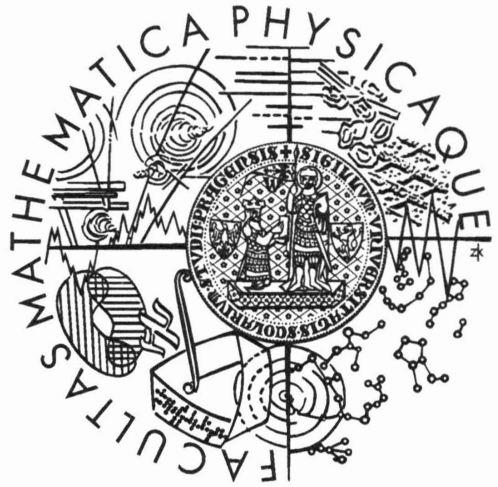
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Lucie Dřizgová

Multikolinearita

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Doc. RNDr. Karel Zvára, CSc.
Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán: Ekonometrie

Poděkování

Děkuji vedoucímu své diplomové práce Doc. RNDr. Karlu Zvárovi, CSc. za cenné rady a připomínky, které mi pomohly k vylepšení textu.

Čestné prohlášení

Prohlašuji, že jsem svoji práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.
V Praze dne 11. prosince 2009

Lucie Dřizgová

Abstrakt

Název práce: Multikolinearita
Autor: Lucie Dřizgová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Doc. RNDr. Karel Zvára, CSc.
Email vedoucího: karel.zvara@mff.cuni.cz
Abstrakt: V naší práci jsme se komplexně zabývali problémem multikolinearity – od metod pro diagnostiku multikolinearity až po metody určené pro překonání problémů multikolinearitou způsobených. Teoreticky jsme porovnali klasickou metodu nejmenších čtverců s alternativními metodami – regresí na hlavních komponentách, regresí pomocí parciálních nejmenších čtverců a hřebenovou regresí. V závěrečné sekci jsme ukázali použití všech metod na praktickém příkladu zpracovaném v programu R.
Klíčová slova: Multikolinearita, regrese na hlavních komponentách, parciální nejmenší čtverce, hřebenová regrese.

Abstract

Title: Multicollinearity
Author: Lucie Dřizgová
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. RNDr. Karel Zvára, CSc.
Supervisor's email address: karel.zvara@mff.cuni.cz
Abstract: In our work, we explored multicollinearity problem from a complex point of view – from diagnostic methods to the solving of the problems which are caused by the multicollinearity. We compared the Least Squares method with some alternative methods – Principal Component Regression, Partial Least Squares Regression and Ridge Regression on the theoretical basis. In the last section, we demonstrated all methods on practical example computed in the program R.
Key words: Multicollinearity, Principal Component Regression, Partial Least Squares, Ridge Regression.

Obsah

Kapitola 1

Úvod

V klasickém lineárním regresním modelu předpokládáme několik základních vlastností. Jednou z nich je lineární nezávislost vysvětlujících proměnných. Pokud dojde k porušení tohoto předpokladu a vysvětlující proměnné budou lineárně závislé, pak některé regresní parametry nebudou odhadnutelné. V případě, že vysvětlující proměnné budou téměř lineárně závislé, mohou mít odhady parametrů modelu metodou nejmenších čtverců značný rozptyl a být numericky nestabilní (při malé změně v hodnotách vysvětlujících proměnných se velmi změní odhad parametru), což snižuje oprávněnost jejich použití.

Právě problematickou situací, kdy máme vysvětlující proměnné vzájemně téměř lineárně závislé, se budeme v naší práci zabývat a ukážeme si metody, které si s multikolinearitou vysvětlujících proměnných poradí – těmi jsou například regrese na hlavních komponentách (Principal Component Regression), regrese parciálních nejmenších čtverců (Partial Least Squares Regression) a hřebenová regrese (Ridge Regression).

Jako první se uvažované metody začaly široce uplatňovat v chemometrii, v oboru zabývajícím se aplikacemi statistických metod na analýzu chemických dat. Jak upozorňují Franková a Friedman v [7], odhady pomocí metody parciálních nejmenších čtverců, regrese na hlavních komponentách nebo hřebenové regrese jsou výhodně aplikovány v analýze dat z potravinových výzkumů, z experimentů analytické chemie nebo z environmentálních studií. Jedním z důvodů je fakt, že díky vysokým cenám jednotlivých experimentů zde často dochází k situacím, kdy měřených veličin ovlivňujících pokus bylo mnohem více než počet pozorování. Využití těchto metod se samozřejmě neomezuje jen na analýzy dat zmíněných výše. Stone a Brooks v článku [18] zkoumají počet nehod v Minnesotě, Kidwellová a Brownová v článku [11] pomocí hřebenové regrese analyzují data o spokojenosti v manželství. My si v naší práci ukážeme použití těchto metod na lékařských datech. Metody

zpracované v naší práci si v kapitole 8 ilustrujeme na příkladu dat pořízených pro potřeby soudního lékařství a pokusíme se najít vhodný model pro určení stáří lidského plodu v závislosti na délkách a šířkách vybraných kostí.

V jednotlivých kapitolách nejprve zavedeme značení používané v celé práci a popíšeme si důsledky multikolinearity. Poté si poradíme s diagnostikou multikolinearity. Dále odvodíme střední kvadratickou chybu pro predikci metodou nejmenších čtverců a budeme se postupně zaobírat regresí na hlavních komponentách, metodou parciálních nejmenších čtverců a hřebenovou regresí a jednotlivé metody srovnáme s metodou nejmenších čtverců. Ve výkladu budeme občas přecházet mezi náhodnými proměnnými a mezi naměřenými hodnotami proměnných, jako tomu často bývá u technické literatury, ze které jsme vycházeli. Věřím, že pro čtenáře nebude problém se ve výkladu orientovat.

K metodám používaným pro regresi na téměř lineárně závislých datech existuje mnoho literatury. Citujeme ji obvykle přímo u popisovaných metod. Základní myšlenky všech metod jsou uvedeny v článcích Stone a Brooks [18], Franková a Friedman [7] nebo knize Rao, Toutenburg, Shalabh, Heumann [16]. Pro rozšíření naší práce mohou sloužit články Sundberg [20] a Stone a Brooks [18], kde autoři rozebírají spojité rozšíření předchozích metod, tzv. spojitou regresi CR (Continuum Regression), která zahrnuje všechny metody jako speciální (limitní) případy. K rozšíření může dále sloužit i článek Garthwaite [8], ve kterém je popsána metoda odhadů pomocí parciálních nejmenších čtverců aplikovaná na modely s více vysvětlovanými proměnnými.

Kapitola 2

Problém multikolinearity

Zajímáme se o závislost vysvětlované proměnné (závisle proměnné) na skupině vysvětlujících proměnných (nezávisle proměnných, regresorů) – snažíme se odhadnout vektor středních hodnot vysvětlované proměnné pomocí vysvětlujících proměnných. Máme k dispozici n pozorování $(\dot{\mathbf{x}}(i), \dot{y}(i)), i = 1, \ldots, n$. Značíme $\dot{\mathbf{x}}(i) = (\dot{x}_1(i), \dot{x}_2(i), \ldots, \dot{x}_p(i))'$ naměřené hodnoty p vysvětlujících proměnných pro pozorování i a $\dot{y}(i)$ je naměřená hodnota vysvětlované proměnné pro pozorování i.

Naměřené hodnoty vysvětlujících proměnných standardizujeme

$$x_j(i) = \frac{\dot{x}_j(i) - \bar{x}_j}{s_{x_j}}, \tag{2.1}$$

kde

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} \dot{x}_j(i), \; s_{x_j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} [\dot{x}_j(i) - \bar{x}_j]^2}. \tag{2.2}$$

Hodnoty vysvětlované proměnné pouze centrujeme

$$y(i) = \dot{y}(i) - \bar{y}, \tag{2.3}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \dot{y}(i). \tag{2.4}$$

Veličiny $x_j(i)$ bývají v některých odborných textech označovány jako tzv. z-skóry a mohou být definovány s hodnotou n namísto $n-1$ ve vzorci pro s_{x_j}. V dalším textu budeme (pokud nebude výslovně uvedeno jinak) pracovat s písmeny bez tečky, které označují hodnoty transformovaných proměnných, $\mathbf{X}$ označuje matici hodnot vysvětlovaných proměnných po standardizaci, $\mathbf{y}$ označuje centrované hodnoty vysvětlované proměnné.

V celé naší práci uvažujeme normální lineární model s plnou hodností (tj. matice $\mathbf{X}$ má plnou sloupcovou hodnost)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.5}$$

kde hodnoty vysvětlujících proměnných tvoří matici $\mathbf{X}$ rozměrů $n \times p$

$$\mathbf{X} = \begin{pmatrix} x_1(1) & \dots & x_p(1) \\ \vdots & \ddots & \vdots \\ x_1(n) & \dots & x_p(n) \end{pmatrix} \tag{2.6}$$

a vektory

$$\mathbf{y} = \begin{pmatrix} y(1) \\ \vdots \\ y(n) \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2.7}$$

označují (po řadě) hodnoty vysvětlované proměnné $(n \times 1)$, vektor příslušných regresních parametrů $(p \times 1)$, vektor jedniček $(n \times 1)$ a vektor chybových členů $(n \times 1)$. O vektoru chybových členů $\boldsymbol{\varepsilon}$ předpokládáme, že se jedná o realizaci náhodného vektoru s rozdělením

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \mathbf{I}_n\right), \tag{2.8}$$

kde $\mathbf{I}_n$ značí jednotkovou matici $(n \times n)$ a σ^2 je neznámý parametr.

Dále zavedeme značení

$$\mathbf{R}_{XX} = \frac{1}{n-1}\mathbf{X}'\mathbf{X} \tag{2.9}$$

pro výběrovou korelační matici vysvětlujících proměnných a

$$\mathbf{r}_{Xy} = \frac{1}{n-1}\frac{\mathbf{X}'\mathbf{y}}{s_y} \tag{2.10}$$

pro vektor výběrových korelačních koeficientů mezi vektorem $\mathbf{y}$ a maticí $\mathbf{X}$, kde definujeme s_y jako

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[\dot{y}(i) - \bar{y}\right]^2}. \tag{2.11}$$

Z předpokladu, že matice $\mathbf{X}$ má plnou hodnost, vyplývá, že i matice $\mathbf{X}'\mathbf{X}$ má plnou hodnost a matice $\mathbf{X}'\mathbf{X}$ i $\mathbf{R}_{XX}$ jsou tedy regulární a pozitivně definitní matice.

Odhad $\mathbf{b}_{OLS}$ metodou nejmenších čtverců je řešení soustavy normálních rovnic (viz např. Anděl [2] na straně 80), tj.

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b}_{OLS} &= \mathbf{X}'\mathbf{y} \\ \mathbf{b}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned} \tag{2.12}$$

a pomocí výběrové korelační matice vysvětlujících proměnných (2.9) a výběrového korelačního koeficientu mezi vektorem $\mathbf{y}$ a maticí $\mathbf{X}$ (2.10) jej můžeme vyjádřit jako

$$\begin{aligned} (n-1)\,\mathbf{R}_{XX}\mathbf{b}_{OLS} &= (n-1)\,s_y\mathbf{r}_{Xy} \\ \mathbf{R}_{XX}\mathbf{b}_{OLS} &= s_y\mathbf{r}_{Xy} \\ \mathbf{b}_{OLS} &= \mathbf{R}_{XX}^{-1}s_y\mathbf{r}_{Xy}. \end{aligned} \tag{2.13}$$

Podle Gaussovy-Markovovy věty (důkaz např. Zvára [24] věta 2.1 na straně 13) dostaneme, že $\mathbf{b}_{OLS}$ je nejlepší (myšleno odhad s nejmenším rozptylem) nestranný lineární odhad parametru $\boldsymbol{\beta}$ v normálním lineárním modelu (2.5). Pro rozptyl odhadu $\mathbf{b}_{OLS}$ platí vztah

$$\begin{aligned} \mathrm{Var}\,(\mathbf{b}_{\mathrm{OLS}}) &= \mathrm{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}\,(\mathbf{y})\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{2.14}$$

Nyní na chvíli odbočme a porovnejme výše uvedené s regresním modelem bez standardizace prvků matice $\dot{\mathbf{X}}$ a bez centrování $\dot{\mathbf{y}}$

$$\dot{\mathbf{y}} = \mathbf{1}\dot{\beta}_0 + \dot{\mathbf{X}}\dot{\boldsymbol{\beta}} + \dot{\boldsymbol{\varepsilon}}. \tag{2.15}$$

Upravíme pravou stranu rovnice tak, že dostáváme závislost $\dot{\mathbf{y}}$ na standardizovaných vysvětlujících proměnných s upravenými regresními koeficienty (s koeficienty bez tečky pro standardizované vysvětlující proměnné) jako

$$\begin{aligned} \dot{\mathbf{y}} &= \sum_{i=1}^{p}\left(\dot{\beta}_0 + \bar{x}_i\dot{\beta}_i\right) + \sum_{i=1}^{p}\frac{(\dot{\mathbf{x}}_\mathbf{i} - \bar{x}_i)\,\dot{\beta}_i s_{x_i}}{s_{x_i}} + \dot{\boldsymbol{\varepsilon}} \\ &= \beta_0^\star + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \end{aligned} \tag{2.16}$$

Pro tento model budou normální rovnice vypadat následovně

$$\begin{pmatrix} \mathbf{1}' \\ \mathbf{X}' \end{pmatrix}(\mathbf{1},\mathbf{X})\begin{pmatrix} b_0^\star \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \mathbf{1}' \\ \mathbf{X}' \end{pmatrix}\dot{\mathbf{y}} \tag{2.17}$$

$s_{x_i}\dot{\beta}_i = \beta_i$

a s využitím vztahu $\mathbf{1}'\mathbf{X} = \mathbf{0}'$ (a tedy i $\mathbf{X}'\dot{\mathbf{y}} = \mathbf{X}'(\mathbf{y} + \bar{y}\mathbf{1}) = \mathbf{X}'\mathbf{y}$) je můžeme dále upravit jako

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} \end{pmatrix} \begin{pmatrix} b_0^\star \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \mathbf{X}'\dot{\mathbf{y}} \end{pmatrix}$$
$$\begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}'\mathbf{X} \end{pmatrix} \begin{pmatrix} b_0^\star \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \mathbf{X}'\mathbf{y} \end{pmatrix}, \tag{2.18}$$

odkud už vidíme, že odhad $b_0^\star = \bar{y}$. Dostáváme

$$\begin{aligned} \dot{\mathbf{y}} - \mathbf{1}\bar{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \end{aligned} \tag{2.19}$$

Vidíme tedy, že absolutní člen je v modelu (2.5) pouze skrytý, a musíme uvažovat, že v modelu je vlastně $p+1$ odhadovaných parametrů, a zohlednit tento fakt zejména u stupňů volnosti při statistických testech. U příkladu budeme postupovat tedy tak, že absolutní člen v modelu ponecháme (bude stejně odhadován jako nula) a získáme tak správné stupně volnosti ve výstupu programu R.

Vraťme se zpět k původnímu výkladu. Vzhledem k tomu, že matice $\mathbf{X}'\mathbf{X}$ je pozitivně definitní matice hodnosti p, existuje spektrální rozklad matice $\mathbf{X}'\mathbf{X}$ jako (viz např. Zvára [24], věta A.3, strana 226)

$$\mathbf{X}'\mathbf{X} = \mathbf{Q}\boldsymbol{\Lambda}^2\mathbf{Q}', \tag{2.20}$$

kde $\boldsymbol{\Lambda}^2$ je diagonální matice s vlastními čísly $\lambda_1^2, \ldots, \lambda_p^2$ na diagonále a matice $\mathbf{Q}$ je matice vlastních vektorů matice $\mathbf{X}'\mathbf{X}$. Všechna vlastní čísla $\lambda_i^2, i = 1, \ldots p$, této pozitivně definitní matice jsou kladná (např. Anděl [2] v kapitole A.2 na straně 321) a můžeme bez újmy na obecnosti předpokládat, že platí

$$\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_p^2. \tag{2.21}$$

Spektrální rozklad matice $\mathbf{X}'\mathbf{X}$ můžeme přepsat jako

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{p} \lambda_i^2 \mathbf{q}_i \mathbf{q}_i', \tag{2.22}$$

kde $\mathbf{q}_i$ jsou příslušné vlastní vektory matice $\mathbf{X}'\mathbf{X}$, $i = 1, \ldots p$ (sloupce matice $\mathbf{Q}$). Pro inverzní matici (připomeňme, že $\mathbf{X}'\mathbf{X}$ je regulární) platí

$$(\mathbf{X}'\mathbf{X})^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \mathbf{q}_i \mathbf{q}_i' \tag{2.23}$$

a rozptyl odhadu v rovnici (2.14) pak můžeme vyjádřit jako

$$\mathrm{Var}\left(\mathbf{b}_{\mathrm{OLS}}\right) = \sigma^2 \sum_{\mathrm{i=1}}^{\mathrm{p}} \frac{1}{\lambda_i^2} \mathbf{q}_i \mathbf{q}_i'. \tag{2.24}$$

Vzhledem k tomu, že $\mathbf{q}_i$ jsou ortonormální vektory $\mathbf{q}_i'\mathbf{q}_i = 1$, pak i matice $\mathbf{q}_i\mathbf{q}_i'$ má omezené prvky, neboť všechny souřadnice ortonormálního vektoru musí být v absolutní hodnotě menší než 1. Ze vztahu (2.24) a z omezenosti prvků $\mathbf{q}_i\mathbf{q}_i'$ nahlédneme, jaký efekt budou mít malá vlastní čísla matice $\mathbf{X}'\mathbf{X}$ na rozptyl odhadu $\mathbf{b}_{OLS}$. Pokud se bude vlastní číslo λ_p^2 blížit nule a všechny prvky matice $\mathbf{q}_p\mathbf{q}_p'$ budou nenulové, pak se bude $\mathrm{Var}\left(\mathbf{b}_{OLS}\right)$ blížit nekonečnu.

Pro další výklad se nám bude hodit rozklad matice $\mathbf{X}$ podle singulárních hodnot (Singular Value Decomposition, viz např. Zvára [24], věta A.4, strana 226)

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{Q}', \tag{2.25}$$

kde $\mathbf{\Lambda}$ je diagonální matice, která má na diagonále singulární čísla matice $\mathbf{X}$, $\lambda_1 \geq \ldots \geq \lambda_p > 0$, $\mathbf{U}$ je ortonormální matice, jejíž sloupce jsou tvořeny levými singulárními vektory matice $\mathbf{X}$ příslušejícími singulárním číslům $\lambda_i, i = 1, \ldots p$, a $\mathbf{Q}$ je ortonormální matice, jejíž sloupce jsou tvořeny pravými singulárními vektory matice $\mathbf{X}$ příslušejícími singulárním číslům $\lambda_i, i = 1, \ldots p$. Pro matici pravých singulárních vektorů jsme použili záměrně stejné označení jako pro matici vlastních vektorů matice $\mathbf{X}'\mathbf{X}$, neboť díky ortonormalitě matice $\mathbf{U}$ platí

$$\begin{aligned}
\mathbf{X}'\mathbf{X} &= \left(\mathbf{U}\mathbf{\Lambda}\mathbf{Q}'\right)'\left(\mathbf{U}\mathbf{\Lambda}\mathbf{Q}'\right) \\
&= \mathbf{Q}\mathbf{\Lambda}'\mathbf{U}'\mathbf{U}\ \mathbf{\Lambda}\mathbf{Q}' \\
&= \mathbf{Q}\mathbf{\Lambda}'\mathbf{\Lambda}\mathbf{Q}' \\
&= \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}'.
\end{aligned} \tag{2.26}$$

Podobně matice $\mathbf{\Lambda}$ je matice, která má na diagonále odmocniny z vlastních čísel matice $\mathbf{X}'\mathbf{X}$.

Kapitola 3

Diagnostika multikolinearity

V kapitole 2 jsme naznačili důsledky multikolinearity a nyní se budeme věnovat tomu, jak ji v našich datech rozpoznat a jak zjistit, které vysvětlující proměnné problémy s multikolinearitou způsobují.

Multikolinearitu definujeme jako odchýlení od ortogonality v množině vysvětlujících proměnných, jak se uvádí v článku Farrara a Glaubera [5]. Pomocí této definice můžeme formulovat statistickou hypotézu a testovat přítomnost multikolinearity v datech.

Označme $|\mathbf{R}_{XX}|$ determinant výběrové korelační matice $\mathbf{R}_{XX}$ vysvětlujících proměnných a označme $r_{ij}, i, j = 1, \ldots p$, prvky matice $\mathbf{R}_{XX}$. Předpokládáme, že matice $\mathbf{X}$ má plnou hodnost, a jak už jsme si uvedli v kapitole 2, je matice $\mathbf{R}_{XX}$ pozitivně definitní. Potom můžeme použít tvrzení (např. v knize Anděl [2], vzorec (A.7), str. 324)

$$|\mathbf{R}_{XX}| \leq r_{11} r_{22} \ldots r_{pp}. \tag{3.1}$$

Vzhledem k tomu, že se jedná o determinant normalizované (výběrové korelační) matice (všechny diagonální prvky $r_{ii} = 1, i = 1, \ldots p$) a vzhledem k tomu, že determinant pozitivně definitní matice je vždy kladné číslo (např. Hebák a Hustopecký [10], str. 28), platí pro něj

$$0 < |\mathbf{R}_{XX}| \leq 1. \tag{3.2}$$

Dokážeme si, že determinant matice $\mathbf{R}_{XX}$ má hodnotu $|\mathbf{R}_{XX}| = 1$ právě tehdy, když jsou sloupce matice $\mathbf{X}$ ortonormální. Jestliže matice $\mathbf{X}$ má ortonormální sloupce, pak $\mathbf{R}_{XX}$ je jednotková matice a tedy její determinant je 1, $|\mathbf{R}_{XX}| = 1$. Důkaz opačné implikace vychází z nerovnosti mezi aritmetickým a geometrickým průměrem a ze spektrálního rozkladu (2.23). Matice $\mathbf{R}_{XX}$

má na diagonále samé jedničky a její stopa je rovna p. Zároveň pro ni platí

$$\begin{aligned} tr\left(\mathbf{R}_{XX}\right) &= tr\left(\mathbf{Q}\boldsymbol{\Lambda}^2\mathbf{Q}'\right) \\ &= \sum_{i=1}^{p} \lambda_i^2. \end{aligned} \tag{3.3}$$

Aritmetický průměr z kladných čísel λ_i^2, $i = 1, ...p$, je podle nerovnosti mezi aritmetickým a geometrickým průměrem větší nebo roven geometrickému průměru, pro který platí

$$\begin{aligned} \sqrt[p]{|\mathbf{R}_{XX}|} &= \sqrt[p]{|\mathbf{Q}\boldsymbol{\Lambda}^2\mathbf{Q}'|} \\ &= \sqrt[p]{|\boldsymbol{\Lambda}^2|} \\ &= \sqrt[p]{\lambda_1^2 \ldots \lambda_p^2}, \end{aligned} \tag{3.4}$$

a rovnost nastává právě tehdy, když všechna vlastní čísla matice $\mathbf{R}_{XX}$ jsou stejná (a tudíž rovna jedné), $\lambda_1^2 = \ldots = \lambda_p^2 = 1$, tj. když matice $\mathbf{R}_{XX}$ je jednotková a tedy matice $\mathbf{X}$ má ortonormální sloupce. Vidíme tedy, že determinant výběrové korelační matice může být dobrým indikátorem přítomnosti multikolinearity v datech.

Předpokládejme na chvíli, že vysvětlující proměnné mají sdružené normální rozdělení. Připomeňme, že n značí počet pozorování a p značí počet vysvětlujících proměnných. Testujeme hypotézu, že populační korelační matice vysvětlujících proměnných je jednotková, tedy že všechny vysvětlující proměnné jsou stochasticky nezávislé. Jak uvádějí Farrar a Glauber v [5] nebo Hebák a Hustopecký v knize [10] (na straně 310), za platnosti nulové hypotézy platí, že statistika

$$\chi^2_{R_{XX}}(\nu) = -\left(n - 1 - \frac{1}{6}(2p+5)\right) \ln |\mathbf{R}_{XX}| \tag{3.5}$$

má přibližně rozdělení $\chi^2(\nu)$ s $\nu = \frac{1}{2}p(p-1)$ stupni volnosti, kde ν je počet prvků nad diagonálou, neboli počet korelačních koeficientů pro dvojice různých vysvětlujících proměnných. Pokud hodnota statistiky $\chi^2_{R_{XX}}(\nu)$ překročí $(1-\alpha)\,100$ procentní kvantil $\chi^2(\nu)$ rozdělení, považujeme multikolinearitu v našich datech za podstatnou na hladině významnosti $(1-\alpha)\,100\%$.

Nyní se podívejme, na které odhady regresních koeficientů vysvětlujících proměnných má multikolinearita velký vliv. Nechť nás zajímá (bez újmy na obecnosti) např. proměnná $\mathbf{x}_1$. Rozdělíme matici $\mathbf{X}$ na dvě části – na sloupec odpovídající proměnné $\mathbf{x}_1$ a na podmatici odpovídající ostatním vysvětlujícím proměnným, kterou označíme $\mathbf{X}_{-1}$. Potom výběrová korelační matice

bude mít tvar

$$\mathbf{R}_{XX} = \begin{pmatrix} 1 & \mathbf{r}'_{X_{-1}x_1} \\ \mathbf{r}_{X_{-1}x_1} & \mathbf{R}_{X_{-1}X_{-1}} \end{pmatrix}, \tag{3.6}$$

kde $\mathbf{R}_{X_{-1}X_{-1}}$ značí výběrovou korelační matici vysvětlujících proměnných bez proměnné $\mathbf{x}_1$ a $\mathbf{r}_{X_{-1}x_1}$ představuje vektor výběrových korelačních koeficientů mezi proměnnou $\mathbf{x}_1$ a proměnnými $\mathbf{x}_j$, $j = 2, \ldots p$. Označíme $r^{ij}, i, j = 1, \ldots p$, prvky inverzní matice $\mathbf{R}_{XX}^{-1}$ ke korelační matici. Pro diagonální prvek r^{11} inverzní matice k výběrové korelační matici (3.6) platí vztah (viz Anděl [2], str. 323)

$$r^{11} = (1 - \mathbf{r}'_{X_{-1}x_1}\mathbf{R}^{-1}_{X_{-1}X_{-1}}\mathbf{r}_{X_{-1}x_1})^{-1}. \tag{3.7}$$

Pokud $\mathbf{r}_{X_{-1}x_1} = \mathbf{0}$, tak $r^{11} = 1$. Prvek r^{11} můžeme přepsat jako

$$\begin{aligned} r^{11} - 1 &= \left(1 - \mathbf{r}'_{X_{-1}x_1}\mathbf{R}^{-1}_{X_{-1}X_{-1}}\mathbf{r}_{X_{-1}x_1}\right)^{-1} - 1 \\ &= \left(1 - r^2_{X_{-1}x_1}\right)^{-1} - 1 \\ &= \frac{r^2_{X_{-1}x_1}}{1 - r^2_{X_{-1}x_1}}, \end{aligned} \tag{3.8}$$

kde $r_{X_{-1}x_1}$ je výběrový koeficient mnohonásobné korelace mezi proměnnou $\mathbf{x}_1$ a ostatními vysvětlujícími proměnnými, jak je definován např. v knize Anděl [2] v kapitole 6.3 na straně 96.

Můžeme tedy přejít ke statistickému testu uvedenému v knize Anděl [2] věta 6.4 na straně 96 nebo v článku Farrara a Glaubera [5]. Opět předpokládejme, že vysvětlující proměnné mají sdružené normální rozdělení. Jestliže populační koeficient mnohonásobné korelace $\rho_{X_{-1}x_1} = 0$ a $n > p + 1$, má statistika

$$\omega_i = \frac{n - p - 1}{p} \frac{r^2_{X_{-i}x_i}}{1 - r^2_{X_{-i}x_i}} \tag{3.9}$$

(přesně) F rozdělení s p a $n - p - 1$ stupni volnosti. Tento test není nic jiného než test hypotézy o podmodelu a platí, že čtverec koeficientu mnohonásobné korelace $r^2_{X_{-i}x_i}$ je totéž co koeficient determinace R_i^2 v regresním modelu závislosti $\mathbf{x}_i$ na ostatních vysvětlujících proměnných. Definici R^2 můžeme najít např. v knize Zvára [24] v kapitole 3.5 na stranách 36-37, nebo v knize Anděl [2] na straně 83. Statistiku ω_i tedy můžeme přepsat jako

$$\omega_i = \frac{n - p - 1}{p} \frac{R_i^2}{1 - R_i^2}. \tag{3.10}$$

S tím již úzce souvisí další indikátor multikolinearity v datech – tzv. inflační faktor – označovaný zkratkou VIF (Variance Inflation Factor), který definu-

jeme jako

$$VIF_i = \frac{1}{1 - R_i^2}. \tag{3.11}$$

Rozptyl odhadu regresních parametrů podle (2.14) rozepíšeme jako

$$\begin{aligned} \mathrm{Var}\left(b_{OLS}\left(i\right)\right) &= \sigma^2 r^{ii} \\ &= \sigma^2 \frac{1}{1 - R_i^2} \\ &= \sigma^2 VIF_i \end{aligned} \tag{3.12}$$

a inflační faktor tedy můžeme interpretovat jako hodnotu, kolikrát se zhorší rozptyl odhadu regresního koeficientu pro proměnnou $\mathbf{x}_i$ v důsledku korelovanosti této vysvětlující proměnné s ostatními vysvětlujícími proměnnými. Jestliže jsou vysvětlující proměnné vzájemně lineárně nezávislé, je koeficient determinace R_i^2 ve vzorci (3.12) nulový a tedy rozptyl odhadu koeficientu $b_{OLS}\left(i\right)$ je σ^2. Pokud se koeficient determinace ve vzorci (3.12) blíží 1, zhoršuje nám vzájemný vztah mezi vysvětlovanými proměnnými rozptyl odhadu $b_{OLS}\left(i\right)$. Jak uvádí Hebák a Hustopecký [10] (str. 309), pokud je inflační faktor větší než 5 až 10, považujeme inflační číslo za vysoké a vliv multikolinearity na rozptyl odhadu regresních koeficientů za podstatný.

Zajímáme-li se o to, jaká je závislost mezi proměnnými $\mathbf{x}_i$ a $\mathbf{x}_j$ s vyloučením vlivu ostatních proměnných, budou nás zajímat výběrové koeficienty parciální korelace. Tyto dostaneme tak, že transformujeme prvky inverzní matice k výběrové korelační matici na výběrové koeficienty parciální korelace, jak uvádí Anděl [2] v kapitole 6.4 na straně 97, a výběrové koeficienty parciální korelace budou rovny

$$r_{ij\bullet} = \frac{-r^{ij}}{\sqrt{r^{ii}}\sqrt{r^{jj}}},\ i, j = 1, \ldots p. \tag{3.13}$$

Opět předpokládáme, že vysvětlující proměnné mají sdružené normální rozdělení. Jestliže platí, že populační koeficient parciální korelace $\rho_{ij\bullet} = 0$ a $n > p + 2$, má statistika

$$t_{ij\bullet} = \frac{r_{ij\bullet}}{\sqrt{1 - r_{ij\bullet}^2}}\sqrt{n - p - 2} \tag{3.14}$$

Studentovo rozdělení o $n - p - 2$ stupních volnosti.

Pro všechny uvedené statistické testy jsme předpokládali, že vysvětlující proměnné mají sdružené normální rozdělení. Farrar a Glauber v článku [5] doporučují se na tyto indikátory multikolinearity podívat i bez splnění tohoto předpokladu.

Jiný přístup k odhalení multikolinearity uvádí např. Zvára [24] v kapitole 11.2 na stranách 159-162 a Naes a Mevik v článku [14]. Vycházejí přitom z vlivu multikolinearity v datech na hodnoty vlastních čísel matice $\mathbf{X}'\mathbf{X}$. Definujeme indexy podmíněnosti matice $\mathbf{X}'\mathbf{X}$ jako poměry odmocniny největšího vlastního čísla matice $\mathbf{X}'\mathbf{X}$ ku odmocninám ostatních vlastních čísel

$$\eta_i = \frac{\lambda_1}{\lambda_i},\ 1 < i \leq p, \tag{3.15}$$

a číslo podmíněnosti η_p matice $\mathbf{X}'\mathbf{X}$ jako poměr odmocnin největšího a nejmenšího vlastního čísla matice $\mathbf{X}'\mathbf{X}$

$$\eta_p = \frac{\lambda_1}{\lambda_p}. \tag{3.16}$$

Pokud je číslo podmíněnosti matice $\mathbf{X}'\mathbf{X}$ velké, naznačuje to přítomnost multikolinearity v datech. V knize Hebák a Hustopecký [10] (příklad na str. 311) můžeme najít doporučení, že pokud poměr největšího a nejmenšího vlastního čísla přesáhne hodnotu 900, $\eta_p > \sqrt{900}$, pak vliv multikolinearity na kvalitu odhadu regresních koeficientů považujeme za vážný.

Kapitola 4

Metoda nejmenších čtverců

Popis normálního lineárního modelu a odvození odhadů parametrů pomocí metody nejmenších čtverců (Ordinary Least Squares) jsme si ukázali v kapitole 2, případně jej můžeme nalézt např. v knize Anděl [2] v kapitole 5 od strany 79 nebo v knize Zvára [24] od strany 20. Zde ještě doplníme odvození střední kvadratické chyby pro predikci střední hodnoty vysvětlované proměnné. Tuto predikci označíme $\hat{\mathbf{y}}$ s dolním indexem podle metody, kterou jsme tuto predikci vytvořili.

Střední kvadratickou chybu odhadu $\mathbf{T}$ parametru $\boldsymbol{\theta}$ si definujeme jako (např. v knize Hebák a Hustopecký [10] na str. 79 nebo Anděl [2] na str. 35)

$$\begin{aligned} MSE\left(\mathbf{T}\right) &= \mathrm{E}\left(\left(\mathbf{T}-\boldsymbol{\theta}\right)\left(\mathbf{T}-\boldsymbol{\theta}\right)'\right) \\ &= \mathrm{Var}(\mathbf{T}) + \left(\mathrm{E}\mathbf{T}-\boldsymbol{\theta}\right)\left(\mathrm{E}\mathbf{T}-\boldsymbol{\theta}\right)' \\ &= \mathrm{Var}(\mathbf{T}) + \left(\mathrm{bias}(\mathbf{T})\right)\left(\mathrm{bias}(\mathbf{T})\right)', \end{aligned} \tag{4.1}$$

kde první člen představuje rozptyl odhadu $\mathbf{T}$ a druhý člen kvadrát vychýlení odhadu $\mathbf{T}$.

Máme-li odhad $\mathbf{b}_{OLS}$ vektoru regresních parametrů $\boldsymbol{\beta}$ v modelu (2.5), potom pro predikci $\hat{\mathbf{y}}_{OLS}$ střední hodnoty vysvětlované proměnné platí vztah

$$\hat{\mathbf{y}}_{OLS} = \mathbf{X}\mathbf{b}_{OLS}. \tag{4.2}$$

Protože v modelu (2.5) je $\mathbf{b}_{OLS}$ nejlepší nestranný lineární odhad parametru $\boldsymbol{\beta}$, pak i $\hat{\mathbf{y}}_{OLS}$ je nejlepší nestranný lineární odhad střední hodnoty vysvětlované proměnné $\mathbf{X}\boldsymbol{\beta}$. Nestrannost odhadu znamená, že vychýlení odhadu je nulové, tj. $\mathrm{bias}\left(\hat{\mathbf{y}}_{OLS}\right) = 0$. Střední kvadratickou chybu predikce

$\hat{\mathbf{y}}_{OLS}$ můžeme pomocí rovnice (2.14) zapsat jako

$$\begin{aligned} MSE\left(\hat{\mathbf{y}}_{OLS}\right) &= \mathrm{Var}\left(\hat{\mathbf{y}}_{OLS}\right) + \left(\mathrm{bias}\left(\hat{\mathbf{y}}_{OLS}\right)\right)\left(\mathrm{bias}\left(\hat{\mathbf{y}}_{OLS}\right)\right)' \\ &= \mathrm{Var}\left(\hat{\mathbf{y}}_{OLS}\right) + 0 \\ &= \mathrm{Var}\left(\mathbf{X}\mathbf{b}_{OLS}\right) \\ &= \mathbf{X}\mathrm{Var}\left(\mathbf{b}_{OLS}\right)\mathbf{X}' \\ &= \sigma^2\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'. \end{aligned} \tag{4.3}$$

Použijeme značení zavedené pro spektrální rozklad matice $\mathbf{X}'\mathbf{X}$ v (2.23) a dostáváme

$$\begin{aligned} MSE\left(\hat{\mathbf{y}}_{OLS}\right) &= \sigma^2\mathbf{X}\left(\sum_{i=1}^{p}\frac{1}{\lambda_i^2}\mathbf{q}_i\mathbf{q}_i'\right)\mathbf{X}' \\ &= \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}\mathbf{X}\mathbf{q}_i\mathbf{q}_i'\mathbf{X}', \end{aligned} \tag{4.4}$$

z čehož můžeme snadno nahlédnout, že pokud jsou vlastní čísla matice $\mathbf{X}'\mathbf{X}$ malá, bude se střední kvadratická chyba velmi zvětšovat. Odpovězme ještě na otázku, co by se stalo, pokud by se malé vlastní číslo λ_i^2 setkalo s nulovým $\mathbf{X}\mathbf{q}_i$. Taková situace nemůže nastat, neboť z $\mathbf{X}\mathbf{q}_i = \mathbf{0}$ plyne, že $\mathbf{X}'\mathbf{X}\mathbf{q}_i = \mathbf{0}$, a z vlastností vlastních čísel a vlastních vektorů matice $\mathbf{X}'\mathbf{X}$ by potom platilo

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{q}_i &= \lambda_i^2\mathbf{q}_i \\ \lambda_i^2\mathbf{q}_i &= \mathbf{0} \\ \lambda_i^2\mathbf{q}_i'\mathbf{q}_i &= 0, \end{aligned} \tag{4.5}$$

čímž dostáváme spor, neboť platí $\mathbf{q}_i'\mathbf{q}_i = 1$ a všechna vlastní čísla λ_i^2 jsou kladná. Nebo jinak – pokud platí $\mathbf{X}\mathbf{q}_i = \mathbf{0}$, tak to neznamená nic jiného, než že matice $\mathbf{X}$ má lineárně závislé sloupce, což je ale spor s předpokladem, že matice $\mathbf{X}$ má plnou sloupcovou hodnost.

V případě, že některá vlastní čísla matice $\mathbf{X}'\mathbf{X}$ jsou velmi malá, může být z hlediska střední kvadratické chyby výhodnější použít vychýlený odhad parametru $\boldsymbol{\beta}$ a právě těmto možnostem se budeme věnovat v následujících kapitolách. Střední kvadratická chyba nám v kapitole 8 pomůže vyhodnotit, která metoda bude nejvhodnější pro naše data. Rozhodovat se budeme právě na základě odhadnutých středních kvadratických chyb pro metody regrese na hlavních komponentách, parciálních nejmenších čtverců a hřebenové regrese a budeme tyto odhady střední kvadratické chyby porovnávat s odhadem střední kvadratické chyby odhadu metodou nejmenších čtverců. Porovnávat odhadnuté střední kvadratické chyby dvou odhadů (tj. porovnávat dvě kvadratické

formy) budeme následovně – pokud je matice rozdílu pozitivně definitní

$$MSE\left(\hat{\mathbf{y}}_{OLS}\right) - MSE\left(\hat{\mathbf{y}}\right) > 0, \tag{4.6}$$

pak řekneme, že odhad $\hat{\mathbf{y}}$ má menší střední kvadratickou chybu než odhad metodou nejmenších čtverců a vybereme si jej jako odhad z tohoto pohledu výhodnější.

Kapitola 5

Regrese na hlavních komponentách

Metodou hlavních komponent (Principal Component Analysis) hledáme k nových regresorů $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k$, $k \leq p$, které vystihují podstatnou část variability vysvětlujících proměnných (informace z vysvětlujících proměnných) a které budou vzájemně ortogonální. Zpravidla chceme zredukovat počet nových regresorů na méně než byl počet původních vysvětlujících proměnných p a zároveň zachytit co největší množství informace z původních vysvětlujících proměnných. Nové regresory (v metodě hlavních komponent nazývané hlavní komponenty) následně použijeme v lineárním modelu pro predikci vysvětlované proměnné. V závěru kapitoly si popíšeme metody pro volbu vhodného počtu hlavních komponent.

Cílem metody hlavních komponent je nalézt nové regresory jako lineární kombinace původních vysvětlujících proměnných, které jsou vzájemně ortogonální a vysvětlují maximum celkového rozptylu původních vysvětlujících proměnných. Získáme je postupně tak, že první hlavní komponenta vystihuje co nejvíce celkové variability vysvětlujících proměnných. Druhá komponenta bude ortogonální na první komponentu a vystihuje maximum zbývající variabity vysvětlujících proměnných (část variability vysvětlujících proměnných, která nebyla vysvětlena první hlavní komponentou), atd., vždy každá další komponenta je ortogonální s předchozími komponentami a vystihuje maximální množství variability neobsažené v předchozích komponentách. Hledáme proto vektory koeficientů $\mathbf{c}_i, \mathbf{c}_i'\mathbf{c}_i = 1$, lineárních kombinací sloupců matice $\mathbf{X}$ takové, aby pro hlavní komponenty $\mathbf{t}_i = \mathbf{X}\mathbf{c}_i$ platila ortogonalita

$$\mathbf{t}_i'\mathbf{t}_j = 0, \ i < j, \tag{5.1}$$

a zároveň aby kvadrát euklidovské normy hlavní komponenty $\mathbf{t}_i$

$$\|\mathbf{t}_i\|_2^2 = \mathbf{c}_i'\mathbf{X}'\mathbf{X}\mathbf{c}_i$$

byl maximální. Podmínku na ortogonalitu nových regresorů (5.1) můžeme rozepsat jako

$$\begin{aligned} \mathbf{c}_i'\mathbf{X}'\mathbf{X}\mathbf{c}_j &= 0, \ i < j \\ \mathbf{c}_i'\mathbf{R}_{XX}\mathbf{c}_j &= 0, \ i < j. \end{aligned} \tag{5.2}$$

Použijeme-li značení zavedené v rovnici (2.23) (spektrální rozklad matice $\mathbf{X}'\mathbf{X}$), pak z extremálních vlastností vlastních čísel vyplývá (viz Hebák a Hustopecký [10], str. 374), že optimální volba je $\mathbf{c}_i = \mathbf{q}_i$. Tedy optimální první hlavní komponenta $\mathbf{t}_i$ je získána jako lineární kombinace matice vysvětlujících proměnných $\mathbf{X}$ a prvního vlastního vektoru matice $\mathbf{X}'\mathbf{X}$, tedy $\mathbf{q}_1$ ($\mathbf{q}_1$ značí vlastní vektor matice $\mathbf{X}'\mathbf{X}$ odpovídající největšímu vlastnímu číslu λ_1^2) a zapíšeme ji jako

$$\mathbf{t}_1 = \mathbf{X}\mathbf{q}_1. \tag{5.3}$$

Analogicky druhá hlavní komponenta bude $\mathbf{t}_2 = \mathbf{X}\mathbf{q}_2$, atd. Tak transformujeme matici $\mathbf{X}$ téměř lineárně závislých vysvětlujících proměnných na matici ortogonálních hlavních komponent $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_p)$ $(n \times p)$, která má ve sloupcích jednotlivé komponenty $\mathbf{t}_i$ s vlastnostmi popsanými výše, a maticově ji zapíšeme jako

$$\mathbf{T} = \mathbf{X}\mathbf{Q}, \tag{5.4}$$

kde $\mathbf{Q}$ je ortonormální transformační matice, jejíž sloupce tvoří vlastní vektory matice $\mathbf{X}'\mathbf{X}$. Matice $\mathbf{Q}$ se při této příležitosti někdy nazývá matice zátěží (z anglického Loadings Matrix).

Využijeme spektrální rozklad matice $\mathbf{X}'\mathbf{X}$ a z rovnice (2.22) snadno dostaneme

$$\begin{aligned} \|\mathbf{T}\|_2^2 &= \mathbf{T}'\mathbf{T} \\ &= \mathbf{Q}'\mathbf{X}'\mathbf{X}\mathbf{Q} \\ &= \mathbf{Q}'\mathbf{Q}\boldsymbol{\Lambda}^2\mathbf{Q}'\mathbf{Q} \\ &= \boldsymbol{\Lambda}^2. \end{aligned} \tag{5.5}$$

Jinými slovy, euklidovská norma na druhou (což je pro nenáhodný centrovaný vektor analogie rozptylu) hlavní komponenty $\mathbf{t}_i$ je rovna vlastnímu číslu λ_i^2.

Dokud předpokládáme $k = p$, můžeme lineární model (2.5) přepsat jako

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{T}\mathbf{Q}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{T}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}. \end{aligned} \tag{5.6}$$

Odhad parametru $\boldsymbol{\beta}^*$ metodou nejmenších čtverců v modelu (5.6) označíme $\mathbf{b}_{PCR}$ a získáme jej jako

$$\mathbf{b}_{PCR} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y} \tag{5.7}$$

a platí pro něj

$$\begin{aligned}\mathbf{b}_{PCR} &= (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}\\ &= (\mathbf{Q}'\mathbf{X}'\mathbf{X}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X}'\mathbf{y}\\ &= \mathbf{Q}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Q}'^{-1}\mathbf{Q}'\mathbf{X}'\mathbf{y}\\ &= \mathbf{Q}'\mathbf{b}_{OLS}.\end{aligned} \tag{5.8}$$

Odhad střední hodnoty vysvětlované proměnné pomocí regrese na hlavních komponentách označíme $\hat{\mathbf{y}}_{PCR}$ a spočteme jej jako

$$\hat{\mathbf{y}}_{PCR} = \mathbf{T}\mathbf{b}_{PCR}, \tag{5.9}$$

což v případě, že matice $\mathbf{T}$ má ve sloupcích p hlavních komponent, není nic jiného než odhad střední hodnoty vysvětlované proměnné metodou nejmenších čtverců

$$\begin{aligned}\hat{\mathbf{y}}_{PCR} &= \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}\\ &= \mathbf{X}\mathbf{Q}(\mathbf{Q}'\mathbf{X}'\mathbf{X}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X}'\mathbf{y}\\ &= \mathbf{X}\mathbf{Q}\mathbf{Q}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Q}'^{-1}\mathbf{Q}'\mathbf{X}'\mathbf{y}\\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\ &= \hat{\mathbf{y}}_{OLS}.\end{aligned} \tag{5.10}$$

Analogicky pro rozptyl tohoto odhadu dostáváme

$$\begin{aligned}\mathrm{Var}(\hat{\mathbf{y}}_{PCR}) &= \sigma^2\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\\ &= \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}\mathbf{X}\mathbf{q}_i\mathbf{q}_i'\mathbf{X}'\\ &= \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}\mathbf{t}_i\mathbf{t}_i'\end{aligned} \tag{5.11}$$

a platí, že $MSE(\hat{\mathbf{y}}_{PCR}) = \mathrm{Var}(\hat{\mathbf{y}}_{PCR}) = \mathrm{Var}(\hat{\mathbf{y}}_{OLS}) = MSE(\hat{\mathbf{y}}_{OLS})$. Vidíme tedy, že o využití regrese na hlavních komponentách má smysl uvažovat jedině pro počet hlavních komponent k menší než počet původních vysvětlujících proměnných, $k < p$.

Označíme $\mathbf{Q} = (\mathbf{Q}_k, \mathbf{Q}_{p-k})$, kde matice $\mathbf{Q}_k = (\mathbf{q}_1, \ldots, \mathbf{q}_k)$ $(p \times k)$ má ve sloupcích prvních k vlastních vektorů matice $\mathbf{X}'\mathbf{X}$ odpovídajících k největším vlastním číslům $\lambda_1^2, \ldots, \lambda_k^2$ a matice $\mathbf{Q}_{p-k} = (\mathbf{q}_{k+1}, \ldots, \mathbf{q}_p)$ $(p \times (p-k))$ má ve sloupcích vlastní vektory $\mathbf{q}_{k+1}, \ldots, \mathbf{q}_p$ matice $\mathbf{X}'\mathbf{X}$ odpovídající vlastním číslům $\lambda_{k+1}^2, \ldots, \lambda_p^2$. Podobně rozdělíme matici $\mathbf{T} = (\mathbf{T}_k, \mathbf{T}_{p-k})$ na matici $\mathbf{T}_k = (\mathbf{t}_1, \ldots, \mathbf{t}_k)$ $(n \times k)$, která má ve sloupcích prvních k hlavních komponent, a na matici $\mathbf{T}_{p-k} = (\mathbf{t}_{k+1}, \ldots, \mathbf{t}_p)$ $(n \times (p-k))$, která má ve sloupcích zbylých $p-k$ hlavních komponent.

Odhad střední hodnoty vysvětlované proměnné pomocí regrese na prvních k hlavních komponentách získáme jako

$$\hat{\mathbf{y}}_{PCR_k} = \mathbf{T}_k \left(\mathbf{T}_k'\mathbf{T}_k\right)^{-1} \mathbf{T}_k'\mathbf{y} \tag{5.12}$$

a jeho rozptyl si podle (5.11) snadno odvodíme jako

$$\mathrm{Var}\left(\hat{\mathbf{y}}_{PCR_k}\right) = \sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i^2} \mathbf{X}\mathbf{q}_i\mathbf{q}_i'\mathbf{X}'. \tag{5.13}$$

Vychýlení tohoto odhadu spočteme podle věty 7.1 o vychýlení odhadů, platí-li širší model, v knize Zvára [24] na stranách 82 až 83 nebo jednoduše podle (4.1), (5.8) a z nestrannosti odhadu střední hodnoty $\mathbf{y}$ metodou nejmenších čtverců dostáváme

$$\begin{aligned}
\mathrm{bias}\left(\hat{\mathbf{y}}_{PCR_k}\right) &= \mathrm{E}\left(\mathbf{T}_k\mathbf{b}_{PCR_k}\right) - \mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{T}_k\mathrm{E}\left(\mathbf{b}_{PCR_k}\right) - \mathbf{X}\boldsymbol{\beta} \qquad (5.14) \\
&= \mathbf{T}_k\mathrm{E}\left(\mathbf{Q}_k'\mathbf{b}_{OLS}\right) - \mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{T}_k\mathbf{Q}_k'\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{X}\mathbf{Q}_k\mathbf{Q}_k'\boldsymbol{\beta} - \mathbf{X}\mathbf{Q}\mathbf{Q}'\boldsymbol{\beta} \\
&= -\sum_{i=k+1}^{p} \mathbf{X}\mathbf{q}_i\mathbf{q}_i'\boldsymbol{\beta}. \qquad (5.15)
\end{aligned}$$

Použít regresi na hlavních komponentách místo odhadu metodou nejmenších čtverců se nám vyplatí pouze v případě, že matice rozdílu

$$MSE\left(\hat{\mathbf{y}}_{OLS}\right) - MSE\left(\hat{\mathbf{y}}_{PCR}\right) > 0 \tag{5.16}$$

bude pozitivně definitní matice. To nastane v případě, že matice

$$\sigma^2 \sum_{i=k+1}^{p} \frac{1}{\lambda_i^2} \mathbf{X}\mathbf{q}_i\mathbf{q}_i'\mathbf{X}' - \left(\sum_{i=k+1}^{p} \mathbf{X}\mathbf{q}_i\mathbf{q}_i'\boldsymbol{\beta}\right)\left(\sum_{i=k+1}^{p} \mathbf{X}\mathbf{q}_i\mathbf{q}_i'\boldsymbol{\beta}\right)' \tag{5.17}$$

bude pozitivně definitní.

Volba vhodného počtu nových regresorů k se v praxi provádí pomocí křížového ověření. Popis metody a doporučení, jakým způsobem vybírat počet regresorů v modelu regrese na hlavních komponentách a regrese pomocí parciálních nejmenších čtverců můžeme nalézt v článku [13]. V tomto článku Mevik a Cederkvist považují za nejvhodnější metodu „Leave One Out", případně metodu desetisložkového křížového ověření. V kapitole 8 věnované praktickým příkladům zpracovaným v programu R proto budeme používat metodu křížového ověření „Leave One Out". Data si rozdělíme na dvě části – jedno pozorování i ponecháme stranou a všechna ostatní pozorování použijeme pro odhad parametrů modelů s různým počtem hlavních komponent. Potom spočteme předpovědi modelů pro jedno vynechané pozorování i a spočteme chybu předpovědi pro toto pozorování. Vše provedeme postupně pro jednotlivá pozorování a podle střední kvadratické chyby předpovědí (pro vynechaná pozorování) jednotlivých modelů vybereme model s nejmenší střední kvadratickou chybou.

Kapitola 6

Metoda parciálních nejmenších čtverců

Přestože je multikolinearita vlastnost vysvětlujících proměnných, následující metoda se s ní bude vyrovnávat i s přihlédnutím k hodnotám vysvětlované proměnné. Postupná konstrukce hlavních komponent je velmi blízká hledání nových regresorů v následující metodě – při odhadech regresních parametrů pomocí parciálních nejmenších čtverců (Partial Least Squares Regression). U regrese na hlavních komponentách jsme hledali koeficienty lineární kombinace $\mathbf{c}_i, \mathbf{c}_i'\mathbf{c}_i = 1$, sloupců matice $\mathbf{X}$ takové, aby pro nové ortogonální regresory $\mathbf{t}_i = \mathbf{X}\mathbf{c}_i$ platilo, že vystihují maximum variability vysvětlujících proměnných. To nám ovšem nezaručí, že nové prediktory budou vhodné i pro predikci střední hodnoty vysvětlované proměnné. U metody parciálních nejmenších čtverců proto nové regresory hledáme s přihlédnutím k hodnotám vysvětlované proměnné a doufáme, že tak získáme lepší predikci střední hodnoty vysvětlované proměnné (lepší ve smyslu menší střední kvadratické chyby).

Cílem metody parciálních nejmenších čtverců je nalézt nové regresory $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k$, $k \leq p$, jako lineární kombinace původních vysvětlujících proměnných (sloupců matice $\mathbf{X}$)

$$\mathbf{t}_i = \mathbf{X}\mathbf{c}_i, \tag{6.1}$$

které jsou vzájemně ortogonální (stejně jako (5.1))

$$\mathbf{t}_i'\mathbf{t}_j = 0, \; i < j, \tag{6.2}$$

a jsou užitečné pro predikci střední hodnoty vysvětlované proměnné. Postupujeme tak, jak je uvedeno v článku Garthwaite [8] (v kapitole 2 na stranách 123 až 124) a ještě lépe vysvětleno v knize Rao, Toutenburg, Shalabh, Heumann [16] (v kapitole 3.14.4 na stranách 84-87).

Nové regresory vytváříme postupně, nejprve se podívejme na $\mathbf{t}_1$. Vycházíme z myšlenky, že z každé vysvětlující proměnné chceme získat maximální možnou informaci pro predikci střední hodnoty vysvětlované proměnné. První nový regresor získáme jako

$$\mathbf{t}_1 = \sum_{j=1}^{p} w_{1j} b_{OLS_{1j}} \mathbf{x}_j, \tag{6.3}$$

kde w_{1i} jsou váhy (váhy mohou být stanoveny různě podle zvoleného druhu metody parciálních nejmenších čtverců a nejčastěji používané váhy si rozebereme níže), $\mathbf{x}_j$ je j-tá vysvětlující proměnná (j-tý sloupec matice $\mathbf{X}$) a $b_{OLS_{1j}}$ je regresní koeficient z lineární regrese $\mathbf{y}$ na jediné vysvětlující proměnné $\mathbf{x}_j$. Jinými slovy: nový regresor jsme získali jako váženou lineární kombinaci p odhadů střední hodnoty vysvětlované proměnné z modelů lineární regrese $\mathbf{y}$ na jediné vysvětlující proměnné $\mathbf{x}_j$.

Druhý nový regresor získáme tak, že využijeme odhadu zbývající informace z matice $\mathbf{X}$ a vektoru $\mathbf{y}$ po odečtení informace vysvětlené prvním novým regresorem $\mathbf{t}_1$. Do druhého kroku vstupujeme s proměnnými $\mathbf{x}_{2j}$, které se rovnají vektoru reziduí z regrese $\mathbf{x}_j$ na $\mathbf{t}_1$. Namísto vysvětlované proměnné $\mathbf{y}$ do druhého kroku vstupuje vektor reziduí z regrese $\mathbf{y}$ na $\mathbf{t}_1$, který označíme $\mathbf{y}_2$. Dále postupujeme podobně jako pro první nový regresor – regresí $\mathbf{y}_2$ na $\mathbf{x}_{2j}$ odhadneme $b_{OLS_{2j}}$ a druhý regresor se bude rovnat

$$\mathbf{t}_2 = \sum_{j=1}^{p} w_{2j} b_{OLS_{2j}} \mathbf{x}_{2j}. \tag{6.4}$$

Další nové regresory sestrojíme analogickým postupem. Namísto vektoru $\mathbf{y}_i$ použijeme vektor $\mathbf{y}_{i+1}$, který je roven reziduím z regrese $\mathbf{y}_i$ na $\mathbf{t}_i$ a jako nové vysvětlující proměnné $\mathbf{x}_{i+1j}, j = 1, \ldots p$, použijeme vektory reziduí z regrese $\mathbf{x}_{ij}$ na $\mathbf{t}_i$.

Vzhledem k tomu, že vektor reziduí je v klasickém lineárním modelu nekorelovaný s prediktory (např. věta 2.2 o reziduích v knize Zvára [24] na straně 14), je splněna podmínka ortogonality nových regresorů.

Nejčastější volba vah je $w_{ij} = \frac{1}{p}$ nebo jako násobek převrácené hodnoty rozptylu regresního koeficientu $\frac{1}{\mathrm{Var} b_{OLS_{ij}}}$ (viz Garthwaite [8]).

Při odhadu pomocí parciálních nejmenších čtverců obvykle potřebujeme menší počet regresorů k, než potřebujeme u regrese na hlavních komponentách. Vyplývá to ze způsobu, jakým byly tyto regresory sestrojeny. Hlavní komponenty byly konstruovány nezávisle na vysvětlované proměnné a tedy jich obvykle potřebujeme větší nebo stejný počet pro vyjádření vztahu s vysvětlovanou proměnnou než je tomu u umělých regresorů získaných metodou

parciálních nejmenších čtverců. Optimální počet nových regresorů dostaneme pomocí křížového ověření, jak bylo popsáno v závěru kapitoly 5.

Na závěr uvedeme literaturu rozšiřující uvedený text. V článku [1] Abdi popisuje situaci, kdy máme více než jednu vysvětlovanou proměnnou. V článku [9] Goutis ilustruje geometrii parciálních nejmenších čtverců a uvádí geometrický důkaz, že odhady parametru β pomocí parciálních nejmenších čtverců mají menší střední kvadratickou chybu než odhady pomocí nejmenších čtverců v případě, že skutečné hodnoty parametru jsou malé. V článku [4] Butler a Denham odhalují podstatu zmenšení rozptylu odhadu pomocí parciálních nejmenších čtverců a zaobírají se situací, kdy odhady pomocí parciálních nejmenších čtverců fungují špatně.

Kapitola 7

Hřebenová regrese

Hřebenová regrese (Ridge Regression) je metoda odvozená od metody nejmenších čtverců. Od regrese na hlavních komponentách a regrese metodou parciálních nejmenších čtverců se liší tím, že nespočívá na principu transformace vysvětlujících proměnných na nové proměnné. Je založená na jednoduché myšlence, že pokud máme problémy s hledáním inverzní matice k matici $\mathbf{X}'\mathbf{X}$, zkusíme nepatrně změnit její diagonální prvky (přičteme kladnou hodnotu δ ke každému diagonálnímu prvku matice $\mathbf{X}'\mathbf{X}$) a výpočet inverzní matice pak bude numericky stabilnější.

Odhad parametrů v modelu hřebenové regrese můžeme vyjádřit jako

$$\mathbf{b}_{RR} = (\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p)^{-1}\,\mathbf{X}'\mathbf{y}, \tag{7.1}$$

kde $\delta > 0$ je parametr hřebenové regrese a $\mathbf{I}_p$ označuje jednotkovou matici rozměrů $p \times p$. Ze vzorce (7.1) snadno můžeme nahlédnout, že pokud by δ bylo nulové, $\delta = 0$, odhad $\mathbf{b}_{RR}$ bude roven odhadu pomocí metody nejmenších čtverců. Pokud $\delta \to \infty$, pak pro odhad $\mathbf{b}_{RR}$ platí $\mathbf{b}_{RR} \to 0$.

Vyjádříme vzájemný vztah odhadu metodou nejmenších čtverců (2.12) a odhadu pomocí hřebenové regrese (7.1) a dostáváme

$$\begin{aligned}\mathbf{b}_{RR} &= (\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p)^{-1}\,\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p)^{-1}\,\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p)^{-1}\,\mathbf{X}'\mathbf{X}\mathbf{b}_{OLS}.\end{aligned} \tag{7.2}$$

Odhad parametrů v modelu hřebenové regrese je vychýlený. Jeho střední kvadratickou chybu odvodíme podle vzorce (4.1) jako

$$MSE\,(\mathbf{b}_{RR}) = \mathrm{Var}\,(\mathbf{b}_{RR}) + \mathrm{bias}\,(\mathbf{b}_{RR})\,\mathrm{bias}\,(\mathbf{b}_{RR})', \tag{7.3}$$

kde jednotlivé členy jsou – rozptyl

$$\begin{aligned}\mathrm{Var}\left(\mathbf{b}_{RR}\right) &= \mathrm{Var}\left(\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{y}\right)\\ &= \left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathrm{Var}\left(\mathbf{y}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1} \qquad (7.4)\end{aligned}$$

a vychýlení

$$\begin{aligned}\mathrm{bias}\left(\mathbf{b}_{RR}\right) &= \mathrm{E}\left(\mathbf{b}_{RR}\right)-\boldsymbol{\beta}\\ &= \left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathrm{E}\mathbf{y}-\boldsymbol{\beta}\\ &= \left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}\\ &= \left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\left(\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-\delta\mathbf{I}_p\boldsymbol{\beta}\right)\\ &= \left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\left(-\delta\mathbf{I}_p\boldsymbol{\beta}\right)\\ &= -\delta\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\boldsymbol{\beta}. \qquad (7.5)\end{aligned}$$

Po dosazení do vzorce (7.3) dostáváme

$$\begin{aligned}MSE\left(\mathbf{b}_{RR}\right) &= \sigma^2\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}+\\ &\quad+\delta^2\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\\ &= \left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\left(\sigma^2\mathbf{X}'\mathbf{X}+\delta^2\boldsymbol{\beta}\boldsymbol{\beta}'\right)\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}. \qquad (7.6)\end{aligned}$$

Porovnáme odhad metodou nejmenších čtverců (2.12) s odhadem pomocí hřebenové regrese (7.1). Pokud bude matice rozdílu

$$MSE\left(\mathbf{b}_{OLS}\right)-MSE\left(\mathbf{b}_{RR}\right)>0 \qquad (7.7)$$

pozitivně definitní, bude výhodnější využít odhad metodou hřebenové regrese. V opačném případě je z hlediska střední kvadratické chyby výhodnější použít odhad metodou nejmenších čtverců. Výraz (7.7) rozepíšeme podle (2.14) a využijeme fakt, že odhad $\mathbf{b}_{OLS}$ je nestranný (vychýlení odhadu $\mathrm{bias}\left(\mathbf{b}_{OLS}\right)=0$, tedy $MSE\left(\mathbf{b}_{OLS}\right)=\mathrm{Var}\left(\mathbf{b}_{OLS}\right)$ a dále s využitím přepisu $\mathrm{Var}\left(\mathbf{b}_{OLS}\right)$ na

$$\begin{aligned}&\mathrm{Var}\left(\mathbf{b}_{OLS}\right)\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\left(\mathbf{I}_p+\delta\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1}\left(\mathbf{X}'\mathbf{X}+2\delta\mathbf{I}_p+\delta^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\left(\mathbf{X}'\mathbf{X}+\delta\mathbf{I}_p\right)^{-1} \qquad (7.8)\end{aligned}$$

dostáváme

$$\begin{aligned}
&MSE\left(\mathbf{b}_{OLS}\right) - MSE\left(\mathbf{b}_{RR}\right) \\
&= \mathrm{Var}\left(\mathbf{b}_{OLS}\right) - \left(\mathrm{Var}\left(\mathbf{b}_{RR}\right) + \mathrm{bias}\left(\mathbf{b}_{RR}\right)\mathrm{bias}\left(\mathbf{b}_{RR}\right)'\right) \\
&= \left(\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p\right)^{-1}\left(\sigma^2\mathbf{X}'\mathbf{X} + 2\delta\sigma^2\mathbf{I}_p + \delta^2\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)\left(\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p\right)^{-1} - \\
&\quad - \left(\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p\right)^{-1}\left(\sigma^2\mathbf{X}'\mathbf{X} + \delta^2\boldsymbol{\beta}\boldsymbol{\beta}'\right)\left(\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p\right)^{-1}\left(2\delta\sigma^2\mathbf{I}_p + \delta^2\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} - \delta^2\boldsymbol{\beta}\boldsymbol{\beta}'\right)\left(\mathbf{X}'\mathbf{X} + \delta\mathbf{I}_p\right)^{-1} \qquad (7.9)
\end{aligned}$$

Výraz (7.9) je pozitivně definitní, pokud

$$2\delta\sigma^2\mathbf{I}_p + \delta^2\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} - \delta^2\boldsymbol{\beta}\boldsymbol{\beta}' \qquad (7.10)$$

je pozitivně definitní. Dospěli jsme tedy k tomu, co uvádí Theobald [22] ve větě 2 na straně 105 – (7.10) je pozitivně definitní, pokud

$$2\sigma^2\mathbf{I}_p - \delta\boldsymbol{\beta}\boldsymbol{\beta}' \qquad (7.11)$$

je pozitivně definitní (neboť $\delta^2\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ je pozitivně definitní).

Odhad $\hat{\mathbf{y}}_{RR}$ v lineárním modelu hřebenové regrese můžeme zapsat jako

$$\hat{\mathbf{y}}_{RR} = \mathbf{X}\mathbf{b}_{RR} \qquad (7.12)$$

a použijeme jej v případě, že rozdíl matic (7.7) bude pozitivně definitní po dosazení odhadů za neznámé parametry. Střední čtvercovou chybu odhadu $\hat{\mathbf{y}}_{RR}$ odhadneme z (7.12) a (7.6) jednoduše jako

$$MSE\left(\hat{\mathbf{y}}_{RR}\right) = \mathbf{X}MSE\left(\mathbf{b}_{RR}\right)\mathbf{X}'. \qquad (7.13)$$

Existuje několik způsobů, jak zvolit hodnotu parametru δ. V knihovně „MASS“ programu R máme v současné době k dispozici tři metody – „modified HKB estimator“, „modified L-W estimator“ a „smallest value of GCV“. Podrobnosti o těchto metodách lze najít v dokumentaci k programu R a hodnoty spočtené těmito metodami se mohou vzájemně značně lišit. Navíc je zde uvažovaná jiná parametrizace matice $\dot{\mathbf{X}}$, a to její z-skóry s hodnotou n namísto $n-1$ ve vzorci (2.1) a s hodnotou n na diagonále. Odhady parametru δ v naší parametrizaci dostaneme jako $1/n$-násobek odhadů výše. Proto je lepší brát tyto hodnoty pouze jako orientační a hodnotu parametru δ zvolit podle grafu tzv. hřebenových stop (Ridge traces), což je graf hodnot odhadů jednotlivých regresních parametrů $b_{RR}\left(i\right), i = 1, \ldots p$, v závislosti na hodnotách parametru δ. Snažíme se podle grafu zvolit takové nejmenší δ, pro které jsou odhady parametru $\boldsymbol{\beta}$ stabilní.

Základní informace o hřebenové regresi jsme čerpali z knih Zvára [24] kapitola 11.5 na stranách 166-169 a Zvára [25] kapitola 9.3 na stranách 139-145, přednášek Stuetzle [19] a článku Vinod [23] na stranách 121-124. V přednáškách Stuetzle [19] můžeme nalézt popis optimalizační úlohy s omezeními a řešení Lagrangeovy rovnice pro hřebenovou regresi. V článku Vinod [23] je uvedena Bayesovská interpretace hřebenové regrese, diskuse k použití standardizace dat a další možná rozšíření – zobecněná hřebenová regrese (Generalized Ridge Regression), iterativní zobecněná hřebenová regrese (Iterative Generalized Ridge Regression) a Steinův-Rulův odhad (Stein-Rule Estimator). Vysvětlení geometrie hřebenových stop (Ridge Traces) můžeme nalézt v článku Swindel [21] na stranách 13-14.

Kapitola 8

Zpracování dat

Metody vysvětlené v předchozích kapitolách si ilustrujeme na příkladu. K dispozici máme data obsažená v knize Fazekas a Kósa [6] – jedná se o naměřené délky a šířky kostí lidských plodů, pohlaví, délku a stáří plodu (v lunárních měsících). Data byla pořízena pro potřeby soudního lékařství a závěry z nich jsou využívány např. během antropologického rozboru kosterních pozůstatků objevených při archeologických výzkumech.

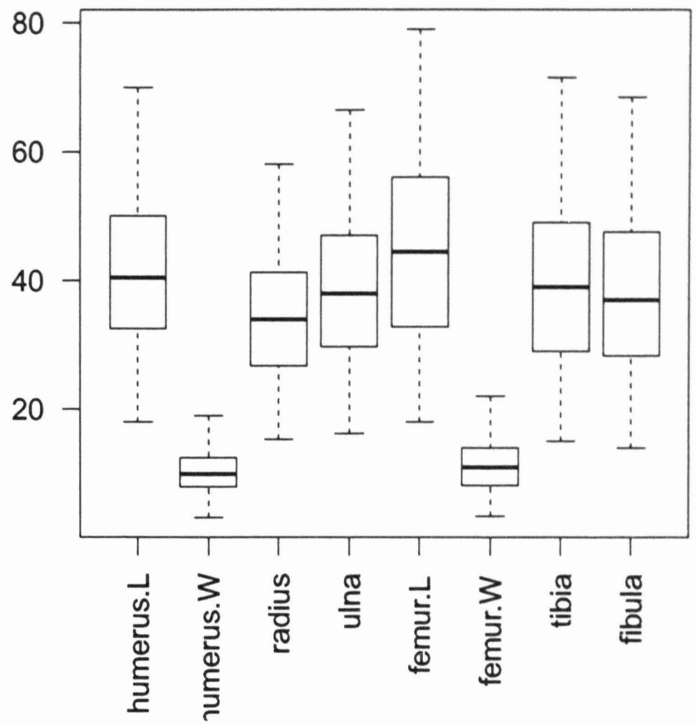
Z původního souboru jsme vybrali údaje o plodech starších než čtyři měsíce a pouze ty proměnné, jejichž hodnoty může naměřit např. archeolog při terénním výzkumu. Hledáme model pro odhad stáří plodu na základě informací o délkách a šířkách vybraných kostí.

8.1 Popisné statistiky

Náš soubor obsahuje následující údaje naměřené na 133 lidských plodech:

lunar	věk plodu v lunárních měsících
humerus.L	délka pažní kosti v mm
humerus.W	šířka pažní kosti v mm
radius	délka vřetenní kosti v mm
ulna	délka loketní kosti v mm
femur.L	délka stehenní kosti v mm
femur.W	šířka stehenní kosti v mm
tibia	délka holenní kosti v mm
fibula	délka lýtkové kosti v mm
skupina	rozdělení na trénovací a testovací skupinu.

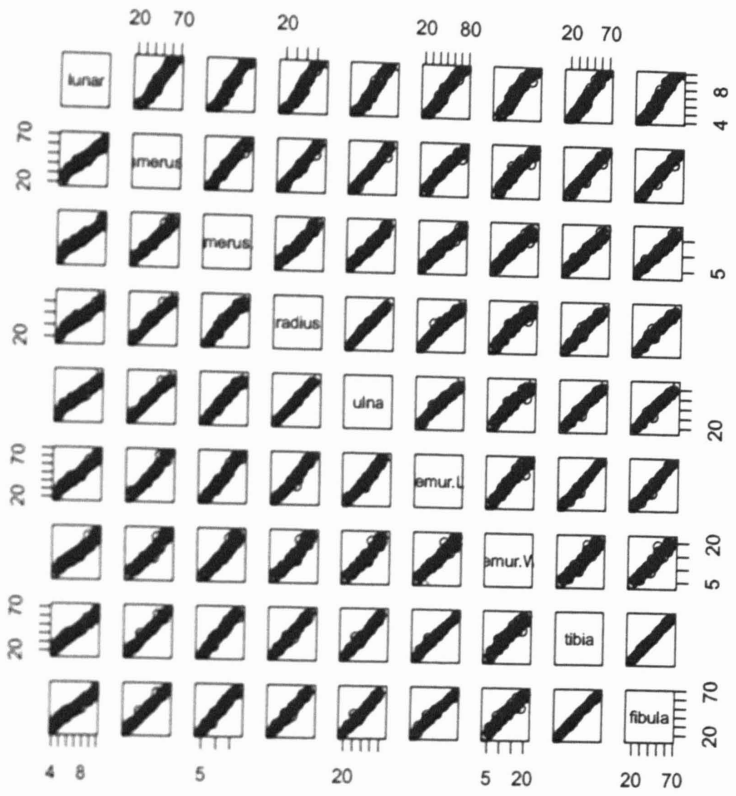
Prohlédněme si naše data. Podíváme se na popisné statistiky jednotlivých proměnných, prohlédneme si krabicové grafy vysvětlujících proměnných a bodové grafy dvojic proměnných, spočteme výběrové korelační koeficienty.



Obrázek 8.1: Krabicové grafy délek a šířek vybraných kostí v mm.

```
> dim(Fazekas)
[1] 133  10
> summary(Fazekas)
     lunar          humerus.L       humerus.W         radius
 Min.   : 4.000   Min.   :18.00   Min.   : 3.20   Min.   :15.30
 1st Qu.: 5.000   1st Qu.:32.50   1st Qu.: 8.00   1st Qu.:26.70
 Median : 6.500   Median :40.50   Median :10.00   Median :34.00
 Mean   : 6.677   Mean   :41.35   Mean   :10.36   Mean   :34.02
 3rd Qu.: 8.000   3rd Qu.:50.00   3rd Qu.:12.50   3rd Qu.:41.20
 Max.   :10.000   Max.   :70.00   Max.   :19.00   Max.   :58.00
      ulna          femur.L         femur.W          tibia
 Min.   :16.20   Min.   :18.00   Min.   : 3.40   Min.   :15.00
 1st Qu.:29.70   1st Qu.:32.80   1st Qu.: 8.20   1st Qu.:29.00
 Median :38.00   Median :44.50   Median :11.00   Median :39.00
 Mean   :38.30   Mean   :44.76   Mean   :11.37   Mean   :39.63
 3rd Qu.:47.00   3rd Qu.:56.00   3rd Qu.:14.00   3rd Qu.:49.00
 Max.   :66.50   Max.   :79.00   Max.   :22.00   Max.   :71.50
     fibula           skupina
 Min.   :14.00   testovaci: 27
 1st Qu.:28.30   trenovaci:106
 Median :37.00
 Mean   :37.99
 3rd Qu.:47.50
 Max.   :68.50
```

Podle obrázku 8.2 soudíme, že jednotlivé vysvětlující proměnné zřejmě budou mít mezi sebou téměř lineární vztah. Pokud zkoumáme výběrové ko-



Obrázek 8.2: Graf závislostí dvojic proměnných.

relační koeficienty v tabulce korelací, vidíme, že jednotlivé regresory jsou mezi sebou silně korelovány. Proto se naše data jeví jako vhodný soubor pro vysvětlení problému multikolinearity.

```
> cor(Fazekas[,-10])
             lunar humerus.L humerus.W    radius      ulna   femur.L   femur.W
lunar     1.0000000 0.9821156 0.9790271 0.9805686 0.9814967 0.9837665 0.9739751
humerus.L 0.9821156 1.0000000 0.9839989 0.9903806 0.9925928 0.9891235 0.9766389
humerus.W 0.9790271 0.9839989 1.0000000 0.9821703 0.9849222 0.9841657 0.9755190
radius    0.9805686 0.9903806 0.9821703 1.0000000 0.9961953 0.9856142 0.9731483
ulna      0.9814967 0.9925928 0.9849222 0.9961953 1.0000000 0.9899855 0.9781465
femur.L   0.9837665 0.9891235 0.9841657 0.9856142 0.9899855 1.0000000 0.9816976
femur.W   0.9739751 0.9766389 0.9755190 0.9731483 0.9781465 0.9816976 1.0000000
tibia     0.9815953 0.9902995 0.9850348 0.9891289 0.9911987 0.9945395 0.9777393
fibula    0.9810271 0.9900833 0.9856983 0.9898545 0.9925745 0.9928285 0.9778311
              tibia    fibula
lunar     0.9815953 0.9810271
humerus.L 0.9902995 0.9900833
humerus.W 0.9850348 0.9856983
radius    0.9891289 0.9898545
ulna      0.9911987 0.9925745
femur.L   0.9945395 0.9928285
femur.W   0.9777393 0.9778311
tibia     1.0000000 0.9979991
fibula    0.9979991 1.0000000
```

Jednotlivé regresory standardizujeme a naměřené hodnoty nezávisle pro-

měnné centrujeme a dále postupujeme tak, jak jsme uvedli v kapitole 3.

8.2 Diagnostika multikolinearity

Testujeme nulovou hypotézu, že populační korelační matice je jednotková, tedy že vysvětlující proměnné jsou nezávislé.

```
> print(D <- det(R_XX))
[1] 7.497584e-14
> print(chi_det_R_XX<-((-1)*(n-1-(2*p+5)/6)*log(D)))
[1] 3067.493
> qchisq(1-alpha, df=(0.5*p*(p-1)))
[1] 41.33714
```

Hodnota determinantu matice $\mathbf{R}_{XX}$ je blízká nule a proto nás nepřekvapí, že na hladině spolehlivosti 95% zamítáme nulovou hypotézu, že populační korelační matice je jednotková. Tedy zamítáme hypotézu, že vysvětlující proměnné jsou nezávislé.

Nejmenší vlastní číslo matice $\mathbf{R}_{XX}$ je blízké nule a číslo podmíněnosti matice $\mathbf{X}'\mathbf{X}$ je větší než 30 (doporučení Hebáka a Hustopeckého [10] v příkladu na straně 311) a tedy i indexy podmíněnosti indikují přítomnost multikolinearity v našich datech.

```
> Lambda2 <- eigen(R_XX)$values
> round(Lambda2,4)
[1] 7.8945 0.0374 0.0233 0.0206 0.0108 0.0081 0.0039 0.0016

> print(indexy_podminenosti <- sqrt(Lambda2[1])/sqrt(Lambda2))
[1]  1.00000 14.53784 18.40751 19.58408 27.09025 31.28304 45.23550 70.60542
> print(cislo_podminenosti <- indexy_podminenosti[p])
[1] 70.60542
```

Podobný závěr dostaneme, podíváme-li se na inflační faktory VIF. Např. rozptyl odhadu regresního koeficientu pro proměnnou tibia je 336-krát větší než by byl bez vzájemného vztahu mezi vysvětlovanými proměnnými.

```
> vif(lm(y~X))
Xhumerus.L Xhumerus.W    Xradius      Xulna   Xfemur.L   Xfemur.W     Xtibia
  79.05448   40.50683  134.73598  206.08277  115.04144   29.01471  336.27286
   Xfibula
 285.28476
```

Žádný diagnostický prostředek (ať uvážíme graf 8.2, výběrovou korelační matici nebo test hypotézy, že populační korelační matice je jednotková) neukazuje, že by proměnné byly vzájemně nezávislé, a naopak podle indexů podmíněnosti, čísla podmíněnosti matice $\mathbf{X}'\mathbf{X}$ a podle inflačních faktorů soudíme, že multikolinearita je v našich datech přítomná.

8.3 Metoda nejmenších čtverců

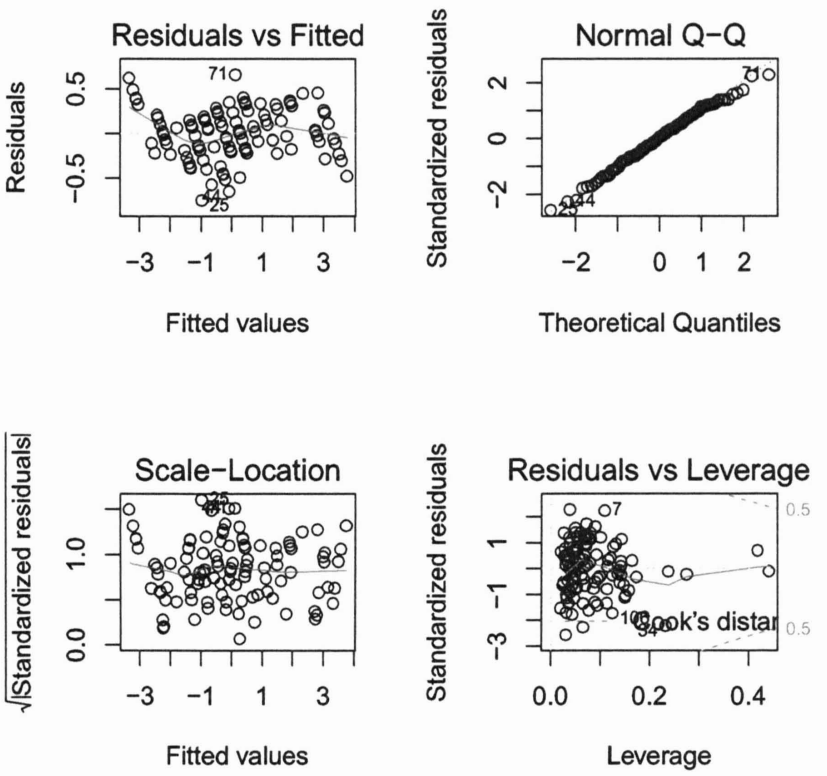
V předchozí sekci jsme identifikovali přítomnost multikolinearity v našich datech. Pojďme se nyní podívat, jaký bude její vliv na odhady metodou nejmenších čtverců.

Absolutní člen ve standardizovaném modelu bude nulový, my jej ale ponecháme, abychom zachovali správný počet stupňů volnosti v F testu užitečnosti modelu pro predikci střední hodnoty proměnné **y** a t-testech jednotlivých proměnných. Koeficient determinace je 0,97 a tedy jsme 97% variability **y** vysvětlili pomocí závislosti na vysvětlujících proměnných.

```
> summary(OLS8 <- lm(y~X))

Call:
lm(formula = y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74480 -0.18763  0.01419  0.20886  0.65745

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.044e-16  2.857e-02 7.15e-15   1.0000
Xhumerus.L   3.672e-01  2.553e-01    1.439   0.1535
Xhumerus.W   4.338e-01  1.827e-01    2.374   0.0196 *
Xradius      5.162e-01  3.333e-01    1.549   0.1246
Xulna       -2.509e-01  4.121e-01   -0.609   0.5441
Xfemur.L     6.924e-01  3.079e-01    2.248   0.0268 *
Xfemur.W     2.931e-01  1.546e-01    1.896   0.0610 .
Xtibia      -2.104e-01  5.265e-01   -0.400   0.6903
Xfibula     -4.829e-02  4.849e-01   -0.100   0.9209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2942 on 97 degrees of freedom
Multiple R-squared: 0.9754,Adjusted R-squared: 0.9734
F-statistic: 480.5 on 8 and 97 DF,  p-value: < 2.2e-16
```

Pro ověření normality reziduí z modelu OLS8 (regresní model s plným počtem vysvětlujících proměnných) použijeme Shapirův-Wilkův test normality z knihovny „stats“. Nezamítáme nulovou hypotézu, že rozdělení reziduí je normální, což odpovídá histogramu reziduí i diagramu normality v obrázku 8.4.

```
> shapiro.test(rstandard(OLS8))

Shapiro-Wilk normality test

data:  rstandard(OLS8)
```



Obrázek 8.3: Grafy reziduí v regresním modelu obsahujícím všechny vysvětlující proměnné.

```
W = 0.9927, p-value = 0.8426
```

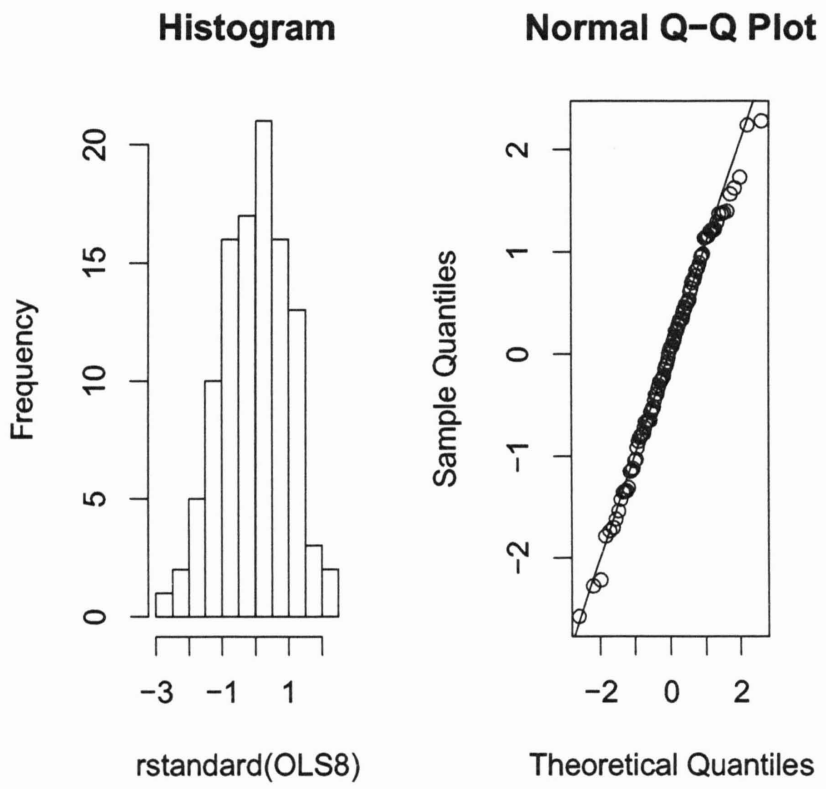
Střední kvadratická chyba predikce odhadu metodou nejmenších čtverců je 0,096504250.

```
> predikce_y_OLS8 <- bar_y_test + X_test %*% OLS8$coef[2:9]
> print(MSEP_OLS8 <- 1/n_test * sum((predikce_y_OLS8 - y_test_dot)^2))
[1] 0.09650425
```

8.4 Regrese na hlavních komponentách

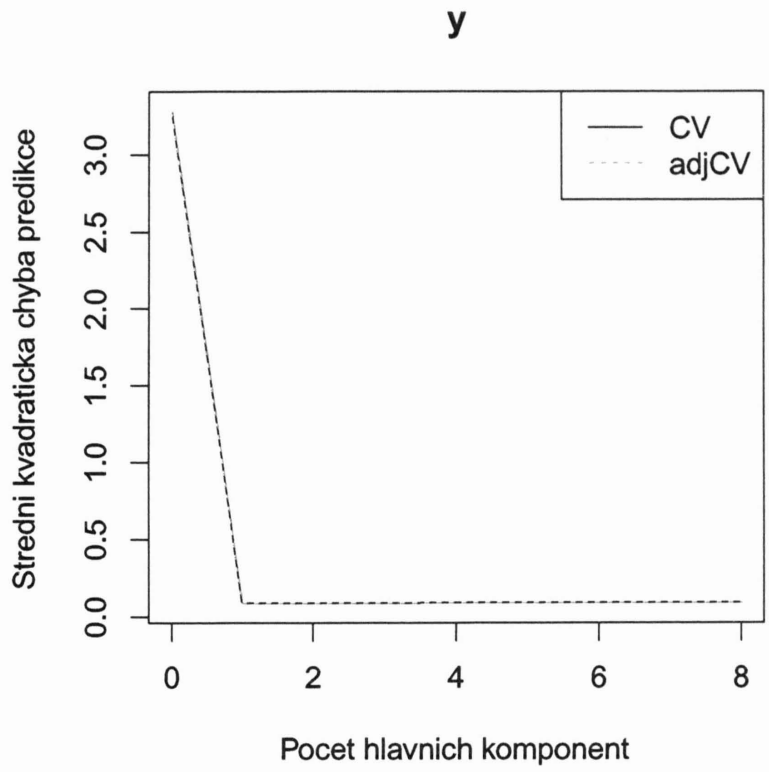
Pro metodu regrese na hlavních komponentách jsme se rozhodli využít knihovny „pls“ v programu R. Podívejme se nejprve na regresi na všech osmi hlavních komponentách, i když již podle vlastních čísel matice $\mathbf{R}_{XX}$ uvedených v sekci 8.2 tušíme, že pro vysvětlení variability $\mathbf{y}$ postačovat bude menší počet.

```
> T<-X%*%Q
> summary(pcr(y~T, validation="LOO"))
Data:  X dimension: 106 8
Y dimension: 106 1
Fit method: svdpc
Number of components considered: 8
```



Obrázek 8.4: Histogram a normální diagram reziduí v regresním modelu obsahujícím všechny vysvětlující proměnné.

```
VALIDATION: RMSEP
Cross-validated using 106 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           1.811   0.2993   0.2984   0.3003   0.3019   0.3016   0.3015
adjCV        1.811   0.2993   0.2984   0.3003   0.3018   0.3016   0.3014
       7 comps  8 comps
CV      0.3037   0.3056
adjCV   0.3036   0.3055

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X    98.68    99.15    99.44    99.70    99.83    99.93    99.98   100.00
y    97.32    97.37    97.37    97.42    97.44    97.50    97.53    97.54
> summary(PCR8<-lm(y~T))

Call:
lm(formula = y ~ T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74480 -0.18763  0.01419  0.20886  0.65745

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
```



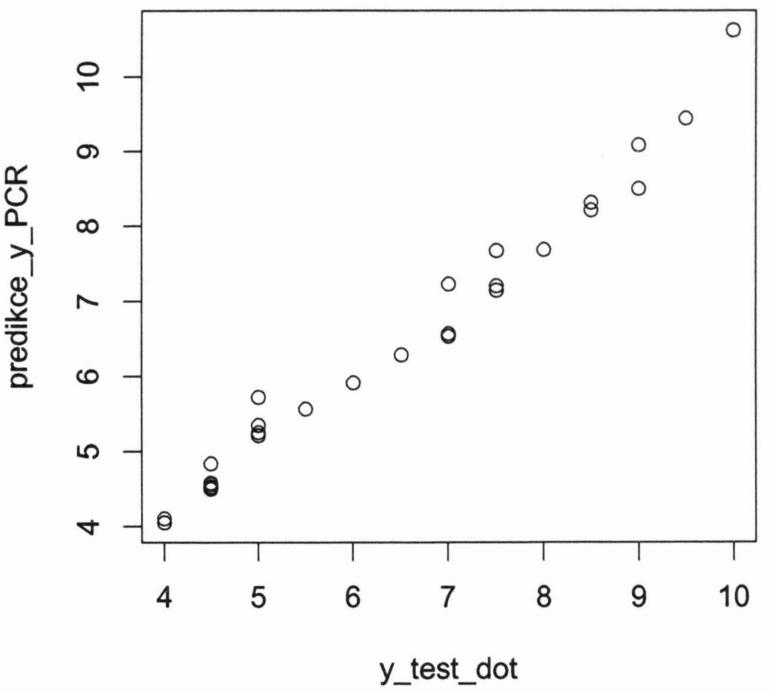
Obrázek 8.5: Výběr počtu hlavních komponent podle MSEP.

```
(Intercept)  2.137e-16  2.857e-02 7.48e-15    1.000
T1          -6.328e-01  1.022e-02  -61.931   <2e-16 ***
T2           2.142e-01  1.485e-01    1.442    0.153
T3           9.872e-03  1.881e-01    0.052    0.958
T4          -2.602e-01  2.001e-01   -1.300    0.197
T5           2.826e-01  2.768e-01    1.021    0.310
T6           4.599e-01  3.197e-01    1.439    0.153
T7          -5.300e-01  4.622e-01   -1.147    0.254
T8           4.219e-01  7.215e-01    0.585    0.560
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2942 on 97 degrees of freedom
Multiple R-squared: 0.9754,Adjusted R-squared: 0.9734
F-statistic: 480.5 on 8 and 97 DF,  p-value: < 2.2e-16
```

Jak vidíme i z obrázku 8.5, výběr počtu komponent je jasný – vybereme právě jednu (první) hlavní komponentu. Vzhledem k tomu, že vysvětlující proměnné jsou spolu silně korelovány, je tato volba podle očekávání.

Příslušný vlastní vektor $\mathbf{q}_1$ nám ukazuje, kolik která původní proměnná přispívá k vybrané hlavní komponentě $\mathbf{t}_1$. Každá proměnná přispívá přibližně stejně, můžeme tedy první hlavní komponentu interpretovat přibližně jako záporně vzatý normovaný součet všech regresorů.

```
> print(Q[,1])
```



Obrázek 8.6: Predikce vybraného modelu regrese na první hlavní komponentě.

```
[1] -0.3539810 -0.3524621 -0.3536314 -0.3545480 -0.3542573 -0.3504713 -0.3544997
[8] -0.3545561

> summary(PCR1<-lm(y~T[,1]))

Call:
lm(formula = y ~ T[, 1])

Residuals:
     Min       1Q   Median       3Q      Max
-0.77282 -0.20017  0.01787  0.17524  0.66553

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.350e-16  2.879e-02 8.16e-15        1
T[, 1]      -6.328e-01  1.029e-02   -61.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2964 on 104 degrees of freedom
Multiple R-squared: 0.9732,Adjusted R-squared: 0.973
F-statistic:  3779 on 1 and 104 DF,  p-value: < 2.2e-16

> predikce_y_PCR <- bar_y_test + X_test %*% Q[,1] %*% PCR1$coef[2]
> print(MSEP_PCR <- 1/n_test * sum((predikce_y_PCR-y_test_dot)^2))
[1] 0.0940853
```

Střední kvadratická chyba predikce pro testovací skupinu bude 0,0940853.

8.5 Metoda parciálních nejmenších čtverců

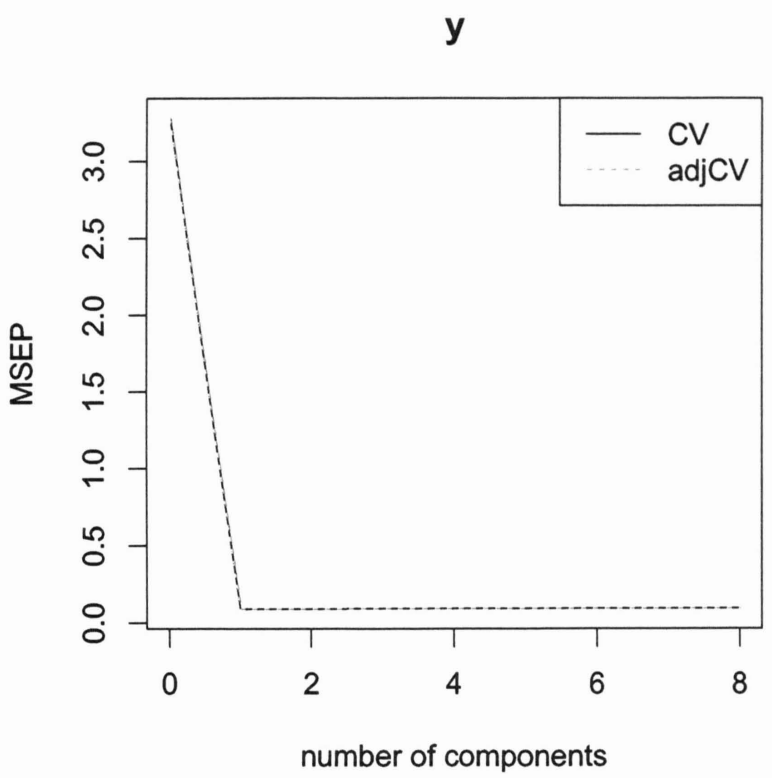
Využijeme knihovny „pls“ z programu R a podíváme se, jaký počet nových regresorů zvolit.

```
> summary(PLSR8<-plsr(y~X, validation="LOO"))
Data:  X dimension: 106 8
Y dimension: 106 1
Fit method: kernelpls
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 106 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           1.811   0.2993   0.2996   0.3030   0.3033   0.3052   0.3045
adjCV        1.811   0.2993   0.2996   0.3029   0.3032   0.3051   0.3044
       7 comps  8 comps
CV      0.3054   0.3056
adjCV   0.3053   0.3055

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X    98.68    99.08    99.32    99.50    99.58    99.70    99.93   100.00
y    97.32    97.46    97.51    97.53    97.54    97.54    97.54    97.54
```

Podobně jako u metody regrese na hlavních komponentách podle obrázku 8.7 vybereme právě jeden nový regresor t_1. Dále postupujeme, jak bylo popsáno v kapitole 6.

Střední kvadratická chyba predikce bude 0,09411276.

```
> # postup podle knihy Rao, Toutenburg, Shalabh, Heumann
> b_1 <- rep(0, p)
> for (i in 1:p) {
+  b_1[i]<-lm(y~X[,i])$coefficients[2]
+  }
> t_1 <- 1/p* X %*% b_1
> summary(lm(y~t_1))

Call:
lm(formula = y ~ t_1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77279 -0.20024  0.01815  0.17509  0.66560

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 2.610e-16  2.878e-02 9.07e-15        1
```



Obrázek 8.7: Výběr počtu hlavních komponent podle MSEP.

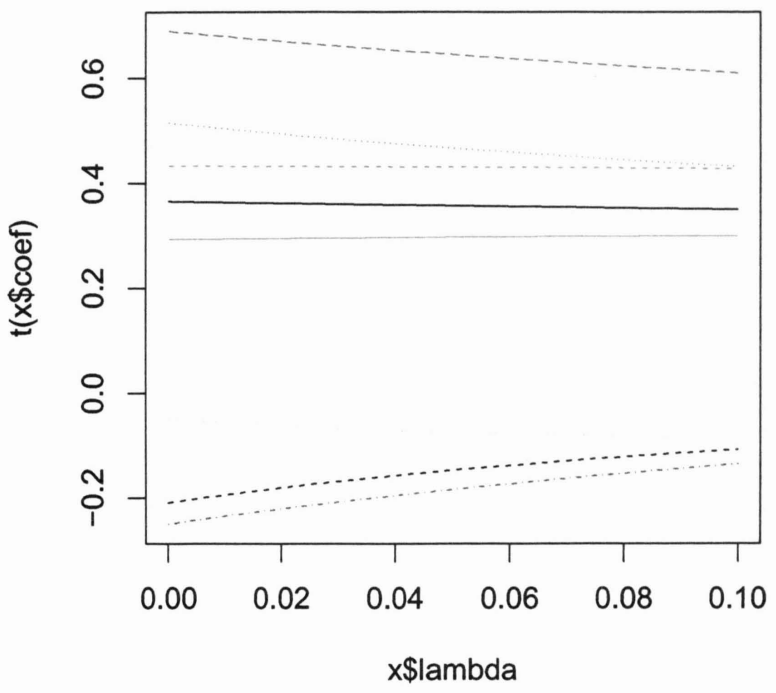
```
t_1          1.013e+00  1.648e-02    61.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2963 on 104 degrees of freedom
Multiple R-squared: 0.9732,Adjusted R-squared: 0.973
F-statistic:  3781 on 1 and 104 DF,  p-value: < 2.2e-16

> predikce_y_PLSR <- bar_y_test + lm(y~t_1)$coef[2] * 1/p * X_test %*% b_1
> print(MSEP_PLSR <- 1/n_test * sum((predikce_y_PLSR-y_test_dot)^2))
[1] 0.09411276
```

8.6 Hřebenová regrese

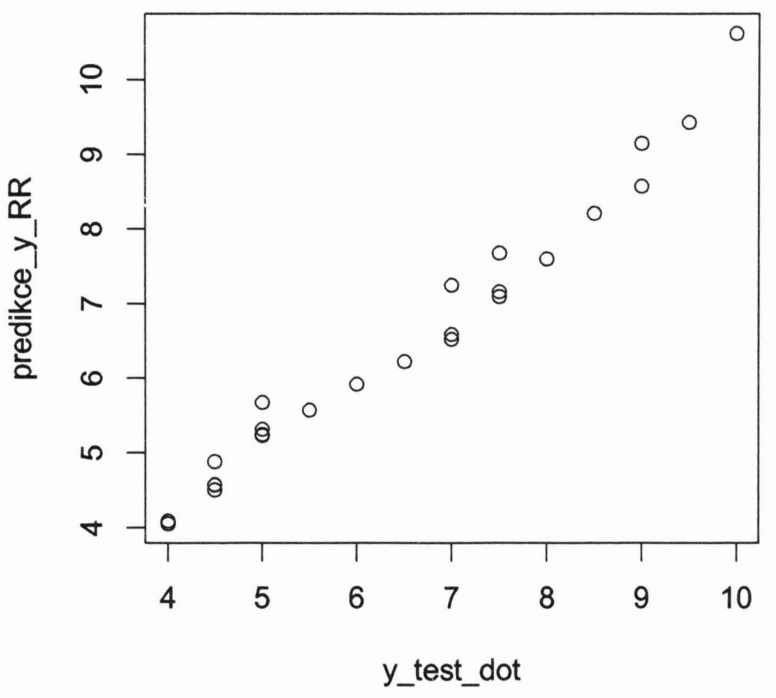
V této sekci nejprve vybereme vhodnou hodnotu parametru δ v rovnici (7.1) podle obrázku hřebenových stop 8.8 a hodnot „modified HKB estimator“ a „modified L-W estimator“.

```
> select(lm.ridge(y ~ X, lambda = seq(0,0.1,0.0001)))
modified HKB estimator is 0.4146032
modified L-W estimator is 0.1654517
smallest value of GCV  at 0.1
```

Rozhodli jsme se vybrat podle „modified HKB estimator“ a odhad parametru δ v naší parametrizaci dostaneme jako $1/n$-násobek odhadu výše,



Obrázek 8.8: Graf hřebenových stop vhodný pro výběr parametru δ.



Obrázek 8.9: Predikce vybraného modelu hřebenové regrese.

$\delta = 0.4146032/n$, kde n je počet pozorování v trénovací skupině. Odhadneme parametry trénovacího modelu a podíváme se na hodnoty vybraného modelu pro testovací skupinu dat. Střední kvadratická chyba predikce bude 0,09655111 a podle obrázku 8.9 soudíme, že model dobře vystihuje naše data.

```
> b_RR<- solve(t(X)%*% X+ 0.4146032/n *diag(rep(1, p))) %*% t(X) %*% y
> predikce_y_RR <- bar_y_test + X_test %*% b_RR
> print(MSEP_RR <- 1/n_test * sum((predikce_y_RR-y_test_dot)^2))
[1] 0.09655111
```

8.7 Srovnání metod

Ve srovnání metod podle střední kvadratické chyby jsou minimální rozdíly díky tomu, že proměnné velmi dobře vysvětlují závislou proměnnou a koeficient determinace v modelu regrese pomocí nejmenších čtverců je téměř roven jedné. Přesto můžeme konstatovat, že jako nejlepší model se nám jeví model regrese na jedné hlavní komponentě, těsně následován modelem parciálních nejmenších čtverců s jedním regresorem.

Metoda	Střední kvadratická chyba predikce
OLS	0,09650425
PCR	0,09408530
PLSR	0,09411276
RR	0,09655111

Jako model nejvhodnější pro naše data jsme tedy vybrali model regrese na jedné hlavní komponentě.

Kapitola 9

Závěr

V naší práci jsme se zabývali problémem multikolinearity v klasickém lineárním regresním modelu – její diagnostikou a metodami, které si s multikolinearitou v našich datech poradí. Použití všech popsaných metod jsme si v kapitole 8 podrobně ukázali na příkladu zpracování lékařských dat v programu R.

Regrese na hlavních komponentách, metoda parciálních nejmenších čtverců a hřebenová regrese, kterými jsme se zabývali, jsou implementovány nejen v programu R, ale i v komerčních programech Statistica, Statgraphics, SPSS, NCSS (mimo PLSR) aj. Prezentované výpočty a přiložené zdrojové kódy na CD byly zpracovány v programu R (verzi 2.9.2 pro Linux) zejména kvůli jeho snadné dostupnosti (viz [15]) a čtenář si tudíž může všechny postupy sám zopakovat nebo upravit.

Multikolinearita je jistě široké téma, které může zahrnovat mnoho rozšíření nad rámec naší práce – zejména další metody, které se s jejími důsledky vyrovnávají, např. Lasso Regression, Steinův-Rulův odhad (Stein-Rule Estimator), případně jiné modifikace zařazených metod jako jsou zobecněná hřebenová regrese (Generalized Ridge Regression), iterativní zobecněná hřebenová regrese (Iterative Generalized Ridge Regression), metoda parciálních nejmenších čtverců pro více než jednu vysvětlovanou proměnnou a další metody.

Přínos naší práce spočívá v podrobném sepsání metod pro diagnostiku multikolinearity v klasickém lineárním modelu, v přehledném představení hlavních principů metod navržených pro regresi se silně korelovanými regresory a v praktickém ukázání jednotlivých metod na příkladu.

Literatura

[1] Abdi H. (2003): Partial least squares regression (PLS-regression). In M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. 792-795.

[2] Anděl J. (2002): Základy matematické statistiky. Univerzita Karlova v Praze, Praha, preprint.

[3] Bennett K. P., Embrechts M. J. (2003): An Optimization Perspective on Kernel Partial Least Squares Regression. Verze ze dne 10.4.2007, http://www.rpi.edu/~bennek/papers/KB-ME-PLS.pdf.

[4] Butler N. A., Denham M. C. (2000): The Peculiar Shrinkage Properties of Partial Least Squares Regression. J. R. Statist. Soc. B (2000) 62, Part 3, 585-593.

[5] Farrar D. E., Glauber R. R. (1967): Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics (February 1967), Vol. 49, No. 1, 92-107.

[6] Fazekas I. G., Kósa F. (1978): Forensic Fetal Osteology. Akadémiai Kiadó, Budapest.

[7] Frank I. E., Friedman J. H. (1993): A Statistical View of Some Chemometrics Regression Tools. Technometrics (May 1993), Vol. 35, No. 2.

[8] Garthwaite P. H. (1994): An Interpretation of Partial Least Squares. Journal of the American Statistical Association (March 1994), Vol. 89, No. 425.

[9] Goutis C. (1996): Partial Least Squares Algorithm Yields Shrinkage Estimators. The Annal of Statistics 1996, Vol. 24, No. 2, 816-824.

[10] Hebák P., Hustopecký J. (1987): Vícerozměrné statistické metody s aplikacemi. SNTL, Praha.

[11] Kidwell J.S., Brown L.H. (1982): Ridge Regression as a Technique for Analyzing Models with Multicollinearity. Journal of Marriage and the Family (May 1982), Vol. 44, No. 2, 287-299.

[12] Longley, J. W. (1967): An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. Journal of the American Statistical Association (September 1967), Vol. 62, No. 319, 819-841.

[13] Mevik B.-H., Cederkvist H. R. (2005): Mean Squared Error of Prediction (MSEP) Estimates for Principal Component Regression (PCR) and Partial Least Squares (PLSR). Preprint - verze ze dne 20.10.2006, http://mevik.net/work/publications/MSEP_estimates.pdf.

[14] Naes T., Mevik B.-H. (2001): Understanding the Collinearity Problem in Regression and Discriminant Analysis. Journal of Chemometrics (2001) 4, Vol. 15, 413-426. Preprint je možné stáhnout na adrese http://mevik.net/work/publications/understanding_collinearity.pdf.

[15] R Foundation: R. Dokumentace k programu R a samotný program ke stáhnutí, http://www.r-project.org/.

[16] Rao C.R., Toutenburg H., Shalabh, Heumann C. (2008): Linear Models and Generalizations: Least Squares and Alternatives. Springer-Verlag New York, LLC, 3rd Edition.

[17] Silvey S. D. (1969): Multicollinearity and Imprecise Estimation. Journal of the Royal Statistical Society, Series B, Vol. 31, No.3, 539-552.

[18] Stone M., Brooks R. J. (1990): Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. J. R. Statist. Soc. B (1990) 52, No. 2, 237-269.

[19] Stuetzle W. (2005): Notes on Ridge Regression. BioStat 538, Winter 2005, verze ze dne 20.3.2007, www.stat.washington.edu/wxs/Stat538-w05/Notes/ridge-regression-1-20-05.pdf.

[20] Sundberg R. (2002): Continuum Regression. Article for 2nd. ed. of Encyclopedia of Statistical Sciences, May 2002, verze ze dne 12.5.2006, http://www.math.su.se/~rolfs/Publikationer/CR-resrep-2202-4.pdf.

[21] Swindel B. F. (1981): Geometry of Ridge Regression Illustrated. The American Statistician, February 1981, Vol. 35, No. 1.

[22] Theobald C. M. (1974): Generalizations of Mean Square Error Applied to Ridge Regression. Journal of the Royal Statistical Society, Series B, Vol. 36, No. 1, 103-106

[23] Vinod H. K. (1978): A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares. The Review of Economics and Statistics, February 1978, Vol. 60, No. 1, 121-131.

[24] Zvára K. (2008): Regrese. Matfyzpress, Praha 2008.

[25] Zvára K. (1989): Regresní analýza. Academia, Praha 1989.