N* – (*k* + 1)].

Předpoklady o náhodné složce u_t modelu (2.1), které jsme definovali vztahy (2.5) až (2.7), jsou jen teoretickými předpoklady, protože hodnoty náhodné složky nemůžeme měřit. Testování teoretických předpokladů je založeno na zkoumání reziduí e_t, které po kvantifikaci modelu můžeme určit jako $e_t = y_t - \hat{y}_t$.

Test náhodné složky na normální rozdělení. Regresní model předpokládá, že jeho náhodná složka má normální rozdělení s nulovou střední hodnotou. Posouzení, zda náhodná složka má normální rozdělení, lze provést na základě grafického průběhu reziduí e_t v závislosti na normovaných reziduích. Normovaná rezidua získáme tak, že rezidua e_t podělíme jejich směrodatnou odchylkou. Pokud má náhodná složka normální rozdělení, průběh v závislosti reziduí od normovaných reziduí musí být přibližně přímočarý. Ilustraci tohoto grafického průběhu (tzv. QQ plot) možno nalézt např. v práci (Marček a kol. (2009).

Testování na normální rozdělení náhodné složky je formálně podloženo na Jarque–Beraovy testovací statistice χ^2, jejíž hodnota má χ^2 rozdělení s 2 stupni volnosti. Hodnota Jarque–Beraovy statistiky se vypočítá výrazem

$$\chi^2 = \frac{N-k}{6}\left[S^2 + \frac{1}{4}(K-3)\right], \tag{2.18}$$

kde *N* je počet pozorování, *k* je počet vysvětlujících proměnných, *S* je šikmost (čtvrtý centrální moment normovaných reziduí), *K* je špičatost. Nulová hypotéza

předpokládá, že rezidua jsou rozdělená, alternativní hypotéza zamítá nulovou hypotézu. Pokud vypočítaná veličina χ^2 podle vztahu (2.18) je větší než tabelovaná $\chi^2_{\alpha,2}$ hodnota na hladině významnosti α, nulovou hypotézu zamítneme, rezidua nemají normální rozdělení.

Předpoklad (2.7) o náhodné složce modelu vyjadřuje, že náhodná složka regresního modelu je náhodnou veličinou pouze tehdy, pokud její hodnoty z různých pozorování nejsou vzájemně závislé. Pokud by byly hodnoty náhodné složky vzájemně závislé, mělo by to za následek podhodnocení směrodatných odchylek parametrů modelu. Vedlo by to ke zvýšení hodnot testovacích veličin na významnost parametrů a v konečném důsledku k chybným optimistickým závěrům o dobré modelové aproximaci dat časové řady vysvětlované proměnné. Především modely založené na časových řadách často porušují předpoklad o nekorelovanosti náhodných složek modelu.

Pokud není splněn předpoklad, že kovariance náhodné složky modelu a obvykle i kovariance reziduí nejsou nulové, variačně-kovarianční matice Σ podle (2.8) náhodné složky modelu není diagonální (všechny nediagonální prvky této matice nejsou nulové). Hodnoty náhodné složky modelu mohou být v tomto případě různým způsobem korelované. Důvodem k tomu může být např. nezahrnutí některé proměnné při specifikaci modelu, časově opožděný efekt přechodových vlivů, volba nesprávného funkčního tvaru modelu, chyby pozorování apod.

Nejčastěji používaným testem zjišťování autokorelace prvního řádu je statistika *d* Durbin–Watsonova testu (Durbin a Watson, 1950), která je definována jako

$$d = \frac{\sum_{t=2}^{n} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2}. \tag{2.19}$$

Durbin–Watsonův test umožňuje statisticky určit, zda existuje nebo neexistuje (je přítomna nebo není přítomna) autokorelace reziduí. Kritické hodnoty statistiky byly tabelované a jejich význam s praktickým příkladem čitatel najde např. v práci (Marček a Marček, 2001).

Pokud jde o testování a o vzájemné srovnávání vhodnosti modelů, ve kterých se používá odhadová metoda ML (Maximum Likelihood), tj. u nelineárních modelů, využívají se testy založené na log likelihood LL-skóre, resp. LL-ratio testu, *p-hodnotách,* nebo *p*(*F*) *hodnotě*. Detaily a aplikaci k této problematice možno nalézt v práci Marček (2014).

2.1.4 Konstrukce prognóz a predikční intervaly

Regresní modely časových řad se často používají k určení bodových předpovědí, tj. pro odhad hodnot vysvětlované veličiny v období $p = N + 1, N + 2, \ldots$. Bodové

předpovědi, resp. jejich predikční intervaly je možné určit, pokud je model kvantifikován, pokud lze předpokládat stabilitu parametrů modelu a pokud jsou k dispozici očekávané hodnoty vysvětlujících proměnných modelu. Za těchto předpokladů lze vypočítat bodové prognózy vysvětlované veličiny jako

$$\hat{y}_p = \mathbf{x}'_p \hat{\mathbf{b}} \text{ pro } p = N + 1, N + 2, ..., \tag{2.20}$$

kde $\mathbf{x}'_p$ je vektor $k + 1$ vysvětlujících proměnných v obdobích p, $\hat{\mathbf{b}}$ je vektor odhadovaných parametrů modelu. Vidíme, že určování prognóz, tj. prognózování vysvětlované veličiny, znamená určování jejích hodnot na období následující po období kvantifikace.

Pokud ve vztahu (2.20) dosadíme za $\hat{\mathbf{b}}$ jeho odhadový výraz, lze vypočítat bodové prognózy vysvětlované veličiny jako

$$\hat{y}_p = \mathbf{x}'_p \hat{\mathbf{b}} = \mathbf{x}'_p \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}, \tag{2.21}$$

kde $\mathbf{X}$ je matice pozorovaných hodnot všech vysvětlujících proměnných v obdobích $t = 1, 2, ..., N$ rozměru $N \times (k + 1)$. Symbol y vyjadřuje sloupcový vektor pozorovaných hodnot vysvětlované veličiny rozměru $N \times 1$. Skutečnou hodnotu vysvětlované proměnné v období $p = N + 1, N + 2$, označíme y_p. Pak chybu prognózy e_p určíme jako rozdíl skutečné hodnoty vysvětlované proměnné a vypočtené prognózy, tj.

$$e_p = y_p - \hat{y}_p. \tag{2.22}$$

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

$$\sigma_p^2 = \mathbf{x}'_p \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p + \sigma^2, \tag{2.23}$$

kde σ^2 je skutečný rozptyl náhodné složky modelu. Skutečný rozptyl náhodné složky modelu neumíme vyčíslit, nahradíme ho odhadem s^2, podle výrazu daný výrazem (2.11), tj.

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

Po dosazení odhadu σ^2 jako s^2 do (2.23) označme odhad rozptylu chyby bodové prognózy σ_p^2 jako s_p^2, tj.

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

Po úpravě posledního výrazu je odhad rozptylu chyby bodové prognózy daný výrazem

$$s_p^2 = s^2 \left[1 + \mathbf{x}_p' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p\right]. \tag{2.24}$$

Směrodatná odchylka chyby předpovědi je daná odmocninou z jejího rozptylu, tj.

$$s_p = s\sqrt{1 + \mathbf{x}_p' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p}. \tag{2.25}$$

Pokud je známý odhad směrodatné odchylky chyby prognózy, za předpokladu normálního rozdělení chyb prognóz, můžeme vypočítat 100 $(1 - \alpha)$ procentuální predikční interval bodové prognózy $\hat{y}_p$ jako

$$\hat{y}_{p\min}^{\max} = \hat{y}_p \pm t_{\alpha,N-(k+1)} s_p = \hat{y}_p \pm t_{\alpha,N-(k+1)} s\sqrt{1 + \mathbf{x}_p' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_p}. \tag{2.26}$$

2.2 Model logistické regrese

V současnosti se lze stále častěji setkat s průnikem metody logistické regrese do ekonomiky, manažerských věd a do oblasti umělé inteligence jako jedné z těžiskových metod získávání znalostí z dat. Obecně se ve většině případů modely logistické regrese používají pro datovou analýzu a jako inferenční nástroj pro pochopení úlohy vstupních veličin pro vysvětlení výstupu. Při vysvětlování podstaty modelu se omezíme na model jednoduché závislosti mezi kvalitativními znaky a na model vícenásobné závislosti z publikace (Marček, 2009). Budeme se věnovat analytickým formám modelů logistické regrese, jejich analytickým formám odhadu parametrů, testování významnosti parametrů a testování významnosti modelů jako celku, přičemž budeme upozorňovat na jejich analogii, příp. odlišnost od standardního modelu regresní analýzy, s kterým jsme se zabývali v předcházející kapitole. Vzhledem k podobnosti standardního regresního modelu s modelem logistické regrese budeme pro větší orientaci v dalším textu označovat parametry standardního modelu regresní analýzy písmeny, $b_0, b_1, \ldots$, příp. jako vektor **b** a parametry modelů logistické regrese jako, $\beta_0, \beta_1, \ldots$, příp. vektorově jako **β**.

Podobně jako v jednoduchém standardním regresním modelu s jednou nezávislou (vysvětlující) veličinou (x) a výstupní (závislou) veličinou (y) i v jednoduchém modelu logistické regrese (MLR) jde o analytické vyjádření vztahu mezi těmito dvěma veličinami na základě pozorování dat $\{y_t, x_t\}$, t = 1, 2, ..., n těchto veličin ve tvaru rovnosti (2.1), tj. $y_t = b_0 + b_1 x_t + u_t$, resp. $f(x_t) = b_0 + b_1 x_t + u_t$ s tím rozdílem, že v MLR hodnoty závislé veličiny (y), popř. i

nezávislých veličin, jsou dichotomické. Budeme předpokládat, že veličina y_t může mít jen dvě hodnotové obměny většinou označované jako 1 nebo 0. Vysvětlování principů MLR uvedeme na následujícím příkladu.

Příklad 2.1

Pro ilustraci MLR předpokládejme, že ve finanční instituci v rámci sociologického výzkumu byla pro 50 pracovníků provedena anketa o sebeuplatnění ve vykonávané práci (y) v závislosti na věku zaměstnance (x). V rámci anketového zjišťování zaměstnanci mohli odpovědět, buď že jsou spokojeni se sebeuplatnění ($y_t = 1$), nebo jsou nespokojeni ($y_t = 0$). Výsledek tohoto šetření se setříděnými odpověďmi podle věku zaměstnanců je zaznamenán v tabulce 2–1 o spokojenosti sebeuplatnění zaměstnanců s vykonávanou prací (SVP).

Pokud výsledky odpovědí na otázku sebeuplatnění zaměstnanců v závislosti na věku vyneseme do grafu, kde osa *x* reprezentuje věk zaměstnanců a osa *y* hodnoty odpovědí, získáme graf na obrázku 2–1. Na obrázku 2–1 je bodový graf, v němž na úrovni 1 jsou vyznačeny body s odpověďmi spokojenost a na úrovni 0 jsou body reprezentující odpovědi vyjadřující nespokojenost se sebeuplatněním. Z tohoto grafu je téměř nemožné vyjádřit závislost sebeuplatnění zaměstnanců od jejich věku, protože variabilita ve všech věkových kategoriích je velká.

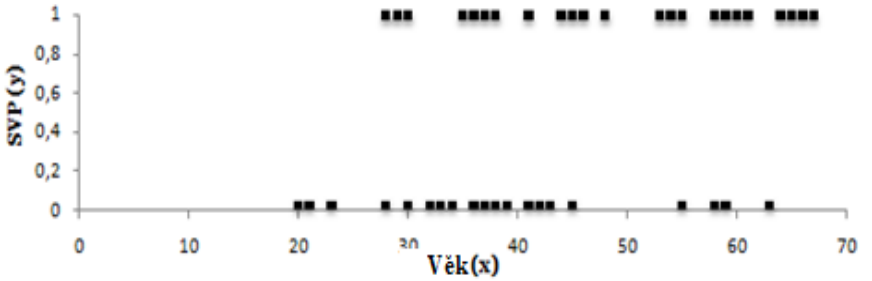
Na pravé straně tabulky 2–2 je nakreslen bodový graf závislosti podílu SVP (y) v příslušné věkové skupině (x). Jak je z tohoto grafu vidět, tato závislost není lineární. Jednotlivé body v grafu ukazují na závislost ve tvaru logistické křivky (S-křivky). Střední hodnota závislé veličiny Y v jednotlivých bodech tohoto grafu je podmíněna nezávislou veličinou, v našem případě *x*, což se zapisuje jako $E(Y \mid x)$. Pro zjednodušení dalšího výkladu tento zápis podmíněné střední hodnoty budeme označovat jako Π(x). Ve standardním jednoduchém lineárně regresním modelu se regresní vztah vyjadřuje vzorcem

$$\hat{y} = b_0 + b_1 x, \quad x \in (-\infty, \infty).$$

Z průběhu logistické křivky v tabulce 2–2 je dále vidět, že podmíněná střední hodnota $\Pi(x)$ pro binární hodnoty je ohraničena intervalem $0 \leq \Pi(x) \leq 1$, a současně tato křivka připomíná i kumulativní rozdělení pravděpodobnosti náhodné veličiny pro model $\Pi(x)$ v případě, že hodnoty *Y* jsou dichotomické. Pro analýzu dichotomické závislé veličiny, kromě logistické kumulované pravděpodobnosti, lze použít i jiné rozdělení. Avšak právě kumulativní rozdělení ve tvaru logistické funkce poskytuje větší flexibilitu a smysluplnou interpretaci ve většině aplikací (Hosmer a Lemenshow, 1989).

Nejčastější tvar vyjádření závislosti logistické regrese je dán výrazem

$$\Pi(x_t) = \frac{e^{\beta_0 + \beta_1 x_t}}{1 + e^{\beta_0 + \beta_1 x_t}}. \tag{2.27}$$



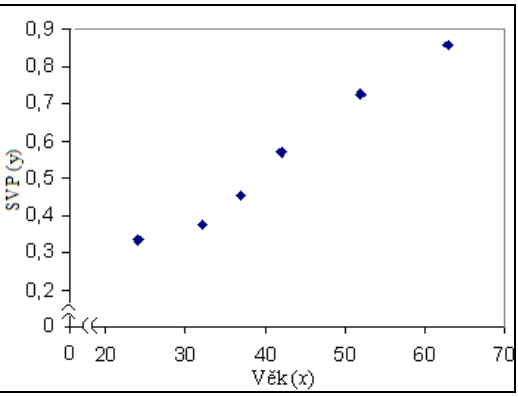
Obrázek 2–1 Graf SVP (spokojenost: $y = 1$, nespokojenost: $y = 0$) v závislosti od věku pracovníků (x)

Tabulka 2–1 Odpovědi na sebeuplatnění zaměstnanců (1 – spokojenost, 0 – nespokojenost)

P. č.	Věk	SVP	P. č.	Věk	SVP
1	20	0	26	41	1
2	21	0	27	41	0
3	23	0	28	42	0
4	28	1	29	43	0
5	28	0	30	44	1
6	29	1	31	45	1
7	30	0	32	45	1
8	30	1	33	46	0
9	32	0	34	48	1
10	32	1	35	53	1
11	33	0	36	54	1
12	34	0	37	55	1
13	35	0	38	55	0
14	35	1	39	58	1
15	36	1	40	58	1
16	36	0	41	58	0
17	37	1	42	59	1
18	37	0	43	59	0
19	37	0	44	60	1
20	37	1	45	61	1
21	38	1	46	63	0
22	38	0	47	64	1
23	38	1	48	65	1
24	39	0	49	66	1
25	39	0	50	67	1

Tabulka 2–2 Kontingenční (frekvenční) tabulka variant znaku sebeuplatnění ve vykonávané práci (SVP) podle věkových skupin (vlevo) a bodový graf hodnot SVP (označené jako *y*) v závislosti od středy věkových intervalů (*x*) je vpravo

		SVP		(y)
Věková skupina	Počet *n*	S	N	Podíl: S/*n*
20–29	6	2	4	0,333
30–35	8	3	5	0,376
36–39	11	5	6	0,455
40–45	7	4	3	0,571
46–59	11	8	3	0,727
60–67	7	6	1	0,857
Spolu:	50	28	22	0,56



S – spokojenost, N – nespokojenost

Potom model logistické regrese má tvar

$$y_t = \Pi(x_t) + u_t,\ t = 1, 2, ..., n. \tag{2.28}$$

Další podstatný rozdíl mezi standardním regresním modelem (2.1) a modelem logistické regrese (2.28) tkví v předpokladech o vlastnostech náhodné složky (chybového člena). Ve standardním regresním modelu (2.1) se o náhodné složce *u* předpokládá, že má normální rozdělení s nulovou střední hodnotou a konstantním rozptylem ve všech pozorováních nezávislé veličiny. V případě modelu logistické regrese (2.28) při dichotomických hodnotách dat výstupu y_t může náhodná složka u_t nabývat pouze dvou hodnot. Pokud $y_t = 1$, pak přímo z modelu (2.28) vyplývá, že $u_t = 1 - \Pi(x_t)$ s pravděpodobností $\Pi(x_t)$, a pokud $y_t = 0$, pak $u_t = -\Pi(x_t)$ s pravděpodobností $1 - \Pi(x_t)$. Je vidět, že v MLR má náhodná složka také nulovou střední hodnotu, ale rozptyl je dán výrazem

$$\sigma_t^2 = \Pi(x_t)\,[1 - \Pi(x_t)], \tag{2.29}$$

a tedy podmíněné rozdělení výstupní veličiny má binomické rozdělení s pravděpodobností $\Pi(x)$.

Velmi důležitou transformací logistické funkce (2.27) je tzv. *logit transformace*. Jak je z tvaru logistické funkce (2.27) vidět, logistická funkce $\Pi(x)$ není lineární s ohledem na její parametry β_0 a β_1. Logit transformace nebo zkráceně jen logit $g(x)$ je definován následujícím výrazem

$$g(x) = \ln\left(\frac{\Pi(x)}{1-\Pi(x)}\right) = \beta_0 + \beta_1 x. \tag{2.30}$$

Je vidět, že funkce logit je již lineární s ohledem na parametry β_0 a β_1. Její hodnoty mohou být spojité v intervalu $(-\infty, \infty)$ v závislosti na hodnotě nezávislé veličiny *x*. Je třeba si uvědomit, jak je to z výrazu (2.30) vidět, že logit je funkcí závislé veličiny MLR, které nahrazuje lineární funkce nezávislých veličin. Proto se v MLR tato funkce nazývá také *spojovací (link) funkcí*.

Logit trnsformací se model logistické regrese v mnohém přiblížil vlastnostem modelu standardní regrese. Tím i metody, které doprovázejí standardní lineární regresi, budou s určitými modifikacemi aplikovatelné i pro logistickou regresi. Platí to i pro metody odhadu parametrů modelu logistické regrese, kterými se budeme zabývat v další podkapitole.

2.2.1 Odhad parametrů

Odhad parametrů β_0 a β_1 modelu logistické regrese (2.27) je v porovnání se standardním lineárním regresním model složitější, protože model logistické regrese je modelem nelineárním v parametrech. Není možné zde pro odhad parametrů použít jednoduchou metodu nejmenších čtverců (Ordinary Least Squares – OLS), tj. takový odhad parametrů β_0 a β_1, kterým se minimalizuje suma čtverců odchylek pozorovaných hodnot Y od odhadnutých hodnot produkovaných tímto modelem. V předcházející podkapitole jsme uvedli, že estimátor OLS pro standardní lineární regresní model má žádoucí vlastnosti, pokud jde o nevychylenost, vydatnost a konzistentnost.

Nejvíce používanou metodou pro odhad parametrů nelineárních modelů za předpokladu, že chybový člen u_t modelu má normální rozdělení, tedy i modelu logistické regrese, je metoda maximální věrohodnosti (Maximum Likelihood – ML). ML metodou se určí takové hodnoty parametrů β_0 a β_1 modelu, kterými se maximalizuje pravděpodobnost získání pozorovaných dat. Metoda ML je založená na vytvoření funkce věrohodnosti *L* a hledají se takové parametry modelu, aby tato pravděpodobnost byla největší, tj. hledají se takové parametry, při kterých dosáhne *L* maxima.

V logistickém regresním modelu budeme stále předpokládat, že závislá veličina je dichotomická dvouhodnotová s hodnotami nula nebo jedna. Pak pro danou hodnotu nezávislé veličiny *x* výraz (2.27) pro $\Pi(x)$ reprezentuje pravděpodobnost, že hodnota *y* je rovna jedné, což se zapisuje $P(y=1|x)$. Opačně, výraz 1 – Π(x) reprezentuje podmíněnou pravděpodobnost, že veličina *y* nabude hodnotu nula pro dané *x*, tj. $P(y=0|x)$. Pro danou dvojici pozorování (x_t, y_t), v němž $y_t = 1$, je příspěvek pro hodnotu funkce věrohodnosti $\Pi(x_t)$, a tu dvojici pozorování, kde $y_t = 0$, je příspěvek pro hodnotu funkce věrohodnosti $1 - \Pi(x_t)$, kde $\Pi(x_t)$ reprezentuje hodnotu $\Pi(x_t)$ vypočtenou v bodě x_t. Potom příspěvek $\xi(x_t)$ pro funkci věrohodnosti *L* od jedné dvojice pozorování je

$$\xi(x_t) = \Pi(x_t)^{yt}\left[1 - \Pi(x_t)\right]^{1-yt}. \tag{2.31}$$

Za předpokladu, že jednotlivá pozorování jsou nezávislá, se funkce věrohodnosti $L(\beta_0, \beta_1)$ pro model logistické regrese získá jako součin příspěvků za všechna pozorování dat, tj.

$$L(\beta_0, \beta_1) = \prod_{t=1}^{n} \xi(x_t) = \prod_{t=1}^{n} \Pi(x_t)^{yt}\left[1 - \Pi(x_t)\right]^{1-yt}. \tag{2.32}$$

Její logaritmus $\ell(\beta_0, \beta_1)$ (log likelihood) má tvar

$$\ell(\beta_0, \beta_1) = \ln\left[L(\beta_0, \beta_1)\right] = \sum_{t=1}^{n}\left\{y_t \ln\left[\Pi(x_t)\right] + (1-y_t)\ln\left[1 - \Pi(x_t)\right]\right\}. \tag{2.33}$$

Postavíme-li první derivace funkce (2.33) podle β_0 a β_1 rovny nule, získáme dvě normální rovnice, také nazývané věrohodnostní (likelihood) rovnice pro odhad parametrů β_0 a β_1. Tyto rovnice jsou:

$$\sum_{t=1}^{n}\left[y_t - \Pi(x_t)\right] = 0 \tag{2.34}$$

a

$$\sum_{t=1}^{n} x_t\left[y_t - (x_t)\right] = 0. \tag{2.35}$$

Řešení posledních dvou rovnic poskytne odhady (estimátory) maximální věrohodnosti pro parametry β_0 a β_1 modelu (2.27). Analogicky jako v standardním regresním modelu (2.1) tyto odhady označíme se stříškou, tj. $\hat{\beta}_0$ a $\hat{\beta}_1$. Jak jsme se již zmínili, estimátory maximální věrohodnosti parametrů jsou konzistentní a asymptoticky vydatné. Pokud je počet pozorování dostatečně velký, pak estimátory $\hat{\beta}_0$ a $\hat{\beta}_1$ lze považovat za normálně rozdělené se střední hodnotou $E(\hat{\beta}_0) = \beta_0$ a $E(\hat{\beta}_1) = \beta_1$.

Jestli ve výrazu pro vyjádření závislosti logistické regrese (2.27) skutečné hodnoty parametrů β_0 a β_1 zaměníme za jejich odhady maximální věrohodnosti $\hat{\beta}_0$ a $\hat{\beta}_1$, získáme odhad hodnot $\hat{\Pi}(x_t) = \dfrac{e^{\hat{\beta}_0+\hat{\beta}_1 x_t}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_t}}$, který současně poskytuje odhad podmíněné pravděpodobnosti veličiny Y = 1, pro $x = x_t$ založený na metodě maximální věrohodnosti, co je zároveň vyrovnaná, resp. predikovaná hodnota výstupu pro model logistické regrese (2.28). Logický důsledek, který

vyplývá z rovnice (2.34) je, že suma pozorovaných hodnot y_t je rovná sumě odhadovaných (předikovaným) hodnot $\hat{\Pi}(x_t)$, tj.

$$\sum_{t=1}^{n} y_t = \sum_{t=1}^{n} \hat{\Pi}(x_t). \tag{2.36}$$

Pro řešení nelineárních rovnic (2.34) a (2.35) založených na funkci věrohodnosti neexistuje všeobecně explicitní výraz. Minimalizace součtu čtverců odchylek se provádí některou iterační metodou, nejčastěji metodou iterační linearizace, např. metodou Gauss–Newtonovou s Marquardovou modifikací (Marquard, 1963). Její předností je to, že je založena na posloupnosti lineárních odhadů linearizovaných rovnic pomocí Taylorova rozvoje. To umožňuje použít statistické testy aplikované při lineárním odhadu jako určité aproximace testovacích postupů. V Gauss–Newtonově metodě iteračního postupu lze použít známé testovací statistiky, jako jsou: koeficient determinace, *t*-test, *F*-test pro ověřování kvality linearizonaného modelu. Proto také většina statistických a ekonometrických programových systémů realizuje nelineární odhad metodou iterační linearizace. Součástí výstupů z nich jsou i hodnoty směrodatných odchylek odhadnutých parametrů, příp. poměr odhadnutých hodnot parametrů a jejich směrodatných odchylek ($\hat{\beta}_0 / \hat{\sigma}_{\hat{\beta}_0}$ a $\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}1}$). Tyto statistiky se používají při testování významnosti parametrů modelů, což je uvedeno v další podkapitole.

V literatuře existují i jiné postupy (metody) odhadu parametrů modelu logistické regrese. Jeden postup založený na diskriminační funkci (6.1) uvedeme podle (Hosmer a Lemenshow, 1989). Předpokládejme, že nezávislá veličina x pochází z normálního rozdělení v rámci každé skupiny dvou výběrů dvojhodnotově definované veličiny. Každá skupina výběrů má rozdílné střední hodnoty a stejný rozptyl. Potom podmíněné rozdělení veličiny Y pro X = x je logistický regresní model. Formálně to lze zapsat takto: pokud X|Y = $j \sim N(\mu_j, \sigma^2)$, j = 1, 2, potom $P(Y = 1 | x) = \Pi(x)$.

Při těchto podmínkách lze vyjádřit parametry modelu logistické regrese jako

$$\beta_0 = \ln\left(\frac{\theta_2}{\theta_1}\right) - 0{,}5\,(\mu_2 - \mu_1)^2 / \sigma^2 \tag{2.37}$$

a

$$\beta_1 = (\mu_2 - \mu_1)^2 / \sigma^2, \tag{2.38}$$

kde $\theta_{j+1} = P(Y = j)$, j = 0, 1. Estimátory pro parametry β_0 a β_1 založené na diskriminační funkci se určí tak, že ve výrazech (2.37) a (2.38) se za μ_{j+1}, θ_{j+1}, a σ^2 dosadí jejich odhady. Obvykle se za μ_{j+1} dosadí aritmetické průměry veličiny

Tabulka 2–3 Možná volba kódování příslušnosti obyvatel pro tři úrovně hodnot

Příslušnost obyvatelů	Umělé proměnné	
	D_1	D_2
Městská	0	0
Venkovní	1	0
Jiná	0	1

x ve skupinách definovaných hodnotami závislé veličiny $y = j$, j = 0, 1, $\hat{\theta}_1 = n_1 / n$ a $\hat{\theta}_0 = 1 - \hat{\theta}_1$. Odhad rozptylu $\hat{\sigma}^2$ se určí podle vztahu.

$$\hat{\sigma}^2 = \left[(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 \right] / (n_1 + n_2 - 2),$$

který koresponduje s výrazem (6.10) v kapitole o diskriminační analýze.

Jednoduchou logistickou závislost ve tvaru výrazu (2.27) s modelem podle výrazu (2.28) lze analogicky přímočaře rozšířit podle standardního vícenásobného regresního modelu na vícenásobný model logistické regrese, tj. pro k nezávislých veličin, které zapíšeme jako vektor $\mathbf{x}_t^T = (1, x_{1t}, x_{2t}, ..., x_{kt})$, t = 1, 2, ..., n. Nejprve budeme předpokládat, že každá nezávislá veličina je zadána intervalovými rozsahy hodnot. Podmíněná pravděpodobnost, že závislá (výstupní) veličina Y nabývá hodnoty 1 za předpokladu vstupu $\mathbf{x}_t$, je $P(Y = 1 | \mathbf{x}_t) = \Pi(\mathbf{x}_t)$ a logit transformace je dána výrazem

$$g(\mathbf{x}_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} = \mathbf{x}_t^T \boldsymbol{\beta}, \tag{2.39}$$

kde $\Pi(\mathbf{x_t}) = \dfrac{e^{g(\mathbf{x}_t)}}{1 + e^{g(\mathbf{x}_t)}}$, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_k)$, $\mathbf{x}_t^T = (1, x_{1t}, x_{2t}, \ldots, x_{kt})$, pro t = 1, 2, ..., n.

Předpokládejme dále, že některé nezávislé veličiny jsou diskrétní a svou povahou kvalitativní veličiny, např. profesionální příslušnost pracovníků, stav, zacházení, politická příslušnost, postoje atd., a jsou vyjádřeny v binárními nominálně (kvalitativně) škálovými hodnotami. V takových případech se používá umělá nezávislá veličina s umělými (dummy) proměnnými. Například veličina příslušnost obyvatel může být kódována třemi proměnnými jako: městská, venkovská nebo jiná. Obecně platí, pokud nominálně škálovaná veličina má d možných hodnot, pak pro jejich zakódování v MLR je potřebných d – 1 umělých proměnných. Např. pro kódování příslušnosti obyvatel pro tři úrovně hodnot jsou zapotřebí dvě umělé proměnné D_1 a D_2 s kódováním podle tabulky 2–3.

V poslední tabulce je vidět možný (nikoli jediný) způsob kódování. Pokud např. respondent patří do skupiny městských obyvatel, obě umělé proměnné jsou kódovány nulami. Pokud respondent patří do skupiny venkovských obyvatel, D_1 má hodnotu 1 a D_2 má hodnotu 0. Pokud respondent nepatří ani do městského ani do venkovského obyvatelstva, tak $D_1 = 0$, $D_2 = 1$. Vzhledem k uvedené možnosti

kódování nezávislých umělých proměnných označíme parametry stojící při *j*-té umělé proměnné symbolem β_{jd}. Pak logit pro logistický model, vzhledem k existenci *j*-té diskrétní veličiny zakódované prostřednictvím $d_j - 1$ umělých proměnných, má tvar

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ + \beta_k x_k + \sum_{d=1}^{d_{j-1}} \beta_{jd} D_{jd}. \qquad (2.40)$$

Ve výrazu (modelu) (2.40) jsou vynechány indexy '*t*' označující jednotlivá pozorování veličin.

Pro odhad parametrů modelu (2.40) jednotlivé parametry a nezávislé veličiny vyjádříme jako vektory ve tvaru $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_k, \beta_{j1}, \beta_{j2}, ...)$, $\mathbf{x}_t^T = (1, x_{1t}, x_{2t}\ \ldots\ x_{kt}, D_{j1t}, D_{j2t}, ...)$. Je vidět, že umělými proměnnými se zvětší počet parametrů modelu. Umělé proměnné můžeme také označit symboly *x*, se *zvýšením* hodnoty *k* počtu nezávislých veličin zahrnutých do modelu. Analogicky to platí i pro symboly parametrů modelu. Předpokládejme, že k dispozici jsou příslušná pozorování nezávislé veličiny, hodnoty kódování umělých proměnných a hodnoty závislé veličiny ($\boldsymbol{x}_t$, y_t), $t = 1, 2, ..., n$, přičemž jednotlivá pozorování jsou navzájem nezávislá. Potom logaritmus funkce věrohodnosti $\ell(\boldsymbol{\beta})$ v (2.33) je analogický jako v případě jednoduchého modelu logistické regrese s jednou nezávislou veličinou s tím rozdílem, že regresní závislost má v tomto případě tvar výrazu (2.39). Postavíme-li první derivace logaritmu funkce věrohodnosti podle $\boldsymbol{\beta}$ rovné nule, získáme $k + 1$ normálních rovnic. Tyto rovnice mají tvar

$$\sum_{t=1}^{n} (y_t - \boldsymbol{\Pi}(\mathbf{x_t})) = \mathbf{0}$$

a (2.41)

$$\sum_{t=1}^{n} \mathbf{x}_{tj} \left[y_t - \Pi(\mathbf{x}_t) \right] = 0$$

pro $j = 1, 2, ..., k$.

Odhadnuté hodnoty vícenásobného modelu logistické regrese jsou dány následujícím výrazem

$$\hat{\Pi}(\mathbf{x}_t) = \frac{e^{g(\mathbf{x}_t)}}{1 + e^{g(\mathbf{x}_t)}},$$

kde se hodnoty logit transformace $g(\mathbf{x}_t)$ podle (2.39), resp. (2.40) vypočítají jako $\mathbf{x}_t^T \hat{\boldsymbol{\beta}}$, pro $t = 1, 2, ..., n$.

2.2.2 Testování významnosti parametrů

Při posuzování kvality odhadnutých parametrů, podobně jako ve standardním regresním modelu, tak v logistické regresi, jsou důležité informace o směrodatných odchylkách těchto parametrů. V případě odhadu parametrů nelineárního MLR a testování jejich významnosti zde přistupují navíc informace o hodnotách funkce věrohodnosti (likelihood function).

Základní princip testování významnosti parametrů standardního i logistického regresního modelu je založen na porovnání přesnosti odhadnutých (vypočítaných, vyrovnaných) hodnot závislé veličiny vypočítaných modelem, v němž je zahrnuta testovaná nezávislá veličina modelu a modelu, v němž je testovaná veličina vynechána. Zatímco ve standardním lineárně regresním modelu je toto porovnání založeno na analýze sumy čtverců odchylek, v modelu logistické regrese porovnání pozorovaných a vyrovnaných hodnot je založen na log funkci věrohodnosti (2.33). K tomu se využívá *D* statistika, která se vypočte jako

$$D = -2 \ln \left[\frac{\text{hodnota log funkce věrohodnosti aktuálního modelu}}{\text{hodnota log funkce věrohodnosti saturovaného modelu}} \right], \quad (2.42)$$

přičemž pod pojmem *saturovaný model* se rozumí model, který obsahuje tolik parametrů (nezávislých veličin), kolik je počet pozorování. Např. pro regresní model s dvěma nezávislými veličinami, pokud existují jen 2 pozorování.

Zlomek v hranatých závorkách ve výrazu (2.42) se nazývá poměr hodnot funkcí věrohodnosti (likelihood ratio). Pokud se příslušné hodnoty funkcí věrohodnosti vyjádří ve tvaru zápisu (2.33), pak po úpravě lze pro hodnotu *D* statistiku napsat tvar

$$D = -2\sum_{t=1}^{n} \left[y_t \ln\left(\frac{\hat{\Pi}(x_t)}{y_t} \right) + (1-y_t) \ln\left(\frac{1-\hat{\Pi}(x_t)}{1-y_t} \right) \right]. \quad (2.43)$$

Výraz (2.43) v MLR regresi je analogií kritéria, jakým je suma čtverců nevysvětlených odchylek ($\sum (y_t - \hat{y}_t)^2$) ve standardním lineárně regresním modelu.

Poslední důležitou statistikou, označovanou jako *G* statistika, pro testování významnosti parametru modelu, a tedy i významnosti zařazení veličiny stojící při tomto parametru do MLR je statistika, která je odvozena z výrazu (2.42) jako

$$G = -D \text{ (pro model s vynecháním posuzované veličiny)}$$
$$-D \text{ (pro model se zařazením posuzované veličiny)}, \quad (2.44)$$

což je ekvivalentní s výrazem

$$G = -2(L_0 - L_1) \quad (2.45)$$

kde L_0 maximalizovaná logaritmovaná funkce s omezením na $\beta_k = 0$ a L_1 je maximalizovaná logaritmovaná funkce bez omezení, která je ohodnocena na **β.**

Nulová hypotéza testu významnosti parametru β_1 daným nulovou hypoézou a $L_1\beta_1$ je formulována jako H_0: $\beta_1 = 0$. Pak statistika G má χ^2 rozdělení s 1 stupněm volnosti.

Významnost parametru β_1 založeného na odhadu ML je možné testovat Waldovym testem, který je označován jako *W*-test. Hodnota Waldova testu se vypočítá jako

$$W = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}, \tag{2.46}$$

kde $\hat{\sigma}_{\hat{\beta}_1}$ je odhad směrodatné odchylky parametru $\hat{\beta}_1$.

V případě jednoduchého MLR, ale především v případě většího počtu nezávislých veličín lze pro testování významnosti parametru β_1 použít i Score-test (ST), pro jehož výpočet nejsou potřebné výstupy z ML. Testovací statistika pro Score-test má tvar

$$\mathrm{ST} = \frac{\sum_{t=1}^{n} x_t \left(y_t - \bar{y}\right)}{\sqrt{\bar{y}\left(1-\bar{y}\right)\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2}}. \tag{2.47}$$

Pokud jde o testování významnosti parametrů vícenásobného logistického regresního modelu, i v tomto případě je možné testovat významnost modelu jako celku (významnost všech parametrů modelu). Testování významnosti modelu jako celku s *k* parametry je založeno na *D* a *G* statistických kritériích, daných výrazy (2.42) a (2.44).

Pro testování významnosti jednotlivých parametrů vícenásobného logistického regresního modelu odhadnutých metodou ML je možno využít Waldovu testovací statistiku *W*. Její hodnota pro *j*-tý parametr je dána v souladu s výrazem (2.46), jako

$$W = \frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}}. \tag{2.48}$$

Za předpokladu platnosti nulové hypotézy H_0: $\beta_j = 0$, W statistika má standardní normální rozdělení. Podobně jako v případě testování jednotlivých parametrů modelu přibližné rozhodování či jednotlivý parametr je nevýznamný, lze uplatnit přibližnou kritickou hodnotu rovnu v absolutní hodnotě 2 na 0,05 hladině významnosti. Pokud $W > 2$, H_0 zamítne. Ve jmenovateli Waldova testu jsou odhady směrodatných odchylek odhadnutých parametrů, které se také nazývají standardními chybami odhadu parametrů (Standard Errors). Tyto

poskytují zpravidla statistické programové systémy pro logistickou regresi ve fázi odhadu parametrů.

Jak jsme uvedli, odhad rozptylu a kovariancí odhadovaných parametrů se získá z iteračního řešení matice druhých parciálních derivací logaritmu funkce věrohodnosti. Obecný tvar této matice má formu (Hosmer a Lemenshow, 1989)

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{t=1}^{n} x_{tj}^2 \Pi(\mathbf{x}_t)\left[1 - \Pi(\mathbf{x}_t)\right]$$

a (2.49)

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_u} = -\sum_{t=1}^{n} x_{tj}^2 x_{tu}^2 \Pi(\mathbf{x}_t)\left[1 - \Pi(\mathbf{x}_t)\right],$$

pro $j, u = 0, 1, 2, ..., k$.

Pokud jednotlivé termy pravých stran rovnic (2.49) setřídíme do $(k + 1) \times (k + 1)$ matice, získáme matici, která se nazývá *informační matice* s označením $\mathbf{I}_{\boldsymbol{\beta}}$. Rozptyly a kovariance odhadovaných parametrů se získají inverzí informační matice, která se nazývá variačně-kovarianční matice parametrů modelu a má označení $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, tj. $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbf{I}_{\beta}^{-1}$. Matice $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ obsahuje na hlavní diagonále rozptyly parametrů $\boldsymbol{\beta}_j$, $j = 0, 1, 2, ..., k$. Její nediagonální prvky obsahují kovariance mezi $\boldsymbol{\beta}_j$ a $\boldsymbol{\beta}_u$. Informační matici $\mathbf{I}_{\hat{\beta}}$, která je vytvořena z odhadů parametrů, je vhodné zapsat ve tvaru

$$\mathbf{I}_{\hat{\beta}} = \mathbf{X}^T \mathbf{V} \mathbf{X}, \tag{2.50}$$

kde matice $\mathbf{X}$ má tvar

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & .. & x_{1k} \\ 1 & x_{21} & x_{22} & .. & x_{2k} \\ .. & .. & .. & .. & .. \\ 1 & x_{n1} & x_{n2} & .. & x_{nk} \end{pmatrix}$$

a matice $\mathbf{V}$ je

$$\mathbf{V} = \begin{pmatrix} \hat{\Pi}_1(1-\hat{\Pi}_1) & 0 & 0 & 0 \\ 0 & \hat{\Pi}_2(1-\hat{\Pi}_2) & ... & 0 \\ & \vdots & & \\ 0 & 0 & 0 & \hat{\Pi}_n(1-\hat{\Pi}_n) \end{pmatrix},$$

kde $\hat{\Pi}_t$ znamená $\hat{\Pi}(\mathbf{x}_t)$, $t = 1, 2, ..., n$.

2.2.3 Interpretace parametrů

Jakmile byly parametry MLR odhadnuty a ověřeny jako významné, následuje nejpodstatnější fáze celého vývoje MLR, tj. vyvozování závěrů a interpretace hodnot jeho parametrů predikovaných hodnot.

Ve standardním jednoduchém regresním modelu (s jednou nezávislou veličinou) se parametr stojící při nezávislé veličině interpretuje vztahem účinku nezávislé veličiny na závislou veličinu. Hodnota parametru vyjadřuje velikost změny hodnoty závislé veličiny při jednotkové změně nezávislé veličiny. Pokud si uvědomíme, že data nezávislé veličiny v MLR mohou být dichotomická, číselná, kategorická, spojitá a různě hodnotově specifikovaná např. při umělých proměnných, taková interpretace parametru v MLR nemusí být vhodná, protože v MLR nejsou rovnocenné měřicí podmínky v jednotlivých typech nezávislých veličin. Jednotková změna hodnoty nezávislé veličiny má proto různé projevy, v závislosti jako jsou její hodnoty škálované. V MLR při interpretaci parametrů, namísto interpretace projevu účinku jednotkové změny nezávislé veličiny na změnu závislé veličiny, sepři dichotomické nezávislé veličině zavedl *poměr pravděpodobností* označovaný symbolem ψ, popř. jeho logaritmus $\ln\psi$. Pro nezávislou dichotomickou veličinu, s kódováním jejích hodnot nula a jedna, je tento poměr vyjádřen následujícími výrazy

$$\psi = \frac{\Pi(1)/[1\text{-}\Pi(1)]}{\Pi(0)/[1\text{-}\Pi(0)]} = \frac{\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\dfrac{1}{1+e^{\beta_0}}}{\dfrac{e^{\beta_0}}{1+e^{\beta_0}}\dfrac{1}{1+e^{\beta_0+\beta_1}}} = e^{\beta_1} \tag{2.51}$$

$$\ln\psi = \ln\frac{\Pi(1)/[1\text{-}\Pi(1)]}{\Pi(0)/[1\text{-}\Pi(0)]} = \ln\left(e^{\beta_1}\right) = \beta_1. \tag{2.52}$$

Platnost rovnosti ve výrazech (2.51) a (2.52) ve tvaru $\psi = e^{\beta_1}$ a $\ln(\psi) = \beta_1$ je splněna pouze za předpokladu, že nezávislá veličina je kódována s hodnotami nula a jedna. Při jiném kódování pro stanovení hodnot a ln (ψ) je třeba použít a vyčíslit jejich zlomkové tvary.

Pokud si uvědomíme, že při dichotomické nezávislé veličině s kódovanými hodnotami $x = 0$ nebo $x = 1$ se vyskytují i dvě podmíněné pravděpodobnosti $1 - \Pi(x)$ a ekvivalentní $\Pi(x)$, pak čitatel zlomku *poměru pravděpodobností* ψ, resp. $\ln\psi$ vyjadřuje poměr pravděpodobnosti výstupní dichotomické veličiny (s kódovanými hodnotami nula a jedna) pro vstup s $x = 1$, k pravděpodobnosti výstupní dichotomické veličiny Y pro vstup s $x = 0$. Takto *poměr pravděpodobností* nebo *šance* ψ je svým významem mírou asociace, kterou se určuje, kolikrát je pravděpodobnější (častější) výskyt sledované hodnoty výstupního znaku s hodnotou vstupu $x = 1$ než s hodnotou $x = 0$. Konkrétní interpretaci šance uvedeme v následujícím příkladu.

Příklad 2.2

V rámci zjišťování využívání kvantitativních prognostických metod při odhadech budoucího vývoje ekonomických procesů firem byl proveden průzkum ve využívání kvantitativních metod. Jeden požadavek v tomto výzkumu byl specifikovat a kvantifikovat závislost využívání kvantitativních metod od získaného druhu vysokoškolsky vzdělaných manažerů. Pilotáž byla provedena korespondenčně dotazníkovou formou.

V tomto průzkumu byli manažeři (respondenti) požádáni, aby uvedli údaje o absolvování druhu vysoké školy (X), který byl kódovaný hodnotami: $x = 1$ – ekonomické vzdělání, $x = 0$ – jiné vzdělání, a údaje o využívání kvantitativních metod při rozhodování (Y), které byly kódovány hodnotami: $y = 1$ – využívají se kvantitativní metody a $y = 0$ – nevyužívají se kvantitativní metody. Od oslovených respondentů byly získány odpovědi, které jsou uvedeny v tabulce 2–4.

Na základě hodnot z tabulky 2–4 máme kvantifikovat (odhadnout) MLR závislost využívání kvantitativních prognostických metod na nabytém druhu vzdělání.

Kontingenční tabulka 2–4 poskytuje informace o tom, že 59 respondentů má hodnoty znaku $x = 1$ a $y = 1$, 19 respondentů má hodnoty znaku $x = 0$ a $y = 1$, 11 respondentů má hodnoty znaků $x = 1$ a $y = 0$ a také 11 respondentů má hodnoty znaků $x = 0$ a $y = 0$.

Odpovídající funkce věrohodnosti pro tato data, v souladu s výrazem (2.32), je

$$L(\beta_0, \beta_1) = \Pi(1)^{59} [1 - \Pi(1)^{11}] \Pi(0)^{19} [1 - \Pi(0)^{11}].$$

Aplikace programu pro logistickou regresi v systému R poskytne odhad MLR podle tabulky 2–5.

V našem příkladu *poměr pravděpodobností* $\hat{\psi} =$ 3,105 znamená, že výskyt využívání kvantitativních prognostických metod je 3,1krát častější pro pracovníky s ekonomickým vysokoškolským vzděláním než pracovníků, kteří nemají ekonomické vysokoškolské vzdělání.

Z výrazů (2.51) a (2.52) je vidět, že hodnotu ψ, příp. $\ln \psi$ je možno snadno vypočítat. Odhad ψ má při velkém výběru normální rozdělení, na základě čehož je možné vypočítat i jeho interval spolehlivosti. Tento interval se vypočítá na základě konečných hodnot intervalu spolehlivosti parametru $\hat{\beta}_1$ s následným umocněním jeho hodnot se základem e, tj. jako $\exp(\hat{\beta}_1 \pm z_{1-\alpha/2}\ \hat{\sigma}_{\hat{\beta}_1})$.

V posledním příkladu byl prezentovaný postup výpočtu *poměru pravděpodobností* ψ pro dichotomickou nezávislou veličinu druh vzdělání (ekonomické a jiné) s kódováním její hodnoty jedna a nula. Rozšíříme zadání příkladu tak, že budeme uvažovat, kromě ekonomického vzdělání, zohlednění i jiného druhu vysokoškolského vzdělání, např. management.

Tabulka 2–4 Kontingenční tabulka klasifikace druhu vzdělání (DV) a využívání kvantitativních prognostických metod (VKPM) pro 100 manažerů

VKPM (y)	DV (x)		Celkem
	Ekonomické (1)	Jiné (0)	
Využívání (1)	59	19	78
Nevyužívaní (0)	11	11	22
Spolu	70	30	100

Tabulka 2–5 Výsledky odhadu parametrů jejich směrodatných odchylek, testovacích charakteristik a ψ

Veličina	Odhad β_1, β_0	Směrodatná odchylka: $\hat{\sigma}_{\beta 1}, \hat{\sigma}_{\beta 0}$	$\hat{\beta}_1 / \hat{\sigma}_{\beta_1}, \hat{\beta}_0 / \hat{\sigma}_{\beta_0}$	$\hat{\psi}$
DV	1,1331	0,5014	2,259872	3,105268
Konstanta	0,5465	0,3789	1,323304	

Tabulka 2–6 Kontingenční tabulka dat pro VKPM s třemi druhy vzdělání (DV) pro 100 respondentů

VKPM (y)	DV (x)			Celkem
	Ekonomické (3)	Management (2)	Jiné (1)	
Využívání (1)	56	7	15	78
Nevyužívání (0)	14	4	4	22
$\hat{\psi}$	0,0	0,79	0,98	

Potom v pilotáži využívání kvantitativních metod může být druh dosaženého vysokoškolského vzdělání kódovaný trojúrovňově: ekonomické, management a jiné. V tomto případě budeme hypoteticky předpokládat, že předchozí kontingenční tabulka 2–4 rozšířená o sledování vstupu i manažerského vzdělání bude mít tvar tabulky 2–6.

V tabulce 2–5 jsou *poměry pravděpodobností* $\hat{\psi}$ vypočteny vzhledem k referenční (srovnávací) skupině pořízeného ekonomického vzdělání. Volba referenční skupiny závisí na cíli prováděné pilotáže. Např. hodnota ψ pro skupinu nabytého druhu vzdělání management se vypočítá jako $(7 \cdot 14) / (56 \cdot 4) = 0{,}79$.

Rozšíření dichotomické (dvouhodnotové) nezávislé veličiny na víceúrovňové kódování se realizuje v MLR umělými proměnnými. Volba kódování umělých proměnných a specifikace referenční skupiny závisí na zvyklosti konkrétního software. Např. pro systémy SAS, SPSS by mohla být pro druh vzdělání (DV) volba kódování a specifikace referenční skupiny podle následující tabulky 2–7.

Tabulka 2–7 Příklad možného tříúrovňového kódování pomocí umělých prměnných z příkladu 2.2

DV (kódování)	Umělé proměnné (kódování)	
	D_1	D_2
Ekonomické (3)	0	0
Management (2)	1	0
Jiné (1)	0	1

Tabulka 2–8 Výsledky odhadu parametrů jejich směrodatných odchyleka testovacích charakteristik z příkladu 2.3

Veličina	Odhad β_1, β_0	Směrodatná odchylka: $\hat{\sigma}_{\beta 1}, \hat{\sigma}_{\beta 0}$	$\hat{\beta}_1 / \hat{\sigma}_{\beta_1}, \hat{\beta}_0 / \hat{\sigma}_{\beta_0}$
Věk	0,0679	0,0282	2,4078
Konstanta	–2,5822	1,1884	–2,1728

V tabulce 2–6 jsou 2 umělé proměnné D_1, D_2. Specifikace referenční skupiny se provádí podle softvéru SAS tak, že se bere ta skupina, pro kterou byla zvolena největší hodnota kódu nezávislé veličiny. V našem případě je to skupina reprezentující dosažené ekonomické vzdělání.

Pokud jde o diskrétní nominální (kvalitativní) a spojitou nezávislou veličinu pro získání interpretovatelné hodnoty *poměru pravděpodobností* ψ, je nutné, v případě diskrétní nominální veličiny, rozdělit ji do kategorií (skupin), a pak ji nahradit příslušným kódováním umělých proměnných. V případě, že nezávislá veličina má hodnoty spojité, interpretace parametrů v MLR je analogická s interpretací parametrů ve standardním jednoduchém modelu lineární regrese.

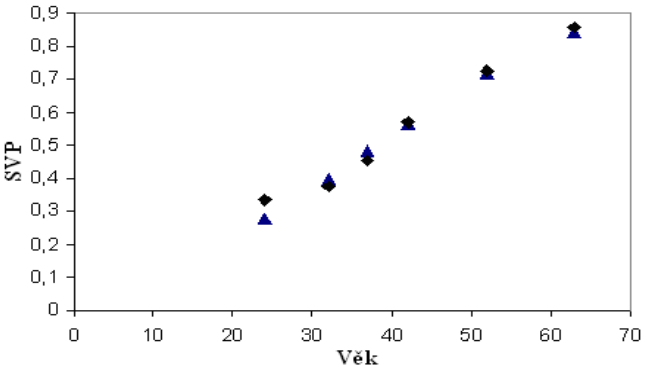
Příklad 2.3

Aplikaci uvedených postupů odhadu parametrů MLR, jejich testování a interpretaci budeme ilustrovat na datech v tabulce 2–1 a 2–2 z pilotáže zjišťování závislostí spokojenosti (nespokojenosti) s vykonávanou prací od věku pracovníků.

Jako první byl proveden odhad parametrů MLR a jejich směrodatných odchylek, který je uveden v následující tabulce 2–8.

Pro testování, zda MLR s odhadnutými parametry může být zjednodušen lze využít G statistiku podle výrazu (2.44). V naší aplikaci hodnota G = 0,0088, co indikuje, že parametr β_1, a tedy i MLR je významný na hladině α = 0,01. Ke stejnému závěru lze dospět Waldovým testem, ve kterém hodnota W-testu = 0,0679 / 0,0282 = 2,4078.

Pokud jde o interpretaci hodnoty odhadu parametru β_1, z tabulky 2–8 je vidět, že data nezávislé veličiny (Věk) jsou ve třech pozorováních lineárně spojitá. S určitým zjednodušením pro interpretace parametru β_1 v MLR uplatníme analogii její interpretace parametrů ve standardním jednoduchém modelu lineární regrese,



Obrázek 2–2 Bodový graf hodnot SVP (◆ skutečnost, ▲ odhad MLR)

tj. bude nás zajímat, jak se změní spokojenost vykonávané práce se změnou věku o 10 let. Odhad *poměru pravděpodobností* prostřednictvím výrazu (2.51) $\hat{\psi} = exp(10 \cdot 0{,}0679) = 1{,}972$, což znamená, že při každém desetiročním zvyšování věku zaměstnanců se zvýší spokojenost s vykonávanou prací téměř dvakrát.

Samozřejmě tento závěr je diskutabilní a neodpovídající skutečnosti, jednak pro uvedené zjednodušení, jednak pro neadekvátní modelování linearizace nezávislé veličiny v logitě. Odstranění poskytne návrh MLR, v němž by jednotlivá pozorování byla rozděleny do skupin, a nezávislá veličina by byla reprezentována umělými proměnnými.

Výsledný tvar odhadnutého modelu má tvar

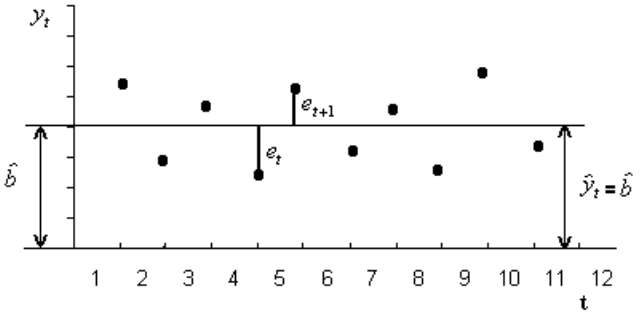
$$\hat{y}_t = \hat{\Pi}(Vek_t) = \frac{\exp(-2{,}5822 + 0{,}0679\ \mathrm{Vek}_t)}{1 + \exp(-2{,}5822 + 0{,}0679\ \mathrm{Vek}_t)}$$

a logit $\hat{g}(Vek_t) = -2{,}5822 + 0{,}0679 Vek_t$. Skutečné a MLR vyrovnané hodnoty SVP jsou vyznačené v grafu na obrázku 2–2.

Složitější situace interpretace parametrů je ve vícenásobném MLR. Její rozbor přesahuje rámec této publikace. Seznámit se s touto problematikou je možné v (Hosmer a Lemenshow, 1989, kap. 3 a 4).

2.3 Modely exponenciálního vyrovnávání

Pravděpodobně nejvíce používanou metodou je metoda exponenciálního vyrovnávání, a to pro její výpočetní jednoduchost a především pro její chybovou přijatelnost. Její autorství je spojeno se jménem Brown, RG; Meyer, RF Metoda je stručně popsána v práci Brown (1963), Meyer (1963).



Obrázek 2–3 Časová řada – konstantní trend

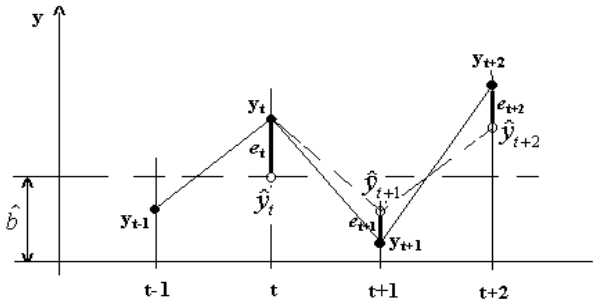
Většina ekonomických dějů nemá ideální vývoj v čase, který by kopíroval jednoduché teoretické funkce, např. přímku, parabolu, hyperbolu apod. V důsledku cenových, technologických, inovačních a jiných vlivů v průběhu některých období se mění jejich plynulý vývoj. Při modelování časových řad v takových případech racionálně nezdůvodníme volbu určitého modelu s konstantními parametry. Je oprávněný požadavek, aby model časové řady zohledňoval náhlé změny a především dokázal s chybovou přijatelností generovat předpovědi pro nejbližší budoucnost.

Jeden ze způsobů modelové aktualizace předpovědí a jejich přizpůsobování posledním informacím je použití proměnlivých parametrů modelu. Proměnlivé parametry se mění s časem, v každém pozorování, na základě průběžného zhodnocování posledního odhadu nebo poslední prognózy veličiny s její skutečností, přičemž v novém, v dalším odhadu hodnoty veličiny se přikládá větší význam posledním (věkově mladším) pozorováním časové řady. Jelikož odhady hodnot časové řady se provádějí v každém pozorování, jedna možná varianta teoretického konceptu exponenciálního vyrovnávání je, že každému sledování budou odpovídat i jiné parametry modelu. Tento koncept se v literatuře označuje jako koncept postupného trendu.

Uvedený princip exponenciálního vyrovnávání, který jsme zatím naznačili jen v generické formě, vyjádříme i formálně. Předpokládejme nejjednodušší případ s pozorováními časové řady $\{y_t\}$, její střední hodnota se v čase nemění, tj. je konstantní během celé historie pozorování (viz obrázek 2–3). Řada má střední hodnotu, kterou označíme symbolem *b*. Při modelování této řady klasickou regresí by mohl být adekvátním modelem model s konstantním trendem, tj.

$$y_t = b + u_t, \tag{2:53}$$

kde u_t je náhodný člen modelu. Nejlepším odhadem parametru *b* – střední hodnoty časové řady je jednoduchý aritmetický průměr $\hat{b} = (1/N)\sum_{t=1}^{N} y_t$. Pak modelem (2.53) lze každou hodnotu pozorování řady vyjádřit jako



Obrázek 2–4 Časová řada s konstantním trendem – jednoduché exponenciální vyrovnávání

$$y_t = \hat{b} + e_t = \hat{y}_t + e_t. \tag{2.54}$$

Při exponenciálním vyrovnávání je nový odhad $\hat{y}_{t+1}$ definován jako kombinace aktuálního odhadu $\hat{y}_t$ a části aktuální chyby e_t, tj.

$$\hat{y}_{t+1} = \hat{y}_t + \alpha e_t, \tag{2.55}$$

kde α je tzv. vyrovnávací konstanta, přičemž $\alpha \in <0,1>$. Porovnáme-li výrazy (2.54) a (2.55), pravou stranu výrazu (2.55) můžeme chápat jako aktualizovanou úroveň časové řady v čase *t*, tj. můžeme ji vyjádřit jako

$$\hat{y}_{t+1} = \hat{b}_t. \tag{2.56}$$

Ze vztahů (2.55), (2.56) je vidět, že úroveň časové řady $\hat{b}_t$ není konstantní, ale je proměnlivá v čase. Její průběh, v našem případě, je dán výrazem (2.54). Průběh $\hat{b}_t$ je zakreslen v obrázku 2–4. V tomto obrázku je zakreslena situace časové řady s konstantním trendem, přičemž odhad pozorování $\hat{y}_t$ v čase $t = 1$ je konstruován klasickým regresním modelem, odhady pozorování $\hat{y}_{t+1}$ a $\hat{y}_{t+2}$ jsou konstruovány technikou exponenciálního vyrovnávání. Z obrázku je jasně vidět, jak technika exponenciálního vyrovnávání zmenšuje chyby odhadů ve srovnání s klasickým regresním modelem.

Rovnice (2.55) se obvykle píše v jiném tvaru, kterým je

$$S_t = S_{t-1} + \alpha\left(y_t - S_{t-1}\right), \tag{2.57}$$

kde S_t je odhad nebo předpověď hodnoty časové řady $\{ y_t \}$ pro další období $t + 1$, které se provádí v aktuálním čase *t*. S_{t-1} je odhad nebo předpověď hodnoty časové řady $\{ y_t \}$ pro aktuální čas *t*, provedený v předchozím období $t - 1$. y_t je

pozorovaná hodnota časové řady v čase *t*. Rozdíl $(y_t - S_{t-1})$ je chyba odhadu, resp. chyba předpovědi v aktuálním čase *t*. Výraz (2.57) se dá po jednoduché úpravě napsat, s přihlédnutím k rovnicím (2.54) a (2.55), jako

$$S_t = \alpha y_t + (1-\alpha) S_{t-1}$$

nebo (2.58)

$$S_t = \alpha y_t + \beta S_{t-1},$$

kde $\beta = 1-\alpha$. Výrazy pro S_t definované podle (2.58), se nazývají jednoduchým exponenciálním vyrovnáváním. S_t se nazývá exponenciálním průměrem v čase *t*, S_{t-1} označuje exponenciální průměr v čase *t* – 1. Z výrazu (2.58) je zřetelně vidět, že exponenciální vyrovnávání je aktualizační (vyrovnávací) procedura, pomocí které je veličina S_t upravována proporcionálně její předchozí hodnotě.

Výraz (2.58), a tedy i výrazy (2.55) až (2.57) jsme nazvali jednoduchým exponenciálním vyrovnáváním. Skutečný význam tohoto pojmenování pochopíme, pokud výraz (2.58) rozepíšeme rekurentním způsobem pro všechna pozorování časové řady { y_t }

$$\begin{aligned} S_t &= \alpha y_t + \beta S_{t-1} = \alpha y_t + \beta\left(\alpha y_{t-1} + \beta S_{t-2}\right) \\ &= \alpha y_t + \alpha\beta y_{t-1} + \beta^2\left(\alpha y_{t-2} + \beta S_{t-3}\right) \\ &= \alpha y_t + \alpha\beta y_{t-1} + \alpha\beta^2 y_{t-2} + \ \ldots\ + \beta^N S_0 \\ &= \alpha \sum_{k=0}^{N-1} \beta^k y_{t-k} + \beta^N S_0, \end{aligned}$$

kde *N* je délka časové řady, *k* = 0, 1, ..., *N* – 1 je věk pozorování. S_0 je veličina charakterizující počáteční podmínky rekurentního výrazu (2.58). Nazývá se počátečním odhadem úrovně časové řady. Pokud časová řada je dostatečně dlouhá, člen $\beta^N S_0$ z posledního výrazu lze vynechat, pak

$$S_t = \alpha \sum_{k=0}^{N-1} \beta^k y_{t-k}. \qquad (2.59)$$

Z výrazu (2.59) je zřejmé, proč se S_t nazývá exponenciálním průměrem. Je to vážený součet všech členů časové řady { y_t } s exponenciálně klesajícími vahami. Např. pokud α = 0,2, potom členy časového řady $y_{N-1}, y_{N-2}, y_{N-3}, \ldots$ budou mít váhy 0,2; 0,16; 0,128; ...

Nezodpovězenými otázkami zůstaly volba vyrovnávací konstanty α a volba počátečního exponenciálního průměru S_0. Volbou vyrovnávací konstanty α, jak je na obrázku 2–4 vidět, je určována rychlost přizpůsobování další prognóze k aktuální chybě předpovědi. Pokud α je velké, v nové prognóze nebo v novém odhadu je zahrnuta velká část posledního pozorování. Pokud α je voleno blízko nuly, nová předpověď se bude téměř shodovat s předchozí a vliv poslední aktuální (pozorované) veličiny je potlačen. Znamená to, pokud budeme požadovat, aby předpovědi modelované veličiny byly stabilní a náhodné chyby (v absolutních hodnotách) nebyly extrémně rozdílné, že budeme volit malou hodnotu α. Pokud budeme požadovat, aby nové odhady, resp. předpovědi časové řady obsahovaly velkou část informací z předchozích pozorování nebo aby byla rychlá odezva na poslední změnu, budeme volit α velké. V různých zdrojích odborné literatury lze nalézt doporučení, že α by nemělo přesáhnout hodnotu 0,3. V jiném případě poruchový člen e_t ve výrazu (2.55) přestává být náhodným. V práci (Gaynor a Kirpatrick, 1994) se při volbě α vychází z průměrné čtvercové chyby předpovědí $(\frac{1}{N}\sum_N (y_t - \hat{y}_t)^2)$. Potom by α měla mít takovou hodnotu, která poskytuje minimální průměrnou čtvercovou chybu. V praktických případech se vyhledávání nejvhodnějšího α podle minimalizačního kritéria průměrné čtvercové chyby provádí iterativně. Na začátku se zvolí malé α, např. 0,01, vypočítá se výraz $\frac{1}{N}\sum_N (y_t - \hat{y}_t)^2$. Pak se malým přírůstkem zvýší α a výpočet se opakuje pro novou hodnotu α. Vyhovující α se vybere takové, které garantuje minimální čtvercovou chybu předpovědí, resp. odhadů.

Co se týče počáteční volby exponenciálního průměru S_0, také neexistuje žádné jednoznačné pravidlo pro jeho určení. Jeho volba, podobně jako volba α, je charakterizovaná silnou dávkou subjektivizmu. Obecně se doporučuje, aby hodnota S_0 při dlouhých časových řadách odrážela střední hodnotu časové řady vypočtenou např. jednoduchým aritmetickým průměrem z *N* pozorování. Jestliže rozsah výběru *N* je malý, hodnota S_0 zvolená kritériem aritmetického průměru bude vychýlena. Tím budou vychýlené i exponenciální průměry S_t. Konstruované předpovědi nebudou přesné ve smyslu nezkreslenosti. Pokud předpovědi budou nesprávné, Montgomery (1990) doporučuje použít při vyrovnávání časové řady pro prvních několik pozorování větší hodnoty α.

Koncept exponenciálního vyrovnávání byl od svého vzniku podrobně teoreticky rozpracován pro časové řady s konstantním, lineárním a kvadratickým trendem a zobecněný pro trend libovolného stupně. V další části rozebereme techniku exponenciálního vyrovnávání pro modely s konstantním, lineárním a kvadratickým trendem.

2.3.1 Jednoduché exponenciální vyrovnání

Naší snahou v předchozí části bylo zdůvodnit, proč při popisech ekonomických časových řad dáváme někdy přednost modelům s proměnlivými parametry. Nejjednodušším modelem s proměnlivými parametry je stacionární model. Na popis stacionárního modelu lze aplikovat metodu exponenciálního vyrovnávání. Pro tento účel uvažujme příklad časové řady s 20 pozorováními kurzů akcí firmy IBM, jejichž hodnoty jsou uvedeny v tabulce 2–9. Data této časové řady nevykazují v pozorováních výběru vzestupný nebo sestupný trend. V časové řadě jsou zjevné fluktuace v datech. Jeho popis klasickým modelem regresní analýzy s konstantním trendem by nebyl adekvátní, neboť jeho všechny vyrovnané hodnoty by byly na úrovni aritmetického průměru $\hat{y}_t = \bar{y} = (1/20)\sum_{t=1}^{20} y_t$. Klasický model regresní analýzy s konstantním trendem by nebyl adekvátním modelem ani pro prognózování jeho hodnot. Na konstrukci prognóz použijeme model exponenciálního vyrovnávání s konstantním trendem. Jak jsme uvedli, modely exponenciálního vyrovnávání zohledňují skutečnost, že vzdálenějším pozorováním, tj. pozorováním věkově starším se přisuzují nižší váhy než pozorováním věkově mladším. Tato skutečnost se využívá při odhadu proměnlivých parametrů b_t modelu v jednotlivých pozorováních časové řady. Odhad proměnlivých parametrů modelu $\hat{b}_t$ určíme váženou metodou nejmenších čtverců, tj. požadujeme, aby

$$\sum_{k=0}^{N-1} \alpha\beta^k \left(y_{t-k} - \hat{b}_t\right)^2 \quad \text{... min.}$$

Požadavek bude splněn pro hodnoty odhadů proměnlivých parametrů $\hat{b}_t$, pokud derivaci posledního výrazu podle $\hat{b}_t$ položíme rovnou nule, čím dostaneme

$$\sum_{k=0}^{N-1} \alpha\beta^k y_{t-k} - \hat{b}_t \alpha \sum_{k=0}^{N-1} \beta^k = 0.$$

Pro dostatečně velké *N* můžeme nahradit součet nekonečné geometrické řady

$$\sum_{k=0}^{\infty} \beta^k \approx \sum_{k=0}^{N=1} \beta^k = \frac{1}{1-\beta} = \frac{1}{\alpha}.$$

Potom z předcházejícího výrazu získáme odhadovanou funkci pro proměnlivé parametry $\hat{b}_t$

$$\hat{b}_t = \alpha \sum_{k=0}^{N-1} \beta^k y_{t-k}. \tag{2.60}$$

Porovnáním odhadovaného výrazu pro proměnlivé parametry modelu konstantního trendu (2.60) s výrazem pro exponenciální průměr S_t podle (2.59) můžeme napsat

$$\hat{b}_t = S_t. \tag{2.61}$$

Tabulka 2–9 Kurzy akcií firmy IBM: jednoduché exponenciální vyrovnávání, α = 0,3

t	y_t	S_t	$\hat{y}_t(1)$	e_t
1	510	507,27	506,1	3,9
2	497	504,189	507,27	–10,27
3	504	504,1323	504,189	–0,189
4	510	505,89261	504,1323	5,8677
5	509	506,82483	505,89261	3,10739
6	503	505,67738	506,82483	–3,824827
7	500	503,97417	505,67738	–5,677379
8	500	502,78192	503,97417	–3,974165
9	500	501,94734	502,78192	–2,781916
10	495	499,86314	501,94734	–6,947341
11	494	498,1042	499,86314	–5,863139
12	499	498,37294	498,1042	0,8958029
13	502	499,46106	498,37294	3,627062
14	509	502,32274	499,46106	9,5389434
15	525	509,12592	502,32274	22,67726
16	512	509,98814	509,12592	2,8740823
17	510	509,9917	509,98814	0,0118576
18	506	508,79419	509,9917	–3,9917
19	515	510,65593	508,79419	6,2058102
20	522	514,05915	510,65593	11,344067
21			514,05915	

Vidíme exponenciální průměry, které jsou současně odhady proměnlivých parametrů a tím i odhady vyrovnaných hodnot časové řady { y_t }.

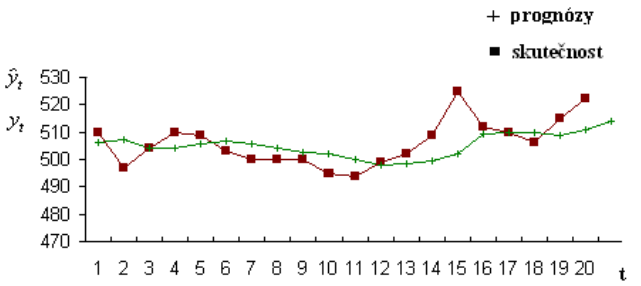
Pokud výrazy (2.56), (2.60) a (2.61) aplikujeme na časovou řadu akcií firmy IBM a při volbě hodnoty vyrovnávací konstanty α = 0,3, získáme hodnoty prognóz jednotlivých pozorování časové řady, které jsou uvedeny v tabulce 2–9. Konkrétně, nejdříve určíme počáteční hodnotu exponenciálního průměru na úrovni aritmetického průměru řady, tj. S_0 = 506,1. Pro výpočet prognóz časové řady jsme použili výrazy (2.55) a (2.57)

$$\hat{y}_{t+1} = S_t = \alpha y_t + (1-\alpha) S_{t-1}$$

pro každou periodu. Předpovědi byly vypočteny následovně:

$$\hat{y}_0(1) = \hat{y}_1 = S_0 = \hat{b} = 506{,}10.$$

V posledním předpisu symbolem $\hat{y}_0(1)$ jsme zvýraznili skutečnost, že předpověď o jedno období dopředu $\hat{y}_{t+1}$ je konstruována z počátku (z úrovně)



Obrázek 2–5 Vývoj prognóz a skutečností IBM akcí (jednoduché exponenciální vyrovnávání, α = 0,3)

prognózy v čase *t*, s předpovědí τ = 1, což pro *t* = 0 a τ = 1 můžeme napsat jako $\hat{y}_t(\tau) = \hat{y}_0(1)$ nebo jednoduše jako $\hat{y}_1$.

Předpověď s počátkem předpovědí *t* = 1.

$$\hat{y}_1(1) = \hat{y}_2 = S_1 = \alpha y_1 + (1-\alpha) S_0 = 0.3(510) + (1-0,3)(506,1) = 507,27.$$

Pokračováním tohoto výpočetního postupu pro periody *t* = 3, 4, ..., 21 získáme prognózy $\hat{y}_3, \hat{y}_4, \ldots, \hat{y}_{21}$. Jejich hodnoty jsou uvedeny v tabulce 2–9. Grafický průběh prognóz se skutečnostmi je na obrázku 2–5.

2.3.2 Dvojité exponenciální vyrovnání

Dvojité exponenciální vyrovnávání nebo Brownův model exponenciálního vyrovnávání pro lineární trend je podobné jako jednoduché exponenciální vyrovnávání. Je jeho rozšířením pro časové řady s lineárními trendy, jejichž hodnoty fluktují kolem tohoto trendu. Jinými slovy řečeno a formálně vyjádřeno, jde o časové řady, ve kterých se střední hodnota lineárně mění s časem v souladu s modelem

$$y_t = b_0 + b_1 t + u_t. \tag{2.62}$$

V modelu (2.62) očekávaná hodnota $E(y_t) = b_0 + b_1 t$ v čase t je lineární funkcí času, u_t je náhodný člen modelu s nulovou střední hodnotou a konstantním rozptylem ve všech pozorováních. Modely jednoduchého, dvojitého a modely vyšších stupňů exponenciálního vyrovnávání, ve kterých se střední hodnota mění s časem, jsou odvozeny od obecného modelu parabolického trendu stupně *N* v tvaru

$$y_t = b_{0t} + b_{1t} t + b_{2t} \frac{t^2}{2!} + \ldots + b_{Nt} \frac{t^N}{N!}. \tag{2.63}$$

V souladu s tvarem modelu (2.63) namísto modelu (2.62) pro lineární trend budeme uvažovat tvar (2.63), tj.

$$y_t = b_{0t} + b_{1t}t + u_t. \tag{2.64}$$

Odhady parametrů modelu lineárního trendu (2.64) označíme $\hat{b}_{0t}, \hat{b}_{1t}$. Jejich odhadované hodnoty získáme metodou nejmenších čtverců, ve které požadujeme, aby platilo

$$\alpha \sum_{k=0}^{N-1} \beta^k \left(y_{t-k} - \hat{b}_{0t} - k\hat{b}_{1t} \right)^2 \text{ ... min.}$$

Pro odhad parametrů modelu (2.64) v minimalizačním kritériu nejmenších čtverců podle posledního výrazu se předpokládá, že model exponenciálního vyrovnávání má počátek času umístěný na konci časové řady. Podmínka vážené metody nejmenších čtverců bude splněna, pokud parciální derivace posledního výrazu podle $\hat{b}_{0t}, \hat{b}_{1t}$ položíme rovné nule. Tím obdržíme systém dvou normálních rovnic

$$\begin{aligned} \sum_{k=0}^{N-1} \beta^k y_{t-k} &= \hat{b}_{0t} \alpha \sum_{k=0}^{N-1} \beta^k - \hat{b}_{1t} \alpha \sum_{k=0}^{N-1} k\beta^k \\ \sum_{k=0}^{N-1} k\beta^k y_{t-k} &= \hat{b}_{0t} \alpha \sum_{k=0}^{N-1} k\beta^k - \hat{b}_{1t} \alpha \sum_{k=0}^{N-1} k^2\beta^k. \end{aligned} \tag{2.65}$$

Pokud budeme uvažovat časovou řadu s $N \to \infty$, potom s využitím asymptotických vlastností platí

$$\sum_{k=0}^{N-1} \beta^k \approx \sum_{k=0}^{\infty} \beta^k = \frac{1}{1-\beta}$$

$$\sum_{k=0}^{N-1} k\beta^k \approx \sum_{k=0}^{\infty} k\beta^k = \frac{\beta}{1-\beta}$$

$$\sum_{k=0}^{N-1} k^2\beta^k \approx \sum_{k=0}^{\infty} k^2\beta^k = \frac{\beta(1+\beta)}{1-\beta}.$$

V reálných podmínkách máme k dispozici dlouhé, ale konečné výběry. Proto jsme indexy sčítání $k = 0, 1, 2, ..., N-1$ označili s přibližnou platností. Pokud dosadíme tyto sumační vztahy do soustavy rovnic (2.65), získáme jednodušší tvar systému rovnic

$$\alpha \sum_{k=0}^{N-1} \beta^k y_{t-k} = \hat{b}_{0t} - \hat{b}_{1t} \frac{\beta}{\alpha}$$

$$\alpha^2 \sum_{k=0}^{N-1} k\beta^k y_{t-k} = \beta \hat{b}_{0t} - \hat{b}_{1t} \frac{\beta}{\alpha}(1+\beta). \tag{2.66}$$

Pokud poslední soustavu rovnic (2.66) vyjádříme ve tvaru exponenciálních průměrů, můžeme ji po úpravě přepsat na tvar

$$S_t = \hat{b}_{0t} - \hat{b}_{1t} \frac{\beta}{\alpha}$$

$$S_t^{[2]} = \hat{b}_{0t} - 2\frac{1-\alpha}{\alpha} \hat{b}_{1t} \tag{2.67}$$

kde

$$S_t^{[2]} = \alpha S_t + \beta S_{t-1}^{[2]} = \alpha \sum_{k=0}^{N-2} \beta^k S_{t-k}. \tag{2.68}$$

Ve výraze (2.68) je symbolem $S_t^{[2]}$ označený exponenciální průměr druhého stupně, který jsme získali aplikací exponenciálního vyrovnávání v časové řadě exponenciálních průměrů $\{ S_t \}$ prvního stupně. Nakonec ze soustavy rovnic (2.67) vypočteme proměnlivé parametry $\hat{b}_{0t}, \hat{b}_{1t}$ modelu (2.62) technikou exponenciálního vyrovnávání jako

$$\hat{b}_{0t} = 2S_t - S_t^{[2]}$$

$$\hat{b}_{1t} = \frac{\alpha}{\beta}\left(S_t - S_t^{[2]}\right). \tag{2.69}$$

Soustava rovnic (2.69) poskytuje odhadované výrazy proměnlivých parametrů modelu (2.64). Pro jejich výpočet, podle těchto výrazů v čase *t*, potřebujeme zjistit hodnoty exponenciálních průměrů $S_t, S_t^{[2]}$. Avšak jejich výpočet vyžaduje znalost hodnot exponenciálních průměrů v čase *t* – 1. Pak na odhad proměnlivých parametrů $\hat{b}_{0t}, \hat{b}_{1t}$ v čase *t* = 1 musíme znát hodnoty počátečních exponenciálních průměrů $S_0, S_0^{[2]}$. Hodnoty $S_0, S_0^{[2]}$ získáme ze soustavy rovnic (2.67), pokud do nich dosadíme např. odhady parametrů z jednoduchého lineárního trendu (2.62), tj.

$$S_0 = \hat{b}_0 - \hat{b}_1 \frac{\beta}{\alpha}$$

$$S_0^{[2]} = \hat{b}_0 - 2\frac{1-\alpha}{\alpha} \hat{b}_1. \tag{2.70}$$

Pokud máme určeny počáteční hodnoty exponenciálních průměrů $S_0, S_0^{[2]}$, můžeme z vyhodnocovacích výrazů (2.69) určit hodnoty proměnlivých parametrů

v čase t = 1, tj. $\hat{b}_{0t}, \hat{b}_{1t}$ a následně z modelu (2.64) předpověď o τ = 1 období dopředu, kterou označíme $\hat{y}_t(\tau) = \hat{y}_1(1)$ nebo jednodušeji $\hat{y}_2$, tj.

$$\hat{y}_1(1) = \hat{y}_2 = \hat{b}_{0t} + \hat{b}_{1t}\tau. \tag{2.71}$$

Ostatní prognózy $\hat{y}_3, \hat{y}_4, \ldots, \hat{y}_{N+1}$ určíme stejným způsobem jako $\hat{y}_2$, tj. postupně, vždy z úrovně předchozí prognózy, o jedno období dopředu, výrazem (2.71).

Pro ilustraci výpočtu prognóz dvojitým exponenciálním vyrovnáváním uvažujme jednoduchý příklad. Předpokládejme, že máme k dispozici pozorování y_t pro t = 1, 2, ..., 18 o časovém využití strojního zařízení. Tato pozorování jsou uvedena v tabulce 2–10. Chceme vypočítat všechny prognózy y_t modelem dvojitého exponenciálního vyrovnávání s vyrovnávací konstantou α = 0,5. Celý postup výpočtu shrneme do následujících kroků:

1. Odhadneme lineární trend klasickým regresním modelem, tj. modelem

   $y_t = b_0 + b_1 t + u_t$ pro t = 1, 2, ..., 18,

   Odhady parametrů provedeme jednoduchou metodou nejmenších čtverců, která poskytne $\hat{b}_0 = 161{,}098,\ \hat{b}_1 = -1{,}9226.$

2. Vypočítáme počáteční prognózu $\hat{y}_1$. Její hodnotu určíme jako vyrovnanou hodnotu z předchozího klasického regresního modelu, tj.

   $$\hat{y}_1 = b_0 + b_1 t = 161{,}098 + (-1{,}9226)\ 1 = 159{,}18.$$

3. Pomocí výrazů (2.67) určíme počáteční hodnoty exponenciálních průměrů $S_0, S_0^{[2]}$

   $$S_0 = \hat{b}_0 - \hat{b}_1 \frac{1-\alpha}{\alpha} = 161{,}098 - (-1{,}9226)\frac{1-0{,}5}{0{,}5} = 163{,}02$$

   $$S_0^{[2]} = \hat{b}_0 - 2\frac{1-\alpha}{\alpha}\hat{b}_t = 161{,}098 - 2\frac{1-0{,}5}{0{,}5}(-1{,}9226) = 161{,}1.$$

4. Z výrazů (2.58) a (2.68) určíme exponenciální průměry $S_t, S_t^{[2]}$ pro t = 1

   $$S_t = \alpha y_t + (1-\alpha) S_{t-1} = 0{,}5\ 163 + (1-0{,}5)163{,}02 = 163{,}01$$

   $$S_t^{[2]} = \alpha S_t + \beta S_{t-1}^{[2]} = 0{,}55\ 163{,}01 + (1\text{-}0{,}55)161{,}1 = 162{,}05\ .$$

5. Z výrazů (2.69) určíme hodnoty proměnlivých parametrů modelu (2.64)

   $$\hat{b}_{0t} = 2S_t - S_t^{[2]} = 2 \cdot 163{,}01 - 162{,}05 = 163{,}97$$

Tabulka 2–10 Dvojité exponenciální vyrovnání – vytížení strojového zařízení, $\alpha = 0{,}5$

T	y_t	S_t	$S_t^{[2]}$	$\hat{b}_{0t}$	$\hat{b}_{1t}$	$\hat{y}_t$	e_t
1	163	163,01	162,05	163,97	0,96	159,18	3,82
2	159	161,01	161,05	160,97	–0,04	164,92	–5,92
3	136	149,51	155,30	143,72	–5,79	160,92	–24,92
4	158	160,51	160,80	160,22	–0,29	137,92	20,08
5	146	154,51	157,80	151,22	–3,29	159,92	–13,92
6	146	154,51	157,80	151,22	–3,29	147,92	–1,92
7	155	159,01	160,05	157,97	–1,04	147,92	7,08
8	158	160,51	160,80	160,22	–0,29	156,92	1,08
9	149	156,01	158,55	153,47	–2,54	159,92	–10,92
10	130	146,51	153,80	139,22	–7,29	150,92	–20,92
11	158	160,51	160,80	160,22	–0,29	131,92	26,08
12	136	149,51	155,30	143,72	–5,79	159,92	–23,92
13	138	150,51	155,80	145,22	–5,29	137,92	0,08
14	129	146,01	153,55	138,47	–7,54	139,92	–10,92
15	129	146,01	153,55	138,47	–7,54	130,92	–1,92
16	130	146,51	153,80	139,22	–7,29	130,92	–0,92
17	127	145,01	153,05	136,97	–8,04	131,92	–4,92
18	124	143,51	152,30	134,72	–8,79	128,92	–4,92
19						125,92	

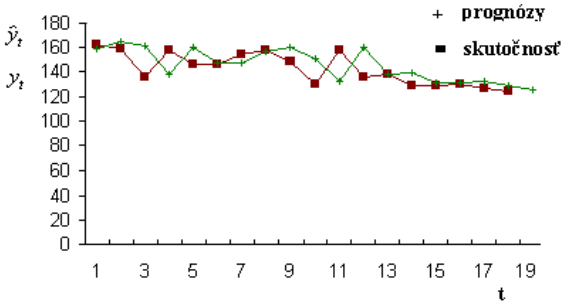
$$\hat{b}_{1t} = \frac{\alpha}{\beta}\left(S_t - S_t^{[2]}\right) = \frac{0{,}55}{1-0{,}55}\left(163{,}01 - 162{,}05\right) = 0{,}96$$

a pomocí výrazu (2.71) předpověď

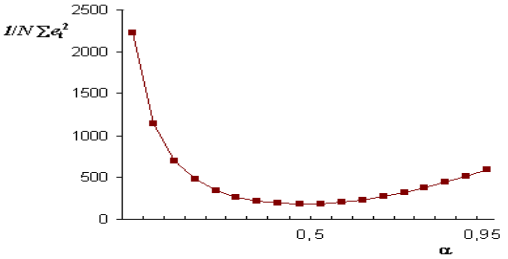
$$\hat{y}_1(1) = \hat{y}_2 = \hat{b}_{0t} + \hat{b}_{1t}\tau = 163{,}97 + 0{,}96\ (1) = 164{,}92.$$

6. Zbývající předpovědi $\hat{y}_3, \hat{y}_4, \ldots, \hat{y}_{N+1}$ určíme stejným způsobem, tj. opakováním kroků 4 a 5. Příslušné exponenciálně průměry $S_t, S_t^{[2]}$ a proměnlivé parametry $\hat{b}_{0t}, \hat{b}_{1t}$ s prognózami $\hat{y}_t$ jsou uvedena v tabulce 2–10. Graf vývoje aktuálních a vyrovnaných hodnot je zobrazen v obrázku 2–6.

Dvojité exponenciální vyrovnávání jsme ilustrovali pro časovou řadu vytíženosti strojového zařízení s vyrovnávací konstantou $\alpha = 0{,}5$. Její volba nebyla náhodná. Vychází z minimalizačních kritérií průměrné čtvercové chyby předpovědí $(\frac{1}{N}\sum_N (y_t - \hat{y}_t)^2 = \frac{1}{N}\sum_N e_t^2)$. Průběh průměrné čtvercové chyby předpovědí v závislosti na hodnotách α je zobrazen na obrázku 2–7. Počáteční



Obrázek 2–6 Dvojité exponenciálně vyrovnávání – vytížení strojového zařízení α = 0,5



Obrázek 2–7 Průběh průměrné čtvercové chyby předpovědí pro dvojité exponenciálně vyrovnávání – vytížení strojového zařízení

hodnotu jsme zvolili α = 0,05. Postupně jsme ji zvyšovali o přírůstky 0,05. Na obrázku 2–7 je vidět, že nejmenší hodnota průměrné čtvercové chyby předpovědí je při α = 0,5.

2.3.3 Trojité exponenciální vyrovnávání a exponenciální vyrovnávání modelů vyšších stupňů

Uvažujme časovou řadu, její trend je kvadratický, tj. jeho střední hodnota se mění v závislosti na času kvadratickou funkcí času, tj.:

$$y_t = b_{0t} + b_{1t}t + b_{2t}t^2 + u_t, \tag{2.72}$$

kde u_t je náhodný člen modelu se standardními předpoklady $u_t \sim \mathrm{N}(0, \sigma^2)$.

Výchozím modelem pro trojité exponenciální vyrovnávání, v souladu s (2.63), je model kvadratického trendu ve tvaru

$$y_t = b_{0t} + b_{1t}t + \frac{1}{2}b_{2t}t^2 + u_t. \tag{2.73}$$

Odhady proměnlivých parametrů modelu kvadratického trendu (2.73) označíme $\hat{b}_{0t}, \hat{b}_{1t}, \hat{b}_{2t}$. Získáme je jednoduchou metodou nejmenších čtverců, při níž se požaduje, aby platilo

$$\alpha\sum_{k=0}^{N-1}\beta^k\left(y_{t-k}-\hat{b}_{0t}-k\hat{b}_{1t}-\frac{k^2}{2}\hat{b}_{1t}\right)^2 \text{ ... min.}$$

Na tomto místě nebudeme rozepisovat celý postup odvození výrazů pro odhad parametrů, neboť postup je analogický jako u modelů exponenciálního vyrovnávání nižších stupňů. Uvedeme jen výsledné výrazy, které jsou nezbytné, abychom mohli odhadovat prognózy technikou trojitého exponenciálního vyrovnávání.

V první řadě, v případě trojitého exponenciálního vyrovnávání, určíme tři počáteční exponenciální průměry $S_0, S_0^{[2]}, S_0^{[3]}$, které jsou definovány

$$\begin{aligned} S_0 &= \hat{b}_0 - \hat{b}_1\frac{\beta}{\alpha}+\frac{\beta(2-\alpha)}{\alpha^2}\hat{b}_2 \\ S_0^{[2]} &= \hat{b}_0 - 2\frac{\beta}{\alpha}\hat{b}_1+\frac{\beta(3-2\alpha)}{\alpha^2}\hat{b}_2 \\ S_0^{[3]} &= \hat{b}_0 - 3\frac{\beta}{\alpha}\hat{b}_1+\frac{3\beta(4-3\alpha)}{\alpha^2}\hat{b}_2. \end{aligned} \tag{2.74}$$

Vidíme, že počáteční exponenciální průměry $S_0, S_0^{[2]}, S_0^{[3]}$ jsou funkcemi odhadů počátečních parametrů $\hat{b}_0, \hat{b}_1, \hat{b}_2$ modelu (2.73).

Dále určíme exponenciální průměry $S_t, S_t^{[2]}, S_t^{[3]}$ v čase *t*, které jsou definovány

$$\begin{aligned} S_t &= \alpha y_t + \beta S_{t-1} \\ S_t^{[2]} &= \alpha S_t + \beta S_{t-1}^{[2]} \\ S_t^{[3]} &= \alpha S_t^{[2]} + \beta S_{t-1}^{[3]}. \end{aligned} \tag{2.75}$$

Prostřednictvím exponenciálních průměrů (2.75) jsou definovány odhadované výrazy proměnlivých parametrů modelu (2.73) jako

$$\begin{aligned} \hat{b}_{0t} &= 3S_t - 3S_t^{[2]} + S_t^{[3]} \\ \hat{b}_{1t} &= \frac{\alpha}{2\beta^2}\left[(6-5\alpha)S_t - 2(5-4\alpha)S_t^{[2]} + (4-3\alpha)S_t^{[3]}\right] \\ \hat{b}_{2t} &= \frac{\alpha^2}{\beta^2}\left(S_t - 2S_t^{[2]} + S_t^{[3]}\right). \end{aligned} \tag{2.76}$$

Nakonec pomocí výrazu (2.73) určíme předpověď o $\tau = 1$ periodu dopředu, s počátkem předpovědi $t = 1$, kterou opět označíme jako $\hat{y}_t(\tau) = \hat{y}_t(1)$ nebo jednodušeji $\hat{y}_{t+1} = \hat{y}_2$

$$\hat{y}_{t+1} = \hat{b}_{0t} + \hat{b}_{0t}\tau + \hat{b}_{2t}\frac{\tau^2}{2}. \tag{2.77}$$

Pro ilustraci použití uvedených výrazů trojitého exponenciálního vyrovnávání uvažujeme příklad, ve kterém máme k dispozici časovou řadu $\{ y_t \}$ s délkou $N = 7$ pozorování o odevzdaných bytech do obecního vlastnictví na Slovensku za roky 1991 až 1997. Tyto údaje jsou uvedeny v tabulce 2–11. Naším úkolem bude vypočítat prognózy dokončených bytů do obecního vlastnictví pro roky 1998 až 2002.

Protože trend časové řady $\{ y_t \}$ pro 7 pozorování nejlépe vystihuje kvadratická funkce $y_t = b_0 + b_1 t + b_2 t^2$ budeme určovat prognózy technikou trojitého exponenciálního vyrovnávání. Vyrovnávací konstantu zvolíme $\alpha = 0{,}11$. Nejprve jednoduchou metodou nejmenších čtverců odhadneme parametry tohoto kvadratického trendu. Odhady parametrů jsou:

$$\hat{b}_0 = 4866, \qquad \hat{b}_1 = -1578{,}88, \qquad \hat{b}_2 = 166{,}48. \tag{2.78}$$

Vypočítáme hodnotu $\hat{y}_t$ z regresního modelu pro $t = 1$

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 t + \hat{b}_2 t^2 = 4866 + (-15788) + 1664{,}8 = 3453{,}6,$$

kterou současně budeme považovat za prognózu skutečné hodnoty y_t pro $t = 1$. Označíme ji $\hat{y}_t(\tau) = \hat{y}_0(1) = \hat{y}_{t+\tau} = \hat{y}_1$, neboť jde o prognózu, která je vytvářena v počátku $t = 0$, s horizontem $\tau = 1$.

Odhady parametrů (2.78) využijeme pro výpočet počátečních exponenciálních průměrů $S_0, S_0^{[2]}, S_0^{[3]}$, které vypočítáme podle výrazů (2.74), tj. pro exponenciální průměr prvního stupně.

$$\begin{aligned} S_0 &= \hat{b}_0 - \hat{b}_1\frac{\beta}{\alpha} + \frac{\beta(2-\alpha)}{\alpha^2}\hat{b}_2 \\ &= 4886 - (-1578{,}88)\frac{0{,}89}{0{,}11} + \frac{0{,}89(2-0{,}11)}{0{,}11^2}166{,}48 = 17643{,}97. \end{aligned}$$

Stejným způsobem určíme počáteční exponenciální průměry druhého a třetího stupně.

$$S_0^{[2]} = \hat{b}_0 - 2\frac{\beta}{\alpha}\hat{b}_1 + \frac{\beta(3-2\alpha)}{\alpha^2}\hat{b}_2 = 3045{,}13$$

$$S_0^{[3]} = \hat{b}_0 - 3\frac{\beta}{\alpha}\hat{b}_1 + \frac{3\beta(4-3\alpha)}{\alpha^2}\hat{b}_2 = 43209{,}49.$$

Tabulka 2–11 Předané byty do obecního vlastnictví v ČR v r. 1991–1997 a prognózy do r. 2002

Rok	t	y_t	$\hat{b}_{0t}$	$\hat{b}_{1t}$	$\hat{b}_{2t}$	$\hat{y}_t$
91	1	3689	3405,7036	–1565,047	0,584	3453,595
92	2	1806	2850,1602	–1629,641	–1,923	1840,948
93	3	2217	2971,4179	–1615,542	–1,376	1219,558
94	4	614	2498,4832	–1670,53	–3,509	1355,188
95	5	1548	2774,0422	–1638,491	–2,266	826,198
96	6	1428	2738,6385	–1642,607	–2,426	1134,418
97	7	1858	2865,5018	–1627,857	–1,853	1094,818
98	8	1236,7183	2682,2044	–1649,169	–2,680	1236,718
99	9	1031,6953	2621,7163	–1656,202	–2,953	1031,695
2000	10	964,03771	2601,7552	–1658,523	–3,043	964,038
2001	11	941,7107	2595,1681	–1659,289	–3,073	941,711
2002	12					934,343

Pokračování tabulky 2–11 Předané byty do obecního vlastnictví v ČR v r. 1991–1997 a prognózy do r. 2002

Rok	t	y_t	S_t	$S_t^{[2]}$	$S_t^{[3]}$
91	1	3689	16108,92	28850,35	41629,98
92	2	1806	15901,79	28827,57	41627,47
93	3	2217	15947,00	28832,54	41628,02
94	4	614	15770,67	28813,14	41625,89
95	5	1548	15873,41	28824,44	41627,13
96	6	1428	15860,21	28822,99	41626,97
97	7	1858	15907,51	28828,19	41627,54
98	8	1236,7183	15839,17	28820,68	41626,72
99	9	1031,6953	15816,62	28818,20	41626,44
2000	10	964,03771	15809,18	28817,38	41626,35
2001	11	941,7107	15806,72	28817,11	41626,32
2002	12				

Z počátečních exponenciálních průměrů můžeme určit pomocí výrazů (2.75) aktuální exponenciální průměry $S_t, S_t^{[2]}, S_t^{[3]}$ pro $t = 1$, tj.

$$S_1 = \alpha y_1 + \beta S_0 = 0,11(3689) + 0,89(17643,97) = 16108,92$$

$$S_1^{[2]} = \alpha S_1 + \beta S_0^{[2]} = 0,11(16108,92) + 0,89(3045,13) = 28850,35$$

$$S_1^{[3]} = \alpha S_1^{[2]} + \beta S_0^{[3]} = 0{,}11\left(28850{,}92\right) + 0{,}89\left(43209{,}49\right) = 41629{,}98.$$

Z aktuálních exponenciálních průměrů pro $t = 1$, pomocí výrazů (2.76), můžeme vypočítat proměnlivé parametry modelu (2.74), tj.

$$\hat{b}_{01} = 3S_1 - 3S_1^{[2]} + S_1^{[3]} = 3\left(16108{,}92\right) - 3\left(28850{,}35\right) + 41629{,}98 = 3405{,}704.$$

Podobně

$$\hat{b}_{11} = \frac{\alpha}{2\beta^2}\left[\left(6-5\alpha\right)S_1 - 2\left(5-4\alpha\right)S_1^{[2]} + \left(4-3\alpha\right)S_1^{[3]}\right] = -1565{,}05$$

$$\hat{b}_{21} = \frac{\alpha^2}{\beta^2}\left(S_1 - 2S_1^{[2]} + S_1^{[3]}\right) = 0{,}584.$$

Pokud máme vypočítané proměnlivé parametry v čase $\hat{b}_{0t}, \hat{b}_{1t}, \hat{b}_{2t}$ v čase $t = 1$, můžeme na základě výrazu (2.77) určit prognózu o jedno období dopředu, tj. s horizontem $\tau = 1$, konstruovanou na začátku předpovědi $t = 1$, tj.

$$\hat{y}_t\left(\tau\right) = \hat{y}_1\left(1\right) = \hat{y}_{1+1} = \hat{y}_2 = \hat{b}_{01} + \hat{b}_{11}\tau + \hat{b}_{21}\frac{\tau^2}{2} \qquad (2.79)$$

$$= 3405{,}706 + \left(-1565{,}05\right)\ 1 + 0{,}584\,\frac{1}{2} = 1840{,}948.$$

Vidíme, že v modelech exponenciálního vyrovnávání pro aktuální hodnotu y_t je předpověď $\hat{y}_t$ konstruována výrazem (2.79) vždy pro hodnoty proměnlivých parametrů v čase $t - 1$ a vždy s horizontem $\tau = 1$.

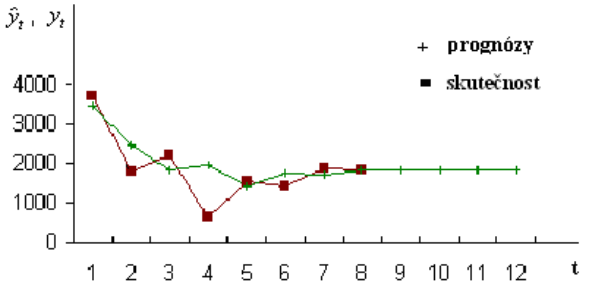
Ze srovnání výrazů (2.55) a (2.57) vyplývá, že $\hat{y}_t$ je současně S_{t-1}. Konkrétně v našem příkladu $\hat{y}_2 = 1\,840{,}948 = S_1$ je nová hodnota, na jejímž základě, z výrazů (2.75), můžeme určit všechny exponenciální průměry v čase $t = 2$ a celý cyklus aktualizace opakovat a určit prognózy až do periody $t = 8$ (tj. včetně roku 1998).

Abychom mohli konstruovat předpovědi pro $t = 9$ až 12, tj. pro roky 1999 až 2002, potřebujeme znát skutečnosti od roku 1998 až do r. 2001, tj. od periody $t = 8$ až do periody $t = 11$. V konstrukci předpovědí můžeme pokračovat, pokud nahradíme neznámé skutečnosti jejich prognózami, tj. prognózami $\hat{y}_8$ až $\hat{y}_{11}$. Tento postup ukážeme na příkladu výpočtu prognózy pro periodu $t = 9$.

Prvním krokem je výpočet $\hat{b}_{08}, \hat{b}_{18}, \hat{b}_{28}$, co vyžaduje hodnoty $S_8, S_8^{[2]}, S_8^{[3]}$. Tyto vypočítáme jako

$$S_8 = \alpha\hat{y}_8 + \beta S_7 = 0{,}11\left(1236{,}718\right) + 0{,}89\left(15907{,}51\right) = 15839{,}17$$

$$S_8^{[2]} = \alpha S_8 + \beta S_7^{[2]} = 0{,}11\left(15839{,}17\right) + 0{,}89\left(28828{,}19\right) = 28820{,}68$$



Obrázek 2–8 Trojité exponenciální vyrovnání – počty odevzdaných bytů do obecného vlastnictví (α = 0,11)

$$S_8^{[3]} = \alpha S_8^{[2]} + \beta S_7^{[3]} = 0{,}11\left(28820{,}68\right) + 0{,}89\left(41627{,}54\right) = 41626{,}72$$

a $\hat{b}_{08}, \hat{b}_{18}, \hat{b}_{28}$ vypočítáme jako

$$\hat{b}_{08} = 3S_8 - 3S_8^{[2]} + S_8^{[3]} = 3\left(15839{,}17\right) - 3\left(28820{,}68\right) + 41626{,}72 = 2682{,}2044.$$

Stejným způsobem určíme

$$\hat{b}_{18} = \frac{\alpha}{2\beta^2}\left[\left(6-5\alpha\right)S_8 - 2\left(5-4\alpha\right)S_8^{[2]} + \left(4-3\alpha\right)S_8^{[3]}\right] = -1649{,}169$$

$$\hat{b}_{28} = \frac{\alpha^2}{\beta^2}\left(S_8 - 2S_8^{[2]} + S_8^{[3]}\right) = -2{,}680.$$

Nakonec předpověď o jedno období dopředu s počátkem předpovědí t = 8 s horizontem τ = 1 je

$$\hat{y}_8(\tau) = \hat{y}_8(1) = \hat{y}_9 = \hat{b}_{08} + \hat{b}_{18}\tau + \hat{b}_{28}\frac{\tau^2}{2}$$

$$= 2682{,}2044 + \left(-1649{,}169\right)\ 1 + \left(-2{,}680\right)\frac{1}{2} = 1031{,}695.$$

Výsledky postupu jsou uvedené v tabulce 2–11 a na obrázku 2–8. Na obrázku 2–8 je vidět, že trojité exponenciální vyrovnání pro časovou řadu odevzdaných bytů ve veřejném sektoru dobře vystihuje vývoj pozorování tohoto časové řady.

Co se týče techniky exponenciálního vyrovnávání pro modely vyšších stupňů, je třeba říci, že se v ekonomice téměř nevyskytují. V případě zájmu čtenář nalezne jejich řešení v pracích Brown (1963).

2.4 ARMA modely a modely přenosových funkcí

Doposud seodely regresní analýzy zakládaly na předpokladech o nezávislosti náhodných poruch a tím i pozorování vysvětlování proměnné zkoumané časové řady. Uvedený předpoklad o nezávislosti náhodných poruch a nezávislosti

pozorování časové řady $\{y_t\}$, například od předchozích pozorováních, není v mnoha praktických případech dodržen.

V této podkapitole se budeme zabývat v první části modelováním závislostí jednoduchých procesů, tj. identifikováním závislosti zkoumané veličiny na pozorování v předchozích obdobích a na časové závislosti náhodné složky procesu. Tyto modely lze chápat jako jednoproměnné modely časových řad, protože na popis vývoje zkoumané veličiny se požívají jen její minulé pozorovaní. Tyto modelovací techniky byly předmětem výzkumu v nedávné minulosti a byly prvotně konsolidovány a prezentovány podle Boxe a Jenkinse (Box a Jenkins, 1970) pod názvem ARMA, resp. ARIMA modely. Ve druhé části budeme předpokládat, že časová řada hodnot výstupní veličiny y_t je v relaci s jednou nebo více hodnotami časových řad jiných (vstupních) veličin $x_{j,t}$. Tyto modely nazýváme modely přenosových funkcí.

2.4.1 ARMA modely

Výchozím bodem modelování diskrétní stacionární časové řady $\{ y_t \}$ s nulovou střední hodnotou v každém bodě je předpoklad o lineární závislosti jeho hodnoty y_t v čase *t* na jeho předchozích hodnotách a na aktuální hodnotě a předchozích hodnotách náhodné složky, tj.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ ... \ + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ .... \tag{2.80}$$

V modelu (2.80) se koeficienty $(\varphi_1, \varphi_2, \ ...)$ nazývají autoregresní parametry, koeficienty $(\theta_1, \theta_2, ...)$ se nazývají parametry procesu klouzavých průměrů a $(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2} \ ...)$ jsou nezávislé náhodné poruchy, o nichž se předpokládá, že jsou generovány procesem s nulovou střední hodnotou, konstantním rozptylem σ^2 s normálním rozdělením. Posloupnost $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2} \ ...\}$ s uvedenými vlastnostmi se nazývá proces bílého šumu. První část modelu (2.80) se nazývá deterministická nebo autoregresní část a druhá se nazývá náhodná část nebo část klouzavých průměrů modelu. Počet členů (proměnných) autoregresní části *p* určuje stupeň autoregresního modelu (procesu) a označuje se jako AR (*p*) model, počet členů náhodné části *q* určuje stupeň náhodné části a označuje se jako MA (q) model nebo proces. Model podle (2.80) určuje smíšený ARMA(*p, q*) = ARMA(∞, ∞) proces. V praktických aplikacích většinou stupně *p, q* nebývají větší než 2.

Identifikace ARMA modelů

Důležitou a jednou ze základních charakteristik používaných při analýze časových řad, zejména při vyhledávání mechanismu, který generuje jeho hodnoty, kromě kovariační funkce, je autokorelační funkce (ACF) ρ_k a parciální autokorelační funkce (PACF) ϕ_{kk}. Výběrová ACF je definována

Tabulka 2–12 Charakteristiky teoretické ACF a PACF

ACF	PACF	Model
pozvolně utlumována	Náhle utlumena po posunu p	AR(p)
náhle utlumena po posunu q	pozvolně utlumována	MA(q)
pozvolně utlumována	pozvolně utlumována	ARMA(p, q)

$$r_k = \frac{\sum_{t=1}^{N-k}\left(y_t - \bar{y}\right)\left(y_{t+k} - \bar{y}\right)}{\sum_{t=1}^{N}\left(y_t - \bar{y}\right)^2}, \; k = 0, 1, ..., K. \tag{2.81}$$

Druhou důležitou charakteristikou ARMA procesů je PACF. Na usuzování skutečného průběhu PACF použijeme parciální autokorelační funkci, jejíž hodnoty $\hat{\phi}_{kk}$ pro $k = 1, 2, ..., K$ určíme z Yule-Walkerovych rovnic (2.82).

$$r_j = \hat{\varphi}_{k1} r_{j-1} + \hat{\varphi}_{k2} r_{j-2} + \; ... + \hat{\varphi}_{kk} r_{j-k}, \;\; \text{pro } j = 1, 2, ..., k. \tag{2.82}$$

Parciální autokorelační funkce (PACF) je definována jako korelace mezi dvěma náhodnými veličinami, přičemž vliv všech ostatních veličin je buď odstraněn, nebo je konstantní (fixován). Parciální korelační koeficient ϕ_{kk}, jako konkrétní hodnota PACF, je mírou vztahu mezi dvěma stacionárními časovými řadami $\{y_t\}$ a $\{y_{t+k}\}$, pokud je eliminován vliv ostatních proměnných $y_{t+1}, y_{t+2}, ..., y_{t+k-1}$. Studium průběhu PACF umožňuje identifikovat počet členů autokorelační části modelu (2.80). Pokud jsme vypočítali hodnoty ACF a PACF z výběru dat a případně graficky zobrazili jejich průběhy, poté na základě poznatků o teoretických průbězích konkrétních typů ARMA procesů a získaných charakteristik procesů z výběru, můžeme předběžně usuzovat na typ ARMA procesu. Charakteristiky ACF a PACF nejčastěji vyskytujících se ARMA procesů jsou shrnuty v tabulce 2–12.

V tabulce 2–12 nejsou uvedeny všechny možné kombinace průběhy výběrových ACF a PACF. Naskýtá se otázka, jak postupovat v takových případech. Řešení poskytují některé numerické algoritmy či rozhodovací kritéria, mezi která patří např. Akaikovo informační kritérium a bayesiánské informační kritérium.

Odhad parametrů

Nejvíce používanou metodou pro odhad parametrů nelineárních modelů je metoda maximální věrohodnosti. Algoritmus pro určení odhadů parametrů metodou maximální věrohodnosti je iterativní a jako každý iterativní algoritmus, vyžaduje počáteční odhad parametrů. Odhad parametrů typu ARMA se proto provádí ve 2 krocích. V prvním kroku se odhadnou počáteční (předběžné) hodnoty parametrů.

Ve druhém kroku se počáteční hodnoty použijí k iterativnímu vyčíslení konečných hodnot parametrů pomocí metody maximální věrohodnosti.

Počáteční odhad parametrů je nejjednodušší pro AR(p) model

$$y_t = \xi + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ \ldots\ + \varphi_p y_{t-p} + \varepsilon_t,\ t = 1, 2, \ldots, N,$$

který je lineární z hlediska parametrů. Na počáteční odhad parametrů lze použít několik metod. Jelikož AR(p) model je lineární, na odhad jeho parametrů lze použít jednoduchou metodu nejmenších čtverců – OLS (Ordinary Least Squares). Estimátor OLS používá přímo výběry dat časových řad. Např. pro model AR (2) ve tvaru $y_t = \xi + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$.

Estimátor na odhad parametrů má tvar $\boldsymbol{\phi} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, kde $\boldsymbol{\phi}' = (\ \hat{\xi}\ \ \hat{\varphi}_1\ \ \hat{\varphi}_2\ \ldots, \hat{\varphi}_p\ )$ je vektor parametrů, $\mathbf{y}' = (y_3, y_4, \ldots, y_N)$ je vektor pozorovaní závislé proměnné (pravá strana AR(2) rovnice) a $\mathbf{X}$ je matice pozorovaní nezávislých proměnných (proměnných na pravé straně AR(p) rovnice).

Pro počáteční odhad parametrů pro MA modely se nejčastěji používá tzv. *dlouhá* AR *metoda*. Předpokladem jejího uplatnění je, aby MA(p) proces byl invertibilní do AR procesu. Invertibilní MA (q) proces může být aproximován AR(j) procesem pro dostatečně velké j, tj.

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ \ldots\ + \varphi_j y_{t-j} + \varepsilon_t. \tag{2.83}$$

Počáteční odhad parametrů MA(q) modelu se provádí ve dvou krocích. V prvním kroku se MA(q) model zamění (invertuje) za model AR(j) a odhadnou se jeho parametry OLS metodou. Pokud známe odhady parametrů $\{\varphi_j\}$, potom z dat časové řady $\{ y_t \}$ můžeme určit odhady hodnot náhodného člena ε_t AR(j) modelu (2.83) jako rezidua e_t, tj.

$$e_t = y_t - \left(\hat{\varphi}_1 y_{t-1} + \hat{\varphi}_2 y_{t-2} + \ \ldots\ + \hat{\varphi}_j y_{t-j}\right). \tag{2.84}$$

Ve druhém kroku dosadíme rezidua $\{ e_t \}$ do MA(q) procesu, čímž získáme model

$$y_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ \ldots\ + \theta_q e_{t-q} + \varepsilon_t \tag{2.85}$$

a znovu v regresní rovnici (2.85) na odhad parametrů lze použít OLS metodu, čímž získáme počáteční odhady $\{\hat{\theta}_i\}$ MA(q) procesu.

Na počáteční odhad parametrů ARMA(p, q) modelu je také možné použít dlouhou AR metodu, neboť ARMA(p, q) proces může být nahrazen AR(j) procesem pro dostatečně velké j. Analogicky jako při odhadu parametrů MA(q)

modelu v jejím prvním kroku výrazem (2.84) získáme rezidua, které ve druhém kroku dosadíme do ARMA(*p*, *q*) modelu, čímž získáme regresní rovnici ve tvaru

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} + \varepsilon_t . \quad (2.86)$$

Uplatněním metody OLS na regresní rovnici (2.86) získáme odhady parametrů $\{\hat{\phi}_j\}$ a $\{\hat{\theta}_i\}$, čím je ARMA(*p*, *q*) aproximovaný na data časové řady $\{ y_t \}$.

Konečný odhad parametrů jak pro AR, tak pro MA a ARMA modely se provádí metodou maximální věrohodnosti (Maximum Likelihood – ML).

Diagnostická kontrola

V modelovacím přístupu, ve kterém se vyhledává mechanismus, kterým jsou nejlépe reprodukována pozorována data, tj. v modelech typu ARMA, kromě uvedené základní míry těsnosti závislosti, přistupují další kritéria, založené na analýze a testování reziduí. Rezidua by měla být časově stabilní, náhodné proměnné s přibližně normálním rozdělením s nulovou střední hodnotou a konstantním rozptylem. Znamená to, že dříve než se kvantifikovaný model použije na konstrukci prognóz, je nezbytné provést testy, které s konečnou platností potvrdí jeho adekvátnost. Označíme rezidua $\{e_t\}$, můžeme je určit z kvantifikovaného ARMA modelu jako

$$e_t = y_t - \hat{y}_t = \hat{\theta}^{-1}(B)\hat{\varphi}(B) y_t, \text{ pro } t = 1, 2, ..., N, \quad (2.87)$$

kde $\hat{\theta}(B)$ a $\hat{\phi}(B)$ jsou odpovídající polynomy z odhadnutých parametrů $\{\hat{\phi}\}$ a $\{\hat{\theta}\}$.

Označme výběrovou ACF reziduí (2.87) jako $\{r_e(k)\}$, kde $r_e(k)$ je

$$r_e(k) = \frac{\sum_{t=1}^{N-k} e_t e_{t+k}}{\sum_{t=1}^{N} e_t^2}, \quad k = 0, 1, ..., K. \quad (2.88)$$

Pokud ARMA model byl správně identifikován, pak $\{r_e(k)\}$ nemohou vykazovat žádnou skrytou strukturu v závislostech reziduí, tj. hodnoty $\{r_e(k)\}$ musí být statisticky nulové, a pak model můžeme použít na konstrukci předpovědí. V opačném případě se celý vývoj počínaje identifikací modelu musí přehodnotit.

Konstrukce předpovědí

V průběhu všech předchozích modelovacích etap jsme použili časovou historii jejich hodnot. Identifikovaný, kvantifikovaný a především ověřený model můžeme při určitých pravidlech úspěšně použít na konstrukci předpovědí.

Jak jsme uvedli v první kapitole, při konstrukci předpovědí lze vycházet z různých předpokladů. Jiné předpoklady, podmínky a metody budou dávat jiné

hodnoty předpovědí. Naším cílem bude aplikovat ARMA modely na odhady budoucích hodnot časové řady, které budou mít minimální průměrnou čtvercovou chybu. Pro konstrukci předpovědí jsou k dispozici procedury, které jsou obsaženy téměř ve všech statistických programových balících. S jednotlivými způsoby a detailními postupy výpočtů předpovědí na konkrétních aplikačních příkladech je možné se seznámit v práci (Marček a Marček, 2001).

2.4.2 Modely přenosových funkcí

Jak jsme uvedli, při modelech přenosové funkce se předpokládá, že časová řada hodnot výstupní veličiny y_t je v relaci s jednou nebo více hodnotami časových řad jiných (vstupních) veličin $x_{j,t}$. Jednoduchý relační model mezi dvěma časovými řadami může mít formu

$$y_t = w_0 x_t + w_1 x_{t-1} + \ .... \ + w_k x_{t-k} + u_t \tag{2.89}$$

a v případě, že existuje *m* vstupů

$$\begin{aligned} y_t = {} & w_0 x_{1,t} + w_1 x_{1,t-1} + \ .... \ + w_k x_{1,t-k} + v_0 x_{2,t} + v_1 x_{2,t-1} + \ .... \ + v_l x_{2,t-l} + \ .... \ + \\ & p_0 x_{m,t} + p_1 x_{m,t-1} + \ .... \ + p_i x_{m,t-h} + u_t, \end{aligned} \tag{2.90}$$

v kterých $\boldsymbol{w'} = (w_0, w_1, ..., w_k)$, $\boldsymbol{v'} = (v_0, v_1, ..., v_l)$, $\boldsymbol{p'} = (p_0, p_1, ..., p_i)$ jsou vektory vah nebo neznámých parametrů modelu, u_t je náhodná složka s normálním a nezávislým rozdělením, nulovou střední hodnotou a konstantním rozptylem ve všech pozorováních. Rovnice (2.89) a (2.90) se nazývají modely přenosových funkcí. Pokud rovnici (2.89) napíšeme pomocí obvyklého zpětného-posunového operátora *B*, dostaneme tvar

$$y_t = w_0 x_t + w_1 x_{t-1} + \ .... \ + w_k x_{t-k} = \left(w_0 + w_1 B + w_2 B^2 + ...\right) x_t = w(B) x_t. \tag{2.91}$$

Metodologie a základní principy modelování a prognózování časových řad pomocí přenosových funkcí jsou popsány v Montgomery a kol. (1990), Granger a Newbold (1986), Helmer a Johansen (1977). V další části uvedeme některé způsoby zápisů lineárních modelů přenosových funkcí.

Rovnice (2.91) se nazývá lineární filtr. Polynom $w(B)$ se nazývá přenosová funkce tohoto filtru. V rovnici (2.91) aktuální hodnota výstupní veličiny je závislá od aktuální hodnoty vstupní veličiny a nekonečného počtu předchozích hodnot vstupní veličiny. Přenosová funkce, resp. její charakteristiky jsou determinovány váhami $w_0, w_1, w_2, ...$ přenosové funkce. Nazývají se impulsní odezvy systému. Po úpravě z rovíc (2.91) lze získat model přenosové funkce v zápisu ve tvaru

$$y_t = \frac{\omega_0 - \omega_1 B - \omega_1 B^2 - \ ... \ - \omega_s B^s}{1 - \delta_1 B - \delta_2 B^2 - \ ... \ - \delta_r B^r} x_{t-b} \ = \ \delta_r^{-1}(B) \omega_s (B) x_{t-b}. \tag{2.92}$$

Identifikace modelů přenosových funkcí

Hodnoty výstupní stacionární časové řady jsou náhodné proměnné, což je vyjádřeno náhodnou složkou u_t v modelu (2.92)

$$y_t = \delta_r^{-1}(B)\omega_s(B)x_{t-b} + u_t, \tag{2.93}$$

kde y_t a x_t jsou stacionární časové rady výstupu, resp. vstupu s nulovými středními hodnotami. Pro časovou řadu náhodné složky u_t modelu (2.93) budeme požadovat, aby jeho hodnoty byly generované ARMA(*p, q*) procesem, tj.

$$\varphi_p(B)u_t = \theta_q(B)\varepsilon_t, \tag{2.94}$$

kde ε_t je náhodná složka tohoto procesu s normálním rozdělením, s nulovou střední hodnotou a konstantním rozptylem ve všech pozorováních (bílý šum). Pokud dosadíme do rovnice (2.93) u_t, které určíme z modelu (2.94), získáme výraz pro hodnoty výstupního časové řady

$$y_t = \delta_r^{-1}(B)\omega_s(B)x_{t-b} + \varphi_p^{-1}(B)\theta_q(B)\varepsilon_t. \tag{2.95}$$

Pod identifikací modelů přenosových funkcí rozumíme specifikaci přenosové funkce, přesněji určení její parametrických hodnot *b, r, s* funkcí $\delta_r^{-1}(B)\omega_s(B)_{t-b}$ v modelu (2.95). Postup získání hodnot ukazatelů je složitější a čtenáře odkazujeme na lit. Mongomery a kol. (1990), resp. Marček a Marček (2001).

Odhad parametrů, diagnostická kontrola

Je vidět, že modely přenosových funkcí jsou modely s velkým počtem parametrů a s velkým rozsahem výběru. Model přenosové funkce modelu (2.93) a model jeho náhodné složky (2.94) je nelineární z hlediska svích parametrů. Budeme předpokládat, že kvantifikací modelu získáme odhady těchto parametrů. Vychází se z předpokladu, že náhodná složka ε_t schématu (2.94) má normální rozdělení s nulovou střední hodnotou a konstantním rozptylem $(\varepsilon_t \sim N(0, \sigma^2))$. Odhad parametrů modelu (2.95) můžeme získat metodou maximální (Gaussovy) věrohodnosti (Brockwel a Davis, 1987). Při používání metody maximální věrohodnosti je důležitá volba počátečních hodnot parametrů. Je známá skutečnost o časové náročnosti odhadu parametrů touto metodou při větším počtu parametrů a velkých výběrech.

Pro odhad parametrů je výhodnější použít metodu podmíněné maximální věrohodnosti, kterou se minimalizuje podmíněna částka čtverců funkce

$$\ell(\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{N} \varepsilon_t^2 \,(\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta} \mid b, \boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{\varepsilon}_0). \tag{2.96}$$

Konstrukce předpovědí

Výchozí model pro konstrukci předpovědí časové řady $\{ y_t \}$ pomocí modelu přenosových funkcí je model daný výrazem (2.93), který je možno vyjádřit i ve tvaru

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \sum_{j=0}^{\infty} w_j \alpha_{t-j}$$

kde x_t je dané výrazem $x_t = \theta_x(B)\phi_x^{-1}(B)\alpha_t$ a $\psi(B) = \varphi_p^{-1}(B)\theta_q(B)$.

Předpověď založená na nekonečné řadě pozorování konstruovanou z počátku předpovědí *N* pro horizont τ možno určit jako očekávanou hodnotu (Brockwel a Davis, 1987), tj.

$$\hat{y}_N(\tau) = \sum_{j=\tau}^{\infty} \psi_j^* \varepsilon_{N+\tau-j} + \sum_{j=\tau}^{\infty} w_j \alpha_{N+\tau-j}.$$

2.5 Support Vector (SV) regresní model

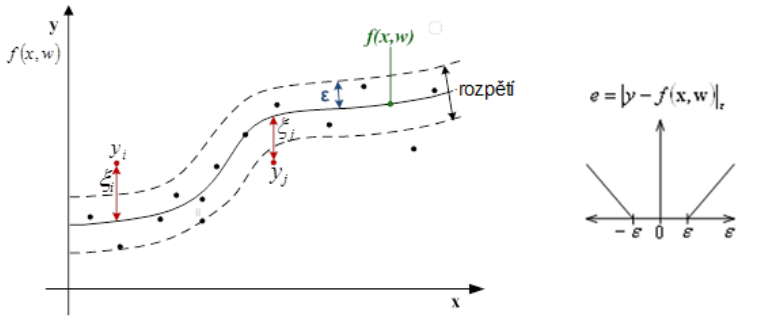
SV regresní model patří do kategorie metod statistického, resp. strojového kontrolovaného (supervizovaného) učení s názvem SVM (Support Vector Machine learning). V porovnání s klasickými regresními modely poskytuje přijatelnější řešení (ve smyslu přesnosti aproximačních funkcí a lepších zevšeobecňujících vlastností), a to zvláště pro malé trénovací množiny. SVM je též metoda jak na učení (trénování) neuronových sítí, modelů založených na teorii fuzzy množin i na specifikaci a kvantifikaci klasických polynomiálních (regresních) modelů. SVM na rozdíl od vícevrstvé sítě perceptronového typu poskytuje globální a jedinečné řešení prostřednictvím algoritmu konvexního kvadratického programování. Metoda byla uvedena v práci Vapnika (1995). V této kapitole se budeme věnovat SV regresnímu modelu, tj. modelu pro vyhledání a aproximace vstupních-výstupních funkcí systémů.

2.5.1 Model SV regrese

Jak jsme se zmínili v úvodu této kapitoly, SVM metoda může být aplikována i pro lineární a nelineární regresi s postupy, zásadami a podmínkami metody SVM pro klasifikaci dat. Mezi podmínky patří existence trénovací množiny dat a výstupních vzorů $D = \{ [x_i, y_i] \in \Re^p \times \Re, \quad i = 1,2,...,n \} \subseteq (X \times Y)^n$, kde $\mathbf{x}$ jsou *p*-dimenzionální vstupní vektory, $\mathbf{y} \in \Re$ jsou odezvy systému (výstupy) se spojitými hodnotami a vyhledání lineární/nelineární regresní funkce ve tvaru

$$f(\mathbf{x}, \mathbf{w}, b) = \begin{cases} \sum_{i=1}^{n} w_i \mathbf{x} + b & \text{pro lineární regresi,} \\ K(\mathbf{x}_i, \mathbf{x}_j)\mathbf{w} + b & \text{pro nelineární regresi,} \end{cases} \tag{2.97}$$

kde $K(\mathbf{x}_i, \mathbf{x}_j)$ jsou příslušné jádrové funkce. Mezi nejčastější kandidátní funkce pro patří:



Obrázek 2–9 Vlevo SVM nelineární regrese (viz text pro detaily), vpravo ztrátová funkce zavedená Vapnikem

$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ lineární pro lineární SVM,

$K(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i^T \mathbf{x}_j + 1)]^d$ polynomiální stupně *d* pro polynomické SVM,

$K(\mathbf{x}_i, \mathbf{x}_j) = \exp[-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2]$ Gaussova pro RBF SVM,

$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh[(\mathbf{x}_i^T \mathbf{x}_j + b)]$ síť perceptronového typu s jednou skrytou vrstvou.

Symboly w_i, b jsou označené parametry, které jsou předmětem učení pomocí lineárního učicího stroje. Chyba regrese nebo reziduum, které označíme symbolem *e* je dána rozdílem mezi naměřenou nebo pozorovanou hodnotou y_i a výstupem, tj. odhadovanou hodnotou z regresní funkce (viz. obrázek 2–9). V levé části obrázku 2–9 jsou odhadnuté hodnoty zakresleny plnou čarou. Na ohodnocení kvality regresní funkce bylo zavedeno nad a pod teoretickou regresní přímku pásmo o šířce označenou symbolem ε.

Jde o pásma malých chyb (ε-rezidua). Tato pásma jsou nakresleny na obrázku 2–9 přerušovanou čárou. Chyby spadající do pásma malých chyb se zanedbávají, v důsledku čehož dochází k ztrátě přesnosti aproximace, zatímco velké hodnoty reziduálů se snaží metoda odstranit. Na ohodnocení míry ztráty přesnosti aproximace se využívá tzv. ztrátová funkce zavedená Vapnikem (1995), která je definovaná

$$|y - f(\mathbf{x}, \mathbf{w})|_\varepsilon = \begin{cases} 0 & \text{jestli } |y - f(\mathbf{x}, \mathbf{w})|_\varepsilon \le \varepsilon, \\ |y - f(\mathbf{x}, \mathbf{w})|_\varepsilon - \varepsilon & \text{jinak,} \end{cases} \quad (2.98)$$

kde $|y - f(\mathbf{x}, \mathbf{w})|_\varepsilon$ je reziduum vzhledem na okraje pásma, které graficky tvar je vidět na obrázku 2–9 vpravo.

2.5.2 Odhad parametrů

Při formulování SVM algoritmu pro odhad parametrů se vychází ze současného minimalizování funkce rizika (2.99)

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} |y_i - f(\mathbf{x}, \mathbf{w})|_{\varepsilon}. \tag{2.99}$$

a normy $\|\mathbf{w}\|^2$, kde

$$\|\mathbf{w}\|^2 = \frac{1}{2} \mathbf{w}^T \mathbf{w} \tag{2.100}$$

vzhledem k omezení

$$\begin{cases} y_i - w^T x_i - b \le \varepsilon, \\ w^T x_i + b - y_i \le \varepsilon, \end{cases} \tag{2.101}$$

tj. minimalizování funkce se zápisem

$$\min_{\mathbf{w}, b, \xi, \xi^*} R(w, \xi, \xi^*) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \tag{2.102}$$

vzhledem k omezení

$$\begin{cases} y_i - \mathbf{w}^T \mathbf{x} - b \le \varepsilon + \xi_i & i = 1, 2, \ ..., \ n, \\ \mathbf{w}^T \mathbf{x} + b - y_i \le \varepsilon + \xi_i^* & i = 1, 2, \ ..., \ n, \\ \xi, \xi_i^* \ge 0 & i = 1, 2, \ ..., \ n, \end{cases} \tag{2.103}$$

kde ξ a ξ^* jsou volné veličiny zakreslené na obrázku 2–9 vlevo. Volitelnou hodnotou konstanty C se penalizují chyby ξ a ξ^*. Zvyšováním hodnoty C se penalizují větší chyby ξ a ξ^*, což vede ke snížení chyby aproximace. Avšak toho je možné dosáhnout jen při zvyšování normy váhového vektoru $\|\mathbf{w}\|$. Zároveň zvyšováním normy $\|\mathbf{w}\|$ není všeobecně zaručená dobrá optimalizace a přesnost modelu (Kecman, 2001). Druhou možností ovlivňování přesnosti modelu je volba šířky pásma volitelným parametrem ε.

Při řešení úlohy (2.102) s omezeními (2.103) se vychází z konstrukce Lagrangeovy funkce v primárních proměnných L_p ve tvaru

$$L_p(\mathbf{w}, b, \xi, \xi_i, \alpha_i, \alpha_i^*, \beta_i, \beta_i^*)$$

$$= \frac{1}{2} w^T w + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{n}\alpha_i \left(\varepsilon + \xi_i - y_i + \mathbf{w}^T\mathbf{x}_i + b\right) \tag{2.104}$$

$$-\sum_{i=1}^{n}\alpha_i^* \left(\varepsilon + \xi_i^* + y_i - \mathbf{w}^T\mathbf{x}_i - b\right) - \sum_{i=1}^{n}\left(\beta_i\xi_i + \beta_i^*\xi_i^*\right)$$

s hodnotami $\alpha_i, \alpha_i^* \geq 0$, $\xi_i, \xi_i^* \geq 0$, i = 1, 2, ..., n. Řešení je dané sedlovými body z Lagrangeovyé funkce (Fletcher, 1987)

$$\max_{\alpha,\alpha_i^*\beta_i,\beta_i^*} \min_{\mathbf{w},\mathrm{b},\xi_i,\xi_i^*} L_p\left(\mathbf{w}, b, \xi, \xi_i, \alpha_i, \alpha_i^*, \beta_i, \beta_i^*\right) \tag{2.105}$$

vzhledem k podmínce

$$\begin{cases} \dfrac{\partial L_p}{\partial w} = 0 \;\rightarrow\; \mathbf{w} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\mathbf{x}_i, \dfrac{\partial L_p}{\partial b} = 0 \;\rightarrow\; \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0, i = 1,\ldots,n, \\ \dfrac{\partial L_p}{\partial \xi_i} = 0, \dfrac{\partial L_p}{\partial \beta_i} = 0 \;\rightarrow\; 0 \leq \alpha_i \leq C, \;\; i = 1,\ldots,n, \dfrac{\partial L_p}{\partial \xi_i^*} = 0, \dfrac{\partial L_p}{\partial \beta_i^*} = 0 \;\rightarrow\; 0 \leq \alpha_i^* \leq C, \end{cases} \tag{2.106}$$

$$\max_{\alpha,\alpha_i^*} \; -\frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\mathbf{x}_i^T\mathbf{x}_j - \varepsilon\sum_{i=1}^{n}\left(\alpha_i + \alpha_i^*\right) + \sum_{i=1}^{n} y_i\left(\alpha_i - \alpha_i^*\right) \tag{2.107}$$

pro omezení (2.106).

Po vypočítání α_i, α_i^* a jejich dosazením do první rovnice s omezením (2.107), se získá vektor parametrů **w** jako

$$\mathbf{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{2.108}$$

a optimální bias *b* jako

$$\begin{cases} b = y_k - \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)K\left(\mathbf{x}_i, \mathbf{x}_k\right) - \varepsilon \quad \text{pro} \quad \alpha_k \in (0, C), \\ b = y_k - \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)K\left(\mathbf{x}_i, \mathbf{x}_k\right) + \varepsilon \quad \text{pro} \quad \alpha_k^* \in (0, C). \end{cases} \tag{2.109}$$

2.5.3 Konstrukce předpovědí

Východiskovým modelem pro konstrukci předpovědí časové řady $\{y_t\}$ pomocí modelu SV regrese je tvar modelu (2.110), tj.

$$\begin{cases} \hat{f}\left(\mathbf{x},\mathbf{w},b\right)=K\left(\mathbf{x}_i,\mathbf{x}_j\right)\mathbf{w}+b \quad \text{nebo} \\ \hat{f}\left(\mathbf{x},\alpha,b\right)=\sum_{i=1}^{n+\tau}\left(\alpha_i-\alpha_i^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right)+b. \end{cases} \tag{2.110}$$

kde τ je horizont prognózování, α_i,α_i^* jsou známé konstanty, *b* je známy bias $\hat{f}(\mathbf{x})$ jsou bodové odhady předpovědí časové řady y_i, $K(\mathbf{x}_i,\mathbf{x}_j)$ jsou přípustné kernel funkce.