

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Zdeněk Ambrož

Regresní modely pro analýzu výnosu portfolia

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr.

Studijní program: Matematika, obor Obecná matematika

2007

Rád bych tímto poděkoval RNDr. Jitce Zichové, Dr., za její cenné rady, připomínky a ochotu vyjít vstříc při řešení problémů při psaní této bakalářské práce.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 6. 8. 2007

Zdeněk Ambrož

Obsah

Úvod	6
1 Korelace	7
1.1 Lineární aproximace	7
1.2 Korelační koeficient a výběrový korelační koeficient	7
1.3 Varianční a korelační matice	8
1.4 Odhady varianční a korelační matice	9
1.5 Koeficient mnohonásobné korelace a výběrový koeficient mnohonásobné korelace	10
1.6 Koeficient parciální korelace a výběrový koeficient parciální korelace ...	11
2 Věta o inverzní varianci a její důsledky	13
2.1 Nezávislost a podmíněná nezávislost	13
2.2 Podmíněná střední hodnota a podmíněný rozptyl	13
2.3 Parciální variance, kovariance a korelace	14
2.4 Lemma o inverzní varianci a jeho důsledky	14
2.5 Nezávislost a podmíněná nezávislost v normálním rozdělení	16
3 Lineární regresní model	18
3.1 Lineární regresní model	18
3.2 Testování submodelů	19
3.3 Koeficient determinace	23
3.4 Model CAPM	24
3.5 Model APT (Arbitrážní cenový model)	27
3.6 Mnohorozměrná regrese	27
4 Grafické modely	29
4.1 Základní pojmy	29
4.2 Grafy a podmíněná nezávislost	29
4.3 Řetězové grafy podmíněných nezávislostí	30
4.4 Regrese s více závisle proměnnými a řetězové grafy	32
4.5 Selektce grafického modelu s řetězovým grafem	33
4.6 Popis programu	35
4.6.1 Regresní analýza	35
4.6.2 Korelační analýza	35
4.6.3 Testování submodelu vůči modelu	36
4.6.4 Testování submodelu vůči nadřazenému submodelu	36
4.6.5 Hledání regresních parametrů vícerozměrné regrese	36
4.6.6 Selektce grafického modelu s řetězovým grafem	36
5 Analýza italských finančních dat	38
5.1 Model CAPM	38
5.2 Model APT	40
5.2.1 Regresní analýza	40
5.2.2 Korelační analýza	42
5.3 Grafický model	45
5.3.1 Korelační analýza	45

5.3.2	Regresní analýza	46
5.3.3	Model získaný selekčním algoritmem	48
5.4	Testování submodelů	50
6	Burzovní indexy českého trhu	57
6.1	Regresní analýza	58
6.2	Korelační analýza	59
6.3	Mnohorozměrná regrese	60
6.4	Selekce grafického modelu	60
6.5	Grafický model s řetězovým grafem	62
	Závěr	67
	Literatura	69
	Přílohy	70

Název práce: Regresní modely pro analýzu výnosu portfolia

Autor: Zdeněk Ambrož

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr.

e-mail vedoucího: zichova@karlin.mff.cuni.cz

Abstrakt: Při studiu výnosu portfolia se používá mnoho modelů. V této práci budeme pracovat s modelem oceňování kapitálových aktiv (model CAPM), arbitrážním cenovým modelem (model APT) a grafickými modely. V daných modelech budeme zkoumat zejména to, nakolik se ovlivňuje výše výnosů cenných papírů a portfolia. V první kapitole shrneme námi užívaný aparát z regresní a korelační analýzy, ve druhé seznámíme čtenáře s některými vlastnostmi variančních matic, ve třetí seznámíme čtenáře s předpoklady a vlastnostmi lineárních regresních modelů, ve čtvrté zopakujeme základní terminologii teorie grafů a ukážeme možnosti využití grafů pro znázornění podmíněných nezávislostí náhodných veličin. V páté a šesté kapitole bude demonstrováno využití uvedených analýz na konkrétních datech z italského (pátá kapitola) a českého (šestá kapitola) prostředí. Regresní a korelační analýzu a selekci grafického modelu s řetězovým grafem provádí program napsaný v systému Mathematica 4.0 (CD je součástí práce).

Klíčová slova: regresní model, portfolio, výnos, grafický model, řetězový graf.

Title: Regression models for the analysis of portfolio return

Author: Zdeněk Ambrož

Department: Department of probability and mathematical statistics

Supervisor: RNDr. Jitka Zichová, Dr.

Supervisor's e-mail address: zichova@karlin.mff.cuni.cz

Abstract: To study the portfolio return we can use a lot of models. In this thesis we work with "CAPM", "APT" and with graphical models. By means of these models we investigate the relationship between some variables influencing the portfolio return. In the first chapter we summarize the apparatus of regression and correlation analysis. In the second chapter we introduce some features of variance matrices. In the third chapter we introduce the assumptions and the features of linear regression model. In the fourth chapter we repeat some basic terminology of the graph theory and we show the possibility of using graphs to demonstrate conditional independence of random variables.

In the fifth and the sixth chapter we demonstrate the application of mentioned analysis to real data from italian (the fifth chapter) and czech (the sixth chapter) market. The regression and the correlation analysis and the selection of a suitable graphical chain model is solved by a program processed in the Mathematica 4.0 system (the CD is enclosed).

Keywords: regression model, graphical model, portfolio, return, graphical chain model.

Úvod

Cílem této práce je prezentovat různé modely regresního typu používané v analýze výnosu portfolia a jiných finančních dat. Jsou to jednak tradiční používané modely popisující závislost výnosu na riziku (CAPM, APT), jednak gaussovské grafické modely popisující strukturu podmíněných (ne)závislostí v množině proměnných. Tyto modely jsou založeny na studiu parciálních korelací, což představuje alternativu k lineárně-regresnímu přístupu k datům. Dále jsou to grafické modely s řetězovým grafem, které na rozdíl od modelů předchozích respektují i očekávanou hierarchickou strukturu modelu. Ke zpracování dat byl použit program napsaný v systému Mathematica 4.0. Tento program je přiložen na CD. V první části práce je shrnuta teorie týkající se zmíněných typů modelů. V části druhé jsou zpracována data z knihy Giudici [4], to jest je provedena analýza výnosu portfolia v závislosti na výnosech jednotlivých cenných papírů a indexů finančního trhu prostřednictvím modelu CAPM, APT a grafického modelu. Všechny tři modely jsou uvedeny ve tvaru modelů lineárních. Dále navrhneme optimální grafický model pomocí selekčního algoritmu a nakonec provedeme testování méně rozsáhlých modelů jako submodelů modelů rozsáhlejších. Ve třetí části této práce je pak pomocí regresní a korelační analýzy a hlavně teorie grafických modelů zpracována další sada dat, a to burzovní indexy z českého kapitálového trhu. Pro tato data dále vybereme jednak vhodný grafický model popisující strukturu vzájemných podmíněných nezávislostí v množině odvětvových indexů, jednak vhodný grafický model s řetězovým grafem popisující (na základě blokové struktury dat), ze kterých veličin je vhodné kterou veličinu predikovat.

1. Korelace

1.1 Lineární aproximace

Mějme náhodnou veličinu Y a vektor $\mathbf{X} = (X_1, \dots, X_k)^T$ typu $k \times 1$ s regulární varianční maticí. Předpokládejme, že platí

$$Y = \alpha + \boldsymbol{\beta}^T \mathbf{X} + e,$$

kde

α je neznámý parametr,

$\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$ vektor neznámých parametrů a

e je náhodná chyba splňující podmínku $E e = 0$.

Označme $\hat{Y}(\mathbf{X})$ nejlepší lineární aproximaci Y pomocí X_1, \dots, X_k ve smyslu minimalizace $E(Y - \hat{Y}(\mathbf{X}))^2$:

$$\hat{Y}(\mathbf{X}) = \alpha + \beta_1 X_1 + \dots + \beta_k X_k = \alpha + \boldsymbol{\beta}^T \mathbf{X}.$$

Platí

$$\boldsymbol{\beta} = (\text{var } \mathbf{X})^{-1} \text{cov}(\mathbf{X}, Y),$$

$$\alpha = E Y - \boldsymbol{\beta}^T E \mathbf{X},$$

kde $\text{cov}(\mathbf{X}, Y) = (\text{cov}(Y, X_1), \dots, \text{cov}(Y, X_k))^T$.

Rozptyl odhadu je roven

$$\text{var } \hat{Y}(\mathbf{X}) = \boldsymbol{\beta}^T \text{var } \mathbf{X} \boldsymbol{\beta} = \text{cov}(Y, \mathbf{X})(\text{var } \mathbf{X})^{-1} \text{cov}(\mathbf{X}, Y).$$

Jsou-li veličiny standardizované, platí

$$E Y = E X_1 = \dots = E X_k = 0 \Rightarrow \alpha = 0$$

a vzorec pro výpočet nejlepší lineární aproximace lze zjednodušit na

$$\hat{Y}(\mathbf{X}) = \beta_1 X_1 + \dots + \beta_k X_k.$$

1.2 Korelační koeficient a výběrový korelační koeficient

Mějme X a Y náhodné veličiny s konečnými druhými momenty. Označme

$$\sigma_X^2 = \text{var } X, \quad \sigma_Y^2 = \text{var } Y, \quad \sigma_{X,Y} = \text{cov}(X, Y).$$

Nechť $\sigma_X^2 \sigma_Y^2 > 0$. Pak korelační koeficient definujeme předpisem

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sqrt{\sigma_X^2 \sigma_Y^2}}.$$

Mějme $(X_1, Y_1)^T, \dots, (X_n, Y_n)^T$ výběr z nějakého dvojrozměrného rozdělení. Označíme

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i, & \bar{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i, \\ S_{X,Y} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}), \\ S_X^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, & S_Y^2 &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.\end{aligned}$$

Výběrový korelační koeficient definujeme předpisem

$$r_{X,Y} = \frac{S_{X,Y}}{\sqrt{S_X^2 S_Y^2}}$$

pro $S_X^2 S_Y^2 \neq 0$.

Věta 1.1: Necht' $(X_1, Y_1)^T, \dots, (X_n, Y_n)^T$ je výběr z dvojrozměrného normálního rozdělení, které má kladné rozptyly a korelační koeficient $\rho = 0$. Dále předpokládejme $n \geq 3$. Pak náhodná veličina

$$T = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$$

má rozdělení t_{n-2} .

Důkaz: Viz Anděl [2], strana 217.

Pomocí věty 1.1 můžeme testovat hypotézu

$$H_0 : \rho = 0$$

proti alternativě

$$H_1 : \rho \neq 0.$$

V případě, že

$$|T| \geq t_{n-2}(\alpha)$$

zamítáme H_0 ve prospěch H_1 .

1.3 Varianční a korelační matice

Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$ a jeho varianční matici

$$V = \text{var } \mathbf{X} = \text{E}[(\mathbf{X} - \text{E } \mathbf{X})(\mathbf{X} - \text{E } \mathbf{X})^T] = \text{E}(\mathbf{X}\mathbf{X}^T) - (\text{E } \mathbf{X})(\text{E } \mathbf{X})^T$$

s diagonálními prvky

$$\sigma_{i,i} = \sigma_i^2 = \text{var } X_i$$

a mimodiagonálními prvky

$$\sigma_{i,j} = \text{cov}(X_i, X_j).$$

Dále označme

$$D = V^{-1}.$$

Korelační matici vektoru \mathbf{X} označme P . Platí pro ni

$$P = \text{cor } \mathbf{X} = B^{-1}VB^{-1},$$

kde

$$B = \text{diag}\{\sigma_1, \dots, \sigma_k\}.$$

Diagonální prvky matice P označme

$$\rho_{i,i} = 1,$$

a mimodiagonální

$$\rho_{i,j} = \rho_{X_i, X_j}.$$

Matice P vznikne z matice V takzvaným škálováním.

Definice 1.1: Mějme matici A typu $m \times m$, $A = (a_{i,j})_{i,j=1}^m$. Matici $S(A)$ nazveme škálovanou maticí k matici A , jestliže

$$S(A) = CAC,$$

kde $C = \text{diag}\left\{\frac{1}{\sqrt{a_{1,1}}}, \dots, \frac{1}{\sqrt{a_{m,m}}}\right\}$.

Zřejmě má že $S(A)$ na diagonále jedničky a $S(V) = P$.

1.4 Odhady varianční a korelační matice

Uvažujme n realizací vektoru $\mathbf{X} = (X_1, \dots, X_k)^T$:

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & \dots & X_{1,k} \\ \dots & \dots & \dots \\ X_{n,1} & \dots & X_{n,k} \end{pmatrix},$$

kde

prvky $X_{i,j}$ jsou i -tou realizací j -té složky vektoru \mathbf{X} ,
řádky \mathbf{X}_i matice \mathbf{X} jsou i -tou realizací vektoru \mathbf{X} , $i=1, \dots, n$ a
sloupce \mathbf{X}_j matice \mathbf{X} jsou n realizací j -té složky vektoru \mathbf{X} , $j = 1, \dots, k$.
Označme

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{i,j}$$

průměr j -té složky vektoru \mathbf{X} přes všechny realizace, odhad $E X_j, j = 1, \dots, k$.

Dále označme

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$$

průměr vektoru \mathbf{X} přes všechny jeho realizace, odhad pro $E \mathbf{X}$.

Zřejmě je

$$\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_k).$$

Označíme

$$s_{j,j} = \frac{1}{n-1} \sum_{t=1}^n (X_{t,j} - \bar{X}_j)^2$$

odhad pro $\sigma_{j,j} = \text{var } X_j, j = 1, \dots, k,$

$$s_{i,j} = \frac{1}{n-1} \sum_{t=1}^n (X_{t,j} - \bar{X}_j)(X_{t,i} - \bar{X}_i)$$

odhad pro kovarianci $\sigma_{i,j} = \text{cov}(X_i, X_j)$ i-té a j-té složky vektoru \mathbf{X} ,

$$S = (s_{i,j})_{i,j=1}^k = \frac{1}{n-1} \sum_{t=1}^n (\mathbf{X}_t - \bar{\mathbf{X}})(\mathbf{X}_t - \bar{\mathbf{X}})^T = \frac{1}{n-1} \sum_{t=1}^n (\mathbf{X}_t \mathbf{X}_t^T - n \bar{\mathbf{X}} \bar{\mathbf{X}}^T)$$

odhad pro varianční matici V , to jest výběrovou varianční matici,

S^{-1} odhad pro D ,

$r_{i,i} = 1$ odhad pro $\rho_{i,i}$,

$$r_{i,j} = \frac{s_{i,j}}{\sqrt{s_{i,i} s_{j,j}}}$$

odhad pro $\rho_{i,j}$ - korelační koeficient i-té a j-té složky vektoru \mathbf{X} ,

$$R = (r_{i,j})_{i,j=1}^k = S(S)$$

odhad korelační matice P , to jest výběrovou korelační matici,

R^{-1} odhad matice P^{-1} .

1.5 Koeficient mnohonásobné korelace a výběrový koeficient mnohonásobné korelace

Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$ a náhodnou veličinu Y . Nechť $V = \text{var } \mathbf{X}$ je regulární matice. Uvažujme nejlepší lineární aproximaci veličiny Y pomocí vektoru \mathbf{X} zavedenou v odstavci 1.1, tj. $\hat{Y} = \alpha + \beta^T \mathbf{X}$.

Koeficient mnohonásobné korelace $\rho_{Y,\mathbf{X}}$ se zavádí jako korelační koeficient mezi veličinami Y, \hat{Y} :

$$\rho_{Y,\mathbf{X}}^2 = \rho_{Y,\alpha+\beta^T \mathbf{X}}.$$

V knize Anděl [2] je uveden vztah

$$\rho_{Y,\mathbf{X}}^2 = \text{cor}(Y, \mathbf{X}) P^{-1} \text{cor}(\mathbf{X}, Y).$$

Mějme náhodný výběr $\begin{pmatrix} Y_1 \\ \mathbf{X}_1 \end{pmatrix}, \dots, \begin{pmatrix} Y_n \\ \mathbf{X}_n \end{pmatrix}$ z nějakého $(k+1)$ -rozměrného rozdělení. Označme $r_{i,j}$ výběrový korelační koeficient mezi i-tou a j-tou složkou vektoru \mathbf{X} a $r_{0,i}$ výběrový korelační koeficient mezi veličinami Y a X_i . Zavedme výběrové korelační matice

$$R_{\mathbf{X},\mathbf{X}} = (r_{i,j})_{i,j=1}^k, \quad R_{Y,\mathbf{X}} = (r_{0,i})_{i=1}^k, \quad R_{\mathbf{X},Y} = R_{Y,\mathbf{X}}^T.$$

Matice $R_{\mathbf{X},\mathbf{X}}$ má na diagonále jedničky a je symetrická. Budeme navíc předpokládat, že je regulární. Výběrový koeficient mnohonásobné korelace $r_{Y,\mathbf{X}}$ definujeme vztahem

$$r_{Y,\mathbf{X}}^2 = R_{Y,\mathbf{X}} R_{\mathbf{X},\mathbf{X}}^{-1} R_{\mathbf{X},Y}.$$

Věta 1.2: Mějme $\begin{pmatrix} Y_1 \\ \mathbf{X}_1 \end{pmatrix}, \dots, \begin{pmatrix} Y_n \\ \mathbf{X}_n \end{pmatrix}$ náhodné vektory získané výběrem z $(k+1)$ -rozměrného regulárního normálního rozdělení. Jestliže dále platí $n > k + 1$ a $\rho_{Y,\mathbf{X}} = 0$, pak náhodná veličina

$$Z = \frac{n - k - 1}{k} \cdot \frac{r_{Y,\mathbf{X}}^2}{1 - r_{Y,\mathbf{X}}^2}$$

má $F_{k,n-k-1}$ rozdělení.

Důkaz: Viz Anděl [2], strana 221.

Věta 1.2 se používá k testování hypotézy

$$H_0 : \rho_{Y,\mathbf{X}} = 0$$

proti alternativě

$$H_1 : \rho_{Y,\mathbf{X}} \neq 0.$$

V případě, že

$$Z \geq F_{k,n-k-1}(\alpha)$$

zamítáme H_0 ve prospěch H_1 .

1.6 Koeficient parciální korelace a výběrový koeficient parciální korelace

Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$ a náhodné veličiny Y a W . Předpokládáme, že \mathbf{X} působí na Y i na W , a budeme zkoumat, jaká by byla závislost mezi Y a W , kdyby na ně vektor \mathbf{X} nepůsobil. Nejlepším lineárním přiblížením náhodných veličin Y a W pomocí \mathbf{X} jsou podle odstavce 1.1 veličiny

$$\hat{Y} = \alpha_1 + \beta_1^T \mathbf{X}, \quad \hat{W} = \alpha_2 + \beta_2^T \mathbf{X}.$$

Rozdíl $Y - \hat{Y}$ můžeme interpretovat jako tu část veličiny Y , na kterou vektor \mathbf{X} nepůsobí. Podobně můžeme interpretovat rozdíl $W - \hat{W}$. Závislost mezi Y a W , je-li eliminován vliv vektoru \mathbf{X} , měříme korelačním koeficientem $Y - \hat{Y}$ a $W - \hat{W}$. Nazýváme ho koeficientem parciální korelace mezi Y a W při daném \mathbf{X} a značíme ho $\rho_{Y,W|\mathbf{X}}$.

V knize Anděl [2] je uveden tvar

$$\rho_{Y,W|\mathbf{X}} = \frac{\rho_{Y,W} - \text{cor}(Y, \mathbf{X})P^{-1} \text{cor}(\mathbf{X}, W)}{\sqrt{(1 - \text{cor}(Y, \mathbf{X})P^{-1} \text{cor}(\mathbf{X}, Y))(1 - \text{cor}(W, \mathbf{X})P^{-1} \text{cor}(\mathbf{X}, W))}}.$$

Mějme náhodný výběr $\begin{pmatrix} Y_1 \\ \mathbf{X}_1 \\ W_1 \end{pmatrix}, \dots, \begin{pmatrix} Y_n \\ \mathbf{X}_n \\ W_n \end{pmatrix}$ z nějakého $(k+2)$ -rozměrného rozdělení.

Označíme $r_{i,j}$ výběrový korelační koeficient mezi i -tou a j -tou složkou vektoru \mathbf{X} , $r_{0,i}$ označme výběrový korelační koeficient mezi veličinami Y a X_i a $r_{k+1,i}$ označme výběrový korelační koeficient mezi veličinami W a X_i , $i = 1, \dots, k$.

Označíme

$$R_{\mathbf{X},\mathbf{X}} = (r_{i,j})_{i,j=1}^k,$$

$$R_{Y,X} = (r_{0,i})_{i=1}^k, \quad R_{X,Y} = R_{Y,X}^T,$$

$$R_{W,X} = (r_{k+1,i})_{i=1}^k, \quad R_{X,W} = R_{W,X}^T.$$

Výběrový koeficient parciální korelace $r_{Y,W|X}$ pak definujeme vzorcem

$$r_{Y,W|X} = \frac{r_{Y,W} - R_{Y,X} R_{X,X}^{-1} R_{X,W}}{\sqrt{(1 - R_{Y,X} R_{X,X}^{-1} R_{X,Y})(1 - R_{W,X} R_{X,X}^{-1} R_{X,W})}},$$

jestliže je jmenovatel různý od nuly.

Věta 1.3: *Nechť náhodné vektory $\begin{pmatrix} Y_1 \\ X_1 \\ W_1 \end{pmatrix}, \dots, \begin{pmatrix} Y_n \\ X_n \\ W_n \end{pmatrix}$ jsou výběrem z regulárního $(k+2)$ -rozměrného normálního rozdělení. Jestliže platí $\rho_{Y,W|X} = 0$, pak při $n > k + 2$ má náhodná veličina*

$$T = \frac{r_{Y,W|X}}{\sqrt{1 - r_{Y,W|X}^2}} \sqrt{n - k - 2}$$

rozdělení t_{n-k-2} .

Důkaz: Viz Anděl [2], strana 223.

Věta 1.3 se používá k testování hypotézy

$$H_0 : \rho_{Y,W|X} = 0$$

proti alternativě

$$H_1 : \rho_{Y,W|X} \neq 0.$$

V případě, že

$$|T| \geq t_{n-k-2}(\alpha)$$

zamítáme H_0 ve prospěch H_1 .

2. Věta o inverzní varianci a její důsledky

2.1 Nezávislost a podmíněná nezávislost

Máme-li dva náhodné vektory \mathbf{X} , \mathbf{Y} , definujeme podmíněnou hustotu

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})},$$

pokud je jmenovatel nenulový, a 0 jinak.

Jsou-li náhodné vektory \mathbf{X} a \mathbf{Y} nezávislé, platí

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}).$$

Nezávislost náhodných vektorů \mathbf{X} a \mathbf{Y} značíme $\mathbf{X} \perp \mathbf{Y}$.

Definice 2.1: Náhodné vektory \mathbf{X} a \mathbf{Y} nazveme podmíněně nezávislé při dané hodnotě vektoru \mathbf{Z} , když podmíněné hustoty splňují rovnost

$$f_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x})f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y})$$

pro všechna \mathbf{x} , \mathbf{y} a pro \mathbf{z} taková, že $f_{\mathbf{Z}}(\mathbf{z}) \neq 0$.

Nezávislost náhodných vektorů \mathbf{X} a \mathbf{Y} při dané hodnotě náhodného vektoru \mathbf{Z} značíme $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$.

Věta 2.1: Pro $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$ platí

$$\begin{aligned} 1) f_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\mathbf{y}) &= f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}) \\ 2) f_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= \frac{f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z})}{f_{\mathbf{Z}}(\mathbf{z})} = f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}, \mathbf{z})f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}) = \\ &= f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y})f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x})f_{\mathbf{Z}}(\mathbf{z}) \end{aligned}$$

Důkaz: Viz Whittaker [6], str. 32

2.2 Podmíněná střední hodnota a podmíněný rozptyl

Podmíněná střední hodnota a podmíněný rozptyl náhodné veličiny Y při daných hodnotách náhodného vektoru \mathbf{X} jsou střední hodnota a rozptyl Y počítané pomocí podmíněné hustoty:

$$\mathbf{E}_{\mathbf{Y}|\mathbf{X}} Y = \int_{-\infty}^{\infty} y f_{\mathbf{Y}|\mathbf{X}}(y) dy,$$

$$\text{var}_{\mathbf{Y}|\mathbf{X}} Y = \mathbf{E}_{\mathbf{Y}|\mathbf{X}} (Y - \mathbf{E}_{\mathbf{Y}|\mathbf{X}} Y)^2.$$

Platí

$$\text{var} Y = \mathbf{E} \text{var}_{\mathbf{Y}|\mathbf{X}} Y + \text{var} \mathbf{E}_{\mathbf{Y}|\mathbf{X}} Y.$$

Uvažujme náhodnou veličinu Y a náhodný vektor \mathbf{X} . Má-li $\begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix}$ normální rozdělení, je nejlepší lineární aproximace Y pomocí \mathbf{X} dána vztahem

$$\widehat{Y}(\mathbf{X}) = E_{Y|\mathbf{X}} Y.$$

2.3 Parciální variance, kovariance a korelace

Mějme $\widehat{Y} = \widehat{Y}(\mathbf{X})$ nejlepší lineární aproximaci náhodné veličiny Y pomocí náhodného vektoru \mathbf{X} , který má regulární varianční matici $V = \text{var } \mathbf{X}$. Podobně mějme $\widehat{W} = \widehat{W}(\mathbf{X})$ nejlepší lineární aproximaci náhodné veličiny W .

Parciální rozptyl $\text{var}(Y|\mathbf{X})$ definujeme vztahem

$$\text{var}(Y|\mathbf{X}) = \text{var}(Y - \widehat{Y}),$$

parciální kovarianci $\text{cov}(Y, W|\mathbf{X})$ definujeme vztahem

$$\text{cov}(Y, W|\mathbf{X}) = \text{cov}(Y - \widehat{Y}, W - \widehat{W})$$

a parciální korelaci $\text{cor}(Y, W|\mathbf{X})$ definujeme vztahem

$$\text{cor}(Y, W|\mathbf{X}) = \frac{\text{cov}(Y, W|\mathbf{X})}{\sqrt{\text{var}(Y|\mathbf{X}) \text{var}(W|\mathbf{X})}},$$

pokud je jmenovatel nenulový. Jedná se tedy o korelaci veličin $Y - \widehat{Y}$, $W - \widehat{W}$, to jest o koeficient parciální korelace zavedený v odstavci 1.6.

Věta 2.2: *Parciální kovariance $\text{cov}(Y, W|\mathbf{X})$ je bilineární v proměnných Y, W a splňuje vztah*

$$\text{cov}(Y, W|\mathbf{X}) = \text{cov}(Y, W) - \text{cov}(Y, \mathbf{X})V^{-1} \text{cov}(\mathbf{X}, W).$$

Důkaz: Viz Whittaker [6], strana 135.

Speciálním případem této věty je věta pro parciální rozptyl.

Věta 2.3: *Parciální rozptyl $\text{var}(Y|\mathbf{X})$ je lineární v proměnné Y a splňuje vztahy*

$$\text{var}(Y|\mathbf{X}) = \text{var } Y - \text{var } \widehat{Y},$$

$$\text{var}(Y|\mathbf{X}) = \text{var } Y - \text{cov}(Y, \mathbf{X})V^{-1} \text{cov}(\mathbf{X}, Y).$$

Důkaz: Viz Whittaker [6], strana 136.

2.4 Lemma o inverzní varianci a jeho důsledky

V odstavcích 1.5 a 1.6 jsme definovali výběrový koeficient mnohonásobné korelace a výběrový koeficient parciální korelace a naznačili jsme způsob jejich výpočtu. V následující kapitole ukážeme, jak lze tyto výběrové korelační koeficienty počítat pomocí inverze varianční matice. Základem je tvrzení známé jako lemma o inverzní varianci. Pro přehlednější zápisu budeme dále používat značení $N((\mathbf{X})^T; i, j)$ pro vektor typu $(k-2) \times 1$ vzniklý vynecháním i -té a j -té složky vektoru \mathbf{X} a $N((\mathbf{X})^T; i)$ pro vektor typu $(k-1) \times 1$ vzniklý vynecháním i -té složky vektoru \mathbf{X} .

Věta 2.4: (Lemma o inverzní varianci) Mějme náhodné vektory \mathbf{X} a \mathbf{Y} , necht' \mathbf{X} má regulární varianční matici V . Předpokládejme, že matice $\text{var}(\mathbf{X}, \mathbf{Y})$ je také regulární. Pak existuje matice k ní inverzní $\text{var}^{-1}(\mathbf{X}, \mathbf{Y})$ a platí

$$\text{var}^{-1}(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \text{var}^{-1} \mathbf{X} + \boldsymbol{\beta} \text{var}^{-1}(\mathbf{Y}|\mathbf{X})\boldsymbol{\beta}^T & -\boldsymbol{\beta} \text{var}^{-1}(\mathbf{Y}|\mathbf{X}) \\ -\text{var}^{-1}(\mathbf{Y}|\mathbf{X})\boldsymbol{\beta}^T & \text{var}^{-1}(\mathbf{Y}|\mathbf{X}) \end{pmatrix},$$

kde $\boldsymbol{\beta} = V^{-1} \cdot \text{cov}(\mathbf{X}, \mathbf{Y})$.

Důkaz: Viz Whittaker [6], strany 128, 142-143.

Z této věty plynou následující důsledky.

Věta 2.5: Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$. Necht' existuje matice $D = V^{-1}$. Potom každý diagonální prvek $d_{i,i}$, $i \in \{1, \dots, k\}$, je převrácenou hodnotou parciálního rozptylu, platí tedy

$$d_{i,i} = \text{var}^{-1}(X_i|N(\mathbf{X}; i)).$$

Důkaz: Viz Whittaker [6], strana 143.

Věta 2.6: Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$. Necht' existuje matice $P^{-1} = \text{cor}^{-1} \mathbf{X} = (p_{i,j}^0)$. Pak její diagonální prvky jsou

$$p_{i,i}^0 = \frac{\text{var} X_i}{\text{var}(X_i|N(\mathbf{X}; i))} = \frac{1}{1 - \rho_{X_i, N(\mathbf{X}; i)}^2},$$

kde $\rho_{X_i, N(\mathbf{X}; i)}$ je koeficient mnohonásobné korelace zavedený v odstavci 1.5.

Důkaz: Viz Whittaker [6], strana 143.

Věta 2.7: Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$. Necht' existuje matice $D = V^{-1}$. Pak mimodiagonální prvky škálované matice $S(D) = D^*$ jsou

$$d_{i,j}^* = -\rho_{X_i, X_j|N(\mathbf{X}; i, j)},$$

jestliže $i \neq j$. Jedná se tedy o hodnoty koeficientu parciální korelace zavedeného v odstavci 1.6 násobené (-1) .

Důkaz: Viz Whittaker [6], strany 143 - 144.

Matice D^* s prvky

$$d_{i,j}^* = \frac{d_{i,j}}{\sqrt{d_{i,i}d_{j,j}}}$$

má stejné prvky jako škálovaná inverzní korelační matice.

Věta 2.8: Mějme náhodný vektor $\mathbf{X} = (X_1, \dots, X_k)^T$. Necht' existuje matice $P^{-1} = \text{cor}^{-1} \mathbf{X} = (p_{i,j}^0)$. Potom mimodiagonální prvky škálované matice $S(P^{-1}) = P^*$ jsou

$$p_{i,j}^* = -\rho_{X_i, X_j|N(\mathbf{X}; i, j)}.$$

Důkaz: Tvrzení plyne ze vztahu mezi varianční a korelační maticí.

Věta 2.9: Mějme náhodné vektory \mathbf{X} a \mathbf{Y} s regulárními maticemi $\text{var } \mathbf{X}$, $\text{var } \mathbf{Y}$, $\text{var} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. Matici $D = \text{var}^{-1} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ vyjádříme ve tvaru

$$D = \begin{pmatrix} D_{X,X} & D_{X,Y} \\ D_{Y,X} & D_{Y,Y} \end{pmatrix}.$$

Blok $D_{X,Y}$ je nulový, právě když je nulová matice $\text{cov} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$.

Důkaz: Viz Whittaker [6], strana 144.

Věta 2.10: Mějme náhodné vektory \mathbf{X} , \mathbf{Y} , \mathbf{Z} a necht' jsou matice $\text{var } \mathbf{X}$, $\text{var } \mathbf{Y}$, $\text{var } \mathbf{Z}$, $\text{var} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}$ regulární. Matici $D = \text{var}^{-1} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \\ \mathbf{Z} \end{pmatrix}$ lze rozdělit na bloky

$$D = \begin{pmatrix} D_{X,X} & D_{X,Y} & D_{X,Z} \\ D_{Y,X} & D_{Y,Y} & D_{Y,Z} \\ D_{Z,X} & D_{Z,Y} & D_{Z,Z} \end{pmatrix}.$$

Blok $D_{Y,Z}$ je nulový, právě když je nulová $\text{cov}(\mathbf{Y}, \mathbf{Z} | \mathbf{X})$.

Důkaz: Viz Whittaker [6], strany 144-145.

2.5 Nezávislost a podmíněná nezávislost v normálním rozdělení

Věta 2.11: Náhodné vektory s normálním rozdělením \mathbf{X} , \mathbf{Y} jsou nezávislé ($\mathbf{X} \perp \mathbf{Y}$), právě když platí:

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = 0.$$

Důkaz: Viz Whittaker [6], strana 164.

Věta 2.12: Náhodné vektory s normálním rozdělením \mathbf{X} , \mathbf{Y} jsou nezávislé ($\mathbf{X} \perp \mathbf{Y}$), právě když platí:

$$D_{X,Y} = 0,$$

kde D je matice z věty 2.9.

Důkaz: Tvrzení plyne z vět 2.9 a 2.11.

Věta 2.13: Náhodné vektory s normálním rozdělením \mathbf{X} , \mathbf{Y} jsou nezávislé při pevném \mathbf{Z} ($\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$), právě když platí:

$$\text{cov}(\mathbf{X}, \mathbf{Y} | \mathbf{Z}) = 0.$$

Důkaz: Viz Whittaker [6], strana 164.

Věta 2.14: Náhodné vektory s normálním rozdělením \mathbf{X} , \mathbf{Y} jsou nezávislé při pevném \mathbf{Z} ($\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$), právě když platí:

$$D_{\mathbf{X}, \mathbf{Y}} = 0,$$

kde $D = \begin{pmatrix} D_{\mathbf{X}, \mathbf{X}} & D_{\mathbf{X}, \mathbf{Y}} & D_{\mathbf{X}, \mathbf{Z}} \\ D_{\mathbf{Y}, \mathbf{X}} & D_{\mathbf{Y}, \mathbf{Y}} & D_{\mathbf{Y}, \mathbf{Z}} \\ D_{\mathbf{Z}, \mathbf{X}} & D_{\mathbf{Z}, \mathbf{Y}} & D_{\mathbf{Z}, \mathbf{Z}} \end{pmatrix}$ je matice z věty 2.10.

Důkaz: Tvrzení plyne z vět 2.10 a 2.13.

Tato kritéria se používají k vyšetření nezávislosti (resp. podmíněné nezávislosti) náhodných vektorů z normálního rozdělení.

3. Lineární regresní model

3.1 Lineární regresní model

Při praktickém zpracování dat máme naměřeno n realizací náhodné veličiny Y , budeme je značit Y_1, \dots, Y_n , a n realizací náhodného vektoru $\mathbf{X} = (X_1, \dots, X_k)^T$. Ty budeme značit $\mathbf{X}_1, \dots, \mathbf{X}_n$. Předpokládejme nadále, že data byla standardizována. Realizace vektoru \mathbf{X} vytvoří řádky matice $\underline{\mathbf{X}}$, to jest

$$\underline{\mathbf{X}} = \begin{pmatrix} X_{1,1} & \dots & X_{1,k} \\ \dots & \dots & \dots \\ X_{n,1} & \dots & X_{n,k} \end{pmatrix}.$$

Sloupce matice \mathbf{X} označíme $\underline{X}_1, \dots, \underline{X}_k$, přičemž j -tý sloupec je tvořen n realizacemi j -té složky vektoru \mathbf{X} . První sloupec někdy bývá tvořen vektorem jedniček. Analogicky $\underline{\mathbf{Y}}$ bude sloupec realizací veličiny Y .

Realizace závisle proměnné Y a nezávisle proměnných X_1, \dots, X_k mají často formu časových řad. Lineární regresní model má pak tvar

$$Y_t = \beta_1 X_{t,1} + \dots + \beta_k X_{t,k} + e_t,$$

pro $t = 1, \dots, n$. V maticovém zápisu má tento model tvar

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e},$$

kde

hodnost matice $\underline{\mathbf{X}}$ je $k < n$,

$\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$ je vektor neznámých parametrů a

$\mathbf{e} = (e_1, \dots, e_n)^T$ je vektor náhodných chyb, pro které předpokládáme

$$\mathbf{E} \mathbf{e} = \mathbf{0}, \quad \text{var } \mathbf{e} = \sigma^2 \mathbf{I},$$

kde

\mathbf{I} je jednotková matice.

Z předpokladů plyne, že při daných hodnotách matice $\underline{\mathbf{X}}$ je

$$\mathbf{E} \underline{\mathbf{Y}} = \underline{\mathbf{X}}\boldsymbol{\beta}, \quad \text{var } \underline{\mathbf{Y}} = \sigma^2 \mathbf{I}.$$

Vektor $\boldsymbol{\beta}$ se odhaduje metodou nejmenších čtverců, tj. minimalizuje se výraz

$$S(\boldsymbol{\beta}) = (\underline{\mathbf{Y}} - \underline{\mathbf{X}}\boldsymbol{\beta})^T (\underline{\mathbf{Y}} - \underline{\mathbf{X}}\boldsymbol{\beta}).$$

Odhad vektoru $\boldsymbol{\beta}$ budeme značit $\mathbf{b} = (b_1, \dots, b_k)^T$.

Věta 3.1: *Platí*

$$\mathbf{b} = (\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^T \underline{\mathbf{Y}}.$$

Důkaz: Viz Anděl [2], strana 102.

Definice 3.1: Řekneme, že náhodný vektor $(X_1, \dots, X_k)^T$ má normální rozdělení $N(\boldsymbol{\mu}, V)$, V regulární, jestliže má sdruženou hustotu

$$f(x_1, \dots, x_k) = \frac{1}{(2\pi)^{\frac{k}{2}} |V|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\boldsymbol{\mu})^T V^{-1}(x-\boldsymbol{\mu})},$$

kde symbolem $|V|$ označujeme determinant matice V .

Věta 3.2: Odhad \mathbf{b} má rozdělení $N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$.

Důkaz: Viz Anděl [2], strana 103.

Označme

$$S_e = \mathbf{Y}^T \mathbf{Y} - \mathbf{b}^T \mathbf{X}^T \mathbf{Y}.$$

S_e nazýváme reziduálním součtem čtverců.

Věta 3.3: Nechť náhodný vektor \mathbf{e} v lineárním regresním modelu má normální rozdělení $N(0, \sigma^2 I)$. Označme $U = (\mathbf{X}^T \mathbf{X})^{-1}$ matici s prvky $u_{i,j}$, $s^2 = \frac{S_e}{n-k}$. Pak pro každé $i = 1, \dots, k$ má náhodná veličina

$$T_i = \frac{b_i - \beta_i}{\sqrt{s^2 u_{i,i}}}$$

rozdělení t_{n-k} .

Důkaz: Viz Anděl [2], strana 103.

Větu 3.3 používáme k testování hypotézy, že parametr β_i je nulový. Hypotézu zamítáme ve prospěch oboustranné alternativy, když

$$|T_i| \geq t_{n-k}(\alpha),$$

kde $t_{n-k}(\alpha)$ je kritická hodnota rozdělení t_{n-k} na hladině α .

3.2 Testování submodelů

Rozložme nyní vektor \mathbf{X} na podvektory \mathbf{X}_a a \mathbf{X}_b :

$$\mathbf{X} = \mathbf{X}_a \cup \mathbf{X}_b; \quad \mathbf{X}_a = (X_1, \dots, X_r)^T; \quad \mathbf{X}_b = (X_{r+1}, \dots, X_k)^T.$$

Nechť $\hat{Y}(\mathbf{X})$ značí nejlepší lineární aproximaci Y pomocí X_1, \dots, X_k :

$$\hat{Y}(\mathbf{X}) = \beta_1 X_1 + \dots + \beta_k X_k = \boldsymbol{\beta}^T \mathbf{X},$$

kde

$$\boldsymbol{\beta} = (\text{var } \mathbf{X})^{-1} \text{cov}(\mathbf{X}, Y).$$

Rozptyl tohoto odhadu je roven

$$\text{var } \hat{Y}(\mathbf{X}) = \boldsymbol{\beta}^T \text{var } \mathbf{X} \boldsymbol{\beta} = \text{cov}(Y, \mathbf{X})(\text{var } \mathbf{X})^{-1} \text{cov}(\mathbf{X}, Y)$$

a reziduální (parciální) rozptyl je roven

$$\text{var}(Y - \widehat{Y}(\mathbf{X})) = \text{var} Y - \text{var} \widehat{Y}(\mathbf{X}) = \text{var}(Y|\mathbf{X}).$$

Uvažujme hypotézu H_0 tvrdící, že Y závisí pouze na r prvních složkách vektoru \mathbf{X} , tedy na vektoru \mathbf{X}_a .

Za platnosti H_0 máme:

$$\beta_{r+1} = \dots = \beta_k = \mathbf{0}.$$

Nechť $\widehat{Y}(\mathbf{X}_a)$ značí nejlepší lineární aproximaci Y pomocí X_1, \dots, X_r . Platí

$$\widehat{Y}(\mathbf{X}_a) = \beta_1 X_1 + \dots + \beta_r X_r = \boldsymbol{\beta}_a^T \mathbf{X}_a,$$

kde

$$\boldsymbol{\beta}_a = (\text{var} \mathbf{X}_a)^{-1} \text{cov}(\mathbf{X}_a, Y).$$

Rozptyl tohoto odhadu je roven

$$\text{var} \widehat{Y}(\mathbf{X}_a) = \boldsymbol{\beta}_a^T \text{var} \mathbf{X}_a \boldsymbol{\beta}_a = \text{cov}(Y, \mathbf{X}_a) (\text{var} \mathbf{X}_a)^{-1} \text{cov}(\mathbf{X}_a, Y)$$

a reziduální (parciální) rozptyl je v tomto případě roven

$$\text{var}(Y - \widehat{Y}(\mathbf{X}_a)) = \text{var} Y - \text{var} \widehat{Y}(\mathbf{X}_a) = \text{var}(Y|\mathbf{X}_a).$$

Označme

\underline{Y} n realizací proměnné Y ,

$\underline{\mathbf{X}}$ n realizací vektoru proměnných X_1, \dots, X_k ,

$\underline{\mathbf{X}}_a$ n realizací vektoru proměnných X_1, \dots, X_r ,

$$S_{Y,Y} = \frac{1}{n-1} \underline{Y}^T \underline{Y} \text{ odhad pro var } Y,$$

$$S_{\mathbf{X},Y} = \frac{1}{n-1} \underline{\mathbf{X}}^T \underline{Y} \text{ odhad pro cov}(\mathbf{X}, Y),$$

$$S_{\mathbf{X},\mathbf{X}} = \frac{1}{n-1} \underline{\mathbf{X}}^T \underline{\mathbf{X}} \text{ odhad pro var } \mathbf{X},$$

$$S_{\mathbf{X}_a,Y} = \frac{1}{n-1} \underline{\mathbf{X}}_a^T \underline{Y} \text{ odhad pro cov}(\mathbf{X}_a, Y),$$

$$S_{\mathbf{X}_a,\mathbf{X}_a} = \frac{1}{n-1} \underline{\mathbf{X}}_a^T \underline{\mathbf{X}}_a \text{ odhad pro var } \mathbf{X}_a.$$

Máme tedy

$$\widehat{\text{var}} Y = S_{Y,Y}, \quad \widehat{\text{var}} \widehat{Y}(\mathbf{X}) = S_{Y,\mathbf{X}} S_{\mathbf{X},\mathbf{X}}^{-1} S_{\mathbf{X},Y}, \quad \widehat{\text{var}} \widehat{Y}(\mathbf{X}_a) = S_{Y,\mathbf{X}_a} S_{\mathbf{X}_a,\mathbf{X}_a}^{-1} S_{\mathbf{X}_a,Y}.$$

Uvažujme model

$$Y = \beta_1 X_1 + \dots + \beta_k X_k + e,$$

kde $e \sim N(0, \sigma^2)$.

Chceme testovat hypotézu $H_0 : \beta_{r+1} = \dots = \beta_k = \mathbf{0}$.

Zabývejme se nejprve variabilitou proměnné Y . Můžeme ji popsat následující tabulkou:

Zdroj variability	Součet čtverců	Stupňů volnosti	F-statistika
X_b při pevném X_a	SS_b	$f_b = k - r$	$F = \frac{SS_b/f_b}{SS_e/f_e}$, $F \sim F_{f_b, f_e} = F_{k-r, n-k}$
X_a	SS_a	$f_a = r$	
reziduální	SS_e	$f_e = n - k$	
celkový	SS_T	$f_T = n$	

Tabulka 3.1: Tabulka analýzy rozptylu pro F-test nulovosti skupiny regresních parametrů pro standardizovaná data

SS_b , SS_a , SS_e , SS_T označují v tabulce následující součty čtverců:

$$SS_b = (n - 1)S_{Y, X} S_{X, X}^{-1} S_{X, Y} - (n - 1)S_{Y, X_a} S_{X_a, X_a}^{-1} S_{X_a, Y},$$

$$SS_a = (n - 1)S_{Y, X_a} S_{X_a, X_a}^{-1} S_{X_a, Y},$$

$$SS_e = (n - 1)S_{Y, Y} - (n - 1)S_{Y, X} S_{X, X}^{-1} S_{X, Y},$$

$$SS_T = (n - 1)S_{Y, Y},$$

tedy s využitím výše uvedeného a věty 2.3 máme:

$$SS_b \text{ je odhad pro } (n - 1)[\text{var } \hat{Y}(\mathbf{X}) - \text{var } \hat{Y}(\mathbf{X}_a)] = (n - 1)[(\text{var } \hat{Y}(\mathbf{X}) - \text{var } Y) - (\text{var } \hat{Y}(\mathbf{X}_a) - \text{var } Y)] = (n - 1)[\text{var}(Y | \mathbf{X}_a) - \text{var}(Y | \mathbf{X})],$$

$$SS_a \text{ je odhad pro } (n - 1) \text{var } \hat{Y}(\mathbf{X}_a) = (n - 1)[\text{var} Y - \text{var}(Y | \mathbf{X}_a)],$$

$$SS_e \text{ je odhad pro } (n - 1)[\text{var } Y - \text{var } \hat{Y}(\mathbf{X})] = (n - 1) \text{var}(Y | \mathbf{X}),$$

$$SS_T \text{ je odhad pro } (n - 1) \text{var } Y.$$

Celkovou variabilitu proměnné Y můžeme tedy rozložit na součet reziduální variability, variability způsobené vektorem X_a a variability způsobené vektorem X_b při pevném vektoru X_a :

$$\begin{aligned} \frac{1}{n - 1} SS_T = \widehat{\text{var}} Y &= [\widehat{\text{var}} Y - \widehat{\text{var}} \hat{Y}(\mathbf{X})] + [\widehat{\text{var}} \hat{Y}(\mathbf{X}_a)] + [\widehat{\text{var}} \hat{Y}(\mathbf{X}) - \widehat{\text{var}} \hat{Y}(\mathbf{X}_a)] = \\ &= \frac{1}{n - 1} (SS_e + SS_a + SS_b). \end{aligned}$$

Rovněž lze psát rozklad

$$\begin{aligned} \frac{1}{n - 1} SS_T = \widehat{\text{var}} Y &= \widehat{\text{var}}(Y | \mathbf{X}) + [\widehat{\text{var}} Y - \widehat{\text{var}}(Y | \mathbf{X}_a)] + [\widehat{\text{var}}(Y | \mathbf{X}_a) - \widehat{\text{var}}(Y | \mathbf{X})] = \\ &= \frac{1}{n - 1} (SS_e + SS_a + SS_b). \end{aligned}$$

Často musíme současně se základním modelem vzít do úvahy i jeho submodely. Ve většině případů vznikne submodel vynecháním několika parametrů původního modelu. Často vzniká celý řetěz submodelů. Submodely se volí tak, aby měly při popisu dané situace přímou a snadnou interpretaci.

Mějme náhodný vektor $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ a předpokládejme, že platí model

$$M : \quad E Y = \beta_1 X_1 + \dots + \beta_k X_k$$

s maticí modelu

$$\underline{\mathbf{X}} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,k} \\ \cdots & \cdots & \cdots \\ X_{n,1} & \cdots & X_{n,k} \end{pmatrix}.$$

Nechť hodnost této matice je

$$h(\underline{\mathbf{X}}) = k.$$

Nejlepší nestranný lineární odhad pro $E \underline{Y}$ v případě platnosti modelu M je:

$$\hat{\underline{\mu}} = \underline{\mathbf{X}}(\underline{\mathbf{X}}^T \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^T \underline{Y}$$

Dále uvažujme submodel

$$M_1 : \quad E Y = \beta_1 X_1 + \cdots + \beta_r X_r, \quad r < k,$$

s maticí submodelu

$$\underline{\mathbf{X}}_a = \begin{pmatrix} X_{1,1} & \cdots & X_{1,r} \\ \cdots & \cdots & \cdots \\ X_{n,1} & \cdots & X_{n,r} \end{pmatrix}.$$

Nechť hodnost této matice je

$$h(\underline{\mathbf{X}}_a) = r,$$

pak nejlepší nestranný lineární odhad pro $E \underline{Y}$ v případě platnosti modelu M_1 je

$$\hat{\underline{\nu}} = \underline{\mathbf{X}}_a(\underline{\mathbf{X}}_a^T \underline{\mathbf{X}}_a)^{-1} \underline{\mathbf{X}}_a^T \underline{Y}.$$

Uvažujme další submodel:

$$M_2 : \quad E Y = \beta_1 X_1 + \cdots + \beta_s X_s, \quad s < r,$$

s maticí submodelu

$$\underline{\mathbf{X}}_{aa} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,s} \\ \cdots & \cdots & \cdots \\ X_{n,1} & \cdots & X_{n,s} \end{pmatrix}$$

Nechť hodnost této matice je

$$h(\underline{\mathbf{X}}_{aa}) = s,$$

pak nejlepší nestranný lineární odhad pro $E \underline{Y}$ v případě platnosti modelu M_2 je

$$\hat{\underline{\tau}} = \underline{\mathbf{X}}_{aa}(\underline{\mathbf{X}}_{aa}^T \underline{\mathbf{X}}_{aa})^{-1} \underline{\mathbf{X}}_{aa}^T \underline{Y}$$

Věta 3.4: Jestliže pro Y platí model M_1 , náhodná veličina

$$F_1 = \frac{(\hat{\underline{\mu}}^T \hat{\underline{\mu}} - \hat{\underline{\nu}}^T \hat{\underline{\nu}})/(k - r)}{(\underline{Y}^T \underline{Y} - \hat{\underline{\mu}}^T \hat{\underline{\mu}})/(n - k)},$$

má rozdělení $F_{k-r, n-k}$. Jestliže pro Y platí platí model M_2 , pak náhodná veličina

$$F_2 = \frac{(\hat{\underline{\nu}}^T \hat{\underline{\nu}} - \hat{\underline{\tau}}^T \hat{\underline{\tau}})/(r - s)}{(\underline{Y}^T \underline{Y} - \hat{\underline{\mu}}^T \hat{\underline{\mu}})/(n - k)},$$

má rozdělení $F_{r-s, n-k}$.

Důkaz: Viz Anděl [1], strana 143.

Hypotéza ${}_1H_0: \beta_{r+1} = \dots = \beta_k = 0$ představuje submodel M_1 modelu M . F-statistika pro test M_1 proti M , to jest pro test ${}_1H_0: \beta_{r+1} = \dots = \beta_k = 0$, je v případě standardizovaných dat

$$F_1 = \frac{\frac{S_{Y,X}(S_{X,X})^{-1}S_{X,Y} - S_{Y,X_a}(S_{X_a,X_a})^{-1}S_{X_a,Y}}{k-r}}{\frac{S_{Y,Y} - S_{Y,X}(S_{X,X})^{-1}S_{X,Y}}{n-k}} = \frac{\frac{SS_b}{k-r}}{\frac{SS_e}{n-k}}.$$

Dostali jsme tedy F-statistiku uvedenou v tabulce analýzy rozptylu 3.1.

A podobně F-statistika pro test M_2 proti M_1 jako submodelu M , tzn. pro test ${}_2H_0: \beta_{s+1} = \dots = \beta_r (= \beta_{r+1} = \dots = \beta_k) = 0$ je

$$F_2 = \frac{\frac{S_{Y,X_a}(S_{X_a,X_a})^{-1}S_{X_a,Y} - S_{Y,X_{aa}}(S_{X_{aa},X_{aa}})^{-1}S_{X_{aa},Y}}{r-s}}{\frac{S_{Y,Y} - S_{Y,X}(S_{X,X})^{-1}S_{X,Y}}{n-k}} = \frac{\frac{SS_{ba}}{r-s}}{\frac{SS_e}{n-k}},$$

kde

$$SS_{ba} = (n-1)S_{Y,X_a}S_{X_a,X_a}^{-1}S_{X_a,Y} - (n-1)S_{Y,X_{aa}}S_{X_{aa},X_{aa}}^{-1}S_{X_{aa},Y},$$

je odhad pro

$$\begin{aligned} (n-1)[\text{var } \hat{Y}(\mathbf{X}_a) - \text{var } \hat{Y}(\mathbf{X}_{aa})] &= (n-1)[(\text{var } \hat{Y}(\mathbf{X}_a) - \text{var } Y) - (\text{var } \hat{Y}(\mathbf{X}_{aa}) - \text{var } Y)] = \\ &= (n-1)[\text{var}(Y|\mathbf{X}_{aa}) - \text{var}(Y|\mathbf{X}_a)]. \end{aligned}$$

Speciálním případem je $r = k - 1$:

Model

$$Y = \beta_1 X_1 + \dots + \beta_k X_k,$$

submodel

$$Y = \beta_1 X_1 + \dots + \beta_{k-1} X_{k-1},$$

$$H_0: \beta_k = 0,$$

F-statistika pro test H_0 :

$$F = \frac{SS_b/k - (k-1)}{SS_e/n - k} \sim F_{1, n-k}$$

3.3 Koeficient determinace

Mějme náhodný výběr $\begin{pmatrix} Y_1 \\ \mathbf{X}_1 \end{pmatrix}, \dots, \begin{pmatrix} Y_n \\ \mathbf{X}_n \end{pmatrix}$ z nějakého $(k+1)$ -rozměrného rozdělení. Označme $r_{i,j}$ výběrový korelační koeficient mezi i -tou a j -tou složkou vektoru \mathbf{X} a $r_{0,i}$ označme výběrový korelační koeficient mezi veličinami Y a X_i . V odstavci 1.5 jsme zavedli výběrové korelační matice

$$R_{\mathbf{X},\mathbf{X}} = (r_{i,j})_{i,j=1}^k, \quad R_{Y,\mathbf{X}} = (r_{0,i})_{i=1}^k, \quad R_{\mathbf{X},Y} = R_{Y,\mathbf{X}}^T.$$

a definovali výběrový koeficient mnohonásobné korelace předpisem

$$r_{Y,X}^2 = R_{Y,X} R_{X,X}^{-1} R_{X,Y}.$$

Tento kvadrát výběrového koeficientu mnohonásobné korelace se také nazývá koeficientem determinace lineárního regresního modelu závislosti Y na X_1, \dots, X_k . Lze jej vyjádřit ve tvaru

$$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

kde $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n) = \mathbf{X}\mathbf{b}$, přičemž matice \mathbf{X} a vektor \mathbf{b} jsou dány lineárním regresním modelem z odstavce 3.1.

Hodnoty koeficientu determinace blízké 1 svědčí o tom, že lineární regresní model dobře vysvětluje variabilitu dat Y_1, \dots, Y_n .

3.4 Model CAPM

Model oceňování kapitálových aktiv (CAPM) je lineárním regresním modelem, který byl vyvinut pro popis očekávaného výnosu portfolia případně individuálních cenných papírů v závislosti na riziku.

Výnosem portfolia označme vážený průměr výnosů jednotlivých cenných papírů, kde váhami jsou podíly cenných papírů v portfoliu. Rozptyl má vztah k riziku portfolia, čím větší je rozptyl, tím větší je riziko. Označme rozptyl výnosu portfolia σ_p^2 . Tento rozptyl je definován vztahem

$$\sigma_p^2 = E(r_p - \bar{r}_p)^2 = E\left[\sum_{i=1}^N \theta_i (r_i - \bar{r}_i)\right]^2 = \sum_{i=1}^N \sum_{j=1}^N \theta_i \theta_j \sigma_{i,j} = \sum_{i=1}^N \sum_{j=1}^N \theta_i \theta_j \sigma_i \sigma_j \rho_{i,j},$$

kde

r_p je výnos portfolia,

\bar{r}_p je střední hodnota výnosu portfolia,

N je počet různých cenných papírů v portfoliu,

θ_i je procentní podíl portfolia držený v i -tém cenném papíru,

r_i je výnos i -tého cenného papíru v portfoliu,

\bar{r}_i je střední hodnota výnosu i -tého cenného papíru v portfoliu,

$\sigma_{i,j}$ je kovariance mezi výnosy i -tého a j -tého cenného papíru,

σ_i je směrodatná odchylka výnosu i -tého cenného papíru a

$\rho_{i,j}$ je korelace mezi výnosy i -tého a j -tého cenného papíru.

Investor se bude snažit snížit celkové riziko diverzifikací, což je proces kombinování instrumentů v portfoliu s cílem omezit celkové riziko, aniž by bylo nutné obětovat výnos z portfolia.

Efektivním nazvěme takové portfolio, pro něž na trhu neexistuje portfolio se stejným rozptylem a vyšším výnosem nebo s nižším rozptylem a stejným výnosem. Předpokládejme portfolio složené z jediného rizikového instrumentu x_1 a bezrizikového instrumentu x_f . Očekávaný výnos tohoto portfolia bude

$$\bar{r}_p = \theta \bar{r}_1 + (1 - \theta) r_f,$$

kde

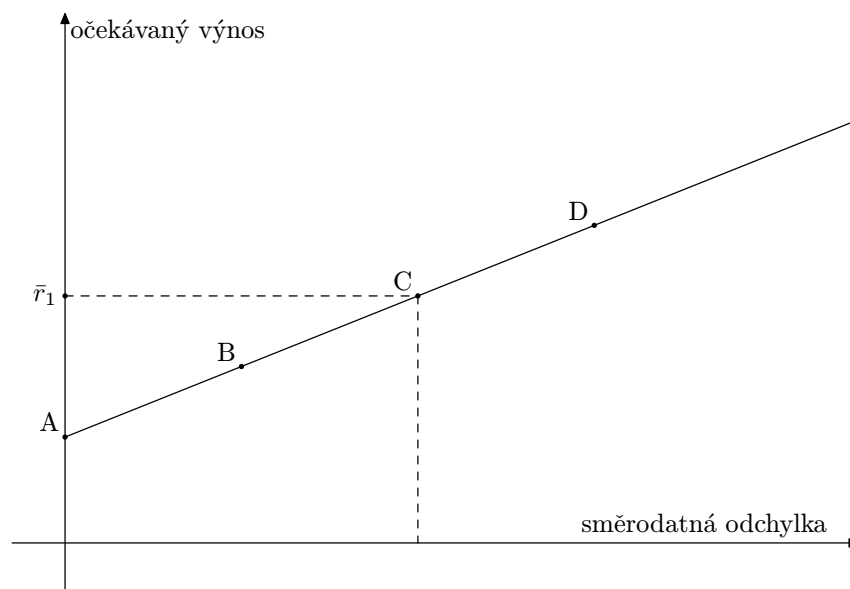
\bar{r}_1 značí očekávaný výnos instrumentu x_1 ,

r_f značí bezrizikovou míru výnosnosti,

θ je podíl z celkové hodnoty portfolia držený v rizikovém instrumentu.

Směrodatná odchylka portfolia je rovna θ násobku směrodatné odchylky rizikového instrumentu.

Předchozí rovnice spolu s úměrou směrodatné odchylky vede k lineární množině portfolií, která je vidět na obrázku 3.1.



Obrázek 3.1

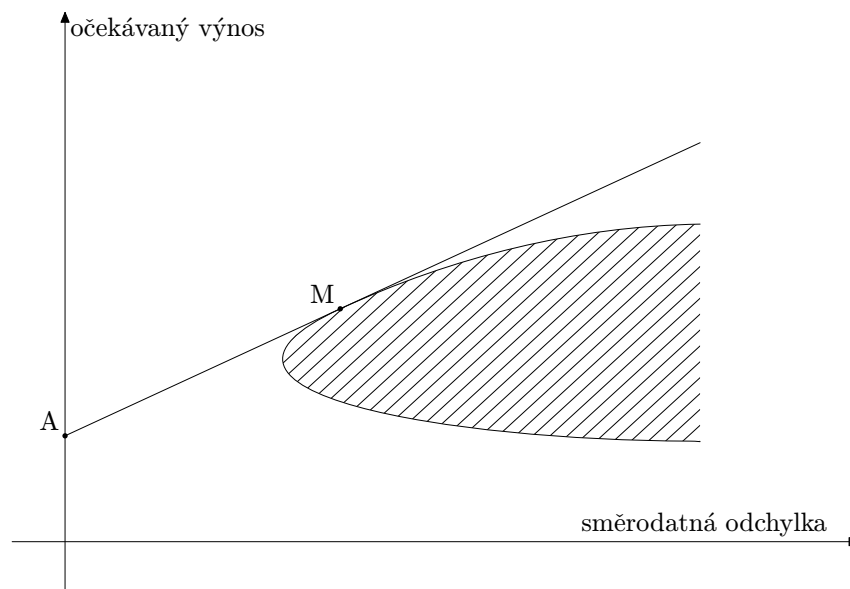
V bodě A investor investuje vše do bezrizikového aktiva. Celkový výnos je r_f , riziko nulové. V bodě C investor investuje vše do rizikového aktiva x_1 s očekávaným výnosem \bar{r}_1 . Bod B je bodem, ve kterém je část majetku investována do bezrizikového a zbytek do rizikového aktiva. V bodě D si investor vypůjčil za bezrizikovou sazbu a tuto výpůjčku vložil spolu s původním majetkem do rizikového aktiva, nakrátko prodal bezrizikové aktivum, aby mohl investovat do rizikového aktiva. Podobně to bude vypadat, i když v portfoliu bude více rizikových aktiv.

Efektivní množinu portfolií získáme kombinací bezrizikového aktiva s rizikovým portfoliem v bodě M, který leží na tečně AM konvexní množiny rizikových portfolií na obrázku 3.2.

Jestliže existuje bezrizikové aktivum, které lze půjčovat a vypůjčovat si za stejnou bezrizikovou úrokovou sazbu, zahrnuje efektivní množina portfolií všechna portfolia, která jsou kombinacemi rizikového portfolia M (označuje se jako tangenciální portfolio rizikových aktiv) a bezrizikového aktiva. Na obrázku 3.2 je efektivní množinou portfolií polopřímka AM. Efektivní množina portfolií je tedy lineární a označuje se jako čára kapitálového trhu. Je dána rovnicí

$$\bar{r}_p = r_f + \frac{\bar{r}_m - r_f}{\sigma_m} \sigma_p,$$

kde



Obrázek 3.2

\bar{r}_p je očekávaný výnos portfolia,

\bar{r}_m je očekávaný výnos tangenciálního portfolia,

r_f je bezriziková úroková sazba,

σ_m je směrodatná odchylka tangenciálního portfolia a

σ_p je směrodatná odchylka portfolia.

Tržní cenou rizika nazvěme směrnici čáry kapitálového trhu. Tato směrnice ukazuje poměr, ve kterém se vyrovnává riziko a výnos:

$$\text{tržní cena rizika} = \frac{\bar{r}_m - r_f}{\sigma_m}.$$

Analyzujeme-li závislost mezi očekávaným výnosem a směrodatnou odchylkou jediného cenného papíru, získáme CAPM jako model

$$\bar{r}_i = r_f + \left(\frac{\bar{r}_m - r_f}{\sigma_m} \right) \left(\frac{\sigma_{i,m}}{\sigma_m} \right)$$

nebo rovnicí

$$\bar{r}_i = r_f + (\bar{r}_m - r_f)\beta_i,$$

kde

\bar{r}_i je očekávaný výnos i-tého cenného papíru,

$\sigma_{i,m} = \sum_{j=1}^N \theta_j \sigma_{i,j}$ je kovariance výnosů i-tého cenného papíru a trhu,

$\sigma_{i,j}$ jsou kovariance mezi výnosy i-tého a j-tého cenného papíru,

θ_j je podíl j-tého cenného papíru v tangenciálním portfoliu M,

$\beta_i = \frac{\sigma_{i,m}}{\sigma_m^2}$ se nazývá beta i-tého cenného papíru. Odhaduje se regresními metodami z historických dat.

Uvedená rovnice se označuje jako čára trhu cenného papíru.

Hodnota $\beta_i \sigma_m^2$ představuje nediverzifikovatelnou složku rizika. Část rizika, kterou lze eliminovat diverzifikací portfolia, model CAPM neoceňuje.

3.5 Model APT (Arbitrážní cenový model)

Model APT byl navržen S. A. Rossem jako alternativa k modelu CAPM. Předpokládá, že poskytuje-li několik cenných papírů a portfolií stejný očekávaný výnos při stejném riziku, musí se prodávat za stejnou cenu. Jestliže tomu tak zpočátku není, investoři eliminují tuto nerovnost arbitráží.

Tento model vychází ze vztahu

$$r_i = \bar{r}_i + \sum_{j=1}^K \delta_j \beta_{i,j} + \varepsilon_i,$$

kde

r_i je skutečný výnos i -tého cenného papíru,

\bar{r}_i očekávaný výnos i -tého cenného papíru,

δ_j je faktor společný pro výnosy všech cenných papírů, jehož střední hodnota je rovna nule,

koeficient $\beta_{i,j}$ měří odezvu i -tého výnosu na j -tý společný faktor, tyto parametry se zpravidla odhadují z historických dat pomocí lineární regrese,

společné faktory δ_j jsou nediverzifikovatelné složky rizika a

ε_i je diverzifikovatelná složka rizika, charakteristická pro samotné aktivum i .

3.6 Mnohorozměrná regrese

Předpokládejme model definovaný rovnicí

$$\underline{Y} = \underline{X}\underline{B} + \underline{U},$$

kde

\underline{Y} je matice typu $n \times p$ reprezentující n pozorování p -rozměrného vektoru závisle proměnných,

\underline{X} je známá matice typu $n \times q$,

\underline{B} je matice neznámých regresních parametrů, tato matice je typu $q \times p$,

\underline{U} je matice náhodných chyb, tato matice je typu $n \times p$ a její řádky jsou nekorelované při dané matici \underline{X} se střední hodnotou $\mathbf{0}$ a s kovarianční maticí Σ .

Jestliže \underline{X} reprezentuje matici n pozorování q -rozměrného vektoru nezávisle proměnných, je uvažovaný model nazýván modelem mnohorozměrné regrese.

Ve speciálním případě $p = 1$ se dostáváme k lineárnímu regresnímu modelu z kapitoly 3.1.

Nejčastěji se předpokládá, že řádky \underline{U} pocházejí z mnohorozměrného normálního rozdělení $N(\mathbf{0}, \Sigma)$, přičemž \underline{U} nezávisí na \underline{X} . Za předpokladu normálního rozdělení chyb je logaritmická věrohodnostní funkce pro data \underline{Y} jako funkční hodnotu matic parametrů \underline{B} , Σ dána vztahem

$$l(\underline{B}, \Sigma) = -\frac{1}{2}n \log |2\pi\Sigma| - \frac{1}{2} \text{tr}(\underline{Y} - \underline{X}\underline{B})\Sigma^{-1}(\underline{Y} - \underline{X}\underline{B})^T.$$

Dále budeme předpokládat $n \geq p + q$

Věta 3.5: Za předpokladu normality jsou maximálně věrohodnými odhady

$$\hat{\underline{B}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$$

a

$$\hat{\underline{\Sigma}} = \frac{\underline{Y}^T \underline{P} \underline{Y}}{n},$$

kde $\underline{P} = \underline{I} - \underline{X}(\underline{X}^T \underline{X})^{-1} \underline{X}^T$.

Důkaz: Viz Mardia a kol. [5], strana 158.

4. Grafické modely

4.1 Základní pojmy

Grafem nazveme objekt skládající se ze dvou množin, množiny vrcholů V a množiny hran E , což je množina dvojic vrcholů z V . Řekneme, že E obsahuje orientovanou hranu mezi body i a j , jestliže E obsahuje uspořádanou dvojici (i, j) , přičemž vrchol i nazveme počátečním a vrchol j koncovým. Graficky znázorníme orientovanou hranu šipkou z vrcholu počátečního do vrcholu koncového. Jestliže E obsahuje uspořádané dvojice (i, j) a (j, i) , řekneme, že mezi bodem i a bodem j je neorientovaná hrana. Graficky znázorníme neorientovanou hranu úsečkou mezi danými vrcholy. Graf nazveme neorientovaným, jestliže všechny hrany jsou neorientované. Vrcholy nazveme sousedními, jestliže je mezi nimi neorientovaná hrana. Cesta z i_1 do i_n je posloupnost navzájem různých vrcholů i_1, \dots, i_n , ve které $(i_k, i_{k+1}) \in E$ pro každé $k = 1, \dots, n - 1$. Cestu z i_1 do i_n nazveme nejkratší, jestliže žádná její podposloupnost není cestou z i_1 do i_n . Cestu nazveme cyklem, jestliže $i_1 = i_n$. Cyklus v neorientovaném grafu nazveme kružnice. Graf je souvislý, jestliže mezi každými dvěma vrcholy existuje cesta. Podmnožina vrcholů separuje vrcholy i a j , jestliže každá cesta z i do j obsahuje nějaký vrchol z dané podmnožiny. Nechť L je podmnožinou V . Množina sousedů L je množina vrcholů nenáležících L , z nichž každý je spojen hranou s nějakým vrcholem z L . Podgrafem grafu G nazveme graf $G_1 = (V_1, E_1)$, jehož množina vrcholů je podmnožinou V a jehož množina hran je podmnožinou E . Jestliže platí $V_1 = V$, nazveme pograf G_1 faktorem grafu G . Podgrafem indukovaným množinou L je podgraf vzniklý z grafu G vynecháním všech vrcholů nenáležících L a všech hran obsahujících některý z vynechaných vrcholů. Graf nazveme úplný, jestliže jsou všechny vrcholy spojeny orientovanými nebo neorientovanými hranami. Klika je maximální úplný podgraf.

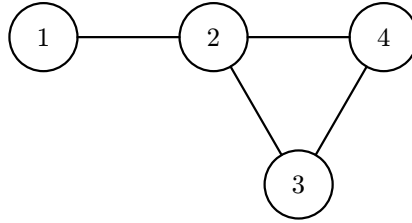
4.2 Grafy a podmíněná nezávislost

Definice 4.1: Mějme náhodné veličiny X_1, \dots, X_k a množinu vrcholů $V = \{1, 2, \dots, k\}$. Neorientovaný graf nazveme grafem podmíněných nezávislostí, jestliže mezi dvěma vrcholy chybí hrana právě v případě, že náhodné veličiny, ke kterým tyto vrcholy přísluší, jsou podmíněně nezávislé při pevných hodnotách ostatních náhodných veličin. Budeme používat např. značení $1 \perp 2 \mid \{3, 4\}$ pro $X_1 \perp X_2 \mid \{X_3, X_4\}$.

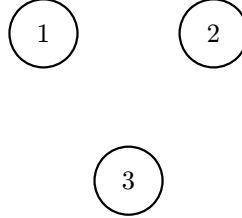
Na obrázku 4.1 je graf nezávislosti, pro $X_1 \perp X_4 \mid \{X_2, X_3\}$ a $X_1 \perp X_3 \mid \{X_2, X_4\}$. Podgraf tvořený vrcholy 2, 3, 4 je klika, vrchol 2 separuje vrcholy 1 a 4 a dále separuje vrcholy 1 a 3.

Obrázek 4.2 zachycuje graf $X_1 \perp X_2 \mid X_3$, $X_1 \perp X_3 \mid X_2$, $X_2 \perp X_3 \mid X_1$, který rovněž zapisujeme ve tvaru $X_1 + X_2 + X_3$.

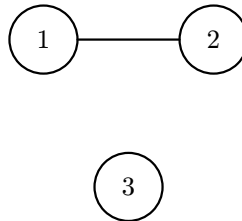
Na obrázku 4.3 je vidět znázornění grafu $X_1 \perp X_3 \mid X_2$, $X_2 \perp X_3 \mid X_1$, který rovněž zapisujeme ve tvaru $X_1 X_2 + X_3$.



Obrázek 4.1



Obrázek 4.2



Obrázek 4.3

Definice 4.2: Grafický model s grafem $G = (V, E)$ je systém pravděpodobnostních rozdělení náhodného vektoru $\mathbf{X} = (X_1, \dots, X_k)^T$, která splňují podmíněné nezávislosti dané k -vrcholovým grafem G .

4.3 Řetězové grafy podmíněných nezávislostí

Uvažujme graf $G = (V, E)$, $V = \{1, \dots, k\}$, jehož vrcholy jsou rozděleny do bloků b_1, \dots, b_r , $r \leq k$, $\bigcup_{i=1}^r b_i = V$, $b_i \cap b_j = \emptyset$, $i \neq j$. Zaveďme na množině bloků úplné uspořádání \prec_b předpisem $b_i \prec_b b_j \Leftrightarrow i < j$. Na množině vrcholů V zaveďme částečné uspořádání \preceq předpisem

$$v \prec w \Leftrightarrow v \in b_i, w \in b_j, i < j,$$

$$v = w \Leftrightarrow v, w \in b_i.$$

Označme dále $K(w) = \bigcup_{i \leq r(w)} b_i$, kde $r(w)$ je index bloku obsahujícího vrchol w .

Definice 4.3: Řetězovým grafem podmíněných nezávislostí náhodných veličin

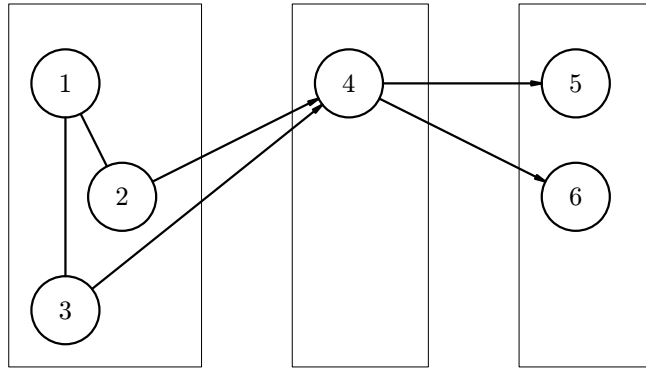
X_1, X_2, \dots, X_k nazveme graf $G = (V, E)$ s výše uvedenými uspořádáními, v němž pro $v \preceq w$ platí

$$(v, w) \notin E \Leftrightarrow X_v \perp X_w | \{X_u, u \in K(w) \setminus \{v, w\}\},$$

$$(v, w) \in E, v \prec w \Rightarrow (w, v) \notin E,$$

$$(v, w) \in E, v = w \Rightarrow (w, v) \in E.$$

V řetězovém grafu podmíněných nezávislostí jsou tedy hrany spojující vrcholy z různých bloků orientované a hrany spojující vrcholy z téhož bloku neorientované, jak vidíme na obrázku 4.4.



Obrázek 4.4

Definice 4.4: Mějme $\mathbf{X}_1, \dots, \mathbf{X}_n$ n realizací z náhodného vektoru $\mathbf{X} = (X_1, \dots, X_k)^T$ s rozdělením $N(0, V)$. Pak logaritmická věrohodnostní funkce má tvar

$$\begin{aligned} l(V) &= \log\left[\left(\frac{1}{(\sqrt{2\pi})^k \sqrt{|V|}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n \mathbf{X}_i^T V^{-1} \mathbf{X}_i}\right] = \\ &= -\frac{nk}{2} \log(2\pi) - \frac{n}{2} \log|V| - \frac{1}{2} \sum_{i=1}^n \mathbf{X}_i^T V^{-1} \mathbf{X}_i. \end{aligned}$$

Definice 4.5: Satureovaný model je grafický model určený úplným grafem.

Definice 4.6: Označme M model s vynechanými hranami. Pak jeho deviancí označíme

$$\text{dev}(M) = 2[l(S) - l(\hat{V})],$$

kde S je výběrová varianční matice v satureovaném modelu a \hat{V} je maximálně věrohodný odhad varianční matice konstruovaný při omezeních daných modelem M . Tento odhad lze počítat iteračně pomocí tzv. IPF algoritmu - viz Whittaker [6], str. 182-184.

Definice 4.7: Diference deviancí Mějme G_0 úplný graf a G_1 a G_2 jeho faktory, ve kterých chybí f_1 a f_2 hran. Pak definujeme diferenci deviancí modelů M_1 a M_2 s grafy $G_2 \subset G_1$ jako $\text{dev}_{G_2, G_1}^* = \text{dev}(G_2) - \text{dev}(G_1)$.

Věta 4.1: Deviance jako testové kritérium Deviance $\text{dev}(G)$ má asymptoticky χ^2 rozdělení. Počet stupňů volnosti je roven počtu vynechaných hran v grafu G . Diference deviancí $\text{dev}_{G_2, G_1}^* = \text{dev}(G_2) - \text{dev}(G_1)$ pro grafy $G_2 \subset G_1$ s f_2 a f_1 vynechanými hranami má asymptoticky χ^2 rozdělení o $f_2 - f_1$ stupních volnosti.

Důkaz: Viz Whittaker [6], strana 245.

Pomocí této věty lze testovat model M_2 s grafem G_2 proti alternativě modelu M_1 s grafem G_1 . Je-li $\text{dev}_{G_2, G_1}^* > \chi_{f_2 - f_1}^2(\alpha)$, zamítáme shodu modelu M_2 s daty ve prospěch M_1 .

Nejčastěji testujeme model s grafem G proti alternativě saturovaného modelu. Testová statistika je $\text{dev}(G)$, neboť deviance v saturovaném modelu je nulová.

Testujeme-li model M_2 s grafem $G_2 = G_1 \setminus (i, j)$ proti modelu M_1 s grafem G_1 , nazýváme testovou statistiku dev_{G_2, G_1}^* deviancí vynechané hrany (i, j) .

Testujeme-li model M_2 s grafem G_2 proti modelu M_1 s grafem $G_2 \cup (i, j)$, nazýváme dev_{G_2, G_1}^* deviancí přidané hrany (i, j) .

Věta 4.2: Rozklad sdružené hustoty *Uvažujme grafický model pro proměnné*

X_1, \dots, X_k *s řetězovým grafem s bloky b_1, \dots, b_r . Sdruženou hustotu $f_{X_1, \dots, X_k}(x_1, \dots, x_k)$ náhodných veličin X_1, \dots, X_k lze zapsat ve tvaru*

$$f_{X_1, \dots, X_k} = f_{b_1} \prod_{i=2}^r f_{b_i | b_1 \cup \dots \cup b_{i-1}},$$

kde

f_{b_1} *je sdružená hustota veličin z bloku b_1 a*

$f_{b_i | b_1 \cup \dots \cup b_{i-1}}$ *jsou příslušné podmíněné hustoty.*

Důkaz: Viz Whittaker [6], strana 81.

Vzhledem k bodu 1) věty 2.1 v podmínkách hustot $f_{b_i | b_1 \cup \dots \cup b_{i-1}}$ vystupují pouze ty vrcholy, které jsou s vrcholy bloku b_i spojeny hranou.

Věta 4.3: *Nechť f_S je hustota rozdělení $N_k(0, S)$ a $f_{\hat{V}}$ je hustota rozdělení $N_k(0, \hat{V})$. Je-li S výběrová varianční matice spočítaná na základě n realizací náhodného vektoru $(X_1, \dots, X_k)^T$ v saturovaném modelu a \hat{V} maximálně věrohodný odhad varianční matice v grafickém modelu M , platí pro devianci modelu M :*

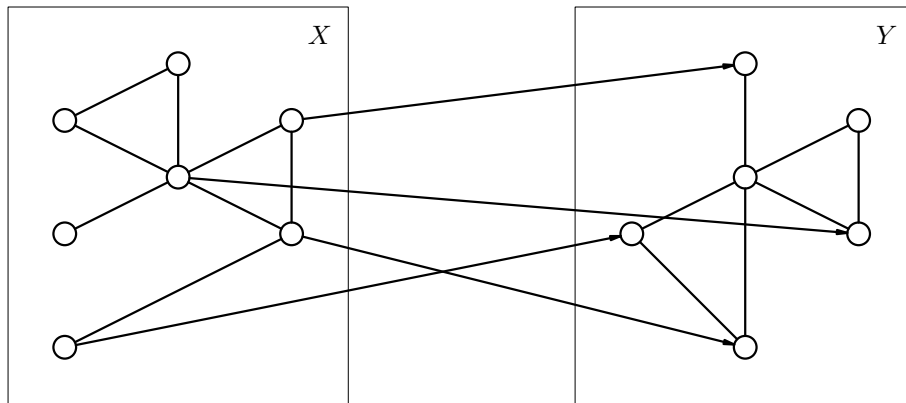
$$\text{dev}(M) = 2n \mathbb{E}_S \log \frac{f_S}{f_{\hat{V}}}.$$

Důkaz: Viz Whittaker [6], str. 185

4.4 Regrese s více závisle proměnnými a řetězové grafy

Předpokládejme dva náhodné vektory \mathbf{X} , \mathbf{Y} , kde $\mathbf{X} = (X_1, \dots, X_p)$ je vektor nezávisle proměnných, zatímco $\mathbf{Y} = (Y_1, \dots, Y_q)$ je vektor závisle proměnných, tedy složky \mathbf{X} ovlivňují složky vektoru \mathbf{Y} , ale ne naopak. Vektor \mathbf{X} reprezentuje množinu příčin a vektor \mathbf{Y} množinu následků. Ignorujeme případné možné podobné chování uvnitř vektoru \mathbf{X} a uvnitř vektoru \mathbf{Y} a předpokládáme, že v případě vzájemného ovlivňování jsou jednotlivé složky daného vektoru rovnocenné. Graficky se tyto situace znázorňují řetězovými grafy podmíněných nezávislostí, v nichž vektory \mathbf{X} , \mathbf{Y} jsou zastoupeny bloky vrcholů b_X , b_Y , přičemž $b_X \prec_b b_Y$ (viz obrázek 4.5).

Popis dat grafickým modelem s řetězovým grafem tedy představuje alternativu k mnohorozměrné regresi zmíněné v odstavci 3.6.



Obrázek 4.5

4.5 Selekcce grafického modelu s řetězovým grafem

Mějme úkol najít model, který co nejlépe popisuje zadaná data. Existují různé selekční algoritmy. Rozhodování mezi jednotlivými grafy činíme na základě deviance nebo difference deviancí. Zmíňme zde tzv. backward algoritmus se stop pravidlem založeným na devianci vynechané hrany. Tento algoritmus na začátku předpokládá úplný graf a v každém kroku zjistí, pro kterou z dosud ponechaných hran je deviance vynechané hrany nejmenší. Pokud je tato deviance menší než kritická hodnota rozdělení χ_1^2 , hrana se z grafu vyjme a postup se opakuje, pokud ne, algoritmus končí.

Selekcí grafického modelu s řetězovým grafem pro dva bloky proměnných můžeme podle knihy Whittaker [6] provádět ve třech krocích.

A_1 : Pomocí backward algoritmu navrhne model s grafem G_1 pro X_1, \dots, X_p .

B_1 : Navrhne model pro $X_1, \dots, X_p, Y_1, \dots, Y_q$ s grafem G_2 obsahujícím úplný podgraf pro vrcholy X_1, \dots, X_p následujícím způsobem:

Vyjdeme z úplného grafu pro vrcholy X_1, \dots, X_p a Y_1, \dots, Y_q . Postupně určíme devianci vynechané hrany pro hrany

$(X_1, Y_1), \dots, (X_1, Y_q),$

$(X_2, Y_1), \dots, (X_2, Y_q),$

\dots

$(X_p, Y_1), \dots, (X_p, Y_q),$

$(Y_1, Y_2), \dots, (Y_1, Y_q),$

$(Y_2, Y_3), \dots, (Y_2, Y_q),$

\dots

$(Y_{q-1}, Y_q).$

Pokud je

$\text{dev}_{i,j} < \chi_1^2(0,05) = 3,84$, pak hranu (i, j) vynecháme z úplného grafu pro proměnné $X_1, \dots, X_p, Y_1, \dots, Y_q$, pokud je

$\text{dev}_{i,j} > \chi_1^2(0,05) = 3,84$, pak hranu (i, j) v grafu ponecháme.

Pro vynechané hrany spočítáme devianci přidané hrany. Pokud pro ni bude platit

$\text{dev}_{i,j} > \chi_1^2(0,05) = 3,84$, pak příslušnou hranu (i, j) do grafu vrátíme, pokud

$\text{dev}_{i,j} < \chi_1^2(0,05) = 3,84$, pak příslušnou hranu (i, j) nevracíme.

Tímto krokem získáme graf, ve kterém budou chybět některé hrany mezi složkami vektoru Y a některé hrany vycházející ze složek vektoru X a končící ve složkách vektoru Y .

C_1 : Výsledný graf G vznikne kombinací modelů s grafy G_1, G_2 z kroků A_1, B_1 .

V grafu G budou neorientované hrany z grafu G_1 , k nim přidáme orientované hrany z vrcholů X_i do vrcholů Y_j a neorientované hrany mezi vrcholy Y_i a Y_j z grafu G_2 .

Na závěr spočítáme celkovou devianci výsledného grafu. Lze ukázat, že $\text{dev}(G) = \text{dev}(G_1) + \text{dev}(G_2)$. Provedeme to na konkrétním případě pro $p = 3$ a $q = 2$, který následuje za algoritmem, na základě vět 4.2, 4.3 a věty 2.1.

V případě tvorby grafického modelu s řetězovým grafem pro více bloků postupujeme následujícím způsobem

A_2 : Kroky A_1, B_1, C_1 vytvoříme model pro bloky b_1, b_2 .

B_2 : Provedeme postup popsáný v B_1 , přičemž za první blok považujeme $b_1 \cup b_2$ a za druhý považujeme b_3 .

C_2 : Pro vrcholy z $b_1 \cup b_2$ ponecháme hrany z A_2 , mezi bloky $b_1 \cup b_2$ a b_3 dáme případné hrany podle kroku B_2 .

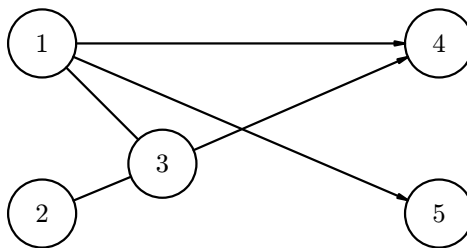
...

A_n : Kroky $A_{n-1}, B_{n-1}, C_{n-1}$ vytvoříme model pro bloky b_1, \dots, b_n .

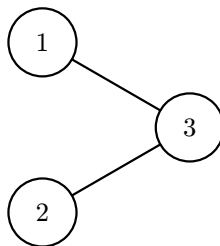
B_n : Provedeme postup popsáný v B_1 , když za první blok považujeme $\bigcup_{i=1, \dots, n} b_i$, za druhý b_{n+1} .

C_n : Sestrojíme graf s hranami podle A_n a B_n

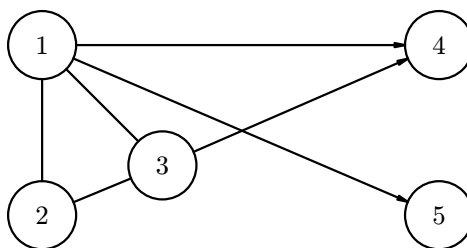
Příklad: Rozklad deviance v řetězovém grafu



Obrázek 4.6 Graf G



Obrázek 4.7 Graf G_1



Obrázek 4.8 Graf G_2

Na obrázcích 4.6, 4.7 a 4.8 vidíme graf G a jeho rozklad na G_1 a G_2 . Podle věty 4.2 máme v modelu s grafem G rozklad sdružené hustoty

$$f_{X_1, X_2, X_3, X_4, X_5} = f_{X_1, X_2, X_3} f_{X_4, X_5 | X_1, X_2, X_3}.$$

To lze podle poznámky za větou 4.2 a podle bodu 2) věty 2.1 psát ve tvaru

$$f_{X_1 | X_3} f_{X_2 | X_3} f_{X_3} f_{X_4, X_5 | X_1, X_3},$$

neboť v modelu s grafem G_1 máme podle 2) ve větě 2.1 rozklad sdružené hustoty

$$f_{X_1, X_2, X_3} = f_{X_1 | X_3} f_{X_2 | X_3} f_{X_3}.$$

V modelu s grafem G_2 máme podle věty 4.2 a následující poznámky rozklad

$$f_{X_1, X_2, X_3, X_4, X_5} = f_{X_1, X_2, X_3} f_{X_4, X_5 | X_1, X_3}.$$

Podle věty 4.3 lze tedy psát

$$\text{dev}(G) = 2n E_S \log \frac{f_{X_1, X_2, X_3, X_4, X_5}}{f_{X_1 | X_3} f_{X_2 | X_3} f_{X_3} f_{X_4, X_5 | X_1, X_2, X_3}},$$

$$\text{dev}(G_1) + \text{dev}(G_2) = 2n E_S \log \left[\frac{f_{X_1, X_2, X_3}}{f_{X_1 | X_3} f_{X_2 | X_3} f_{X_3}} \frac{f_{X_1, X_2, X_3, X_4, X_5}}{f_{X_1, X_2, X_3} f_{X_4, X_5 | X_1, X_2, X_3}} \right] = \text{dev}(G).$$

4.6 Popis programu

Program byl napsán v systému Mathematica 4.0. Po načtení dat ze souboru lze volat procedury pro regresní analýzu, korelační analýzu, testování submodelu vůči modelu, testování submodelu vůči nadřazenému submodelu, hledání regresních parametrů vícerozměrné regrese, selekci grafického modelu s řetězovým grafem a další pomocné procedury. Část programu byla inspirována programem z práce Zelinková [7].

4.6.1 Regresní analýza

Procedura *regresnialyza*[*variancnimaticerean_*, *pocetvyuzitychpozorovanirean_*] provádí regresní analýzu pro vstupní varianční matici standardizovaných dat, ve které je závisle proměnná reprezentována posledním řádkem a posledním sloupcem, nezávisle proměnné řádky a sloupce zbylými. Nejprve je vypočtena hodnota R^2 , poté jsou spočteny regresní parametry a statistiky jejich významnosti. Nakonec je pro každou nezávisle proměnnou proveden výpočet R^2 v modelu redukovaném o tuto proměnnou.

4.6.2 Korelační analýza

Procedura *vyznamnostkorkcoef*[*korelacnimaticekoran_*, *pocetvyuzitychpozorovanikororan_*] provádí korelační analýzu pro vstupní korelační matici, ve které je závisle proměnná reprezentována posledním řádkem a posledním sloupcem, nezávisle proměnné řádky a sloupce

zbylými. Procedura určí výběrové korelační koeficienty, statistiky jejich významnosti, dále pak výběrové koeficienty parciální korelace a opět statistiky jejich významnosti.

4.6.3 Testování submodelu vůči modelu

Procedura *submodely*[*variancnimaticesub_*, *mensisub_*, *vetsisub_*, *zavislepromennasub_*, *pocetvyuzitychpozorovanisub_*] testuje submodel zadaný vstupní varianční maticí s nezávisle proměnnými specifikovanými vstupem *mensisub_* a se závisle proměnnou specifikovanou vstupem *zavislepromennasub_* proti modelu s nezávisle proměnnými specifikovanými sjednocením *mensisub_* a *vetsisub_*. Procedura spočte výslednou testovou statistiku.

4.6.4 Testování submodelu vůči nadřazenému submodelu

Procedura *vicesubmodelu*[*variancnimaticesub_*, *mensisub_*, *vetsisub_*, *nejvetsisub_*, *zavislepromennasub_*, *pocetvyuzitychpozorovanisub_*] se od procedury předchozí liší tím, že submodel zadaný vstupní varianční maticí s nezávisle proměnnými specifikovanými vstupem *mensisub_* a se závisle proměnnou specifikovanou vstupem *zavislepromennasub_* testuje proti modelu s nezávisle proměnnými specifikovanými sjednocením *mensisub_* a *vetsisub_* jako submodelu modelu s nezávisle proměnnými specifikovanými sjednocením *mensisub_*, *vetsisub_* a *nejvetsisub_*. Procedura spočte výslednou testovou statistiku.

4.6.5 Hledání regresních parametrů vícerozměrné regrese

Procedura *viceroz*[*maticevicereg_*, *nezavislevicereg_*, *zavislevicereg_*] najde regresní parametry v regresním modelu určeného vstupní varianční maticí, vstupní množinou nezávisle proměnných a vstupní množinou závisle proměnných.

4.6.6 Selektce grafického modelu s řetězovým grafem

Procedura *backward2*[*puvvarmatice_*, *bloky_*, *pocetrealizaci_*, *nazvypromennych_*] hledá ze vstupní varianční matice nejvhodnější grafický model s řetězovým grafem splňující hierarchickou strukturu zadanou v proměnné *bloky_*.

Po načtení dat ze souboru a rozdělení proměnných do bloků se pomocí backward algoritmu založeného na devianci vynechané hrany provede krok A_1 - selektce hran mezi nezávisle proměnnými z 1. bloku, tj. pro všechny hrany v grafu se spočítá deviance vynechané hrany, vezme se ta nejmenší z deviancí vynechané hrany a pokud je menší nebo rovna kritické hodnotě $\chi_1^2(0,05)$, hrana bude vyjmuta a postup se opakuje. Když bude deviance vynechané hrany všech hran už větší nebo rovna $\chi_1^2(0,05)$, krok A_1 končí a začíná krok B_1 .

Krok B_1 se provede pomocí cyklu, který ve svém i -tém běhu uvažuje model s nezávisle proměnnými z bloků b_1, \dots, b_i a závisle proměnnými z bloku b_{i+1} . Nejprve se doplní data do matice předpřipravených dat pro výpočet, zvětší se matice souslednosti (matice zaznamenávající přítomnost a nepřítomnost hran mezi jednotlivými vrcholy) a vytvoří vektor

všech nových hran na testování (tj. hran netestovaných v předchozích bĕzích vĕpočtu). Pro kaĕdou z tĕchto hran se vypočítá deviance vynechané hrany. Pokud bude menší než $\chi_1^2(0,05)$, hrana bude vyjmuta, tj. na jejím místĕ v matici souslednosti bude '*' a hrana sama bude vložena do množiny vynechaných hran.

Po otestování všech hran na možnost vynechání se spočítá deviance grafu bez vynechaných hran, aby bylo možné postupně pro všechny vynechané hrany spočítat devianci přidané hrany. Pokud tato bude menší než $\chi_1^2(0,05)$, hrana bude opravdu vynechána, tj. vložena do množiny vynechané hrany po opravĕ, v opačném případě bude do grafu vrácena, tj. hvĕzdička v matici souslednosti bude opĕt přepsána jedničkou. Pak se spočítá deviance grafu obsahujícího všechny hrany kromĕ hran vynechaných (resp. vynechaných a nevrácených) v tomto kroku.

Po otestování možného navrácení všech původně vynechaných hran se spočítá deviance výsledného grafu. Deviance tohoto grafu bude rovna součtu posledně spočítané deviance a deviance spočítané buď v kroku A_1 (jestliĕže cyklem procházíme poprvĕ), nebo v předchozím průchodu cyklem. Pokud je tato deviance menší než kritická hodnota příslušného χ^2 rozdělení, nezamítáme shodu modelu navrženého algoritmu s daty. V případě dosud nezpracovaného bloku promĕnných program pokračuje krokem B_{i+1} .

5. Analýza italských finančních dat

V této části práce použijeme dříve uvedené modely k analýze dat převzatých z knihy Giudici [4]. Jsou to údaje o vybraném italském investičním fondu. Proměnné, které značíme X_A, X_B, \dots, X_M , můžeme rozdělit do tří skupin:

X_A, \dots, X_F popisují složení portfolia fondu, to jest jedná se o procentní zastoupení různých typů cenných papírů v portfoliu (italské státní cenné papíry, obligace italských firem, obligace zahraničních firem aj.);

X_G je rozdíl výnosu fondu a bezrizikového výnosu, tuto proměnnou nazveme nadvýnos;

X_H, \dots, X_M jsou indexy finančního trhu, které mohou ovlivňovat výnos fondu

Zajímá nás, pomocí kterých proměnných lze lineárně predikovat nadvýnos. Databázi tvoří časové řady měsíčních pozorování proměnných X_A, \dots, X_M za období od února 1992 do ledna 1998. Jelikož jsou proměnné v různých jednotkách, byla provedena jejich standardizace odečtením průměrů a vydělením směrodatnými odchylkami. Standardizovaná data mají nulovou střední hodnotu a jednotkový rozptyl.

Označme $X_{i,A}$, kde $i = 1, 2, \dots, 72$, prvky standardizované časové řady pro proměnnou X_A a podobně pro ostatní proměnné.

Matice standardizovaných dat je pak

$$\mathbf{X} = \begin{pmatrix} X_{1,A} & X_{1,B} & \dots & X_{1,M} \\ X_{2,A} & X_{2,B} & \dots & X_{2,M} \\ \dots & \dots & \dots & \dots \\ X_{72,A} & X_{72,B} & \dots & X_{72,M} \end{pmatrix}$$

a výběrovou varianční matici lze určit ze vztahu

$$S_{X,X} = \frac{1}{71} \mathbf{X}^T \mathbf{X}$$

Prvky matice $S_{X,X}$ budeme označovat $s_{A,A}, s_{A,B}, \dots, s_{M,M}$.

Výběrová varianční matice má na diagonále jedničky a je rovna výběrové korelační matici $R_{X,X}$. Tato matice je uvedena v knize Giudici [4] na straně 229 a umožní nám provést veškeré odhady a testování v regresních modelech. Nalezneme ji v příloze jako matici 1. Vzhledem k tomu, že nemáme k dispozici časové řady, ale jen matici $S_{X,X}$, předpokládejme, že autoři ověřili potřebné předpoklady normality rozdělení jednotlivých náhodných veličin a nezávislosti členů jednotlivých časových řad. Výpočty byly provedeny v programu popsáném v kapitole 4.6., mezivýsledky budeme uvádět v zaokrouhleném tvaru.

5.1 Model CAPM

V této kapitole se budeme zabývat konstrukcí modelu CAPM pro zvolená data.

Jelikož pracujeme s časovými řadami, budeme potřebovat i časové řady zastupující r_m a r_f . Použijme tedy regresní model se závisle proměnnou $Y = X_G$ a nezávisle proměnnou $X = X_L$, to jest

$$\begin{aligned} Y_t &= r_t - r_{f,t} = X_{t,G}, \\ X_t &= r_{m,t} - r_{f,t} = X_{t,L} \end{aligned}$$

pro $t = 1, 2, \dots, 72$.

Za nezávisle proměnnou jsme v souladu s publikací Giudici [4] zvolili jeden z indexů, konkrétně index Generale della Banca d' Italia.

Regresní rovnice odpovídající modelu čáry kapitálového trhu má tvar

$$Y_t = \beta \cdot X_t + e_t,$$

kde $e_t \sim N(0, \sigma_e^2)$ a e_1, e_2, \dots, e_{72} jsou nezávislé.

Jde o lineární regresi s jednou nezávisle proměnnou. Parametr β odhadneme na základě věty 3.1

$$b = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = 0,888,$$

neboť vzhledem k tvaru výběrové varianční matice $S_{X,X}$ (viz příloha) máme

$$\mathbf{X} = (X_{1,L}, X_{2,L}, \dots, X_{72,L})^T,$$

$$\mathbf{Y} = (X_{1,G}, X_{2,G}, \dots, X_{72,G})^T.$$

Predikce nadvýnosu je tedy

$$\widehat{X}_G = 0,888 X_L.$$

Koeficient determinace R^2 spočítáme pomocí věty 2.6 z prvku $r_{G,G}$ inverzní výběrové korelační matice proměnných X_G a X_L .

$$R^{-1} = \begin{pmatrix} 1 & 0,888 \\ 0,888 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 4,729 & -4,199 \\ -4,199 & 4,729 \end{pmatrix},$$
$$r_{G,G}^0 = \frac{1}{1 - R^2},$$

tedy

$$R^2 = \frac{r_{G,G}^0 - 1}{r_{G,G}^0} = \frac{4,729 - 1}{4,729} = 0,789.$$

Poznámka: Vzhledem k tomu, že jsme pracovali jen se dvěma proměnnými, je koeficient mnohonásobné korelace roven výběrovému korelačnímu koeficientu. Vzhledem ke kapitole 1.5 a k tomu, že pracujeme se standardizovanými daty, lze rovněž psát

$$R^2 = S_{Y,X} \cdot (S_{X,X})^{-1} \cdot S_{X,Y} = s_{L,G}^2.$$

$S_{X,Y} = S_{Y,X} = s_{L,G}$ zde označuje odhad kovariance a korelačního koeficientu standardizovaných proměnných X_G, X_L .

Dále otestujeme nulovost parametru β z rovnice

$$Er = r_f + (Er_m - r_f) \cdot \beta,$$

tj. $EY = \beta X$ pro $Y = X_G, X = X_L$.

Vzhledem k tomu, že předpokládáme

$$Er_m \geq r_f, \quad Er \geq r_f,$$

nebude β nabývat záporných hodnot. Budeme tedy testovat hypotézu

$$H_0 : \beta = 0$$

proti jednostranné alternativě

$$H_1 : \beta > 0$$

na pětiprocentní hladině na základě věty 3.3. Máme

$$\begin{aligned} b &= 0,888, \\ S_e &= Y^T Y - b^T \underline{X}^T \cdot Y = 71 (S_{Y,Y} - b^T \cdot S_{X,Y}) = 71 (S_{Y,Y} - S_{Y,X} \cdot S_{X,X}^{-1} \cdot S_{X,Y}) = \\ &= 71 (s_{G,G} - s_{L,G}^2) = 71 (1 - R^2) = 15,013, \\ s^2 &= \frac{S_e}{72 - 1} = \frac{1 - R^2}{71} = \frac{0,211}{71} = 0,00298, \\ T &= \frac{0,888 - 0}{\sqrt{s^2 \cdot u_{L,L}}} = \frac{0,888}{\sqrt{0,211 \cdot \frac{1}{71}}} = 16,272. \end{aligned}$$

Vzhledem k tomu, že

$$t_{71}(0,10) = 1,6666,$$

hypotézu H_0 zamítáme ve prospěch alternativy H_1 .

5.2 Model APT

5.2.1 Regresní analýza

Chceme-li popsat nadvýnos pomocí více nezávisle proměnných, které máme k dispozici, lze navrhnout model APT.

Rovnice modelu APT má v našem případě tvar

$$r = Er + \sum_{j=1}^6 \delta_j \beta_j + e,$$

kde proměnné mají tyto významy:

- r výnos portfolia fondu,
- Er očekávaný výnos, zvolme za něj bezrizikový výnos r_f ,
- δ_j faktory ovlivňující výnos, tj. nezávisle proměnné,
- β_j regresní koeficienty,
- e náhodná chyba.

V tomto případě zvolíme regresní model se závisle proměnnou $Y = X_G$ a nezávisle proměnnými $X_1 = X_H, X_2 = X_I, \dots, X_6 = X_M$, tedy indexy finančního trhu. Regresní rovnice je

$$Y_t = X_{t,G} = r_t - r_{t,f} = \sum_{j=1}^6 \beta_j X_{t,j} + e_t,$$

kde $t = 1, 2, \dots, 72$ a opět předpokládáme $e_t \sim N(0, \sigma_e^2)$ nezávislé.

Jde o případ lineární regrese s šesti nezávisle proměnnými. V tomto modelu opět odhadneme parametry $\beta_1, \beta_2, \dots, \beta_6$ na základě věty 3.1. Jejich případnou nulovost lze testovat pomocí t -testu uvedeného ve větě 3.3.

Odhad vektoru koeficientů β získáme podle věty 3.1 z rovnice

$$b = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = S_{X,X}^{-1} S_{X,Y}.$$

Řádky matice \mathbf{X} jsou pozorovanými realizacemi vektoru (X_H, X_I, \dots, X_M) .

Podklady pro výpočet odhadu získáme z matice 1 v příloze.

$S_{X,X}$ je typu 6×6 a je odhadem pro varianční matici vektoru (X_H, \dots, X_M) :

$$S_{X,X} = \begin{pmatrix} 1,000 & 0,689 & -0,127 & 0,811 & 0,984 & 0,044 \\ 0,689 & 1,000 & 0,048 & 0,749 & 0,800 & -0,020 \\ -0,127 & 0,048 & 1,000 & -0,015 & -0,088 & 0,684 \\ 0,811 & 0,749 & -0,015 & 1,000 & 0,833 & 0,077 \\ 0,984 & 0,800 & -0,088 & 0,833 & 1,000 & 0,036 \\ 0,044 & -0,020 & 0,684 & 0,077 & 0,036 & 1,000 \end{pmatrix},$$

$S_{X,Y}$ je vektorem typu 6×1 a je odhadem pro vektor kovariancí složek náhodného vektoru $(X_H, \dots, X_M)^T$ a náhodné veličiny Y .

$$S_{X,Y} = \frac{1}{71} \mathbf{X}^T \mathbf{Y} = (0,851; 0,835; -0,049; 0,808; 0,888; -0,013)^T$$

Nyní dosadíme tyto odhady do rovnice pro výpočet odhadu b . Dostáváme

$$b = (2,169; 0,909; 0,041; 0,039; -2,001; -0,050)^T.$$

Náhodnou veličinu X_G budeme tedy predikovat ze vztahu

$$\hat{X}_G = 2,169 \cdot X_H + 0,909 \cdot X_I + 0,041 \cdot X_J + 0,039 \cdot X_K - 2,001 \cdot X_L - 0,050 \cdot X_M.$$

Opět podle věty 2.6 platí

$$r_{G,G}^0 = \frac{1}{1 - R^2},$$

takže po vypočtení $r_{G,G}^0 = 7,063$ invertováním výběrové korelační matice proměnných X_G, X_H, \dots, X_M můžeme vypočítat koeficient determinace $R^2 = 0,858$.

Nyní budeme testovat hypotézy

$$H_{0,N} : \beta_i = 0$$

proti alternativě

$$H_{1,N} : \beta_i \neq 0$$

(pro $N = H, I, J, K, L, M$) na pětiprocentní hladině na základě věty 3.3.

Hodnoty testových statistik jsou

$$T_H = 3,032, \quad T_I = 4,156, \quad T_J = 0,598,$$

$$T_K = 0,417, \quad T_L = -2,379, \quad T_M = -0,732.$$

Vzhledem k tomu, že

$$t_{66}(0,05) \doteq 1,997,$$

hypotézu nulovosti regresního koeficientu nezamítáme pro X_J, X_K, X_M a zamítáme ji pro X_H, X_I, X_L . Model APT bychom tedy mohli zjednodušit na regresní model se třemi nezávisle proměnnými X_H, X_I, X_L .

5.2.2 Korelační analýza

Nyní budeme vyšetřovat korelační strukturu mezi X_G a postupně $X_H, X_I, X_J, X_K, X_L, X_M$.

Nejprve vyšetříme nulovost $\rho_{G,N}$ (opět pro $N = H, I, J, K, L, M$). Na pětiprocentní hladině budeme testovat hypotézu

$${}_{G,N}H_0 : \rho_{G,N} = 0$$

proti alternativě

$${}_{G,N}H_1 : \rho_{G,N} \neq 0.$$

V matici 1 v příloze si vyhledejme hodnoty příslušných výběrových korelačních koeficientů:

$$\begin{aligned} r_{G,H} &= 0,851, & r_{G,I} &= 0,835, & r_{G,J} &= -0,049, \\ r_{G,K} &= 0,808, & r_{G,L} &= 0,888, & r_{G,M} &= -0,013. \end{aligned}$$

Za předpokladu platnosti ${}_{G,H}H_0$ je podle věty 1.1 testová statistika

$$T_{G,H} = \frac{r_{G,H}}{\sqrt{1 - r_{G,H}^2}} \cdot \sqrt{70} = 13,558$$

náhodnou veličinou s rozdělením t_{70} . Vzhledem k tomu, že

$$t_{70}(0,05) \doteq 1,994,$$

zamítáme hypotézu ${}_{G,H}H_0$ ve prospěch alternativy ${}_{G,H}H_1$, čili zamítneme hypotézu nekorelovanosti X_G a X_H . Podobně spočteme následující testové statistiky:

$$\begin{aligned} T_{G,I} &= 12,696, & T_{G,J} &= -0,410, & T_{G,K} &= 11,474, \\ T_{G,L} &= 16,157, & T_{G,M} &= -0,109. \end{aligned}$$

Celkově nezamítáme hypotézy nekorelovanosti X_G postupně na X_J, X_M a zamítáme hypotézy nekorelovanosti X_G postupně na X_H, X_I, X_K, X_L .

Nyní vyšetřujeme parciální korelace mezi veličinou X_G a veličinami $X_H, X_I, X_J, X_K, X_L, X_M$. Nejprve spočteme výběrové koeficienty parciální korelace mezi X_G a postupně $X_H, X_I, X_J, X_K, X_L, X_M$. Ty můžeme získat buď podle věty 2.7 pomocí (-1) -násobku matice získané škálováním inverzní varianční matice, nebo postupem naznačeným v kapitole 1.6:

Označme

$$N(G, H) = (I, J, K, L, M),$$

a máme

$$r_{G,H|N(G,H)} = 0,350.$$

Dále označme

$$N(G, I) = (H, J, K, L, M)$$

atd. a postupně dostaneme

$$r_{G,I|N(G,I)} = 0,455, \quad r_{G,J|N(G,J)} = 0,073, \quad r_{G,K|N(G,K)} = 0,051,$$

$$r_{G,L|N(G,L)} = -0,281 \quad r_{G,M|N(G,M)} = -0,090$$

Z těchto koeficientů budeme pomocí věty 1.3 vyšetřovat možnou nekorelovanost proměnné X_G postupně na $X_H, X_I, X_J, X_K, X_L, X_M$ s vyloučením vlivu zbývajících proměnných. Nejprve testujme na základě této věty na pětiprocentní hladině hypotézu

$${}_{G,H}H_0 : \rho_{G,H|N(G,H)} = 0$$

proti alternativě

$${}_{G,H}H_1 : \rho_{G,H|N(G,H)} \neq 0.$$

Hodnota příslušné testové statistiky je

$$T_{G,H} = \frac{\sqrt{r_{G,H|N(G,H)}}}{\sqrt{1 - r_{G,H|N(G,H)}^2}} \cdot \sqrt{65} = 3,009.$$

Obdobným postupem dostaneme

$$T_{G,I} = 4,125, \quad T_{G,J} = 0,593, \quad T_{G,K} = 0,413,$$

$$T_{G,L} = -2,361, \quad T_{G,M} = -0,726.$$

Vzhledem k tomu, že

$$t_{65}(0,05) \doteq 1,997,$$

hypotézu nulovosti koeficientu nezamítáme pro X_J, X_K, X_M a zamítáme ji pro X_H, X_I, X_L , tedy přímo je veličina X_G významně ovlivněna veličinami X_H, X_I, X_L a méně významně veličinami X_J, X_K, X_M .

Dvojice	Korelace	Parciální korelace
<i>GH</i>	0,851	0,350
<i>GI</i>	0,835	0,455
<i>GJ</i>	-0,049	0,073
<i>GK</i>	0,808	0,051
<i>GL</i>	0,888	-0,281
<i>GM</i>	-0,013	-0,090

Tabulka 5.1: Korelace a parciální korelace X_G s nezávisle proměnnými, statisticky významné korelace a parciální korelace jsou zvýrazněny tučným písmem

V tabulce 5.1 vidíme, že korelace nadvýnosu X_G s proměnnými reprezentujícími finanční indexy se liší od parciálních korelací. Tato odlišnost je způsobena vzájemnou provázaností finančních indexů. Všimněme si například, že korelace nadvýnosu X_G s indexem Generale della Banca d'Italia X_L , jenž byl použit jako vysvětlující proměnná v CAPM modelu, je kladná, zatímco jejich parciální korelace je záporná. S tím souvisí i záporný odhad regresního parametru $\beta_5 : b_5 = -2,001$ u X_L v APT modelu, zatímco odhad regresního koeficientu u X_L v CAPM modelu $b = 0,888$ byl kladný.

Korelace mezi nadvýnosem X_G a indexem X_K je statisticky významná, zatímco jejich parciální korelace nikoliv. Znamená to, že X_K ovlivňuje X_G prostřednictvím ostatních indexů. Podívejme se do matice 1 v příloze na příslušné výběrové korelační koeficienty a spočtěme výběrové parciální korelační koeficienty proměnné X_K s ostatními nezávislými proměnnými:

Dvojice	Korelace	Parciální korelace
KH	0,811	0,377
KI	0,749	0,438
KJ	-0,015	0,011
KL	0,833	-0,314
KM	0,077	0,093

Tabulka 5.2: Korelace a parciální korelace X_K s ostatními nezávisle proměnnými, statisticky významné korelace a parciální korelace jsou zvýrazněny tučným písmem

Z tabulky 5.2 plyne, že veličina X_K je významně ovlivněna veličinami X_H, X_I, X_L , čili veličinami významně ovlivňujícími X_G . Takže veličina X_K ovlivňuje X_G přes X_H, X_I, X_L , a proto ji nezahrneme do výsledného modelu o závisle proměnné X_G a nezávisle proměnných X_H, X_I, X_L . Korelační analýza tedy nalézá stejné zjednodušení APT modelu jako regresní analýza. Vektor parametrů zjednodušeného modelu má odhad

$$b = (2,259, 0,962, -2,104)^T.$$

Otestujme ještě nulovost regresních koeficientů. Testujme hypotézy

$${}_N H_0 : \beta_N = 0$$

proti alternativě

$${}_N H_1 : \beta_N \neq 0$$

(pro $N = H, I, L$) na pětiprocentní hladině na základě věty 3.3. Hodnoty testových statistik jsou

$$T_H = 3,493, \quad T_I = 5,011, \quad T_L = -2,694.$$

Vzhledem k tomu, že

$$t_{69}(0,05) \doteq 1,995,$$

hypotézu nulovosti regresního koeficientu zamítáme pro všechny nezávisle proměnné tohoto zjednodušeného APT modelu.

Vzhledem k tomu, že prvek inverzní korelační matice je

$$r_{G,G}^0 = 6,991,$$

je koeficient determinace

$$R^2 = \frac{r_{G,G}^0 - 1}{r_{G,G}^0} = \frac{6,991 - 1}{6,991} = 0,857$$

V tomto případě je R^2 jen nepatrně menší než v prvním případě u modelu APT s šesti nezávisle proměnnými. Tento model téměř stejně přesně popisuje variabilitu nadvýnosu X_G jako model s šesti nezávisle proměnnými. Nadvýnos X_G lze tedy pomocí indexů finančního trhu predikovat ze vztahu

$$\hat{X}_G = 2,259 \cdot X_H + 0,962 \cdot X_I - 2,104 \cdot X_L.$$

5.3 Grafický model

5.3.1 Korelační analýza

Analýza parciálních korelačních koeficientů nám umožní lépe studovat strukturu vzájemných závislostí v množině navzájem korelovaných proměnných, jak jsme naznačili v modelu APT. Studujeme totiž „čisté závislosti“ ve dvojicích veličin s vyloučením vlivu veličin ostatních. Nadále se proto zaměříme na tento přístup. Rozšíříme přitom množinu nezávisle proměnných tak, že vedle vlivu finančních indexů X_H, X_I, \dots, X_M na nadvýnos připustíme i vliv proměnných X_A, X_B, \dots, X_F popisujících portfolio fondu. Informaci o struktuře parciálních korelací v množině proměnných podává graf podmíněných nezávislostí (viz definice 4.1).

Jestliže budeme uvažovat všech 13 proměnných X_A, X_B, \dots, X_M , můžeme sestavit grafy o maximálně $\binom{13}{2} = 78$ hranách, což dává 2^{78} možných grafů. Každý z nich popisuje grafický model určený za předpokladu normality podmíněnými nezávislostmi dvojic proměnných.

V grafickém modelu je každá proměnná zastoupena právě jedním vrcholem. Máme tedy vrcholy X_A, X_B, \dots, X_M . Označme $\mathbf{X} = (X_A, X_B, \dots, X_M)^T$. Hrana mezi vrcholy X_i a X_j chybí, právě když

$$X_i \perp X_j \mid N(\mathbf{X}; i, j)$$

což podle věty 2.14 platí, právě když

$$d_{i,j} = 0,$$

kde

$$D = V^{-1} \text{ a } V = \text{var}(X_A, X_B, \dots, X_M)^T.$$

To podle věty 2.7 nastává, právě když pro koeficient parciální korelace platí

$$\rho_{X_i, X_j \mid N(\mathbf{X}; i, j)} = 0.$$

Za předpokladu normality dat to znamená, že $X_i - \widehat{X}_i, X_j - \widehat{X}_j$ jsou nezávislé, kde \widehat{X}_i je nejlepší lineární aproximací X_i pomocí $N(\mathbf{X}; i, j)$, \widehat{X}_j je nejlepší lineární aproximací X_j pomocí $N(\mathbf{X}; i, j)$, obojí ve smyslu kapitoly 1.1 a při použití značení zavedeného v kapitole 2.4.

Tedy X_i a X_j jsou nezávislé při daných hodnotách $N(\mathbf{X}; i, j)$.

Je-li naopak $\rho_{X_i, X_j | N(\mathbf{X}; i, j)} \neq 0$, tedy v grafu existuje hrana spojující vrcholy X_i, X_j , znamená to korelaci proměnných $X_i - \widehat{X}_i, X_j - \widehat{X}_j$, to jest jejich lineární závislost. Odtud plyne, že X_i lze modelovat regresní rovnicí v závislosti na X_j . Jelikož vyšetřujeme závislost proměnné $Y = X_G$ na ostatních proměnných $X_A, \dots, X_F, X_H, \dots, X_M$, zajímají nás pouze hrany spojující vrchol X_G s ostatními vrcholy. Přítomnost takových hran v grafu můžeme otestovat na základě věty 1.3, tedy testem nulovosti příslušných koeficientů parciální korelace. Výběrové koeficienty parciálních korelací jsou podle věty 2.7 (-1) -násobky mimodiagonálních prvků škálované inverzní výběrové varianční matice. Statisticky významné parciální korelace jsou vyznačeny tučně:

$$\begin{aligned} r_{G,A|N(\mathbf{X};G,A)} &= 0,152, \\ r_{G,B|N(\mathbf{X};G,B)} &= -\mathbf{0,448}, \\ r_{G,C|N(\mathbf{X};G,C)} &= 0,121, \\ r_{G,D|N(\mathbf{X};G,D)} &= 0,228, \\ r_{G,E|N(\mathbf{X};G,E)} &= -\mathbf{0,293}, \\ r_{G,F|N(\mathbf{X};G,F)} &= -\mathbf{0,324}, \\ r_{G,H|N(\mathbf{X};G,H)} &= \mathbf{0,402}, \\ r_{G,I|N(\mathbf{X};G,I)} &= \mathbf{0,535}, \\ r_{G,J|N(\mathbf{X};G,J)} &= -0,068, \\ r_{G,K|N(\mathbf{X};G,K)} &= -0,119, \\ r_{G,L|N(\mathbf{X};G,L)} &= -\mathbf{0,325}, \\ r_{G,M|N(\mathbf{X};G,M)} &= -0,077. \end{aligned}$$

Statisticky významné výběrové parciální korelace indikují přítomnost jednotlivých hran vycházejících z vrcholu X_G v grafu podmíněných nezávislostí. Na základě tohoto testování navrhneme pro nadvýnos X_G regresní model s nezávisle proměnnými $X_B, X_E, X_F, X_H, X_I, X_L$. Regresory X_B, X_E, X_F se vztahují ke složení portfolia fondu, regresory X_H, X_I, X_L odpovídají finančním indexům, a to právě těm, které byly zařazeny do regresního APT modelu studujícího závislost nadvýnosu X_G pouze na indexech.

5.3.2 Regresní analýza

K vyjádření závislosti použijeme regresní model z kapitoly 3.1.

$$Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde

$Y = (X_{1,G}, X_{2,G}, \dots, X_{72,G})^T$ je vektor pozorování závislé proměnné X_G ,
 \mathbf{X} je matice 72×6 , pozorování nezávislých proměnných v daných časech,
 β je 6-rozměrný vektor parametrů,
 e je 72-rozměrný vektor chyb z rozdělení $N(0, \sigma_e^2 \cdot I)$.

Odhad vektoru koeficientů β získáme podle věty 3.1 z rovnice a z matice 1 v příloze

$$b = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = S_{X,X}^{-1} S_{X,Y} = (-0,157; -0,097; -0,084; 2,053; 0,875; -1,855)^T.$$

Náhodnou veličinu X_G budeme tedy predikovat výrazem

$$\widehat{X}_G = -0,157 \cdot X_B - 0,097 \cdot X_E - 0,084 \cdot X_F + 2,053 \cdot X_H + 0,875 \cdot X_I - 1,855 \cdot X_L.$$

Po vypočtení $r_{G,G}^0 = 8,909$ invertováním příslušné korelační matice můžeme podle věty 2.6 vypočítat R^2 :

$$R^2 = \frac{r_{G,G}^0 - 1}{r_{G,G}^0} = \frac{8,909 - 1}{8,909} = 0,888$$

Nyní budeme (pro $N = B, E, F, H, I, L$) testovat hypotézy

$${}_N H_0 : \beta_N = 0$$

s alternativou

$${}_N H_1 : \beta_N \neq 0$$

na pětiprocentní hladině na základě věty 3.3. Hodnoty testových statistik jsou

$$T_B = -3,509, \quad T_E = -1,838, \quad T_F = -1,700, \\ T_H = 3,474, \quad T_I = 4,958, \quad T_L = -2,602.$$

Při kritické hodnotě

$$t_{66}(0,05) \doteq 1,997$$

tedy nezamítáme hypotézu nulovosti regresního parametru u X_E a u X_F a zamítáme ji u X_B, X_H, X_I, X_L .

Vytvoříme tedy výsledný regresní model o závisle proměnné X_G a nezávisle proměnných X_B, X_H, X_I, X_L .

Odhad vektoru koeficientů β vyjde roven

$$b = (-0,160; 2,153; 0,869; -1,906)^T.$$

Všechny jeho složky jsou již statisticky významné.

Nyní spočtíme R^2 :

$$r_{G,G}^0 = 8,373 \\ R^2 = \frac{r_{G,G}^0 - 1}{r_{G,G}^0} = \frac{8,373 - 1}{8,373} = 0,881.$$

Koeficient determinace R^2 se nám oproti modelu s šesti nezávisle proměnnými zmenšil jen nepatrně, vystačíme tedy s modelem se čtyřmi nezávisle proměnnými X_B, X_H, X_I, X_L a nadvýnos X_G lze predikovat ze vztahu

$$\widehat{X}_G = -0,160 \cdot X_B + 2,153 \cdot X_H + 0,869 \cdot X_I - 1,906 \cdot X_L.$$

Kromě trojice finančních indexů X_H, X_I, X_L se tedy při predikci uplatní pouze proměnná X_B představující zastoupení italských státních cenných papírů v portfoliu fondu.

Porovnejme na závěr použité proměnné a koeficienty determinace v navržených regresních modelech:

Model	Proměnné	R^2
CAPM	X_L	0,789
APT	X_H, X_I, X_L	0,857
grafický	X_B, X_H, X_I, X_L	0,881

Tabulka 5.3: Srovnání jednotlivých regresních modelů

Ve všech třech případech jsme vytvořili regresní model, který podle koeficientu determinace dobře popisuje variabilitu nadvýnosu X_G .

V modelu CAPM jsme pracovali s předem určenou nezávisle proměnnou X_L – indexem Generale della Banca d' Italia, který byl navržen pro CAPM model v knize Giudici [4].

V modelu APT jsme za nezávisle proměnné ovlivňující nadvýnos zvolili množinu finančních indexů X_H, X_I, \dots, X_M . Tuto množinu jsme zkusili zredukovat dvěma způsoby – testováním nulovosti regresních koeficientů a vyšetřením korelační struktury proměnných X_G, X_H, \dots, X_M . Oba postupy vedly k výslednému modelu se stejnými nezávisle proměnnými X_H, X_I, X_L .

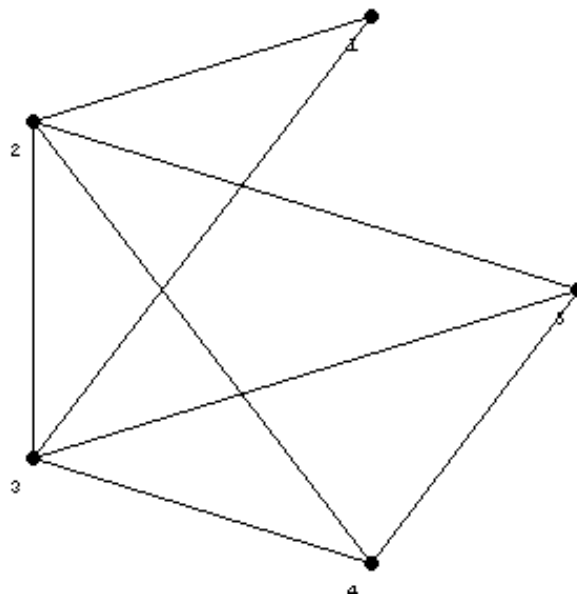
V grafickém modelu jsme uvažovali všechny možné nezávisle proměnné, tj. složky portfolia vyšetřovaného investičního fondu X_A, X_B, \dots, X_F a finanční indexy X_H, X_I, \dots, X_M . Výběr proměnných do regresního modelu byl učiněn na základě analýzy parciálních korelací, což odpovídá ideji grafických modelů. Testovali jsme tedy přítomnost hran vycházejících z vrcholu X_G reprezentujícího nadvýnos v grafu o 13 vrcholech X_A, X_B, \dots, X_M . Takto získaný model byl zjednodušen na základě testování nulovosti regresních koeficientů.

5.3.3 Model získaný selekčním algoritmem

Grafické modely však nebyly vytvořeny jen pro tvorbu modelů regresních, ale především pro zkoumání podmíněných nezávislostí v množině proměnných a tím pro zjišťování souvislostí nejen mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými, ale i mezi vysvětlujícími proměnnými navzájem. Uvažujme tedy proměnné X_B, X_G, X_H, X_I, X_L , které jsme v předešlém textu zařadili do regresního modelu. Chceme vybrat ze všech možných grafických modelů s grafy o pěti vrcholech takový, který nejlépe popisuje strukturu podmíněných nezávislostí v datech. Je tedy třeba testovat shodu grafických modelů s daty a provést výběr vhodného modelu. Pro tento účel byly vyvinuty selekční algoritmy (viz např. kniha Whittaker [6]). K řešení našeho problému použijeme program popsáný v kapitole 4.6, resp. jeho část užívající tzv. backward algoritmus (konkrétně *backward algoritmus se stop pravidlem založeným na devianci vynechané hrany*). Výstupem programu je graf podmíněných nezávislostí na obrázku 5.1.

Počet vynechaných hran = 2

Celková deviance = 1.11737 < $\chi_2^2 = 5,99$



Obrázek 5.1

Bod v grafu	Příslušející proměnná
1	X_B
2	X_G
3	X_H
4	X_I
5	X_L

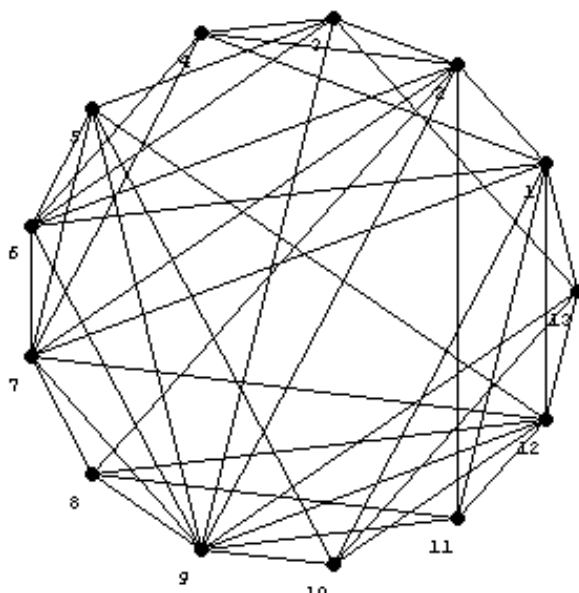
Tabulka 5.4: Označení bodů v grafu na obrázku 5.1

Z obrázku je vidět, že proměnné X_B, X_H, X_I, X_L ovlivňují závažně nejen proměnnou X_G , ale až na dvě výjimky (X_B s X_I a X_B s X_L) i sebe navzájem. Chybějící hrany (1, 4), (1, 5) lze vysvětlit tím, že proměnná X_B patří do skupiny veličin popisujících složení portfolia fondu, zatímco proměnné X_I, X_L jsou finanční indexy. Podgraf tvořený vrcholy 3, 4, 5 je klika znamenající vzájemné ovlivňování finančních indexů X_H, X_I, X_L .

Dalším výstupem je celková deviance, což je testová statistika pro měření shody modelu s daty, jak bylo uvedeno v kapitole 4. Zde shodu výsledného modelu s daty nezamítáme.

Na obrázku 5.2 je graf podmíněných nezávislostí mezi všemi vyšetřovanými proměnnými X_A, X_B, \dots, X_M , reprezentovanými čísly 1, 2, \dots , 13.

Počet vynechaných hran = 35
 Celková deviance = 29.3242 < $\chi_{35}^2 = 49,80$



Obrázek 5.2

S vrcholem 7 (reprezentujícím veličinu X_G) jsou spojeny hranou vrcholy 1 (X_A), 2 (X_B), 4 (X_D), 5 (X_E), 6 (X_F), 8 (X_H), 9 (X_I) a 12 (X_L). Porovnáme-li množinu vrcholů spojených s vrcholem reprezentujícím proměnnou X_G s množinou určenou na základě analýzy parciálních korelací v kapitole 5.3.1, zjistíme, že backward algoritmus vybral navíc vrcholy představující proměnné X_A a X_D , to jest podíly dvou skupin obligací v portfoliu fondu. Tyto proměnné jsme v analýze parciálních korelací do modelu nezařadili při hodnotách výběrových parciálních korelací 0,152 a 0,228. Shodu výsledného modelu s daty opět nezamítáme.

5.4 Testování submodelů

Shodu modelů navržených selekčním algoritmem s daty jsme v kapitole 5.3.3 testovali na základě teorie grafických modelů pomocí deviance. Jiný přístup představuje testování submodelů lineárního regresního modelu zmíněné v kapitole 3.2.

V této kapitole budeme testovat námi navržené modely jako submodely nejširšího možného modelu i jako submodely mezi sebou navzájem. Nejprve testujme námi navržený grafický model získaný selekčním algoritmem proti modelu se všemi dvanácti nezávisle proměnnými.

Vektor $\mathbf{X} = (X_A, X_B, \dots, X_M)^T$ rozložíme na vektory

$$\mathbf{X}_a = (X_A, X_B, X_D, X_E, X_F, X_H, X_I, X_L)^T, \quad \mathbf{X}_b = (X_C, X_J, X_K, X_M)^T$$

a uvažujme hypotézu H_0 tvrdící, že Y závisí pouze na vektoru X_a

$$H_0 : \beta_C = \beta_J = \beta_K = \beta_M = 0,$$

čili hypotézu, že model

$$\begin{aligned} M : \mathbf{E}Y &= \mathbf{E}X_G = \\ &= \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_E X_E + \beta_F X_F + \\ &\quad + \beta_H X_H + \beta_I X_I + \beta_J X_J + \beta_K X_K + \beta_L X_L + \beta_M X_M \end{aligned}$$

s maticí modelu o hodnotě $h(\mathbf{X}) = 12$ lze zjednodušit na submodel

$$M_1 : \mathbf{E}Y = \mathbf{E}X_G = \beta_A X_A + \beta_B X_B + \beta_D X_D + \beta_E X_E + \beta_F X_F + \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí submodelu o hodnotě $h(\mathbf{X}_a) = 8$.

Vypočteme nejprve součty čtverců na základě podkladů z matice 1 v příloze.

$$SS_b = (n-1)S_{Y,X} S_{X,X}^{-1} S_{X,Y} - (n-1)S_{Y,X_a} S_{X_a,X_a}^{-1} S_{X_a,Y} = 64,220 - 63,786 = 0,434,$$

$$SS_a = (n-1)S_{Y,X_a} S_{X_a,X_a}^{-1} S_{X_a,Y} = 63,786,$$

$$SS_e = (n-1)S_{Y,Y} - (n-1)S_{Y,X} S_{X,X}^{-1} S_{X,Y} = 71 - 64,220 = 6,780,$$

$$SS_T = (n-1)S_{Y,Y} = 71.$$

Testová statistika je rovna

$$F = \frac{\frac{SS_b}{f_b}}{\frac{SS_e}{f_e}} = \frac{\frac{0,434}{12-8}}{\frac{6,780}{72-12}} = 0,959.$$

A nyní sestrojme tabulku analýzy rozptylu pro test nulovosti skupiny regresních parametrů příslušejících podvektoru X_b :

Zdroj variability	Součet čtverců	Stupňů volnosti	F-statistika
X_b při pevném X_a	0,434	4	$F = 0,959$
X_a	63,786	8	
reziduální	6,780	60	
celkový	71	72	

Tabulka 5.5: Tabulka analýzy rozptylu pro test M_1 proti M

Jelikož

$$F = 0,959 < F_{4,60} = 2,525,$$

nezamítáme submodel M_1 ve prospěch modelu M .

Dále vektor

$$\mathbf{X}_a = (X_A, X_B, X_D, X_E, X_F, X_H, X_I, X_L)^T$$

rozložíme na vektory

$$\mathbf{X}_{aa} = (X_B, X_H, X_I, X_L)^T \quad \mathbf{X}_{bb} = (X_A, X_D, X_E, X_F)^T.$$

Uvažujme hypotézu $\frac{1}{2}H_0$ tvrdící, že nadvýnos Y závisí pouze na vektoru X_{aa} , to jest na nezávisle proměnných vybraných v kapitole 5.3.2.

$$\frac{1}{2}H_0 : \beta_A = \beta_C = \beta_D = \beta_E = \beta_F = \beta_J = \beta_K = \beta_M = 0$$

čili hypotézu, že model

$$M : \mathbf{E}Y = \mathbf{E}X_G = \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_E X_E + \beta_F X_F + \\ + \beta_H X_H + \beta_I X_I + \beta_J X_J + \beta_K X_K + \beta_L X_L + \beta_M X_M$$

s maticí modelu o hodnotě $h(\underline{\mathbf{X}}) = 12$ lze zjednodušit na submodel

$$M_2 : \mathbf{E}Y = \mathbf{E}X_G = \beta_B X_B + \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí submodelu o hodnotě $h(\underline{\mathbf{X}}_{aa}) = 4$, přičemž nebereme v úvahu model M_1 .

Vypočteme nejprve součty čtverců

$$SS_{bbb} = (n-1)S_{Y,X} S_{X,X}^{-1} S_{X,Y} - (n-1)S_{Y,X_{aa}} S_{X_{aa},X_{aa}}^{-1} S_{X_{aa},Y} = 64,220 - 62,520 = 1,700,$$

$$SS_{aa} = (n-1)S_{Y,X_{aa}} S_{X_{aa},X_{aa}}^{-1} S_{X_{aa},Y} = 62,520,$$

$$SS_e = (n-1)S_{Y,Y} - (n-1)S_{Y,X} S_{X,X}^{-1} S_{X,Y} = 71 - 64,220 = 6,780.$$

Testová statistika je rovna

$$F = \frac{\frac{SS_{bbb}}{f_{bbb}}}{\frac{SS_e}{f_e}} = \frac{\frac{1,700}{12-4}}{\frac{6,780}{72-12}} = \frac{0,212}{0,113} = 1,880.$$

Nyní uveďme tabulku analýzy rozptylu pro test nulovosti skupiny regresních parametrů příslušejících podvektoru X_b :

Zdroj variability	Součet čtverců	Stupňů volnosti	F-statistika
$X_b \cup X_{bb}$ při pevném X_{aa}	1,700	8	$F = 1,880$
X_{aa}	62,520	4	
reziduální	6,780	60	
celkový	71	72	

Tabulka 5.6: Tabulka analýzy rozptylu pro test M_2 proti M

Testová statistika nám vyšla

$$F = 1,880 < F_{8,60} = 2,097,$$

tedy nezamítáme grafický model získaný korelační a regresní analýzou ve prospěch původního modelu M .

Nyní uvažujme hypotézu 2_2H_0 tvrdící, že Y závisí pouze na vektoru X_{aa} proti alternativě, že Y závisí i na dalších složkách vektoru X_a

$${}^2_2H_0 : \beta_A = \beta_D = \beta_E = \beta_F = 0 \quad (= \beta_C = \beta_J = \beta_K = \beta_M),$$

čili hypotézu, že model

$$M : EY = E X_G = \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_E X_E + \beta_F X_F + \\ + \beta_H X_H + \beta_I X_I + \beta_J X_J + \beta_K X_K + \beta_L X_L + \beta_M X_M$$

s maticí modelu o hodnosti $h(\underline{X}) = 12$ se submodelem

$$M_1 : EY = E X_G = \beta_A X_A + \beta_B X_B + \beta_D X_D + \beta_E X_E + \beta_F X_F + \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí submodelu o hodnosti $h(\underline{X}) = 8$ lze dále zjednodušit na submodel

$$M_2 : EY = E X_G = \beta_B X_B + \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí submodelu o hodnosti $h(\underline{X}_{aa}) = 4$.

Vypočtěme nejprve součet čtverců

$$SS_{bb} = (n-1)S_{Y, X_a} S_{X_a, X_a}^{-1} S_{X_a, Y} - (n-1)S_{Y, X_{aa}} S_{X_{aa}, X_{aa}}^{-1} S_{X_{aa}, Y} = 63,786 - 62,520 = 1,266,$$

a poté testovou statistiku

$$F = \frac{\frac{SS_{bb}}{f_{bb}}}{\frac{SS_e}{f_e}} = \frac{\frac{1,266}{8-4}}{\frac{6,780}{72-12}} = \frac{0,316}{0,113} = 2,801.$$

Uvedme tabulku analýzy rozptylu pro test nulovosti skupiny regresních parametrů příslušejících podvektoru X_{bb} :

Zdroj variability	Součet čtverců	Stupňů volnosti	F-statistika
X_{bb} při pevném X_{aa}	1,266	4	$F = 2,801$
X_{aa}	62,520	4	
X_b při pevném X_a	0,434	4	$F = 0,959$
reziduální	6,780	60	
celkový	71	72	

Tabulka 5.7: Tabulka analýzy rozptylu pro test M_2 proti M_1 jako submodelu M

Protože platí

$$F = 2,801 > F_{4,60} = 2,525,$$

zamítáme submodel M_2 ve prospěch M_1 jako submodelu modelu M .

Testovali jsme proti sobě modely navržené v předchozích podkapitolách. Nejprve jsme testovali jako submodely grafický model získaný selekčním algoritmem (tedy model M_1 s nezávisle proměnnými $X_A, X_B, X_D, X_E, X_F, X_H, X_I, X_L$) a grafický model získaný korelační a regresní analýzou (tedy model M_2 s nezávisle proměnnými X_B, X_H, X_I, X_L) proti modelu původnímu (tj. modelu M se všemi dvanácti nezávisle proměnnými). Žádný z uvažovaných submodelů jsme nezamítli ve prospěch modelu původního. Není tedy nutné popisovat nadvýnos X_G pomocí všech proměnných, které jsou k dispozici. Pak jsme testovali model M_2 proti modelu M_1 jako submodelu modelu M , tedy jestli lze po odejmutí proměnných X_C, X_J, X_K, X_M odejmout ještě proměnné X_A, X_D, X_E, X_F . Tímto postupným způsobem jsme zamítli zjednodušení M_1 na model M_2 se čtyřmi nezávisle proměnnými. Původní model M tedy lze zjednodušit na model M_2 se čtyřmi nezávisle proměnnými, ale zjednodušení pouze na model M_1 s osmi nezávisle proměnnými je vhodnější.

Nyní označme jako \mathbf{X} vektor

$$\mathbf{X} = (X_B, X_H, X_I, X_L)^T$$

a předpokládejme platnost modelu

$$M_2 : EY = E X_G = \beta_B X_B + \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí modelu o hodnotě $h(\mathbf{X}) = 4$.

Vektor \mathbf{X} rozdělme na dva subvektory

$$\mathbf{X}_a = (X_H, X_I, X_L)^T, \quad \mathbf{X}_b = (X_B)$$

a uvažujme hypotézu H_0 tvrdící, že Y závisí pouze na vektoru \mathbf{X}_a , to jest na proměnných odpovídajících finančním indexům, nikoli na proměnné související se složením portfolia fondu:

$$H_0 : \beta_B = 0,$$

čili hypotézu, že model M_2 lze zjednodušit na submodel

$$M_3 : EY = E X_G = \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí submodelu o hodnotě $h(\mathbf{X}_a) = 3$.

Součty čtverců jsou rovny

$$SS_b = (n-1)S_{Y,X} S_{\mathbf{X},\mathbf{X}}^{-1} S_{\mathbf{X},Y} - (n-1)S_{Y,\mathbf{X}_a} S_{\mathbf{X}_a,\mathbf{X}_a}^{-1} S_{\mathbf{X}_a,Y} = 62,520 - 60,844 = 1,676,$$

$$SS_a = (n-1)S_{Y,\mathbf{X}_a} S_{\mathbf{X}_a,\mathbf{X}_a}^{-1} S_{\mathbf{X}_a,Y} = 60,844,$$

$$SS_e = (n-1)S_{Y,Y} - (n-1)S_{Y,\mathbf{X}} S_{\mathbf{X},\mathbf{X}}^{-1} S_{\mathbf{X},Y} = 71 - 62,520 = 8,480$$

a testová statistika je rovna

$$F = \frac{\frac{SS_b}{f_b}}{\frac{SS_e}{f_e}} = \frac{\frac{1,676}{4-3}}{\frac{8,480}{72-4}} = \frac{1,676}{0,125} = 13,443.$$

Opět prostudujme tabulku analýzy rozptylu pro test nulovosti skupiny regresních parametrů příslušejících podvektoru \mathbf{X}_b :

Zdroj variability	Součet čtverců	Stupňů volnosti	F-statistika
X_b při pevném X_a	1,676	1	$F = 13,443$
X_a	60,844	3	
reziduální	8,480	68	
celkový	71	72	

Tabulka 5.8: Tabulka analýzy rozptylu pro test M_3 proti M_2

Jelikož

$$F = 13,443 > F_{1,68} = 3,982,$$

zamítáme submodel M_3 ve prospěch modelu M_2 , tedy model APT z tabulky 5.3 ve prospěch modelu grafického z téže tabulky.

Otestujme ještě model CAPM z tabulky 5.3 proti modelu APT.

Označíme jako \mathbf{X} vektor

$$\mathbf{X} = (X_H, X_I, X_L)^T$$

se dvěma subvektory

$$\mathbf{X}_a = (X_L), \quad \mathbf{X}_b = (X_H, X_I),$$

předpokládejme platnost modelu

$$M_3 : \mathbf{E} Y = \mathbf{E} X_G = \beta_H X_H + \beta_I X_I + \beta_L X_L$$

s maticí modelu o hodnosti $h(\mathbf{X}) = 3$ a uvažujme hypotézu H_0 tvrdící, že Y závisí pouze na vektoru X_a

$$H_0 : \quad \beta_H = \beta_I = 0,$$

čili hypotézu, že model M_3 lze zjednodušit na submodel

$$M_4 : \mathbf{E} Y = \mathbf{E} X_G = \beta_L X_L$$

s maticí submodelu o hodnosti $h(\mathbf{X}_a) = 1$.

Součty čtverců jsou

$$SS_b = (n-1)S_{Y,X} S_{\mathbf{X},\mathbf{X}}^{-1} S_{X,Y} - (n-1)S_{Y,X_a} S_{X_a,X_a}^{-1} S_{X_a,Y} = 60,844 - 55,987 = 4,857,$$

$$SS_a = (n-1)S_{Y,X_a} S_{X_a,X_a}^{-1} S_{X_a,Y} = 55,987,$$

$$SS_e = (n-1)S_{Y,Y} - (n-1)S_{Y,X} S_{\mathbf{X},\mathbf{X}}^{-1} S_{X,Y} = 71 - 60,844 = 10,156,$$

$$SS_T = (n-1)S_{Y,Y} = 71,$$

a testová statistika je rovna

$$F = \frac{\frac{SS_b}{f_b}}{\frac{SS_e}{f_e}} = \frac{\frac{4,857}{3-1}}{\frac{10,156}{72-3}} = \frac{2,429}{0,147} = 16,500.$$

Údaje jsou pro zpřehlednění opět uvedeny v tabulce.

Zdroj variability	Součet čtverců	Stupňů volnosti	F-statistika
X_b při pevném X_a	4,857	2	$F = 16,500$
X_a	55,987	1	
reziduální	10,156	69	
celkový	71	72	

Tabulka 5.9: Tabulka analýzy rozptylu pro test M_4 proti M_3

Pro vysokou testovou statistiku

$$F = 16,500 > F_{2,69} = 3,130$$

zamítáme submodel M_4 ve prospěch modelu M_3 , tedy CAPM ve prospěch APT.

Testovali jsme nyní modely získané korelační a regresní analýzou, jejichž přehled je uveden v tabulce 5.3. Předpokládali jsme vždy platnost širšího modelu a testovali proti němu model s méně závisle proměnnými. Nejprve jsme testovali navržený APT model s nezávisle proměnnými X_H , X_I , X_L proti grafickému modelu s nezávisle proměnnými X_B , X_H , X_I , X_L , tedy jsme testovali možnost vynětí proměnné X_B z modelu grafického, pak jsme testovali CAPM model s nezávisle proměnnou X_L proti APT modelu, což odpovídalo testování možnosti vynětí proměnných X_H , X_I z APT modelu. V obou případech jsme zjednodušení zamítli, což odpovídá i hodnotám koeficientu determinace 3 modelů z tabulky 5.3.

6. Burzovní indexy českého trhu

Tvorba regresních modelů pro analýzu výnosu italského investičního fondu v předešlé kapitole byla založena na výběrové varianční matici a provedena bez znalosti původních dat. Na začátku této kapitoly zpracujeme podobným způsobem známá data z české finanční praxe, včetně aplikace vhodných transformací a ověření předpokladů použitých metod. Dále navrhne grafický model podmíněných nezávislostí a nakonec navrhne grafický model s řetězovým grafem podmíněných nezávislostí, ve kterém už bude na rozdíl od grafického modelu s neorientovaným grafem naznačena hierarchická struktura dat a na rozdíl od předchozích postupů tvorby regresních modelů několik hierarchicky uspořádaných vrstev, přičemž vrstvy závisle proměnných nebudou nutně obsahovat pouze jednu proměnnou.

Budeme zkoumat závislosti burzovních indexů, konkrétně celkového indexu kapitálového trhu IBIX a subindexů, které zahrnují pouze cenné papíry společností z jednotlivých odvětví, a to: zemědělství, potravinářství, výroba nápojů, těžební průmysl, textilní průmysl, dřevozpracující průmysl, chemický průmysl, stavebnictví, hutní průmysl, strojírenství, elektrotechnický průmysl, energetika, doprava, obchod, finančnictví, služby, výroba bižuterie, investiční fondy. Poslední skupinu tvoří do předchozích skupin nezařazené cenné papíry. K těmto indexům máme k dispozici časové řady o délce 241 pozorování, vyjadřující denní pozorování z let 1993, 1994. Budeme se snažit pracovat pouze s veličinami, jejichž realizace splňují předpoklad normality a nezávislosti. Veličiny, se kterými budeme dále pracovat, vytvoříme z původních veličin transformací první diference přirozených logaritmů. Testování normality dat bylo provedeno Kolmogorov - Smirnovovým testem v programu Statistica. Na pětiprocentní hladině nezamítáme hypotézu, že daná náhodná veličina má normální rozdělení, u veličin získaných ze subindexů pro *zemědělství, potravinářství, textilní průmysl, dřevozpracující průmysl, chemický průmysl, stavebnictví, hutní průmysl, strojírenství, elektrotechnický průmysl, obchod, služby, výroba bižuterie* a indexu *ostatní průmysl*. Vzhledem k tomu, že regresní model i selekční algoritmy pro grafické modely vyžadují vzájemnou nezávislost realizací každé užívané náhodné veličiny, vyšetříme pomocí znaménkového testu uvedeného v knize Cipra [3] platnost hypotéz o jejich nezávislosti pro všechny subindexy splňující předpoklad normality. Po vyloučení subindexu *služby*, u kterého jsme na rozdíl od zbývajících na pětiprocentní hladině zamítli hypotézu nezávislosti jednotlivých realizací, nám zbývají náhodné veličiny vytvořené transformací subindexů *zemědělství* (proměnné vztahující se k této veličině budou indexovány písmenem Z), *potravinářství* (P), *textilní průmysl* (T), *dřevozpracující průmysl* (D), *chemický průmysl* (Ch), *stavebnictví* (Sta), *hutní průmysl* (H), *strojírenství* (Str), *elektrotechnický průmysl* (E), *obchod* (O), *bižuterii* (B) a *ostatní průmysl* (Os). Z této množiny ještě vyjme proměnné vztahující se k subindexům *ostatní průmysl* (Os) a *bižuterie* (B), které lze těžko zařadit do hierarchické struktury subindexů potřebné ke konstrukci grafického modelu s řetězovým grafem podmíněných nezávislostí. Zbývající veličiny použijeme v regresním modelu jako nezávisle proměnné. Veličina získaná transformací indexu *IBIX* sice na pětiprocentní hladině nespĺňuje předpoklad normality, ale vzhledem k tomu, že histogram je jednovrcholový a na hladině 1 % lze normalitu akceptovat, použijeme ji v regresní a korelační analýze a v selekci grafického modelu jako logicky se nabízející závisle proměnnou $Y = X_I$ a v případě tvorby grafického modelu s řetězovým grafem jako samostatnou složku posledního bloku veličin.

Stejně jako v předešlé kapitole byla provedena standardizace dat, tedy varianční matice je zároveň i maticí korelační.

Výpočty byly provedeny v programu popsaném v kapitole 4.6, mezivýsledky budeme uvádět v zaokrouhleném tvaru.

6.1 Regresní analýza

Z matice dat \underline{X} vypočteme výběrovou varianční matici $S_{X,X}$ (viz příloha – matice 2) s inverzí $S_{X,X}^{-1}$.

Z matice \underline{X} a vektoru Y , což je vektor pozorování transformace indexu *IBIX*, dostaneme kovarianční matici $S_{X,Y}$.

Matice $S_{X,Y}$
0,122
0,191
0,481
0,372
0,360
0,369
0,379
0,521
0,344
0,063

Použitím těchto výběrových matic dostaneme odhad parametru β :

$$b = (-0,062; -0,040; 0,199; 0,090; 0,167; 0,038; 0,133; 0,246; 0,078; -0,123)^T .$$

Náhodnou veličinu Y tedy budeme predikovat pomocí vztahu

$$\hat{Y} = -0,062 \cdot X_Z - 0,040 \cdot X_D + 0,199 \cdot X_{Ch} + 0,090 \cdot X_H + 0,167 X_P + 0,038 \cdot X_T \\ + 0,133 \cdot X_{Sta} + 0,246 \cdot X_{Str} + 0,078 \cdot X_E - 0,123 \cdot X_O .$$

Nyní můžeme pomocí věty 2.6 spočítat R^2 :

$$p_{I,I}^0 = 1,630 , \\ R^2 = \frac{p_{I,I}^0 - 1}{p_{I,I}^0} = \frac{1,630 - 1}{1,630} = 0,386 .$$

Model tedy vystihuje variabilitu indexu *IBIX* zhruba z třiceti devíti procent. V tabulce 6.1 jsou hodnoty testových statistik pro test nulovosti jednotlivých regresních koeficientů v aktuálním modelu, u tučně vyištěných hodnot zamítáme nulovost koeficientu. Dále tam jsou hodnoty R^2 v modelech získaných vynecháním příslušné veličiny.

Proměnná	Index	Významnost regr. koef.	R^2 v reduk. modelu
X_Z	zemědělství	-1,127	0,383
X_D	dřevozprac. průmysl	-0,699	0,385
X_{Ch}	chemický průmysl	2,964	0,363
X_H	hutní průmysl	1,420	0,381
X_P	potravinářství	2,905	0,364
X_T	textilní průmysl	0,587	0,385
X_{Sta}	stavebnictví	2,070	0,375
X_{Str}	strojírenství	3,277	0,357
X_E	elektrotech. průmysl	1,308	0,382
X_O	obchod	-2,200	0,373

Tabulka 6.1: Významnost regresních koeficientů

6.2 Korelační analýza

Zkusme otestovat významnost výběrových korelačních koeficientů a výběrových koeficientů parciálních korelací proměnné IBIX s ostatními proměnnými. Pro dvojici proměnných $X_Z, Y = X_I$ má podle věty 1.1 za předpokladu nekorelovanosti náhodná veličina

$${}_{10}T_Z = \frac{r_{Z,I}}{\sqrt{1 - r_{Z,I}^2}} \cdot \sqrt{238}$$

rozdělení t_{238} .

Podle věty 1.3 má za předpokladu nekorelovanosti náhodná veličina

$${}_{10}T_{Z|N(X;Z)} = \frac{r_{Z,I|N(X;Z)}}{\sqrt{1 - r_{Z,I|N(X;Z)}^2}} \cdot \sqrt{229}$$

rozdělení t_{229} .

Jelikož po dosažení příslušných korelačních koeficientů dostaneme hodnoty

$${}_{10}T_Z = 1,901,$$

$${}_{10}T_{Z|N(X;Z)} = -1,125,$$

čili hodnoty v absolutní hodnotě menší než $t_{238}(0,05) \doteq 1,97$, $t_{229}(0,05) \doteq 1,97$, nezamítáme nekorelovanost indexu zemědělství a indexu IBIX bez ohledu na to, zda připouštíme vliv ostatních indexů či nikoliv. Obdobně lze postupovat u ostatních odvětvových indexů ve dvojici s IBIXem.

V tabulce 6.2 jsou statisticky významné korelace a parciální korelace vyznačeny tučně.

Odvětví	Proměnná	Korelace	Parciální korelace
zemědělství	X_Z	0,122	-0,074
dřevozpracující průmysl	X_D	0,191	-0,046
chemický průmysl	X_{Ch}	0,481	0,192
hutní průmysl	X_H	0,372	0,093
potravinářství	X_P	0,360	0,188
textilní průmysl	X_T	0,369	0,039
stavebnictví	X_{Sta}	0,379	0,135
strojírenství	X_{Str}	0,521	0,211
elektrotechnický průmysl	X_E	0,344	0,086
obchod	X_O	0,063	-0,144

Tabulka 6.2: Významnost korelací a parciálních korelací

Analýzou parciálních korelací jsme dospěli ke stejnému modelu jako analýzou regrese.

6.3 Mnohorozměrná regrese

Uvažujme nyní závisle proměnné X_Z (označme jako X_1), X_D (X_2), X_{Ch} (X_3), X_H (X_4) a nezávisle proměnné X_P (označme jako Y_1), X_T (Y_2), X_{Sta} (Y_3) a X_{Str} (Y_4).

Matici \mathbf{B} neznámých regresních koeficientů odhadneme podle kapitoly 3.6 jako

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \begin{pmatrix} 0,137 & 0,013 & 0,135 & 0,060 \\ 0,135 & 0,214 & 0,184 & 0,083 \\ 0,218 & 0,238 & 0,303 & 0,426 \\ 0,080 & 0,309 & 0,084 & 0,287 \end{pmatrix}.$$

Blok b_1 bychom tedy pomocí bloku b_2 odhadovali předpisem

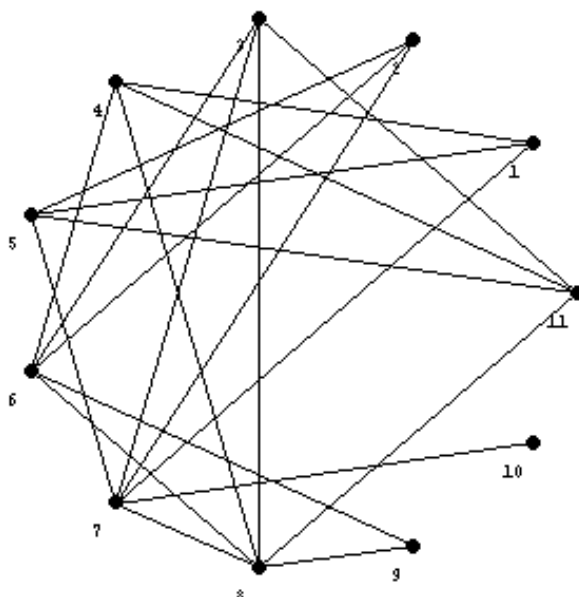
$$\hat{\mathbf{Y}} = \mathbf{X} \begin{pmatrix} 0,137 & 0,013 & 0,135 & 0,060 \\ 0,135 & 0,214 & 0,184 & 0,083 \\ 0,218 & 0,238 & 0,303 & 0,426 \\ 0,080 & 0,309 & 0,084 & 0,287 \end{pmatrix}.$$

6.4 Selekcce grafického modelu

Pro porovnání se teď podívejme na případ, kdy nepracujeme s regresním modelem, ale s grafickým modelem získaným *backward algorithmem se stop pravidlem založeným na devianci vynechané hrany*. Jak již víme, tento model popíše celkovou strukturu vzájemných souvislostí všech proměnných. Graf podmíněných nezávislostí získaný tímto algoritmem najdeme na obrázku 6.1.

Počet vynechaných hran = 34

Celková deviance = $34.9985 < \chi_{34}^2 = 48,60$



Obrázek 6.1

Vrchol v grafu	Proměnná	Index
1	X_Z	zemědělství
2	X_D	dřevozpracující průmysl
3	X_{Ch}	chemický průmysl
4	X_H	hutní průmysl
5	X_P	potravinářství
6	X_T	textilní průmysl
7	X_{Sta}	stavebnictví
8	X_{Str}	strojírenství
9	X_E	elektrotechnický průmysl
10	X_O	obchod
11	$Y = X_I$	IBIX

Tabulka 6.3: Označení bodů v grafu na obrázku 6.1

Tento graf, na rozdíl od analýzy parciálních korelací, která by vedla ke grafu s hranami spojujícími vrchol související s veličinou $Y = X_I$ s vrcholy souvisejícími s veličinami X_{Ch} , X_P , X_{Sta} , X_{Str} , X_O , obsahuje hrany spojující vrchol spjatý s veličinou $Y = X_I$ s vrcholy souvisejícími s veličinami X_{Ch} , X_H , X_P , X_{Str} . Index IBIX je tedy podle modelu navrženého backward algoritmem ovlivněn chemickým, hutním, potravinářským a strojírenským průmyslem. Dále lze z grafu studovat vzájemné ovlivňování jednotlivých odvětvových indexů. Shodu modelu s daty opět nezamítáme.

6.5 Grafický model s řetězovým grafem

Vytvořme nyní grafický model s řetězovým grafem. Z množiny vrcholů reprezentujících veličiny vytvořené transformací první diference logaritmů českých burzovních subindexů, se kterými v této kapitole pracujeme, vytvořme následující bloky:

$$b_1 = \{Z, D, Ch, H\},$$

$$b_2 = \{P, T, Sta, Str\},$$

$$b_3 = \{E, O\},$$

$$b_4 = \{I\}.$$

Program navrhne vhodný grafický model s řetězovým grafem vycházející z námi zadané hierarchické struktury dat a prezentuje ho pomocí matice souslednosti, což je čtvercová matice, jejíž rozměry jsou shodné s počtem vrcholů v grafu. Hodnota na jejím i -tém řádku a j -tém sloupci informuje, zda a jaká hrana je mezi i -tým a j -tým vrcholem. Vystačíme si pouze s dolní trojúhelníkovou maticí, diagonálu a horní trojúhelník tedy vyplníme nulami.

V prvním kroku bereme jako nezávisle proměnné vrcholy z bloku b_1 , čili Z, D, Ch, H a jako závisle proměnné vrcholy z bloku b_2 , čili P, T, Sta, Str . Program navrhl grafický model s následující maticí sousledností

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ * & 1 & * & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & * & 1 & 1 & 0 & 0 \\ * & * & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Přítomnost či nepřítomnost orientované či neorientované hrany mezi i -tým a j -tým vrcholem, přičemž $i < j$ a vrcholy jsou očíslovány v pořadí $Z, D, Ch, H, P, T, Sta, Str$, je vyjádřena prvkem na i -tém řádku a j -tém sloupci matice. V matici byly pro větší názornost naznačeny hranice mezi bloky vrcholů. Tuto matici lze rozdělit na čtyři bloky.

V levém horním bloku jsou zaznamenány informace o hranách mezi vrcholy z bloku b_1 . Tato struktura hran byla získána backward algoritmem podle bodu A_1 z kapitoly 4.5. Jednička znamená přítomnost neorientované hrany, hvězdička nepřítomnost jakékoliv hrany. Jednička znamená přítomnost neorientované hrany, hvězdička nepřítomnost jakékoliv hrany.

V levém dolním bloku můžeme studovat informace o orientovaných hranách vycházejících z vrcholů z bloku b_1 do vrcholů z bloku b_2 .

V pravém dolním bloku jsou zaznamenány informace o hranách mezi vrcholy z bloku b_2 . Jednička znamená přítomnost neorientované hrany, hvězdička nepřítomnost jakékoliv hrany. V našem případě jsou všechny hrany v bloku přítomny, blok je tedy úplným podgrafem grafu celkového, klikou. Struktura hran v dolních blocích matice souslednosti je výsledkem bodu B_1 z kapitoly 4.5.

Z možných nových hran jich bylo vynecháno šest, z toho jedna neorientovaná

$$(Z, D)$$

a pět orientovaných

$$(Z, T), (Z, Str), (D, Str), (Ch, T), (H, Sta).$$

Podle bodu C_1 z kapitoly 4.5 dostáváme, že deviance grafu obsahujícího pouze vrcholy z bloku b_1 byla programem spočítána jako

$$\text{dev } G(Z, D, Ch, H) = 2,84428 < \chi_1^2 = 3,84,$$

Deviance grafu obsahujícího všechny vrcholy z bloků b_1, b_2 , přičemž vrcholy z bloku b_1 tvoří úplný graf, vyšla

$$\text{dev } G((Z.D.Ch.H), P, T, Sta, Str) = 2,89715 < \chi_5^2 = 11,07$$

a konečně deviance celkového navrženého grafu s vrcholy z bloků b_1 a b_2 je

$$\begin{aligned} & \text{dev } G(Z, D, Ch, H, P, T, Sta, Str) = \\ & = \text{dev } G(Z, D, Ch, H) + \text{dev } G((Z.D.Ch.H), P, T, Sta, Str) = 5,74143 < \chi_6^2 = 12,59, \end{aligned}$$

takže nezamítáme výsledný graf ve prospěch grafu úplného s příslušnými vrcholy.

Dále postupujeme podle bodů A_2, B_2, C_2 v kapitole 4.5.

Ve druhém kroku bereme jako nezávisle proměnné vrcholy z bloků b_1 a b_2 , čili $Z, D, Ch, H, P, T, Sta, Str$ a jako závisle proměnné proměnné související s vrcholy z bloku b_3 , čili s vrcholy E, O . Program pro danou situaci navrhl grafický model s následující maticí sousledností

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 1 & * & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & * & 1 & 1 & 0 & 0 & 0 & 0 \\ * & * & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ \\ * & * & * & 1 & * & 1 & * & 1 & 0 & 0 \\ * & * & * & * & * & * & 1 & * & * & 0 \end{pmatrix}.$$

Přítomnost či nepřítomnost orientované či neorientované hrany mezi i -tým a j -tým vrcholem ($i < j$) je vyjádřena jedničkou (hrana přítomna) nebo hvězdičkou (hrana nepřítomna) na v j -tém řádku a i -tém sloupci matice. V matici byly pro větší názornost naznačeny hranice mezi bloky proměnných. Tuto matici můžeme rozdělit na devět částí. Levé horní čtyři bloky matice jsou převzaty z předchozího kroku (záporná čísla byla nahrazena hvězdičkami, které v levém horním rohu jako dříve záporná čísla ukazují na nepřítomnost neorientované hrany), nové jsou bloky vpravo a dole.

V levém dolním bloku je vyjádřena přítomnost (jedničkou) či nepřítomnost (hvězdičkou) orientované hrany vycházející vždy z příslušného vrcholu z bloku b_1 a končící v příslušném vrcholu z bloku b_3 .

V dolním bloku uprostřed je podobně vyjádřena přítomnost, či nepřítomnost orientované hrany vycházející z příslušného vrcholu z bloku b_2 a končící v příslušném vrcholu z bloku b_3 .

V pravém dolním bloku můžeme pozorovat přítomnost, či nepřítomnost hrany (neorientované) mezi vrcholy bloku b_3 .

Z nových možných hran jich bylo vynecháno třináct, z toho jedna neorientovaná

$$(E, O)$$

a dvanáct orientovaných

$$(Z, E), (D, E), (Ch, E), (P, E), (Sta, E),$$

$$(Z, O), (D, O), (Ch, O), (H, O), (P, O), (T, O), (Str, O).$$

Deviance grafu obsahujícího všechny vrcholy z bloků b_1, b_2, b_3 , přičemž vrcholy z bloků b_1 a b_2 tvoří společně úplný graf, je

$$\text{dev } G((Z.D.Ch.H.P.T.Sta.Str), E, O) = 8,46598 < \chi_{13}^2 = 22,36$$

a konečně deviance celkového navrženého grafu je

$$\begin{aligned} & \text{dev } G(Z, D, Ch, H, P, T, Sta, Str, E, O) = \\ & = \text{dev } G(Z, D, Ch, H, P, T, Sta, Str) + \text{dev } G((Z.D.Ch.H.P.T.Sta.Str), E, O) = \\ & = 14,2074 < \chi_{19}^2 = 30,14 \end{aligned}$$

takže nezamítáme výsledný graf ve prospěch grafu úplného s příslušnými vrcholy.

Ve třetím a zároveň posledním kroku bereme jako nezávisle proměnné ty, které souvisejí s vrcholy z bloků b_1, b_2 a b_3 , čili s vrcholy $Z, D, Ch, H, P, T, Sta, Str, E, O$, a jako závisle proměnné ty, které souvisejí s vrcholy z bloku b_4 , čili s vrcholem IBIX . Program pro danou situaci navrhl grafický model s následující maticí sousledností

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & 1 & * & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & * & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ * & * & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ \\ * & * & * & 1 & * & 1 & * & 1 & 0 & 0 & 0 \\ * & * & * & * & * & * & 1 & * & * & 0 & 0 \\ \\ * & * & 1 & * & 1 & * & 1 & 1 & * & 1 & 0 \end{pmatrix}.$$

Přítomnost či nepřítomnost orientované mezei i -tým vrcholem ($i < 11$) a vrcholem $IBIX$ je vyjádřena jedničkou (hrana přítomna) nebo hvězdičkou (hrana nepřítomna) na posledním řádku a i -tém sloupci matice.

Z možných nových orientovaných hran jich bylo vynecháno pět

$$(Z, IBIX), (D, IBIX), (H, IBIX), (T, IBIX), (E, IBIX).$$

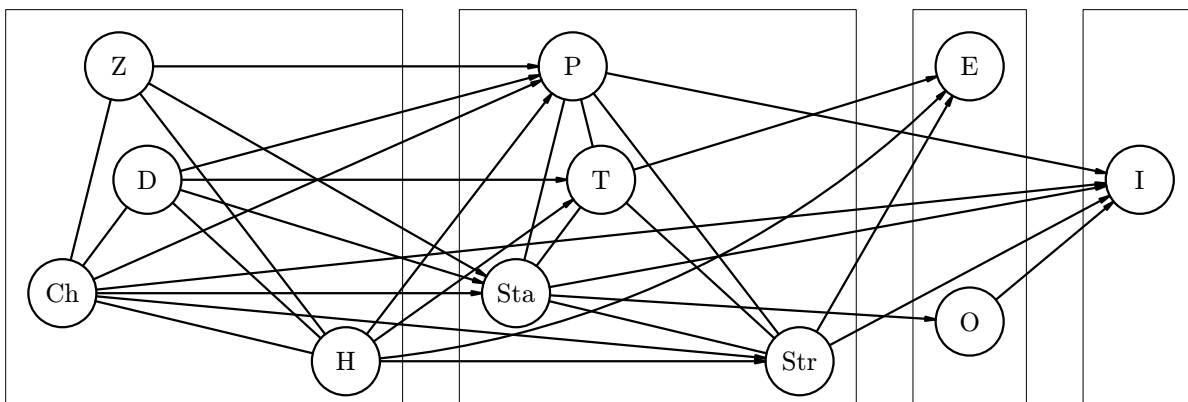
Deviance grafu obsahujícího všechny vrcholy, přičemž všechny kromě vrcholu $IBIX$ tvoří úplný graf, je

$$\text{dev } G((Z.D.Ch.H.P.T.Sta.Str.E.O), IBIX) = 6,74358 < \chi_5^2 = 11,07$$

a konečně deviance celkového navrženého grafu je

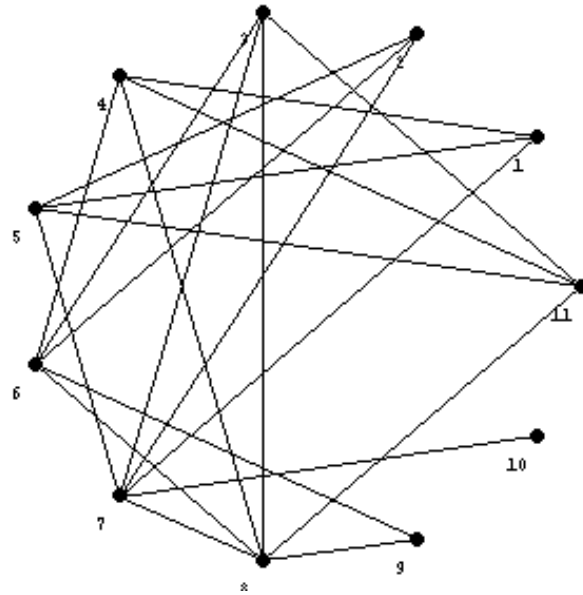
$$\begin{aligned} \text{dev } G(Z, D, Ch, H, P, T, Sta, Str, E, O, IBIX) &= \\ &= \text{dev } G(Z, D, Ch, H, P, T, Sta, Str, E, O) + \\ &+ \text{dev } G((Z.D.Ch.H.P.T.Sta.Str.E.O), IBIX) = 20,951 < \chi_{24}^2 = 36,42 \end{aligned}$$

takže nezamítáme definitivní výsledný graf ve prospěch grafu úplného s příslušnými vrcholy (viz obrázek 6.2).

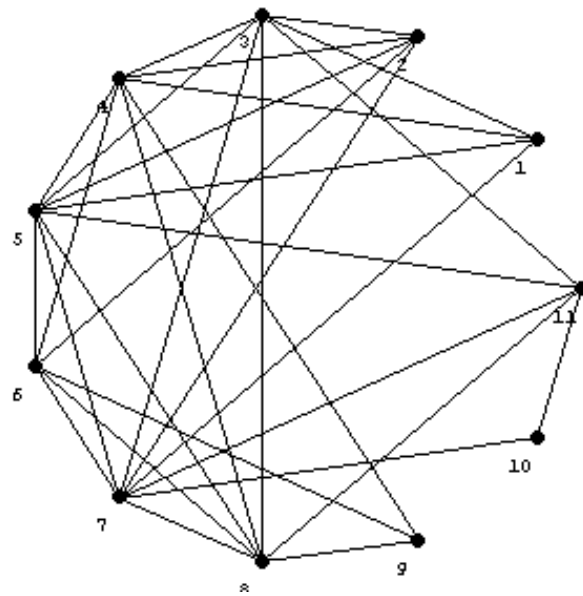


Obrázek 6.2

Tento model dospěl ke stejným proměnným (X_{Ch} , X_P , X_{Sta} , X_{Str} , X_O) přímo ovlivňujícím index $IBIX$ jako regresní a korelační analýza. Zajímavým výsledkem je to, že je v modelu obchod přímo ovlivněn pouze stavebnictvím. Zřejmě poněkud překvapivým výsledkem je vzájemné ovlivňování textilního průmyslu s potravinářstvím, stavebnictvím a strojírenstvím, a že dále ovlivňuje elektrotechnický průmysl. Na obrázcích 6.3 a 6.4 můžeme porovnat neorientovaný graf podmíněných nezávislostí získaný selekčním backward algoritmem s právě získaným grafem, ve kterém pro lepší možnost porovnání není naznačena hierarchická struktura dat ani orientace hran. Vidíme, že předem zadané rozdělení dat na skupiny vzájemně se ovlivňujících proměnných zvětšilo počet hran v grafu.



Obrázek 6.3



Obrázek 6.4

Závěr

Cílem této práce bylo prezentovat různé modely regresního typu používané v analýze výnosu portfolia a jiných finančních dat. Po shrnutí potřebné teorie jsme zpracovali dva soubory dat.

První soubor byl převzat z literatury Giudici [4]. Ukázali jsme na něm, že modely CAPM a APT lze interpretovat jako lineární regresní modely. Závisle proměnnou v nich byl nadvýnos, to jest rozdíl výnosu portfolia fondu a bezrizikového výnosu. Nezávisle proměnné byly zvoleny podle Giudici [4]. V APT modelu jsme ukázali, že analýza regresních koeficientů a analýza koeficientů parciální korelace vedou ke stejnému zjednodušení množiny nezávisle proměnných. Pak jsme uvažovali všechny možné nezávisle proměnné a výběr z nich do regresního modelu jsme učinili na základě testování nulovosti parciálních korelací. Znamená to, že jsme testovali na zvolené hladině přítomnost jednotlivých hran vycházejících z vrcholu pro nadvýnos v grafickém modelu s třináctivrcholovým grafem. Tento postup potvrdil statistickou významnost nezávisle proměnných určených jako významné v CAPM a APT modelu. Koeficienty determinace ukázaly, že všechny vyšetřované modely dobře popisují variabilitu nadvýnosu. Nakonec jsme selekčním algoritmem našli grafický model popisující celkovou strukturu podmíněných nezávislostí v množině vyšetřovaných proměnných. Selekční algoritmus potvrdil významnost faktorů ovlivňujících nadvýnos, které jsme získali při testech nulovosti parciálních korelačních koeficientů. Dobrá shoda vybraného grafického modelu s daty byla potvrzena statistickým testem založeným na devianci modelu. Pak jsme uvažované modely testovali proti sobě jako submodely lineárního regresního modelu. Ukázali jsme, že model se všemi dvanácti nezávisle proměnnými lze zjednodušit na submodel se sedmi nezávisle proměnnými navržený selekčním algoritmem i na submodel se čtyřmi nezávisle proměnnými navržený korelační a regresní analýzou, ovšem pomocí testování druhého submodelu proti submodelu prvnímu jsme zjistili, že vhodnější je zjednodušit původní model na model se sedmi nezávisle proměnnými. Dále jsme testovali submodel APT se třemi proměnnými proti modelu grafickému získanému korelační a regresní analýzou se čtyřmi proměnnými a submodel CAPM s jednou proměnnou proti modelu se třemi proměnnými. V obou případech jsme možné zjednodušení modelu zamítli.

Druhý soubor představují data z české finanční praxe. Hledali jsme souvislost mezi burzovním indexem IBIX a vybranými odvětvovými indexy. Regresní analýza, testování přítomnosti jednotlivých hran z vrcholu pro IBIX neboli testování nulovosti parciálních korelací i selekční algoritmus k nalezení optimálního jedenáctivrcholového grafu potvrdily vliv odvětví potravinářství, chemický průmysl a strojírenství na hodnoty IBIXU. Regresní a korelační analýza nám jako ovlivňující index IBIX navrhly i odvětví stavební průmysl a obchod a nenavrhly hutní průmysl. Testování přítomnosti jednotlivých hran v grafu nám dává možnost výběru nezávisle proměnných do regresního modelu. Selekce grafického modelu vhodným algoritmem pak navíc poskytne přehled o celkové struktuře vzájemných souvislostí v množině vyšetřovaných proměnných. Grafické modely tedy mohou být kombinovány s modely regresními a představují část statistiky aplikovatelnou ve finančních analýzách. Nakonec jsme sestrojili grafický model s řetězovým grafem. Výhodou tohoto modelu je, že při sledování vzájemných závislostí všech jednotlivých proměnných zároveň respektuje předpokládanou hierarchickou strukturu dat, a to dokonce tím způsobem, že nepovoluje jen dvě vrstvy proměnných (tj. nezávisle proměnné a závisle proměnné), ale

povoluje jich libovolně mnoho. Co se týká proměnných přímo ovlivňujících index IBIX, došli jsme pomocí grafického modelu s řetězových grafem ke stejnému výsledku jako u modelu s neorientovaným grafem a u regresní a korelační analýzy.

Literatura

- [1] Anděl J. (1978): Matematická statistika. SNTL, Praha
- [2] Anděl J. (1998): Statistické metody. Matfyzpress, Praha
- [3] Cipra T. (1986): Analýza časových řad s aplikacemi v ekonomii. SNTL, Praha
- [4] Giudici P. (2002): Data Mining. Mc Graw-Hill, Milano
- [5] Mardia K. V., Kent J. T. Bibby J. M. (1979): Multivariate Analysis. Academia Press, London
- [6] Whittaker J. (1990): Graphical models in applied multivariate Statistics. Wiley, New York
- [7] Zelinková J. (2003): Regrese a grafické modely pro finanční analýzu. Diplomová práce MFF UK, Praha

Přílohy