



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Vysoká škola báňská – Technická univerzita Ostrava



PLÁNOVÁNÍ EXPERIMENTŮ

učební text

Josef Tošenovský

Ostrava 20102

Recenze: Ing. Radomír Perzina, Ph.D.
Prof. RNDr. Alena Lukasová, CSc.

Název: Plánování experimentů
Autor: Josef Tošenovský
Vydání: první, 2012
Počet stran: 230
Náklad: 20

Studijní materiály pro studijní obor Řízení jakosti FMMI
Jazyková korektura: nebyla provedena.

Určeno pro projekt:

Operační program Vzdělávání pro konkurenceschopnost
Název: Personalizace výuky prostřednictvím e-learningu
Číslo: CZ.1.07/2.2.00/07.0339
Realizace: VŠB – Technická univerzita Ostrava
Projekt je spolufinancován z prostředků ESF a státního rozpočtu ČR

© Josef Tošenovský
© VŠB – Technická univerzita Ostrava

ISBN 978-80-248-2592-2

	OBSAH	
1.	Úplný faktorový experiment se dvěma úrovněmi	7
1.1	Kódované proměnné	9
1.2	Jednofaktorový plán	10
1.3	Efekt faktoru	11
1.4	Rozptyl odhadu efektu faktoru	17
1.5	Test významnosti efektu	20
1.6	Grafické hodnocení efektu faktorů	21
1.7	Regresní model experimentu	23
1.8	Reziduální odchylky: výpočet, vlastnosti a význam	24
1.9	Metoda nejmenších čtverců v DOE	25
1.10	Grafy interakcí	27
2.	Částečný faktorový experiment se dvěma úrovněmi	37
2.1	Operace s faktory a jejich vlastnosti	38
2.2	Nalezení zaměnitelných dvojic	39
2.3	Stanovení počtu zaměnitelných faktorů	46
2.4	Volba generátorů u středových plánů	47
2.5	Projekce plánu	47
2.6	Doplňkové frakce: tvorba a vlastnosti	49
2.7	Plán experimentu pro osm faktorů	52
3.	Významné body plánu	58
4	Úplné kvadratické modely	61
4.1	Podmínky nalezení modelu	61
4.2	Test křivosti	62
4.3	Metody nalezení kvadratického modelu	63
4.3.1	Kombinovaný plán	63
4.3.2	Výpočet optimální úrovně významných faktorů	64
4.3.3	Box – Behnkenův plán experimentu	70
4.3.4	Plány pro faktory se třemi úrovněmi	71
5.	Bloky	74
5.1	Konstrukce bloků	75
5.2	Hodnocení vlivu bloků	77
6.	ANOVA v DOE	82
6.1	Rozklad reziduálního rozptylu	82
6.2	Test Lack-of-fit	83
7.	Způsoby navrhování robustních výrobků	90
7.1	Minimalizace rozptylu Y	90
7.2	Zařazení šumů do plánu experimentu	96
8.	Dynamické plánování experimentů	107
9.	Kvalitativní proměnné	121
9.1	Konstrukce plánů s kvalitativními proměnnými	121
9.2	Jak působí kvalitativní proměnné v modelu	127
10.	ANOVA	129
10.1	Jednoduché třídění	129
10.2	Dvojné třídění	137
10.3	Trojné třídění (Latinské čtverce)	140

11.	Širší souvislosti plánování experimentů	151
11.1	Regresní analýza	151
11.1.1	Základní pojmy regresní analýzy, metoda nejmenších čtverců	151
11.1.2	Předpoklady regresní analýzy	153
11.1.3	Vybrané věty regresní analýzy	154
11.1.4	Testování předpokladů regresní analýzy	169
11.1.5	Zobecněná metoda nejmenších čtverců	182
11.2	Časové řady	188
11.2.1	Klasická metodika analýzy časových řad	189
11.2.2	Box Jenkinsova metodika analýzy časových řad	201
	Animace	

ÚVOD

Plánování experimentů je jedním z nejúčinnějších nástrojů předvýrobní etapy. Umožňuje nalézt faktory, které nejvýznamněji ovlivňují výrobní proces i jeho výstupy a stanovit také jejich optimální hodnoty. Lze tedy říci, že plánování experimentů je matematickým prostředkem, který umožní výrobcům kvantifikovat významnost vstupů, které jsou na počátku vytypované jako pravděpodobně vlivné. Dále stanoví, jak vybrané vstupy nastavit, aby proces dosahoval požadovaných výstupů při maximální stabilitě (tedy minimální variabilitě) a odolnosti proti tzv. šumům, tj. nepředvídatelným negativním vlivům na výrobní proces.

Předmět plánování experimentů využívá poznatků matematické statistiky i teorie pravděpodobnosti. V tomto textu jsou vyloženy základní pojmy a vztahy. Zájemci o hlubší studium naleznou mnoho dalších poznatků např. v citované literatuře.

1. ÚPLNÝ FAKTOROVÝ EXPERIMENT NA DVOU ÚROVNÍCH

(Full Factorial Design at Two Levels)



Čas ke studiu: 20 hodin



Cílem kapitoly je objasnit

- a) princip sestavení plánu experimentu
- b) několik způsobů výpočtu efektu faktoru
- c) postup při hodnocení významnosti efektů
- d) konstrukci lineárních a kvadratických modelů experimentu
- e) analýzu rozptylu v plánování experimentů
- f) navrhování robustních výrobků
- g) některé speciální postupy plánování experimentů



Výklad

Dříve, než ukážeme na jednoduchém příkladě základy plánování experimentů, vysvětlíme základní pojmy a cíle DOE.

Pod pojmem experimentovat se dále rozumí měnit obvyklé pracovní podmínky s cílem nalézt *nejlepší pracovní* postupy a současně získat hlubší poznatky o vlastnostech výrobku a výrobního procesu.

Experimentální postupy lze rozdělit na

- a) experimenty neplánované (živelné),
- b) experimenty plánované.

Plánované experimenty se řídí plánem experimentu.

Plán experimentu stanovuje 3 charakteristiky („3P“)

- a) *počet* pokusů, ze kterých se experiment skládá,
- b) *podmínky*, za kterých se jednotlivé pokusy uskuteční,
- c) *pořadí* pokusů.

Z uvedeného je zřejmé, že se zde rozlišuje význam pojmů:

pokus = zjištění hodnoty ukazatele kvality za určitých, předem plánovaných, podmínek výroby;

experiment = systém všech pokusů.

Formulaci „nejlepší pracovní postupy“ z úvodního odstavce lze precizovat takto: Označíme-li sledovaný ukazatel kvality Y (resp. ukazatele kvality $Y_1, \dots, Y_k$) a faktory, které jej ovlivňují

A,B,C,D, ... se mohou pohybovat na různých úrovních, řekněme $A_1, A_2, A_3, \dots$ pro faktor A, $B_1, B_2, B_3, \dots$ pro faktor B atd., pak *cílem plánování experimentu* je

- 1) *kvantifikovat míru významnosti* každého z faktorů, což znamená rozhodnout, které z faktorů A,B,C,D, ... rozhodujícím způsobem ovlivňují ukazatel kvality Y,
- 2) *určit úrovně významných faktorů* tak, aby Y bylo *optimální* a *stabilní*.

Příklad 1 ([3], upraveno).

Sleduje se, kolik zmáčknutí (Y) vydrží pružina až do zničení v závislosti na těchto faktorech: L = délka pružiny, G = tloušťka drátu, T = typ materiálu. Má se zjistit, které faktory jsou rozhodující pro životnost pružiny

Řešení:

Sestavíme tabulku faktorů a jejich uvažovaných úrovní (tab.1):

faktor	označení	dolní úroveň	horní úroveň
		-	+
délka pružiny	L	10cm	15cm
tloušťka drátu	G	5mm	7mm
materiál	T	A	B

Tab.1 Seznam faktorů a úrovní

Nyní sestavíme plán experimentu (tab.2)

Existuje více způsobů jak sestavit plán, podle kterého se budou provádět jednotlivé pokusy. Mezi nejpoužívanější plány patří úplný faktorový plán, který v daném případě vypadá takto:

Pokus	L	G	T	Y_i
1	10	5	A	
2	15	5	A	
3	10	7	A	
4	15	7	A	
5	10	5	B	
6	15	5	B	
7	10	7	B	
8	15	7	B	

Tab.2 Plán experimentu

Y_i je výsledek i-tého pokusu.

1.1 Kódované proměnné

Plán experimentu je výhodnější psát pomocí této symboliky: Je-li každý z faktorů uvažován na dvou úrovních, pak dolní úroveň bude značena -1 (resp. jen „-“) a horní úroveň +1 (resp. „+“). Tabulka 2 potom bude mít tvar

Pokus	L	G	T	Y_i
1	-1	-1	-1	
2	+1	-1	-1	
3	-1	+1	-1	
4	+1	+1	-1	
5	-1	-1	+1	
6	+1	-1	+1	
7	-1	+1	+1	
8	+1	+1	+1	

Tab.3 Plán experimentu v kódovaných proměnných

Přepočítání původních proměnných na kódované proměnné se může provést nejen pro krajní hodnoty $x_{\max}$ ($= +1$) a $x_{\min}$ ($= -1$) takto:

$$x_c = \frac{x_o - \frac{x_{\max} + x_{\min}}{2}}{\frac{x_{\max} - x_{\min}}{2}} \quad (1)$$

kde x_o = proměnná x v původních jednotkách,

x_c = kódovaná proměnná,

$x_{\max}$ = horní úroveň x ,

$x_{\min}$ = dolní úroveň x .

Například přepočítání (kódování) L pro dolní hodnotu 10 je

$$L_c = \frac{10 - \frac{15 + 10}{2}}{\frac{15 - 10}{2}} = -1$$

a pro horní úroveň $G = 7$ bude kódovaná hodnota

$$G_c = \frac{7 - \frac{7 + 5}{2}}{\frac{7 - 5}{2}} = +1$$

Pro kvalitativní proměnné se vzorec (1) nepoužívá: tato proměnná nabývá jen dvou hodnot, takže nemá smysl počítat $(x_{\max} + x_{\min})/2$ a ani se nemůže vyskytnout problém s dekodováním jiné hodnoty než +1 resp. -1.

Počet pokusů, ze kterých je sestaven úplný experiment, se vypočítá při k faktorech pomocí vztahu (počet variací k -té třídy ze dvou prvků s opakováním) $n = 2^k$, takže zde, při $k = 3$ faktorech, je počet pokusů (řádků) $n = 2^3 = 8$. Proto má tabulka 8 řádků.

Grafickým znázorněním plánu experimentu je tzv. CPLOT. Pro dva faktory je to čtverec, pro tři faktory krychle se středem v počátku souřadné soustavy. Souřadnice vrcholů odpovídají jednotlivým pokusům (podrobně a graficky viz kap.3).

1.2 Jednofaktorový plán

Jinou možností sestavení plánu je jednofaktorový plán, při kterém se u jednoho faktoru mění úroveň z dolní na horní a u ostatních se drží na střední úrovni, která je průměrem dolní a horní úrovně a značí se symbolem „0“. V našem případě by jednofaktorový plán vypadal takto:

Pokus	L	G	T
1	+	0	0
2	-	0	0
3	0	+	0
4	0	-	0
5	0	0	+
6	0	0	-

Tab.4 Jednofaktorový plán

Počet pokusů je u tohoto typu plánu $n = 2 \cdot k$, zde tedy $n = 2 \cdot 3 = 6$ pokusů (řádků).

Další možností sestavení plánu je částečný faktorový plán, o kterém je pojednáno v kapitole 2.

Jestliže je plánem experimentu stanoveno, za jakých podmínek se provádí jednotlivé pokusy, je možné provést celý experiment a zaznamenat hodnoty sledovaného ukazatele Y . V našem případě byl každý pokus opakován dvakrát. Výsledky jsou v tabulce 5:

pokus	faktor			výsledek		průměr
	L	G	T	Y1	Y2	
1	-	-	-	77	81	79
2	+	-	-	98	96	97
3	-	+	-	76	74	75
4	+	+	-	90	94	92
5	-	-	+	63	65	64
6	+	-	+	82	86	84
7	-	+	+	72	74	73
8	+	+	+	92	88	90

Tab.5 Výsledky opakovaných pokusů

Sestavením tabulky 5 skončily přípravné a experimentální práce. Dále následují už jen výpočty. Jejich cílem bude stanovit, které z faktorů ovlivňují významným způsobem životnost pružiny Y.

Vzhledem k tomu, že některé faktory nemusí být samostatně významné, zatímco v součinnosti s jiným faktorem (tzv. interakci) už významné jsou, doplní se tab. 5 o všechny možné interakce. Teprve potom se počítá významnost faktorů a interakcí.

Znaménka ve sloupcích LG, LT, GT, LGT se získají jako součin znamének v odpovídajících sloupcích (tab.6).

pokus	L	G	T	LG	LT	GT	LGT	$\bar{Y}$
1	-	-	-	+	+	+	-	79
2	+	-	-	-	-	+	+	97
3	-	+	-	-	+	-	+	75
4	+	+	-	+	-	-	-	92
5	-	-	+	+	-	-	+	64
6	+	-	+	-	+	-	-	84
7	-	+	+	-	-	+	-	73
8	+	+	+	+	+	+	+	90

Tab.6 Faktory a jejich interakce

Dále se bude počítat tzv. efekt (vliv) jednotlivých faktorů.

1.3 Efekt faktoru

Efekt faktoru se rozumí změna ukazatele kvality Y, kterou způsobí přechod tohoto faktoru z dolní úrovně (-) na horní úroveň (+).

Výpočet efektu faktoru se může provést různými způsoby.

Uvedeme postupně 5 způsobů, které jsou v různých situacích užitečné:

- a) průměr rozdílů
- b) rozdíl průměrů
- c) znaménková metoda
- d) Yatesova metoda
- e) polovina regresních koeficientů

Průměr rozdílů

Výpočet ukážeme např. pro efekt faktoru L. Podle definice je to změna Y při přechodu L z úrovně „-“ na „+“, přičemž ostatní faktory se nemění. Takové přechody jsou možné čtyři s efekty:

$$97 - 79 = 18$$

$$92 - 75 = 17$$

$$84 - 64 = 20$$

$$90 - 73 = 17$$

Za efekt L vezmeme průměr z těchto hodnot

$$\frac{1}{4}[(97 - 79) + (92 - 75) + (84 - 64) + (90 - 73)] = 18$$

Efekt faktoru L označíme $ef(L) = 18$.

Podobně se vypočítá efekt faktoru G. Efekty jsou

$$\begin{aligned}75 - 79 &= -4 \\92 - 97 &= -5 \\73 - 64 &= 9 \\90 - 84 &= 6\end{aligned}$$

Průměr těchto hodnot je efekt G:

$$ef(G) = \frac{1}{4}(-4 - 5 + 9 + 6) = 1,5$$

Efekt faktoru T: Jsou 4 možné přechody z úrovně „-„ na „+“ s efekty

$$\begin{aligned}84 - 97 &= -13 \\64 - 79 &= -15 \\90 - 92 &= -2 \\73 - 75 &= -2\end{aligned}$$

Průměr těchto hodnot je efekt faktoru T (značí se l_T):

$$ef(T) = \frac{1}{4}(-13 - 15 - 2 - 2) = -8$$

Rozdíl průměrů

Přeskupením hodnot v závorce při výpočtu L můžeme výraz upravit na tvar

$$\frac{1}{4}(97 + 92 + 84 + 90) - \frac{1}{4}(79 + 75 + 64 + 73) = \bar{Y}_+ - \bar{Y}_-$$

V první závorce jsou hodnoty Y pro faktor na horní úrovni. Podobně ve druhé závorce jsou hodnoty Y, když je faktor na své dolní úrovni. Při výpočtu efektu metodou „průměr rozdílů“ se nejprve počítaly rozdíly hodnoty Y na úrovni „+“ a „-“, a z nich pak průměr. Je možné postupovat v opačném pořadí: nejprve vypočítat průměrné Y na horní úrovni ($\bar{Y}_+$) resp. dolní úrovni ($\bar{Y}_-$) a potom jejich rozdíl. Např. pro faktor L

$$\bar{Y}_+ = \frac{97 + 92 + 84 + 90}{4} = 90,75$$

$$\bar{Y}_- = \frac{79 + 75 + 64 + 73}{4} = 72,75$$

$$ef(L) = \bar{Y}_+ - \bar{Y}_- = 90,75 - 72,75 = 18$$

Metoda „rozdíl průměrů“ je užitečná

- při odvození *rozptylu odhadu efektů* faktorů (kap.1.4)
- při konstrukci *grafu efektu* hlavních faktorů (viz obr.1)

Grafy efektů faktorů L,G a T

Nejprve vypočítáme průměrné Y pro dolní (-Y) a horní (+Y) úrovně, potom sestojíme grafy:

	L	G	T	Y
	-	-	-	79
	+	-	-	97
	-	+	-	75
	+	+	-	92
	-	-	+	64
	+	-	+	84
	-	+	+	73
	+	+	+	90
$\bar{Y}_-$	72,75	81	85,75	
$\bar{Y}_+$	90,75	82,5	77,75	

Např. pro faktor L je $\bar{Y}_- = \frac{1}{4}(79 + 75 + 64 + 73) = 72,75$

a $\bar{Y}_+ = \frac{1}{4}(97 + 92 + 84 + 90) = 90,75$.

Efekt L lze vypočítat jako rozdíl $\bar{Y}_+$ a $\bar{Y}_-$, tj. rozdíl Y v koncových bodech grafu: $ef(L) = 90,75 - 72,75 = 18$. Úhel úsečky s osou x ukazuje velikost efektu: nulový efekt odpovídá úsečce rovnoběžné s osou x.

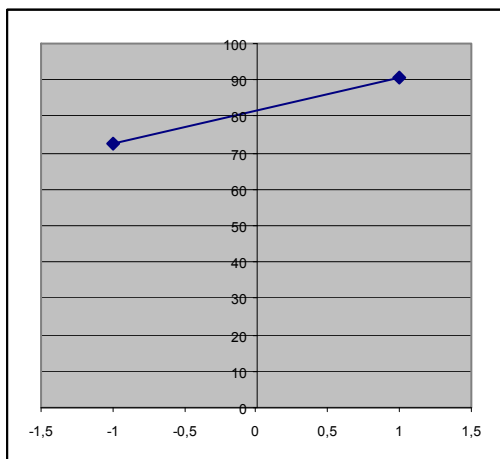
Směr rostoucí znamená, že přechodem z dolní na horní úroveň faktoru vzroste efekt, směr klesající znamená naopak pokles efektu.

Data pro konstrukci grafu:

Faktor	-1	1
	72,75	90,75

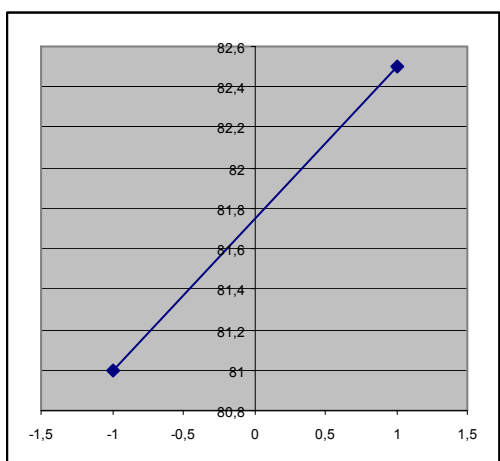
Úsečkový graf pro faktor L: pro L na dolní úrovni je

$\bar{Y}_- = 72,75$ a pro L na horní úrovni $\bar{Y}_+ = 90,75$.



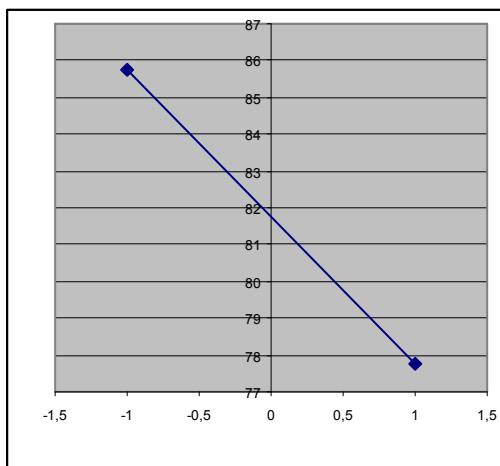
Faktor G

-1	1
81	82,5



Faktor T

-1	1
85,75	77,75



Protože se Y maximalizuje (je to životnost), jsou optimální úrovně +L, +G, -T.

Pozn.:

1. Vždy se odečítá od $ef(\bar{Y}_+)$ efekt $ef(\bar{Y}_-)$.
2. V kapitole o regresních modelech experimentu uvidíme, že u lineárních modelů se hledá optimální úroveň faktorů uvedeným způsobem, zatímco u kvadratických modelů se optimální hodnota faktoru vypočítá.

Znaménková metoda

Metoda výpočtu pomocí průměru z rozdílů Y je při větším počtu faktorů obtížně použitelná. Jednodušší algoritmus získáme (např. při výpočtu efektu T) takto: dosadíme za -13, -15, -2 a -2 výše uvedené rozdíly:

$$ef(T) = \frac{1}{4}[(84 - 97) + (64 - 79) + (90 - 92) + (73 - 75)] = -8$$

Seřadíme-li čísla v závorce tak, jak jsou uvedena v tabulce 5 (ovšem i se znaménky), máme

$$ef(T) = \frac{1}{4}(-79 - 97 - 75 - 92 + 64 + 84 + 73 + 90) = -8$$

Posloupnost těchto čísel a příslušných znamének odpovídá výsledkům pokusů, opatřených znaménky ve sloupci faktoru T. Znaménko „mínus“ znamená, že zvětšováním T se zmenšuje Y.

Obecný návod je tedy tento: sečtou se hodnoty ve sloupci Y, přičemž každá hodnota má znaménko, odpovídající znaménku u příslušného faktoru v odpovídajícím řádku. Součet se vydělí $n/2$, kde n = počet pokusů.

Například pro faktor L bude

$$L = (1/4)(-79+97-75+92-64+84-73+90) = 18.$$

Znaménko „plus“ znamená, že zvětšováním L se zvětšuje Y.

Podobně se postupuje u interakcí, například pro interakci LG znaménkovou metodou bude

$$LG = (1/4)(79-97-75+92+64-84-73+90) = -1.$$

Znaménkovou metodu lze také definovat jako součin dvou vektorů, spočívající ve vynásobení (řádkového) vektoru znamének hodnoceného faktoru se sloupcem (vektorem) výsledků pokusů, např. pro faktor L je to součin vektorů

$$ef(L) = \frac{1}{4}(-1, 1, -1, 1, -1, 1, -1, 1) \cdot \begin{pmatrix} 79 \\ 97 \\ 75 \\ 92 \\ 64 \\ 84 \\ 73 \\ 90 \end{pmatrix} = 18$$

který se dělí polovinou počtu pokusů.

Efekt faktorů a jejich interakcí doplníme do tabulky 7.

čís.	L	G	T	LG	LT	GT	LGT	$\bar{y}$	efekt	faktor
1	-	-	-	+	+	+	-	79	81,75	(1)
2	+	-	-	-	-	+	+	97	18,0	L
3	-	+	-	-	+	-	+	75	1,5	G
4	+	+	-	+	-	-	-	92	-1,0	LG
5	-	-	+	+	-	-	+	64	-8,0	T
6	+	-	+	-	+	-	-	84	0,5	LT
7	-	+	+	-	-	+	-	73	6,0	GT
8	+	+	+	+	+	+	+	90	-0,5	LGT

Tab.7 Efekt faktorů a interakcí

Yatesův algoritmus

Další možností výpočtu efektu faktorů je Yatesův algoritmus.

Postupuje se tak, že ve sloupci (1) je součet 1. a 2. řádku, potom součet 3. a 4. řádku, 5. a 6. řádku, 7. a 8. řádku, následuje rozdíl 2. a 1. řádku, 4. a 3. řádku, 6. a 5. řádku, 8 a 7 řádku. Stejně se postupuje ve sloupci (2) a (3). Číslovaných sloupců je tolik, kolik je faktorů. Efekt je roven podílu čísla v posledním číslovaném sloupci a "dělitele". Posloupnost faktorů a

interakcí v posledním sloupci je pevná a nelze ji změnit. Odtud také pochází seřazení faktorů v posledním sloupci tabulky 7.

Y	(1)	(2)	(3)	dělitel	efekt	faktor
79	176	343	654	8	81,75	průměr
97	167	311	72	4	18,0	L
75	148	35	6	4	1,5	G
92	163	37	-4	4	-1,0	LG
64	18	-9	-32	4	-8,0	T
84	17	15	2	4	0,5	LT
73	20	-1	24	4	6,0	GT
90	17	-3	-2	4	-0,5	LGT

Zde se postupuje tak, že ve sloupci (1) je součet 1. a 2. řádku, potom součet 3. a 4. řádku, 5. a 6. řádku, 7. a 8. řádku, následuje rozdíl 2. a 1. řádku, 4. a 3. řádku, 6. a 5. řádku, 8 a 7 řádku. Stejně se postupuje ve sloupci (2) a (3). Číslovaných sloupců je tolik, kolik je faktorů. Efekt je roven podílu čísla v posledním číslovaném sloupci a "dělitele". Posloupnost faktorů a interakcí v posledním sloupci je pevná a nelze ji změnit. Odtud také pochází seřazení faktorů v posledním sloupci tabulky 7.

1.4 Rozptyl odhadu efektu faktoru

Výsledky jednotlivých pokusů $Y_1, \dots, Y_n$ jsou náhodné veličiny. Vektor $Y = (Y_1, \dots, Y_n)$ je tentýž jako v regresní analýze ($Y = X \cdot \beta + \varepsilon$).

Pro vektor Y bylo dokázáno: $Y \sim N(EY, \text{var } Y)$; předpokládá se, že $\text{var } Y = \sigma^2 I$ což znamená, že všechny náhodné veličiny Y_i mají

- normální rozdělení
- stejný rozptyl σ^2

Předpoklad normality je významný: využívá se dále

- a) při odvození rozptylu odhadu efektu s_e^2 ,
- b) při grafickém hodnocení významnosti efektu.

Nesplnění normality znamená, že není vhodný lineární model. Hledá se tedy kvadratický, kde se jinak se počítají koeficienty (vektor $\mathbf{b}$), jejich rozptyl a jejich testování.

S využitím výpočtu efektu jako rozdílu průměrů můžeme získat charakteristiky efektu. Jestliže každá náhodná veličina Y_i má rozdělení

$$Y_i \sim N(\mu, \sigma^2),$$

pak pro jejich průměr platí

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

a také efekt faktoru (neboť je lineární kombinací $\bar{Y}_+$ a $\bar{Y}_-$) má normální rozdělení s parametry

$$ef = \bar{Y}_+ - \bar{Y}_- \Rightarrow ef \sim N(\mu_e, \sigma_e^2).$$

Střední hodnota

$$\mu_e = E(\bar{Y}_+ - \bar{Y}_-) = \mu - \mu = 0.$$

Odhad rozptylu efektu σ_e^2 (stejný pro všechny faktory) bude:

$$s_e^2 = \frac{4 \cdot \sigma^2}{n} \quad (2)$$

kde n je celkový počet pokusů (včetně opakování).

Rozptylu odhadu efektu vypočítáme takto:

$$D(\text{efektu}) = s_e^2 = D(\bar{Y}_+ - \bar{Y}_-) = D(\bar{Y}_+) + D(\bar{Y}_-) = \frac{\sigma^2}{\frac{n}{2}} + \frac{\sigma^2}{\frac{n}{2}} = 2 \frac{\sigma^2}{\frac{n}{2}} = \frac{4\sigma^2}{n}$$

Máme tedy

$$ef \sim N(0, \frac{4\sigma^2}{n})$$

Rozptyl odhadu efektu se použije k testování významnosti efektu faktorů.

Poznámka:

Při výpočtu charakteristik efektu byly použity vlastnosti disperze náhodné veličiny

$$D(k \cdot X) = k^2 \cdot D(X)$$

a věta o jednom výběru z normálního rozdělení

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n}); i = 1, 2, \dots, n.$$

V případě *opakovaných pokusů* se σ^2 odhadne pomocí veličiny s^2 , která se vypočítá

$$s^2 = \frac{\nu_1 \cdot s_1^2 + \dots + \nu_k \cdot s_k^2}{\nu_1 + \dots + \nu_k} \quad (3)$$

$$\nu_i = n_i - 1,$$

n_i je počet opakování i -tého pokusu, s_i^2 je rozptyl v i -tém pokusu.

s^2 je *vážený rozptyl*.

Pokud se jednotlivé *pokusy neopakují*, je odhad σ^2 prováděn různými způsoby, např.

- a) pomocí tzv. centrálních bodů (kap.3),
- b) jako průměr druhých mocnin efektů nejvyšších interakcí.

Například budou-li sledovány čtyři faktory, označené 1,2,3,4, použije se efekt jejich nejvyšších interakcí takto:

nejvyšší interakce	efekt	druhá mocnina efektu
123	-0,75	0,5625
124	0,50	0,2500
134	-0,25	0,0625
234	-0,75	0,5625
1234	-0,25	0,0625
	součet	1,5

Odtud rozptyl $s^2 = 1,5 / 5 = 0,3$.

V daném případě, protože se každý z pokusů opakoval dvakrát, je možné z dvojic výsledků Y_1, Y_2 vypočítat rozptyl v každém pokusu.

Rozptyl lze při dvou opakováních vypočítat snadněji (než s použitím definice rozptylu) pomocí vztahu

$$s_i^2 = \frac{1}{2} (Y_1 - Y_2)^2 \quad (4)$$

neboť

$$s^2 = \frac{1}{2-1} \left[\left(Y_1 - \frac{Y_1 + Y_2}{2} \right)^2 + \left(Y_2 - \frac{Y_1 + Y_2}{2} \right)^2 \right] = \frac{1}{2} (Y_1 - Y_2)^2$$

Po výpočtu efektu faktorů je potřeba provést vyhodnocení významnosti získaných hodnot, neboť např. není jasné. Je-li $ef(L)$ dostatečná hodnota, abychom L prohlásili za vlivný faktor.

Používají se různé metody hodnocení významnosti efektů, například

- a) testem
- b) graficky (normální resp. poloviční normální p-stní graf)
- c) pomocí ANOVA

Dále se budeme zabývat testem a grafem pro hodnocení významnosti efektů.

1.5 Test významnosti efektu

Výhodou opakování pokusů je možnost testovat významnost efektů nebo provádět analýzu reziduí.

Nevýhodou opakování je zvyšování počtu pokusů a tím nákladů.

Jiné možnosti umožňující totéž bez opakování a tedy při nižších nákladech jsou například

- a) zdvojení měření,
- b) zařazení centrálních bodů do experimentu,
- c) projekce plánu.

Zdvojení měření znamená, že při jednom pokusu se provádí více měření. Zdvojené měření *není totéž* co opakované měření.

O centrálních bodech a projekci plánu je pojednáno v dalších kapitolách.

V této kapitole uvedeme všechny body testu významnosti efektu.

Test významnosti efektů

1. Nulová hypotéza H_0 : efekt faktoru je bezvýznamný

Alternativní hypotéza H_1 : efekt faktoru je významný

2. Testovací kritérium

$$t = \frac{\text{efekt}}{s_e} \quad (5)$$

3. Kritická hodnota

$$t_{n_1+n_2+\dots+n_k-n}(\alpha) \quad (6)$$

kde $n_1, \dots, n_k$ jsou počty opakování pokusů, zde $n_i = 2$; n je počet pokusů bez opakování, zde $n = 8$.

4. Závěr testu: pro

$$|t| > t_{n_1+n_2+\dots+n_k-n}(\alpha)$$

se zamítá nulová hypotéza což znamená, že efekt (a tedy faktor) je významný.

V příkladě „Pružina“ je

$$t_{n_1+n_2+\dots+n_k-n}(\alpha) = t_{16-8}(0,05) = 2,306$$

Rozptyl

$$s^2 = \frac{8+2+2+8+2+8+2+8}{8} = 5$$

a

$$s_e^2 = \frac{4 \cdot s^2}{n} = \frac{4 \cdot 5}{16} = 1,25 ; \quad s_e = 1,12$$

Výpočty jsou uspořádány opět v tabulce

čís.	Y ₁	Y ₂	Y ₁ -Y ₂	s _i ²	efekt	T = efekt/s _e
1	77	81	-4	8		
2	98	96	2	2	L = 18	16,07
3	76	74	2	2	G = 1,5	1,34
4	90	94	-4	8	LG = -1,0	-0,89
5	63	65	-2	2	T = -8,0	-7,14
6	82	86	-4	8	LT = 0,5	0,45
7	72	74	-2	2	GT = 6,0	5,36
8	92	88	4	8	LGT = -0,5	-0,45

Tab.8 Testování významnosti efektu

Kritickou hodnotu převyšuje v absolutní hodnotě testovací kritérium faktorů L, T a interakce GT. Ostatní faktory tedy nemají vliv na životnost pružiny

1.6 Grafické hodnocení efektu faktorů

Pokud se neprovádí opakování jednotlivých pokusů, bývá používána grafická metoda určování významných faktorů, konkrétně *normální pravděpodobnostní graf*. V grafu se na vodorovnou osu vynáší efekt a na svislou osu relativní kumulativní četnost

$$P_i = \frac{100(i - 0,5)}{m}, \quad (7)$$

kde $i = 1, 2, \dots, m$, m je počet faktorů a interakcí.

Za významné se považují ty faktory, které se nacházejí výrazně mimo hlavní linii (znázorněné někdy přímkou, která představuje Gaussovu křivku na normálním pravděpodobnostním papíře). Při použití grafické metody je užitečné sestavit do pomocné tabulky tyto údaje:

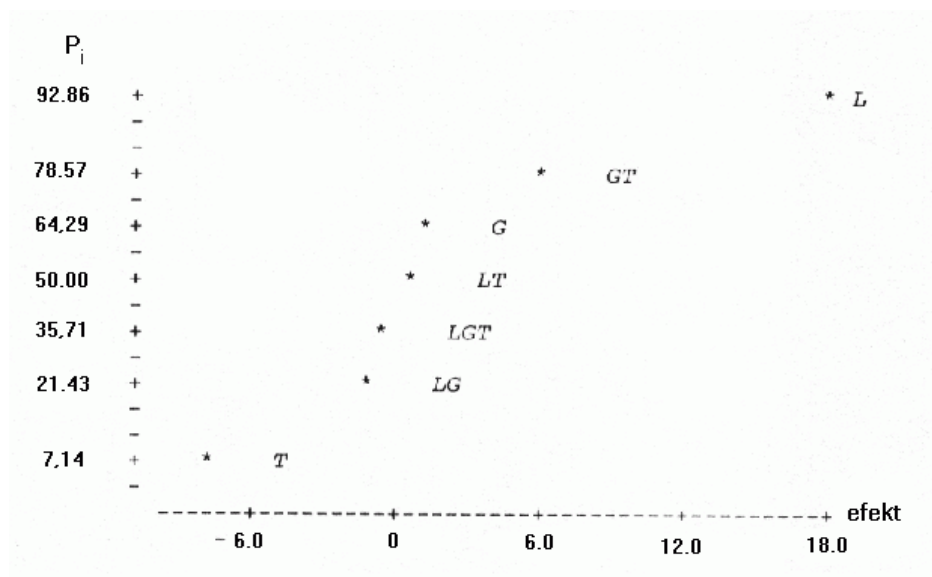
Číslo	1	2	3	4	5	6	7
Efekt	-8,0	-1,0	-0,5	0,5	1,5	6,0	18
Faktor	T	LG	LGT	LT	G	GT	L
P_i	7,14	21,42	35,71	50	64,29	78,57	92,86

Tab. 9 Grafické hodnocení významnosti efektu (Pružina)

Efekty ve druhém řádku jsou uspořádány vzestupně.

Graf (*normal probability plot*), sestrojený z údajů ve druhém řádku (vodorovná osa) a ve čtvrtém řádku (svislá osa) je na obr.1. Hodnoty P_i se vynášejí na normální pravděpodobnostní papír, takže máme normální pravděpodobnostní graf.

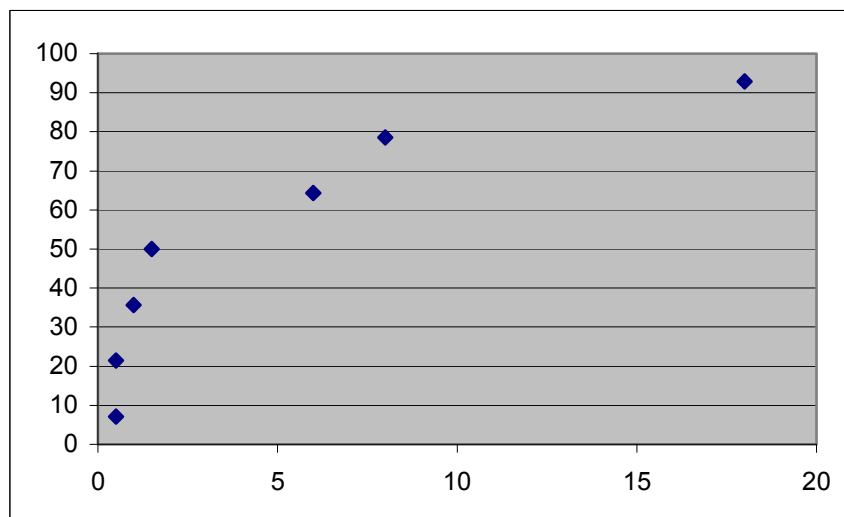
V grafu je vidět, že mimo hlavní linii jsou ty faktory, u kterých testovací kritérium překročilo kritickou hodnotu. Jsou to body L (nejvýrazněji), T a GT.



Obr.1 Grafické určení vlivných faktorů (pružina)

Alternativou k tomuto grafu je *poloviční normální pravděpodobnostní graf (half-normal probability plot)*, kde se k jeho sestrojení použijí absolutní hodnoty efektů:

Číslo	1	2	3	4	5	6	7
Efekt	0,5	0,5	1,0	1,5	6,0	8,0	18
Faktor	LT	LGT	LG	G	GT	T	L
P_i	7,14	21,42	35,71	50	64,29	78,57	92,86



Obr.2 Poloviční pravděpodobnostní graf
(*half-normal probability plot*)

1.7 Regresní model experimentu

Po nalezení vlivných faktorů lze sestavit model experimentu.

Dále budeme rozlišovat tyto modely (pro dva faktory):

a) lineární model

$$Y = b_0 + b_1x_1 + b_2x_2$$

Tento model je postačující pro modelování malých oblastí.

b) úplný kvadratický model

$$Y = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2$$

Je často dostačující i pro popis komplikovaných tvarů závislosti, jeho koeficienty se snadno naleznou, je propracovaná snadná klasifikace typů těchto modelů a existují vzorce pro výpočet optimální úrovně proměnných, které figurují v těchto modelech.

c) neúplný kvadratický model

$$Y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2$$

Výpočet regresních koeficientů u jednotlivých modelů:

U modelu a) a c) se koeficienty mohou počítat různými způsoby, např.

1) Pomocí efektů , kde $b_0 = \bar{Y}$, b_1 ; b_2 ,..., b_k je polovinou příslušného efektu, např. $0,5 \cdot \text{ef}(x_1) = b_1$.

2) Metodou nejmenších čtverců

Pro model b) nelze použít efektů ani MNČ.

Používají se např. tyto způsoby výpočtu:

- 1) Kombinovaný plán
- 2) Tříúrovňový plán
- 3) Box - Behnkenův plán

Neúplný kvadratický model v příkladě „Pružina“ bude

$$Y = 81,75 + 9.L - 4.T + 3.G.T$$

Přepočet na původní proměnné bude (vzorec 1)

$$Y = 81,75 + 9 \cdot \frac{L_o - 12,5}{2,5} - 4T + 3 \cdot \frac{G_o - 6}{1} \cdot T$$

Po úpravě vychází

$$Y = 36,75 + 3,6L_o - 22T + 3 G_o T$$

Model lze využít k výpočtu *teoretických hodnot*.

Například pro $L = G = T = -1$ bude

$$\hat{Y}_1 = 81,75 + 9 \cdot (-1) - 4 \cdot (-1) + 3 \cdot (-1) \cdot (-1) = 79,75$$

Poznámka:

U nekódovaných proměnných je problém při dosazování za kvalitativní proměnné, zde např. typ materiálu: nelze je dosadit.

Problém se řeší se zavedením speciálních proměnných, viz kap.9.

1.8 Reziduální odchylky: výpočet, vlastnosti a význam

Rozdíl empirické hodnoty Y (= výsledek pokusu) a teoretické hodnoty $\hat{Y}$ je *reziduální odchylka* (residual error): $e_i = Y_i - \hat{Y}_i$.

Y		Teoretické hodnoty	e (rezidua)	
1.pokus	2.pokus		1.rezidua	2.rezidua
77	81	79,75	$77 - 79,75 = -2,75$	$81 - 79,75 = 1,25$
98	96	97,75	0,25	-1,75
76	74	73,75	2,25	0,25
90	94	91,75	-1,75	2,25
63	65	65,75	-2,75	-0,75
82	86	83,75	-1,75	2,25
72	74	71,75	0,25	2,25
92	88	89,75	2,25	-1,75

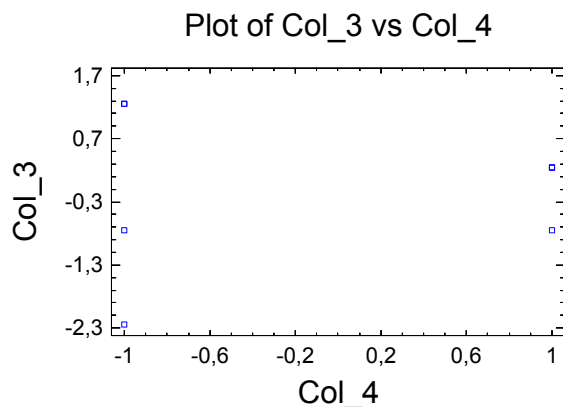
Rezidua by měla být u správného modelu

- a) nezávislá,
- a) normálně rozdělená (nenormalita reziduí znamená, že lineární model není vhodný),
- b) s konstantním rozptylem.

Např. pro faktor L v příkladě „Pružina“ máme tato rezidua:

L	G	T	LG	LT	GT	LGT	Res
-1	-	-	+	+	+	-	-0,75
+1	-	-	-	-	+	+	-0,75
-1	+	-	-	+	-	+	1,25
+1	+	-	+	-	-	-	0,25
-1	-	+	+	-	-	+	-2,25
+1	-	+	-	+	-	-	0,25
-1	+	+	-	-	+	-	1,25
+1	+	+	+	+	+	+	0,25

Pro faktor „L“ bude graf reziduí vers. úrovně faktoru:



Rezidua mají pro dolní úroveň faktoru L větší rozptyl než pro horní úroveň. Nekonstantní rozptyl znamená, že variabilita Y je ovlivňována úrovní faktorů a musí se vzít v úvahu. Takový graf se sestaví pro všechny faktory L, G a T.

1.9 Metoda nejmenších čtverců v DOE

Koeficienty v regresním modelu lze vypočítat také MNC pomocí známého vzorce $b = (X^T X)^{-1} X^T Y$. Jeho použití ukážeme na jednoduchém příkladě.

Příklad 2

Mějme plán 2^2 .

Plán experimentu pak bude:

Pokus	A	B	AB
1	- 1	- 1	1
2	1	- 1	- 1
3	- 1	1	- 1
4	1	1	1

Výpočet regresních koeficientů provedeme dle vztahu

$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, kde matice $\mathbf{X}$ a další potřebné matice jsou:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{4^4} \begin{bmatrix} 4^3 & 0 & 0 & 0 \\ 0 & 4^3 & 0 & 0 \\ 0 & 0 & 4^3 & 0 \\ 0 & 0 & 0 & 4^3 \end{bmatrix} = \frac{1}{4} \mathbf{I}$$

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \frac{1}{4} \mathbf{I} \mathbf{X}^T \mathbf{Y} = \frac{1}{4} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} Y_1 + Y_2 + Y_3 + Y_4 \\ -Y_1 + Y_2 - Y_3 + Y_4 \\ -Y_1 - Y_2 + Y_3 + Y_4 \\ Y_1 - Y_2 - Y_3 + Y_4 \end{bmatrix}$$

V prvním řádku dostáváme průměr z Y_i , v dalších řádcích součet výsledků (Y_i) se stejnými znaménky jako u znaménkové metody, ale čtvrtinu místo očekávané poloviny (obecně $n/2$). Je to proto, že zde počítáme koeficienty, kdežto u znaménkové metody efekty.

Například efekt prvního faktoru je $0,5(Y_1 - Y_2 + Y_3 - Y_4)$, zde ale máme koeficient = $0,25 \cdot (Y_1 - Y_2 + Y_3 - Y_4)$.

Jelikož koeficient = efekt / 2, bude

koeficient = $0,5 [0,5 (Y_1 - Y_2 + Y_3 - Y_4)] = 0,25 (Y_1 - Y_2 + Y_3 - Y_4)$.

Pozn.: U úplného kvadratického modelu jsou sloupce b_{11} a b_{22} lineárně závislé, proto by výpočet vůbec neproběhl.

K čemu slouží regresní modely experimentu:

- 1) k výpočtu $\bar{Y}$ pro libovolnou úroveň faktoru
- 2) k určení směru dynamického plánování experimentů
- 3) k výpočtu optimálních hodnot faktorů (u kvadratických modelů – lineární nemají extrém)
- 4) k výpočtu teoretických hodnot

5) k výpočtu reziduálních odchylek

1.10 Grafy interakcí

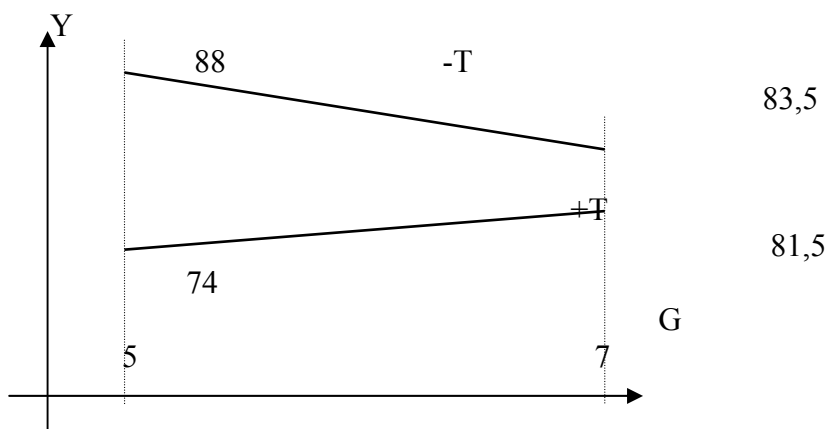
Pro významné interakce se sestavují grafy, umožňující diskusi o optimální úrovni jednotlivých faktorů *v závislosti na druhém faktoru*. Výsledná optimální úroveň může být diametrálně odlišná od té, která se stanovila individuálně.

Tak například pro významnou interakci GT sestavíme graf vlivu G na ukazatel kvality Y v závislosti na faktoru T: V tabulce 6 nalezneme

G	T			průměr
-	-	79	97	88
+	-	75	92	83,5

G	T			průměr
-	+	64	84	74
+	+	73	90	81,5

Tab.10 Určení krajních bodů grafu interakcí



Obr.3 Vliv G na Y v závislosti na T

Na vodorovné ose je první faktor z interakce GT, tedy G, na svislé vždy Y. Sestaví se krajní body grafu; mají souřadnice

-1	88
+1	83,5

-1	74
+1	81,5

V prvním případě bylo T záporné, ve druhém kladné. Proto jsou grafy označeny -T resp. +T. Z grafu je vidět, že pro maximalizaci Y je nejlepší T na dolní úrovni

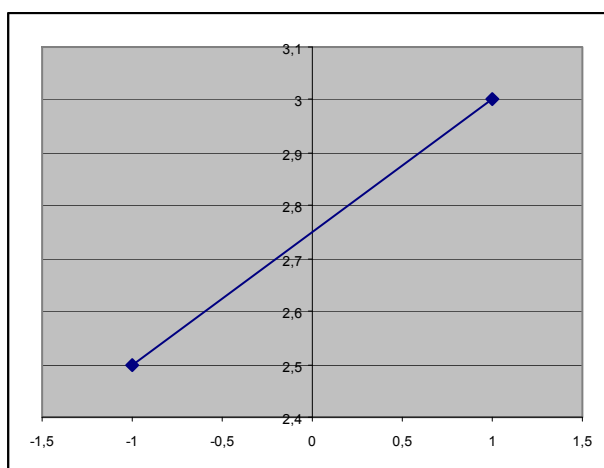
(-T) a že změna G neovlivňuje výrazně Y.

Příklad 3

Sestrojte graf efektu úrovní faktoru B pro plán

	A	B	Y
	-	-	3
	+	-	5
	-	+	2
	+	+	1
Průměr ++	$(5+1)/2 = 3$	$(2+1)/2 = 1.5$	
Průměr -Y	$(3+2)/2 = 2.5$	$(3+5)/2 = 4$	

Efekt A metodou „rozdíl průměrů“ $ef(A) = 3 - 2.5 = 0.5$. Efekt faktoru B je $ef(B) = 1.5 - 4 = -2.5$



Obr.4 Graf efektu úrovní faktoru A



Shrnutí kapitoly 1

Základní informace této kapitole jsou: jak sestavit plán experimentu a vypočítat vliv jednotlivých faktorů. Jelikož vypočítané hodnoty faktorů nelze hodnotit subjektivně, existuje exaktní metoda – test významnosti efektů. Je také možné použít grafickou metodu hodnocení.

Nalezené významné faktory jsou použity k sestavení modelu experimentu. Ten může být lineární nebo kvadratický. Při jejich sestavování jsou významné rozdíly a je proto vhodné nejprve rozhodnout, je-li kvadratický model potřeba. Kvadratický model je přesnější a umožní nalézt ideální nastavení významných parametrů. Dalším zlepšením modelování procesu je použití dynamického plánování.

Všechny tyto poznatky byly objasněny v první kapitole. Jejich zvládnutí si ověřte pomocí kontrolních otázek a úloh k procvičení.



Otázky ke kapitole 1

1. Do které výrobní etapy zejména patří DOE
2. Jaké jsou cíle DOE
3. Co obsahuje plán experimentu
4. Co je pokus a co experiment
5. Napište vzorec pro kódování proměnné
6. Definujte efekt faktoru
7. Vyjmenujte 5 způsobů výpočtu efektu faktoru
8. Jak se sestrojí graf efektu faktoru
9. K čemu slouží graf efektu faktoru
10. Jak se vypočítá počet pokusů v úplném plánu
11. Jak se vypočítá rozptyl odhadu efektu faktoru
12. Jak se odhadne rozptyl veličiny Y
13. Napište vzorec pro vážený rozptyl
14. Jak se počítá rozptyl odhadu efektu faktoru u neopakovaných pokusů
15. Napište jednoduchý vzorec pro rozptyl dvou veličin
16. Napište všechny body testu významnosti efektu
17. Napište tabulku potřebnou ke grafickému hodnocení efektů
18. Co je na osách při grafickém hodnocení efektů
19. Jak se poznají významné faktory při grafickém hodnocení
20. Jak se rozdělují regresní modely experimentu
21. Jak se počítají koeficienty v lineárním modelu
22. Jak se počítají koeficienty v úplném kvadratickém modelu
23. K čemu slouží regresní model experimentu
24. Jak se počítají reziduální odchylky
25. K čemu složí reziduální odchylky
26. Jaké vlastnosti musí mít reziduální odchylky
27. Ukažte použití MNČ v DOE



Řešené úlohy ke kapitole 1

Příklad 1

1. Napište do tabulky poloviční plán pro faktory A,B,C:

A	B	C	Y	efekt	faktor
			22		
			20		A
			17		B
			21		AB

- Výsledky pokusů v krychlových bodech jsou 22,20,17,21. Určit efekty A, B a C.
- Napsat lineární model pro tento plán.
- Najít teoretickou hodnotu Y pro A,B a C na dolní úrovni.
- Doplnit tento plán o hvězdicové body
- Doplnit do plánu tři centrální body a určit pure error S_{PE} , jsou li výsledky v centrálních bodech 26, 28 a 30.

Příklad 2

Sestavte úplný plán pro faktory E,F,G. Výsledky pokusů jsou 5,5,4,5,3,2,3,1.

Vypočítejte efekt faktorů a interakcí dvojic a trojice.

Sestavte tabulku, potřebnou pro konstrukci grafu pro hodnocení významnosti efektů.

Příklad 3

Napište úplný plán pro faktory A,B,C.

Vypočítejte $ef(A)$ a $ef(C)$, jsou li výsledky jednotlivých pokusů 2,6,7,7,4,5,6,3.

Napište tabulky potřebné pro graf interakcí. Proveďte náskres grafu.

Jaká je optimální úroveň, když Y se minimalizuje?

Příklad 4

Pro faktory T,C,K

- napište úplný plán
- vypočítejte efekty
- otestujte efekty
- vyhodnoťte efekty graficky
- napište neúplný kvadratický model
- sestrojte úsečkové grafy a vypočítejte efekt jako rozdíl průměrů
- vypočítejte rezidua

Výsledky pokusů při dvou opakováních jsou:

Y_1	Y_2
59	61
74	70
50	58
69	67
50	54
81	85
46	44
79	81

Příklad 5

Použijte Yatesův algoritmus k určení efektu dat v tabulce, získaných úplným faktorovým plánem 2^3 .

x_1	x_2	x_3	y
-	-	-	13
+	-	-	11
-	+	-	9
+	+	-	5
-	-	+	11
+	-	+	8
-	+	+	9
+	+	+	7

**KLÍČ K ŘEŠENÍ**

Řešení(1):

A	B	C	Y
-1	-1	1	22
1	-1	-1	20
-1	1	-1	17
1	1	1	21
-1,41	0	0	
1,41	0	0	
0	-1,41	0	
0	1,41	0	
0	0	-1,41	
0	0	1,41	
0	0	0	26
0	0	0	28
0	0	0	30

$$ef(A) = 1, ef(B) = -2, ef(C) = 3; \bar{Y} = 20$$

$$Y = 20 + 0,5A - B + 1,5C$$

$$\bar{Y}(-1, -1, -1) = 20 - 0,5 + 1 - 1,5 = 19$$

$$S_{PE} = 2^2 + 0 + 2^2 = 8$$

Řešení(2):

	E	F	G	Y	efekt	faktor
1	-	-	-	5		
2	+	-	-	5	-0,5	E
3	-	+	-	4	-0,5	F
4	+	+	-	5	0	EF
5	-	-	+	3	-2,5	G
6	+	-	+	2	-1	EG
7	-	+	+	3	0	FG
8	+	+	+	1	-0,5	EFG

i	1	2	3	4	5	6	7
Efekt	-2,5	-1	-0,5	-0,5	-0,5	0	0
Faktor	G	EG	E	F	EFG	EF	FG
Pi	7,14	21,43	35,71	50	64,29	78,57	92,86

$$P_i = \frac{100(i - 0,5)}{7}$$

Řešení(3):

A	B	C	Y
-	-	-	2
+	-	-	6
-	+	-	7
+	+	-	7
-	-	+	4
+	-	+	5
-	+	+	6
+	+	+	3
		$\bar{Y}_- = 5,5$	
		$\bar{Y}_+ = 4,5$	

Optimální úroveň pro C je horní úroveň.

Tabulky pro konstrukci grafu interakcí

A	C	Y	Průměrné Y
+	-	6;7	6,5
-	-	2;7	4,5

A	C	Y	Průměrné Y
+	+	5;3	4
-	+	4;6	5

Řešení(4):

Počet pokusů: $n = 2^3 = 8$

pokus	T	C	K	TC	TK	CK	TCK	Y ₁	Y ₂	$\bar{Y}$
1	-	-	-	+	+	+	-	59	61	60
2	+	-	-	-	-	+	+	74	70	72
3	-	+	-	-	+	-	+	50	58	54
4	+	+	-	+	-	-	-	69	67	68
5	-	-	+	+	-	-	+	50	54	52
6	+	-	+	-	+	-	-	81	85	83
7	-	+	+	-	-	+	-	46	44	45
8	+	+	+	+	+	+	+	79	81	80
$\bar{Y}_-$	52,75	66,75	63,5							
$\bar{Y}_+$	75,75	61,75	65							

$$\bar{\bar{Y}} = 64,25$$

Výpočet efektů pomocí znaménkové metody

$$ef(T) = \frac{1}{4}(-60 + 72 - 54 + 68 - 52 + 83 - 45 + 80) = 23$$

$$ef(C) = \frac{1}{4}(-60 - 72 + 54 + 68 - 52 - 83 + 45 + 80) = -5$$

$$ef(K) = \frac{1}{4}(-60 - 72 - 54 - 68 + 52 + 83 + 45 + 80) = 1,5$$

$$ef(TC) = \frac{1}{4}(60 - 72 - 54 + 68 + 52 - 83 - 45 + 80) = 1,5$$

$$ef(TK) = \frac{1}{4}(60 - 72 + 54 - 68 - 52 + 83 - 45 + 80) = 10$$

$$ef(CK) = \frac{1}{4}(60 + 72 - 54 - 68 - 52 - 83 + 45 + 80) = 0$$

$$ef(TCK) = \frac{1}{4}(-60 + 72 + 54 - 68 + 52 - 83 - 45 + 80) = 0,5$$

Test významnosti efektů

1) Stanovení hypotéz: H_0 : efekt faktoru je bezvýznamný

H_1 : efekt faktoru je významný

Pokusy	s_i^2
1	2
2	8
3	32
4	2
5	8
6	8
7	2
8	2

$$s_i^2 = \frac{1}{2}(Y_1 - Y_2)^2$$

Např.: $s_1^2 = \frac{1}{2}(Y_1 - Y_2)^2 = \frac{1}{2}(59 - 61)^2 = 2$

$$s^2 = \frac{v_1 \cdot s_1^2 + v_2 \cdot s_2^2 + v_3 \cdot s_3^2 + v_4 \cdot s_4^2 + v_5 \cdot s_5^2 + v_6 \cdot s_6^2 + v_7 \cdot s_7^2 + v_8 \cdot s_8^2}{v_1 + v_2 + v_3 + v_4 + v_5 + v_6 + v_7 + v_8} =$$

$$= \frac{1 \cdot 2 + 1 \cdot 8 + 1 \cdot 32 + 1 \cdot 2 + 1 \cdot 8 + 1 \cdot 8 + 1 \cdot 2 + 1 \cdot 2}{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1} = \frac{64}{8} = 8$$

$$s_e^2 = \frac{4 \cdot \sigma^2}{n} = \frac{4 \cdot 8}{16} = 2$$

2) Testovací kritérium: $t = \frac{efekt}{s_e}$

faktor	t
T	16,263
C	-3,536
K	1,061
TC	1,061
TK	7,071
CK	0
TCK	0,354

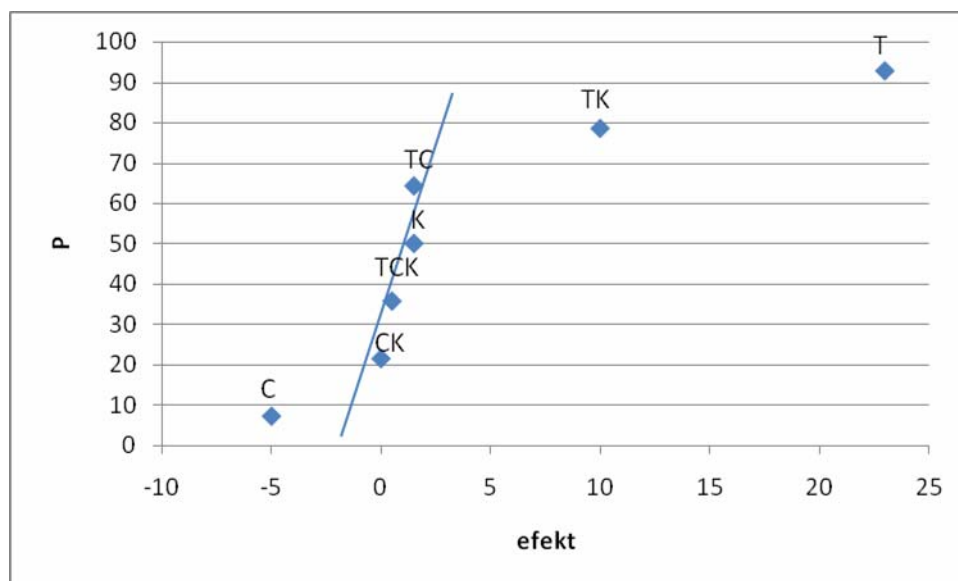
Významné faktory + jedna interakce

3) Kritická hodnota: $K = t_8(0,05) = 2,306$

4) Pokud je testovací kritérium v absolutní hodnotě větší, než kritická hodnota, zamítáme H_0 . Významné faktory jsou vyznačeny červeně. Jsou to T, C a TK.

Grafické hodnocení efektu

Číslo	1	2	3	4	5	6	7
Efekt	-5	0	0,5	1,5	1,5	10	23
Faktor	C	CK	TCK	K	TC	TK	T
P_i	7,143	21,43	35,71	50	64,286	78,57	92,857



Obr.1 Graf hodnocení efektu

Významné efekty leží mimo hlavní linii: C, TK, T.

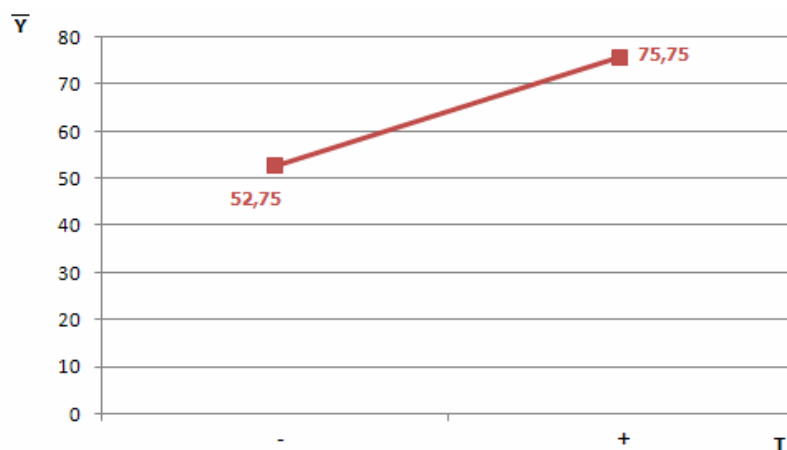
Neúplný kvadratický model

$$y = b_0 + b_T T + b_C C + b_{TK} TK$$

$$Y = 64,5 + 11,5T - 2,5C + 5TK$$

Úsečkový graf pro efekt T

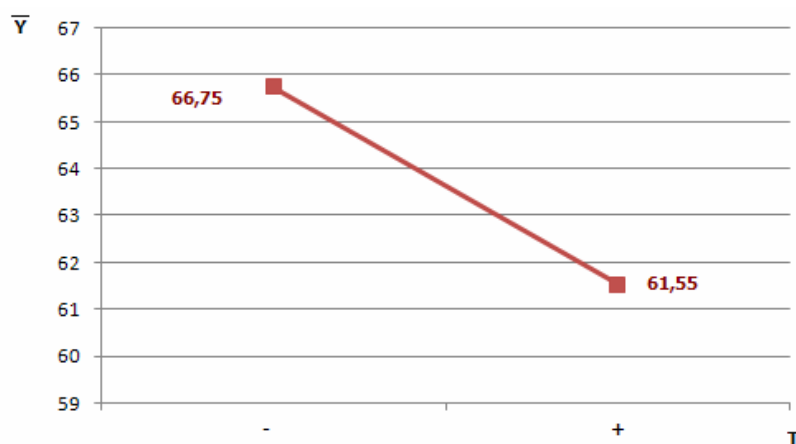
$$ef(T) = \bar{Y}_+ - \bar{Y}_- = 75,75 - 52,75 = 23$$



Obr. 2 Úsečkový graf pro efekt T

Úsečkový graf pro efekt C

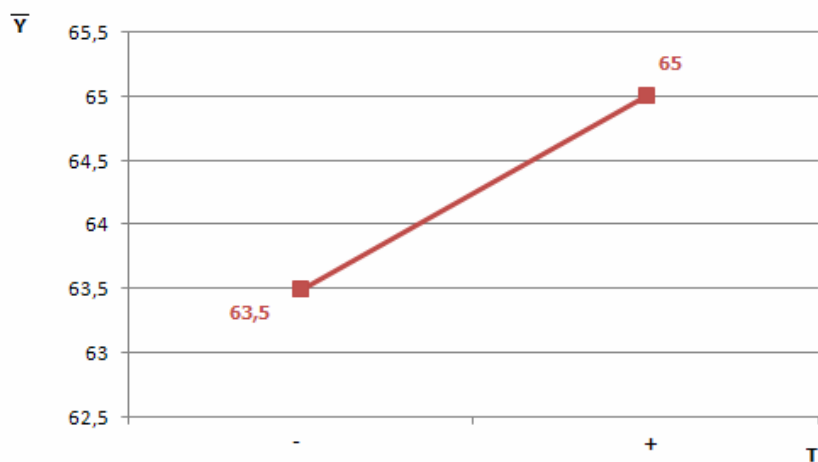
$$ef(C) = \bar{Y}_+ - \bar{Y}_- = 61,55 - 66,75 = -5$$



Obr. 3 Úsečkový graf pro efekt C

Úsečkový graf pro efekt K

$$ef(K) = \bar{Y}_+ - \bar{Y}_- = 65 - 63,5 = 1,5$$



Obr. 4 Úsečkový graf pro efekt K

Výpočet reziduí

$$\hat{Y}_1 = 64,25 - 1,5 + 2,5 + 5 = 60,25 \Rightarrow e_1 = \hat{Y}_1 - \bar{Y}_1 = 0,25$$

$$\hat{Y}_2 = 64,25 + 1,5 + 2,5 - 5 = 73,25 \Rightarrow e_2 = \hat{Y}_2 - \bar{Y}_2 = 1,25$$

$$\hat{Y}_3 = 64,25 - 1,5 - 2,5 + 5 = 55,25 \Rightarrow e_3 = \hat{Y}_3 - \bar{Y}_3 = 1,25$$

$$\hat{Y}_4 = 64,25 + 1,5 - 2,5 - 5 = 68,25 \Rightarrow e_4 = \hat{Y}_4 - \bar{Y}_4 = 0,25$$

$$\hat{Y}_5 = 64,25 - 1,5 + 2,5 - 5 = 50,25 \Rightarrow e_5 = \hat{Y}_5 - \bar{Y}_5 = -1,75$$

$$\hat{Y}_6 = 64,25 + 1,5 + 2,5 + 5 = 83,25 \Rightarrow e_6 = \hat{Y}_6 - \bar{Y}_6 = 0,25$$

$$\hat{Y}_7 = 64,25 - 1,5 - 2,5 - 5 = 45,25 \Rightarrow e_7 = \hat{Y}_7 - \bar{Y}_7 = 0,25$$

$$\hat{Y}_8 = 64,25 + 1,5 - 2,5 + 5 = 78,25 \Rightarrow e_8 = \hat{Y}_8 - \bar{Y}_8 = -1,75$$

T	C	TK	$\hat{Y}$	$\bar{Y}$	e_i
-	-	+	60,5	60	0,25
+	-	-	73,25	72	1,25
-	+	+	55,25	54	1,25
+	+	-	68,25	68	0,25
-	-	-	50,25	52	-1,75
+	-	+	83,25	83	0,25
-	+	-	45,25	45	0,25
+	+	+	78,25	80	-1,75

Řešení(5):

Yatesův algoritmus:

y	(1)	(2)	(3)	dělitel	efekt	faktor	x_1	x_2	x_3
13	24	38	73	8	9,125	průměr	-	-	-
11	14	35	-11	4	-2,75	x_1	+	-	-
9	19	-6	-13	4	-3,25	x_2	-	+	-
5	16	-5	-1	4	-0,25	$x_{1,2}$	+	+	-
11	-2	-10	-3	4	-0,75	x_3	-	-	+
8	-4	-3	1	4	0,25	$x_{1,3}$	+	-	+
9	-3	-2	7	4	1,75	$x_{2,3}$	-	+	+
7	-2	1	3	4	0,75	$x_{1,2,3}$	+	+	+

2. ČÁSTEČNÝ FAKTOROVÝ EXPERIMENT SE DVĚMA ÚROVNĚMI

(Fractional Factorial Design at Two Levels)



Čas ke studiu: 10 hodin



Cílem druhé kapitole je objasnit zejména:

- a) jak lze snížit počet pokusů v experimentu
- b) jaké problémy přináší snížení počtu pokusů a jak se řeší



Výklad

V úplném faktorovém experimentu se sestavuje plán experimentu pro každý faktor. U částečného faktorového experimentu se plán sestaví jen pro několik faktorů. Ty budeme nazývat *hlavní faktory*. Ostatní faktory, které nazveme *vedlejší faktory*, se vyjádří pomocí hlavních. Tím se dosáhne snížení počtu pokusů. Označení hlavní a vedlejší nijak nesouvisí s velikostí jejich vlivu na sledovaný ukazatel Y . Ten ani nelze na počátku experimentu vymezit, neboť se vychází z toho, že o vlivu faktorů se nic neví. Pokud jde o způsob vyjádření vedlejších faktorů pomocí hlavních, je významné, jak se provede. Tomuto problému se budeme věnovat později.

Je-li 2^k označení pro úplný experiment,

2 = počet úrovní faktorů,

k = počet faktorů,

pak 2^{k-p} je označení pro částečný faktorový experiment,

p = stupeň snížení.

Chceme-li například v plánu 2^7 , který představuje $n = 128$ pokusů, snížit počet pokusů na polovinu, tj.

$$\frac{2^7}{2} = 2^{7-1},$$

dostáváme částečný faktorový experiment, který představuje

$$n = 2^{7-1} = 64$$

pokusů, tedy p o l o v i n u. Je to nejmenší možné snížení počtu pokusů. Plány se snížením počtu pokusů na polovinu se nazývají *poloviční plány*.

Stupeň snížení p může být i vyšší než 1, například

$$2^{7-4},$$

kde bude jen $n = 8$ pokusů. To je pro $k = 7$ faktorů nejvyšší možné snížení. U tzv. *lineárních plánů* (plánů prvního řádu), sloužících k nalezení modelů, které obsahují jen faktory bez jejich kombinací nebo mocnin, vychází největší možné snížení z pravidla, podle kterého *počet pokusů nesmí být menší než počet parametrů v regresním modelu*, což je zde totožné s *počtem faktorů*.

Musí tedy platit

$$n \geq k.$$

Pro *kvadratické plány*, které jsou určeny k nalezení kvadratických modelů, je podmínek více, jak uvidíme ve čtvrté kapitole.

Plány s největším snížením jsou *saturované (nasycené) plány*. V uvedeném případě $k = 7$ a $n = 2^{7-4} = 8$, takže $p = 4$ je nejvyšší možný stupeň snížení.

Například pro $k = 15$ faktorů je nejvyšší možné snížení 2^{15-11} , neboť počet pokusů je při $p = 11$ roven $n = 16$ a počet faktorů je $k = 15$. Pokud bychom provedli ještě vyšší snížení, např. 2^{15-12} , potom $k = 15$, ale $n = 8$, takže $n < k$. Mezi polovičními a nasycenými plány může být ještě řada možností snížení. Plány se stupněm snížení mezi maximálním a minimálním se nazývají *středové plány*.

Například mezi 2^{7-1} a 2^{7-4} jsou středové plány 2^{7-2} a 2^{7-3} .

Částečné faktorové plány lze tedy rozdělit do tří skupin:

- a) plány s nejnižším snížením (poloviční)
- b) plány s nejvyšším snížením (saturované)
- c) středové plány

2.1 Operace s faktory a jejich vlastnosti

Označme I faktor, obsahující jen úroveň „+“. Takový faktor se nazývá *jednotkový*.

Pro operace s faktory platí tyto zřejmé vztahy

$$A.A = I \tag{8}$$

$$A.I = I.A = A \tag{9}$$

$$(A.B).C = A(B.C) \tag{10}$$

$$A.B = B.A \tag{11}$$

Při vyhodnocování částečných plánů se používá několik důležitých pojmů, které objasníme v následujícím obecném příkladě.

Příklad 1

A,B,C,D,E jsou faktory, pro které se má sestavit poloviční plán.

Postupujeme tak, že určíme 4 hlavní faktory (například A,B,C,D), pro které se sestaví úplný plán a zbývající (vedlejší) faktor E se vyjádří jako jejich kombinace, například

$$E = ABCD \quad (12)$$

Každá kombinace faktorů tvoří *slovo*. Slovo se skládá z písmen (faktorů). Počet písmen ve slově je *délka slova*. Vztah (12) se nazývá *generátor*. V plánu 2^{k-p} je p vedlejších faktorů, z nichž každý musí být vytvořen (generován) pomocí hlavních, proto je také p generátorů. Vynásobením generátoru (12) levou stranou, tj. faktorem E,

$$E.E = E. ABCD$$

a s použitím vztahů (8) – (11) dostáváme

$$I = ABCDE. \quad (13)$$

Slova, která jsou rovna jednotkovému faktoru I, se nazývají *definiční rovnice*, například (13). Jak uvidíme dále, definičních rovnic může být i více než jedna. Délka *nejkratšího* slova v definičních rovnicích je tzv. *řešení plánu* (resolution of the design) a zapisuje se k typu plánu římskou číslicí jako index. Zde např. 2^{5-1}_V . Z tohoto zápisu čteme: faktory jsou na dvou úrovních, počet faktorů je 5, vedlejší faktor je jeden, jedná se o poloviční plán a stupeň řešení je V. Řešení plánu je V proto, že slovo v definiční rovnici (13) má délku (počet písmen) 5.

2.2 Nalezení zaměnitelných dvojic

Pomocí definiční rovnice lze najít dvojice faktorů (resp.interakcí), které tvoří stejné posloupnosti znamének a které se nazývají *zaměnitelné dvojice*.

Například zaměnitelnou interakci k DE nalezneme vynásobením definiční rovnice touto interakcí

$$I = ABCDE / DE$$

$$DE.I = DE.ABCDE$$

Podle (8) a (9) dostáváme

$$DE = ABC$$

Praktické použití ilustruje následující příklad.

Příklad 2 ([3], upraveno)

Sleduje se množství barviva Y, které zůstane na látce po absolvování testů (ve srovnání se standardním vzorkem), v závislosti na faktorech:

A = pH, B = teplota, C = koncentrace, D = dokončovací teplota, E = dokončovací čas.

Faktory a jejich dvě úrovně jsou v tabulce 1.

Faktor	Symbol	-	+
PH	A	4.5	5.5
teplota	B	70° C	80° C
koncentr.	C	1g/l	3g/l
dok.teplota	D	170° C	190° C
dok.čas	E	50s.	70s.

Tab.1 Faktory a jejich úrovně

Má se sestrojit úplný plán a nalézt nejvýznamnější faktory, sestrojit poloviční plán, nalézt nejvýznamnější faktory a porovnat výsledek polovičního plánu s úplným plánem.

Řešení:

Úplný faktorový plán 2^5

Čís.	A	B	C	D	E	Y	DE	ABC	ACDE
1	-	-	-	-	-	13.1	+	-	+
2	+	-	-	-	-	9.9	+	+	-
3	-	+	-	-	-	8.1	+	+	+
4	+	+	-	-	-	7.5	+	-	-
5	-	-	+	-	-	9.0	+	+	-
6	+	-	+	-	-	9.2	+	-	+
7	-	+	+	-	-	-1.0	+	-	-
8	+	+	+	-	-	-1.0	+	+	+
9	-	-	-	+	-	10.6	-	-	-
10	+	-	-	+	-	8.2	-	+	+
11	-	+	-	+	-	11.0	-	+	-
12	+	+	-	+	-	11.2	-	-	+
13	-	-	+	+	-	5.1	-	+	+
14	+	-	+	+	-	9.7	-	-	-
15	-	+	+	+	-	4.1	-	-	+
16	+	+	+	+	-	2.9	-	+	-
17	-	-	-	-	+	6.4	-	+	-
18	+	-	-	-	+	9.8	-	+	+
19	-	+	-	-	+	9.0	-	+	-
20	+	+	-	-	+	6.6	-	-	+
21	-	-	+	-	+	4.9	-	+	+
22	+	-	+	-	+	5.3	-	-	-
23	-	+	+	-	+	-5.1	-	-	+
24	+	+	+	-	+	-3.7	-	+	-
25	-	-	-	+	+	17.3	+	-	+
26	+	-	-	+	+	12.7	+	+	-
27	-	+	-	+	+	12.9	+	+	+
28	+	+	-	+	+	13.7	+	-	-
29	-	-	+	+	+	12.4	+	+	-
30	+	-	+	+	+	12.4	+	-	+
31	-	+	+	+	+	3.8	+	-	-
32	+	+	+	+	+	4.0	+	+	+

Efekty faktorů v úplném plánu, vypočítané znaménkovou metodou, jsou:

Faktor	A	B	C	D	E
Efekt	-0.2	-4.5	-6.0	4.0	0.3

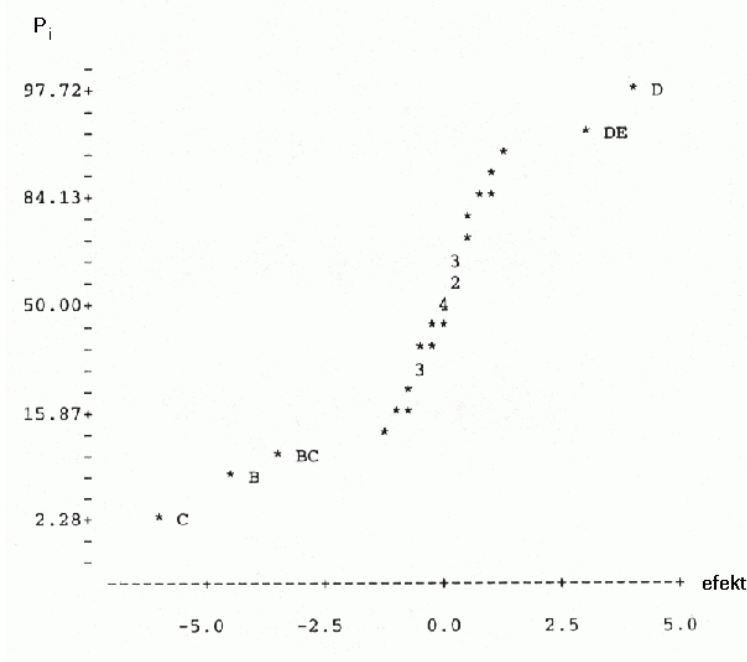
Fakt.	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
Ef.	0.0	0.9	-0.1	0.1	-3.5	1.4	-0.5	0.6	-0.8	3.0

Fakt.	ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE
Ef.	-0.6	0.3	0.1	0.3	-0.3	-0.7	0.4	-0.5	-1.5	0.2

Fakt.	ABCD	ABCE	ABDE	ACDE	BCDE	ABCDE
Ef.	-1.1	0.8	1.0	0.1	0.2	-0.4

Určíme významné faktory.

Protože se jednotlivé pokusy neopakují, nelze použít testu (bez opakování není možné vypočítat směrodatnou odchylku, která je součástí testovacího kritéria). V takových případech se používá grafická metoda. Body, které leží mimo hlavní linii, jsou zde C(efekt = -6.0, $P_i = 1.67$), B(-4.5, 5), BC(-3.5, 8.33), DE(3.0, 95), D(4.0, 98.33), (obr.1).



Obr.1 Grafické určení vlivných faktorů v úplném plánu

Nyní sestavíme poloviční plán 2^{5-1} .

- a) Jako hlavní faktory jsou stanoveny A,B,C,D, vedlejší je E. Ten bude vyjádřen pomocí hlavních. Generátor volíme $E = ABCD$.

Pokus	A	B	C	D	E = ABCD	úplný plán	Y
1	-	-	-	-	+	17	6,4
2	+	-	-	-	-	2	9,9
3	-	+	-	-	-	3	8,1
4	+	+	-	-	+	20	6,6
5		-	+	-	-	5	9,0
6	+	-	+	-	+	22	5,3
7	-	+	+	-	+	23	-5,1
8	+	+	+	-	-	8	-1,0
9	-	-	-	+	-	9	10,6
10	+	-	-	+	+	26	12,7
11	-	+	-	+	+	27	12,9
12	+	+	-	+	-	12	11,2
13	-	-	+	+	+	29	12,4
14	+	-	+	+	-	14	9,7
15	-	+	+	+	-	15	4,1
16	+	+	+	+	+	32	4,0

Tab.2 Poloviční plán (základní frakce)

Poloviční plán tedy má 16 pokusů, ale nikoli prvních 16 z úplného plánu. Které pokusy úplného plánu jsou zde obsaženy, je patrné z předposledního sloupce („úplný plán“), kde jsou čísla řádků plánu 2⁵.

Druhou polovinu úplného plánu, kterou nazýváme *doplňkovou frakci*, tj. zbývajících 16 pokusů nalezneme, zvolíme-li generátor $E = -ABCD$:

č.	A	B	C	D	-ABCD =E	úplný plán	Y
1	-	-	-	-	-	1	13,1
2	+	-	-	-	+	18	9,8
3	-	+	-	-	+	19	9,0
4	+	+	-	-	-	4	7,5
5	-	-	+	-	+	21	4,9
6	+	-	+	-	-	6	9,2
7	-	+	+	-	-	7	-1,0
8	+	+	+	-	+	24	-3,7
9	-	-	-	+	+	25	17,3
10	+	-	-	+	-	10	8,2
11	-	+	-	+	-	11	11,0
12	+	+	-	+	+	28	13,7
13	-	-	+	+	-	13	5,1
14	+	-	+	+	+	30	12,4
15	-	+	+	+	+	31	3,8
16	+	+	+	+	-	16	2,9

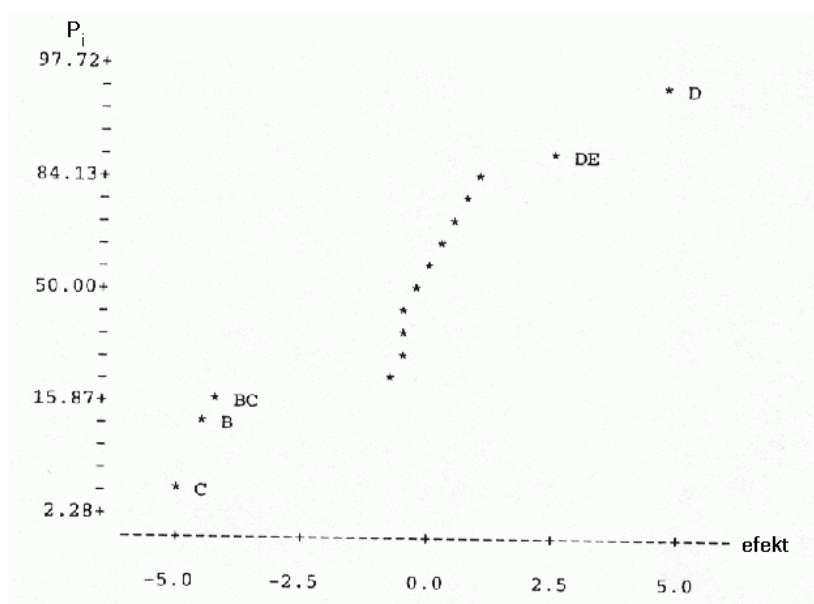
Tab.3 Poloviční plán (doplňková frakce)

Výpočet efektu faktorů pro poloviční plán, se počítá opět znaménkovou metodou. Výsledky jsou v tabulce 4: první dva sloupce pro první polovinu, druhé dva sloupce pro druhou polovinu plánu.

Faktor	Efekt v 1.frakci	Faktor	Efekt ve 2.frakci
A+BCDE	0,0	A-BCDE	-0,4
B+ACDE	-4,4	B-ACDE	-4,6
C+ABDE	-5,0	C-ABDE	-7,0
D+ABDE	4,8	D-ABDE	3,2
E+ABCD	-0,8	E-ABCD	1,4
AB+CDE	0,2	AB-CDE	-0,2
AC+BDE	-0,6	AC-BDE	2,4
AD+BCE	-0,6	ADBCE	0,4
AE+BCD	0,5	AE-BCD	-0,3
BC+ADE	-4,2	BC-ADE	-2,8
BD+ACD	1,1	BD-ACD	1,7
BE+ACD	-0,2	BE-ACD	-0,8
CD+ABE	0,7	CD-ABE	0,5
CE+ABD	-0,5	CE-ABD	-1,1
DE+ABC	2,4	DE-ABC	3,6

Tab.4 Efekt faktorů v obou polovičních plánech

Vypočítaným efektům však neodpovídá faktor, ale dvojice faktorů. V dalším budeme řešit, proč dvojice a které faktory tvoří dvojici. Nejprve však provedeme grafické vyhodnocení významnosti efektů. Je na obrázku 2.



Obr.2 Grafické určení vlivných faktorů v polovičním plánu

Z porovnání grafů na obr.1 a obr.2 je zřejmé, že poloviční plán dává stejné hodnocení jako úplný plán. Znamená to tedy, že snížením počtu pokusů na polovinu nedošlo ke ztrátě informací a tím ke změně výsledků.

Porovnáme-li však efekty faktorů a jejich interakcí v úplném a v polovičním plánu je vidět, že nejsou *přesně* stejné.

Navíc, jak bylo uvedeno, jsou v tabulce 14 uváděny efekty pro *s o u č t y*, například $A + BCDE$, $B + ACDE$ apod.

Součty jsou ve sloupci „faktor“ v tabulce 14 uvedeny proto, že snížením počtu pokusů dochází k tomu, že některé posloupnosti znamének u interakcí faktorů jsou stejné. Například (podle tab.12)

ABC	-	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+
DE	-	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+

Jsou-li u interakcí ABC a DE stejné posloupnosti znamének je jasné, že také efekt (zde počítaný znaménkovou metodou) je stejný. Vypočítaný efekt patří oběma interakcím, proto jsou v součtu. Neznamená to ale, že na každou připadá polovina!

Kdybychom, pouze teoreticky, našli oba poloviční plány, můžeme vypočítat, kolik z celkového efektu připadá na jednotlivé interakce v součtu (tento výpočet se v praxi samozřejmě neprovádí). V tabulce 14 je například efekt

$$B + ACDE = - 4,4$$

a

$$B - ACDE = - 4,6.$$

Řešením těchto dvou rovnic dostáváme efekt připadající na B: -4,5 a na interakci ACDE: 0,1. Z těchto výsledků je vidět dva důležité poznatky:

- „čistý“ efekt B je přesně shodný s efektem B v úplném plánu,
- na interakci připadá zanedbatelně malý podíl celkové hodnoty efektu.

Obecně platí, že čím delší je slovo, tím menší má efekt. Za bezvýznamný je považován efekt (interakce) o délce slova alespoň 3.

Vysvětluje se také, proč efekt $B + ACDE = - 4,4$ není shodný s efektem B v úplném plánu (B: - 4,5). Je to proto, že efekt B je „kontaminován“ efektem interakce ACDE.

Bylo již řečeno, že při sestavování polovičních plánů se jeden z faktorů vyjádří jako interakce ostatních. Říkáme-li „ostatních“ neznamená to, že nutně všech ostatních. Tak například při pěti faktorech A,B,C,D,E je možné vyjádřit E mnoha způsoby.

Porovnejme tyto dva:

- $E = AB$,
- $E = ABCD$

Příslušné definiční rovnice jsou

- $I = ABE$
- $I = ABCDE$

V případě a) máme plán 2_{III}^{5-1} , v případě b) plán 2_V^{5-1} . Tento druhý plán je lepší, neboť má řešení plánu V což znamená, při hledání zaměnitelných dvojic budou v součtu krátká a dlouhá slova; například k A, máme

- a) A + BE
- b) A + BCDE

V případě b) tvoří zaměnitelnou dvojici s A interakce více faktorů, která má proto menší podíl na celkovém efektu, u více než dvou faktorů dokonce tak malý, že se zanedbává a pracuje se jen s krátkým slovem, zde s faktorem A. Tím se usnadňuje diskuse k vypočítanému efektu.

Optimální volby generátorů pro různé počty faktorů (k) a různé stupně řešení jsou uváděny v tabulkách. Jednu z nejkratších navrhl L.W.Robinson (tab.5).

V tabulce jsou 4 šipky: např. pro k = 3 a st.řešení IV: má li být řeš. IV, musí se udělat úplný plán. Tučně jsou uvedeny počty pokusů pro dané k a stupeň řešení. Pro k = 12 až 15 nejsou uvedeny generátory.

	<i>Stupeň řešení</i>			
k	III		IV	V
3	3 = 1.2	4	→	Full 8
4	→		4 = 1.2.3 8	Full 16
5	4 = 1.2	8	→	5 = 1.2.3.4 16
	5 = 1.3			
6	6 = 2.3		5 = 1.2.3 16	6 = 1.2.3.4.5 32
			6 = 1.2.4	
7	7 = 1.2.3		7 = 1.3.4	7 = 1.2.3.4.5.6 64
8	→		8 = 2.3.4	7 = 1.2.3.4 64
				8 = 1.2.5.6
9	5 = 1.2.3	16	6 = 1.2.3 32	8 = 1.2.3.4.5 128
	6 = 1.2.4		7 = 1.2.4	9 = 1.2.3.6.7
	7 = 1.3.4		8 = 1.2.5	
	8 = 2.3.4		9 = 1.3.4.5	
	9 = 1.2.3.4			
10	10 = 1.2		10 = 2.3.4.5	10 = 1.4.5.6
11	11 = 1.3		6 = 1.2.3 32	11 = 2.4.6.7
			7 = 1.2.4	
			8 = 1.2.5	
			9 = 1.3.4	
			10 = 1.3.5	
			11 = 1.4.5	
12	12 = 1.4		12 = 2.3.4	256
13	13 = 2.3		13 = 2.3.5	
14	14 = 2.4		14 = 2.4.5	
15	15 = 3.4		15 = 3.4.5	

Tab.5 Optimální volby generátorů
(L.W.Robinson, QE 15, No3, 2003, s .403-406)

Poznámka

Je vždy lepší udělat poloviční plán a pak, na základě analýzy, ho případně doplnit (o další frakci), než dělat ihned úplný plán.

Stejně zásady jako u polovičních plánů platí pro středové a saturované plány. Saturované plány jsou ty, kde

$$\text{počet parametrů } (p) = \text{počet pokusů } (n)$$

Saturovaný plán lze definovat pomocí p (počet parametrů modelu) nebo pomocí k (počet faktorů). U lineárních modelů je saturovaný plán ten, kde $n = p$ neboli $n = k + 1$, protože $p = k + 1$.

2.3 Stanovení počtu zaměnitelných faktorů

V plánu 2^{k-p} je p vedlejších faktorů. Proto je potřeba p generátorů a z nich lze vytvořit

$$\binom{p}{1} + \binom{p}{2} + \dots + \binom{p}{p} \quad (14)$$

definičních rovnic (dané definiční rovnice + součin dvojic, kterých je p nad dvěma; součin trojic, kterých je p nad třemi atd.). Výraz (14) lze zapsat ve tvaru

$$\sum_{k=1}^p \binom{p}{k} \quad \text{resp.} \quad \sum_{k=0}^p \binom{p}{k} - 1 = \quad \text{resp.} \quad \left(\sum_{k=0}^p \binom{p}{k} 1^{p-k} \cdot 1^k \right) - 1 = (1+1)^p - 1 = 2^p - 1$$

Počet definičních rovnic v plánu 2^{k-p} je $2^p - 1$.

Z každé definiční rovnice lze vytvořit k danému faktoru zaměnitelný faktor, takže v plánu 2^{k-p} je počet zaměnitelných faktorů roven počtu definičních rovnic + daný faktor (plus jedna)

$$(2^p - 1) + 1 = 2^p$$

Počet zaměnitelných faktorů v plánu 2^{k-p} je 2^p .

Příklad 3 (nalezení definičních rovnic)

Uvažujme situaci, kdy je sledováno 7 faktorů A,B,C,D,E,F,G, ovlivňujících ukazatel kvality Y. Jestliže se sestavuje saturovaný plán 2^{7-4} , jsou vybrány hlavní faktory A,B,C, pro které bude sestaven úplný plán. Zbývající 4 faktory (D,E,F,G) jsou vedlejší. Generátory jsou voleny například takto (jejich počet je p tj. 4):

$$D = AB, E = AC, F = BC, G = ABC.$$

Definiční rovnice budou (jejich počet je $2^p - 1 = 2^4 - 1 = 15$):

$$ABD = ACE = BCF = ABCG$$

$$(\text{součin dvou}) = BCDE = ACDF = CDG = ABEF = BEG = AFG$$

$$(\text{součin tří}) = DEF = ADEG = BDFG = CEFG$$

$$(\text{součin čtyř}) = ABCDEFG$$

Protože místo $2^7 = 128$ pokusů jich bude jen $2^{7-4} = 8$, je každý plán frakcí, která je 2^{-4} celého plánu, takže existuje 16 frakcí a v každé je 16 zaměnitelných faktorů.

Například pro faktor A máme tuto skupinu zaměnitelných faktorů

$$A + BD + CE + ABCF + BCG + ABCDE + CDF + ACDG + BEF + ABEG + FG + ADEF + DEG + ABDFG + ACEFG + BDCEFG$$

(jejich počet je $2^p = 2^4 = 16$)

Z nich má smysl ponechat jen interakce dvou, takže pro A to bude

$$A + BD + CE + FG$$

2.4 Volba generátorů u středových plánů

Středové plány by měly spojovat dobré vlastnosti polovičních a nasycených plánů. Mezi středovými plány se hledá takový plán, který má méně pokusů než poloviční a pokud možno nejvýhodnější skupiny zaměnitelných faktorů. Připomeňme, že za nejlepší zaměnitelnou interakci se považuje ta, která je tvořena největším počtem faktorů. Skladba zaměnitelných faktorů souvisí s volbou generátorů.

Příklad 4

Porovnat výsledek různé volby generátorů u středového plánu 2^{7-3} s faktory A,B,C,D,E,F,G.

a) Generátory: E = ABCD, F = ABC, G = BCD

Definiční rovnice budou

$$I = ABCDE = ABCF = BCDG = DEF = AEG = ADFG = BCEFG$$

Stupeň řešení plánu je zde III: 2^{7-3}_{III}

b) Generátory: E = ABC, F = BCD, G = ACD,

Definiční rovnice I = ABCE = BCDF = ACDG = ADEF = BDEG = CEFG = ABFG, kde stupeň řešení je IV: 2^{7-3}_{IV} .

2.5 Projekce plánu

Má-li některý z faktorů nevýznamný efekt, je možné ho vynechat. Získá se tak plán s opakováním. Např. pro tři faktory A,B,C má úplný plán 2^3 tvar

	A	B	C
1	-	-	-
2	+	-	-
3	-	+	-
4	+	+	-
5	-	-	+
6	+	-	+
7	-	+	+
8	+	+	+

Vynecháním faktoru C vznikl plán s opakováním.

Oobecně platí, že máme-li úplný plán 2^k bez opakování a jestliže v něm vynecháme h ($h < k$) faktorů, vznikne úplný plán pro $k - h$ faktorů s počtem opakování 2^h .

Vynechá-li se v polovičním plánu 1 faktor, vznikne úplný plán.

Příklad 5

Mějme plán 2^{3-1} s faktory A, B a $C = AB$.

A	B	$C = AB$
-	-	+
+	-	-
-	+	-
+	+	+

Vynecháním C vznikne úplný plán.

Podobně vynecháním B vznikne úplný plán

A	B	$C = AB$
-	-	+
+	-	-
-	+	-
+	+	+

Vynecháme-li např. v polovičním plánu 2^{5-1} faktory A a E, vznikne částečný plán pro B, C a D se dvěma opakováními:

č.	A	B	C	D	$-ABCD = E$	úplný plán	Y
1	-	-	-	-	-	1	13,1
2	+	-	-	-	+	18	9,8
3	-	+	-	-	+	19	9,0
4	+	+	-	-	-	4	7,5
5	-	-	+	-	+	21	4,9
6	+	-	+	-	-	6	9,2
7	-	+	+	-	-	7	-1,0
8	+	+	+	-	+	24	-3,7
9	-	-	-	+	+	25	17,3
10	+	-	-	+	-	10	8,2
11	-	+	-	+	-	11	11,0
12	+	+	-	+	+	28	13,7
13	-	-	+	+	-	13	5,1
14	+	-	+	+	+	30	12,4
15	-	+	+	+	+	31	3,8
16	+	+	+	+	-	16	2,9

Obecně vynechá-li se d faktorů v plánu 2^{k-p} bez opakování, vznikne buď úplný plán nebo částečný plán s opakováním. Vzniklý plán má tytéž definiční rovnice, které ale neobsahují vynechané faktory.

Při analýze částečných plánů vznikají komplikace v důsledku součtu faktorů, kterým náleží vypočítaný efekt. Při navrhování částečných plánů je proto snaha, aby v součtu

- a) byla krátká a dlouhá slova
- b) bylo co nejméně slov.

První cíl lze řešit správnou volbou generátorů; ta způsobí v konečném důsledku vyšší stupeň řešení; jako příklad jsme uvedli porovnání generátorů $E = AB$ resp. $E = ABCD$, které vedly k plánům 2_{III}^{5-1} resp. 2_V^{5-1} , tedy stupňům řešení III resp. V.

2.6 Doplnkové frakce: tvorba a vlastnosti

Jednou z možností jak dosáhnout cíle b) je doplnění základní frakce plánu o další, tzv. doplnkové frakce, tedy slučování frakcí (sequence).

Doplnková frakce u polovičních plánů

Příklad 6

Mějme faktory 1, 2, 3, 4, 5 a poloviční plán 2^{5-1} .

Generátor má tvar $5 = 1234$. S tímto generátorem se vytvoří první polovina (první frakce) úplného plánu. Spolu se všemi interakcemi dvou faktorů je plán v tabulce:

1	2	3	4	5	12	13	14
-	-	-	-	+	+	+	+
+	-	-	-	-	-	-	-
-	+	-	-	-	-	+	+
+	+	-	-	+	+	-	-
-	-	+	-	-	+	-	+
+	-	+	-	+	-	+	-
-	+	+	-	+	-	-	+
+	+	+	-	-	+	+	-
-	-	-	+	-	+	+	-
+	-	-	+	+	-	-	+
-	+	-	+	+	-	+	-
+	+	-	+	-	+	-	+
-	-	+	+	+	+	-	-
+	-	+	+	-	-	+	+
-	+	+	+	-	-	-	-
+	+	+	+	+	+	+	+
Pokračování							
15	23	24	25	34	35	45	Y
-	+	+	-	+	-	-	56
-	+	+	+	+	+	+	53
+	-	-	-	+	+	+	63
+	-	-	+	+	-	-	65
+	-	+	+	-	-	+	53
+	-	+	-	-	+	-	55
-	+	-	+	-	+	-	67
-	+	-	-	-	-	+	61
+	+	-	+	-	+	-	69

+	+	-	-	-	-	+	45
-	-	+	+	-	-	+	78
-	-	+	-	-	+	-	93
-	-	-	-	+	+	+	49
-	-	-	+	+	-	-	60
+	+	+	-	+	-	-	95
+	+	+	+	+	+	+	82

Efekt jednotlivých faktorů vychází:

1 = -2	14 = -0,75
2 = 20,5	15 = 1,25
3 = 0	23 = 1,5
4 = 12,25	24 = 10,75
5 = -6,25	25 = 1,25
12 = 1,5	34 = 0,25
13 = 0,5	35 = 2,25
	45 = -9,5

V tomto plánu mají některé interakce stejné posloupnosti znamének a proto i stejný efekt, např. 123 a 45. Efekt interakce 45 je $ef(45) = -9,5$. Protože je to společný efekt s interakcí 123, píšeme

$$ef(45) = -9,5 \rightarrow 45 + 123$$

Druhá polovina (druhá frakce) k plánu 2^{5-1} má generátor

$$5 = - 1234$$

takže I = - 12345

Zaměnitelný faktor např. k 2 je v první frakci 1345 a ve druhé frakci - 1345. Efekt faktoru 2 je v první frakci

$$ef(2) = 20,5 \rightarrow 2 + 1345$$

a ve druhé frakci

$$ef(2) = 18,5 \rightarrow 2 - 1345$$

„Čistý“ vliv faktorů 2 a 1345 by se vypočítal jen při znalosti obou frakcí.

Jelikož je k dispozici jen jedna frakce, nelze „čistý“ efekt počítat. Je proto žádoucí, aby v zaměnitelných dvojicích s faktory byly co nejdelší interakce. Toho se dosáhne tehdy, je-li vysoký stupeň řešení.

U polovičních plánů, sleduje-li se k faktorů 1, 2, 3, ..., k, je nejlepší tento postup:

a) sestavit úplný plán pro faktory 1,2,..., k-1

b) k-tý faktor vyjádřit pomocí faktorů 1,2,3,..., (k-1):

$$k = 1.2.3. \dots . (k-1)$$

Doplňkové frakce u středových a saturovaných plánů

V plánu 2^{7-4} budou k dispozici tyto generátory pro vytvoření všech 16 frakcí (každá frakce je tvořena 8 pokusy)

$$D = \pm AB, E = \pm AC, F = \pm BC, G = \pm ABC.$$

Plán 2^{7-4} a výsledky pokusů (pro generátory s kladnými znaménky) jsou v tabulce:

	A	B	C	D = AB	E = AC	F = BC	G = ABC	Y
1	-	-	-	+	+	+	-	69
2	+	-	-	-	-	+	+	52
3	-	+	-	-	+	-	+	60
4	+	+	-	+	-	-	-	83
5	-	-	+	+	-	-	+	71
6	+	-	+	-	+	-	-	50
7	-	+	+	-	-	+	-	59
8	+	+	+	+	+	+	+	88

Druhou frakci plánu 2^{7-4} vytvoříme *například* takto:

$$D = -AB, E = AC, F = BC, G = ABC,$$

takže jen u faktoru D se změnilo znaménko.

Plán této frakce a výsledky pokusů jsou v následující tabulce:

	A	B	C	D = -AB	E = AC	F = BC	G = ABC	Y
Pokus								
9	-	-	-	-	+	+	-	47
10	+	-	-	+	-	+	+	74
11	-	+	-	+	+	-	+	84
12	+	+	-	-	-	-	-	62
13	-	-	+	-	-	-	+	53
14	+	-	+	+	+	-	-	78
15	-	+	+	+	-	+	-	87
16	+	+	+	-	+	+	+	60

U první frakce je například efekt faktoru A a zaměnitelné dvojice

$$ef_1(A) = 3.5 \rightarrow A + BD + CE + FG$$

U druhé vychází

$$ef_2(A) = 0.75.$$

Zaměnitelné dvojice u druhé frakce se získají z první tak, že se změni znaménka v definičních rovnicích tam, kde je faktor D

$$ef_2(A) = 0.75 \rightarrow A - BD + CE + FG.$$

Efekt faktoru A i zaměnitelné dvojice *ve spojeném* plánu lze získat jako průměr z obou frakcí

$$\begin{aligned} \frac{1}{2}(ef_1(A) + ef_2(A)) &= \frac{1}{2}(3.5 + 0.75) \rightarrow \\ \rightarrow \frac{1}{2}[(A + BD + CE + FG) + (A - BD + CE + FG)] &= A + CE + FG = 2,125 \end{aligned}$$

(ověřte výpočtem $ef(A)$ z celého plánu)

Zdůvodnění je jednoduché:

$$\frac{1}{2}\left(\frac{Y_1 + \dots + Y_8}{4} + \frac{Y_9 + \dots + Y_{16}}{4}\right) = \frac{Y_1 + \dots + Y_{16}}{8}$$

Výraz vlevo je průměr 1. frakce a 2. frakce, výraz vpravo je efekt počítaný z obou (sloučených) frakcí.

2.7 Plán experimentu pro osm faktorů

Další cestou ke zkrácení plánu při jednoduché analýze výsledků je konstrukce speciálních plánů. Například pro 8 faktorů je možné sestavit speciální plány v těchto krocích:

- prvních osm pokusů naplánovat stejně jako pro plán 2^{7-4} s tím, že osmý sloupec je tvořen jen znaménky plus
- dalších osm pokusů (tj. pokusy 9 – 16) tvoří doplňkovou frakci s opačnými znaménky

Výhody tohoto postupu:

- nejméně pokusů
- počítá čistý efekt (bez součtů)

Pok	1	2	3	4	5	6	7	8	Y
1	-	-	-	+	+	+	-	+	14,0
2	+	-	-	-	-	+	+	+	16,8
3	-	+	-	-	+	-	+	+	15,0
4	+	+	-	+	-	-	-	+	15,4
5	-	-	+	+	-	-	+	+	27,6
6	+	-	+	-	+	-	-	+	24,0
7	-	+	+	-	-	+	-	+	27,4
8	+	+	+	+	+	+	+	+	22,6
9	+	+	+	-	-	-	+	-	22,3
10	-	+	+	+	+	-	-	-	17,1
11	+	-	+	+	-	+	-	-	21,5
12	-	-	+	-	+	+	+	-	17,5
13	+	+	-	-	+	+	-	-	15,9
14	-	+	-	+	-	+	+	-	21,9
15	+	-	-	+	+	-	+	-	16,7
16	-	-	-	-	-	-	-	-	20,3

Na závěr této kapitoly uvedeme doporučení pro tvorbu částečných plánů:

- a) plán začít s omezeným počtem faktorů, které stačí k řízení procesu
- b) generátory volit podle doporučení v tabulkách
- c) po výpočtu efektů vynechat nevýznamné faktory (projekce)
- d) přidat podle potřeby vhodné doplňkové frakce
- e) na začátku plánování nedělat úplný plán ani plán s opakováním
- f) u většího počtu faktorů, potřebných hned v úvodu, hledat speciální plány



Shrnutí druhé kapitoly

Tato kapitola je rozšířením poznatků z první kapitoly. Přináší zejména informace o tom, jak snížit počet pokusů a které problémy – vedle výhod – toto snížení přináší.



Otázky ke kapitole 2

1. Jak značíme částečný faktorový experiment
2. Jak stanovíme počet pokusů v částečném experimentu
3. Proč se provádí částečný faktorový experiment
4. Jak rozdělujeme částečné faktorové experimenty
5. Co je jednotkový faktor
6. Uveďte vlastnosti operací s faktory
7. Jak se stanoví maximální stupeň snížení
8. Co je generátor, definiční rovnice a stupeň řešení
9. Co jsou a jak se naleznou zaměnitelné dvojice
10. Co ovlivní volba generátoru
11. Jak se zjistí optimální volba generátoru
12. Jak se stanoví počet generátorů a počet definičních rovnic
13. Jak se stanoví počet zaměnitelných faktorů
14. Jaký je optimální generátor u polovičního plánu
15. Co je projekce plánu
16. Co způsobí projekce u úplného plánu
17. Co způsobí projekce u polovičního plánu
18. K čemu slouží doplňkové frakce
19. Napište doplňkovou frakci k polovičnímu plánu s 5 faktory
20. Jak se snadno stanoví zaměnitelné faktory a efekty při sloučení dvou frakcí
21. Vysvětlíte konstrukci plánu pro 8 faktorů



Úlohy ke kapitole 2

Příklad 1

Napište částečný plán 2^{5-2} pro faktory A,B,C,D,E, jsou li výsledky pokusů postupně při prvním provedení 17,16,18,15,17,30,18,29, a při druhém provedení 16,18,17,16,19,31,18,28.

	A	B	C	D	E	Y ₁	Y ₂
1	-	-	+	-	-	17	16
2	+	-	-	-	+	16	18
3	-	+	-	-	+	18	17
4	+	+	+	-	-	15	16
5	-	-	+	+	+	17	19
6	+	-	-	+	-	30	31
7	-	+	-	+	-	18	18
8	+	+	+	+	+	29	28

Dále:

- vypočítejte efekty
- určete všechny možné zaměnitelné faktory

Příklad 2

Udělejte částečný plán 2^{5-2} pro faktory A,B,C,D,E, Vypočítejte efekty a sestrojte tabulku potřebnou ke grafickému hodnocení efektů.

Generátory volte: D = AB a E = AC

Plán a výsledky pokusů jsou v tabulce:

	A	B	C	D	E	Y
1	-	-	-	+	+	8,5
2	+	-	-	-	-	1
3	-	+	-	-	+	5
4	+	+	-	+	-	8,5
5	-	-	+	+	-	9,5
6	+	-	+	-	+	1
7	-	+	+	-	-	7
8	+	+	+	+	+	7,5



Klíč k řešení kap. 2

Řešení(1):

Částečný plán je 2^{4-2}_{III}

	A	B	C = AB	D	E = ABD	Y ₁	Y ₂	$\bar{Y}$
1	-	-	+	-	-	17	16	16,5
2	+	-	-	-	+	16	18	17
3	-	+	-	-	+	18	17	17,5
4	+	+	+	-	-	15	16	15,5
5	-	-	+	+	+	17	19	18
6	+	-	-	+	-	30	31	30,5
7	-	+	-	+	-	18	18	18
8	+	+	+	+	+	29	28	28,5

a) výpočet efektů:

$$\text{ef}(A) = \frac{1}{4} (-16,5 + 17 - 17,5 + 15,5 - 18 + 30,5 - 18 + 28,5) = 5,375$$

$$\text{ef}(B) = \frac{1}{4} (-16,5 + 17 + 17,5 + 15,5 - 18 - 30,5 + 18 + 28,5) = -0,625$$

$$\text{ef}(C+AB) = \frac{1}{4} (16,5 - 17 - 17,5 + 15,5 + 18 - 30,5 - 18 + 28,5) = -1,125$$

$$\text{ef}(D) = \frac{1}{4} (-16,5 - 17 - 17,5 - 15,5 + 18 + 30,5 + 18 + 28,5) = 7,125$$

$$\text{ef}(E+ABD) = \frac{1}{4} (-16,5 + 17 + 17,5 - 15,5 + 18 - 30,5 - 18 + 28,5) = 0,125$$

Generátory jsou:

$$\begin{aligned} E &= ABD \\ C &= AB \end{aligned}$$

Definiční rovnice:

$$I = ABDE = ABC = CDE$$

Zaměnitelné faktory:

$$\begin{aligned} A &+ BC \\ B &+ AC \\ C &+ AB + DE \\ D &+ CE \\ E &+ CD \\ AD &+ BE \\ BD &+ AE \end{aligned}$$

Řešení(2):

Definiční rovnice: $I = ABD, I = ACE, I = BCDE$

Zaměnitelné faktory:

Zaměnitelné faktory s A:

$$A = BD = CE = ABCDE$$

Zaměnitelné faktory s B:

$$B = AD = ABCE = CDE$$

Zaměnitelné faktory s C:

$$C = ABCD = AE = BDE$$

Zaměnitelné faktory s D:

$$D = AB = ACDE = BCE$$

Zaměnitelné faktory s E:

$$E = ABDE = AC = BCD$$

Efekty:

$$ef(A) = (-8,5+1-5+8,5-9,5+1-7+7,5) / 4 = -3$$

$$ef(B) = (-8,5-1+5+8,5-9,5-1+7+7,5) / 4 = 2$$

$$ef(C) = (-8,5-1-5-8,5+9,5+1+7+7,5) / 4 = 0,5$$

$$ef(D) = (8,5-1-5+8,5+9,5-1-7+7,5) / 4 = 5$$

$$ef(E) = (8,5-1+5-8,5-9,5+1-7+7,5) / 4 = -1$$

Tabulka pro grafické určení významnosti faktorů:

	1	2	3	4	5
Efekt	-3	-1	0,5	2	5
Faktor	A	E	C	B	D
P_i	10	30	50	70	90

$$P_i = \frac{100 \cdot (t - 0,5)}{m}$$

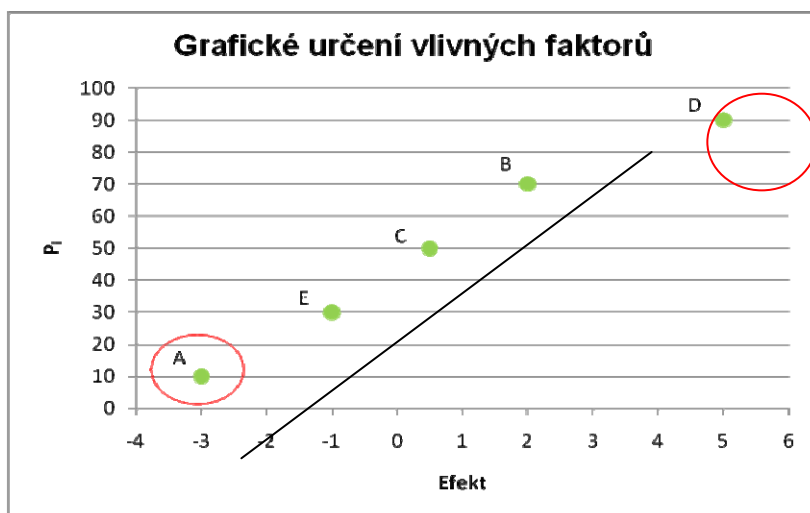
$$P_1 = \frac{100 \cdot (1 - 0,5)}{5} = 10$$

$$P_2 = \frac{100 \cdot (2 - 0,5)}{5} = 30$$

$$P_3 = \frac{100 \cdot (3 - 0,5)}{5} = 50$$

$$P_4 = \frac{100 \cdot (4 - 0,5)}{5} = 70$$

$$P_5 = \frac{100 \cdot (5 - 0,5)}{5} = 90$$



Významnými faktory jsou faktory A a D

3. VÝZNAMNÉ BODY PLÁNU



Čas ke studiu: 1 hodina



Cíl: Definovat tzv. významné body plánu a jejich vlastnosti

Při hledání kvadratických regresních modelů (čtvrtá kapitola) bude potřeba doplnit plán experimentu o body, se kterými jsme prozatím nepracovali. V této kapitole proto potřebné body probereme.

U každého z uváděných bodů je uvedeno jaké má souřadnice, doporučený počet a význam.



Výklad

V plánu experimentu se používají podle potřeby tři typy bodů:

- 1) krychlové body (factorial points)
- 2) centrální body (central points)
- 3) hvězdíkové body (axial = star points)

U každého typu uvedeme počet bodů v plánu, jejich souřadnice a význam.

Krychlové body jsou v plánu vždy.

Počet krychlových bodů je $n = 2^{k-p}$.

Souřadnice pro tři faktory: $(\pm 1, \pm 1, \pm 1)$.

Význam: slouží k výpočtu efektů faktorů.

Centrální body

Doporučený počet centrálních bodů je 3 – 5, nebo se počty určí z tabulky.

Souřadnice pro tři faktory: $(0,0,0)$

Význam:

- a) z Y_i v centrálních bodech se počítá průměr a srovnává se s průměrem Y_i v krychlových bodech (test zda postačuje lineární model = test křivosti)
- b) z Y_i v centrálních bodech se počítá odhad rozptylu σ^2 jako náhrada za odhad σ^2 z opakovaných pokusů (příklad 1)
- c) z e_i v centrálních bodech se počítá čistá chyba měření (pure error)

Centrální body mohou být přidány k plánu 2^{k-p} , pokud jsou všechny faktory *kvantitativní*.

U *kvalitativních* faktorů neexistuje úroveň 0.

Centrální body se nepoužívají k výpočtu efektů faktorů.

Hvězdíkové body

Při dvou faktorech x_1, x_2 leží na kružnici se středem $(0,0)$, která prochází vrcholy čtverce.

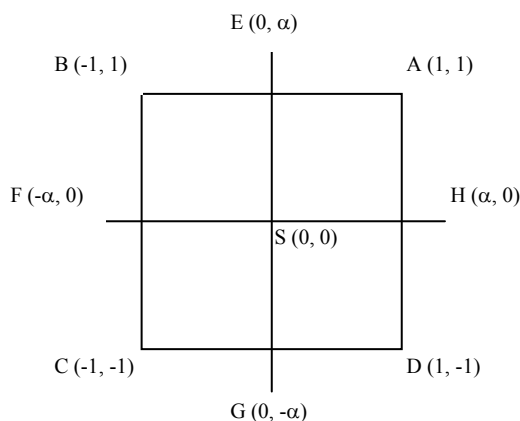
U třech faktorů jsou to body, kde kulová plocha se středem v počátku a procházející krychlovými body plánu, protíná souřadné osy.

Počet hvězdíkových bodů: $n = 2k$

Souřadnice pro tři faktory: $(\pm\alpha, 0, 0), (0, \pm\alpha, 0), (0, 0, \pm\alpha)$, kde *doporučené* $\alpha = \sqrt[4]{n_k}$, n_k je počet krychlových bodů.

Význam:

- umožňují výpočet koeficientů v úplném kvadratickém modelu
- zpřesňují výpočet regresních koeficientů



Obr.1 Významné body plánu 2^2

Schéma plánu 2^2 , obsahujícího uvedené body:

x_1	x_2		Typ bodu
-	-	C	Krychlové body
+	-	D	
-	+	B	
+	+	A	
0	0	S	Centrální bod
α	0	H	Hvězdicové body
$-\alpha$	0	F	
0	α	E	
0	$-\alpha$	G	

Příklad 1 (výpočet s_e^2 v plánu bez opakování)

Vypočítejte rozptyl odhadu efektů pomocí centrálních bodů pro následující plán experimentu.

Pokus	1	2	3	Y	Faktory a úrovně	-	0	+
1	-	-	-	160	1	330	480	700
2	+	-	-	37	2	0,010	0,015	0,022
3	-	+	-	165	3	0,049	0,070	0,100
4	+	+	-	22	Y ukazatel kvality			
5	-	-	+	172				
6	+	-	+	35				
7	-	+	+	120				
8	+	+	+	18				
9	0	0	0	66				
10	0	0	0	83				
11	0	0	0	71				
12	0	0	0	82				

Řešení:

Průměr $\bar{Y}$ v centrálních bodech je

$$\bar{Y}_C = \frac{1}{4}(66 + 83 + 71 + 82) = 75,5$$

Rozptyl s^2 v centrálních bodech

$$s^2 = \frac{(66 - 75,5)^2 + (83 - 75,5)^2 + (71 - 75,5)^2 + (82 - 75,5)^2}{4 - 1} = 69,6$$

Rozptyl odhadu efektu

$$s_e^2 = \frac{4s^2}{n_k} = \frac{4 \cdot 69,6}{8} = 34,83$$

a směrodatná odchylka

$$s_e = 5,9.$$



Shrnutí kapitoly 3

Uvedli jsme tři typy bodů: krychlové, centrální a hvězdicové. U každého z nich je potřeba vědět, jaké má souřadnice, jak je umístěn v grafickém znázornění plánu a jaký je jeho význam.



Otázky ke kapitole 3

1. Vyjmenujte typy významných bodů plánu
2. Uveďte souřadnice a stanovení počtu u každého typu významných bodů
3. K čemu slouží centrální body
4. K čemu slouží hvězdicové body
5. Nakreslete významné body pro dva faktory
6. Jak se počítá rozptyl odhadu efektu pomocí centrálních bodů

4. ÚPLNÉ KVADRATICKÉ MODEL Y



Čas ke studiu: 6 hodin



Cíl: Objasnit postup při konstrukci úplného kvadratického modelu

Dosud jsme se zabývali (v první kapitole) metodikou nalezení lineárního regresního modelu. Protože někdy nestačí lineární model, je potřeba hledat komplexnější kvadratický model. První otázkou je, kdy je potřeba tento model, poté se řeší způsob jeho nalezení a nakonec jeho využití k výpočtu optimálního nastavení procesu. Stručně lze říci, že ve srovnání s postupem u lineárních modelů je u kvadratických modelů všechno jinak. V této kapitole ukážeme, jak se výše uvedené otázky řeší.



Výklad

4.1 Podmínky nalezení modelu

Pro nalezení úplného kvadratického modelu je nutné splnit minimálně tyto podmínky

- plán experimentu musí mít alespoň $(1 + 2k + k(k-1))/2$ bodů (viz počet členů v rovnici (1)),
- každá proměnná v plánu musí mít alespoň tři úrovně (tomu lze vyhovět buď doplněním plánu o hvězdicové body = kombinovaný plán, nebo přímo volit 3 úrovně faktorů = tříúrovňový plán)

Pozn.:

Bod (a) lze formulovat obecněji: počet pokusů nesmí být menší než počet koeficientů v hledaném modelu; u lineárních modelů pak dostáváme podmínku jednodušší: $n \geq k$; n = počet pokusů, k = počet faktorů.

Dříve, než začneme hledat úplný kvadratický model, ověříme, je-li takový model potřeba; to je možné zjistit různými způsoby:

- testem křivosti
- testem lack-of-fit (kap.6)
- pomocí ANOVA (kap.6)

V této kapitole ukážeme jen test křivosti. Další dvě metody jsou obsáhlejší a je jim proto věnována samostatná kapitola.

4.2 Test křivosti (curvature)

K testu křivosti se používají centrální body. Jejich počet označíme n_C , průměrné y v centrálních bodech je $\bar{y}_C$ a v krychlových bodech $\bar{y}_F$ (F-factorial points); počet krychlových bodů označíme n_F .

Nulová hypotéza testu křivosti

$$H_0 : \sum_{j=1}^k \beta_{jj} = 0$$

tzn., není nutný kvadratický model.

Testovací kritérium je

$$T = \frac{\frac{SS_{CPQ}}{1}}{\frac{\sum (y_i - \bar{y}_C)^2}{n_C - 1}}$$

SS_{CPQ} je součet čtverců odchylek *kvadratického zakřivení* a počítá se

$$SS_{CPQ} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_C + n_F}$$

Kritická hodnota

$$F_{1, n_C - 1}(\alpha)$$

Pro

$$T \leq F_{1, n_C - 1}(\alpha)$$

není nutný kvadratický model.

Příklad 1

Y v centrálních bodech je: 40.3, 40.5, 40.7, 40.2, 40.6;

průměr v centrálních bodech je $\bar{y}_C = 40,46$;

Y pro krychlové body: 41.5, 40, 39.3, 40.9;

průměr v krychlových bodech je 40,425.

Má se provést test křivosti

Řešení:

Testovací kritérium

$$T = \frac{\frac{SS_{CPQ}}{1}}{\frac{\sum (y_i - \bar{y}_C)^2}{n_C - 1}} = \frac{\frac{4.5(40,425 - 40,46)^2 / (4 + 5)}{1}}{\frac{(40,3 - 40,46)^2 + \dots + (40,6 - 40,46)^2}{5 - 1}} = 0,063$$

Kritická hodnota

$$F_{1, n_C - 1}(\alpha) = F_{1, 4}(0,05) = 12,22$$

Závěr:

Protože T nepřevyšuje $F(0,05)$, H_0 platí což znamená, že není potřeba kvadratický model.

4.3 Metody nalezení kvadratického modelu

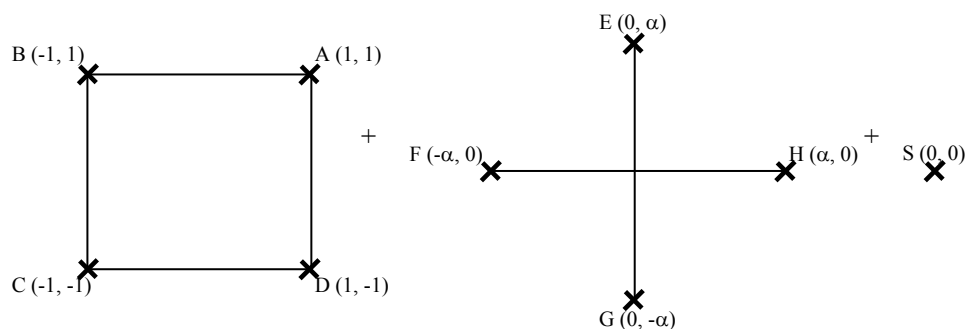
Koeficienty úplného kvadratického modelu nejsou polovinou efektů, jak tomu bylo u lineárního a neúplného kvadratického modelu. Používají se proto jiné metody k jejich nalezení. Dále uvedeme tyto tři:

- Kombinovaný plán (Composite design)
- Tříúrovňový plán (Three – level design)
- Box – Behnkenův plán

4.3.1 Kombinovaný plán

Kombinovaný plán tvoří:

krychlové body + hvězdicové body + centrální body.



U centrálních plánů, kdy je střed v bodě $(0; 0)$, se vytvoří hvězdicový plán snadno, jen pomocí dvou nových úrovní $\pm \alpha$ (čím méně úrovní, tím lépe).

Je-li α počítáno pomocí vztahu $\alpha = \sqrt[n_k]{n_k}$, kde n_k je počet kvadratických bodů v plánu, hovoříme o *rotovatelnosti* plánu.

U rotovatelných plánů je rozptyl Y stejný v bodech stejně vzdálených od středu. Např. rozptyl Y v bodech $(\pm 1; \pm 1)$, $(\pm \alpha; 0)$ a $(0; \pm \alpha)$ je stejný, neboť všechny jsou ve vzdálenosti 1 od středu $(0; 0)$.

Pokud je α počítáno jinak, plán není rotovatelný.

4.3.2 Výpočet optimální úrovně významných faktorů

U kvadratických modelů se hledají

- stacionární body plochy = optimální *hodnoty* faktorů
(u lineárních modelů hovoříme o optimální *úrovni* faktoru: dolní resp. horní; extrémů nejsou - grafem je rovina resp. nadrovina)
- oblasti stacionárních bodů
- typy extrémů (max., min.) resp. sedlový bod

Mějme úplný kvadratický model

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \beta_{ii} x_i^2 + \sum_{i < j} \sum \beta_{ij} x_i x_j, \quad (1)$$

který lze zapsat také pomocí matic

$$y = \beta_0 + x^T b + x^T B x \quad (2)$$

kde B je (symetrická) matice odhadu koeficientů β_{ij} , b je vektor odhadu koeficientů β_i .

Určení stacionárních bodů

K nalezení stacionárních bodů použijeme věty V1 a V2 o derivování podle vektoru:

$$\text{Je-li } \vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_n \end{bmatrix}, \quad \vec{\varpi} = \begin{bmatrix} \varpi_1 \\ \varpi_2 \\ \dots \\ \varpi_n \end{bmatrix}, \quad A = \begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ \dots & & \dots \\ a_{n,1} & \dots & a_{n,n} \end{bmatrix}, \text{ pak platí}$$

Věta 1

$$\frac{\delta \omega^T \alpha}{\delta \varpi} = \alpha$$

Věta 2

$$\frac{\delta \varpi^T A \varpi}{\delta \varpi} = (A + A^T) \varpi$$

Derivujme (2) podle $\mathbf{x}$. Dostáváme

$$\frac{\partial \hat{y}}{\partial \mathbf{x}} = \mathbf{b} + (\mathbf{B} + \mathbf{B}^T)\mathbf{x} = \mathbf{b} + 2\mathbf{B}\mathbf{x} = 0 \Rightarrow \mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}$$

Vektor stacionárních bodů tedy vypočítáme z rovnice

$$\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}.$$

Příslušnou hodnotu $\hat{y}_s$ získáme dosazením $\mathbf{x}_s$ za $\mathbf{x}$ do úplného kvadratického modelu

$$\hat{y}_s = b_0 + \mathbf{x}^T \mathbf{b} + \mathbf{x}^T \mathbf{B} \mathbf{x}$$

Dostáváme

$$\hat{y}_s = b_0 + \frac{1}{2} \mathbf{x}^T \mathbf{b}$$

Typ extrému se určí pomocí kořenů *charakteristické rovnice* matice $\mathbf{B}$

$$\det(\mathbf{B} - \lambda \mathbf{I}) = 0 \quad (4)$$

Kořeny charakteristické rovnice se nazývají *charakteristická čísla* nebo *vlastní čísla* matice $\mathbf{B}$.

Typ extrému:

- a) Jsou-li všechny kořeny $\lambda_1, \dots, \lambda_k$ kladné, jedná se o *minimum*
- b) Jsou-li všechny kořeny $\lambda_1, \dots, \lambda_k$ záporné, jedná se o *maximum*
- c) Jsou-li kořeny $\lambda_1, \dots, \lambda_k$ smíšené, jedná se o *sedlový bod*

V příkladu 2 ukážeme, že stacionární body je možné hledat také pomocí derivací rovnice (1).

Příklad 2

Určit optimální hodnoty významných parametrů pro úplný kvadratický model, $k = 3$:

$$y = b_0 + \sum_{i=1}^3 b_i x_i + \sum_{i=1}^3 b_{ii} x_i^2 + \sum_{i < j} b_{ij} x_i x_j$$

Postup řešení:

- a) derivovat úplný kvadratický model podle všech proměnných
- b) vzniklou soustavu rovnic zapsat maticově a vypočítat matici $\mathbf{X}$
- c) zapsat maticově úplný kvadratický model

d) dosadit za X do (c)

Řešení:

a) Rozepíšeme úplný kvadratický model

$$y = b_0 + (b_1x_1 + b_2x_2 + b_3x_3) + (b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2) + (b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3)$$

Derivujeme

$$\frac{\partial y}{\partial x_1} = b_1 + 2b_{11}x_1 + b_{12}x_2 + b_{13}x_3$$

$$\frac{\partial y}{\partial x_2} = b_2 + 2b_{22}x_2 + b_{12}x_1 + b_{23}x_3$$

$$\frac{\partial y}{\partial x_3} = b_3 + 2b_{33}x_3 + b_{13}x_1 + b_{23}x_2$$

Položíme rovno nule a upravíme

$$2b_{11}x_1 + b_{12}x_2 + b_{13}x_3 = -b_1$$

$$b_{12}x_1 + 2b_{22}x_2 + b_{23}x_3 = -b_2$$

$$b_{13}x_1 + b_{23}x_2 + 2b_{33}x_3 = -b_3$$

b) Soustavu napíšeme v maticovém tvaru

$$2 \begin{bmatrix} b_{11} & \frac{b_{12}}{2} & \frac{b_{13}}{2} \\ \frac{b_{12}}{2} & b_{22} & \frac{b_{23}}{2} \\ \frac{b_{13}}{2} & \frac{b_{23}}{2} & b_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

symbolicky

$$2 \cdot \mathbf{B} \cdot \mathbf{x} = -\mathbf{b}$$

$$\mathbf{x} = \frac{-1}{2} \mathbf{B}^{-1} \mathbf{b}$$

Z porovnání maticového a symbolického zápisu je zřejmé, co je B a b.

Příklad 3 (určení typu stacionárních bodů)

Mají se najít stacionární body a jejich typ pro kvadratický model se dvěma proměnnými x_1, x_2 . Plán experimentu má tvar 2^2 a je doplněn o hvězdicové body a čtyři centrální body:

$$\begin{aligned}x_1(\min) &= 200, \\x_1(\max) &= 250, \\x_2(\min) &= 15, \\x_2(\max) &= 25.\end{aligned}$$

Plán experimentu

	x_1	x_2	Y
1	-1	-1	43
2	1	-1	78
3	-1	1	69
4	1	1	73
5	-1,414	0	48
6	1,414	0	76
7	0	-1,414	65
8	0	1,414	74
9	0	0	76
10	0	0	79
11	0	0	83
12	0	0	81

Řešení:

Vychází model

$$Y = 79,75 + 9,83x_1 + 4,22x_2 - 8,88x_1^2 - 5,13x_2^2 - 7,75x_1x_2$$

Odtud

$$B = \begin{bmatrix} -8,88 & -3,875 \\ -3,875 & -5,13 \end{bmatrix} \text{ a } b = \begin{bmatrix} 9,83 \\ 4,22 \end{bmatrix}$$

$$\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} = \frac{-1}{2} \begin{bmatrix} -0,16798 & 0,12689 \\ 0,12689 & -0,29078 \end{bmatrix} \begin{bmatrix} 9,83 \\ 4,22 \end{bmatrix} = \begin{bmatrix} 0,5579 \\ -0,0101 \end{bmatrix}$$

Optimální *kódované* hodnoty (stacionární body) proměnných x_1, x_2 jsou 0.5579 a 0.0101. Při této volbě vstupů je optimální $Y = 82,47$.

V reálných hodnotách, označených řeckými písmeny ξ_1, ξ_2 , máme reálné optimální hodnoty

$$\begin{aligned}\xi_1 &= 225 + 25x_1 = 225 + 25 \cdot 0,5579 = 238,95 \\ \xi_2 &= 20 + 5x_2 = 20 + 5 \cdot (-0,0101) = 19,95\end{aligned}$$

Typ extrému

V našem případě je charakteristická rovnice

$$\det \begin{bmatrix} -8,88 - \lambda & -3,85 \\ -3,85 & -5,13 - \lambda \end{bmatrix} = 0$$

odtud

$$\lambda^2 + 14,01\lambda + 30,54 = 0$$

a charakteristická čísla

$$\lambda_1 = -11,3098 \text{ a } \lambda_2 = -2,7002.$$

Protože oba kořeny jsou záporné, jedná se o maximum.

Kromě optimální úrovně parametrů je možné najít také *optimální oblast*, ve které je dosahováno nejlepší úrovně sledovaného výstupu Y a současně minimální variability. Tato oblast se nalezne jako průnik vrstevnicových grafů pro optimální Y a pro minimální variabilitu.

Vrstevnicové grafy

Příklad 4 (vrstevnicové grafy)

Pro faktory A,B je sestaven úplný plán, doplněný o 5 centrálních bodů. Protože se má najít úplný kvadratický model, je plán dále doplněn a hvězdicové body. Sledují se celkem tři ukazatele kvality, na které mají odběratelé tyto požadavky: $Y_1 > 78.5$, $62 < Y_2 < 68$ a $Y_3 < 3400$.

Má se najít kvadratický model ke každému ukazateli kvality.

Plán experimentu a výsledky pokusů jsou:

A	B	Y_1	Y_2	Y_3
-1	-1	76,5	62	2940
-1	1	77,0	60	3470
1	-1	78,0	66	3680
1	1	79,5	59	3890
0	0	79,9	72	3480
0	0	80,3	69	3200
0	0	80,0	68	3410
0	0	79,7	70	3290
0	0	79,8	71	3500
1,414	0	78,4	68	3360
-1,414	0	75,6	71	3020
0	1,414	78,5	58	3630
0	-1,414	77,0	57	3150

Byly nalezeny významné faktory pro každý výstup Y_i .

Pro jednotlivé výstupy vychází tyto modely (kódované hodnoty):

$$Y_1 = 79,94 + 0,99X + 0,52Y - 1,38X^2 - 1,00Y^2 + 0,25XY$$

$$Y_2 = 70,00 - 0,16X - 0,95Y - 0,69X^2 - 6,69Y^2 - 1,25XY$$

$$Y_3 = 3386,2 + 205,1X + 177,4Y$$

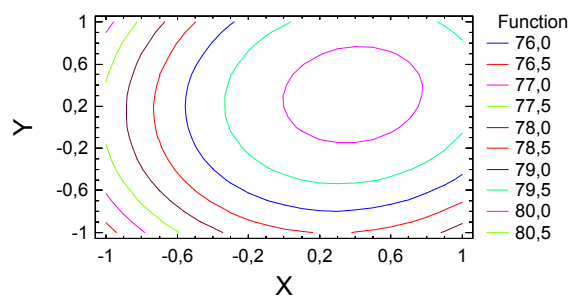
Pozn.

Pro potřeby konstrukce grafu s programem Statgraphics je v modelech místo označení A,B použito X,Y.

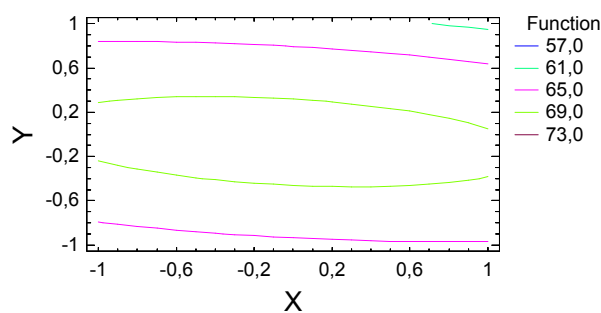
Protože se hledá oblast optimální regulace procesu, je potřeba najít tzv. vrstevnicové grafy. Nejprve se sestrojí grafy těchto ploch, dále pak řezy plochy rovinou (x,y).

Výstup Y₁

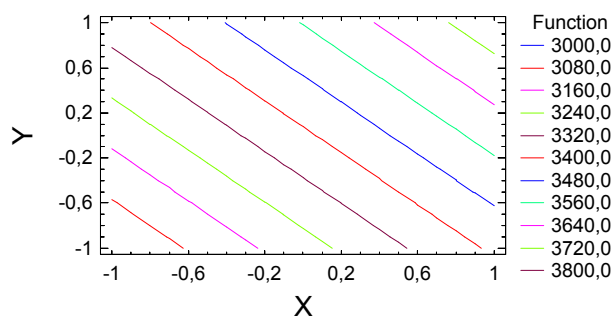
$$9,94 + 0,99 \cdot X + 0,52 \cdot Y - 1,38 \cdot X^2 - 1,00 \cdot Y^2 + 0,25 \cdot X \cdot Y$$

**Výstup Y₂**

$$70,00 - 0,16 \cdot X - 0,95 \cdot Y - 0,69 \cdot X^2 - 6,69 \cdot Y^2 - 1,25 \cdot X \cdot Y$$

**Výstup Y₃**

$$3386,2 + 205,1 \cdot X + 177,4 \cdot Y$$



Optimální režim výroby představuje průnik všech vrstevnicových grafů s vymezením oblastí, které vyhovují výchozím požadavkům.

Při konstrukci úplného kvadratického plánu jsme uvedli, že jsou různé možnosti sestavení plánu: kombinovaný plán, tříúrovňový plán a Box Benkenův plán. Poslední dva probereme v této kapitole.

4.3.3 Box - Behnkenův plán experimentu (BBD)

Sestavení plánu experimentu se provádí *pro tři faktory* takto:
jeden faktor je na úrovni 0 (zde první) a pro zbývající dva faktory se sestaví úplný plán (modře):

Pok.	x	y	z
1	0	-1	-1
2	0	1	-1
3	0	-1	1
4	0	1	1
5	-1	0	-1
6	1	0	-1
7	-1	0	1
8	1	0	1
9	-1	-1	0
10	1	-1	0
11	-1	1	0
12	1	1	0
13	0	0	0

Další faktor je na úrovni 0 (zde druhý) a pro zbývající dva faktory se sestaví úplný plán, atd.
Poslední řádek je *vektor* centrálních bodů.

Pro čtyři faktory: vytvoří se úplný plán 2^2 postupně pro všechny dvojice faktorů, přičemž ostatní jsou na úrovni 0.

Pro pět faktorů:

±1	±1	0	0	0
±1	0	±1	0	0
±1	0	0	±1	0
±1	0	0	0	±1
0	±1	±1	0	0
0	±	0	±1	0
0	±	0	0	±1
0	0	±1	±1	0
0	0	±	0	±1
0	0	0	±1	±1
0	0	0	0	0

Tento zápis je zkrácená podoba plánu: např. v prvním řádku $\pm 1, \pm 1, 0, 0, 0$ zápisu znamená, že pro první a druhou proměnnou se dělá úplný plán, ostatní 0. Zápis $\pm 1, \pm 1, 0, 0, 0$ tedy představuje 4 řádky plánu.

Poslední řádek je *vektor* centrálních bodů.

Počet pokusů pro $k = 3, 4, 5$: $n = \binom{k}{2} \cdot 4$

Pro dva faktory se BBD nedělá.

4.3.4 Plány pro faktory se třemi úrovněmi

Úplný faktorový plán pro tři faktory na třech úrovních:

Plán se sestaví podobně jako pro dvouúrovňové faktory:

- stanovíme počet pokusů: $n = 3^k$
- s tvorbou začneme např. u faktoru x_3 : střídáme úrovně -1, 0, 1; pokračujeme u x_2 : obměna úrovně je trojnásobek předchozí; nakonec u x_3 -obměna je trojnásobek předchozí, tedy po devíti stejných úrovních:

Pokus	x_1	x_2	x_3	y
1	- 1	- 1	- 1	
2	- 1	- 1	0	
3	- 1	- 1	1	
4	- 1	0	- 1	
5	- 1	0	0	
6	- 1	0	1	
7	- 1	1	- 1	
8	- 1	1	0	
9	- 1	1	1	
10	0	- 1	- 1	
11	0	- 1	0	
12	0	- 1	1	
13	0	0	- 1	
14	0	0	0	
15	0	0	1	
16	0	1	- 1	
17	0	1	0	
18	0	1	1	
19	1	- 1	- 1	
20	1	- 1	0	
21	1	- 1	1	
22	1	0	- 1	
23	1	0	0	
24	1	0	1	
25	1	1	- 1	
26	1	1	0	
27	1	1	1	

Příklad 5

Sestavte úplný plán pro dva faktory A,B, které mají tři úrovně.
S tvorbou začněte u faktoru A.

Řešení:

	A	B	Y
1	-1	-1	
2	0	-1	
3	1	-1	
4	-1	0	
5	0	0	

6	1	0	
7	-1	1	
8	0	1	
9	1	1	

Pozn.: Délka tohoto plánu oproti dvouúrovňovému není příliš odlišná. U tří faktorů však je $n = 3^3 = 27$, kdežto pro dvě úrovně je počet pokusů $n = 2^3 = 8$. Proto, pokud je to možné, upřednostňujeme dvouúrovňové plány.



Shrnutí kapitoly 4

V této kapitole bylo vysvětleno, kdy je potřeba kvadratický model a jak se sestojí. Celkem byly uvedeny tři způsoby možného postupu a také důvod, proč nelze uplatnit metodiku, použitou u lineárních modelů.



Otázky ke kapitole 4

1. Jaké jsou podmínky sestavení plánu pro nalezení úplného kvadratického modelu
2. Kolik je koeficientů v úplném kvadratickém modelu se dvěma faktory
3. K čemu slouží test křivosti
4. Napište všechny body testu křivosti
5. Uveďte tři metody nalezení úplného kvadratického modelu
6. Jak se stanoví α u hvězdicových bodů
7. Co je rotovatelnost plánu
8. Napište maticově úplný kvadratický model
9. Proveďte odvození vzorce pro výpočet stacionárních bodů a uveďte potřebné pomocné věty
10. Napište prvky matice B a vektoru b
11. Napište charakteristickou rovnici matice B
12. Jak poznáme typ extrému u kvadratického modelu
13. Jak se nalezne a k čemu slouží vrstevnicový graf
14. Napište Box – Behnkenův plán pro 3, 4 a 5 faktorů
15. Napište plán pro 3 faktory se třemi úrovněmi
16. Napište plán pro 2 faktory se třemi úrovněmi



Cvičení ke kapitole 4

Příklad

Najděte úplný kvadratický model a matice B a b k plánu:

	x_1	x_2	Y
1	-1	-1	12
2	1	-1	12
3	-1	1	13

4	1	1	15
5	-1,414	0	10
6	1,414	0	11
7	0	-1,414	14
8	0	1,414	18
9	0	0	17
10	0	0	15
11	0	0	16
12	0	0	16



KLÍČ K ŘEŠENÍ KAP. 4

c	X_1	X_2	X_{12}	X_{11}	X_{22}
16,0	0,42	1,20	0,50	-2,82	-0,06

$$b = \begin{bmatrix} 0,42 \\ 1,20 \end{bmatrix}, \quad B = \begin{bmatrix} -2,82 & 0,25 \\ 0,25 & -0,06 \end{bmatrix}$$

5. BLOKY



Čas ke studiu: 3 hodiny



Cíle:

- objasnit smysl používání bloků
- ukázat postup konstrukce bloků
- vysvětlit způsob vyhodnocení vlivu bloků



Výklad

Při realizaci experimentu musí být provedeny všechny pokusy za stejných podmínek. Mění se jen to, co je stanoveno plánem. Pokud není možné dodržet tuto zásadní podmínku (například nelze provést celý experiment během jednoho dne, nebo z jedné dávky suroviny), provádí se rozdělení pokusů do tzv. bloků. Vytvořit bloky prakticky znamená rozdělit pokusy do skupin, ve kterých jsou stejné podmínky pro jejich provedení (např. pokusy provedené první den tvoří první blok, pokusy z druhého dne pak druhý blok). Bloky se vytvářejí tak, aby

- při výpočtu efektu byl eliminován případný vliv toho, že během experimentu nelze zajistit pro všechny pokusy stejné podmínky,
- při hodnocení výsledků experimentu bylo možné stanovit, zda bloky mají vliv na výsledek.

Bloky se tvoří pomocí *generátorů bloků*, které jsou uvedeny v tabulkách. Máme tedy kromě generátorů vedlejších faktorů a generátorů frakcí ještě generátory bloků.

Výběr z tabulky generátorů bloků:

Počet proměnných k	Velikost bloku	Generátor bloku
3	4	B1=123
	2	B1=12, B2=13
4	8	B1=1234
	4	B1=124, B2=134
	2	B1=12, B2=23, B3=34
5	16	B1=12345
	8	B1=123, B2=345
	4	B1=125, B2=235, B3=345
	2	B1=12, B2=13, B3=34, B4=45
6	32	B1=123456
	16	B1=1236, B2=3456
	8	B1=135, B2=1256, B3=1234
	4	B1=126, B2=136, B3=346, B4=456
	2	B1=12, B2=23, B3=34, B4=45, B5=56

Tab.1 Generátory bloků

5.1 Konstrukce bloků

Příklad 1 (tři faktory, čtyři bloky)

Rozdělte do 4 bloků pokusy v plánu 2^3 .

Řešení:

Je potřeba rozdělit $2^3 = 8$ pokusů do 4 bloků, které tvoří $8/4 = 2$ pokusy. V tabulce pro $k = 3$ a velikost bloku 2 jsou generátory bloků $B1=12$ a $B2 = 13$.

Existují 4 varianty znamének, které vytvoří B1, B2. Příslušné bloky očíslováme:

1	2	3	12	13	blok
-1	-1	-1	1	1	IV
1	-1	-1	-1	-1	I
-1	1	-1	-1	1	III
1	1	-1	1	-1	II
-1	-1	1	1	-1	II
1	-1	1	-1	1	III
-1	1	1	-1	-1	I
1	1	1	1	1	IV

Máme tedy tyto bloky:

Blok	I	II	III	IV
(B1,B2)	(-, -)	(+, -)	(-, +)	(+, +)

Konečné rozdělení experimentu do bloků bude

Č.pokusu	Experimentální proměnné (faktory)			Blokové proměnné		Č.bloku
	1	2	3	B1=12	B2=13	
1	-	-	-	+	+	IV
2	+	-	-	-	-	I
3	-	+	-	-	+	III
4	+	+	-	+	-	II
5	-	-	+	+	-	II
6	+	-	+	-	+	III
7	-	+	+	-	-	I
8	+	+	+	+	+	IV

Příklad 2 (tři faktory, dva bloky)

Mějme faktory A,B,C, pro které je proveden úplný plán:

	A	B	C	AB	AC	BC	ABC	Blok
1	-	-	-	+	+	+	-	I
2	+	-	-	-	-	+	+	II
3	-	+	-	-	+	-	+	II
4	+	+	-	+	-	-	-	I
5	-	-	+	+	-	-	+	II
6	+	-	+	-	+	-	-	I
7	-	+	+	-	-	+	-	I
8	+	+	+	+	+	+	+	II

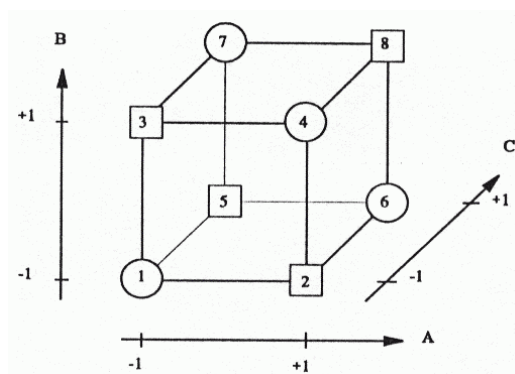
Pokusy rozdělte do dvou bloků. Porovnejte efekty vypočítané v nerozděleném plánu, plánu rozděleném do bloků správně a nesprávně.

Řešení:

Pro $k = 3$ a velikost bloku 4 je podle tab.1 generátor bloků $B1 = 123$.

Do prvního bloku patří ty pokusy, kde $ABC = -1$ a do druhého ty, kde $ABC = +1$. První blok znamená provedení pokusu první den, druhý blok provedení druhý den. Při grafickém znázornění tohoto plánu pomocí kvádry se ukazuje, že na každé hraně je pokus z I. a II. bloku (obr.1).

Obr.1 Rozdělení plánu 2^3 do dvou bloků



Takto naplánované rozdělení do bloků eliminuje vliv té nepříznivé skutečnosti, že pokusy není možné provést během jednoho dne.

Vliv správného a nesprávného rozdělení do bloků ukazují následující 3 tabulky, které představují

- experiment provedený během jednoho dne (= kontrolní experiment),
- experiment provedený dle výše uvedeného plánu,
- experiment rozdělený do bloků bez generátorů.

Tabulka (2a):

pokus	A	B	C	Y
1	-	-	-	42
2	+	-	-	49
3	-	+	-	37
4	+	+	-	46
5	-	-	+	32
6	+	-	+	41
7	-	+	+	33
8	+	+	+	44

Tabulka (2b):

Blok/ pokus	A	B	C	Y
I/6	+	-	+	41
I/4	+	+	-	46
I/1	-	-	-	42
I/7	-	+	+	33
II/5	-	-	+	38
II/3	-	+	-	43
II/8	+	+	+	50
II/2	+	-	-	55

Efekty faktorů počítaných pro tabulky a) a b) jsou stejné:

$A = 9.0$, $B = -1.0$, $C = -6.0$, $AB = 1.0$, $AC = 1.0$, $BC = 3.0$, $ABC = 6.0$.

To znamená, že rozdělení pokusů do dvou dnů neovlivnilo výsledek. Je ovšem otázka, zda rozdělení do bloků vůbec mělo na pokusy vliv. Odpověď je pozitivní: interakce ABC je významná, přitom lze předpokládat, že u vyšších interakcí je efekt zanedbatelný. Proto tuto významnou hodnotu vysvětlujeme jako vliv bloků. A to je také důvod, proč se rozdělení do bloků provádí pomocí interakcí více faktorů. Ve třetím případě, kdy byly pokusy rozděleny neplánovitě, se ukazuje, že výsledek interakce BC byl ovlivněn.

Tabulka (2c):

Blok/pok.	A	B	C	Y
I/4	+	+	-	46
I/6	+	-	+	41
I/5	-	-	+	32
I/3	-	+	-	37
II/8	+	+	+	50
II/1	-	-	-	48
II/2	+	-	-	55
II/7	-	+	+	39

Efekt faktorů je zde:

$A = 9.0$, $B = -1.0$, $C = -6.0$, $AB = 1.0$, $AC = 1.0$, $BC = 9.0 (!)$, $ABC = 0.0$.

5.2 Hodnocení vlivu bloků

Jestliže se experiment dělí na bloky, je nutné stanovit, zda rozdíly ve výsledcích pokusů v jednotlivých blocích jsou významné. K tomuto účelu se z reziduálního součtu čtverců S_R vyčlení část, připadající na bloky a provede se ANOVA.

Pokud rozdělení experimentu do bloků ovlivňuje výsledky pokusů, zavádí se bloková proměnná x_B , takže model má potom tvar

$$Y = f(\mathbf{x}) + d \cdot x_B + e$$

U dvou bloků se koeficient blokové proměnné d vypočítá takto: kódované hodnoty blokové proměnné x_B se označí v prvním bloku -1 a ve druhém bloku +1. Koeficient blokové proměnné bude

$$d = \frac{\sum_{j=1}^n x_{Bj} Y_j}{\sum_{j=1}^n x_{Bj}^2}$$

n = počet všech pokusů

Příklad 3 (hodnocení vlivu bloků)

Úplný plán pro tři faktory je doplněn o dva centrální body a hvězdicové body s volbou $\alpha = \pm 2$. Pokusy jsou rozděleny do dvou bloků. Prvních 8 pokusů tvoří první blok (-1), druhých 8 pokusů = druhý blok (+1).

X1	X2	X3	Y	Blok
-1	-1	-1	25,74	-1
1	-1	-1	48,98	-1
-1	1	-1	42,78	-1
1	1	-1	35,94	-1
-1	-1	1	41,50	-1
1	-1	1	50,10	-1
-1	1	1	46,06	-1
1	1	1	27,70	-1
0	0	0	57,52	1
0	0	0	59,68	1
-2	0	0	35,50	1
2	0	0	44,18	1
0	-2	0	38,58	1
0	2	0	28,46	1
0	0	-2	33,50	1
0	0	2	42,02	1

Mají se vypočítat koeficient blokové proměnné a provést ANOVA k vyhodnocení vlivu bloků. Uvažovaný model je úplný kvadratický.

Řešení:

Regresní koeficient blokové proměnné je

$$d = \frac{-25,74 - \dots + 42,02}{(-1)^2 + \dots + (-1)^2} = \frac{20,64}{16} = 1,29$$

Základní ANOVA

	SS	Df	rozptyl	F	p-value
Model	$SS_M = 1451,62$	$k-1 = 9$	161,29	23,29	0,0005
Rezidua	$SS_R = 41,6$	$n-k = 6$	6,93		
Součet	$SS_T = 1493,17$	15			

k = počet parametrů: 1 absolutní člen + 3 lineární + 3 smíšené + 3 ryze kvadratické = 10
 n = počet bodů = 16

Výpočet koeficientů (Statgraphics)

parametr	odhad	s	T	p-value
C	41,14	2,7	15,1	0,00
Blok	1,29	2,7	0,47	0,64
X1	1,5	2,7	0,55	0,59
X2	-2,13	2,7	-0,78	0,45
X3	1,81	2,7	0,66	0,52

U dvou bloků se součet čtverců odchylek, připadající na bloky, určí

$$SS(blocks) = n \cdot d^2$$

$$SS = 16 \cdot 1,29^2 = 26,6256$$

Obecně pro m bloků je

$$SS(blocks) = \sum_{i=1}^m \frac{B_i^2}{n_i} - \frac{G^2}{n} = \frac{318,8^2}{8} + \frac{339,44^2}{8} - \frac{658,24}{16} = 26,6256$$

m = počet bloků

B_i = součet výsledků pokusů (Y) v i – tém bloku

n_i = počet pokusů v i – tém bloku

G součet výsledků všech pokusů (Y)

n = celkový počet pokusů

Reziduální součet čtverců, „očištěný“ o vliv bloků, se vypočítá

$$s_R = \sum e_i^2 - SS(blocks)$$

Při vyhodnocování SS_B (blok) testem ANOVA se použije stupně volnosti $m - 1$.

Testovací kritérium

$$T = \frac{SS(blocks)/(m-1)}{S_R/[df_R - (m-1)]}$$

m = počet bloků

ANOVA pro vyhodnocení vlivu bloků

	SS	df	SS/df	T	Kritic. hodnota
bloky	26,6	$m-1 = 1$	$26,6/1 = 26,6$	$26,6/3 = 8,9$	$F_{1,5}(0,05) = 6,6$
rezidua bez bloků	$41,6-26,6 = 15$	$df_R-(m-1) = 6-1$	$15/5 = 3$		

T překračuje $F_{1,5}(0,05)$ což znamená, že bloky ovlivňují výsledek pokusu.



Shrnutí kapitoly 5

Této kapitole byl objasněn smysl zavádění bloků do plánu experimentu, postup při jejich tvorbě a způsob vyhodnocení vlivu bloků na výsledky pokusů.



Otázky ke kapitole 5

1. Kdy se pokusy rozdělují do bloků
2. Jak se tvoří bloky
3. Jaké generátory jste již poznali
4. Jak se zjistí, zda rozdělení do bloků ovlivnilo výsledky
5. Co následuje, jestliže rozdělení do bloků má vliv na výsledek pokusu
6. Jak se počítá koeficient blokové proměnné u dvou bloků
7. Jak se počítá součet čtverců u dvou bloků
8. Jak se počítá součet čtverců u m bloků
9. Napište testovací kritérium testu vlivu bloku; jaké má rozdělení



Úlohy k řešení 1.1.

Příklad

Pro 4 faktory (označené 1,2,3,4) vytvořte úplný plán a rozdělte ho do 8 bloků.



KLÍČ K ŘEŠENÍ KAP. 5

Řešení:

Generátory podle tab.1 jsou $B_1 = 12$, $B_2 = 23$, $B_3 = 34$.

Tyto generátory mohou vytvořit tyto varianty znamének:

			blok
-1	-1	-1	I
1	-1	-1	II
-1	1	-1	III
1	1	-1	IV
-1	-1	1	V
1	-1	1	VI
-1	1	1	VII
1	1	1	VIII

Nyní rozdělíme pokusy do bloků:

1	2	3	4	12	23	34	blok
-1	-1	-1	-1	1	1	1	VIII
1	-1	-1	-1	-1	1	1	VII
-1	1	-1	-1	-1	-1	1	V
1	1	-1	-1	1	-1	1	VI
-1	-1	1	-1	1	-1	-1	II
1	-1	1	-1	-1	-1	-1	I
-1	1	1	-1	-1	1	-1	III
1	1	1	-1	1	1	-1	IV
-1	-1	-1	1	1	1	-1	IV
1	-1	-1	1	-1	1	-1	III
-1	1	-1	1	-1	-1	-1	I
1	1	-1	1	1	-1	-1	II
-1	-1	1	1	1	-1	1	VI
1	-1	1	1	-1	-1	1	V
-1	1	1	1	-1	1	1	VII
1	1	1	1	1	1	1	VIII

6. ANOVA V DOE



Čas ke studiu: 4 hodiny



Cíl: Ukázat způsob hodnocení kvality modelu na základě analýzy rozptylu

S pojmem analýza rozptylu (ANOVA) se setkáme ještě v desáté kapitole v poněkud jiné souvislosti.



Výklad

V plánování experimentů se matice X sestavuje nejen s ohledem na nalezení nejlepšího modelu, ale také tak, aby bylo možné provést co nejlepší analýzu nalezené regresní funkce. Proto se buď plánovitě provádí opakované pokusy, nebo jsou do modelu zařazeny centrální body. Pak je možné provést ANOVA podrobněji, než je běžné. Počítá se součet čtverců odchylek pro

a) **Model:** S_M

Tento součet se rozkládá, při použití úplného kvadratického modelu, na složku lineární,

ryze kvadratickou a smíšenou: $S_M = S_{Li} + S_Q + S_S$.

(pro tento rozklad je třeba doplnit plán o hvězdicové body);

b) **Rezidua:** S_R

Opakování pokusů resp. centrální body umožňují další rozklad S_R :

$S_R = S_P + S_L$, kde S_P je tzv. čistá chyba a S_F je chyba modelu;

tento rozklad vyžaduje opakované resp. zdvojené pokusy nebo centrální body;

c) **Bloky:** S_B

Je-li plán rozdělen na bloky, počítá se, jaká část S_R připadá na bloky (kap.5).

6.1 Rozklad reziduálního rozptylu

Reziduální odchylky jsou rozdílem empirické a teoretické hodnoty $e_i = Y_i - \hat{Y}_i$.

Příčinou neshody Y_i a $\hat{Y}_i$ je nepřesnost měření a nedokonalost modelu.

Při opakování pokusů bude výraz pro výpočet e_i doplněn o index s číslem opakování.

Označme Y_{ij} j -té pozorování pro bod $\mathbf{x}_i$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n_i$

m = počet úrovní (všech různých tj. těch, které se neopakují); n_i = počet opakování $\mathbf{x}_i$.

Reziduální odchylka je $Y_{ij} - \hat{Y}_i = (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \hat{Y}_i)$

Umocněním a po úpravě máme

$$\sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^m n_i (\bar{Y}_i - \hat{Y}_i)^2$$

První člen na pravé straně označíme S_P (*pure error*): představuje chybu měření; druhý člen S_L (*lack of fit*) je mírou neshody empirického modelu s teoretickým.

V další kapitole ukážeme, jak rozklad $S_R = \sum_i e_i^2$ slouží k testování modelu.

6.2 Test Lack-of-fit

Test je založen na rozkladu $S_R = S_P + S_L$ a posouzení, je-li významný rozdíl mezi S_L a S_P . Pokud je rozdíl významný, považujeme model za nedostatečný.

Test lack of fit se může provést

- a) pro opakované pokusy
- b) pro centrální body

a to pro jednu proměnnou (x) i obecně pro k proměnných ($x_1, \dots, x_k$). To pak vyžaduje sledovat Y pro tytéž volby regresorů ($x_1, \dots, x_k$), neboli stejně řádky v matici $\mathbf{X}$.

Hodnocení významnosti S_L

$H_0: S_L = S_P, \quad H_1: S_L \neq S_P$

Shoda S_P a S_L se hodnotí F testem s testovacím kritériem

$$F = \frac{\frac{S_L}{df_L}}{\frac{S_P}{df_P}} = \frac{s_L^2}{s_P^2}$$

Kritická hodnota pro hladinu významnosti 0,05: $F_{m,n}(0,05)$, kde $m = df_L$, $n = df_P$.

Ho se zamítá při překročení kritické hodnoty.

Výpočet stupňů volnosti df_R a df_L

$$\begin{aligned} df_R &= n-p \\ df_P &= \sum_{i=1}^m (n_i - 1) = n - m = n_C - 1 \\ df_R &= df_L + df_P \end{aligned}$$

$$df_L = n-p-(n-m) = m-p$$

n_i = počet opakování i -tého pokusu

p = počet parametrů v regresní funkci

n_c = počet centrálních bodů
 m = počet *různých* pokusů
 n = počet bodů

Výpočet S_L a S_P ukážeme v následujících příkladech pro různé situace:

- a) pro opakované pokusy
- b) pro plán s centrálními body

Příklad 1(test lack-of-fit pro *plán s opakováním*)

K bodům (x,Y) se má najít regresní přímka a provést ANOVA, rozložit S_R na S_P a S_L a provést test lack-of-fit.

x	Y
1,3	2,3
1,3	1,8
2,0	2,8
2,0	1,5
2,7	2,2
3,3	3,8
3,3	1,8
3,7	3,7
3,7	1,7
4,0	2,8
4,0	2,8
4,0	2,2
4,7	3,2
4,7	1,9
5,0	1,8
5,3	3,5
5,3	2,8
5,3	2,1
5,7	3,4
6,0	3,2
6,0	3,0
6,3	3,0
6,7	5,9

Řešení:

Regresní přímka má rovnici $\hat{Y} = 1,426 + 0,316x$.

ANOVA, která je součástí regresní analýzy, je v tab.1.

	SS	df	Rozptyl	Testovací kriterium	Kritická hodnota
S_M	5,499	$p-1 = 2 - 1$	$S_M/1=5,499$	$T = 5,499/0,728$ $T = 7,56$	$F_{1,21}(0,05) = 4,325$
S_R	15,278	$n-p = 23-2$	$S_R/21=0,728$		
S_T	20,777				

Tab.1 ANOVA

Přehled vzorců pro výpočet součtu čtverců odchylek (z nich některé jsou součástí ANOVA v tabulce):

$$1) S_T = \sum (Y_i - \bar{Y})^2$$

$$2) S_R = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$$

$$3) S_M = S_T - S_R$$

$$4) S_P = \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2 \quad (\text{viz tab.2})$$

$$5) S_L = S_R - S_P$$

Výpočet S_P :

V prvním sloupci jsou hodnoty x , ve kterých se prováděl *opakovaně* pokus.

Opakované x_j	Y_{ju}	$\bar{Y}_j$	$\sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2$	$df = n_j - 1$
1,3	2.3, 1.8	2,05	0,125	1
2,0	2.8, 1.5	2,15	0,845	1
3,3	3.8, 1.8	2,80	2,000	1
3,7	3.7, 1.7	2,70	2,000	1
4,0	2.8, 2.8, 2.2	2,60	0,240	2
4,7	3.2, 1.9	2,55	0,845	1
5,3	3.5, 2.8, 2.1	2,80	0,980	2
6,0	3.2, 3	3,10	0,020	1
suma			$S_P = 7,055$	$df_P = 10$

Tab.2 Výpočet S_P a df_P

	SS	df	Rozptyl	Testovací kritérium F
S_M	5,499	1	5,499	F = 0,748/0,706 = 1,06
S_R	15,278	$df_R = 23 - 2 = 21$	0,728	
S_L	$S_L = S_R - S_P = 8,223$	$df_L = 21 - 10 = 11$	$8,223/11 = 0,748$	
S_P	7,055	$df_P = 10$	$7,055/10 = 0,706$	
S_T	20,777			

Tab.3 ANOVA s rozkladem S_R

$$df_R = n - p = 23 - 2 = 21$$

$$df_P = \sum_{j=1}^m (n_j - 1) = 10 \quad (\text{tab.2})$$

$$df_L = df_R - df_P = 21 - 10 = 11$$

Kritická hodnota: $F_{11,10}(0,05) = \mathbf{2,94}$

Protože $F = 1,061 < F_{11,10}(0,05) = \mathbf{2,94}$, není položka S_L významná; proto je model vyhovující (dobře shodný s teoretickým).

Pokud by S_L bylo významné, je nalezený model nedostatečně shodný s teoretickým. Pak je potřeba pokusit se zjistit, kde a čím je tato neshoda způsobena. Příčinou může být autokorelace nebo heteroskedasticita. Ta se zjišťuje pomocí analýzy reziduí.

Příklad 2 (test lack-of-fit pro *plán s opakováním*)Vypočítat S_P a S_L , je-li $S_R = 250,134$.

Data:

x	Y
1,0	10,84
1,0	9,30
2,0	16,35
3,3	22,88
3,3	24,35
4,0	24,56
4,0	25,86
4,0	29,16
4,7	24,59
5,0	22,25
5,6	25,90
5,6	27,20
5,6	25,61
6,0	25,45
6,0	26,56
6,5	21,03
6,9	21,46

*Řešení:*Výpočet S_P

X	$(Y_{ij} - \bar{Y}_j)^2$	d.f.
1	1,1858	1
3,3	1,0805	1
4	11,2467	2
5,6	1,4341	2
6	0,6161	1
suma	$S_P = 15,5632$	$df_P = 7$

$$S_L = S_R - S_P = 250,134 - 15,5632 = 234,5708$$

Příklad 3 (test lack-of-fit pro plán s *centrálními body*)Najít S_R pro úplný kvadratický model a provést test lack-of-fit.

x_1	x_2	Y
-1	-1	43
1	-1	78
-1	1	69
+1	1	73
-1,44	0	48
1,44	0	76
0	-1,44	65
0	1,44	74
0	0	76
0	0	79
0	0	83
0	0	81

Řešení:

$$\hat{Y} = 79,7 + 9,7x_1 + 4,1x_2 - 8,7x_1^2 - 5,1x_2^2 - 7,8x_1x_2$$

$$S_R = 36,5; S_M = 1732,46; \bar{Y}_C = 79,75$$

$$S_P = (76-79,75)^2 + (79-79,75)^2 + (83-79,75)^2 + (81-79,75)^2 = 26,75 \sim 26,8$$

$$S_L = S_R - S_P = 36,5 - 26,8 = 9,7$$

		df	rozptyl	T	
S_R	36,5	$df_R = n-p = 12-6 = 6$			$F_{3,3}(0,05) = 9,27$
S_L	9,7	$df_L = df_R - df_P = 3$	$9,7/3 = 3,2$	$3,2/8,9 = 0,36$	
S_P	26,8	$df_P = n_c - 1 = 3$	$26,8/3 = 8,9$		

$T < 9,27$; ponecháme H_0 .

$n = 12$ počet bodů

$p = 6$ počet parametrů regresní funkce

$n_c = 4$ počet centrálních bodů

Příklad 4 (test lack-of-fit pro plán s *centrálními body*)

Pro plán 2^{6-2} se 4 centrálními body, generátory $E = ABC$, $F = BCD$; provést test lack-of-fit.

Pro plán experimentu

A	B	C	D	E =	F =	Y
				ABC	BCD	
-1	-1	-1	-1	-1	-1	6
1	-1	-1	-1	1	-1	10
-1	1	-1	-1	1	1	32
1	1	-1	-1	-1	1	60
-1	-1	1	-1	1	1	4
1	-1	1	-1	-1	1	15
-1	1	1	-1	-1	-1	26
1	1	1	-1	1	-1	60
-1	-1	-1	1	-1	1	8
1	-1	-1	1	1	1	12
-1	1	-1	1	1	-1	34
1	1	-1	1	-1	-1	60
-1	-1	1	1	1	-1	16
1	-1	1	1	-1	-1	5
-1	1	1	1	-1	1	37
1	1	1	1	1	1	52
0	0	0	0	0	0	29
0	0	0	0	0	0	34
0	0	0	0	0	0	26
0	0	0	0	0	0	30

byla nalezena lineární regresní funkce a provedena ANOVA:

ANOVA				Test.krit.	
	<i>df</i>	<i>SS</i>	<i>SS/df</i>	<i>F</i>	<i>p-val</i>
S_M	$p-1 = 6$	5858,375	976,3958	14,88365	3,79E-05
S_R	$n-p = 13$	852,825	65,60192		
Celkem	19	6711,2			

Řešení:

$$p = 7; n = 20$$

$$\bar{Y}_C = 29,75$$

$$S_P = 32,75$$

$$S_L = S_R - S_P = 852,825 - 32,75 = 820,075$$

	S(Součet čtverců)	df	S/df	T	Kritická hod.
S _R	852,825	df _R = n-p = 20-7 = 13			F _{10,3} (0,05) = 8,78
S _L	820,075	df _L = df _R -df _P = 13 – 3 = 10	820,075/10=82,0075	7,51	
S _P	32,75	df _P = n.-1= 3	32,75/3=10,92		

$T < F_{10,3}(0,05)$, S_L není významně větší než S_P , model se ponechává (H_0 se nezamítá).



Shrnutí kapitoly 6

V této kapitole jsme ukázali možnosti testování vhodnosti modelu na základě analýzy rozptylu. Ta byla provedena pro dvě základní situace, vyskytující se v praxi.



Otázky ke kapitole 6

1. Proveďte základní rozklad rozptylu S_T při ANOVA
2. Rozepište S_R
3. Napište všechny body testu lack-of-fit
4. K čemu slouží test lack-of-fit
5. Jak se počítají stupně volnosti df_L a df_P
6. Uveďte dvě situace pro výpočet df_P



Úlohy ke kapitole 6

Příklad 1

Vypočítejte S_p pomocí centrálních bodů pro následující plán experimentu.

Pokus	1	2	3	Y
1	-	-	-	140
2	+	-	-	150
3	-	+	-	170
4	+	+	-	160
5	-	-	+	120
6	+	-	+	120
7	-	+	+	140
8	+	+	+	180
9	0	0	0	150
10	0	0	0	140
11	0	0	0	160
12	0	0	0	150

$$S_R = 1116,67; S_P = 200,0; S_L = 916,67$$

Příklad 2

Proveďte test křivosti v příkladě 2



KLÍČ K ŘEŠENÍ KAP. 6

Řešení(1): [$S_p = 200$]

7. ZPŮSOBY NAVRHOVÁNÍ ROBUSTNÍCH VÝROBKŮ



Čas ke studiu: 4 hodiny



Cíl: Vysvětlí různé přístupy k navrhování robustních výrobků



Výklad

Cílem plánování experimentů je nejen nalézt faktory, které nejvýznamněji ovlivňují sledovaný ukazatel kvality. To by bylo pro výrobce málo. Neméně důležité je zjistit, jakou hodnotu mají vybrané faktory mít, aby ukazatel kvality Y byl optimální. Optimalita se chápe jako dosažení maxima Y (například výkon motoru), minima Y (spotřeba motoru) resp. určité cílové hodnoty. Optimální hodnotu ukazatele Y by měl produkt mít i při různých podmínkách užívání, při různém zacházení a dalších vnějších vlivech, které výrobce nemůže ovlivnit. Výrobky s takovými vlastnostmi se nazývají *robustní* a dosažení robustnosti je také jedním z cílů plánování experimentů.

V této kapitole se budeme zabývat různými způsoby zajišťování stability ukazatele kvality Y .

7.1 Minimalizace rozptylu Y

V této kapitole se budeme zabývat metodami nalezení takové úrovně významných faktorů, při kterých je Y co nejstabilnější, tedy variabilita Y minimální, současně s udržením nejlepší možné hodnoty Y .

Příklad 1 ([2], upraveno)

Sleduje se vrstva oxidu (Y) na membráně v závislosti na teplotě (A), čase (B), tlaku (C) a proudu plynu (D). Požadavek technologů je, aby vrstva oxidu byla mezi 390 a 410. Současně se žádá, aby C bylo na dolní úrovni a variabilita sledovaného ukazatele Y aby nepřesáhla hodnotu 2,0. Provedlo se 2^4 pokusů bez opakování, ale se čtyřmi měřeními u každého pokusu. Je potřeba

- najít faktory, které významně ovlivňují Y a model závislosti Y na uvedených faktorech,
- najít faktory, které významně ovlivňují rozptyl Y a model závislosti $s^2(Y)$ na uvedených faktorech,
- sestrojit vrstevnicové grafy pro a) a b).

Nejprve se vypočítají průměry a rozptyly Y v jednotlivých pokusech:

Výsledky měření v jednotlivých pokusech
(jedná se o zdvojené měření)

x_1	x_2	x_3	x_4	1.měř.	2.měř.	3.měř.	4.měř.
-1	-1	-1	-1	378	376	379	378
1	-1	-1	-1	415	416	417	416
-1	1	-1	-1	380	379	383	381
1	1	-1	-1	450	446	447	448
-1	-1	1	-1	375	371	369	372
1	-1	1	-1	391	390	391	390
-1	1	1	-1	384	385	385	385
1	1	1	-1	426	433	431	430
-1	-1	-1	1	381	381	383	380
1	-1	-1	1	416	420	412	415
-1	1	-1	1	371	372	370	371
1	1	-1	1	445	448	448	446
-1	-1	1	1	377	377	379	378
1	-1	1	1	391	391	400	392
-1	1	1	1	375	376	377	376
1	1	1	1	430	430	428	429

Tab.1 Úplný plán a opakovaná měření Y

Průměr a rozptyl opakovaných měření. Logaritmus rozptylu je počítán proto, aby se srovnaly řádové rozdíly.

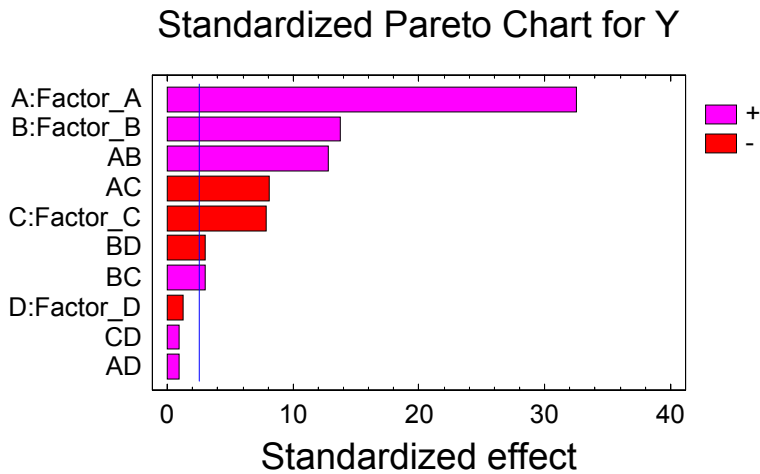
Průměr Y	$s^2(\text{rozptyly Y})$	$\ln s^2$
378	2	0,693
416	0,67	-0,4
381	3,33	1,203
448	3,33	1,203
372	6,67	1,898
390	2	0,693
385	0,67	-0,4
430	8,67	2,153
380	12	2,485
415	14,67	2,686
371	0,67	-0,4
446	6	1,792
378	1,33	0,285
392	34	3,526
376	0,67	-0,4
429	1,33	0,285

Tab.2 Zpracování výsledků opakovaných měření

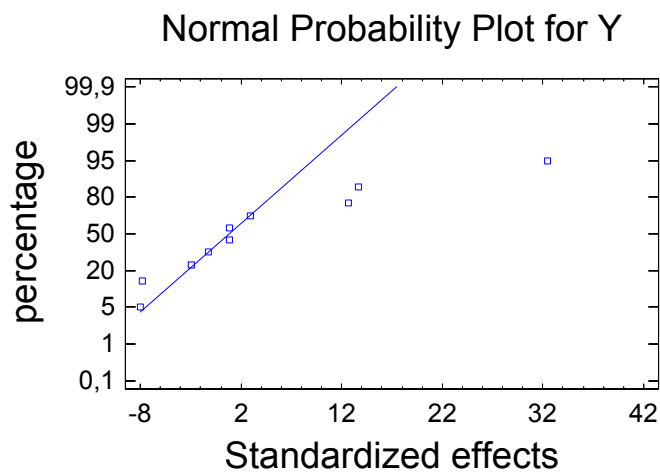
Nyní je potřeba určit významné faktory zvláště pro Y a zvláště pro rozptyl $s^2(Y)$.

Ukazatel Y

Nalezneme (dvěma způsoby) významné faktory pro Y s použitím programu Statgraphics. První použitá metoda je Paterova analýza, druhá normální pravděpodobnostní graf.



Obr.1 Paretova analýza k určení vlivných faktorů



Obr.2 Normální pravděpodobnostní graf k určení vlivných faktorů

Významné faktory a interakce jsou A,B,C,AB,AC.

Ve výsledcích Statgraphics je provedena ANOVA a jsou počítány koeficienty regresního modelu pro všechny faktory a interakce dvojic.

Hodnoty p-value potvrzují, že významnými faktory jsou A, B, C, AB a AC.

Analysis of Variance for Y

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	7439,06	1	7439,06	1053,32	0,0000
B:Factor_B	1314,06	1	1314,06	186,06	0,0000
C:Factor_C	430,563	1	430,563	60,96	0,0006
D:Factor_D	10,5625	1	10,5625	1,50	0,2758
AB	1139,06	1	1139,06	161,28	0,0001
AC	451,563	1	451,563	63,94	0,0005
AD	5,0625	1	5,0625	0,72	0,4358
BC	60,0625	1	60,0625	8,50	0,0332
BD	60,0625	1	60,0625	8,50	0,0332
CD	5,0625	1	5,0625	0,72	0,4358
Total error	35,3125	5	7,0625		
Total (corr.)	10950,4	15			

Tab.3 ANOVA

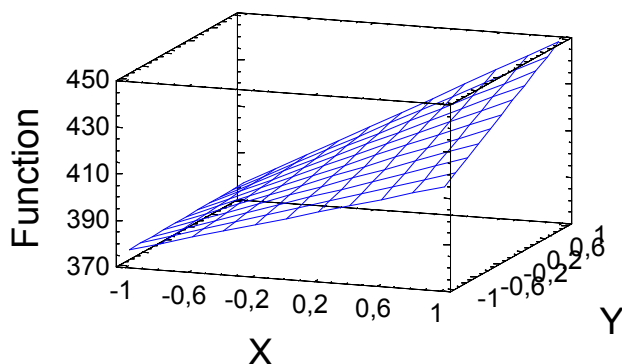
Rovnice pro konstrukci grafu závislosti Y na těchto faktorech

$$Y = 399,19 + 21,56A + 9,06B - 5,19C + 8,44AB - 5,31AC$$

Sestrojíme graf nalezení funkce.

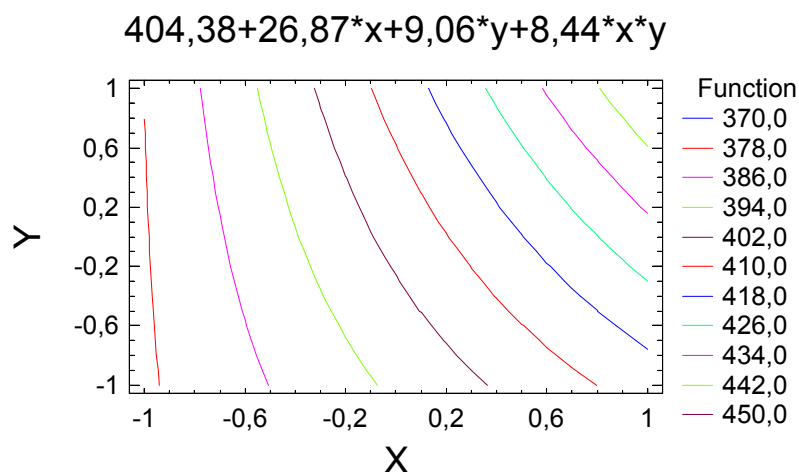
Pro konstrukci grafu mohou být jen dvě proměnné. Proto volíme v souladu s požadavkem C na dolní úrovni (-1). Pro C = -1 má rovnice a graf tvar :

$$404,38 + 26,87 \cdot x + 9,06 \cdot y + 8,44 \cdot x \cdot y$$



Obr.3 Závislost Y na dvou faktorech

Vrstevnicový graf pro určení optimální technologické oblasti vznikne, provedou-li se řezy nalezení plochy souřadnou rovinou (xOy).



Obr.4 Vrstevnicový graf k obr.3

První a poslední barva vrstevnice (stupnice vpravo) není v grafu.

Pro požadované výstupy Y mezi 390 a 410 jsou vrstevnice zelená a červená uprostřed obrázku.

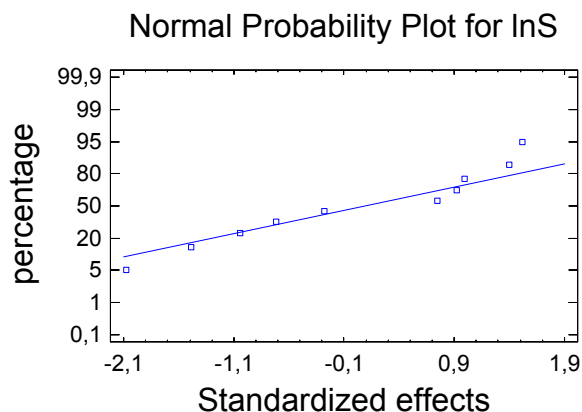
2. Variabilita

Dále zopakujeme uvedený postup pro *ukazatel variability* $\ln(s^2)$.

Uurčíme faktory významné pro variabilitu.

Rovnice modelující závislost $s^2(Y)$ na faktorech se nalezne stejně jako pro výstup Y ; protože výsledné rozptyly v jednotlivých pokusech se řádově liší, je vhodné použít logaritmus $s^2(Y)$ k dosažení větší homogenity výsledků; závisle proměnnou tedy bude $\ln s^2(Y)$.

Vlivné faktory je možné stanovit pomocí normálního pravděpodobnostního grafu nebo ANOVA.

Obr.5 Určení vlivných faktorů pro $\ln s^2(Y)$

Analysis of Variance for lns2

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	2,70109	1	2,70109	2,31	0,1891
B:Factor_B	2,58406	1	2,58406	2,21	0,1973
C:Factor_C	0,0933303	1	0,0933303	0,08	0,7889
D:Factor_D	0,646416	1	0,646416	0,55	0,4907
AB	1,14811	1	1,14811	0,98	0,3673
AC	0,987042	1	0,987042	0,84	0,4004
AD	2,29826	1	2,29826	1,96	0,2199
BC	0,59985	1	0,59985	0,51	0,5060
BD	5,04003	1	5,04003	4,31	0,0926
CD	1,27238	1	1,27238	1,09	0,3447
Total error	5,84843	5	1,16969		
Total (corr.)	23,219	15			

Tab.4 ANOVA pro faktory ovlivňující $\ln s^2(Y)$

Z normálního pravděpodobnostního grafu a také ANOVA je vidět, že žádný z faktorů neovlivňuje $\ln s^2(Y)$ významně. Největší vliv dle grafu i ANOVA má A (vpravo nahoře) a BD (vlevo dole) v grafu.

Těmto faktorům také odpovídají nejnižší hodnoty p-val v tab.4.

Sestrojíme model pro A, B, D a BD a graf plochy a vrstevnic, ve kterém bude C na dolní úrovni (-1), (neboť se žádá malá koncentrace) a D na horní úrovni (+1).

Rovnice modelu je

$$\ln s^2 = 1,08 + 0,41.A - 0,40.B + 0,20.D - 0,56.BD$$

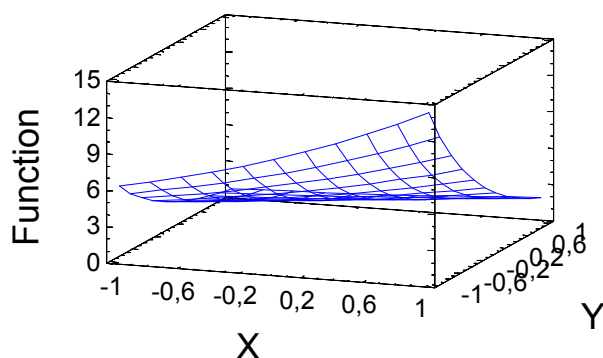
odtud

$$s^2 = \exp(1,08 + 0,41.A - 0,40.B + 0,20.D - 0,56.BD)$$

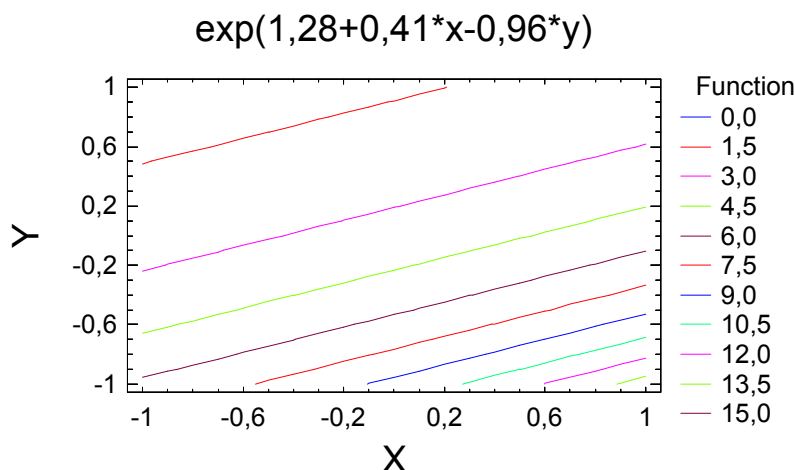
Pro D = +1 (a označení proměnných A = x, B = y pro Statgraphics) má rovnice, ke které hledáme graf, tvar

$$\ln s^2 = 1,28 + 0,41*x - 0,96*y$$

$$\exp(1,28 + 0,41*x - 0,96*y)$$

Obr.6 Graf závislosti $s^2(Y)$ na faktorech A a B

Příslušný vrstevnicový graf bude



Obr.7 Vrstevnicový graf k obr.6

Variabilita pod požadovanou hodnotu 2,0 je v oblasti nad horní fialovou vrstevnicí.

7.2 Zařazení šumů do plánu experimentu

V příkladě 1 byly počítány takové úrovně významných faktorů, při kterých je co nejstabilnější Y. Neuvažovaly se však vlivy, působící nezávisle na řídicím pracovníkovi, tzv. šumy.

Další snižování variability ukazatele kvality Y je možné zařazením šumů do plánu experimentu. Potom se navrhnu takové úrovně *kontrolovatelných* vstupních parametrů (ovlivňujících Y), při kterých je výrobek resp. proces minimálně citlivý (*robustní*) na *nekontrolovatelné* vstupy (šum, noise).

Šumy jsou sice nekontrolovatelné vlivy, působící negativně na kvalitu sledovaného ukazatele Y, lze je však minimalizovat prostřednictvím plánování experimentů vhodným navrhováním kontrolovatelných vstupů.

Jsou dvě základní možnosti zařazení šumů do plánu:

- 1) křížovým uspořádáním plánů pro faktory a šumy (crossed array) – příklad 2,
- 2) zařazení do plánu jako ostatní faktory – příklad 3.

Křížové uspořádání je uvedeno v obecné podobě v tabulce 5.

Faktory jsou značeny x_1 , x_2 a šumy z_1 , z_2 .

		z_2	-	-	+	+	$\bar{Y}$	s
x_1	x_2	z_1	-	+	-	+		
-	-							
+	-							
-	+							
+	+							

Tab.5 Křížové uspořádání faktorů a šumů

Faktory x_1, x_2 se při křížovém uspořádání pro každou dvojici (x_1, x_2) vystřídají se všemi variantami šumů (z_1, z_2) .

V takovém plánu se každý bod (x_1, x_2) vyskytuje vícekrát i bez opakování pokusů a je proto možné počítat průměr a rozptyl z Y_i . Dále se nalezne model závislosti $\bar{Y}$ a s^2 na významných faktorech, jejich grafy a nakonec vrstevnicové grafy.

V příkladě 2 je robustnost řešena pomocí křížového uspořádání faktorů a šumů.

Příklad 2 (návrh robustního výrobku, [2], upraveno)

Při výrobě změkčovače prádla, který se používá při praní, je posuzována jeho kvalita podle stupně viskozity (Y). Ta podle vyjádření odborníků v daném provozu závisí na řadě faktorů:

- a) typ vstupní suroviny (A),
- b) množství ustalovač (B),
- c) pH (C),
- d) typ přísady (D),
- e) množství přísady (E).

Vedle těchto *kontrolovatelných faktorů* je viskozita vyráběného změkčovače prádla ovlivňována také faktory, které nemůže výrobce ovlivnit, neboť závisí na uživateli (*šumy*). Zde je to například

- f) doba vyvětrávání změkčovače (M),
- g) typ vody na praní (N),
- h) teplota vody při praní (O).

Změkčovač má být vyroben tak, aby

1. Měl co nejlepší užitné vlastnosti.
Těch bude dosaženo při *minimální viskozitě*.
2. Tuto vlastnost zachoval pokud možno u každého uživatele, tedy nezávisle na šumech.
Změkčovač má tedy být *robustní výrobek*.

Faktor		-	+
A	Vstupní surovina	M1	M2
B	Množství ustalovače	málo	mnoho
C	pH	2.5	3.5
D	Typ přísady	S1	S2
E	Množství přísady	málo	mnoho

Tab.6 Úroveň kontrolovatelných faktorů.

Šum		-	+
M	Doba vyvětrávání	pod 10 dnů	nad 10 dnů
N	Typ vody	měkká	tvrdá
O	Teplota vody	studená	vlažná

Tab.7 Úroveň šumů

Plán experimentu s uvážením šumů

					M	-	+	-	+
					N	-	-	+	+
A	B	C	D=AB	E=BC	O=MN	+	-	-	+
-	-	-	+	+		3200	4500	175	1560
+	-	-	-	+		37,5	42,5	300	242,5
-	+	-	-	-		1600	475	137,5	60
+	+	-	+	-		1900	2200	302,5	3660
-	-	+	+	-		125	112,5	965	1900
+	-	+	-	-		250	325	325	1920
-	+	+	-	+		50	112,5	445	2050
+	+	+	+	+		175	97,5	492,5	340
						Y1	Y2	Y3	Y4

Tab.8 Plán experimentu pro faktory a šumy

V tabulce 4 jsou dva částečné plány: jeden pro ovlivnitelné faktory a druhý pro šumy. V následující tabulce 5 je z hodnot $Y_1, \dots, Y_4$ vypočítán průměr a ukazatel variability $\log(s)$ (logaritmus směrodatné odchylky je výhodný proto, že zmenšuje rozpětí, ve kterém se často s pohybuje):

A	B	C	D	E	$\bar{Y}$	$\log(s)$
-	-	-	+	+	2358,7	3,28
+	-	-	-	+	155,6	2,13
-	+	-	-	-	568,1	2,85
+	+	-	+	-	2015,6	3,14
-	-	+	+	-	775,6	2,92
+	-	+	-	-	705,0	2,90
-	+	+	-	+	664,4	2,97
+	+	+	+	+	276,3	2,25

Tab.9 Průměr a variabilita kontrolovatelných faktorů

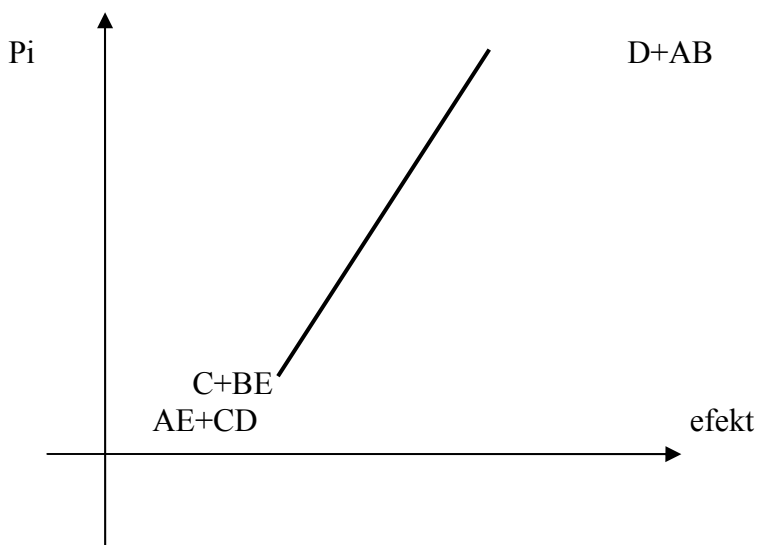
Dále je vypočítán efekt faktorů a to jak vzhledem k průměrné viskozitě $\bar{Y}$ tak vzhledem k ukazateli variability $\log(s)$.

Faktor	Efekt	
	$\bar{Y}$	$\log(s)$
A+BD	-303,6	-0,40
B+AD+CE	-117,7	-0,01
C+BE	-669,2	-0,09
D+AB	833,3	0,18
E+BC	-152,3	-0,30
AC+DE	74,2	0,03
AE+CD	-992,0	-0,53

Tab.10 Efekty faktorů

Definiční rovnice zde jsou: $I = ABD = BCE = ACDE$. Odtud se získají zaměnitelné faktory. Následuje grafické určení nejvýznamnějších faktorů a nalezení jejich optimální úrovně vzhledem k $\bar{Y}$ a $\log(s)$. To je asi nejpodstatnější rozdíl oproti dosud uvedeným částečným

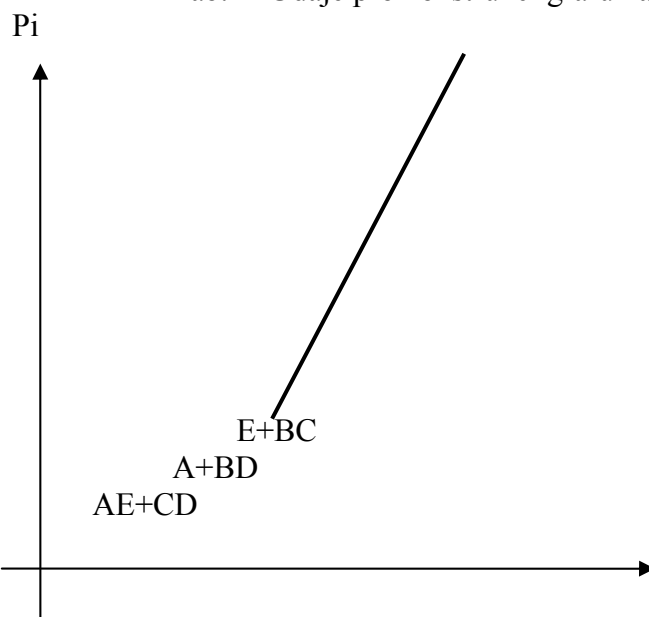
faktorovým plánům, neboť optimalizace faktorů vzhledem k ukazateli variability $\log(s)$ řeší otázku: při jakých úrovních hlavních parametrů je minimální variabilita Y , jinými slovy, za jakých podmínek je největší stabilita ukazatele Y vůči šumům.

Obr.8 Určení vlivných faktorů pro Y

Z grafického hodnocení efektu faktorů vycházejí jako významné C a D . Je tedy pro ukazatel kvality (viskozitu Y) rozhodující pH (faktor C) a typ přísady (D). Také jejich interakce se jeví jako významná a bude proto sestrojen graf, vyjadřující vliv C na Y v závislosti na D (obr. 10).

i	1	2	3	4	5	6	7
efekt	-992	-669	-303	-152	-117	74	833
faktor	AE+CD	C+BE	A+BD	E+BC	B+AD+CE	AC+DE	D+AB
Pi	7,14	21,42	35,71	50	64,28	78,57	92,85

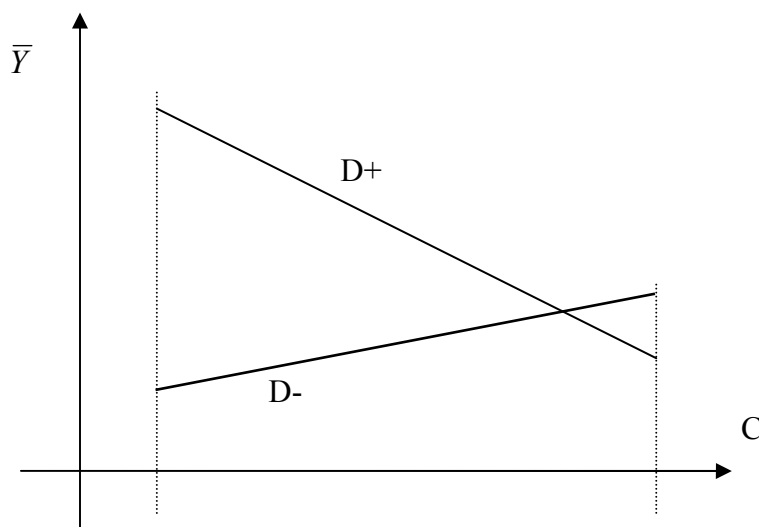
Tab.11 Údaje pro konstrukci grafu na obr.8

Obr.9 Určení vlivných faktorů pro $\log(s)$

Z grafu v obr. 9 vychází jako faktory, významně ovlivňující variabilitu vstupní surovina (A) a množství přísady (E) a také jejich interakce AE, která bude vyhodnocena v grafu, vyjadřujícím vliv vstupní suroviny A na viskozitu Y v závislosti na množství přísady E (obr. 11).

i	1	2	3	4	5	6	7
efekt	-0,53	-0,40	-0,30	-0,09	-0,01	0,03	0,18
faktor	AE+CD	A+BD	E+BC	C+BE	B+AD+CE	AC+DE	D+AB
Pi	7,14	21,42	35,71	50	64,28	78,57	92,85

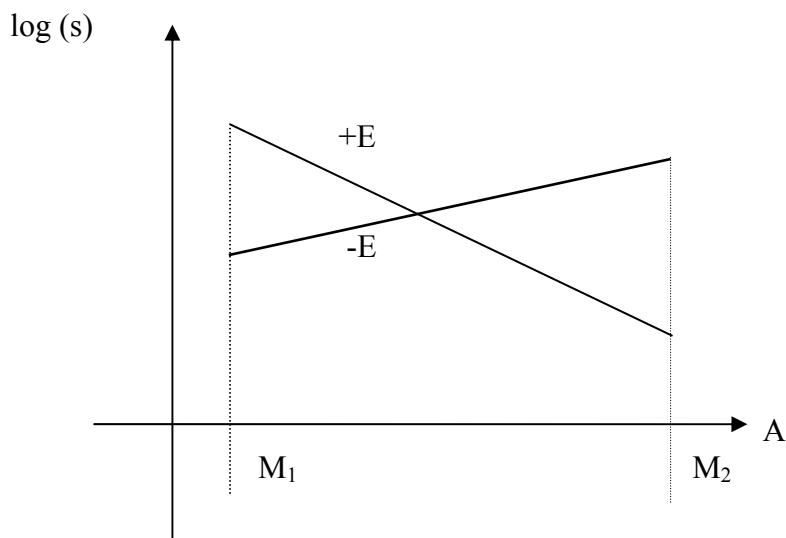
Tab.12 Údaje pro konstrukci grafu na obr. 9



Obr.10 Vliv faktoru C na $\bar{Y}$ v závislosti na D

Z obrázku 10 je zřejmé, že minimální průměrná viskozita $\bar{Y}$ bude dosažena při pH = 2,5 a přísady – D (dolní úroveň).

Je pravda, že při právě opačných podmínkách, tj. pH = 3,5 a přísadě + D není viskozita o mnoho větší. Jestliže by však během výrobního procesu došlo ke změně Ph z 3,5 na 2,5 (například v důsledku technologické nekázně), pak dojde k obrovskému nárůstu $\bar{Y}$. Naopak při režimu 2,5 a -D se $\bar{Y}$ nezmění tak výrazně, bude-li místo pH = 2,5 z nějakého důvodu náhle pH = 3,5.

Obr.11 Vliv faktoru A na $\log(s)$ v závislosti na E

Z obrázku 11 je vidět, že minimální $\log(s)$ (a tudíž maximální stability vzhledem k šumům) se dosáhne při těchto úrovních ovlivnitelných faktorů: vstupní surovina M2 a velkém množství přísad (E+).

Celkově se tedy doporučuje tento technologický režim při výrobě změkčovače prádla:

pH = 2,5 (C-)

typ přísady S1 (D-)

vstupní surovina M2 (A+)

velké množství přísad (E+)

Další možností práce se šumy je jejich zařazení do plánu jako faktory.

Příklad 3 ([3], upraveno)

Při hledání oblasti, která vyhovuje všem požadavkům technologů, se sleduje

- optimální hodnota výstupu Y,
- minimální variabilita výstupu (robustní proces),

Je potřeba najít model pro variabilitu v závislosti na faktorech

$\text{var}(\mathbf{x}, \mathbf{z})$. To lze provést například tak, že se nalezne rovnice $y(\mathbf{x}, \mathbf{z})$ a z ní se vypočítá $\text{var}[y(\mathbf{x}, \mathbf{z})]$.

Úplný plán se zařazeným šumem je v tab.13.

pokus	šum	faktory			Y
	Z1	X1	X2	X3	
1	-	-	-	-	45
2	+	-	-	-	71
3	-	+	-	-	48
4	+	+	-	-	65
5	-	-	+	-	68
6	+	-	+	-	60
7	-	+	+	-	80
8	+	+	+	-	65
9	-	-	-	+	43
10	+	-	-	+	100
11	-	+	-	+	45
12	+	+	-	+	104
13	-	-	+	+	75
14	+	-	+	+	86
15	-	+	+	+	70
16	+	+	+	+	96

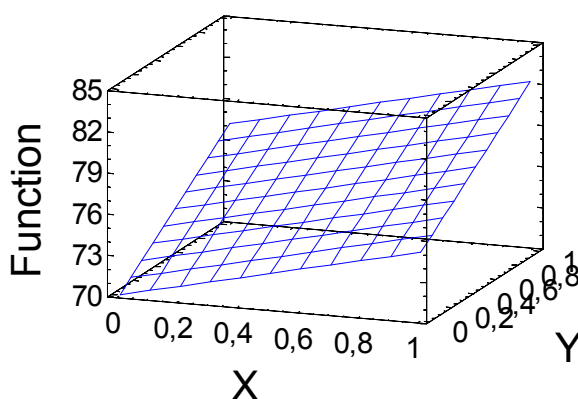
Tab.13 Úplný plán pro faktory a šumy
Koeficienty v neúplném kvadratickém modelu vypočítáme jako polovinu efektů

$$y(x, z_1) = 70,06 + \left(\frac{21,65}{2}\right)z_1 + \left(\frac{9,875}{2}\right)x_2 + \left(\frac{14,625}{2}\right)x_3 - \left(\frac{18,125}{2}\right)x_2z_1 + \left(\frac{16,625}{2}\right)x_3z_1$$

$$y(x, z_1) = 70,06 + 10,81z_1 + 4,94x_2 + 7,31x_3 - 9,06x_2z_1 + 8,31x_3z_1$$

Protože $E(z_1) = 0$, je střední hodnota $y(x, z_1)$

$$E[y(x, z_1)] = \beta_0 + \beta_2x_2 + \beta_3x_3 = 70,06 + 4,94x_2 + 7,31x_3$$



Obr.12 Závislost $y(x, z)$ na X a Y (x_2 a x_3)

Model variance $y(x, z_1)$ je

$$\text{var}[y(x, z)] = \text{var}[70,06 + 4,94x_2 + 7,31x_3 + (10,81 - 9,06x_2 + 8,31x_3)z]$$

$$\text{var}[y(x, z_1)] = \sigma_z^2 (10,81 - 9,06x_2 + 8,31x_3)^2$$

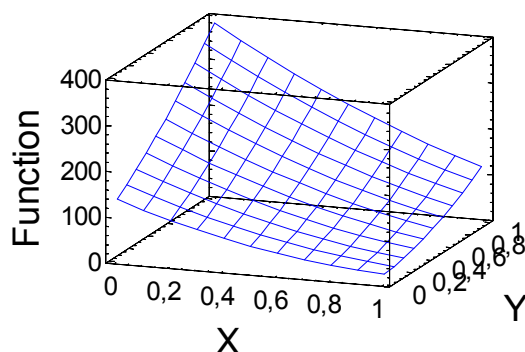
$$\text{a } \sigma_z^2 = 1$$

Model variance $y(x, z_1)$ získáme výpočtem

$$\text{var}[y(x, z_1)] = \text{var}[70,06 + 4,94x_2 + 7,31x_3 + (10,81 - 9,06x_2 + 8,31x_3)z]$$

$$\text{var}[y(x, z_1)] = \sigma_z^2 (10,81 - 9,06x_2 + 8,31x_3)^2 \text{ a } \sigma_z^2 = 1.$$

Při konstrukci grafu se Statgraphics byly opět označeny proměnné x_2 a x_3 jako X a Y. Graf je na obrázku 13.



Obr.13 Závislost variabilit $\text{var } y(x, z_1)$ na x_2 a x_3

V příkladě 3 byl zařazen jeden šum. Pro dva a více šumů může být kvadratický model, doplněný o šumy, vyjádřen v rozepsaném tvaru, nebo maticově. Dostaneme pak obecnější podobu kvadratického modelu.

Kvadratický model s proměnnými x_1, x_2 , doplněný o šumy z_1, z_2 , má rovnici

$$\begin{aligned} y(x, z) = & \beta_0 + (x_1\beta_1 + x_2\beta_2) + (\beta_{12}x_1x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2) + \\ & + (\gamma_1z_1 + \gamma_2z_2) + (\delta_{11}x_1z_1 + \delta_{12}x_2z_1 + \delta_{12}x_1z_2 + \sigma_{22}x_2z_2) \end{aligned} \quad (1)$$

Vyjádření v maticovém tvaru

$$y(x, z) = \beta_0 + x^T \beta + x^T Bx + z^T \gamma + x^T \Delta z + \varepsilon, \quad (2)$$

ve kterém je

- absolutní člen β_0
- lineární členy x : $x^T \beta = (x_1, x_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$

- smíšené a ryze kvadratické členy $\mathbf{x}$

$$\mathbf{x}^T B \mathbf{x} = (x_1, x_2) \begin{pmatrix} \beta_{11} & \beta_{12}/2 \\ \beta_{12}/2 & \beta_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

- lineární členy pro šumy $\mathbf{z} : \mathbf{z}^T \boldsymbol{\gamma} = (z_1, z_2) \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$

- dvoufaktorové interakce faktorů - šumů:

$$\begin{aligned} \mathbf{x}^T \Delta \mathbf{z} &= (x_1, x_2) \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \\ &= \delta_{11}x_1z_1 + \delta_{21}x_2z_1 + \delta_{12}x_1z_2 + \delta_{22}x_2z_2 \end{aligned}$$

Předpokládá se, že $E(\mathbf{z}) = 0$ a $\text{var}(\mathbf{z}) = \sigma_z^2 \mathbf{I}$:

Výpočet variance $y(\mathbf{x}, \mathbf{z})$ (rovnice 2) podle $\mathbf{z}$:

$$\begin{aligned} \text{var}_z[y(\mathbf{x}, \mathbf{z})] &= 0 + \text{var}_z(\mathbf{z}^T \boldsymbol{\gamma} + \mathbf{x}^T \Delta \mathbf{z}) = \text{var}_z(\boldsymbol{\gamma}^T \mathbf{z} + \mathbf{x}^T \Delta \mathbf{z}) = \text{var}_z(\boldsymbol{\gamma}^T + \mathbf{x}^T \Delta) \mathbf{z} = \\ &= (\boldsymbol{\gamma}^T + \mathbf{x}^T \Delta) \sigma_z^2 (\boldsymbol{\gamma}^T + \mathbf{x}^T \Delta)^T = (\boldsymbol{\gamma}^T + \mathbf{x}^T \Delta) \sigma_z^2 (\boldsymbol{\gamma} + \Delta^T \mathbf{x}) = l^T(\mathbf{x}) l(\mathbf{x}) \sigma_z^2 \end{aligned}$$

$$\text{var}_z[y(\mathbf{x}, \mathbf{z})] = l^T(\mathbf{x}) l(\mathbf{x}) \quad (3)$$

kde

$$l(\mathbf{x}) = (\boldsymbol{\gamma} + \Delta^T \mathbf{x}), \quad (4)$$

$l(\mathbf{x})$ je sloupcový vektor *parciálních derivací* $y(\mathbf{x}, \mathbf{z})$ podle vektoru $\mathbf{z} = (z_1, z_2)$:

$$l(\mathbf{x}) = \begin{pmatrix} \frac{\partial y}{\partial z_1} \\ \frac{\partial y}{\partial z_2} \end{pmatrix} = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix}$$

Ověříme rovnost (3).

Vypočítáme-li $l(\mathbf{x})$ jako derivaci (1) podle $\mathbf{z}$, máme

$$\frac{\partial y}{\partial z_1} = \gamma_1 + \delta_{11}x_1 + \delta_{12}x_2$$

$$\frac{\partial y}{\partial z_2} = \gamma_2 + \delta_{12}x_1 + \delta_{22}x_2$$

takže

$$l(\mathbf{x}) = \begin{pmatrix} \gamma_1 + \delta_{11}x_1 + \delta_{12}x_2 \\ \gamma_2 + \delta_{12}x_1 + \delta_{22}x_2 \end{pmatrix}$$

Rozepíšeme-li (4), dostáváme stejné $l(x)$:

$$l(x) = (\gamma + \Delta^T x) = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{12} & \delta_{22} \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \gamma_1 + \delta_{11}x_1 + \delta_{12}x_2 \\ \gamma_2 + \delta_{12}x_1 + \delta_{22}x_2 \end{pmatrix}$$

$$z^T = (z_1, z_2); \text{ (platí: } z^T \gamma = \gamma^T z; \sigma_z^2 = 1 \text{)}.$$

Tento výsledek naznačuje další možnost výpočtu $\text{var } y_z(x, z)$: pomocí funkcí $l(x)$.

Příklad 4 ([2], upraveno)

Předpokládá se, že y je ovlivňováno dvěma regulovatelnými faktory x_1, x_2 a třemi šumy z_1, z_2, z_3 , které jsou zařazeny do plánu 2^{5-1} jako faktory.

Úplný kvadratický model vychází ve tvaru

$$y(x, z) = 30,37 - 2,92x_1 - 4,13x_2 + 2,60x_1^2 + 2,18x_2^2 + 2,87x_1x_2 + 2,73z_1 - 2,33z_2 + 2,33z_3 - 0,27x_1z_1 + 0,89x_1z_2 + 2,58x_1z_3 + 2,01x_2z_1 - 1,43x_2z_2 + 1,56x_2z_3$$

Model střední hodnoty $y(x, z)$ je

$$E[y(x, z)] = 30,37 - 2,92x_1 - 4,13x_2 + 2,60x_1^2 + 2,18x_2^2 + 2,87x_1x_2$$

Model variance $y(x, z)$ je

$$\text{var}[y(x, z)] = l^T(x) l(x) = [l_1, l_2, l_3] \begin{bmatrix} l_1 \\ l_2 \\ l_3 \end{bmatrix} = l_1^2 + l_2^2 + l_3^2$$

$$\frac{\partial y}{\partial z_1} = 2,73 - 0,27x_1 + 2,01x_2 = l_1$$

$$\frac{\partial y}{\partial z_2} = -2,33 + 0,89x_1 - 1,43x_2 = l_2$$

$$\frac{\partial y}{\partial z_3} = 2,33 + 2,58x_1 + 1,56x_2 = l_3$$

$$l_1^2 = 7,45 - 1,46x_1 + 10,98x_2 + 0,073x_1^2 + 4,04x_2^2 - 1,08x_1x_2$$

$$l_2^2 = 5,42 - 4,14x_1 + 6,66x_2 + 0,79x_1^2 + 2,04x_2^2 - 2,54x_1x_2$$

$$l_3^2 = 5,42 + 12,02x_1 + 7,26x_2 + 6,56x_1^2 + 2,43x_2^2 + 8,04x_1x_2$$

$$\text{var}[y(x, z)] = l(x)^T l(x) = 18,29 + 6,42x_1 + 24,9x_2 + 7,42x_1^2 + 5,51x_2^2 + 4,42x_1x_2$$



Shrnutí kapitoly 7

V této kapitole byly uvedeny dva zásadní způsoby navrhování výrobků, odolných vůči tzv. šumům, tedy robustních výrobků.



Otázky ke kapitole 7

1. Uveďte dvě vlastnosti, které musí mít vstupní parametry procesu
2. Uveďte dvě metody k určování významnosti faktorů
3. Co nahrazuje opakování pokusů
4. Jak se vyrovnávají řádové rozdíly v datech
5. Jak se nalezne vrstevnicový graf
6. K čemu slouží vrstevnicový graf
7. Ke kterým modelům se hledá vrstevnicový graf
8. Co jsou šumy
9. Co je to robustní výrobek
10. Uveďte dvě metody zařazení šumů do plánu experimentu
11. Jak lze získat $\text{var}[y(x, z)]$ bez opakování pokusů
12. Napište kvadratický model doplněný o šumy v maticovém tvaru
13. Rozepište matice v otázce 12 (pro dva faktory a dva šumy)
14. Napište vzorec pro výpočet $\text{var}[y(x, z)]$ pomocí parciálních derivací $y(\mathbf{x}, \mathbf{z})$

8. DYNAMICKÉ PLÁNOVÁNÍ EXPERIMENTŮ



Čas ke studiu: 6 hodin



Cíl: Ukázat způsob nalezení absolutně nejlepšího nastavení vstupních parametrů na principu dynamického plánování experimentů



Výklad

Originální název této metody plánování experimentů je Steepest Ascent (SA) (metoda nejstrmějšího výstupu).

Při plánování experimentů tak, jak bylo dosud popsáno, se sestavil plán experimentu určitého typu, vypočítal vliv jednotlivých faktorů a našel model. Ten umožní nalézt optimální hodnoty vlivných faktorů. Jedná se však pouze o *lokální optimum*. Pro nalezení *absolutního optima* faktorů je potřeba pokračovat v konstrukci dalších plánů a to na ploše, jejíž rovnice je vyjádřena nalezeným lineárním modelem.

Hlavní etapy metody dynamického plánování jsou:

1. Sestavit první plán pro lineární model, najít první rovnici lineárního modelu a z ní určit směr posuvu prvního plánu.
2. V daném směru určit body „průzkumných“ pokusů a nalezení středu druhého plánu.
3. Sestavení druhého plánu pro lineární model, nalezení rovnice druhého lineárního modelu, určení nového směru posuvu plánu, provedení průzkumných pokusů v novém směru, atd. Postup se opakuje, dokud přináší zlepšení výstupu Y.
4. Poslední plán se doplní o hvězdicové body pro nalezení úplného kvadratického modelu.
5. Vypočítá se optimální nastavení faktorů.

Uvedený postup ukážeme v příkladě 1.

Příklad 1 (data podle [1], upraveno)

Sledovaný výstup Y závisí na dvou faktorech: x_1 a x_2 . Úplný plán, doplněný třemi centrálními body, je v tab.1.

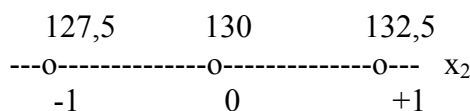
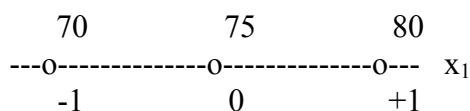
	Provozní hodnoty		Kódované hodnoty		Y
	x_1	x_2	ξ_1	ξ_2	
1	70	127,5	-1	-1	54,3
2	80	127,5	1	-1	60,3
3	70	132,5	-1	1	64,6
4	80	132,5	1	1	68,0
5	75	130,0	0	0	60,3
6	75	130,0	0	0	64,3
7	75	130,0	0	0	62,3

Tab.1 První plán metody SA

Najděte lineární model, směr postupu a metodou dynamického plánování nalezněte nový střed pro další plán. Cílem je maximalizace výstupu Y .

Řešení:

Pro potřeby kódování a dekódování hodnot faktorů je dobrou pomůckou grafické znázornění: na číselné ose jsou dolní a horní hranice a střední hodnota; nad osou jsou reálné hodnoty, pod osou v kódovém označení -1, +1 a 0:



Kódovací a dekódovací rovnice
pro x_1

$$\xi_1 = \frac{x_1 - 75}{5} \Rightarrow x_1 = 5\xi_1 + 75$$

a pro x_2

$$\xi_2 = \frac{x_2 - 130}{2,5} \Rightarrow x_2 = 2,5\xi_2 + 130.$$

Výpočet efektu faktorů:

$$ef(x_1) = \frac{1}{2}(-54,3 + 60,3 - 64,6 + 68) = 4,7$$

$$ef(x_2) = \frac{1}{2}(-54,3 - 60,3 + 64,6 + 68) = 9,0$$

Průměrná hodnoty Y v prvním plánu je (počítá se i z Y v centrálních bodech)

$$\bar{Y}_1 = \frac{1}{7}(54,3 + 60,3 + 64,6 + 68 + 60,3 + 64,3 + 62,3) = 62,01$$

Odhad rozptylu v centrálních bodech v prvním plánu:

$$s_1^2 = \frac{1}{3-1}[(60,3 - 62,3)^2 + (62,3 - 62,3)^2 + (64,3 - 62,3)^2] = 4$$

Lineární model má rovnici

$$Y = 62,01 + \frac{4,7}{2}x_1 + \frac{9}{2}x_2 = 62,01 + 2,35x_1 + 4,5x_2$$

Směr pohybu plánu je určen vektorem, jehož souřadnice jsou koeficienty faktorů x_1 a x_2 v lineárním modelu: (2,35; 4,5).

Souřadnice upravíme pro zjednodušení dalších výpočtů:

$$\left(\frac{2,35}{2,35} = 1; \frac{4,5}{2,35} = 1,91\right)$$

Násobky těchto souřadnic jsou kódované hodnoty faktorů. V těchto bodech budou provedeny průzkumné pokusy.

Násobky základního směrového vektoru, tj. souřadnice pro průzkumné pokusy s výsledkem Y_i :

$$\left(\frac{2,35}{2,35} = 1; \frac{4,5}{2,35} = 1,91\right) \dots Y_1 = 73,3$$

$$(2; 2 \cdot 1,91 = 3,82)$$

$$(3; 3 \cdot 1,91 = 5,73) \dots \mathbf{Y_3 = 86,8}$$

$$(4; 4 \cdot 1,91 = 7,64)$$

$$(5; 5 \cdot 1,91 = 9,55) \dots Y_5 = 58,2$$

Protože cílem je maximalizace Y , je novým středem bod s kódovanými hodnotami souřadnic (3;5,73).

Nové centrum v reálných hodnotách tedy je:

$$\text{Pro } x_1: 5 \cdot 3 + 75 = 90,$$

$$\text{Pro } x_2: 2 \cdot 5,5,73 + 130 = 144,3; \text{ zaokrouhlíme na } x_2 = 145.$$

Kolem nového centra stanovíme dolní a horní hranice faktorů širší než u prvního plánu. Získá se tak větší prostor k analýze.

$$\begin{array}{ccccccc} 80 & & 90 & & 100 & & \\ \text{---o---} & \text{---o---} & \text{---o---} & & x_1 & & \\ -1 & & 0 & & +1 & & \end{array}$$

$$\begin{array}{ccccccc} 140 & & 145 & & 150 & & \\ \text{---o---} & \text{---o---} & \text{---o---} & & x_2 & & \\ -1 & & 0 & & +1 & & \end{array}$$

Kódovací a dekódovací rovnice
pro x_1

$$\xi_1 = \frac{x_1 - 90}{10} \Rightarrow x_1 = 10\xi_1 + 90$$

a pro x_2

$$\xi_2 = \frac{x_2 - 145}{5} \Rightarrow x_2 = 5\xi_2 + 145.$$

Ve druhém plánu jsou dva centrální body

	Provozní hodnoty		Kódované hodnoty		Y
	x_1	x_2	ξ_1	ξ_2	
1	80	140	-1	-1	78,8
2	100	140	1	-1	84,5
3	80	150	-1	1	91,2
4	100	150	1	1	77,4
5	90	145	0	0	89,7
6	90	145	0	0	86,8

Tab.2 Druhý plán metody SA

Průměrná hodnota v centrálních bodech je $\bar{Y}_C = 88,25$

Odhad rozptylu v centrálních bodech ve druhém plánu:

$$s_2^2 = \frac{1}{2-1} [(89,7 - 88,25)^2 + (86,8 - 88,25)^2] = 4,21$$

Spojením odhadů v centrálních bodech upřesníme odhad σ_e^2 :

$$\sigma_e^2 \approx \frac{2,4 + 1,4,21}{3} = 4,07$$

Výpočet efektů ve druhém plánu:

$$ef(x_1) = \frac{1}{2} (-78,8 + 84,5 - 91,2 + 77,4) = -4,05$$

$$ef(x_2) = \frac{1}{2} (-78,8 - 84,5 + 91,2 + 77,4) = 2,65$$

Průměrná hodnota Y ve druhém plánu je $\bar{Y} = 84,73$

Model ve druhém plánu má rovnici

$$Y = 84,73 - 2,025x_1 + 1,325x_2.$$

Stejným způsobem bychom postupovali ve vytyčování nového směru dále. Postup se opakuje tak dlouho, dokud nový směr přináší zlepšení výstupu Y, resp. pokud se nedostaneme za hranice možností technologického nastavení faktorů.

Postup považujeme v tomto případě za ukončený.

Druhý plán nyní doplníme o hvězdicové body (a dva centrální body k dalšímu zlepšení odhadu σ_e^2) tak, aby bylo možné hledat úplný kvadratický model:

	Provozní hodnoty		Kódované hodnoty		Výstup
	x_1	x_2	ξ_1	ξ_2	Y
1	80	140	-1	-1	78,8
2	100	140	1	-1	84,5
3	80	150	-1	1	91,2
4	100	150	1	1	77,4
5	90	145	0	0	89,7
6	90	145	0	0	86,8
7	76*	145	-1,41	0	83,3
8	104*	145	1,41	0	81,2
9	90	138*	0	-1,41	81,2
10	90	152*	0	1,41	79,5
11	90	145	0	0	87,0
12	90	145	0	0	86,0

Tab.3 Třetí plán metody SA

*) Výpočet reálných hodnot:

Pro x_1

$$\xi_1 = -1,41: 10 \cdot x_1 + 90 = 10(-1,41) + 90 = 75,9$$

$$\xi_1 = 1,41: 10 \cdot 1,41 + 90 = 104,1$$

Pro x_2

$$\xi_2 = -1,41: 5 \cdot x_2 + 145 = 5(-1,41) + 145 = 137,95$$

$$\xi_2 = 1,41: 5 \cdot 1,41 + 145 = 152,05$$

Průměrná hodnoty Y v centrálních bodech ve třetím plánu

$$\bar{Y}_C = \frac{1}{6}(87 + 86) = 86,5$$

Odhad rozptylu v centrálních bodech σ_e^2 ze třetího plánu:

$$s_3^2 = \frac{1}{2-1}[(87 - 86,5)^2 + (86 - 86,5)^2] = 0,5$$

Spojením odhadů s_1^2, s_2^2, s_3^2 upřesníme odhad σ_e^2 :

$$\sigma_e^2 \approx \frac{2,4 + 1,4,21 + 1,0,5}{4} = 3,18$$

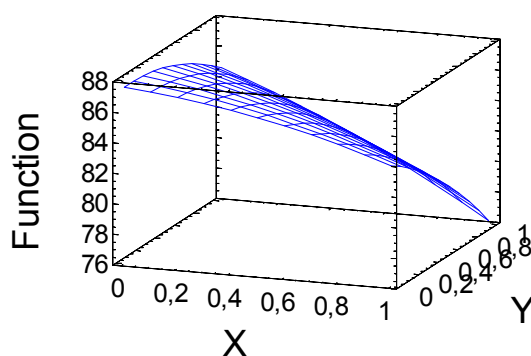
Úplný kvadratický model má rovnici

$$Y = 87,36 - 1,38x_1 + 0,36x_2 - 2,14x_1^2 - 3,10x_2^2 - 4,87x_1x_2$$

Koeficient determinace $r^2 = 0,882577$.

Graf modelu je na obr. 1.

Z důvodu konstrukce grafu se Statgraphics jsou použity proměnné x, y místo x_1 a x_2 .



Vypočítáme optimální hodnoty vstupů x_1 a x_2 .

Z koeficientů kvadratického modelu sestavíme matice b , B a vypočítáme vektor optimálních vstupů x_s :

$$b = \begin{bmatrix} -1,38 \\ 0,36 \end{bmatrix}, \quad B = \begin{bmatrix} -2,14 & -2,44 \\ -2,44 & -3,10 \end{bmatrix}, \quad x_s = \frac{-1}{2} B^{-1} b = \begin{bmatrix} -3,78 \\ 3,04 \end{bmatrix}$$

V reálných hodnotách je $x_1 = 10(-3,78) + 90 = 52,2$ a $x_2 = 5,3,04 + 145 = 160,2$

Pokud se neprovádí dynamické plánování, vychází optimální x_1 a x_2 z prvního plánu:

	X1	X2	Y
	-	-	54,3
	+	-	60,3
	-	+	64,6
	+	+	68,0
Y-	59,45	57,3	
Y+	64,15	66,3	

Při maximalizaci Y je tedy optimální volba pro horní úroveň x_1 a x_2 ; v reálných hodnotách je to $x_1 = 80$ a $x_2 = 132,5$.

Příklad 2 (data [2], upraveno)

Sledovaný znak Y je ovlivněn dvěma faktory: x_1 a x_2 . Byl sestaven úplný plán se čtyřmi centrálními body. Má se najít model, určit směr postupu a nové centrum. Průzkumné pokusy jsou dvojnásobky a trojnásobky základního plánu (tab.4). Y se maximalizuje.

X_1	X_2	ξ_1	ξ_2	Y
-	-	1,2	275	775
+	-	1,6	275	670
-	+	1,2	325	890
+	+	1,6	325	730
0	0	1,4	300	745
0	0	1,4	300	760
0	0	1,4	300	780
0	0	1,4	300	720

Tab.4 Plán experimentu

Řešení:

Model

$$\hat{Y} = 758,75 - 66,25x_1 + 43,75x_2$$

$$r^2 = 0,89$$

Směr postupu: (-66,25;43,75)

Úprava směrového vektoru:

$$\left(\frac{-66,25}{66,25} = -1; \frac{43,75}{66,25} = 0,66 \right)$$

K základnímu směru vypočítáme dvojnásobky a trojnásobky, přepočítáme je na reálné hodnoty. Pak je možné provést průzkumné pokusy a v bodě, kde je max.Y, je stanoveno nové centrum.

Kódovací a dekódovací rovnice

pro x_1

$$\xi_1 = \frac{x_1 - 1,4}{0,2}$$

a pro x_2

$$\xi_2 = \frac{x_2 - 300}{25}.$$

Přepočet kódovaných hodnot na reálné:

$$x_1 = 0,2\xi_1 + 1,4$$

a pro x_2

$$x_2 = 25\xi_2 + 300.$$

Výsledky průzkumných pokusů jsou v tabulce č.5:

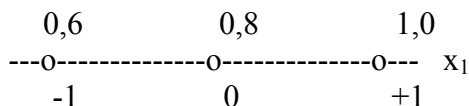
Faktory		Y
Kódované hodnoty (ξ_1, ξ_2)	Reálné hodnoty (x_1, x_2)	
(-1;0,66)	(1,2;318)	845
(-2;1,32)	(1,0;333)	950
(-3;1,98)	(0,8;349,5)	1040

Tab.5 Průzkumné pokusy

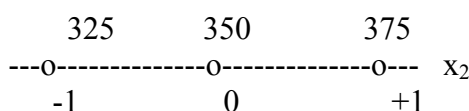
Nové centrum v reálných hodnotách tedy je kolem bodu (0,8;350)

Kolem nového centra stanovíme dolní a horní hranice faktorů:

Pro x_1 :



Pro x_2 :



Pro další postup je sestaven nový plán se dvěma centrálními body:

x_1	x_2	ξ_1	ξ_2	Y
-	-	0,6	325	
+	-	1,0	325	
-	+	0,6	375	
+	+	1,0	375	
0	0	0,8	350	
0	0	0,8	350	

Tab.6 Druhý plán experimentu v novém centru

Příklad 3 (data [3], upraveno)

Sleduje se pórovitost materiálu(Y) v závislosti na teplotě ($T = x_1$) a tlaku ($P = x_2$). Má se najít takové T a P, aby pórovitost byla minimální. Postup je rozdělen do několika kroků.

1. Byl proveden úplný plán pro dva faktory 2^2 , doplněný dvěma centrálními body T_o, P_o (jsou to průměrné hodnoty z $x_{\max}$ a $x_{\min}$), které představují dosud používané parametry $T = 650^\circ C$ a $P = 975 \text{ kg/cm}^2$.

Intervaly pro faktory: $T \in \langle 640, 660 \rangle$, $P \in \langle 950, 1000 \rangle$ stanovil technolog.

	Provozní hodnoty		Kódované hodnoty		Y
	T	P	ξ_1	ξ_2	
1	640	950	-1	-1	6,09
2	660	950	1	-1	5,53
3	640	1000	-1	1	6,78
4	660	1000	1	1	6,16
5	650	975	0	0	5,93
6	650	975	0	0	6,12

Tab.7 První plán experimentu

Centrální body (T_o, P_o) umožní dále

- a) vypočítat rozptyl Y v pokusech (s^2),
- b) testovat, zda je lineární model

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2$$

vyhovující, nebo je potřeba (úplný) kvadratický model

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

2. Odhady koeficientů $\beta_o, \beta_1, \beta_2$ lineárního modelu se vypočítají z efektu faktorů:
(b_o je průměr z Y_i , b_1, b_2 je polovina efektu – v hranaté závorce)

$$b_o = 1/4 [6.09 + 5.53 + 6.78 + 6.16] = 6.14$$

$$b_1 = \frac{1}{2} \left[\frac{1}{2} (-6.09 + 5.53 - 6.78 + 6.16) \right] \approx -0.29$$

$$b_2 = \frac{1}{2} \left[\frac{1}{2} (-6.09 - 5.53 + 6.78 + 6.16) \right] = 0.33$$

Lineární model má rovnici

$$\hat{Y} = 6.14 - 0.29x_1 + 0.33x_2$$

resp.

$$\hat{Y} = 6.14 - 0.29T + 0.33P$$

3. Test, zda lineární model je dostačující (test křivosti)

Ukazatelem zakřivení plochy $f(x_1, x_2)$ je rozdíl průměrů $\bar{y}_k - \bar{y}_c$, který může být vyhodnocen testem.

Kompletní test rozdílu $\bar{y}_k - \bar{y}_c$ bude:

a) $H_o: \bar{y}_k - \bar{y}_c = 0$,

$\bar{y}_k$ = průměrný efekt z pokusů v kvádru (6,14),

$\bar{y}_c$ = průměrný efekt z pokusů v centru (6,025).

c) Testovací kritérium

$$T = \frac{\bar{y}_k - \bar{y}_c}{s \sqrt{\frac{1}{n_k} + \frac{1}{n_c}}}$$

(1)

n_k = počet pokusů v kvádru (4),

n_c = počet pokusů v centrálním bodě (2).

Potřebné s se vypočítá pomocí s_k a s_c :

$$s_k^2 = \frac{1}{4-1} [(6.09 - 6.14)^2 + (5.53 - 6.1)^2 + (6.78 - 6.14)^2 + (6.16 - 6.14)^2]$$

$$s_k^2 = 0.2615$$

$$s_c^2 = \frac{1}{2-1} [(5.93 - 6.025)^2 + (6.12 - 6.025)^2]$$

$$s_c^2 = 0.01805$$

$$s^2 = \frac{3 \cdot 0.2615 + 1 \cdot 0.01805}{3 + 1} = 0.2006$$

$$s = 0.45$$

Dosazením do (1) dostáváme

$$T = \frac{6.14 - 6.025}{0.45 \sqrt{\frac{1}{4} + \frac{1}{2}}} = 0.29$$

c) Kritická hodnota

$$K = t_{n_k + n_c - 2}(\alpha) = t_{4+2-2}(0.05) = 4.303$$

(2)

d) Protože $T < 4.303$, je rozdíl $\bar{y}_k - \bar{y}_c$ bezvýznamný a tedy lineární model postačující.

Použité hodnoty: $n_k = 2^2 = 4$, $n_c = 2$, $\bar{y}_k = 6.14$, $\bar{y}_c = 1/2 (5.93 + 6.12) = 6.025$.

4. Určení směru průzkumu plochy:

$$\frac{\delta Y}{\delta T} = -0.29, \quad \frac{\delta Y}{\delta P} = 0.33$$

(3)

Směr maximálního růstu Y je určen vektorem $(-0.29, 0.33)$, takže směr maximálního poklesu Y je obrácený, tj. $(0.29, -0.33)$.

Kroky pro realizaci dalších pokusů v určeném směru lze stanovit různě, například

1. krok: $(0.29/0.29 = 1, -0.33/0.29 = -1.14)$
2. krok je dvojnásobek prvního: $(2, -2.28)$
3. krok: $(3, -3.42)$
4. krok: $(4, -4.56)$
5. krok: $(5, -5.70)$

Přepoččet kódovaných hodnot na reálné podle vztahů

$$T_c = \frac{T - 650}{10}, \quad P_c = \frac{P - 975}{25}$$

dostáváme $T = 650 + 10 T_c$, $P = 975 + 25 P_c$

Pro jednotlivé kroky vychází

- 1) $650 + 10 \cdot 1 = 660$, $975 + 25 \cdot (-1.14) = 946.5 \approx 950$
- 2) $650 + 10 \cdot 2 = 670$, $975 + 25 \cdot (-2.28) \approx 920$
- 3) $650 + 10 \cdot 3 = 680$, $975 + 25 \cdot (-3.42) \approx 890$
- 4) 690, 860
- 5) 700, 830

Jiná možnost stanovení kroků pro další provádění pokusů na zkoumané ploše:

vypočítá se délka vektoru $(0.29, -0.33)$, která je $\sqrt{0.29^2 + 0.33^2} = 0.44$. Potom jednotkový vektor směru postupu bude $u = (0.29/0.44, -0.33/0.44) = (0.66, -0.75)$. Postupuje se v násobcích $2u$, $4u$, $6u$, ... resp. u , $3u$, $5u$, ... tohoto vektoru. Kódované hodnoty se nakonec dekodují. V našem případě máme

Čís. pokusu	Násobek u	$T(^{\circ})$	$P(^{\circ})$	Výsl.pok.
1	$3u = (1.98, -2.25)$	670	920	4.53
2	$5u = (3.30, -3.75)$	685	880	3.28
3	$7u = (4.62, -5.25)$	700	845	2.54
4	$9u = (5.94, -6.75)$	710	805	4.15

(*) výsledky zaokrouhlovány na ± 5

Y je minimální pro $(T, P) = (700, 845)$. To bude nové centrum pro plánování experimentu 2^2 .

5. Nový plán se třemi centrálními body:

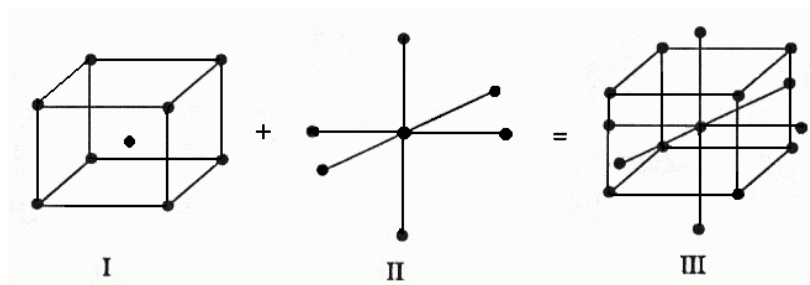
	Provozní hodnoty		Kódované hodnoty		Y
	T	P	ξ_1	ξ_2	
1	690	820	-1	-1	2,20
2	710	820	1	-1	3,71
3	690	870	-1	1	2,86
4	710	870	1	1	3,49
5	700	845	0	0	2,53
6	700	845	0	0	2,30
7	700	845	0	0	2,54

Tab.8 Plán s novým centrem

Pro další plán experimentu nalezneme kvadratický model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

6. K výpočtu koeficientů β_{11}, β_{22} je nutné doplnit plán z bodu 5) na hvězdicový plán.



Obr.2 Plán v kvádru (I) a hvězdicový plán (II)

Pro dvě proměnné (T, P) má hvězdicový plán s centrálními body toto obecné schéma:

$-\alpha$	0
α	0
0	$-\alpha$
0	α
0	0
0	0
.....	
0	0

Tab.9 Obecné schéma hvězdicového plánu

Hodnota α a počet centrálních bodů n_{oe} se vypočítá

$$\alpha = \sqrt[4]{N_c} \quad (4)$$

kde N_c = počet pokusů prvního experimentu (v kvádru resp. čtverci) bez centrálních bodů, tj. 2^k resp. 2^{k-p} . Další vzorec, který je možné použít, je

$$\alpha = \sqrt{\frac{N_c(2k + n_{oe})}{2(N_c + n_{oc})}} \quad (5)$$

k = počet faktorů,

n_{oc} = počet centrálních bodů kvádru (volí se),

n_{oe} = počet centrálních bodů hvězdicového plánu (počítá se).

Společným řešením obou vzorců se vypočítá alfa a n_{oe} . V našem příkladě je

$$\alpha = \sqrt[4]{4} = \sqrt{2},$$

$$\alpha = \sqrt{\frac{4(4+n_{oe})}{2(4+3)}} = \sqrt{2} \Rightarrow n_{oe} = 3.$$

Hvězdicové body + centrální body má tvar

	Provozní hodnoty		Kódované hodnoty		Y
	T	P	ξ_1	ξ_2	
1	685	845	-1,41	0	2,87
2	715	845	1,41	0	4,12
3	700	810	0	-1,41	4,02
4	700	880	0	1,41	4,05
5	700	845	0	0	3,30
6	700	845	0	0	3,40
7	700	845	0	0	3,17

Tab.10 Hvězdicový plán

7. Z plánů v tab. 8 a 10 (jejich spojení) se vypočítají koeficienty úplného kvadratického modelu:

$$Y = 2,8 + 0,48T + 0,06P + 0,13T^2 + 0,40P^2 - 0,22TP$$

8. Určení extrémů modelu:

a) Z parciálních derivací modelu podle T a P

$$\frac{\partial Y}{\partial T} = 0, \frac{\partial Y}{\partial P} = 0 \Rightarrow T^*, P^*$$

$$\frac{\partial Y}{\partial T} = 0,48 + 0,26T - 0,22P = 0$$

$$\frac{\partial Y}{\partial P} = 0,048 + 0,26T - 0,22P$$

Řešením této soustavy dostáváme optimální kódované hodnoty faktorů T a P:

$T^* = -2,49$ a $P^* = -0,76$.

b) Pomocí vzorce $x_s = \begin{bmatrix} T^* \\ P^* \end{bmatrix}_s = \frac{-1}{2} B^{-1} b = \frac{-1}{2} \begin{bmatrix} 0,13 & -0,11 \\ -0,11 & 0,40 \end{bmatrix}^{-1} \begin{bmatrix} 0,48 \\ 0,06 \end{bmatrix} = \begin{bmatrix} -2,49 \\ -0,76 \end{bmatrix}$



Shrnutí kapitoly 8

V této kapitole byl vysvětlen způsob nalezení nejlepšího možného nastavení významných vstupů procesu metodou Steepest Ascent, zde označenou jako dynamické plánování-nejedná se o překlad, ale jiné pojmenování, které naznačuje podstatu metody: plán experimentu je posouván prostorem za účelem vyhledávání nejlepších podmínek nastavení faktorů, ovlivňujících proces. Dosud uváděné způsoby optimalizace vstupů byly lokálními optimy.



Otázky ke kapitole 8

1. Uveďte hlavní kroky dynamického plánování experimentu
2. Co je cílem dynamického plánování
3. Znázorněte graficky a napište rovnici pro kódování pro $x_{\min} = 70$, $x_{\max} = 80$.
4. Napište rovnici pro dekódování k úloze 3
5. Napište lineární model, je-li $ef(x_1) = 4,7$ a $ef(x_2) = 9$. $\bar{Y} = 62,01$.
6. Určete směr postupu k modelu v bodě 5
7. Určete další kroky postupu k úloze 6
8. Napište dva vzorce pro výpočet α (hvězdicový bod)
9. Jak se určí směr posuvu plánu pomocí lineárního modelu a proč
10. Kolikrát se opakuje posun plánu
11. Jak se prověří potřeba kvadratického modelu

9. KVALITATIVNÍ PROMĚNNÉ



Čas ke studiu: 3 hodiny



Cíl Ukázat způsob použití kvalitativních proměnných v plánech experimentu a v modelu



Výklad

Jsou-li v regresním modelu zařazeny proměnné, které nelze měřit, tedy kvalitativní proměnné (atributy), vznikají při sestavování plánu a nalezení modelu problémy.

Obecně lze říci, že *kvalitativní* proměnné jsou modelovány pomocí spojitých *kvantitativních* proměnných. V dalším budeme řešit

- Jak se sestaví plán s kvalitativní proměnnou
- Jak se sestaví model s kvalitativní proměnnou
- Jak působí kvalitativní proměnná v modelu

Způsob modelování kvalitativní proměnné závisí na počtu těchto proměnných, na počtu úrovní a na typu modelu (lineární nebo kvadratický).

9.1 Konstrukce plánů s kvalitativními proměnnými

Příklad 1

Mějme dvě kvantitativní proměnné x_1 , x_2 a jednu kvalitativní proměnnou z ; všechny proměnné mohou nabývat dvě úrovně.

Uvažujme lineární model:

$$y = \beta_0 + \sum_{i=1}^2 \beta_i x_i + \gamma \cdot z ;$$

Modelování kvalitativní proměnné z při sestavení matice X :

$z = 0$, když je v plánu dolní úroveň a $z = 1$ pro horní úroveň.

Konstrukce plánu:

Sestaví se dva úplné plány pro x_1, x_2 a každý se doplní dolní resp. horní úrovní z :

x_1	x_2	z
-1	-1	0
1	-1	0
-1	1	0
1	1	0
-1	-1	1
1	-1	1
-1	1	1
1	1	1

Matice X:

$$X = \begin{bmatrix} 1 & -1 & -1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Jinou možností modelování kvalitativní proměnné je pomocí kódování $z = -1$, když je v plánu dolní úroveň a $z = +1$ pro horní úroveň.

$$X = \begin{bmatrix} & x_1 & x_2 & z \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Žádný z uvedených způsobů není preferován.

Příklad 2

Mějme dvě kvantitativní proměnné x_1, x_2 a jednu kvalitativní z , která může nabývat *tří úrovně* A, B a C. Má se sestavit plán pro nalezení modelu, ve kterém jsou i všechny možné smíšené členy, tedy i členy, obsahující kvantitativní a kvalitativní proměnné

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma \cdot z + \delta_1 z \cdot x_1 + \delta_2 z \cdot x_2 + \beta_{12} x_1 x_2$$

V plánu mohou být kvalitativní proměnné:

-1	-1	A
-1	-1	B
-1	-1	C
-1	1	A
-1	1	B
-1	1	C
1	-1	A
1	-1	B
1	-1	C
1	1	A
1	1	B
1	1	C
0	0	A
0	0	B
0	0	C

V plánu jsou dále tři centrální body spolu s kvalitativní úrovní A, B a C

Matice X nemůže obsahovat kvalitativní proměnné, neboť s ní budou prováděny potřebné maticové operace.

Proto kvalitativní proměnnou z modelujeme pomocí dvou kvantitativních proměnných z_1, z_2 definovaných takto:

$z_1 = 1$ pro úroveň A, jinak 0

$z_2 = 1$ pro úroveň B, jinak 0.

Odtud plyne, že pro úroveň C je: $z_1 = 0$ (není A), $z_2 = 0$ (není B), takže při této volbě zbývá úroveň C. Přehledně v tabulce:

z_1	z_2	Úroveň z
1	0	A
0	1	B
0	0	C

Matice X (a plán také) je sestavena tak, že úroveň $(-, -)$ pro x_1 a x_2 je s každou úrovní z (A, B a C), podobně úroveň $(-, +)$ s A, B a C atd.

$$X = \begin{bmatrix} -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Schématicky můžeme zapsat matici X v tabulce i s označením sloupců a řádků:

X1	X2	Z1	Z2	
-1	-1	1	0	A
-1	-1	0	1	B
-1	-1	0	0	C
-1	1	1	0	A
-1	1	0	1	B
-1	1	0	0	C
1	-1	1	0	A
1	-1	0	1	B
1	-1	0	0	C
1	1	1	0	A
1	1	0	1	B
1	1	0	0	C
0	0	1	0	A
0	0	0	1	B
0	0	0	0	C

Jinou možností je

z_1	z_2	Úroveň z
1	-1	A
-1	1	B
-1	-1	C

Matice X v tabulkovém vyjádření:

X1	X2	Z1	Z2	
-1	-1	1	-1	A
-1	-1	-1	1	B
-1	-1	-1	-1	C
-1	1	1	-1	A
-1	1	-1	1	B
-1	1	-1	-1	C
1	-1	1	-1	A
1	-1	-1	1	B
1	-1	-1	-1	C
1	1	1	-1	A
1	1	-1	1	B
1	1	-1	-1	C
0	0	1	-1	A
0	0	-1	1	B
0	0	-1	-1	C

Kromě standardního postupu při sestavování plánu se provádí generování s pomocí programů, které vytváří speciální zkrácené plány. Při generování takových plánu jsou různé možnosti, které však nejsou rovnocenné, např. pro dvě kvantitativní proměnné x_1, x_2 a jednu kvalitativní proměnnou z:

a)

z	x ₁	x ₂	x ₁ x ₂	x ₁ z	x ₂ z
-	-	-	+	+	+
-	+	-	-	-	+
-	-	+	-	+	-
-	+	+	-	-	-
+	0	0	0	0	0

b)

z	x ₁	x ₂	x ₁ x ₂	x ₁ z	x ₂ z
-	-	-	+	+	+
-	+	-	-	-	+
-	-	+	-	+	-
+	+	+	+	+	+
-	0	0	0	0	0

Oba plány obsahují jako základ plán 2^2 pro x_1 a x_2 s jedním centrálním bodem a umožňují odhad koeficientů u všech třech proměnných v lineárním modelu.

Plány se však liší úrovněmi z , které odpovídají jednotlivým bodům plánu. To způsobí, že pokud bychom místo lineárního modelu použili neúplný kvadratický, se členy x_1x_2 , x_1z a x_2z ,

$$y = \beta_0 + \sum_{i=1}^2 \beta_i x_i + \gamma \cdot z + \beta_{12} x_1 x_2 + \delta_1 x_1 z + \delta_2 x_2 z$$

nebyly by plány a), b) rovnocenné:

- oba plány umožní odhad b_{12} členu x_1x_2
- plán b) umožní odhad d_1 členu x_1z a d_2 členu x_2z
- plán a) neumožní odhad členů s interakcemi $z \cdot x_1$, $z \cdot x_2$ proto, že v plánu je přibližně $x_2z = (-1)x_2$
a $x_1z = (-1)x_1$ (silná korelace těchto sloupců)

Příklad 3

Pro dvě kvantitativní proměnné x_1, x_2 a dvě kvalitativní proměnné v_1, v_2 , každá se dvěma úrovněmi, se použijí dvě pomocné kvantitativní proměnné z_1, z_2 k modelování v_1 a v_2 v plánu:

Úrovně v_1 a v_2 jsou modelovány takto:

z_1	z_2	Úroveň v_1, v_2
1	0	(2,1)
0	1	(1,2)
0	0	(1,1)
1	1	(2,2)

Matice X pro lineární model (bez prvního sloupce rovněž plán experimentu) bude:

$$X = \begin{bmatrix} & x_1 & x_2 & z_1 & z_2 \\ 1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 & 1 \\ 1 & -1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

V tomto plánu již není obvyklé schéma střídání znamének, plán je generován speciálním programem automaticky. Zadávají se pouze vstupní parametry a požadavky na plán, např. max. počet pokusů a typ zvoleného modelu.

Příklad 4

Má se sestavit plán experimentu pro dvě kvalitativní proměnné z , w , které jsou na 3 úrovních a jedna kvantitativní proměnná x se 2 úrovněmi. Počet pokusů bude $3^2 \times 2^1$

Kódované hodnoty z_1 a z_2 modelují první kvalitativní proměnnou (z), kódované hodnoty w_1 a w_2 druhou proměnnou (w).

z_1	z_2	z
1	0	A
0	1	B
0	0	C

w_1	w_2	w
1	0	A
0	1	B
0	0	C

Plán:

Úrovně z, w		x	z_1	z_2	w_1	w_2
1,3		± 1	1	0	0	0
1,2		± 1	1	0	0	1
1,1		± 1	1	0	1	0
2,3		± 1	0	1	0	0
2,2		± 1	0	1	0	1
2,1		± 1	0	1	1	0
3,3		± 1	0	0	0	0
3,2		± 1	0	0	0	1
3,1		± 1	0	0	1	0

Obecně pro jednu kvalitativní proměnnou s k úrovněmi se použije $k-1$ pomocných kvantitativních proměnných;

k -tá úroveň je reprezentována k nulami $(0, \dots, 0)$.

Např. kvalitativní proměnná se čtyřmi úrovněmi bude vyjádřena třemi kvantitativními proměnnými z_i , $i = 1, 2, 3$:

z_1	z_2	z_3	Úroveň z
1	0	0	A
0	1	0	B
0	0	1	C
0	0	0	D

Nalezení plánu pro kvadratický model s kvalitativními proměnnými je obtížnější než u lineárních.

9.2 Jak působí kvalitativní proměnné v modelu

- není-li kvalitativní proměnná v interakci s proměnnou kvantitativní, ovlivňuje pouze absolutní člen
- je-li kvalitativní proměnná v interakci s proměnnou kvantitativní, ovlivňuje koeficienty kvantitativních proměnných

Příklad 5 (jak působí kvalitativní proměnná)

Mějme 2 kvantitativní proměnné a jednu kvalitativní.

- Model bez interakcí x , z :

$$\hat{y} = b_o + b_1 x_1 + b_2 x_2 + c \cdot z .$$

Dosazením za z např. $z = -1$ bude mít rovnice tvar

$$\hat{y} = (b_o - c) + b_1 x_1 + b_2 x_2 ,$$

takže dosazením za z se změnil jen *absolutní člen*.

- Model s interakcemi x , z :

$$\hat{y} = b_o + b_1 x_1 + b_2 x_2 + d_1 x_1 z + d_2 x_2 z .$$

Dosazením za $z = -1$ máme

$$\hat{y} = b_o + (b_1 - d_1) x_1 + (b_2 - d_2) x_2 ,$$

takže se *změnily koeficienty* u kvantitativních proměnných.



Shrnutí kapitoly 9

V této kapitole jsme ukázali, jak pracovat s kvalitativními proměnnými při sestavování plánů experimentu a při jejich zařazení do regresního modelu. Byly probrány různé počty kvantitativních a kvalitativních proměnných s různým počtem úrovní, které jsou zařazeny do modelu.



Otázky ke kapitole 9

1. Co jsou kvalitativní proměnné
2. Napište plán pro dvě kvantitativní a jednu kvalitativní proměnnou, všechny na dvou úrovních, model je lineární.
3. Napište matici X pro plán z bodu 2
4. Napište model pro dvě kvantitativní proměnné (x_1, x_2) a jednu kvalitativní (z), ve kterém jsou i smíšené členy kvalitativní proměnné s kvantitativními
5. Jak působí kvalitativní proměnné v modelu

10. ANALÝZA ROZPTYLU (ANOVA)



Čas ke studiu: 6 hodin



Cíl

- ukázat analýzu rozptylu jako zobecnění dvouvýběrového t-testu
- objasnit možnosti využití ANOVA k hodnocení závislosti výstupu na různém počtu vstupních faktorů



Výklad

10.1 Jednoduché třídění

Test významnosti rozdílu mezi *dvěma* výběrovými průměry (dvouvýběrový t-test) má nulovou hypotézu $H_0 : \mu_1 = \mu_2$. Pokud se prověřuje rovnost více než dvou středních hodnot, není správné použít opakovaně dvouvýběrový t-test, neboť dochází k výraznému narůstání chyby prvního druhu. Použije se analýza rozptylu (Analysis of Variance), která je označována zkratkou ANOVA.

ANOVA je test s těmito kroky:

Předpokládáme, že k dispozici je k nezávislých výběrů z rozdělení $N(\mu_i, \sigma^2)$, $i = 1, 2, \dots, k$.

- 1) $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
 H_1 : Alespoň jedna rovnost neplatí.

- 2) Testovací kritérium

$$F = \frac{S_T / (k - 1)}{S_R / (n - k)}$$

kde S_T a S_R jsou meziskupinový resp. vnitroskupinový *součet čtverců odchylek*. Vzorec pro jejich výpočet je uveden dále.

- 3) Kritická hodnota $F_{k-1, n-k}(\alpha)$

- 4) Je-li $T > F_{k-1, n-k}(\alpha)$, zamítá se H_0 .

Jak je patrné ze zadání (výběry z rozdělení $N(\mu_i, \sigma^2)$), podmínkou použití ANOVA je normalita dat a rovnost rozptylů ve vzorcích.

Protože velikost vzorků nebývá velká, je vhodný Shapiro Wilkův test pro ověření normality. Pro rovnost rozptylů, tedy testování hypotézy $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$, lze použít např. Bartlettův test. Data v příkladu 1 splňují oba předpoklady.

Veličinu, která bude předmětem sledování (např. sledovaný znak kvality), budeme dále označovat y a rozlišovat, kolik faktorů ovlivňuje její hodnotu. Podle toho se hovoří o jednoduchém třídění (y ovlivňuje jeden faktor), dvojném třídění, atd.

V této kapitole se předpokládá, že sledovaný ukazatel y závisí na jednom znaku (faktoru), který může mít k úrovní. Všechny potřebné vzorce i postup ANOVA uvedeme v následujícím příkladě.

Příklad 1 (jednoduché třídění)

Sleduje se množství emisí (y) automobilu v závislosti na typu přísady do benzínu (A,B,C,D). Celkem byly provedeny 4 pokusy s každým typem přísady. Zjištěné emise jsou v tabulce 1.

A	B	C	D
21	26	25	20
20	27	26	23
16	15	16	13
15	20	17	20
$\bar{y}_1 = 18$	$\bar{y}_2 = 22$	$\bar{y}_3 = 21$	$\bar{y}_4 = 19$
$\bar{y} = 20$			

Tab.1 Výsledky měření emisí s různými typy přísad

Má se rozhodnout, je-li významný rozdíl mezi výběrovými průměry emisí, které odpovídají jednotlivým typům přísad.

Řešení:

Z uvedených 16 výsledků je vypočítán součet čtverců odchylek S_D od celkového průměru $\bar{y}$ podle vzorce

$$S_D = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ji} - \bar{y})^2 \quad (1)$$

k = počet skupin (zde jsou skupiny A,B,C a D, proto $k = 4$)

i = číslo pokusu (zde $i = 1,2,3,4$)

j = číslo skupiny (zde $j = 1,2,3,4$)

n_j = počet hodnot v j -té skupině (zde všechny skupiny stejné, $n_j = 4$)

$$S_D = (21 - 20)^2 + (20 - 20)^2 + \dots + (13 - 20)^2 + (20 - 20)^2 = 296$$

S_D lze rozložit na součet čtverců odchylek mezi skupinami S_T a vnitroskupinový součet čtverců odchylek (reziduální) S_R . Platí

$$S_D = S_T + S_R \quad (2)$$

$$S_T = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 \quad (3)$$

n_j = četnost v j-té skupině ($j = 1, 2, \dots, k$)

$\bar{y}_j$ = průměr v j-té skupině

$$S_T = 4(18 - 20)^2 + 4(22 - 20)^2 + 4(21 - 20)^2 + 4(19 - 20)^2 = 40$$

$$S_R = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2 \quad (4)$$

$$S_R = (21 - 18)^2 + \dots + (15 - 18)^2 + (26 - 22)^2 + \dots + (20 - 22)^2 + (25 - 21)^2 + \\ + \dots + (17 - 21)^2 + (20 - 19)^2 + \dots + (20 - 19)^2 = 256$$

Výsledky rozkladu podle (2) jsou v tabulce 2.

Součet čtverců	d.f.	Rozptyl	Testovací krit	Sig.lev.
S _T = 40*	k – 1 4 – 1 = 3	$s_T^2 = 40/3 = 13.3$	$s_T^2 / s_R^2 = 0,62$	0,61
S _R = 256*	n – k 16 – 4 = 12	$s_R^2 = 256/12 = 21.3$		
S _D = 296*	3 + 12 = 15	Pozn.: Počet pokusů n = 16. d.f. = degrees of freedom = stupně volnosti		

Tab. 2 ANOVA (na PC)

Pozn. (*):

Indexy T, R a D souvisí s pojmy **T**reatment, **R**esidual a **D**ifference.

Závěr: Průměry se významně neliší, takže nezáleží na tom, jakou přísadu do paliva použijeme.

Výpočet s programem Excel:

Nástroje/Analýza dat/ANOVA:jeden faktor

Data: jako v tab.1

ANOVA						
Zdroj variability	Součet čtverců	d.f.	Rozptyl	F	Hodnota P	F krit
Mezi výběry	40	3	13,33333	0,625	0,61244	3,490295
Uvnitř výběrů	256	12	21,33333			
Celkem	296	15				

Poznámka: Ve výsledcích ANOVA bývá místo kritické hodnoty uváděna hodnota p-val (jiné označení *sig.lev.* nebo *hodnota p*). Vyhodnocení testu spočívá v porovnání p-val se zvolenou hladinou významnosti α : je-li

$p\text{-val} < \alpha$, byla překročena kritická hodnota a H_0 se zamítá. Jiné hodnocení testu: je-li $T >$ kritická hodnota, zamítá se H_0 .

Zrychlená ANOVA

Existuje i jiný, početně pohodlnější, rozklad součtu čtverců odchylek (resp. rozptylu). Vyjádříme-li

$$S_D = \sum_{j=1}^k \sum_{i=1}^{n_j} y_{ji}^2 - n \cdot \bar{y}^2 = S - S_A \quad (5)$$

a dosadíme do (2), bude

$$S - S_A = S_T + S_R$$

a odtud

$$S = S_A + S_T + S_R \quad (6)$$

Výpočet S a S_A je snadnější než S_D . S_R se „dopočítá“ nakonec z rovnice

$$S_R = S - S_A - S_T.$$

Rozklad S podle (6) je v tabulce 3.

$$S_A = n \cdot \bar{y}^2 = 16 \cdot 20^2 = 6400.$$

Součet čtverců	d.f.	Rozptyl	Test. kritérium	Kritická hodnota
S _A = 6400	1		$s_T^2 / s_R^2 = 0.62$	*) F _{3,12} (0.05) = 3,49
S _T = 40	3	$s_T^2 = 40/3 = 13.3$		
S _R = 256	12	$s_R^2 = 256/12 = 21.3$		
S = 6696	16	Pozn.: S _A se nevyhodnocuje.		

Tab. 3 ANOVA (ruční výpočet)

Protože testovací kritérium nepřekračuje kritickou hodnotu, není důvod k zamítnutí nulové hypotézy H_0 : Rozdíl mezi přísadami je bezvýznamný. Závěr je stejný v obou tabulkách.

* Poznámka: $F_{3,12}(0.05)$ se hledá v tabulkách, nebo s Excelem: **fx/FINV**

Pravděpod.	0,05
volnost1	3
volnost2	12

Schefeho kritérium

V případě zamítnutí H_0 se konstatuje, že průměry si nejsou rovny. Z testu však není jasné, které dvojice nesplňují rovnost. Máme-li z nějakého důvodu zájem to zjistit, je možné použít např. Schefeho kritérium, které tak tvoří pokračování testu. Postupuje se tak, že rozdíly výběrových průměrů v absolutní hodnotě se porovnávají s kritickou hodnotou (na pravé straně vztahu 7)

$$|\bar{y}_i - \bar{y}_j| > \left[(n_i^{-1} + n_j^{-1}) \cdot (k-1) \cdot s^2 F_{k-1, n-k}(\alpha) \right]^{\frac{1}{2}}, \quad (7)$$

$$\text{kde } s^2 = \frac{S_R}{n-k}$$

a je-li větší, prohlásíme rozdíl mezi těmito dvojicemi za významný. Takto se musí prověřit všechny dvojice průměrů. V software bývá příslušná dvojice nějak označena, např. * .

Příklad 2 (data z lit.[4])

Sleduje se výnos různých odrůd brambor (A,B,C,D). Výsledky jsou v tabulce

A	B	C	D
0,9	1,3	1,3	1,1
0,8	1	1,5	1,2
0,6	1,3	1,6	1
0,9		1,1	
		1,5	

Výpočet s programem [Excel](#):

Anova: jeden faktor

Faktor				
Výběr	Počet	Součet	Průměr	Rozptyl
Řádek 1	4	3,2	0,8	0,02
Řádek 2	3	3,6	1,2	0,03
Řádek 3	5	7	1,4	0,04
Řádek 4	3	3,3	1,1	0,01

ANOVA						
Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Mezi výběry	0,816	3	0,272	9,973333	0,001805	3,587431
Všechny výběry *	0,3	11	0,027273			
Celkem	1,116	14				

* takto je v Excelu označen součet čtverců odchylek uvnitř skupin, tedy S_R .

Pomocí ANOVA bylo zjištěno, že mezi průměrnými výnosy z trsu jsou významné rozdíly. Je třeba dále zjistit, mezi kterými odrůdami. V tabulce jsou dále uvedeny rozdíly průměrů a příslušná kritická hodnota. Část kritické hodnoty

$$\left[(k-1) \cdot s^2 F_{k-1, n-k}(\alpha) \right]$$

je stejná pro všechny dvojice průměrů a má hodnotu $F_{3,11}(0,05) = 3,59$.

$$\left[(k-1) \cdot s^2 F_{k-1, n-k}(\alpha) \right] = (4-1) \cdot 0,02727 \cdot 3,59$$

k (počet odrůd) = 4

n (počet trsů) = 15

Tabulka výsledků

Např. pro A-C:

$$|\bar{y}_1 - \bar{y}_3| > \left[(n_1^{-1} + n_3^{-1}) \cdot (k-1) \cdot s^2 F_{k-1, n-k}(\alpha) \right]^{\frac{1}{2}} = \left[(0,25 + 0,20) \cdot 3 \cdot \frac{0,3}{15-4} \cdot 3,59 \right]^{0,5} = 0,363$$

Dvojice	Levá strana (7)	Pravá strana (7)
A-B	0,4	0,41381
A-C	0,6	0,36345 *
A-D	0,3	0,41381
B-C	0,2	0,39568
B-D	0,1	0,44238
C-D	0,3	0,39568

Tab.4 Schefeho metoda

U dvojice odrůd A a C je levá strana vzorce (7) větší než pravá. Skutečně také výnosy odrůdy A(0.8) a C(1.4) jsou největším rozdílem ve výnosech.

Výpočet s programem STATGRAPHICS

Compare/ ANOVA/ One Way ...

Ikona: Tabulator Options:

Multiple range test = **Scheffe**:

Data: první sloupec Y_i , druhý sloupec : ke kterému faktoru patří Y_i

ANOVA Table for Col_1 by Col_2

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	0,816	3	0,272	9,97	0,0018
Within groups	0,3	11	0,0272727		
Total (Corr.)	1,116	14			

Číslo úrovně faktorů (contrast)	$\bar{y}_i - \bar{y}_j$ (difference)	Kritická hodnota (limits)
1 - 2	-0,4	0,413786
1 - 3	-0,6	0,363432 *
1 - 4	-0,3	0,413786
2 - 3	-0,2	0,395655
2 - 4	0,1	0,442356
3 - 4	0,3	0,395655

* označení statisticky významného rozdílu

Model ANOVA

Teoretická četnost v i-tém řádku a j-tém sloupci y_{ij} se modeluje jako průměr v této skupině + náhodná odchylka od skupinového průměru.

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

Průměr v j-té skupině je modelován jako součet celkového průměru a vlivu této úrovně faktoru

$$\mu_j = \mu + \tau_j$$

μ_j = průměr v j-té skupině (sloupci)

τ_j = vliv j-té úrovně faktoru

μ = celkový průměr

Dosazením dostáváme

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

ε_{ij} = náhodná odchylka

Odhad modelovaných veličin označíme stříškou:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i$$

MNČ se vypočítá:

$\hat{\mu} = \bar{y}_{..}$ (celkový průměr)

$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$ ($\bar{y}_{i.}$ je skupinový průměr)

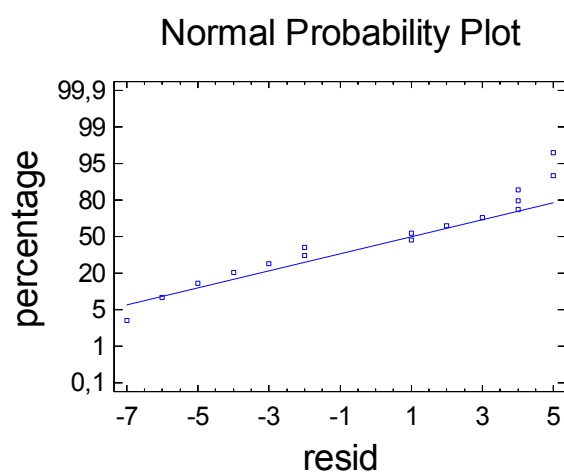
Potom

$$\hat{y}_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) = \bar{y}_{i.}$$

Model y_{ij} je potřeba k výpočtu reziduí: $e_{ij} = y_{ij} - \hat{y}_{ij}$

y_{ij}	$\hat{y}_{ij} = \bar{y}_{i.}$	e_{ij}
21	18	3
20	18	2
16	18	-2
15	18	-3
26	22	4
27	22	5
15	22	-7
20	22	-2
25	21	4
26	21	5
16	21	-5
17	21	-4
20	19	1
23	19	4
13	19	-6
20	19	1

Rezidua jsou potřeba ke konstrukci diagnostických grafů, kterými se ověřují předpoklady ANOVA. Normalita se ověří např. graficky pomocí normálního pravděpodobnostního grafu:



10.2 Dvojné třídění

Je-li sledovaný znak ovlivňován dvěma faktory, hovoříme o dvojném třídění. Rozklad součtu čtverců S se provede analogicky jako v případě jednoduchého třídění, pouze přibude jeden sčítanec. Označíme ho S_B . Rozklad součtu čtverců S potom bude

$$S = S_A + S_B + S_T + S_R \quad (8)$$

resp. bez použití S_A (provádí se na PC)

$$S_D = S_B + S_T + S_R \quad (9)$$

kde

$$S_B = k \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \quad (10)$$

$$S_T = n \sum_{j=1}^k (\bar{y}_j - \bar{y})^2 \quad (11)$$

$$S_A = n \cdot k \cdot \bar{y}^2 \quad (12)$$

$$S = \sum_{j=1}^k \sum_{i=1}^n y_{ij}^2 \quad (13)$$

Reziduální součet čtverců, který je nejpracnější, se vypočítá z rovnice (8)

$$S_R = S - S_A - S_B - S_T$$

Příklad 3 (pokračování z kap. 10.1)

Sledují se emise výfukových plynů (y) v závislosti na

- a) typu přísady (A,B,C,D) = 1. faktor, ovlivňující y ,
- b) vlivu řidiče (I,II,III,IV) = 2. faktor, ovlivňující y .

Výsledky experimentu jsou v tabulce 5.

	Přísada				
Řidič	A	B	C	D	průměr
I	21	26	25	20	
II	20	27	26	23	
III	16	15	16	13	
IV	15	20	17	20	
průměr					

Tab.5 Plán a výsledky experimentu

Výpočet skupinových průměrů a součtu čtverců odchylek manuálně provedeme v tabulce:

					průměry
	21	26	25	20	23
	20	27	26	23	24
	16	15	16	13	15
	15	20	17	20	18
průměry	18	22	21	19	20

Analýza rozptylu je provedena oběma způsoby: manuálním a na PC.
Rozklad podle (8):

Součet čtverců	d.f.	Rozptyl	Test.krit.	Krit.hod.
S _A = 6400	1		$s_B^2 / s_R^2 = 16,3$	$F_{3,9}(0,05) = 3,86$
S _B = 216	n - 1 = 3	$s_B^2 = 72$		
S _T = 40	k - 1 = 3	$s_T^2 = 13,3$	$s_T^2 / s_R^2 = 3,0$	
S _R = 40	9	$s_R^2 = 4,4$		
S = 2696	16	Pozn.:d.f. u S _R je obecně (n -1)(k -1)		

Tab. 6a ANOVA (manuální výpočet)

Protože jen testovací kritérium u faktoru B překračuje kritickou hodnotu, je pouze u faktoru B významný rozdíl průměrů.

Rozklad podle (9) je v tabulce 6.

Součet čtverců	d.f.	Rozptyl	Test. krit.	Sig. level
$S_B = 216$	3	$s_B^2 = 216 / 3 = 72$	$s_B^2 / s_R^2 = 16,3$ $s_T^2 / s_R^2 = 3,02$	0,0006
$S_T = 40$	3	$s_T^2 = 40 / 3 = 13,3$		0,0877
$S_R = 40$	9	$s_R^2 = 40 / 9 = 4,4$		
$S_D = 296$	15			

Tab. 6b ANOVA (výpočet s PC)

Pouze u B je významný rozdíl.

Lineární model pozorování

$$y_{ijk} = \mu + \beta_i + \tau_j + (\beta\tau)_{ij} + \varepsilon_{ijk}$$

$\beta_i, \tau_j, (\beta\tau)_{ij}$ vyjadřují vliv faktorů A, B a interakce AB

β_i = řádkový faktor

τ_j = sloupcový faktor

$(\beta\tau)_{ij}$ = smíšený faktor (interakce)

$\mu, \beta_i, \tau_j, (\beta\tau)_{ij}$ se odhadnou :

$\hat{\mu} = \bar{y}...$ ($\bar{y}...$ = celkový průměr)

$\beta_i = \bar{y}_{i..} - \bar{y}...$ ($\bar{y}_{i..}$ = průměr v i-tém řádku)

$\tau_j = \bar{y}_{.j.} - \bar{y}...$ ($\bar{y}_{.j.}$ = průměr v j-tém sloupci)

$(\hat{\beta}\tau)_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...$ ($\bar{y}_{ij.}$ = průměr v pokusu ij = okno ij)

Odhad modelu (řecká písmena se stříškou)

$$\begin{aligned}\hat{y}_{ijk} &= \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + (\hat{\beta}\tau)_{ij} = \\ &= \bar{y}... + (\bar{y}_{i..} - \bar{y}...) + (\bar{y}_{.j.} - \bar{y}...) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...) = \bar{y}_{ij.}\end{aligned}$$

10.3 Trojné třídění (Latinské čtverce)

Je-li sledovaný znak y ovlivňován třemi faktory, hovoříme o tzv. Latinských čtvercích (trojném třídění). Rozklad součtu čtverců odchylek zde bude

$$S = S_A + S_B + S_C + S_T + S_R \quad (14-7)$$

resp. při použití PC

$$S_D = S_B + S_C + S_T + S_R \quad (15-8)$$

kde

$$S_A = k \cdot k \cdot \bar{y}^2, \quad (16)$$

k je rozměr tabulky s výsledky experimentu.
 S_B, S_C, S_T se počítají stejně podle vzorce

$$k \sum_{i=1}^k (\bar{y}_i - \bar{y})^2, \quad (17)$$

$\bar{y}_i$ je průměr v příslušné (testované) skupině.

$$S_R = S - S_A - S_B - S_C - S_T$$

$$S = \sum_{i=1}^k \sum_{j=1}^k y_{ij}^2 \quad (18)$$

$$S_D = \sum_{i=1}^k \sum_{j=1}^k (y_{ij} - \bar{y})^2 \quad (19)$$

Příklad 4

Sleduje se množství emisí (y) v závislosti na

- typu přísady do benzínu (A,B,C,D),
- vlivu řidiče (I,II,III,IV),
- použitým vozidlem (1,2,3,4).

Výsledky experimentu jsou v tabulce 7. Protože 1. řádek a 1. sloupec jsou již obsazené, bude třetí faktor zaznamenáván do tabulky. Tato filosofie zápisu se uplatňuje i u čtyř a více faktorů.

	1	2	3	4
I	A 21	B 26	D 20	C 25
II	D 23	C 26	A 20	B 27
III	B 15	D 13	C 16	A 16
IV	C 17	A 15	B 20	D 20

Tab. 7 Výsledky experimentu s uvážením tří faktorů

Výpočty průměrů pro jednotlivá třídění jsou v pomocné tabulce 8.

Řidiči	Vozidla	Přísady
I:23	1:19	A:18
II:24	2:20	B:22
III:15	3:19	C:21
IV:18	4:22	D:19

Tab.8 Výpočet skupinových průměrů

$$S = 21^2 + 26^2 + 20^2 + \dots + 20^2 = 6696$$

$$S_B = 4[(23-20)^2 + (24-20)^2 + (15-20)^2 + (18-20)^2] = 216$$

$$S_C = 4[(19-20)^2 + (20-20)^2 + (19-20)^2 + (22-20)^2] = 24$$

$$S_T = 4[(18-20)^2 + (22-20)^2 + (21-20)^2 + (19-20)^2] = 40$$

$$S_A = 4 \cdot 4 \cdot 20^2 = 6400$$

$$S_R = 6696 - 6400 - 216 - 24 - 40 = 16$$

	d.f.	Rozptyl	Test.krit.	Krit.hod.
$S_A = 6400$	1	$s_A^2 = 6400$		
$S_B = 216$	$k-1=3$	$s_B^2 = 216/3 = 72$	$s_B^2 / s_R^2 \approx 27$	$F_{3,6}(0,05) = 4,76$
$S_C = 24$	$k-1=3$	$s_C^2 = 24/3 = 8$	$s_C^2 / s_R^2 \approx 3,0$	
$S_T = 40$	$k-1=3$	$s_T^2 = 40/3 = 13.3$	$s_T^2 / s_R^2 \approx 5,0$	
$S_R = 16$	$(k-1)(k-2)=6$	$s_R^2 = 16/6 = 2.67$		
$S = 6696$	16			

Tab. 9 ANOVA (manuální výpočet)

Závěr: Nejvýznamnějším vlivem je řidič (bloková proměnná B).
Rozklad podle (14):

Součet čtv.	d.f.	Rozptyl	Test.krit.	Sig.lev.
$S_B = 216$	$(k-1) = 3$	$s_B^2 = 72$	$s_B^2 / s_R^2 \approx 27$	0,0007
$S_C = 24$	$(k-1) = 3$	$s_C^2 = 8$	$s_C^2 / s_R^2 \approx 3,0$	0,1170
$S_T = 40$	$(k-1) = 3$	$s_T^2 = 13,3$	$s_T^2 / s_R^2 \approx 5,0$	0,0452
$S_R = 16$	$(k-1)(k-2) =$	$s_R^2 = 2,67$		
$S_D = 296$	15			

Tab. 10 ANOVA (na PC)

Jediným vlivným faktorem je faktor B.



Shrnutí kapitoly 10

V této závěrečné kapitole byl vysvětlen princip analýzy rozptylu za účelem hodnocení závislosti výstupu Y na vstupech x_1 (jednoduché třídění) resp. x_1, x_2 dvojné třídění atd.



Otázky ke kapitole 10

1. Jaká je souvislost dvouvýběrového t-testu a ANOVA
2. Zkuste jedním slovem odpovědět na otázku: „Co je ANOVA?“
3. Napište H_0 , testovací kritérium a kritickou hodnotu ANOVA
4. Co znamená jednoduché a dvojné třídění v ANOVA?
5. Rozložte součet čtverců S_D a S pro jednoduché třídění
6. V čem se liší ruční a počítačový výpočet rozkladu součtu čtverců odchylek
7. Napište tabulku ANOVA pro ruční výpočet (jednoduché třídění)
8. Napište tabulku ANOVA výpočet na PC (jednoduché třídění)
9. Co přesně znamená zamítnutí H_0 v ANOVA
10. Napište tabulku ANOVA pro ruční výpočet (dvojné třídění)
11. Napište tabulku ANOVA pro výpočet na PC (dvojné třídění)
12. K čemu slouží Schefeho kritérium; napište ho
13. Napište teoretický model ANOVA pro jednoduché třídění
14. Napište empirický model ANOVA pro jednoduché třídění
15. Jaké vlastnosti reziduí se v ANOVA předpokládají; jak se ověří
16. Rozložte součet čtverců S_D a S pro dvojné třídění
17. Nakreslete tabulku pro záznam výsledků pokusů u dvojného třídění
18. Napište teoretický model ANOVA pro dvojné třídění
19. Napište empirický model ANOVA pro dvojné třídění
20. Jak je nazýváno trojné třídění
21. Rozložte součet čtverců S_D a S pro trojné třídění
22. Nakreslete tabulku pro záznam výsledků pokusů u trojného třídění
23. Vytvořte tabulku 2x2 pro trojné třídění, doplňte libovolně „naměřené“ hodnoty Y a ukažte, jak budou vaše data zapsána v programu Statgraphics pro potřebu ANOVA



Úlohy ke kapitole 10

Příklad 1

Sleduje se závislost síly výbuchu výbušné směsi Y na čtyřech faktorech:

- a) pracovník, který směs připravil (1,2,3,4,5), b) dávka (I,II,III,IV,V), c) výbušná směs (A, B, C, D, E), d) vzorek ($\alpha, \beta, \gamma, \delta, \epsilon$). Výsledek experimentu je v tabulce č. 1.
Má se provést ANOVA na PC a manuálně.

Tabulka č. 1 Výsledek experimentu

Dávky čistého materiálu	Pracovník				
	1	2	3	4	5
I	A α 24	B γ 20	C ϵ 19	D β 24	E δ 24
II	B β 17	C δ 24	D α 30	E γ 27	A ϵ 36
III	C γ 18	D ϵ 38	E β 26	A δ 27	B α 21
IV	D δ 26	E α 31	A γ 26	B ϵ 23	C β 22
V	E ϵ 22	A β 30	B δ 20	C α 29	D γ 31

Příklad 2

Výrobce hliníkových slitin vyrábí přípravek na zjemnění zrna v ingotu. Produkt je vyráběn ve 4 pecích. Každá pec má své vlastní jedinečné výrobní charakteristiky, takže každý experiment, který je prováděn ve slévárně na více pecích, má pec jako problémovou proměnnou. Technici určili, že zásadní vliv na velikost zrna má také rychlost míchání. Každá pec má 4 rychlosti míchání.

Výsledné velikosti zrn jsou v tabulce. Proveďte ANOVA.

Rychlost	Pec			
míchání	1	2	3	4
5	8	4	5	6
10	14	5	6	9
15	14	6	9	2
20	17	9	3	6

Příklad 3

Posuďte, zda se významně liší průměrné výdaje v rodinách s různým počtem dětí, máte li tyto údaje:

Počet dětí v rodině	Výdaje v tis.Kč
0	36.1, 36.6, 34.7, 37.9, 34
1	37.9, 37.4, 38.8, 37.1, 35.1
2	36.2, 42.2, 39.5, 38.8, 40
3	43.6, 41.4, 39.2, 42.2, 38.8
4	38.6, 43.7, 39.1, 44.2, 44.4

Příklad 4

Proveďte ANOVA u jednoduchého třídění:

A	B	C
48	50	51
49	50	52
50	48	50
	52	

**KLÍČ K ŘEŠENÍ KAP. 10**

Řešení(1):

Vyhodnocení bylo provedeno v programu Statgraphics. Výsledky testu jsou v tabulce č. 2, Analýza rozptylu v PC.

Tabulka č. 2 Analýza rozptylu v PC

Analysis of Variance for hod_mereni - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:davky_cist_mat	68,0	4	17,0	2,06	0,1783
B:pracovnik	150,0	4	37,5	4,55	0,0329
C:vybusna_smes	330,0	4	82,5	10,00	0,0033
D:test_vzorek	62,0	4	15,5	1,88	0,2076
RESIDUAL	66,0	8	8,25		
TOTAL (CORRECTED)	676,0	24			

All F-ratios are based on the residual mean square error.

Vyhodnocení dle tabulky č. 2:

Porovnává se hodnota P-Value se zvolenou hladinou významnosti $\alpha = 0.05$. Je-li $P\text{-Value} < \alpha$, zamítá se H_0 o rovnosti průměrů a přijímá se alternativa H_1 .

Pro faktor – dávky čistého materiálu je $P\text{-Value } 0,1783 > \alpha \rightarrow$ přijímá se H_0

Pro faktor – pracovník je $P\text{-Value } 0,0329 < \alpha \rightarrow$ zamítá se H_0 a přijímá se H_1

Pro faktor – výbušná směs je $P\text{-Value } 0,0033 < \alpha \rightarrow$ zamítá se H_0 a přijímá se H_1

Pro faktor – testovaný vzorek je $P\text{-Value } 0,2076 > \alpha \rightarrow$ přijímá se H_0

Ruční výpočet:

Tabulka č. 3 Výsledek experimentu s výběrovými průměry

Dávky čistého materiálu	Pracovník					$\bar{y}_i$
	1	2	3	4	5	
1	A α 24	B γ 20	C ϵ 19	D β 24	E δ 24	22,2
2	B β 17	C δ 24	D α 30	E γ 27	A ϵ 36	26,8
3	C γ 18	D ϵ 38	E β 26	A δ 27	B α 21	26,0
4	D δ 26	E α 31	A γ 26	B ϵ 23	C β 22	25,6
5	E ϵ 22	A β 30	B δ 20	C α 29	D γ 31	26,4
$\bar{y}_i$	21,4	28,6	24,2	26,0	26,8	$\bar{y} = \mathbf{25,4}$

Další průměry y_i :

A = 28,6	$\alpha = 27,0$
B = 20,2	$\beta = 23,8$
C = 22,4	$\gamma = 24,4$
D = 29,8	$\delta = 24,2$
E = 26,0	$\epsilon = 27,6$

Označení: S_B – dávka, S_C – pracovník, S_E – výbušná směs, S_T – testovací vzorky.

$$S = S_A + S_B + S_C + S_E + S_T + S_R$$

$$S_R = S - S_A + S_B + S_C + S_E + S_T$$

$$\bar{y} = \frac{1}{25} \cdot \sum_{i=1}^k y_i = \mathbf{25,4}$$

$$S = \sum_j^k \sum_i^k y_{ij}^2 = (24^2 + 17^2 + 18^2 + 26^2 + 22^2 + 20^2 + 24^2 + 38^2 + 31^2 + 30^2 + 19^2 + 30^2 + 26^2 + 26^2 + 20^2 + 24^2 + 27^2 + 27^2 + 23^2 + 29^2 + 24^2 + 36^2 + 21^2 + 22^2 + 31^2) = \mathbf{16805}$$

$$S_A = k \cdot k \cdot \bar{y}^2 = 5 \cdot 5 \cdot 25,4^2 = \mathbf{16129}$$

$$S_B = S_C = S_E = S_T = k \cdot \sum_{i=1}^k (\bar{y}_i - \bar{y})^2,$$

kde $\bar{y}_i$ je průměr v příslušné testované skupině.

$$S_B = 5 \cdot [(22,2 - 25,4)^2 + (26,8 - 25,4)^2 + (26,0 - 25,4)^2 + (25,6 - 25,4)^2 + (26,4 - 25,4)^2] = 5 \cdot 13,6 = \mathbf{68}$$

$$S_C = 5 \cdot [(21,4 - 25,4)^2 + (28,6 - 25,4)^2 + (24,2 - 25,4)^2 + (26,0 - 25,4)^2 + (26,8 - 25,4)^2] = 5 \cdot 50 = \mathbf{150}$$

$$S_E = 5 \cdot [(28,6 - 25,4)^2 + (20,2 - 25,4)^2 + (22,4 - 25,4)^2 + (29,8 - 25,4)^2 + (26,0 - 25,4)^2] = 5 \cdot 66 = \mathbf{330}$$

$$S_T = 5 \cdot [(27,0 - 25,4)^2 + (23,8 - 25,4)^2 + (24,4 - 25,4)^2 + (24,2 - 25,4)^2 + (27,6 - 25,4)^2] = 5 \cdot 12,4 = 62$$

$$S_R = S - S_A + S_B + S_C + S_E + S_T = 16805 - 16129 - 68 - 330 - 62 - 150 = 66$$

Tabulka č 4 ANOVA – ruční výpočet

Součet čtverců	Stupně volnosti	Rozptyl	Testovací kritérium	Kritická hodnota
$S_A = 16129$	1			
$S_B = 68$	$k - 1 = 4$	$S_B^2 = \frac{68}{4} = 17$	$\frac{S_B^2}{S_R^2} = \frac{17}{8,25} = 2,06$	$F_{4,8}(0,05) = 3,838$
$S_C = 150$	$k - 1 = 4$	$S_C^2 = \frac{150}{4} = 37,5$	$\frac{S_C^2}{S_R^2} = \frac{37,5}{8,25} = 4,55$	$F_{4,8}(0,05) = 3,838$
$S_E = 330$	$k - 1 = 4$	$S_E^2 = \frac{330}{4} = 82,5$	$\frac{S_E^2}{S_R^2} = \frac{82,5}{8,25} = 10,0$	$F_{4,8}(0,05) = 3,838$
$S_T = 62$	$k - 1 = 4$	$S_T^2 = \frac{62}{4} = 15,5$	$\frac{S_T^2}{S_R^2} = \frac{15,5}{8,25} = 1,88$	$F_{4,8}(0,05) = 3,838$
$S_R = 66$	$(k-1)(k-3) = 8$	$S_R^2 = \frac{66}{8} = 8,25$		
$S = 16805$	25			

Vyhodnocení tabulky č. 4:

Při ručním výpočtu se porovnává hodnota testovacího kritéria s kritickou hodnotou. Je-li hodnota testovacího kritéria menší než kritická hodnota, není důvod k zamítnutí nulové hypotézy (přijímá se hypotézu H_0). Je-li hodnota testovacího kritéria větší než kritická hodnota, přijímá se hypotéza H_1 .

Pro faktor S_B – dávka čistého materiálu:

$$T = 2,06 < K = 3,838 \rightarrow \text{přijímá se } H_0$$

Pro faktor S_C – pracovník:

$$T = 4,55 > K = 3,838 \rightarrow \text{přijímá se } H_1.$$

Pro faktor S_E – výbušná směs (A, B, C, D, E):

$$T = 10,00 > K = 3,838 \rightarrow \text{přijímá se } H_1.$$

Pro faktor S_T

$T = 1,88 < K = 3,838 \rightarrow$ přijímá se H_0 .

Oběma postupy (v PC i ručním výpočtem) jsme došli ke stejnému závěru.

Řešení(2):

Řešení na PC

K analýze použijeme SW STATGRAPHICS Plus.

Výpis programu:

Compare – Analysis of Variance – Multifactor ANOVA

Analysis of Variance for Col_1 - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Col_2	165,188	3	55,0625	6,35	0,0133
B:Col_3	22,1875	3	7,39583	0,85	0,4995
RESIDUAL	78,0625	9	8,67361		
TOTAL	265,438	15			

Data se do tabulky SW Statgraphics ukládají takto:

Měření	Dávky	Pracovník	Směs
24	1	1	1
17	2	1	2
18	3	1	3
26	4	1	4
22	5	1	5
20	1	2	2
24	2	2	3
38	3	2	4
31	4	2	5
30	5	2	1
19	1	3	3
30	2	3	4
26	3	3	5
26	4	3	1
20	5	3	2
24	1	4	4
27	2	4	5
27	3	4	1
23	4	4	2
29	5	4	3
24	1	5	5
36	2	5	1
21	3	5	2
22	4	5	3
31	5	5	4

Manuální řešení

Rozklad součtů čtverců S se provede:

$$S = S_A + S_B + S_T + S_R,$$

$$\text{kde } S_A = n \cdot \bar{y}^2 = 4 \cdot 5 \cdot 7,69^2 = \underline{\underline{946,178}}$$

$$S = \sum_{I=1}^k \sum_{j=1}^k y_{ij}^2 = \underline{\underline{1211,61}}$$

$$S_B = k \cdot \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 = 4 \cdot [(5,75 - 7,69)^2 + (8,5 - 7,69)^2 + (7,75 - 7,69)^2 + (8,75 - 7,69)^2]$$

$$S_B = \underline{\underline{22,18}}$$

$$S_T = k \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 = 4 \cdot [(13,25 - 7,6875)^2 + (6 - 7,6875)^2 + (5,75 - 7,6875)^2 + (5,75 - 7,6875)^2]$$

$$S_T = \underline{\underline{165,19}}$$

$$S_R = S - S_A - S_B - S_T = 1211,65 - 946,178 - 22,18 - 165,19 = \underline{\underline{78,0625}}$$

Kritická hodnota z tabulek: $F_{3,9}(0,05) = 5,08$

Tabulka ANOVA

	d.f.	Rozptyl	Testovací kritérium	Kritická hodnota
S _A = 946,178	1			
S _B = 22,18	3	$S_B^2 = 22,18 / 3 = 7,39$	$S_B^2 / S_R^2 = 0,85$	$F_{3,9}(0,05) = 3,86$
S _T =165,19	3	$S_T^2 = 165,19 / 3 = 55,06$	$S_T^2 / S_R^2 = 6,35$	
S _R = 78,0625	9	$S_R^2 = 78,0625 / 9 = 8,67$		
S = 1211,61	16			

Výsledky (3):

($S_T = 120.372$, $S_R = 77.588$, $T = 7.757$)

Řešení(4):

A	B	C
48	50	51
49	50	52
50	48	50
	52	
49	50	51

$\bar{Y} = 50$

$n = 10$, $k = 3$

S		d.f.	s^2	T	Kritic. hodnota
S_M	6	2	3	1,75	
S_V	12	7	1,71		$F_{2,7}(0,05) = 6,54$
S_T	18				

$$S_M = 3.(49-50)^2 + 4.(50-50)^2 + 3.(51-50)^2 = 6$$

$$S = 25018$$

$$S_A = 10.50^2 = 25000$$

$$S = 2500 + S_M + S_V; S_V = 25018 - 25000 - 6 = 12$$

11. ŠIRŠÍ SOUVISLOSTI PLÁNOVÁNÍ EXPERIMENTŮ



Čas ke studiu: 20 hodin



Cíl kapitoly

- a) ukázat souvislosti DOE, regresní analýzy a ekonometrie
- b) uvést základní vzorce regresní analýzy
- c) definovat předpoklady regresní analýzy včetně způsobu jejich testování a překonání problému při jejich nesplnění



Výklad

Tato skripta jsou věnována plánování experimentů. Jedná se o disciplínu s širokými možnostmi aplikací mj. také v řízení jakosti. Současně je to však téma, související velmi úzce s dalšími matematickými disciplínami, jmenovitě regresní analýzou a také ekonometrií. V této závěrečné kapitole bychom rádi ukázali souvislosti uvedených disciplín.

V matematické statistice byla probrána kapitola regresní analýzy, zabývající se modelováním závislosti sledovaných výstupů (např. ukazatelů kvality) na proměnných, označovaných v řízení procesů jako vstupy, regulovatelné faktory resp. nezávislé proměnné.

Abychom mohli dále probírat problémová místa regresní analýzy, připomeneme nejvýznamnější pojmy a vztahy.

11.1 Regresní analýza

11.1.1 Základní pojmy regresní analýzy, metoda nejmenších čtverců

Tvar závislosti y na $\vec{x} = (x_1, \dots, x_k)$ je v nejjednodušším případě, kdy $\vec{x} = x$, vyjádřen obecně $y = f(x)$. Funkční předpis f může mít nejrůznější tvar, např. $y = \beta_1 + \beta_2 x$, $y = \beta_0 + \beta_1 x + \beta_2 x^2$, $y = \ln(x)$, atp. Při závislosti y na k proměnných máme rovnici $y = f(x_1, \dots, x_k) = f(\vec{x})$. Je-li tato funkce uvažována ve tvaru

$$y = \beta_1 f_1(\vec{x}) + \beta_2 f_2(\vec{x}) + \dots + \beta_k f_k(\vec{x})$$

hovoříme o lineární regresi (lineární podle parametrů).

Nejčastěji uvažujeme (1) ve tvaru:

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

V nejjednodušším případě, kdy y je lineární funkcí x , je $y = \beta_1 + \beta_2 x_2$.

K dispozici jsou experimentálně získané body $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$.

Podstatné při hledání těchto bodů je, že x_i se *volí*, zatímco Y_i se získá *měřením* a je tedy zatíženo chybou měření.

Y_i = empirická (měřená) hodnota závisle proměnné

y_i = teoretická hodnota závisle proměnné

ε_i = reziduální odchylka

Vztah mezi empirickou hodnotou Y_i a teoretickou hodnotou y_i je vyjádřen rovnicí

$$Y_i = y_i + \varepsilon_i; \quad (1)$$

ε_i je reziduální odchylka

Má-li teoretická regresní funkce tvar

$$y = \beta_1 x_1 + \dots + \beta_k x_k, \quad (2)$$

je

$$y_i = \beta_1 x_{i1} + \dots + \beta_k x_{ik}, \quad i = 1, 2, \dots, n \quad (3)$$

a

$$Y_i = \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i \quad (4)$$

Odhadem teoretické regresní funkce (3) je empirická regresní funkce

$$\hat{Y} = b_1 x_{i1} + \dots + b_k x_{ik}. \quad (5)$$

Vztah (1) potom můžeme zapsat ve tvaru

$$Y_i = \hat{Y}_i + e_i \quad (6)$$

e_i je empirická reziduální odchylka.

Po dosazení za $\hat{Y}_i$ do (6) máme

$$Y_i = b_1 x_{i1} + \dots + b_k x_{ik} + e_i; \quad i = 1, 2, \dots, n. \quad (7)$$

Soustavu rovnic (7) lze zapsat pomocí matic

$$Y = X\vec{b} + \vec{e} \quad (8)$$

Matice X má tvar

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad (9)$$

První řádek je první *volbou* proměnných $x_1, \dots, x_k$, atd. až po n -tou volbu v posledním řádku. Zdůrazňujeme volba: proměnné x_i se volí, hodnoty Y_i jsou získány měřením.

Nyní budeme hledat způsob výpočtu koeficientů b_i , tedy vektoru $\vec{b}$; empirický regresní koeficient b_i je bodovým odhadem teoretického regresního koeficientu β_i .

Kriteriem pro posouzení kvality nalezených odhadů b_i je součet čtverců odchylek reziduí

$$S_e = \sum_{i=1}^n e_i^2;$$

nejlepší odhad je ten, pro který je *součet čtverců odchylek* $\sum_{i=1}^n e_i^2$ minimální, tedy

$$\sum_{i=1}^n e_i^2 = \min. \quad (10)$$

Vztah (10), nazývaný *metoda nejmenších čtverců* (MNC resp. LS), je výchozí pro nalezení vektoru $\vec{b}$;

v maticovém vyjádření má tvar

$$S_e = \sum_{i=1}^n e_i^2 = \vec{e}^T \cdot \vec{e} = \min.$$

11.1.2 Předpoklady regresní analýzy

O reziduálních odchylkách, které z pohledu teorie pravděpodobnosti představují náhodné proměnné, se činí tyto předpoklady:

- 1) Střední hodnota je nula: $E(\varepsilon_i) = 0$
- 2) Rozptyl je konstantní, nezávislý na i : $D(\varepsilon_i) = \sigma^2$
- 3) Veličiny $\varepsilon_i, \varepsilon_j$ jsou nezávislé: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$
- 4) Veličiny ε_i mají normální rozdělení: $\varepsilon_i \sim N(0, \sigma^2)$

Také o matici (9) se činí jisté předpoklady. Uvedeme je v návaznosti k předpokladům o reziduích:

- 5) matice X je nenáhodná

6) sloupce v $\mathbf{X}$ jsou lineárně nezávislé

Předpoklady o reziduích lze vyjádřit také v maticovém tvaru

$$1b) E(\vec{\varepsilon}) = \vec{0}$$

$$2b) \text{var}(\vec{\varepsilon}) = \sigma^2 I$$

$$3b) \vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$$

Zápis $\text{var}(\vec{\varepsilon}) = \sigma^2 I$ obsahuje body (2) a (3), neboť

$$\text{var}(\vec{\varepsilon}) = \sigma^2 I = \sigma^2 \begin{bmatrix} D(\varepsilon_1) & \text{cov}(\varepsilon_1, \varepsilon_2) & \text{cov}(\varepsilon_1, \varepsilon_3) \\ \text{cov}(\varepsilon_2, \varepsilon_1) & D(\varepsilon_2) & \text{cov}(\varepsilon_2, \varepsilon_3) \\ \text{cov}(\varepsilon_3, \varepsilon_1) & \text{cov}(\varepsilon_3, \varepsilon_2) & D(\varepsilon_3) \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

Odtud je zřejmé, že rozptyly jsou konstantní (bod 2) a kovariance nulové (bod 3).

Při hledání modelu a následném testování jeho vlastností budou použity věty regresní analýzy, které nyní uvede.

Věty zásadního významu jsou uvedeny s důkazem, ostatní pomocné věty bez důkazu.

11.1.3 Vybrané věty regresní analýzy

Věta 1

Odhad regresních koeficientů $\vec{b} = (b_1, \dots, b_k)^T$ metodou nejmenších čtverců se vypočítá pomocí vztahu

$$\vec{b} = (X^T X)^{-1} X^T Y. \quad (12)$$

Dosazením $Y - Xb$ za $\vec{e}$ do (11) dostáváme

$$S_e = (Y - Xb)^T (Y - Xb) = (Y^T - b^T X^T)(Y - Xb) = Y^T Y - Y^T Xb - b^T X^T Y + b^T X^T Xb$$

Protože

$$Y^T Xb = (b^T X^T Y)^T$$

a výraz v závorce je skalár, je $Y^T Xb = (b^T X^T Y)^T = b^T X^T Y$, můžeme psát

$$S_e = Y^T Y - Y^T Xb - b^T X^T Y + b^T X^T Xb = Y^T Y - 2b^T X^T Y + b^T X^T Xb$$

Hledá se minimum této vektorové funkce. Pro derivace S_e podle vektoru $\vec{b}$ použijeme věty z kapitoly 4.3.2: Jsou-li ϖ a α vektory, A je matice, potom

$$\frac{\delta \omega^T \alpha}{\delta \omega} = \alpha$$

$$\frac{\delta \omega^T A \omega}{\delta \omega} = (A + A^T) \omega$$

Derivováním S_e dostáváme

$$\frac{\partial S_e}{\partial \vec{b}} = -2X^T Y + 2X^T X \vec{b} = 0$$

Odtud

$$\vec{b} = (X^T X)^{-1} X^T Y.$$

Výpočet $\vec{b}$ podle Věty 1 je bodovým odhadem regresních koeficientů $\beta_1, \dots, \beta_k$.

U bodových odhadů je potřeba zjistit, zda se jedná o nevychýlený odhad, tj. zda $E\vec{b} = \vec{\beta}$ a dále stanovit varianční matici $\text{var}(\vec{b})$.

Věta 2

Odhad $\vec{b}$ je nestranný, tzn. $E(\vec{b}) = \vec{\beta}$ a má varianční matici $\text{var}(\vec{b}) = \sigma^2 (X^T X)^{-1}$.

$$\text{a) } E\vec{b} = E(X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T E(X\vec{\beta} + \vec{\varepsilon}) = (X^T X)^{-1} X^T X\vec{\beta} = \vec{\beta}$$

$$\begin{aligned} \text{b) } \text{var}(\vec{b}) &= \text{var}(X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T \text{var}(X\vec{\beta} + \vec{\varepsilon}) X [(X^T X)^{-1}]^T \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

kde σ^2 se počítá podle V4.

Věta 3

Pro reziduální součet čtverců S_e platí:

$$\text{a) } S_e = Y^T [I - X(X^T X)^{-1} X^T] Y$$

$$\text{b) } S_e = Y^T Y - \vec{b}^T X^T Y$$

Věta 4

Nestranným odhadem parametru σ^2 je náhodná veličina $s^2 = \frac{S_e}{n-k}$

Věta 5

Má-li náhodný vektor $\vec{\varepsilon}$ normální rozdělení, tj. $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$, potom

a) $Y \sim N(X\vec{\beta}, \sigma^2 I)$

b) $\vec{b} \sim N\left[\vec{\beta}, \sigma^2 (X^T X)^{-1}\right]$

Věta 6

Náhodná veličina $\frac{Se}{\sigma^2}$ má rozdělení χ_{n-k}^2 .

Věta 7

Označme v_{ij} prvky matice $V = (X^T X)^{-1}$. Potom

$$\frac{b_i - \beta_i}{s\sqrt{v_{ii}}} \sim t_{n-k}$$

S použitím věty: Má-li náhodná veličina X_1 rozdělení $N(0,1)$ a náhodná veličina X_2 rozdělení

χ_n^2 , pak $\frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t_{n-k}$, sestrojíme takový zlomek pomocí V_2, V_4, V_5 :

$$\frac{\frac{b_i - \beta_i}{\sigma\sqrt{v_{ii}}}}{\sqrt{\frac{\frac{S_e}{\sigma^2}}{n-k}}} = \frac{\frac{b_i - \beta_i}{\sigma\sqrt{v_{ii}}}}{\sqrt{\frac{s^2(n-k)}{\sigma^2}}} = \frac{b_i - \beta_i}{s\sqrt{v_{ii}}} \sim t_{n-k}$$

Věta 7 se používá

a) pro testování hypotézy $H_o : \beta_i = \beta_i^o$:

Jestliže

$$\left| \frac{b_i - \beta_i^o}{s\sqrt{v_{ii}}} \right| > t_{n-k}(\alpha),$$

zamítá se H_0 . Nejčastěji se volí $\beta_i^o = 0$.

b) k intervalovému odhadu koeficientů β_i :

$$P\left[\left| \frac{b_i - \beta_i}{s\sqrt{v_{ii}}} \right| \leq t_{n-k}(\alpha) \right] = 1 - \alpha.$$

Odtud

$$P\left[b_i - t_{n-k}(\alpha)\sqrt{s^2 v_{ii}} \leq \beta_i \leq b_i + t_{n-k}(\alpha)\sqrt{s^2 v_{ii}} \right]$$

Věta 8

Jestliže $\vec{\beta} = \vec{0}$, potom náhodná veličina $\frac{S_{\bar{y}} / k}{S_e / (n - k)}$ má rozdělení $F_{k, n-k}$.

Tato věta se používá v regresní analýze k testování modelu.

Příklad 1

Sledujeme závislost spotřeby elektrické energie (y) na délce elektrického vedení (x_2) a odběru energie (x_3). Má se najít regresní funkce ve tvaru $\hat{y} = b_1 + b_2 x_2 + b_3 x_3$ a provést predikci pro $x_1=1.8$, $x_2=4.3$ a $x_1=1.9$, $x_2=4.4$.

x_2	x_3	Y
1,2	3,6	3,2
1,3	3,7	3,3
1,3	3,8	3,4
1,4	3,8	3,5
1,4	3,9	3,6
1,5	3,9	3,6
1,5	4	3,7
1,6	4	3,8
1,6	4,1	3,9
1,7	4,2	4

Řešení:

Z tabulky získáme potřebné matice X a Y

$$X = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} \quad Y = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix}$$

Jelikož v regresní funkci je absolutní člen, je $x_1 = 1$ a matice X doplněna o první sloupec jedniček. Proto také x -ové hodnoty v tabulce jsou označeny x_2 a x_3 .

Vztah (12) vytvoříme postupně:

$$X^T \cdot X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} = \begin{bmatrix} 10 & 14,5 & 39 \\ 14,5 & 21,25 & 56,8 \\ 39 & 56,8 & 152,4 \end{bmatrix}$$

$$(X^T \cdot X)^{-1} = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix}$$

$$X^T \cdot Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1,2 & 1,3 & 1,3 & 1,4 & 1,4 & 1,5 & 1,5 & 1,6 & 1,6 & 1,7 \\ 3,6 & 3,7 & 3,8 & 3,8 & 3,9 & 3,9 & 4 & 4 & 4,1 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,8 \\ 3,9 \\ 4 \end{bmatrix} = \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix}$$

Nyní můžeme podle (12) vypočítat:

$$\vec{b} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} \cdot \begin{bmatrix} 36 \\ 52,56 \\ 140,82 \end{bmatrix} = \begin{bmatrix} -0,78 \\ 0,60 \\ 0,90 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Hledaná regresní funkce má tedy rovnici $\hat{Y} = -0,78 + 0,60x_2 + 0,90x_3$.

Výpočet teoretických hodnot

Teoretické hodnoty $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$ lze vypočítat dosazením za x_2 a x_3 z prvního, druhého, ..., n -tého řádku matice X do regresní funkce, např.

$$\hat{Y}_1 = -0,78 + 0,60 \cdot 1,2 + 0,90 \cdot 3,6 = 3,18$$

$$\hat{Y}_2 = -0,78 + 0,60 \cdot 1,3 + 0,90 \cdot 3,7 = 3,33$$

$$\dots$$

$$\hat{Y}_{10} = -0,78 + 0,60 \cdot 1,7 + 0,90 \cdot 4,2 = 4,02$$

Tyto hodnoty lze získat rychleji a všechny najednou takto

$$\hat{Y} = X \cdot \vec{b} = \begin{bmatrix} 1 & 1,2 & 3,6 \\ 1 & 1,3 & 3,7 \\ 1 & 1,3 & 3,8 \\ 1 & 1,4 & 3,8 \\ 1 & 1,4 & 3,9 \\ 1 & 1,5 & 3,9 \\ 1 & 1,5 & 4 \\ 1 & 1,6 & 4 \\ 1 & 1,6 & 4,1 \\ 1 & 1,7 & 4,2 \end{bmatrix} \cdot \begin{bmatrix} -0,78 \\ 0,6 \\ 0,9 \end{bmatrix} = \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \hat{Y}_{10} \end{bmatrix}$$

Výpočet reziduálních odchylek

Odečtením vektorů Y a $\hat{Y}$ získáme vektor empirických reziduálních odchylek $\vec{e}$:

$$\vec{e} = Y - \hat{Y} = \begin{bmatrix} 3,2 \\ 3,3 \\ 3,4 \\ 3,5 \\ 3,6 \\ 3,6 \\ 3,7 \\ 3,9 \\ 3,9 \\ 4 \end{bmatrix} - \begin{bmatrix} 3,18 \\ 3,33 \\ 3,42 \\ 3,48 \\ 3,57 \\ 3,63 \\ 3,72 \\ 3,78 \\ 3,87 \\ 4,02 \end{bmatrix} = \begin{bmatrix} 0,02 \\ -0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ -0,03 \\ 0,02 \\ 0,02 \\ 0,03 \\ -0,02 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ e_{10} \end{bmatrix}$$

Rozptyl odhadu regresních koeficientů

Protože při výpočtu regresních koeficientů se jedná o odhady, je účelné také nalézt rozptyly těchto odhadů. Získáme je jako prvky hlavní diagonály matice:

$$\text{var}(\vec{b}) = s^2 \cdot (X^T \cdot X)^{-1} \quad (13)$$

kde

$$s^2 = \frac{\vec{e}^T \cdot \vec{e}}{n - k} \quad (14)$$

$\vec{e}$ = vektor reziduálních odchylek

$\vec{e}^T$ = vektor transponovaný k $\vec{e}$

n = počet bodů

k = počet parametrů

V našem případě

$$\vec{e}^T \cdot \vec{e} = \begin{bmatrix} 0,02 & -0,03 & -0,02 & 0,02 & 0,03 & 0,03 & -0,02 & 0,02 & 0,03 & -0,02 \end{bmatrix} \cdot \begin{bmatrix} 0,02 \\ -0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ 0,03 \\ -0,02 \\ 0,02 \\ 0,03 \\ -0,02 \end{bmatrix} = 0,006$$

$$s^2 = \frac{\vec{e}^T \cdot \vec{e}}{n - k} = \frac{0,006}{10 - 3} = 0,0008571$$

$$\text{var}(\vec{b}) = s^2 \cdot (X^T \cdot X)^{-1} = 0,0008571 \cdot \begin{bmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{bmatrix} = \begin{bmatrix} 0,2102 & 0,0926 & -0,0883 \\ 0,0926 & 0,0514 & -0,0429 \\ 0,0883 & -0,0429 & 0,0386 \end{bmatrix}$$

$$D(b_1) = 0,2102 \Rightarrow s(b_1) = 0,4584$$

Podobně:

$$D(b_2) = 0,0514 \Rightarrow s(b_2) = 0,2267$$

$$D(b_3) = 0,0386 \Rightarrow s(b_3) = 0,1965$$

Po nalezení rovnice regresní funkce a rozptylu odhadů regresních koeficientů píšeme obvykle výsledné řešení tak, že pod regresní koeficienty uvádíme příslušné směrodatné odchylky. V našem případě máme:

$$\hat{Y} = -0,78 + 0,60 x_2 + 0,90 x_3 \\ (0,4584) \quad (0,2267) \quad (0,1965)$$

Při výpočtu regresních koeficientů $b_1, b_2, \dots, b_k$ se stává, že mezi koeficienty jsou až řádové rozdíly, např. $b_1 = 200$ a $b_2 = 0,02$. V takových případech stojíme před problémem, zda má smysl zařadit např. zde b_2 do regresní funkce. K objektivnímu posouzení významnosti regresních koeficientů lze použít následující test.

Test významnosti regresních koeficientů

1) Nulová a alternativní hypotéza

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

2) Testovací kritérium

$$T_i = \frac{b_i}{s(b_i)} \quad (15)$$

kde b_i je odhad parametru β_i ,
 $s(b_i)$ je směrodatná odchylka odhadu tohoto parametru.

3) Kritická hodnota $t_{n-k}(\alpha)$ 4) Porovnáme T a $t_{n-k}(\alpha)$

Je-li $|T| > t_{n-k}(\alpha)$, zamítáme H_0 a přijmeme alternativní hypotézu H_1 , podle které vypočítaný koeficient je možné považovat za různý od nuly a je proto důvod pro jeho zařazení do regresní funkce.

Zde máme

$$T_2 = \frac{b_2}{s(b_2)} = \frac{0,60}{\sqrt{0,0514}} = 2,65$$

$$T_3 = \frac{b_3}{s(b_3)} = 4,58$$

$$t_{n-k}(\alpha) = t_{10-3}(0,1) = 1,895$$

Protože $T_2 > 1,895$ a také $T_3 > 1,895$, zařadíme oba koeficienty do regresní funkce.

Jak bylo řečeno, výchozím předpokladem pro nalezení odhadu $\vec{b} = (b_1, \dots, b_k)$ koeficientů $\vec{\beta} = (\beta_1, \dots, \beta_k)$ je znalost matic X a Y , což znamená znalost bodů $(x_{i1}, x_{i2}, \dots, x_{ik}, Y_i)$. Praktik, stojící před úlohou formulovanou v úvodu této kapitoly, však musí tyto hodnoty získat. V této souvislosti musí rozhodnout, jak volit hodnoty $(x_{i1}, x_{i2}, \dots, x_{ik})$ a kolik bodů je potřeba pro seriózní odhad regresních koeficientů.

Uvážíme-li, že jak volba bodů x_{ij} , tak jejich počet mají rozhodující význam pro kvalitu odhadu, jsou obě otázky klíčové. Další zpracování těchto údajů je už rutinní záležitostí.

Uvedenou problematikou se zabývá plánování experimentů.

Intervaly spolehlivosti

Pro parametry $\beta_1, \dots, \beta_k$, hledáme intervaly, ve kterých lze očekávat tyto parametry s pravděpodobností $1-\alpha$; vypočítáme je pomocí vztahu:

$$\langle b_i - t_{n-k}(\alpha) \cdot s(b_i), b_i + t_{n-k}(\alpha) \cdot s(b_i) \rangle \quad (16)$$

kde:

b_i = odhad parametru β_i

$s(b_i)$ = směrodatná odchylka odhadu

$t_{n-k}(\alpha)$ = kritická hodnota Studentova rozdělení

n = počet bodů

k = počet parametrů

α = hladina významnosti

Při hladině významnosti α je stupeň spolehlivosti $1 - \alpha$. S touto pravděpodobností se nachází neznámý parametr β_i v intervalu (16).

Např. pro b_2 a $\alpha = 0,05$

$$(0,60 - 2,365 \cdot 0,2267; 0,60 + 2,365 \cdot 0,2267) \quad (17)$$

Test modelu (ANOVA)

Tímto testem se prověřuje, je-li možné vyjádřit Y jako lineární kombinaci vybraných proměnných. Zamítnutí H_0 znamená, že aspoň jedna z proměnných přispívá významně v modelování Y .

1.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

resp. ve tvaru

$$H_0 : \vec{\beta} = \vec{0} \text{ a } H_1 : \vec{\beta} \neq \vec{0}$$

2. Testovací kritérium

$$T = \frac{S_{\hat{Y}}/(k-1)}{S_e/(n-k)}, \quad S_{\hat{Y}} = \sum_1^n (\hat{Y}_i - \bar{Y})^2$$

3. Kritická hodnota

$$F_{k-1, n-k}(\alpha)$$

4. Je-li $T > F_{k-1, n-k}(\alpha)$, zamítá se H_0 .

Zde

$$T = \frac{0,594/(3-1)}{0,006/(10-3)} = 346,5$$

$$F_{3-1, 10-3}(0,05) = 4,73$$

Protože T překročilo kritickou hodnotu, zamítá se H_0 což znamená, že alespoň jedno β_i je nenulové, Y lze modelovat pomocí příslušné proměnné x_i a model se považuje za vyhovující.

Predikce

Jedním z hlavních poslání regresních modelů je odhad očekávaného výstupu Y_i pro zvolené vstupy x_i , tedy predikce Y_i .

Např. predikce v bodě $X_0 = (1, x_{1,0}, x_{2,0}) = (1, 1,8, 4,3)$ se vypočítá

$$\hat{Y}_0 = \hat{Y}(X_0) = \hat{Y}(1, x_{1,0}, x_{2,0}) = -0,78 + 0,60x_2 + 0,90x_3 = -0,78 + 0,60 \cdot 1,8 + 0,90 \cdot 4,3 = 4,17$$

Maticově

$$\hat{Y}_0 = (1,1.8,4.3) \begin{pmatrix} -0,78 \\ 0,60 \\ 0,90 \end{pmatrix} = 4,17$$

Predikce pro dva body

$$\hat{Y}_o = \begin{pmatrix} 1 & 1,8 & 4,3 \\ 1 & 1,9 & 4,4 \end{pmatrix} \begin{pmatrix} -0,78 \\ 0,60 \\ 0,90 \end{pmatrix} = \begin{pmatrix} 4,17 \\ 4,32 \end{pmatrix} = \begin{pmatrix} \hat{Y}_{1,0} \\ \hat{Y}_{2,0} \end{pmatrix}$$

Obecně

$$\hat{Y}_0 = X_0 b$$

Rozptyl predikce (pro dva body)

$$\text{var } \hat{Y}_0 = X_0 \text{var}(b) X_0' = X_0 \frac{Se}{n-k} (X'X)^{-1} X_0' = \begin{pmatrix} s^2(Y_{1,0}) & \\ & s^2(Y_{2,0}) \end{pmatrix}$$

Intervalový odhad predikce

Predikované výstupy jsou bodovým odhadem. Je možný také intervalový odhad středních hodnot $E(\hat{Y}_{1,0}), E(\hat{Y}_{2,0})$. K jejich výpočtu je potřeba mj. také rozptyl predikce.

Např.

$$s^2(\hat{Y}_0) = \frac{0,006}{10-3} (1,1.8,4.3) \begin{pmatrix} 245,2 & 108 & -103 \\ 108 & 60 & -50 \\ -103 & -50 & 45 \end{pmatrix} \begin{pmatrix} 1 \\ 1,8 \\ 4,3 \end{pmatrix} = 5,5711 \cdot 10^{-4}$$

$$s(\hat{Y}_0) = 0,023603$$

$$E(\hat{Y}_{1,0}) \in < \hat{Y}_{1,0} - t_{n-k}(\alpha) \cdot s(\hat{Y}_{1,0}); \hat{Y}_{1,0} + t_{n-k}(\alpha) \cdot s(\hat{Y}_{1,0}) >$$

$$E(\hat{Y}_{1,0}) \in < 4,17 - 2,365 \cdot 0,023603; 4,17 + 2,365 \cdot 0,023603 > = < 4,12; 4,23 >$$

$$t_{10-3}(0,05) = 2,365 \text{ (Excel).}$$

Příklad 1 – pokračování

Má se predikovat Y se Statgraphics pro $x_1 = 1.8$, $x_2 = 4.3$ a $x_1 = 1.9$, $x_2 = 4.4$

Řešení:

- zapsat do dalších řádků (= řádky 11 a 12) tyto proměnné (Y chybí),
- najít model (počítá jen s úplnými řádky 1 až 10),
- ikona : [Tabular Options/Reports](#): výsledky jsou na obrazovce:

Regression Results for Y

Row	Fitted Value	Std. Error for Forecast	Lower 95,0% CL for Forecast	Upper 95,0% CL for Forecast	Lower 95,0% CL for Mean	Upper 95,0% CL for Mean
11	4,17	0,037607	4,08107	4,25893	4,11419	4,22581
12	4,32	0,0414039	4,2221	4,4179	4,25077	4,38923

V tabulce jsou také v posledních dvou sloupcích intervalové odhady $E(Y)$.

Vlivné body

Podobně jako u jednorozměrného souboru, také v regresní analýze se mohou vyskytovat body, které se výrazně liší od ostatních, zde konkrétně s extrémní y-onovou resp. x-ovou souřadnicí. Souhrnně se nazývají vlivné body.

Jsou dvě cesty, jak se vyhnout problémům s vlivnými body

1. Použít robustní metody, tj. metody, které nejsou ovlivněny vlivnými body
2. Odhalit, vyhodnotit a odstranit vlivné body

To je možné dvěma způsoby:

- a) testem
- b) grafickými metodami

Oba přístupy používají speciální matici $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ jejíž prvky na hlavní diagonále h_{ii} se určí pro i-tý řádek

$$h_{ii} = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T$$

$\mathbf{x}_i$ je i-tý řádek v $\mathbf{X}$.

i	h_{ii}
1	0,4
2	0,25
3	0,4
4	0,2
5	0,25
6	0,25
7	0,2
8	0,4
9	0,25
10 suma	0,4 3,0 = k

Velikost h_{ii} závisí

- a) na plánu
- b) na modelu

Vlastnosti h_{ii}

a) $0 \leq h_{ii} \leq 1$

b) $\sum_1^n h_{ii} = k$

k = počet parametrů

h_{ii} (lverage): ukazuje velikost vlivu Y_i na $\hat{Y}_i$; protože $\sum h_{ii} = k$, je průměr prvků na hlavní diagonále $\sum h_{ii} / n = k/n$; přesáhnutí dvojnásobku průměru, tj. $2k/n$ (k je počet parametrů) se považuje za významné.

Je tedy průměrné $h_{ii} = \text{počet parametrů} / \text{počet pokusů}$

Pro identifikaci vlivných bodů jsou dále potřebná různě modifikovaná rezidua e_i

- a) **normovaná** (též standardizovaná) rezidua e_i/σ : má přibližně $\sigma=1$, proto má být v int. $(-3,3)$. Za σ se dosadí odhad s .

Pozn.: σ je směrodatná odchylka *teoretických* reziduí ε_i ; u e_i je poněkud jiná – viz bod c)

- b) **studentizovaná rezidua** r_i : (Internally Studentized Residua)

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sigma \sqrt{1 - h_{ii}}};$$

r_i má být v intervalu $(-3,3)$.

Pozn.: $s(e_i)$ se vypočítá takto:

$$\bar{e} = Y - \hat{Y} = Y - Xb = Y - X.(X'X)^{-1}X'Y = Y - HY = (I - H)Y$$

$$\text{var}(\bar{e}) = (I - H) \text{var} Y (I - H)' = (I - H)\sigma^2 I$$

protože $(I-H)$ je symetrická a idempotentní.

Potom $D(e_i) = (1 - h_{ii})\sigma^2$ takže $s(e_i) = \sigma \sqrt{1 - h_{ii}}$.

Jestliže se vynechá i -tý pokus (řádek matice X) nazýváme r_i Externally Studentized Residua:

$$r_i = \frac{e_i}{s_{-i} \sqrt{1 - h_{ii}}}$$

c) **Cookova vzdálenost** D_i : má být menší než 1 proto, že $F_{p,n-p}(0,5) \approx 1$

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{(1 - h_{ii})}$$

d) **predikovaná rezidua** $e_{(i)}$, **PRESS** (prediction error sum of squares)

Vynecháme v matici Y (a také X) i -tý řádek, $i = 1, 2, \dots, n$ a vypočítáme $\hat{Y}$.

$$\text{Potom } e_{(i)} = Y_i - \hat{Y}_{(i)}$$

Např.: $e_{(1)} = Y_1 - \hat{Y}_{(1)}$, kde $\hat{Y}_{(1)}$ bylo počítáno bez bodu $(\mathbf{x}_1, Y_1)$

Charakteristika PRESS pro posouzení vlivných bodů

$$PRESS = \sum e_{(i)}^2 = \sum \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Je-li v datech vlivný bod, projeví se to velkou odchylkou e_i a $e_{(i)}$ a tedy velkou hodnotou PRESS.

Velkému h_{ii} odpovídá i velká hodnota PRESS.

PRESS se používá k predikované korelaci

$$R_{PRED}^2 = 1 - \frac{\sum e_{(i)}^2}{S_Y} = 1 - \frac{PRESS}{S_Y}$$

Jako ukázkou zmiňovaných robustních metod uvedeme výpočet regresních koeficientů pro přímkou.

Příklad 2 (robustní odhad přímky)

Zvolíme několik bodů na přímce $Y = a + b \cdot x$, kde $a = 0$ a $b = 2$:

x	1	2	3	4	5	6	7	8	9	10
Y	2(200)	4	6	8	10	12	14	16	18	20

Postupně hledejte koeficienty regresní přímky pro zadané hodnoty a pro úmyslně chybný údaj $Y = 200$ (modře) a to klasicky a robustní metodou odhadu. Výsledky porovnejte. Vzhledem k tomu, že rovnice hledané přímky je známa, je možné ihned posoudit správnost výpočtů.

Řešení:

a) MNČ budeme nyní k těmto bodům hledat a, b a ověříme, je-li $a = 0$, $b = 2$:

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix} = \begin{pmatrix} 10 & 55 \\ 55 & 385 \end{pmatrix}$$

$$\det(X^T X) = 3850 - 3025 = 825$$

$$(X^T X)^{-1} = \frac{1}{825} \begin{pmatrix} 385 & -55 \\ -55 & 10 \end{pmatrix}, \quad (X^T Y) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ 6 \\ 8 \\ 10 \\ 12 \\ 14 \\ 16 \\ 18 \\ 20 \end{pmatrix} = \begin{pmatrix} 110 \\ 770 \end{pmatrix}$$

$$b = \frac{1}{825} \begin{pmatrix} 385 & -55 \\ -55 & 10 \end{pmatrix} \begin{pmatrix} 110 \\ 770 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

b) Změňme $Y_1 = 2.00$ na chybný údaj $Y_1 = 200$. MNČ (Statgraphics) nyní nalezne $a = 79.2$, $b = -8.8$

c) **Robustní** odhad koeficientů ze *správných* dat bude:

Kvartily $x_{25} = 3$, $x_{50} = 5.5$ a $x_{75} = 8$; $y_{25} = 6$, $y_{50} = 11$ a $y_{75} = 16$. Koeficienty

$$b = \frac{y_{75} - y_{25}}{x_{75} - x_{25}} = \frac{16 - 6}{8 - 3} = 2; \quad a = y_{50} - b \cdot x_{50} = 11 - 2 \cdot 5.5 = 0$$

Pro správná data vychází tedy shodně s předchozím výpočtem.

c) robustní odhad z *chybných* dat:

nové uspořádání Y pro výpočet kvartilů:

4	6	8	10	12	14	16	18	20	200
---	---	---	----	----	----	----	----	----	-----

Kvartily: $x_{25} = 3$, $x_{50} = 5,5$ a $x_{75} = 8$; $y_{25} = 8$, $y_{50} = 13$ a $y_{75} = 18$.

Koeficienty

$$b = (y_{75} - y_{25}) / (x_{75} - x_{25}) = (18 - 8) / (8 - 3) = 2, \quad a = y_{50} - b \cdot x_{50} = 13 - 2 \cdot 5,5 = 2$$

Robustní výpočet je podstatně méně ovlivněn chybným údajem než v případě b).

Na závěr této kapitoly uvádíme přehled nejdůležitějších vzorců regresní analýzy.

PŘEHLED VZORCŮ PRO REGRESNÍ ANALÝZU

1. $\vec{b} = (\mathbf{X}^T \mathbf{X})^{-1} \cdot \mathbf{X}^T \mathbf{Y}$
2. $\hat{\mathbf{Y}} = \mathbf{X} \cdot \vec{b}$
3. $\vec{e} = \mathbf{Y} - \hat{\mathbf{Y}}$
4. $S_e = \sum_i (\mathbf{Y}_i - \hat{\mathbf{Y}})^2 = \sum_i \mathbf{e}_i^2, \quad S_{\hat{\mathbf{Y}}} = \sum_i (\hat{\mathbf{Y}}_i - \bar{\mathbf{Y}})^2$
5. $\text{var}(\vec{b}) = \frac{\sum_i \mathbf{e}_i^2}{n - k} \cdot (\mathbf{X}^T \mathbf{X})^{-1}$
6. $T_i = \frac{b_i}{s(b_i)} \sim t_{n-k}$
7. $T = \frac{\frac{S_{\hat{\mathbf{Y}}}}{k-1}}{\frac{S_e}{n-k}} \sim F_{k-1, n-k}$
8. $\hat{\mathbf{Y}}_0 = \mathbf{X}_0 \vec{b}$
9. $s^2(\hat{\mathbf{Y}}_0) = \frac{S_e}{n-k} \cdot \mathbf{X}_0 (\mathbf{X}^T \mathbf{X})^{-1} \cdot \mathbf{X}_0^T$
10. $\hat{\mathbf{Y}}_0 \pm t_{n-k}(\alpha) \cdot s(\hat{\mathbf{Y}}_0)$
11. $\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \mathbf{X})^{-1} \cdot \mathbf{X}^T$
 $\mathbf{H}_{ii} = \vec{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_i^T$

Vzorce, uvedené v přehledu, je možné použít jen jsou-li splněny předpoklady regresní analýzy z kapitoly 1.1.2.

V této souvislosti se řeší celá řada přirozených otázek, zejména

- a) jak lze ověřit splnění uvedených předpokladů
- b) jak postupovat v regresní analýze, nejsou-li splněny
- c) co způsobí použití vzorců 1-11 při nesplnění předpokladů

Těmito otázkami se budeme zabývat v následující kapitole.

11.1.4 Testování předpokladů regresní analýzy

Předpoklady o reziduích jsou vysloveny pro *teoretická* rezidua ε_i , která jsou však nedostupná. Je tedy potřeba ověřit alespoň vlastnosti *empirických* reziduů e_i . Ta však získáme až po nalezení regresní funkce, takže vzorec pro nalezení vektoru $\vec{b}$ je nutné použít dříve, než budou předpoklady ověřeny. Teprve potom se ověří. Jsou-li splněny, nalezená regresní funkce se ponechá, nejsou-li splněny, je nutné buď upravit data tak, aby byl problém odstraněn, nebo použít jiné postupy, které jsou schopné problém překonat bez úpravy dat.

Vraťte se ke kapitole 11.1.2, připomeňte si předpoklady a v dalším textu se zamyslete nad způsobem jejich ověření.

Předpoklad č. 1 se ověří snadno: je-li součet reziduů roven nule, je také jejich průměr roven nule a podmínka je splněna.

Nesplnění předpokladu č. 2 se nazývá **heteroskedasticita**. Nesplnění znamená, že $D(\varepsilon_i) = \sigma_i^2$ nebo jinak, $D(\varepsilon_i) \neq \sigma^2$.

Následky heteroskedasticity:

odhad je nestranný a konzistentní, ale není eficientní.

Testování reziduů na heteroskedasticitu se provádí mnoha způsoby a existuje také grafický projev heteroskedasticity.

Test 1 (párový koeficient korelace)

Nulová hypotéza

$$H_0: \rho(|\varepsilon|, t) = 0$$

$$H_1: \rho(|\varepsilon|, t) \neq 0$$

Testovacím kritériem je koeficient korelace mezi absolutní hodnotou reziduů a časem měření:

$$r(|\varepsilon|, t).$$

Jeho významnost je potřeba také otestovat pomocí kritéria

$$t = \frac{|r| \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

(18)

s kritickou hodnotou $t_{n-2}(\alpha)$

Závěr celého testu heteroskedasticity: je-li $t < t_{n-2}(\alpha)$ je korelace nevýznamná a H_0 platí. Heteroskedasticita se tedy nepotvrdila.

Příklad 3 (Statg, data [Nowak](#)):

- Nalezněte regresní funkci $y = a + b \cdot x$ a rezidua
- Znázorněte v grafu body $[t, \text{abs}(e_t)]$
- Vypočítejte $r(|e|, t)$ a otestujte významnost r

Řešení: $\hat{Y} = 36,8 + 0,34$

Rezidua a jejich absolutní hodnoty jsou v tabulce. Je vidět nárůst reziduí v čase.

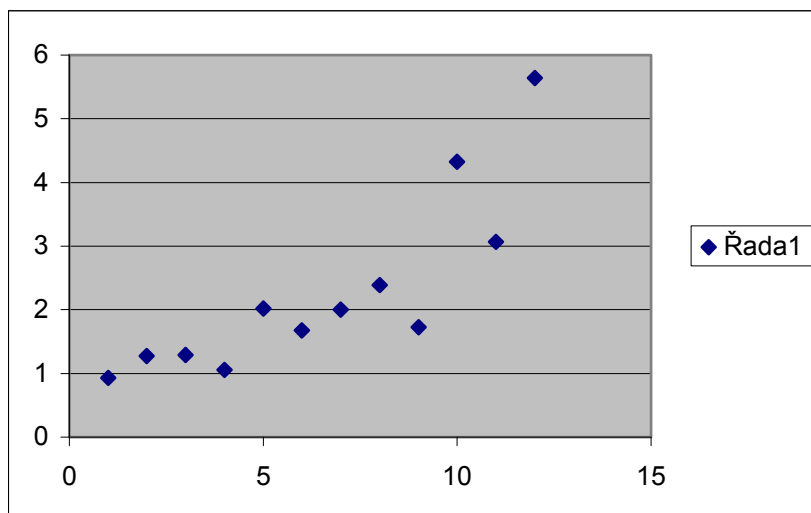
t	y_t	x_t	e_t	ABS(e)
1	43	21	-0,93	0,93
2	43	22	-1,27	1,27
3	44	25	-1,29	1,29
4	43	23	-1,06	1,06
5	48	27	2,02	2,02
6	46	32	-1,68	1,68
7	49	30	2,00	2,00
8	47	37	-2,39	2,39
9	48	38	-1,73	1,73
10	52	32	4,32	4,32
11	47	39	-3,07	3,07
12	54	34	5,64	5,64

Jak se projevuje heteroskedasticita v grafu reziduí:

Osa x: čas t ; osa y: absolutní hodnoty reziduálních odchylek

Excel: ikona/XYbodový/č.1/Dokončit

Data: modré sloupky tabulky.



$r(|e|, t) = 0,83$

Excel: [Nástroje/Analýza dat/Korelace](#)

Test korelačního koeficientu (není v Excelu):

$$t = \frac{0,83 \cdot \sqrt{12-2}}{\sqrt{1-0,83^2}} = 4,71$$

$t_{12-2}(0,05) = 2,228139$ (Excel: [fx/Statistické/TINV](#)]

$t > t_{10}(0,05) \Rightarrow H_0$ se zamítá, tj. rozptyl není konstantní

Je-li y závislé na více proměnných, tj. $y = f(x_1, x_2, \dots, x_k)$, nezjistí se tímto testem, která proměnná je zdrojem heteroskedasticity. To umožní test GQ.

Test 2 (Goldfeld-Quandt, GQ)

Základní kroky testu

Vynechat p prostředních hodnot (je-li $n = 30$, je doporučená volba $p = 8 \Rightarrow p = (n/30) \cdot 8$); p upravit tak, aby obě části byly stejné.

Nalézt regresní funkci pro 1. a 2. část, rezidua u obou funkcí, stanovit $\sum e_1^2$ (= součet reziduí 1. poloviny), $\sum e_2^2$ (= součet reziduí 2. poloviny)

3. Testovací kritérium

$$T = \frac{\sum e_2^2}{\sum e_1^2}$$

4. Kritická hodnota $F_{G,G}(\alpha)$, kde $G = (n - p - 2k) / 2$

k = počet parametrů v regresní funkci

p = počet vynechaných hodnot

5. $T > F_{G,G}(\alpha) \Rightarrow$ zamítá se H_0 (H_0 : rozptyl je konstantní)

Výhodou testu G-Q je, že odhalí, která proměnná je zdrojem heteroskedasticity.

Příklad 4

Proveďte test G-Q pro data z příkladu 3.

Řešení:

Pro test G-Q se data uspořádají podle x :

zadání		uspořádat dle x		vynechat p = 4	
y	x	y	x	y	x
43	21	43	21	43	21
43	22	43	22	43	22
44	25	43	23	43	23
43	23	44	25	44	25
48	27	48	27		
46	32	49	30		
49	30	46	32		
47	37	52	32		
48	38	54	34	54	34
52	32	47	37	47	37
47	39	48	38	48	38
54	34	47	39	47	39

Hlavní kroky testu:

a) najít 2 regresní funkce a příslušné součty čtverců reziduí:

$$\sum_1^4 e_1^2 = 0,17; \sum_9^{12} e_2^2 = 5,42$$

b) vypočítat testovací kritérium $T = 5,42 / 0,17 = 31,88$;

c) kritická hodnota* $F_{2,2}(0,05) = 19$ (v Excelu: [fx/Statistické/FINV](#))

d) závěr testu: $T = 31,88 > 19$; jedná se o heteroskedasticitu

Příklad 5 (data Koutsoyianis)

Pro daný soubor proveďte test G-Q podle vzoru v příkladě 4).

Regresní funkce má tvar $S = a + bx$.

S	x	S(pokrač.)	x
90	9954	1578	24127
131	10508	1654	25604
122	10979	1400	26500
107	11912	1829	27670
406	12747	2200	28300
503	13499	2017	27430
431	14269	2105	29560
588	15522	1600	28150
898	16730	2250	32100
950	17663	2420	32500
779	18575	2570	35250
819	19635	1720	33500
1222	21163	1900	36000
1702	22880	2100	36200
		2300	38200

Odstranění heteroskedasticity

- a) Předpokládáme-li, že $D(\varepsilon_t) = c \cdot x_{kt}^2$, je potřeba najít proměnnou x_k způsobující heteroskedasticitu.

Např. je-li $Y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + \varepsilon_t$ a $D(\varepsilon_t) = c \cdot x_{2t}^2$, $t = 1, 2, \dots, n$

Pak úprava rovnice (dat) bude

$$\frac{Y_t}{x_{2t}} = \beta_1 \frac{x_{1t}}{x_{2t}} + \beta_2 + \beta_3 \frac{x_{3t}}{x_{2t}} + \dots + \beta_k \frac{x_{kt}}{x_{2t}} + \frac{\varepsilon_t}{x_{2t}}$$

Tato úprava vede k odstranění heteroskedasticity, neboť

$$D\left(\frac{\varepsilon_t}{x_{2t}}\right) = \frac{1}{x_{2t}^2} D(\varepsilon_t) = c$$

Příklad 6 (data Nowak; Excel)

Odstraňte heteroskedasticitu v příkladě 3.

Řešení:

Excel: nejprve data seřadit podle x, poté upravit

odstranění H	
y/x	1/x
2,04761905	0,047619
1,95454545	0,045455
1,86956522	0,043478
1,76	0,04
1,77777778	0,037037
1,63333333	0,033333
1,4375	0,03125
1,625	0,03125
1,58823529	0,029412
1,27027027	0,027027
1,26315789	0,026316
1,20512821	0,025641

Výpočet koeficientů pro upravená data

$$Y/x = 0,42 + 1/x \cdot 34,38 \cdot x$$

Odtud

$$Y = 0,42x + 34,38$$

Pro srovnání regresní funkce pro původní data má rovnici

$$Y = 0,34x + 36,7$$

Nesplnění předpokladu č. 3 se nazývá **autokorelace**.

V případě autokorelace je $\text{cov}(\varepsilon_t, \varepsilon_{t-1}) \neq 0$

Autokorelace se projeví v grafu reziduí jejich systematickým průběhem v čase. Naopak není-li autokorelace v reziduích, kolísají náhodně hodnoty reziduí kolem hodnoty 0.

Příčiny autokorelace mohou být různé, např.

- a) Povaha dat: např. u časových řad s trendem vykazují trend také rezidua
- b) Vynechání proměnné v modelu
- c) Existence endogenní zpožděné proměnné na místě regresoru
- d) Úpravou dat
- e) Chybná volba typu regresní funkce

Následky autokorelace jsou stejné jako u heteroskedasticity:

Odhad je nestranný a konzistentní, ale není eficientní.

Autokorelaci lze odhalit testem Durbin-Watson.

Test Durbin-Watson

1. $H_0: \text{cov}(\varepsilon_t, \varepsilon_{t-1}) = 0$ (tj. $\rho = 0$)

2a. $r_1 = r(e_t, e_{t-1})$

2b.

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

3. Kritická hodnota z tabulek: d_L, d_U

4. Je-li $d < d_L \Rightarrow H_0$ se zamítá
 $d > d_U \Rightarrow H_0$ se přijímá
 $d_L < d < d_U$ nelze tímto testem rozhodnout.

Za testovací kritérium pak použijeme

$$t = \frac{|r_1| \cdot \sqrt{n-1-2}}{\sqrt{1-r_1^2}} \sim t_{n-1-2}$$

Pro $r_1 < 0$ bereme místo d testovací kritérium: $d' = 4 - d$

Příklad 7

Na základě uvedených údajů proveďte diskusi o autokorelaci reziduí.

$r_1 = -0,417$; $d = 2,61$; $n = 20$, $\alpha = 0,05$; $d_L = 1,2$, $d_U = 1,41$.

Řešení:

Protože $r_1 < 0$, bereme $d' = 4 - d = 1,39$

Jelikož $1,2 < d' < 1,41$, nelze pomocí d' rozhodnout.

Použijeme proto hodnotu r . Testem ověříme významnost

$$t = \frac{|-0,417| \cdot \sqrt{20-1-2}}{\sqrt{1-(-0,417)^2}} = 1,89;$$

$t_{20-1-2}(0,05) = 2,1 \Rightarrow H_0$ se přijímá, tj. $\rho = 0$.

Úprava testovacího kritéria DW na tvar, umožňující jeho snadný odhad

$$d = \frac{\sum_1^n (e_t - e_{t-1})^2}{\sum_1^n e_t^2} = \frac{\sum_1^n (e_t^2 - 2e_t e_{t-1} + e_{t-1}^2)}{\sum_1^n e_t^2} = \frac{\sum_1^n e_t^2 - 2 \sum_1^n e_t e_{t-1} + \sum_1^n e_{t-1}^2}{\sum_1^n e_t^2}$$

$$\sum_2^n e_{t-1}^2 \approx \sum_2^n e_t^2 \approx \sum_1^n e_t^2 \Rightarrow d = \frac{2 \sum_1^n e_t^2 - 2 \sum_1^n e_t e_{t-1}}{\sum_1^n e_t^2} \approx 2(1 - \rho_1)$$

Pozn.:

$$\rho_1 = \frac{\text{cov}(e_t, e_{t-1})}{s_e^2} = \frac{\frac{1}{n} \sum_2^n (e_t - \bar{e})(e_{t-1} - \bar{e})}{\frac{1}{n} \sum_1^n (e_t - \bar{e})^2}$$

Odstranění autokorelace

Mějme regresní funkci

$$Y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + \varepsilon_t$$

(a)

V případě autokorelace se předpokládá, že

$$\varepsilon_t = \rho \cdot \varepsilon_{t-1} + a_t \quad (a_t \dots \text{nekorelované s } a_{t-1})$$

Rovnici (a) napíšeme pro čas (t-1) a vynásobíme ρ :

$$Y_{t-1} = \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \dots + \beta_k x_{k,t-1} + \varepsilon_{t-1} \quad / \cdot \rho$$

(b)

Od rovnice (a) odečteme rovnici (b)

$$Y_t - \rho \cdot Y_{t-1} = \beta_1 (x_{1t} - \rho \cdot x_{1,t-1}) + \beta_2 (x_{2t} - \rho \cdot x_{2,t-1}) + \dots + (\varepsilon_t - \rho \cdot \varepsilon_{t-1})$$

$$t = 2, 3, \dots, n$$

Poslední člen je reziduální odchylka: $(\varepsilon_t - \rho \cdot \varepsilon_{t-1}) = a_t$

a_t je podle předpokladu nekorelované s a_{t-1} .

Za ρ je zjednodušeně možno zvolit $\rho = 1$, takže úprava dat bude spočívat jen ve výpočtu diferencí.

Při výpočtu máme k dispozici místo teoretické hodnoty ρ její odhad r .

Úprava dat tedy bude:

$$Y_2^* = Y_2 - r_1 Y_1, \quad x_{21}^* = x_{21} - r_1 x_{11}, \quad x_{31}^* = x_{31} - r_1 x_{21}, \text{ atd.}$$

ve všech sloupcích.

V prvním řádku, kde nelze provést předchozí úpravu, je transformace dat následující:

$$Y_1^* = Y_1 \sqrt{1 - r_1^2}, \quad x_{11}^* = x_{11} \sqrt{1 - r_1^2}, \quad x_{12}^* = x_{12} \sqrt{1 - r_1^2}$$

Příklad 8 (data Koutsoyiannis)

Sleduje se závislost $z = a + bx$. Provéřte rezidua na autokorelaci

- nalezením regresní funkce ve Statgraphics, jehož součástí je i kritérium DW
- výpočtem testovacího kritéria v Excelu (rezidua jsou dána)

Řešení:

Statgraphics

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	3,26457E7	1	3,26457E7	1025,40	0,0000
Residual	573069,0	18	31837,2		
Total (Corr.)	3,32188E7	19			

Correlation Coefficient = 0,991337

R-squared = 98,2749 percent

R-squared (adjusted for d.f.) = 98,179 percent

Standard Error of Est. = 178,43

Mean absolute error = 128,84

Durbin-Watson statistic = 0,937394 (P=0,0018)

Lag 1 residual autocorrelation = 0,347367

z	x	e	$r = e_t - e_{t-1}$	r^2	e^2
3748	21777	122,1739			14884
4010	22418	204,9988	83	6889	42025
3711	22308	-63,2535	-268	71824	3969
4004	23319	-52,8526	10	100	2809
4151	24180	-146,523	-94	8836	21609
4569	24893	72,17611	219	47961	5184
4582	25310	-31,3855	-103	10609	961
4697	25799	-53,0729	-22	484	2809
4753	25886	-21,3916	32	1024	441
5062	26868	13,11555	34	1156	169
5669	28134	266,2377	253	64009	70756
5628	29091	-42,2671	-308	94864	1764
5736	29450	-34,6163	7	49	1225
5946	30705	-175,419	-140	19600	30625
6501	32372	-86,3864	89	7921	7396
6549	33152	-256,415	-170	28900	65536
6705	33764	-271,484	-15	225	73441
7104	34411	-53,3365	218	47524	2809
7609	35429	167,1077	220	48400	27889
8100	36200	442,5944	276	76176	196249
				536551	572550
				dw	0,937125

Příklad 9 (data Alonso)**Sleduje se závislost $y = a + bx$. Provéřte rezidua na autokorelaci**

- nalezením regresní funkce ve Statgraphics, jehož součástí je i kritérium DW
- výpočtem testovacího kritéria v Excelu (rezidua jsou dána)

Řešení:

Statgraphics

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	8,71999	1	8,71999	1017,17	0,0000
Residual	0,128592	15	0,0085728		
Total (Corr.)	8,84858	16			

Correlation Coefficient = 0,992707

R-squared = 98,5468 percent

R-squared (adjusted for d.f.) = 98,4499 percent

Standard Error of Est. = 0,0925894

Mean absolute error = 0,0773324

Durbin-Watson statistic = 0,634431 (P=0,0001)

Lag 1 residual autocorrelation = 0,57101

Excel

y	x	e	(e _t -e _{t-1}) ²	e ²
6,656	6,557	0,115862		0,013424
6,692	6,599	0,053194	0,003927	0,00283
6,667	6,643	-0,07517	0,016478	0,005651
6,709	6,624	0,011464	0,007506	0,000131
6,825	6,647	0,073432	0,00384	0,005392
7,091	6,759	0,076319	8,34E-06	0,005825
7,183	6,848	-0,04076	0,013708	0,001662
7,306	6,932	-0,1151	0,005526	0,013247
7,518	6,992	-0,04405	0,005048	0,00194
7,666	7,061	-0,05815	0,000199	0,003381
7,762	7,139	-0,14539	0,007611	0,021137
7,896	7,180	-0,1077	0,00142	0,0116
8,077	7,236	-0,05826	0,002445	0,003394
8,294	7,309	-0,01275	0,002071	0,000163
8,507	7,367	0,063993	0,00589	0,004095
8,688	7,412	0,139278	0,005668	0,019398
8,844	7,485	0,123784	0,00024	0,015322
			0,081583	0,128592
			DW	0,634431

Pro zjednodušení transformace dat se volí $r = 1$, takže se počítají jen difference.

Často vede ke zlepšení ještě jednodušší úprava: na pravou stranu se zařadí $y(-1)$, tj. zpožděná endogenní proměnná. Při kontrole autokorelace upravených dat (na pravé straně je zpožděná proměnná) se musí provést místo DW test Durbin

$$h = \left(1 - \frac{DW}{2}\right) \sqrt{\frac{n}{1 - n \cdot s^2(b_i)}}$$

$s^2(b_i)$ je rozptyl koeficientu endogenní zpožděné proměnné

Příklad 10 (Statg.; data Nowak)

Proveďte úpravu dat vedoucí k odstranění autokorelace; je dáno $r_1 = -0,403$

Pozn. Ve Statgraphics :

Lag 1 residual autocorrelation = $-0,418766$

Řešení:

Původní data (x,y), upravená data (y^* , x^*)

Upravená data		Data	
y^*	x^*	y	x
15,47	33,4	16,9	36,5
26,71	53,81	19,9	39,1
27,42	59,86	19,4	44,1
27,72	63,27	19,9	45,5
26,52	67,34	18,5	49
27,56	76,15	20,1	56,4
29,3	89,13	21,2	66,4
30,44	107,16	21,9	80,9
33,03	125,92	24,2	93,4
33,75	147,14	24	109,5
32,87	167,73	23,2	123,6
35,85	181,41	26,5	131,6
35,78	202,92	25,1	149,1
39,72	217,69	29,6	157,6
43,63	237,11	31,7	173,6
41,08	251,87	28,3	181,9
42,7	266,61	31,3	193,3
41,51	267,48	28,9	189
44,15	266,47	32,5	190,3
40,1	265,59	27	188,9

Předpoklad č. 4 o normalitě reziduí lze prověřit pohodlně grafickými metodami, například v normálním pravděpodobnostním grafu.

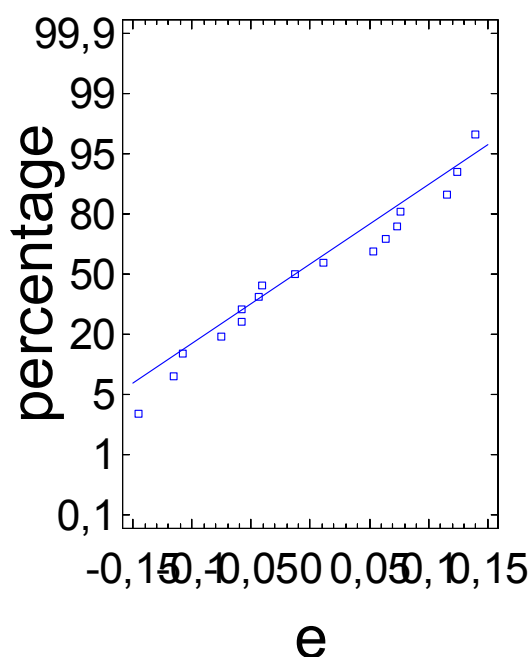
Tak např. pro data (x,y) v tabulce byla nalezena regresní přímka a reziduální odchylky:

y	x	e
6,656	6,557	0,115862
6,692	6,599	0,053194
6,667	6,643	-0,07517
6,709	6,624	0,011464
6,825	6,647	0,073432
7,091	6,759	0,076319
7,183	6,848	-0,04076
7,306	6,932	-0,1151
7,518	6,992	-0,04405
7,666	7,061	-0,05815
7,762	7,139	-0,14539
7,896	7,180	-0,1077

8,077	7,236	-0,05826
8,294	7,309	-0,01275
8,507	7,367	0,063993
8,688	7,412	0,139278
8,844	7,485	0,123784

Príslušný normální pravděpodobnostní graf

Normal Probability Plot



Pokud body „dobře přiléhají“ k přímce, mají data N rozdělení.

Předpoklad č. 5 o tom, že matice X je nenáhodná, tedy že data v ní uložená nejsou zatížena náhodnými chybami, se neprověřuje.

Předpoklad č. 6 o lineární nezávislosti sloupců v matici X je velmi závažný a měl by se testovat vždy. Nesplnění předpokladu č. 6 se nazývá **multikolinearita**.

Rozlišuje se:

a) multikolinearita přesná: jedna exogenní proměnná je **přesně** lineární kombinací ostatních.

Pak $X^T X$ je singulární $\Rightarrow$ detekce je snadná: výpočet neproběhne

b) multikolinearita zdánlivá: jedna exogenní proměnná je **přibližně** lineární kombinací ostatních

Důsledek: extrémně velký rozptyl odhadu β_i

Jednou z možností odhalení multikolinearity je použití testu modelu, který je dostupný vždy. Jeho základní verze má tvar

1. $H_0: \vec{\beta} = \vec{0}$
 $H_1: \vec{\beta} \neq \vec{0}$
2. $T = \frac{S_{\bar{y}}/(k-1)}{S_e/(n-k)}$
3. $F_{k-1, n-k}(\alpha)$
4. Je-li $T > F_{k-1, n-k}(\alpha)$, zamítá se H_0

Dosadíme-li do tohoto testu za závisle proměnnou postupně argumenty $x_1, x_2, \dots, x_k$ bude testem prověřeno, zda je možné modelovat příslušné x_i pomocí zbývajících. Je-li to možné, pak je v matici multikolinearita.

Typickým projevem multikolinearity je, že při testu významnosti koeficientů ($H_0: \beta_i = 0$) platí H_0 (= všechny koeficienty jsou nulové) a současně v testu modelu H_0 neplatí ($H_0: \vec{\beta} \neq \vec{0}$).

Příklad 11 (projev multikolinearity) (data Alonso)

Sleduje se závislost výdajů (Y) na příjmech (x_2) a úsporách (x_3).

Nalezněte regresní funkci, proveďte test koeficientů a test modelu a výsledky porovnejte (projev multikolinearity).

Y	32	11	15	17	16	13	18	20	14	17	41	17	33	20	18
x_2	36	12	16	18	17	14	20	23	15	18	50	19	37	22	19
x_3	144	47	63	70	67	52	79	90	58	70	204	76	149	86	76

Uvažuje se regresní funkce: $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$

Řešení:

Statgraphics

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	1,54174	0,82426	1,87045	0,0860
x_2	1,40731	0,724192	1,94329	0,0758
x_3	-0,145742	0,174853	-0,833515	0,4208

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	990,048	2	495,024	1044,85	0,0000
Residual	5,6853	12	0,473775		

Excel

ANOVA						
	<i>Rozdíl</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Významnost F</i>	
Regrese	2	990,048	495,024	1044,851	3,46466E-14	
Rezidua	12	5,685297	0,473775			
Celkem	14	995,7333				

	<i>Koeficienty</i>	<i>Chyba stř. hodnoty</i>	<i>t stat</i>	<i>Hodnota P</i>	<i>Dolní 95%</i>	<i>Horní 95%</i>
Hranice Soubor X	1,54174	0,82426	1,870453	0,086001	0,254168305	3,337649
1 Soubor X	1,407312	0,724192	1,943287	0,075809	-0,17056609	2,98519
2	-0,14574	0,174853	-0,83352	0,420836	-0,52671324	0,235229

S poznatků uvedených dosud o regresní analýze je patrné, že výpočet regresních koeficientů je rutinní záležitostí, avšak provede se až po sérii testů, které se týkají zejména

- předpokladů regresní analýzy
- výsledků měření Y_i
- regresních koeficientů
- regresního modelu

11.1.5 Zobecněná metoda nejmenších čtverců (GLS)

Při nesplnění druhého resp. třetího předpokladu regresní analýzy je možné postupovat dvěma způsoby

- odstranit heteroskedasticitu resp. autokorelaci postupem, uvedeným v kap. 11.1.4 a potom použít metodu nejmenších čtverců (LS) z kap. 11.1.1
- neupravovat data, tj. ponechat je s problémem heteroskedasticity resp. autokorelace a použít zobecněnou metodu nejmenších čtverců (GLS).

Zobecněné metodě nejmenších čtverců je věnována tato kapitola.

Klasický model lineární regrese má tvar

$$\mathbf{Y} = \mathbf{X} \cdot \vec{\beta} + \vec{\epsilon}, \text{var}(\vec{\epsilon}) = \sigma^2 \cdot \mathbf{I}$$

Zobecněný model lineární regrese (Aitkenův) je vyjádřen ve tvaru

$$\mathbf{Y} = \mathbf{X} \cdot \vec{\beta} + \vec{\epsilon}, \text{var}(\vec{\epsilon}) = \sigma^2 \cdot \mathbf{V}, \mathbf{V} \neq \mathbf{I},$$

$\mathbf{V}$ je pozitivně definitní matice, z čehož plynou dále potřebné vlastnosti

$$a) \quad \mathbf{V}^{-1} = \mathbf{P}^T \cdot \mathbf{P}$$

$$b) \quad \mathbf{PVP}^T = \mathbf{I}$$

Vynásobením klasického modelu maticí $\mathbf{P}$ a po výpočtu varianční matice nových (transformovaných) reziduí dostáváme

$$\mathbf{P} \cdot \mathbf{Y} = \mathbf{P} \cdot \mathbf{X} \cdot \vec{\beta} + \mathbf{P} \cdot \vec{\epsilon}$$

$$\text{var}(\mathbf{P} \cdot \vec{\epsilon}) = \mathbf{P} \cdot \text{var}(\vec{\epsilon}) \cdot \mathbf{P}^T$$

$$\text{var}(\mathbf{P} \cdot \vec{\epsilon}) = \mathbf{P} \cdot \sigma^2 \cdot \mathbf{V} \cdot \mathbf{P}^T = \sigma^2 \cdot \mathbf{I}$$

Označíme-li

$$\mathbf{P} \cdot \mathbf{Y} = \mathbf{Y}^*, \quad \mathbf{P} \cdot \mathbf{X} = \mathbf{X}^*, \quad \mathbf{P} \cdot \vec{\epsilon} = \vec{\epsilon}^*,$$

pak můžeme psát

$$\mathbf{Y}^* = \mathbf{X}^* \cdot \vec{\beta} + \vec{\epsilon}^*$$

Při odhadu $\vec{\beta}$ v této rovnici je možné postupovat dvěma způsoby:

a) transformovat data (tj. vypočítat $\mathbf{X}^*$, $\mathbf{Y}^*$) a použít LS:

$$\vec{\beta} = (\mathbf{X}^{*T} \cdot \mathbf{X}^*)^{-1} \cdot \mathbf{X}^{*T} \cdot \mathbf{Y}^{*T};$$

Zde je potřeba matice $\mathbf{P}$.

b) ponechat původní data a použít GLS, kterou získáme po dosazení za $\mathbf{X}^*$, $\mathbf{Y}^*$

$$\vec{\beta} = [(\mathbf{PX})^T \cdot (\mathbf{PX})]^{-1} \cdot (\mathbf{PX})^T \cdot \mathbf{PY} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \cdot \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y} = \mathbf{GLS};$$

Zde je potřeba matice $\mathbf{V}$.

V dalším se soustředíme na sestavení matice $\mathbf{P}$ resp. $\mathbf{V}$. Matice $\mathbf{V}$ má různý tvar pro heteroskedasticitu a pro autokorelaci.

Pro heteroskedasticitu, kdy $D(\epsilon_i) \neq \sigma^2$, máme místo $\mathbf{I}$ matici $\mathbf{V} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ 0 & 0 & \sigma_3^2 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix};$

Potom inverzní matice k $\mathbf{V}$ a matice $\mathbf{P}$ zde budou

$$\mathbf{V}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & 0 & \dots \\ 0 & 0 & \frac{1}{\sigma_3^2} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2} & 0 & \dots \\ 0 & 0 & \frac{1}{\sigma_3} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Pro autokorelaci, kdy $\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0$, máme místo $\mathbf{I}$ matici

$$\mathbf{V} = \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} & \dots \\ \sigma_{21} & \sigma^2 & \sigma_{23} & \dots \\ \sigma_{31} & \sigma_{32} & \sigma^2 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma^2} & \frac{\sigma_{13}}{\sigma^2} & \dots \\ \frac{\sigma_{21}}{\sigma^2} & 1 & \frac{\sigma_{23}}{\sigma^2} & \dots \\ \frac{\sigma_{31}}{\sigma^2} & \frac{\sigma_{32}}{\sigma^2} & 1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \stackrel{(1)}{=} \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots \\ \rho_1 & 1 & \rho_1 & \dots \\ \rho_{21} & \rho_1 & 1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \stackrel{(2)}{=} \begin{bmatrix} 1 & \rho_1 & \rho_1^2 & \dots \\ \rho_1 & 1 & \rho_1 & \dots \\ \rho_1^2 & \rho_1 & 1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Inverzní matice k $\mathbf{V}$ a matice $\mathbf{P}$ zde budou

$$\mathbf{V}^{-1} = \frac{1}{1 - \rho_1^2} \cdot \begin{bmatrix} 1 & -\rho_1 & 0 & \dots \\ -\rho_1 & 1 + \rho_1^2 & -\rho_1 & \dots \\ 0 & -\rho_1 & 1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}, \quad \mathbf{P} = \frac{1}{\sqrt{1 - \rho^2}} \begin{bmatrix} \sqrt{1 - \rho_1^2} & 0 & 0 & \dots \\ -\rho_1 & 1 & 0 & \dots \\ 0 & -\rho_1 & 1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Například pro matici $\mathbf{V}_{3 \times 3}$ bude v případě heteroskedasticity počítána matice $\mathbf{V}^{-1}$ v těchto krocích:

$$\mathbf{V} = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix},$$

$$|\mathbf{V}| = 1 + \rho^4 + \rho^4 + \rho^4 - \rho^2 - \rho^2 = 1 - 2 \cdot \rho^2 + \rho^4 = (1 - \rho^2)^2$$

$$\begin{aligned} A_{11} &= 1 - \rho^2 \\ A_{12} &= \rho - \rho^3 \\ A_{13} &= \rho^2 - \rho^2 = 0 \end{aligned}$$

$$\begin{aligned} A_{21} &= \rho - \rho^3 \\ A_{22} &= 1 - \rho^4 \\ A_{23} &= \rho - \rho^3 \end{aligned}$$

$$\begin{aligned} A_{31} &= \rho_1^2 - \rho_1^2 = 0 \\ A_{32} &= \rho - \rho^3 \\ A_{33} &= 1 - \rho^2 \end{aligned}$$

$$\mathbf{V}^{-1} = \frac{1}{(1 - \rho^2)^2} \cdot \begin{bmatrix} 1 - \rho^2 & -(\rho - \rho^3) & 0 \\ -(\rho - \rho^3) & 1 - \rho^4 & -(\rho - \rho^3) \\ 0 & -(\rho - \rho^3) & 1 - \rho^2 \end{bmatrix} = \frac{1}{1 - \rho^2} \cdot \begin{bmatrix} 1 & -\rho & 0 \\ -\rho & 1 + \rho^2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix}$$

$$\mathbf{P}^T \cdot \mathbf{P} = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} \sqrt{1-\rho^2} & -\rho & 0 \\ 0 & 1 & -\rho \\ 0 & 0 & 1 \end{bmatrix} \cdot \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 \\ -\rho & 1 & 0 \\ 0 & -\rho & 1 \end{bmatrix} =$$

$$= \frac{1}{1-\rho^2} \begin{bmatrix} 1-\rho^2+\rho^2 & -\rho & 0 \\ -\rho & 1+\rho^2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix} = \mathbf{V}^{-1}$$

Metoda GLS zdánlivě překoná problém heteroskedasticity resp. autokorelace bez potřebné transformace dat, jak byla uvedena v kap.11.1.4. Porovnáme-li ovšem kroky GLS s transformací dat je zřejmé, že v GLS proběhne stejná úprava dat. Porovnáme nyní obě metody na příkladě.

Porovnání výsledku dat po úpravě rovnice a po násobení maticí P

- pro [heteroskedasticitu](#)

Nechť $D(\varepsilon_t) = k \cdot x_{t2}^2$

Nejprve úprava dat:

$$Y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} \quad / \cdot 1/x_{t2}, t = 1, 2, 3, \dots, n$$

$$\frac{Y_t}{x_{t2}} = \beta_1 \cdot \frac{x_{t1}}{x_{t2}} + \beta_2 \cdot 1 + \dots + \beta_k \cdot \frac{x_{tk}}{x_{t2}}$$

$$\begin{aligned} t=1: \quad \frac{Y_1}{x_{12}} &= \beta_1 \cdot \frac{x_{11}}{x_{12}} + \beta_2 + \dots + \beta_k \cdot \frac{x_{1k}}{x_{12}} \\ t=2: \quad \frac{Y_2}{x_{22}} &= \beta_1 \cdot \frac{x_{21}}{x_{22}} + \beta_2 + \dots + \beta_k \cdot \frac{x_{2k}}{x_{22}} \\ &\dots \dots \dots \\ t=n: \quad \frac{Y_n}{x_{n2}} &= \beta_1 \cdot \frac{x_{n1}}{x_{n2}} + \beta_2 + \dots + \beta_k \cdot \frac{x_{nk}}{x_{n2}} \end{aligned}$$

Vstupní data mají tvar

i	x ₁	x₂	...	x _k	Y _i
1	x ₁₁	x ₁₂	...	x _{1k}	Y ₁
2	x ₂₁	x ₂₂	...	x _{2k}	Y ₂
3	x ₃₁	x ₃₂	...	x _{3k}	Y ₃
...	...	...	...	...	...
n	x _{n1}	x _{n2}	...	x _{nk}	Y _n

Červeně je označen proměnná, jež je příčinou heteroskedasticity.

Nyní provedeme násobení matice X vektorem P :

$$X^* = P \cdot X = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ x_{12} & 1 & 0 & \dots & 0 \\ 0 & x_{22} & 1 & \dots & 0 \\ 0 & 0 & x_{32} & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{1}{x_{n2}} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ x_{31} & x_{32} & \dots & x_{3k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ x_{12} & x_{12} & x_{12} & \dots & x_{12} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ x_{22} & x_{22} & x_{22} & \dots & x_{22} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \\ x_{n2} & x_{n2} & x_{n2} & \dots & x_{n2} \end{bmatrix}$$

Oba způsoby dávají tytéž výsledky.

- pro **autokorelaci** dostáváme pro Y :

$$Y^* = P \cdot Y = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \cdot \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \end{bmatrix} = \begin{bmatrix} Y_1 \cdot \sqrt{1-\rho^2} & \dots & Y_1^* \\ -\rho \cdot Y_1 + Y_2 & \dots & Y_2^* \\ -\rho \cdot Y_2 + Y_3 & \dots & Y_3^* \\ \dots & \dots & \dots \end{bmatrix}$$

což je totéž jako při výpočtu diferencí.

Podobně $X^* = PX$:

$$X^* = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 \\ -\rho & 1 & 0 \\ 0 & -\rho & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} = \begin{bmatrix} x_{11}\sqrt{1-\rho^2} & x_{12}\sqrt{1-\rho^2} & x_{13}\sqrt{1-\rho^2} \\ -\rho x_{11} + x_{21} & -\rho x_{12} + x_{22} & -\rho x_{13} + x_{23} \\ -\rho x_{21} + x_{31} & -\rho x_{22} + x_{32} & -\rho x_{23} + x_{33} \end{bmatrix} \dots \begin{matrix} x_1^* \\ x_2^* \\ x_3^* \end{matrix}$$

Počínaje druhým řádkem jsou v matici X^* v každém sloupci rozdíly 2, 3, ..., n-tého řádku s ρ -násobkem předchozího. Z prvního řádku je nyní vidět, proč byla v kapitole *Odstranění autokorelace* provedena tato úprava.



Shrnutí kapitoly 11.1

V této kapitole byly uvedeny základní pojmy regresní analýzy, prezentováno osm nejdůležitějších vět regresní analýzy včetně předpokladů, za kterých lze uvedené věty použít.

Ke všem předpokladům jsou také uvedeny způsoby jejich prověření, důsledky nesplnění předpokladů a způsob úpravy dat, který vede k odstranění problému s nesplněním nejzávažnějších předpokladů č. 2, 3 a 5. Byla ukázána souvislost DOE s problematikou regresní analýzy.



Otázky ke kapitole 11.1

1. Napište teoretickou a empirickou regresní funkci. V čem se liší?
2. Jak se značí empirické a teoretické reziduální odchylky?
3. Znázorněte v obrázku empirickou hodnotu závisle proměnné, příslušnou teoretickou hodnotu a reziduální odchylku.
4. Napište soustavu regresních rovnic v rozepsaném a maticovém tvaru
5. Napište matici X
6. Vyjádřete MNČ v rozepsaném tvaru a pomocí vektorů
7. Napište předpoklady regresní analýzy
8. Napište předpoklady regresní analýzy v maticovém vyjádření
9. Napište a dokažte větu 1 o výpočtu regresních koeficientů b_i .
10. Napište a dokažte větu 2 o vlastnostech odhadu regresních koeficientů
11. Napište pomocné věty 3 – 6. K čemu slouží?
12. Napište a dokažte větu 7. Které věty použijete k důkazu?
13. K čemu slouží věta 7?
14. Napište větu 8. K čemu slouží?
15. Jak provedete intervalový odhad β_i ?
16. Uveďte dva způsoby výpočtu teoretických hodnot. Jak se značí teoretické hodnoty?
17. Napište všechny body testu regresních koeficientů
18. Napište v maticovém tvaru vztah pro výpočet predikce.
19. Jak se počítá rozptyl predikce?
20. Jak se počítá matice H ?
21. Jak se počítá i -tý řádek matice H ?
22. K čemu slouží matice H ?
23. Uveďte dvě vlastnosti prvku h_{ii}
24. Jak se počítají normovaná a studentizovaná rezidua?
25. Jak se počítá $D(e_i)$? Dokažte.
26. Jak se počítá robustní odhad regresních koeficientů v regresní přímce?
27. Co je heteroskedasticita?
28. Napište všechny body testu Goldfeld Quandt
29. Jak se počítá a upravuje p v testu Goldfeld Quandt?
30. Jak se počítá G v testu Goldfeld Quandt?
31. Jak se odhalí heteroskedasticita pomocí r ?
32. Jak se projeví heteroskedasticita v grafu reziduí?
33. Co způsobí heteroskedasticita?
34. Jak se odstraní heteroskedasticita?
35. Co je autokorelace, jak se projev v grafu reziduí?
36. Napište všechny body testu Durbin Watson
37. Upravte testovací kritérium testu DW na tvar $DW \approx 2(1 - r)$
38. V jakých mezích se vyskytuje DW a proč?
39. Jaká hodnota DW je ideální?
40. Co způsobuje autokorelace?

41. Jak se odstraní autokorelace?
42. Napište maticově klasický a zobecněný regresní model
43. Proveďte odvození metody GLS
44. Ukažte, že pro $V = I$ je $GLS = LS$
45. Nalezněte matici V^{-1} a P pro heteroskedasticitu
46. Nalezněte matici V^{-1} a P pro autokorelaci

11.2 Časové řady



Cíl kapitoly

- a) vysvětlit základní pojmy teorie časových
- b) ukázat metody modelování časových řad klasickou metodikou a Box - Jenkinsovou metodikou



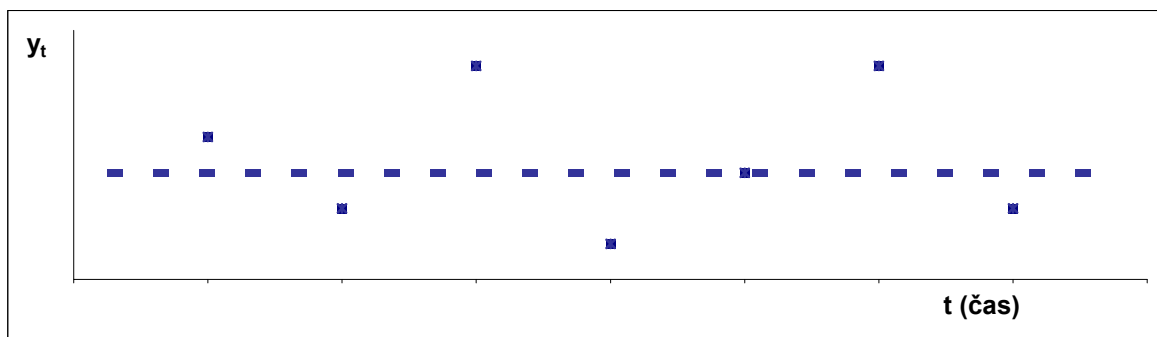
Výklad

Po nalezení modelu, a po provedení výše uvedených testů, jsou pak možné jeho různé aplikace. Regresní modely se v ekonomických aplikacích používají zejména proto, že umožňují predikci (extrapolaci) vývoje sledovaných ukazatelů. V technických aplikacích slouží ovšem také pro potřeby interpolace, regulace procesů apod.

Pokud jde o predikci, je další často používanou cestou využití časových řad. Jelikož při hledání modelů časových řad je významnou součástí také regresní analýza, uvádíme tuto problematiku v rámci širších souvislostí DOE v této kapitole.

Časová řada (ČŘ) je posloupnost hodnot uspořádaných v čase od nejstarších k nejnovějším. Značí se $\{y_t\}, \{x_t\}$ atd.

Graf ČŘ:



U ČŘ se zejména

- a) počítají charakteristiky ČŘ
- b) hledá model ČŘ

Jsou dvě základní strategie studia časových řad:

- a) klasická metodika
- b) Box - Jenkinsova metodika

V obou se řeší výše uvedené dvě úlohy, avšak zcela odlišným přístupem.

Časové řady	
Klasická metodika	Box Jenkinsova metodika
1. Charakteristiky	1. Charakteristiky
2. Model: $y_t = T_t + C_t + S_t + e_t$	2. Modely: AR, MA, ARMA
Nalezení trendové složky je možné	Další modely např. ARIMA, SARIMA aj.
a) MNČ	
b) modifikací MNČ	
c) jinak než s použitím MNČ	
Nalezení C_t a S_t : stejně	

Studium ekonomických aplikací časových řad a regresní analýzy (RA) a to i za situace nesplnění předpokladů klasické regresní analýzy (modifikovaná RA) je předmětem disciplíny, která se nazývá ekonometrie.

Ekonometrie	
Časové řady	Regresní analýza (modifikovaná)

Ta užívá poněkud jinou terminologii a symboliku pro proměnné.

Můžeme říci, že proměnné v ekonometrii mají, ve srovnání s matematikou, tyto dvě odlišnosti

- a) Opatřují se časovým indexem, tzn. sleduje se, z jakého pochází času.
Např. x_t, x_{t-1}, y_{t-2}
- b) Jsou jinak pojmenované:
 - 1. endogenní (= závisle proměnná): y_t resp. y_{it} např. $y_{1t}, y_{2t}, \dots$
 - 2. predeterminované (= nezávisle proměnná)
 - a) exogenní: $x_{it}, x_{it-1},$
 - b) endogenní zpožděné $y_{i,t-1}, y_{i,t-2}, \dots$ např. $y_{1,t-1}, y_{2,t-2},$ atd.

11.2.1 Klasická metodika analýzy časových řad

Charakteristiky časové řady

Popisné charakteristiky

1. Průměr ČŘ $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$

Časová řada, která nevykazuje v čase trend, ale kolísá náhodně kolem své střední hodnoty, se nazývá *stacionární podle střední hodnoty*.

Charakteristika $\bar{y}$ má smysl jen u stacionárních ČŘ.

$$2. \text{ Rozptyl ČŘ } s_y^2 = \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y})^2 \quad (20)$$

Časová řada, která nevykazuje v čase *trend rozptylu* (zvětšování nebo zmenšování), se nazývá *stacionární podle rozptylu*.

Přítomnost obou stacionarit v ČŘ se posuzuje nejčastěji vizuálně.

Dynamické charakteristiky

$$3. \text{ Absolutní přírůstek: } \Delta y_t = y_t - y_{t-1} \quad (21)$$

4. Koefficient růstu:

$$K_t = \frac{y_t}{y_{t-1}} \quad (22)$$

Například pro obecné údaje bude

Rok	y_t	Δy_t	K_t
1985	y_1		
1986	y_2	$y_2 - y_1$	y_2 / y_1
1987	y_3	$y_3 - y_2$	y_3 / y_2
1988	y_4	$y_4 - y_3$	y_4 / y_3
1989	y_5	$y_5 - y_4$	y_5 / y_4

Průměrný koefficient růstu

$$\bar{k} = \sqrt[4]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \cdot \frac{y_5}{y_4}} = \sqrt[4]{\frac{y_5}{y_1}} \quad (23)$$

Obecně

$$\bar{k} = \sqrt[n-1]{\frac{y_n}{y_1}} \quad (24)$$

Průměrný absolutní přírůstek

$$\bar{\Delta} = \frac{(y_2 - y_1) + (y_3 - y_2) + (y_4 - y_3) + (y_5 - y_4)}{4} = \frac{y_5 - y_1}{4} \quad (25)$$

Obecně

$$\bar{\Delta} = \frac{y_n - y_1}{n - 1} \quad (26)$$

Příklad 12

Pro dané údaje vypočítat průměrný absolutní přírůstek a průměrný koeficient růstu.

Rok	těžba y_t	Δy_t	K_t
1985	1286		
1986	1363	77	1,059
1987	1393	30	1,022
1988	1495	102	1,073
1989	1571	76	1,051
1990	1610	39	1,025

Řešení:

$$\bar{\Delta} = \frac{1610 - 1286}{5} = 64,8 \Rightarrow \text{za 6 let rostla těžba průměrně o 64,8}$$

$$\bar{k} = \sqrt[5]{\frac{1610}{1286}} = 1,045 \Rightarrow \text{v průměru rostla těžba o 4,5\%}$$

Modely časové řady

Modely slouží zejména k predikci (prognóze, předpovědi) vývoje sledovaného znaku y .

Při modelování ČŘ se u klasické metodiky analýzy řad vychází z předpokladu, že ČŘ může obsahovat 4 druhy pohybu:

$$y_t = T_t + C_t + S_t + \varepsilon_t$$

T_t = trend; jeho určení je zásadní úkol v analýze ČŘ

C_t = cyklická složka = dlouhodobé pravidelné kolísání okolo trendu
(jeho sledování vyžaduje dlouhou ČŘ)

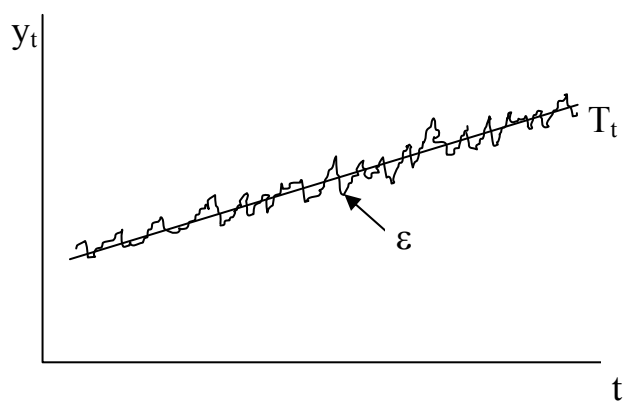
S_t = sezónní složka = kolísání s periodou menší než 1 rok

ε_t = náhodné kolísání (tato složka je vždy)

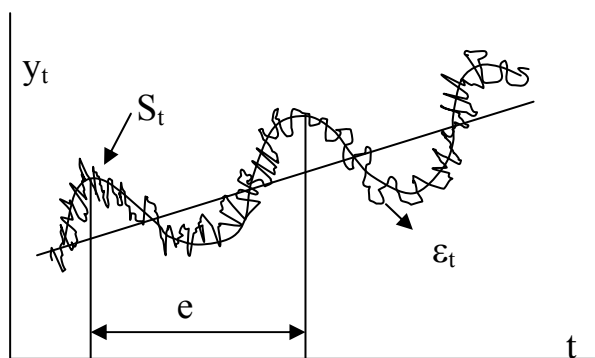
Následující příklad ukazuje, jak se projeví v grafu ČŘ nepřítomnost některé ze čtyř složek. Poznamenejme ovšem, že náhodná složky musí být přítomna vždy, má-li se jednat o ČŘ.

Příklady průběhu časových řad, když chybí některá složka rozkladu y_t :

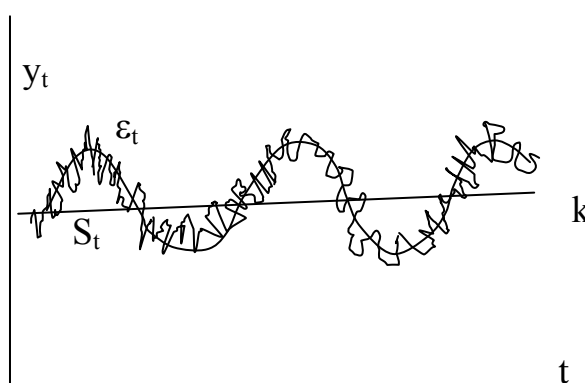
1. $y_t = T_t + \varepsilon_t$



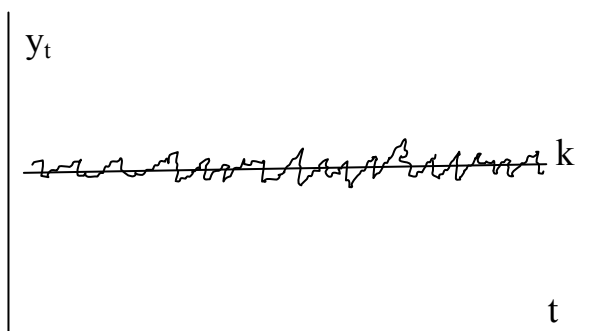
2. $y_t = T_t + S_t + \varepsilon_t$



3. $y_t = k + S_t + \varepsilon_t$ (nemá trend)



$$4. y_t = k + \varepsilon_t$$



Nalezení trendové složky T_t je úloha, která může být řešena řadou metod. Jednou z nich je regresní analýza, která byla probírána v předchozí kapitole a nebude proto dále řešena. Další možnosti jsou speciální postupy, z nichž uvedeme tři: logistický trend, metodu klouzavých průměrů a metodu exponenciálního vyrovnání ČŘ.

a) Logistický trend

$$y = \frac{k}{1 + b_o b_1^t} \quad (27)$$

patří mezi tzv. S křivky.

Použití: modelování poptávky po zboží dlouhodobé spotřeby a modelování prodeje některých druhů výrobků.

Místo MNČ se zde používá **metoda vybraných bodů**:

Pro tři časy t_1, t_2 a t_3 se zjistí y_1, y_2 a y_3 . Tím se získá soustava rovnic

$$y_1 = \frac{k}{1 + b_o b_1^{t_1}}$$

$$y_2 = \frac{k}{1 + b_o b_1^{t_2}}$$

$$y_3 = \frac{k}{1 + b_o b_1^{t_3}}$$

Výhodné je volit časy t_1, t_2 a t_3 takto: $t_1 = 0, t_2 = \Delta$ a $t_3 = 2\Delta$,

kde Δ je určitý časový úsek.

Potom

$$b_o = \frac{k - y_1}{y_1}, \quad b_1 = \left[\frac{y_1(k - y_2)}{y_2(k - y_1)} \right]^{\frac{1}{t_2}}, \quad k = \frac{2y_1y_2y_3 - y_2^2(y_1 + y_3)}{y_1y_3 - y_2^2}$$

b) Klouzavé průměry:

Posloupnost hodnot ČŘ nahradíme novou řadou průměrů, vypočítaných z kratších úseků ČŘ, při čemž tyto úseky postupně posouváme (kloužeme).

Délka úseku, ze kterého se počítá průměr, se u neperiodických ČŘ volí jako lichý počet hodnot y_t , např. 3, 5, 7. Lichý proto, že vypočítaný průměr se přiřadí prostřední hodnotě.

U periodických ČŘ se volí délka klouzavé části shodná s délkou periody.

c) Exponenciální vyrovnaní:

Hodnota y z času t , tzn. y_t , se tzv. vyrovná (vyjádří) pomocí hodnot $y_1, y_2, y_3, \dots, y_t$ tak, že každé hodnotě se přiřadí tzv. váha w . Čím starší je y_t , tím menší váha je jí přiřazena.

Pro první vyrovnanou hodnotu $\hat{y}_1$ se volí různé strategie, např.

- a) položí se rovna y_1 : $\hat{y}_1 = y_1$
- b) za $\hat{y}_1$ se dosadí průměr předchozí části ČŘ

Další vyrovnané hodnoty definujeme rekurentním vztahem

$$\hat{y}_t = wy_t + (1 - w)\hat{y}_{t-1} \quad (28)$$

$t = 2, 3, \dots, n$

Například

$$\hat{y}_2 = wy_2 + (1 - w)\hat{y}_1$$

$$\hat{y}_3 = wy_3 + (1 - w)\hat{y}_2$$

atd.

Jak se dělá predikce u klouzavých průměrů a exp. vyrovnaní:
v obou případech

- a) dělá se jen na 1 krok dopředu
- b) poslední klouzavý průměr resp. exponenciální vyrovnaní je predikce na 1 krok dopředu

$$\hat{y}_t(1) = \hat{y}_t$$

Chyba predikce

$$\hat{a}_{t+1} = y_{t+1} - y_t(1)$$

Příklad 13 (logistický trend)

V tabulce jsou údaje o počtu vyrobených polovodičů v letech 1985 – 95:

t	85	86	87	88	89	90	91	92	93	94	95
t'	0	1	2	3	4	5	6	7	8	9	10
y	5	6	9	16	22	25	32	34	41	44	45

Tendenci vývoje popište logistickou trendovou funkcí

$$y = \frac{k}{1 + \beta_o \beta_1^t} \quad (29)$$

Řešení:

Zavedeme novou časovou proměnnou t' (druhý řádek). Z časové řady vybereme tři hodnoty y_1, y_2 a y_3 , ve kterých položíme teoretické hodnoty rovny empirickým: $y = \hat{y}$.

$y_1(0) = 5$, $y_2(5) = 25$ a $y_3(10) = 45$, kde $t'_1 = 0$, $t'_2 = 5$, $t'_3 = 10$.

Nyní odhadneme parametry trendové funkce

$$k = \frac{2y_1y_2y_3 - y_2^2(y_1 + y_3)}{y_1y_3 - y_2^2} = \frac{2 \cdot 5 \cdot 25 \cdot 45 - 25^2(5 + 45)}{5 \cdot 45 - 25^2} = 50$$

$$b_o = \frac{k - y_1}{y_1} = \frac{50 - 5}{5} = 9$$

$$b_1 = \left[\frac{y_1(k - y_2)}{y_2(k - y_1)} \right]^{\frac{1}{t'_2}} = \left[\frac{5(50 - 25)}{25(50 - 5)} \right]^{\frac{1}{5}} = 0,644$$

Odhad logistického trendu je

$$\hat{y} = \frac{50}{1 + 9 \cdot 0,644^{t'}}$$

Například pro rok 1996 tj. $t' = 11$ bude

$$\hat{y} = \frac{50}{1 + 9.0,644^{11}} = 49,61$$

Příklad 14 (klouzavé průměry, exponenciální vyrovnání)

Údaje o spotřebě pitné vody (y_i) za 3 týdny:

den	y_i		kl.průměry	ex.vyrov.
Po	0,64			0,640
U	0,78			0,738
St	0,93			0,872
Č	0,66		0,896	0,724
Pá	0,99		0,911	0,910
So	1,22		0,890	1,127
N	1,05		0,874	1,073
Po	0,75		0,870	0,847
U	0,63		0,914	0,695
St	0,82		0,833	0,783
Č	0,63		0,869	0,676
Pá	1,30		0,839	1,113
So	0,65		0,836	0,789
N	1,30		0,819	1,147
Po	0,54		0,809	0,722
U	0,61		0,736	0,644
St	0,70		0,829	0,683
Č	0,56		0,820	0,597
Pá	0,79			0,732
So	1,30			1,130
N	1,24			1,207

- Vyrovnat řadu prostými klouzavými průměry
- Vyrovnat ČŘ metodou exponenciálního vyrovnání; volte $w = 0,7$.

Řešení:

- Klouzavé průměry** budou počítány z týdenních hodnot a přiřazeny prostřednímu dni, tj. Čt:

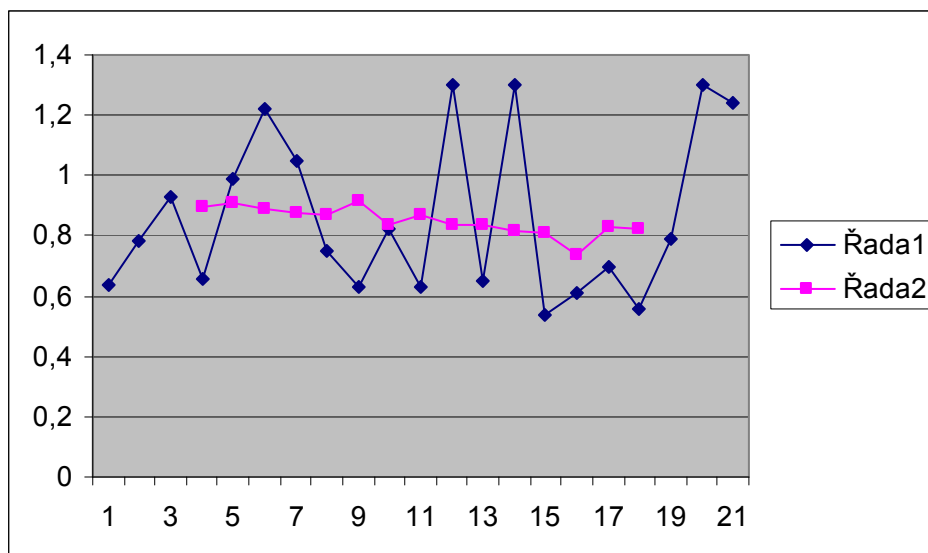
Výpočet (Excel):

$$\bar{y}_1 = \frac{0,64 + 0,78 + 0,93 + 0,66 + 0,99 + 1,22 + 1,05}{7} = 0,896$$

$$\bar{y}_2 = \frac{0,78 + 0,93 + \dots + 1,05 + 0,75}{7} = 0,911$$

atd.

Graf Excel: / Spojnicový (typ č.4)



Výpočet STATGRAPHICS (klouzavé průměry)

1) Nalezení modelu: [Special/ Time Series/ Smoothing](#)

2) Predikce: ikona [Tabular Option/ Forecast Table](#)

Pozn.: změna klouzavé části: pravé tlačítko myši(na graf): [Pane/Option](#)

b) **exponenciální vyrovňání:** $w = 0,7$

Excel:

$$\hat{y}_1 = y_1$$

$$\hat{y}_t = w \cdot y_t + (1 - w) \cdot \hat{y}_{t-1}$$

$t = 2, 3, \dots, n :$

$t = 1: \hat{y}_1 = 0,64$

$t = 2: \hat{y}_2 = 0,7 \cdot y_2 + (1 - 0,7) \cdot \hat{y}_1 = 0,7 \cdot 0,78 + 0,3 \cdot 0,64 = 0,738$

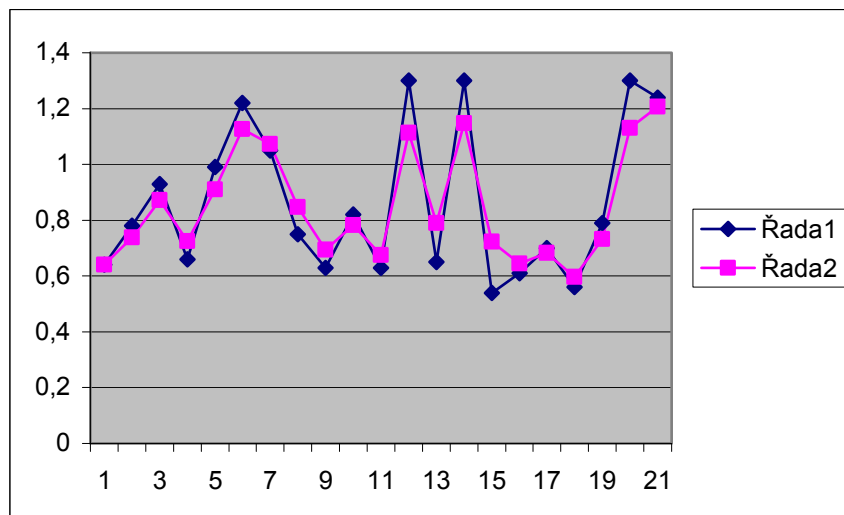
$t = 3: \hat{y}_3 = w \cdot y_3 + (1 - w) \cdot \hat{y}_2 = 0,7 \cdot 0,93 + 0,3 \cdot 0,738 = 0,872$

$t = 4: \hat{y}_4 = w \cdot y_4 + (1 - w) \cdot \hat{y}_3 = 0,7 \cdot 0,66 + 0,3 \cdot 0,872 = 0,724$

atd.

Excel: [Graf / Spojnicový \(typ č.4\)](#)

Řada 1 = y_i , řada 2 = exp. vyrovňání



STATGRAPHICS (exponenciální vyrovňání)

a) [Special/ Time Series/ Forecasting](#)

b) Výběr modelu: Pravé tlačítko myši:

[Analysis Option/ Simple Exp. Smoothing](#)

Regresní metoda modelování **sezónní** složky

Předpokládejme aditivní dekompozici $y_t = T_t + S_t + e_t$

Sezónní složku vyjadřujeme pomocí **umělých nula jedničkových proměnných**, které nabývají jedničku, nachází li se ČŘ v uvažované sezóně a jinak nulu. Např. u dvanáctiměsíční sezónnosti budou pomocné proměnné D_j pro $j = 1, 2, \dots, 12$:

$D_1 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ pro leden,

$D_2 = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ pro únor,

atd.

$D_{12} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$ pro prosinec.

Je-li v modelu absolutní člen, **vynechá se D_1** (nebo jiné D_i) kvůli multikolinearitě (součet v řádcích pro D_i dává jedna = sloupci jedniček u absolutního členu).

Model sezónní složky ČŘ

- bez trendu (a bez absolutního členu):

$$y_t = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \dots + \beta_{12} D_{12} + a_t$$

Pozn.: v tomto případě se nepoužije sloupec t (s časy 1, 2, 3, ...). Mohou být všechna D_i .

- s lineárním trendem bez absolutního členu:

$$y_t = \beta_0 t + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \dots + \beta_{12} D_{12} + a_t$$

Pozn.: mohou být všechna D_i

- s lineárním trendem a absolutním členem: zde je max. s-1 sezón (bez D_1)

$$y_t = \beta_0 + \beta_1 t + \beta_2 D_2 + \beta_3 D_3 + \dots + \beta_{12} D_{12} + a_t$$

Pozn.1: **Nelze** použít všechna D_i !

Pozn.2: Umělé proměnné se v regresní analýze používají např. u

- kvalitativní proměnné (viz DOE)
- složené regresní funkce

Příklad 15

Řada má *čtvrtletní* sezónnost.

Najděte *sezónní* model s trendovou (lineární) částí a s absolutním členem metodou využívající **pomocnou proměnnou**

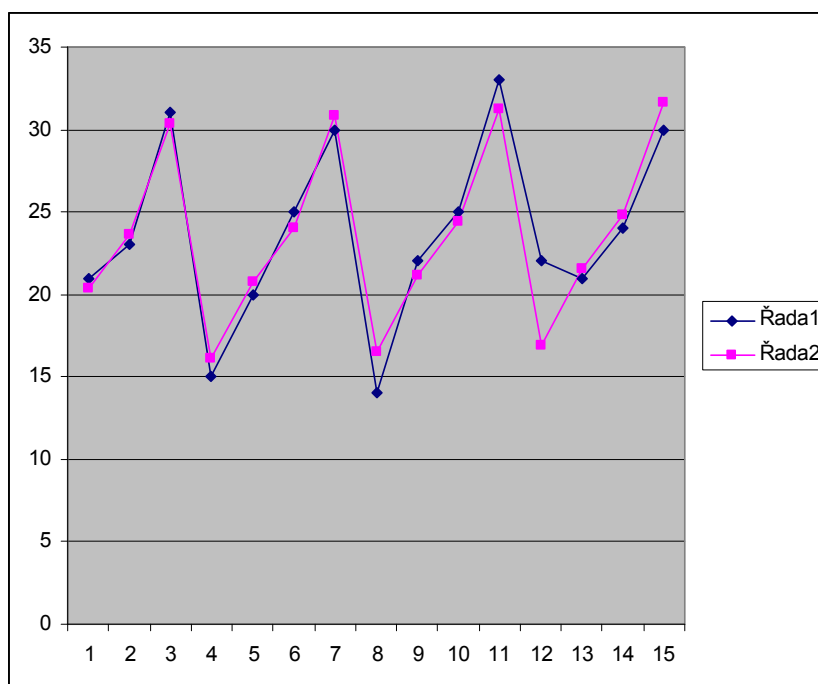
Znázorněte původní (y) a vyrovnanou (yt) č.ř (s [Excelem](#)).

t	y
1	21
2	23
3	31
4	15
5	20
6	25
7	30
8	14
9	22
10	25
11	33
12	22
13	21
14	24
15	30
16	16

Data pro modelování sezónní složky:

Č	as	y	d1	d2	d3	d4	PREDICTED
1	1	21	1	0	0	0	20,4
2	2	23	0	1	0	0	23,65
3	3	31	0	0	1	0	30,4
4	4	15	0	0	0	1	16,15
5	5	20	1	0	0	0	20,8
6	6	25	0	1	0	0	24,05
7	7	30	0	0	1	0	30,8
8	8	14	0	0	0	1	16,55
9	9	22	1	0	0	0	21,2
10	10	25	0	1	0	0	24,45
11	11	33	0	0	1	0	31,2
12	12	22	0	0	0	1	16,95
13	13	21	1	0	0	0	21,6
14	14	24	0	1	0	0	24,85
15	15	30	0	0	1	0	31,6
16	16	16	0	0	0	1	17,35

Graf časové řady a jejího modelu



11.2.2 Box – Jenkinsova metodika analýzy časových řad

Charakteristiky časové řady

Vzájemná kovarianční funkce časových řad x_t a y_t

$$\begin{aligned} c_k &= E \cdot (x_t - \bar{x}) \cdot (y_{t+k} - \bar{y}) \\ c_k &= 1/n \cdot \sum (x_t - \bar{x}) \cdot (y_{t+k} - \bar{y}) \end{aligned} \quad (30)$$

kde střední hodnota

$$\bar{x} = 1/n \cdot \sum x_t \quad (31)$$

Disperze

$$s_x^2 = c_o = 1/n \cdot \sum (x_t - \bar{x})^2$$

Autokovarianční funkce

$$\begin{aligned} c_k &= E (x_t - \bar{x}) \cdot (x_{t+k} - \bar{x}) \\ c_k &= 1/n \cdot \sum (x_t - \bar{x}) \cdot (x_{t+k} - \bar{x}) \end{aligned} \quad (32)$$

Autokorelační funkce

$$r_k = c_k / c_o \quad (33)$$

Vzájemná korelační funkce časových řad $\{x_t\}$ a $\{y_t\}$

$$r_k = c_k / (s_x \cdot s_y) \quad (33a)$$

Rozptyl r_k

Nejdůležitější charakteristikou je bezpochyby autokorelace. Složí k určení typu modelu, řádu modelu i koeficientů v tomto modelu. Počítá se pomocí vzorce 33.

Protože r_k je *bodovým odhadem* teoretické autokorelační funkce ρ_k , má smysl zabývat se rozptylem tohoto odhadu. Odhaduje se

$$s^2(r_k) \approx \frac{1}{n} \left(1 + 2 \sum_{i=1}^{k-1} r_i^2 \right) \quad (33b)$$

Například

$$s^2(r_1) = \frac{1}{n}, \quad s^2(r_2) = \frac{1}{n} [1 + 2r_1^2], \quad s^2(r_3) = \frac{1}{n} [1 + 2r_1^2 + 2r_2^2].$$

Intervalový odhad ρ_k

Vedle bodového odhadu je možný také intervalový odhad ρ_k . Pro 5% hladinu významnosti je

$$P[\rho_k \in (-1,96 \cdot s(r_k); 1,96 \cdot s(r_k))] = 0,95 \quad (33c)$$

Test významnosti r_k :

Vypočítanou hodnotu r_k je nutné vyhodnotit co do významnosti. Použijeme tento test:

1) Nulová hypotéza

$$H_0 : \rho_k = 0$$

$$H_1 : \rho_k \neq 0$$

2) Testovací kritérium $T = \frac{|r_k|}{s(r_k)}$

3) Kritická hodnota pro $\alpha = 0,05$ je 1,96

4) Závěr: pro $T > 1,96$ se zamítá H_0 .

Některé vlastnosti r_k

a) $r_k = r_{-k}$ což znamená, že r_k je sudá funkce

b) $r_0 = 1$, takže všechny autokorelační funkce procházejí bodem 1 na svislé ose.

Box – Jenkinsovy modely časové řady

(V dalším pro jednoduchost předpokládáme, že střední hodnota časové řady je nulová. Toho lze dosáhnout centrováním všech hodnot řady).

Autoregresní model řádu p: AR(p)

Model AR(p) vychází se z předpokladu, že hodnotu y_t lze vyjádřit jako lineární kombinaci hodnot $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ z časů $t-1, t-2, \dots, t-p$:

$$y_t = f_1 \cdot y_{t-1} + f_2 \cdot y_{t-2} + \dots + f_p \cdot y_{t-p} + a_t \quad (34)$$

kde $f_1, f_2, \dots, f_p$ jsou neznámé koeficienty, a_t je náhodná veličina vyrovnávající rozdíl mezi y_t a $f_1 \cdot y_{t-1} + f_2 \cdot y_{t-2} + \dots + f_p \cdot y_{t-p}$, u které předpokládáme tyto vlastnosti:

$$E(a_t) = 0, \text{ var}(a_t) = s_a^2, \text{ cov}(a_t, a_{t+k}) = 0 \text{ pro } k \neq 0 \quad (35)$$

$\{a_t\}$ tvoří rovněž časovou řadu, která bývá v technické praxi nazývána bílým šumem.

Vynásobením vztahu (34) veličinou y_{t-k} dostaneme

$$y_t \cdot y_{t-k} = f_1 \cdot y_{t-1} \cdot y_{t-k} + f_2 \cdot y_{t-2} \cdot y_{t-k} + \dots + f_p \cdot y_{t-p} \cdot y_{t-k} + a_t \cdot y_{t-k} \quad (36a)$$

Vypočítáme-li střední hodnotu tohoto výrazu, máme

$$c_k = f_1 \cdot c_{k-1} + f_2 \cdot c_{k-2} + \dots + f_p \cdot c_{k-p} \quad (36b)$$

a po vydělení c_0

$$r_k = f_1 \cdot r_{k-1} + f_2 \cdot r_{k-2} + \dots + f_p \cdot r_{k-p} \quad (36c)$$

Odtud pro $k = 1, 2, \dots, p$, uvažíme-li, že $r_0 = 1$ a $r_{-k} = r_k$, dostaneme

$$\begin{aligned} r_1 &= f_1 + f_2 \cdot r_1 + \dots + f_p \cdot r_{p-1} \\ r_2 &= f_1 \cdot r_1 + f_2 + \dots + f_p \cdot r_{p-2} \\ &\dots\dots\dots \\ r_p &= f_1 \cdot r_{p-1} + f_2 \cdot r_{p-2} + \dots + f_p \end{aligned} \quad (37)$$

Tyto tzv. Yule-Walkerovy rovnice slouží k výpočtu koeficientů $f_1, f_2, \dots, f_p$ modelu (34), přičemž r_k vypočítáme podle (33).

Pro $k = 0$ máme z (36a)

$$y_t \cdot y_t = f_1 \cdot y_t \cdot y_{t-1} + f_2 \cdot y_t \cdot y_{t-2} + \dots + f_p \cdot y_t \cdot y_{t-p} + a_t \cdot y_t$$

Vypočítáme-li nyní střední hodnotu, přičemž $E(y_t \cdot y_{t-k}) = c_k$ a $E(a_t \cdot y_t) = s_a^2$, máme

$$c_o = s_y^2 = f_1 \cdot c_1 + f_2 \cdot c_2 + \dots + f_p \cdot c_p + s_a^2$$

Po vydělení s_y^2 a úpravě

$$1 - f_1 \cdot r_1 - f_2 \cdot r_2 - \dots - f_p \cdot r_p = \frac{s_a^2}{s_y^2}$$

a odtud

$$s_y^2 = \frac{s_a^2}{1 - f_1 \cdot r_1 - \dots - f_p \cdot r_p}$$

(38)

Tento vztah se spíše využívá k odhadu s_a^2

$$s_a^2 = s_y^2 (1 - f_1 \cdot r_1 - \dots - f_p \cdot r_p)$$

tj.

$$s_a^2 = s_y^2 (1 - r' \cdot f)$$

kde

$$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} \quad \text{a} \quad f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_p \end{bmatrix}$$

Rozptyl s_a^2 se počítá také ze vztahu

$$s_a^2 = \frac{\sum_{t=1}^n a_t^2}{n - p - q}$$

Prakticky nepoužívanějšími AR modely jsou:

Autoregresní model prvního řádu: AR(1)

Rovnice tohoto modelu:

$$y_t = f_1 \cdot y_{t-1} + a_t,$$

nebo také např.

$$y_{t+1} = f_1 \cdot y_t + a_{t+1}$$

$$y_{t+2} = f_1 \cdot y_{t+1} + a_{t+2}, \text{ atd.}$$

Z rovnic (37) pro $p = 1$ máme

$$f_1 = r_1$$

Rozptyl časové řady bude

$$s_y^2 = \frac{s_a^2}{1 - f_1^2}$$

Rozptyl odhadu koeficientu f_1

$$\text{var}(f_1) = s^2(f_1) = n^{-1} \cdot (1 - f_1^2)$$

Je-li y_{t+k} hodnota y z času $t + k$, pak její predikci budeme dále značit $y_t(k)$ a počítáme ji pro tento model

$$y_t(k) = f_1^k \cdot y_t$$

Pro $k = 1$

$$y_t(1) = f_1 \cdot y_t$$

Při odvození této a dalších predikčních rovnic se využívá vztahu $y_t(k) = E(y_{t+k})$, za $E(a_{t+k})$ dosazujeme pro $k > 0$ nulu (neznámé hodnoty s předpokládanou střední hodnotou 0), takže u AR(1) vychází

$$y_t(1) = E(y_{t+1}) = f_1 y_t + E(a_{t+1}) = f_1 \cdot y_t$$

$$y_t(2) = f_1 \cdot y_t(1) = f_1^2 \cdot y_t$$

atd.

$$y_t(k) = f_1^k \cdot y_t$$

Rozptyl chyby predikce e_{t+k} se počítá jako střední hodnota kvadrátu chyby predikce a má u tohoto modelu tvar

$$\text{var}(e_{t+k})^2 = E(y_{t+k} - y_t(k))^2 = \frac{1 - f_1^{2k}}{1 - f_1^2} \cdot s_a^2$$

Pro $k = 1$

$$\text{var}(e_{t+1})^2 = s_a^2$$

Autoregresní model druhého řádu: AR(2)

Rovnice tohoto modelu

$$y_t = f_1 \cdot y_{t-1} + f_2 y_{t-2} + a_t$$

Nebo také

$$y_{t+1} = f_1 \cdot y_t + f_2 y_{t-1} + a_{t+1}, \text{ atp.}$$

Pro výpočet f_1, f_2 dostáváme z (37) pro $p = 1, 2$

$$r_1 = f_1 + f_2 \cdot r_1$$

$$r_2 = f_1 \cdot r_1 + f_2$$

a odtud

$$f_1 = r_1 \cdot \frac{1 - r_2}{1 - r_1^2}, \quad f_2 = \frac{r_2 - r_1^2}{1 - r_1^2}$$

Rozptyl odhadu koeficientů modelu

$$s^2(f_1) = s^2(f_2) = \left(\frac{1 - (f_2)^2}{n} \right)$$

Rozptyl časové řady

$$s_y^2 = \frac{s_a^2}{1 - f_1 \cdot r_1 - f_2 \cdot r_2}$$

Z (19) bude predikční rovnice pro $k = 1$

$$y_t(1) = E(y_{t+1}) = f_1 \cdot y_t + f_2 \cdot y_{t-1} + E(a_{t+1})$$

$$y_t(1) = f_1 \cdot y_t + f_2 \cdot y_{t-1}$$

a $\text{var}(e_{t+1})^2 = s_a^2$

Zavedeme-li operátor B:

$$By_t = y_{t-1}$$

pak

$$y_{t-1} = By_t$$

$$y_{t-2} = By_{t-1} = B(By_t) = B^2 y_t$$

$$y_{t-3} = B^3 y_t$$

...

$$y_{t-p} = B^p y_t$$

S použitím operátoru B můžeme (34) upravit

$$y_t = f_1 \cdot B \cdot y_t + f_2 \cdot B^2 \cdot y_t + f_3 \cdot B^3 \cdot y_t + \dots + f_p \cdot B^p \cdot y_t + a_t$$

Úpravou

$$y_t(1 - f_1 \cdot B - f_2 \cdot B^2 - \dots - f_p \cdot B^p) = a_t$$

dostáváme

$$y_t \cdot f(B) = a_t, \quad f(B) = 1 - f_1 \cdot B - \dots - f_p \cdot B^p$$

Odtud

$$y_t = f^{-1}(B) \cdot a_t$$

Tento vztah vede k myšlence vyjádřit y_t pomocí jiné časové řady $\{a_t\}$ a tedy k modelům označovaným MA.

Model klouzavých průměrů řádu q: MA(q)

Model MA(q) vyjadřuje hodnoty studované časové řady pomocí hodnot jiné řady – bílého šumu.

Má základní rovnici

$$y_t = a_t - t_1 \cdot a_{t-1} - t_2 \cdot a_{t-2} - \dots - t_q \cdot a_{t-q} \quad (39)$$

Kovariance u tohoto modelu bude

$$c_k = E(y_t \cdot y_{t-k}) = E(a_t - t_1 \cdot a_{t-1} - \dots - t_q \cdot a_{t-q}) \times (a_{t-k} - t_1 \cdot a_{t-k-1} - \dots - t_q \cdot a_{t-q-k})$$

Pro $k = 0$ dostáváme disperzi

$$c_0 = (1 + t_1^2 + t_2^2 + \dots + t_q^2) \cdot s_a^2$$

Pro $k = 1, 2, \dots, q$

$$c_k = (-t_k + t_1 \cdot t_{k+1} + t_2 \cdot t_{k+2} + \dots + t_{q-k} \cdot t_q) \cdot s_a^2$$

Autokorelační funkce potom

$$r_k = \frac{c_k}{c_0} = \frac{-t_k + t_1 \cdot t_{k+1} + t_2 \cdot t_{k+2} + \dots + t_{q-k} \cdot t_q}{1 + t_1^2 + t_2^2 + \dots + t_q^2} \quad (40)$$

Nejčastěji používanými modely MA jsou:

MA (1):

$$y_{t+1} = a_{t+1} - t_1 \cdot a_t$$

pro který se neznámý koeficient t_1 počítá ze vztahu (40);

pro $k = 1$ ($q = 1$)

$$r_1 = \frac{-t_1}{1 + t_1^2}$$

a rozptyl jeho odhadu stejně jako u AR modelů

$$s^2(t_1) = n^{-1} \cdot (1 - t_1^2)$$

Rozptyl časové řady

$$s_y^2 = (1 + t_1^2) \cdot s_a^2$$

Predikční rovnice pro $k = 1$

$$y_t(1) = E(a_{t+1}) - t_1 \cdot E(a_t)$$

$$y_t(1) = -t_1 \cdot a_t$$

a

$$\text{var}(e_{t+1})^2 = s_a^2$$

MA (2):

Rovnice

$$y_{t+1} = a_{t+1} - t_1 \cdot a_t - t_2 \cdot a_{t-1}$$

Rozptyl časové řady

$$s_y^2 = (1 + t_1^2 + t_2^2) \cdot s_a^2$$

Z rovnice (40) pro $k = 1$ a $k = 2$ ($q = 2$) dostáváme rovnice

$$r_1 = \frac{-t_1 + t_1 \cdot t_2}{1 + t_1^2 + t_2^2}$$

$$r_2 = \frac{-t_2}{1 + t_1^2 + t_2^2}$$

pro výpočet t_1, t_2 .

Rozptyl odhadu koeficientů t_1 a t_2

$$s^2(t_1) = s^2(t_2) = n^{-1} \cdot (1 - t_2^2)$$

Z rovnice (39) bude pro $k = 1$ predikční rovnice

$$y_t(1) = E(y_{t+1}) = E(a_{t+1}) - t_1 \cdot E(a_t) - t_2 \cdot E(a_{t-1})$$

$$y_t(1) = -t_1 \cdot a_t - t_2 \cdot a_{t-1}$$

a

$$\text{var}(e_{t+1})^2 = s_a^2$$

Kombinovaný model ARMA (p, q):

Tento model je kombinací předchozích dvou. Má rovnici

$$y_t = f_1 \cdot y_{t-1} + f_2 \cdot y_{t-2} + \dots + f_p \cdot y_{t-p} + a_t - t_1 \cdot a_{t-1} - \dots - t_q \cdot a_{t-q}$$

resp. po úpravě

$$y_t - f_1 \cdot y_{t-1} - f_2 \cdot y_{t-2} - \dots - f_p \cdot y_{t-p} = a_t - t_1 \cdot a_{t-1} - t_2 \cdot a_{t-2} - \dots - t_q \cdot a_{t-q}$$

a zavedení operátoru B

$$y_t(1 - f_1 \cdot B - f_2 \cdot B^2 - \dots - f_p \cdot B^p) = a_t(1 - t_1 \cdot B - t_2 \cdot B^2 - \dots - t_q \cdot B^q)$$

tj.

$$y_t \cdot f(B) = a_t \cdot t(B)$$

Prakticky se používá tento model pro $p = q = 1$: ARMA (1, 1)

$$y_{t+1} = f_1 \cdot y_t + a_{t+1} - t_1 \cdot a_t$$

Disperze a kovariance tohoto procesu budou

$$\begin{aligned} c_0 &= E(y_t \cdot y_t) = E(f_1 \cdot y_{t-1} + a_t - t_1 \cdot a_{t-1})^2 = \\ &= f_1^2 \cdot c_0 - 2 \cdot f_1 \cdot t_1 \cdot E(y_{t-1} \cdot a_{t-1}) + s_a^2 + t_1^2 \cdot s_a^2 \end{aligned}$$

Protože

$$E(y_{t-1} \cdot a_{t-1}) = s_a^2$$

máme

$$c_0(1 - f_1^2) = s_a^2(1 + t_1^2 - 2 \cdot f_1 \cdot t_1)$$

odkud

$$c_0 = \frac{1 + t_1^2 - 2 \cdot t_1 \cdot f_1}{1 - f_1^2} \cdot s_a^2$$

Nyní určíme kovariance $c_1, c_2, \dots$

$$\begin{aligned} c_1 &= E(y_{t-1} \cdot (f_1 \cdot y_{t-1} + a_t - t_1 \cdot a_{t-1})) = f_1 \cdot c_0 - t_1 \cdot s_a^2 = \\ &= \frac{(1 - f_1 \cdot t_1)(f_1 - t_1)}{1 - f_1^2} \cdot s_a^2 \end{aligned}$$

$$c_2 = E(y_{t-2} \cdot (f_1 \cdot y_{t-1} + a_t - t_1 \cdot a_{t-1})) = f_1 \cdot c_1$$

a podobně

$$c_k = f_1 \cdot c_{k-1}$$

Autokorelační funkce pro $k = 1$

$$r_1 = \frac{c_1}{c_0} = \frac{(1 - f_1 \cdot t_1)(f_1 - t_1)}{1 + t_1^2 - 2 \cdot f_1 \cdot t_1}$$

a obecně

$$r_k = f_1 \cdot r_{k-1} \quad \text{pro } k = 2, 3, \dots$$

takže

$$r_2 = f_1 \cdot r_1$$

a z rovnic (45), (46) určíme koeficienty f_1 a t_1 .

Rozptyl odhadu je

$$s^2(t_1) = \frac{(1 - t_1^2)(1 - f_1 \cdot t_1)^2}{n \cdot (f_1 - t_1)^2}$$

$$s^2(f_1) = \frac{(1 - f_1^2)(1 - f_1 \cdot t_1)^2}{n \cdot (f_1 - t_1)^2}$$

Predikční rovnice pro $k = 1$ se získá ze vztahu (44):

$$\begin{aligned} y_t(1) &= E(y_{t+1}) = f_1 \cdot y_t + E(a_{t+1}) - t_1 \cdot E(a_t) \\ y_t(1) &= f_1 \cdot y_t - t_1 \cdot a_t \end{aligned}$$

$$\text{var}(e_{t+1})^2 = s_a^2$$

Při použití modelů AR, MA a ARMA se předpokládá, že střední hodnota řady nevykazuje trend (řada stacionární podle střední hodnoty) a rovněž disperze nemá trend v čase (stacionární podle disperze). Splnění těchto předpokladů je potřeba ověřit a to buď testem nebo velmi často vizuálně z realizace časové řady.

Při nesplnění některého z těchto požadavků je nutné nestacionaritu odstranit.

Nestacionaritu řady podle střední hodnoty lze odstranit výpočtem prvních diferencí členů řady, tj. řadu $\{y_t\}$ nahradíme řadou

$$\{y_t - y_{t-1}\}$$

a není-li ani nyní stacionární, vypočítáme opět difference členů této nové řady, tedy druhé difference. To je zpravidla postačující, difference vyššího řádu se prakticky nepočítají.

Mějme například model

$$z_t = f_1 \cdot z_{t-1} + a_t$$

tj.

$$z_t (1 - f_1 \cdot B) = a_t$$

Je-li $z_t = y_t - y_{t-1}$, čili $\{z_t\}$ je řada prvních diferencí, můžeme psát

$$(y_t - y_{t-1})(1 - f_1 \cdot B) = a_t$$

$$y_t (1 - B)(1 - f_1 \cdot B) = a_t$$

$$y_t (1 - B) f(B) = a_t$$

Je-li dále $y_t = x_t - x_{t-1}$, máme

$$y_t (1 - B) f(B) = a_t$$

$$(x_t - x_{t-1})(1 - B) f(B) = a_t$$

$$x_t (1 - B)^2 f(B) = a_t$$

Je vidět, že zařazením členu $1 - B$ do rovnice $y_t \cdot f(B) = a_t$ je totéž, jako počítat s řadou prvních diferencí, zařazení $(1-B)^2$ znamená počítat se druhými diferencemi atd.

Obecně u d-tých diferencí bude

$$y_t (1 - B)^d f(B) = a_t$$

Pro model ARMA (p,q) pak můžeme psát

$$y_t (1 - B)^d \cdot f(B) = a_t \cdot t(B)$$

který je označen **ARIMA (p,d,q)** a zahrnuje všechny předchozí modely.

Například model ARIMA(1,0,0), zkráceně píšeme (1,0,0), odpovídá modelu AR(1), model (1,1,0) je model AR(1) pro první diference, model (0,0,2) je model MA(2) apod.

Při nestacionaritě podle disperze se provádí některá z následujících transformací:

$$y_t^* = \ln(y_t), \quad y_t^* = (y_t)^{1/2} \quad \text{resp.} \quad y_t^* = 1 / (y_t)^{1/2}$$

Významnost nalezených koeficientů testujeme podobně jako v regresní analýze:

Nulová a alternativní hypotéza

$$H_0 : f_k = 0 \text{ resp. } t_k = 0$$

$$H_1 : f_k \neq 0 \quad t_k \neq 0$$

Testovací kritérium

$$T = f_k / s(f_k) \text{ resp. } T = t_k / s(t_k)$$

Kritická hodnota se nalezne v tabulkách kritických hodnot Studentova rozdělení pro n-p-q stupně volnosti

$$t_{n-p-q}(\alpha), \alpha = \text{hladina významnosti}$$

Protože při správné volbě modelu časové řady jsou reziduální odchylky nekorelované, posuzuje se vhodnost modelu nepřímo – *testováním reziduí*:

Test Box-Pierce

Je-li y_t (empirická) hodnota časové řady a $\hat{y}_t$ její (teoretický) odhad, pak difference $y_t - \hat{y}_t = a_t$ tvoří také časovou řadu, můžeme vypočítat autokorelace

$$r_a(t) \text{ pro } t = 1, 2, \dots, k = n/10$$

H_0 : Všechny autokorelační koeficienty jsou nulové

$$(r_1 = r_2 = \dots = r_k = 0)$$

Testovací kritérium bude

$$Q = n \cdot \sum_{t=1}^k r_a^2(t)$$

Q má rozdělení χ^2_{k-p-q} v případě, že reziduální odchylky jsou nekorelované.

Předpoklad o nekorelovatelnosti hodnot řady $\{a_t\}$ je ověřován u každého modelu také

testem Durbin-Watson:

$$H_0 : \text{cov}(a_t, a_{t-1}) = 0$$

$$H_1 : \text{cov}(a_t, a_{t-1}) \neq 0$$

Testovací kritérium

$$DW = \sum_{t=2}^n (a_t - a_{t-1})^2 / \sum_{t=1}^n a_t^2$$

Protože přibližně platí

$$DW = 2 - 2 \cdot r_a(1)$$

je vidět, že při DW blízce hodnotě 2 nejsou a_t korelované. Nesplnění tohoto předpokladu signalizuje, že regresní model nebyl nalezen správně a v konečném důsledku je příčinou nekvalitní predikce.

Při praktickém hledání vhodného modelu jsme postupovali v těchto etapách:

1) Určení typu a řádu modelu:

K tomuto účelu je pro první rozhodování sestavena tabulka 1 a to podle Box-Jenkinsovy metodiky. Je připravena pro modely AR (1), AR (2), MA (1), MA (2) a ARIMA (1,1). V prvním sloupci každého modelu je znázorněn typický průběh grafu autokorelační funkce, ve druhém typické parciální korelační funkce.

K objasnění pojmu parciální korelace napíšeme několik AR modelů, např.

$$\begin{aligned} y_t &= f_{11} \cdot y_{t-1} + a_t \\ y_t &= f_{21} \cdot y_{t-1} + f_{22} \cdot y_{t-2} + a_t \\ y_t &= f_{31} \cdot y_{t-1} + f_{32} \cdot y_{t-2} + f_{33} \cdot y_{t-3} + a_t \end{aligned}$$

Pro odlišení koeficientů v jednotlivých modelech používáme dva indexy.

Koeficienty f_{ij} představují parciální korelace mezi y_t a y_{t-i} . Např. f_{22} je parciální korelace mezi proměnnými y_t a y_{t-2} .

Máme-li např. model AR (2), je zřejmě $f_{33} = f_{44} = \dots = 0$. Vypočítáme-li tedy z Y-W rovnice koeficienty $f_{11}, f_{22}, f_{33}, \dots, f_{kk}$ lze očekávat, že u modelu řádu p bude f_{kk} pro $k > p$ prakticky nulové.

Při testování významnosti koeficientů f_{kk} použijeme přibližného vztahu

$$s(f_{kk}) = 1 / n^{1/2}$$

Prakticky se doporučuje:

je-li f_{kk} mimo interval $(-2 \cdot s(f_{kk}), 2 \cdot s(f_{kk}))$, považujeme jej za nulový.

Průběh grafů atokorelací a parciálních korelací pro nejpoužívanější modely AR(1), AR(2), MA(1), MA(2) a ARMA(1,1) jsou v tabulce. Podle průběhu se může rozhodnout o volbě typu a řádu modelu.

2) Výpočet koeficientů vybraného modelu.

Provede se podle výše uvedených vzorců podle typu modelu.

Příklad 16

Časová řada spotřeby vody v domácnosti je v tabulce.

Vypočítat autokorelace a parciální korelace pro posuv $p=1,2,3,4$ a 5.

den	y_t
Po	0,64
U	0,78
St	0,93
Č	0,66
Pá	0,99
So	1,22
N	1,05
Po	0,75
U	0,63
St	0,82
Č	0,63
Pá	1,30
So	0,65
N	1,30
Po	0,54
U	0,61
St	0,70
Č	0,56
Pá	0,79
So	1,30
N	1,24

Řešení:

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. .	. .	1	-0.006	-0.006	0.0009	0.977
. *	. *	2	0.085	0.085	0.1756	0.916
*** .	*** .	3	-0.372	-0.374	3.7627	0.288
. ** .	. ** .	4	-0.272	-0.317	5.7906	0.215
. * .	. * .	5	-0.185	-0.183	6.7942	0.236

AC = autokorelace

PAC = parciální korelace

Sestavení predikčního modelu

Po nalezení vhodného modelu časové řady se sestaví predikční model.

Odhad y_{t+k} označme $y_t(k)$.

Věta 9

$$y_t(k) = E(y_{t+k})$$

Postup:

1. Pro dané parametry (p, d, q) sestavit model $y_t(1-B)^d f(B) = a_t t(B)$
2. Vypočítat y_t

3. Vypočítat analogicky y_{t+k}
4. $E(y_{t+k})$

Při výpočtu $E(a_{t+k})$ je

pro $k = \dots, -2, -1, 0$ $E(a_{t+k})$ rovno výrazu v závorce (střední hodnota minulých a současných hodnot), pro $k = 1, 2, \dots$ je $E(a_{t+k})$ rovno nule (střední hodnota budoucích časů).

Příklad 17

Napsat predikční rovnici řady $\{z_t\}$ pro model

- a) ARIMA(0,1,1)
- b) ARIMA(1,0,1)
- c) ARIMA(1,1,1)
- d) ARIMA(1,1,0)
- e) ARIMA(0,2,2)

Řešení pro (a):

$$z_t(1-B) = a_t(1-t_1B)$$

$$z_t = z_{t-1} + a_t - t_1 a_{t-1}$$

$$*) z_{t+k} = z_{t+k-1} + a_{t+k} - t_1 a_{t+k-1}$$

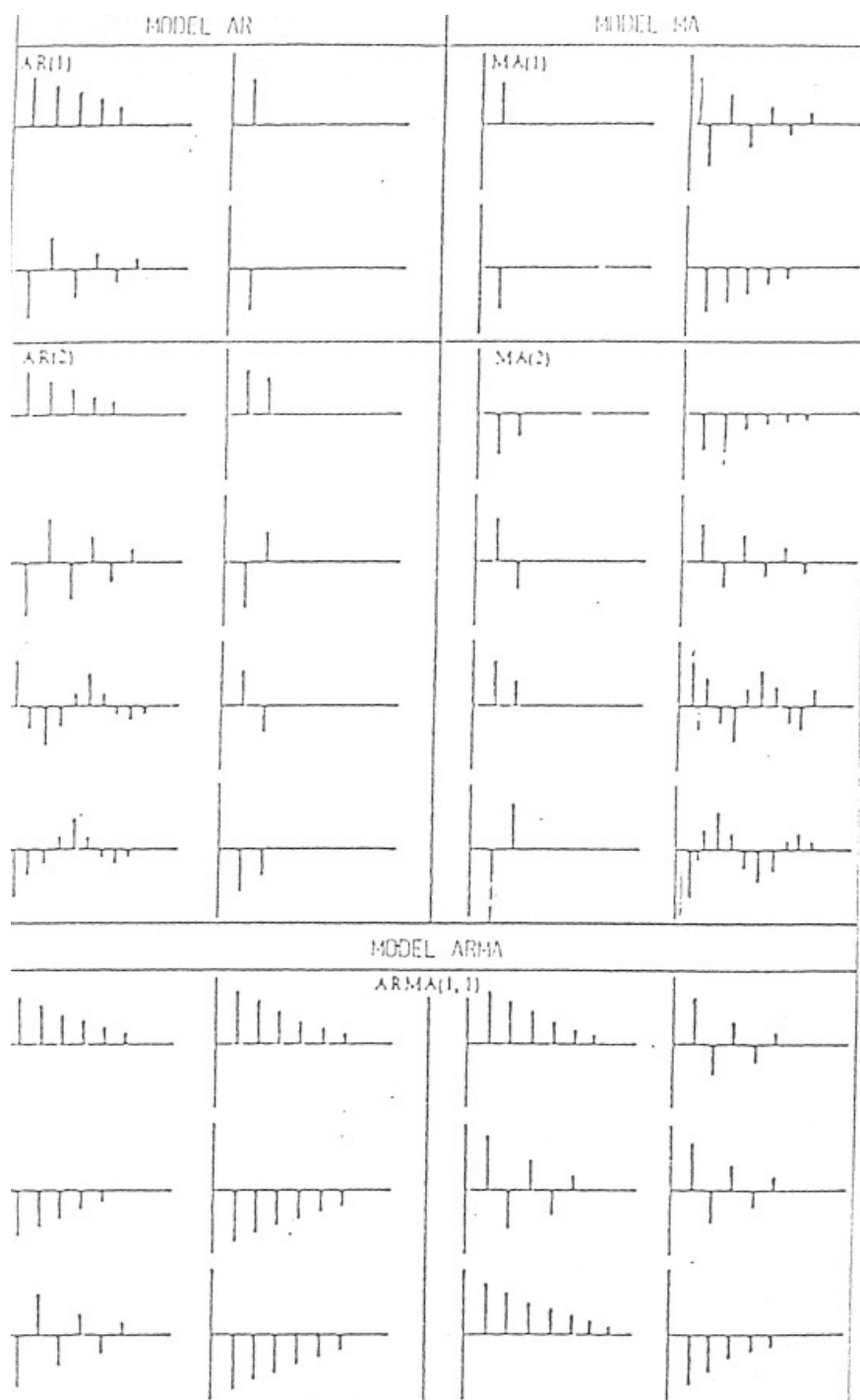
$$z_t(k) = E(z_{t+k}) = z_t(k-1) \text{ pro } k \geq 2$$

$k = 1$ dosadíme do *)

$$z_{t+1} = z_t + a_{t+1} - t_1 a_t$$

$$z_t(1) = z_t - t_1 a_t$$

Tabulka1 (Box-Jenkins) Grafy AC a PAC funkce pro vybrané modely k určení typu a řádu modelu.





Shrnutí pojmů kapitoly 11.2

V kapitole 11.2 byly vysvětleny dva způsoby analýzy časových řad: klasická metodika a Box Jenkinsova metodika. U obou přístupů jsou počítány charakteristiky časové řady a hledány modely časové řady. Strategie jsou však naprosto odlišné. Při modelování trendové složky časové řady jsou uplatňovány poznatky regresní analýzy.



Otázky ke kapitole 11.2

1. Definujte časovou řadu
2. Uveďte dva přístupy k analýze časových řad
3. Uveďte dvě základní úlohy při studiu časových řad
4. Jak se značí proměnné v ekonometrii?
5. Jaké je pojmenování proměnných v ekonometrii ?
6. Uveďte charakteristiky č.ř. klasické metodiky
7. Definujte stacionaritu č.ř. podle střední hodnoty (rozptylu
8. Napište základní rozklad hodnoty y_t u klasické metodiky
9. Jak se určí logistický trend?
10. Jak se počítají klouzavé průměry
11. Uveďte vzorec pro exponenciální vyrovnání č.ř.
12. Jak se modeluje sezónní složka?
13. Uveďte charakteristiky č.ř. v Box jenkinsově metodice
14. Uveďte vlastnosti autokorelační funkce
15. jak se počítá rozptyl r_k ?
16. Jak se stanoví intervalový odhad ρ_k
17. Jak se testuje r_k ?
18. Jaké jsou základní modely v B-J metodice?
19. Napište model $AR(p)$, $MA(q)$, $ARMA(p,q)$
20. Proveďte odvození Y-W rovnic
21. Napište Y.W rovnice pro $AR(1)$, $AR(2)$, $AR(3)$
22. Definujte operátor B
23. Napište model $ARIMA(p,d,q)$
24. Jak se odstraní trend č.ř.?
25. Co představují koeficienty f_{ii} ? Jaký je jejich význam?
26. Jak se ověří, jsou li nenulové?
27. Uveďte základní východisko pro nalezení predikčního modelu
28. Uveďte základní kroky při nalezení predikčního modelu
29. Nakreslete tabulku pro určení typu a řádu modelu



Úlohy k procvičení

1. Pro danou časovou řadu vypočítejte

a) průměrný absolutní přírůstek

b) průměrný koeficient růstu

t	1	2	3	4	5	6	7	8
yt	50	53	48	47	50	59	60	58

2. Pro danou časovou řadu vypočítejte

a) průměrný absolutní přírůstek

b) průměrný koeficient růstu

t	1	2	3	4	5	6
yt	5	9	11	10	12	14

3. Pro danou časovou řadu vypočítejte

a) exponenciální vyrovnání, je-li $w = 0,7$. Napište první 3 hodnoty.

b) průměrný koeficient růstu

t	1	2	3	4	5	6	7
yt	11	15	14	12	16	18	19

4. Pro danou časovou řadu vypočítejte

a) exponenciální vyrovnání, je-li $w = 0,8$. Napište první 3 hodnoty.

b) vyrovnání klouzavými průměry s délkou klouzavé části 5

t	1	2	3	4	5	6	7	8
yt	20	22	19	20	18	17	20	22

5. Vypočítejte testovací kritérium DW pro posouzení autokorelace, je-li dáno

$$e^T = (0.02; -0.04; -0.02; -0.10; 0.04)$$

6. Vypočítejte testovací kritérium DW pro posouzení autokorelace, je-li dáno

$$e^T = (-0.1; 0.2; -0.3; -0.1; 0.4)$$

7. Vypočítejte testovací kritérium DW pro posouzení autokorelace, je-li dáno

$$e^T = (-0.01; 0.08; -0.02; -0.03; 0.05)$$

8. Vypočítejte testovací kritérium DW pro posouzení autokorelace, je-li dáno

$$e^T = (-0.3; 0.4; -0.1; -0.2; 0.4; -0.1)$$

9. Napište rovnici pro model ARMA(4,2) a zjednodušte jej pomocí operátoru B.
10. Napište rovnici pro model ARMA(3,4) a zjednodušte jej pomocí operátoru B.
11. Napište rovnici pro model ARIMA(2,1,3) ve zkráceném tvaru.
12. Napište Yule Walkerovy rovnice pro model AR(3)
13. Otestujte koeficient f_{11} modelu AR(1), je-li dáno:
 $n = 5$, $f_{11} = -0,909$ a $\alpha = 0,05$.
14. Jsou dány autokorelace $r_1 = -0,577$ a $r_2 = 0,458$. Vypočítejte $s^2(r_1)$ a $s^2(r_2)$; $n = 10$.
15. Odstraňte heteroskedasticitu v modelu $\hat{Y}_t = b_1 x_{t1} + b_2 x_{t2} + \dots + b_k x_{tk} + e_t$, je-li příčinou heteroskedasticity proměnná x_k .
16. Napište vzorec pro výpočet r_k a $s^2(r_k)$.
17. Vypočítejte $s^2(r_3)$ je-li $r_1 = 0.8$, $r_2 = 0.2$ a $n = 10$.
18. Řešte všechny body příkladu 1 pro vstupní data v tabulce:

Y	x_1	x_2
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2



KLÍČ K ŘEŠENÍ

1) ($\bar{\Delta} = 1,14; \bar{k} = 1,02$)

2) ($\bar{\Delta} = 18; \bar{k} = 1,22$)

3) ($\hat{y}_1 = 11; \hat{y}_2 = 13,8; \hat{y}_3 = 13,94; \bar{k} = 1,095$)

4) ($\hat{y}_1 = 20; \hat{y}_2 = 21,6; \hat{y}_3 = 19,52; \bar{k} = 1,095$)

5) (DW = 2.14)

6) (DW = 2.032)

7) (DW = 2.38)

8) (DW = 2.89)

9) ($y_t = f_1 \cdot y_{t-1} + f_2 \cdot y_{t-2} + f_3 y_{t-3} + f_4 \cdot y_{t-4} + a_t - t_1 \cdot a_{t-1} - t_2 \cdot a_{t-2}; f(B^4)y_t = t(B^2)a_t$)

10) ($y_t = f_1 \cdot y_{t-1} + f_2 \cdot y_{t-2} + f_3 y_{t-3} + a_t - t_1 \cdot a_{t-1} - t_2 \cdot a_{t-2} - t_3 \cdot a_{t-3} - t_4 a_{t-4}; f(B^3)y_t = t(B^4)a_t$)

11) ($f(B^2)(1-B)y_t = t(B^3) \cdot a_t$)

$$(r_1 = f_1 + f_2 r_1 + f_3 r_2$$

12) $r_2 = f_1 r_1 + f_2 + f_3 r_1$

$$r_3 = f_1 r_2 + f_2 r_1 + f_3)$$

13) ($T = |\sqrt{5} \cdot 0,909| = 2,03 > 1,96$)

14) ($s^2(r_1) = 0,1; s^2(r_2) = 0,166$)

17) ($s^2(r_3) = 0.236$)

18)

a) Odhad regresních koeficientů

$$X^T Y = \begin{bmatrix} 368 \\ 2710 \\ 35 \end{bmatrix}$$

$$\vec{b} = \frac{1}{4664} \begin{bmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ 240 & -40 & 608 \end{bmatrix} \begin{bmatrix} 368 \\ 2710 \\ 35 \end{bmatrix} = \begin{bmatrix} 6,47 \\ 6,59 \\ 0,26 \end{bmatrix}$$

b) Teoretické hodnoty

$$\hat{Y} = (13.06, 25.98, 32.83, 39.68, 52.34, 59.19, 72.11, 72.89)$$

c) Rezidua

$$\vec{e}^T = (-3.06, -0.98, -0.83, 3.32, 5.66, 2.81, -5.11, -1.89)$$

d) Rozptyl koeficientů $\sum_i e_i^2 = 91.65$

$$\text{var}(\vec{b}) = \frac{91.65}{8-3} \cdot \frac{1}{4664} \begin{bmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ 240 & -40 & 608 \end{bmatrix} = \begin{bmatrix} 11.35 & & \\ & 0.25 & \\ & & 2.39 \end{bmatrix}$$

e)

$$\bar{Y} = 46, Se = 91.65, S_{\hat{Y}} = 3318.21$$

f) Test koeficientů

$$T_1 = \frac{b_1}{s(b_1)} = \frac{6.47}{3.37} = 1.92, T_2 = \frac{6.59}{0.5} = 13.18, T_3 = 0.17$$

$$t_{n-k}(\alpha) = t_{8-3}(0.05) = 2.571$$

$$T_3 < 2.571 \Rightarrow \beta_3 = 0$$

g) Test modelu

$$T = \frac{3318/(3-1)}{91.65/(8-3)} = \frac{16.59}{18.33} = 90.51$$

$$F_{k-1, n-k}(\alpha) = F_{3-1, 8-3}(0.05) = 5.79$$

h) Predikce

$$X_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix}$$

$$\hat{Y}_0 = \begin{bmatrix} 1 & 11 & 3 \\ 1 & 12 & 5 \end{bmatrix} \begin{bmatrix} 6.47 \\ 6.59 \\ 0.26 \end{bmatrix} = \begin{bmatrix} 79.74 \\ 86.85 \end{bmatrix}$$

Literatura

1. Alonso,J.H.: Ejercicios de econometria. ESIC, Madrid 1989.
2. Anděl,J.: Matematická statistika. SNTL Praha 1978.
3. Koutsoyiannis,A.: Theory of Econometrics. MACMILLAN 1977.
4. Nowak,E.: Zarys metod ekonometrii. Warszawa 1990.
5. Box,G.-Hunter,W.-Hunter,S.:*Statistics for experimenters*. Wiley 1978.
6. Myers,R.H., Montgomery,D.C. *Response Surface Methodology*,2nd Ed., Wiley,NY 2002.
7. Prat,A.-Tort,X.- Grima,P.-Pozutea,L.: *Métodos estadísticos. Control y mejora de la calidad*. UPC, Barcelona 1992.
8. Anděl,J.: *Matematická statistika*. Praha, SNTL 1978.
9. Box,G.E.P.-Jenkins,G.M.: *Time series analysis*. San Francisco, 1970.