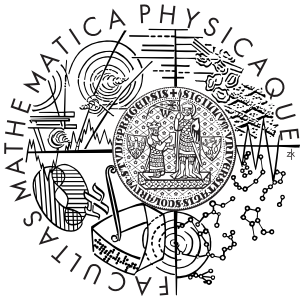
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Barbora Zuzáková

Mnohorozměrná lineární regrese

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Peter Bubelíny
Studijní program: Matematika, obecná matematika

2010

Na tomto místě bych ráda poděkovala všem, kteří mě jakkoliv podpořili při psaní této bakalářské práce. Zejména děkuji svému vedoucímu Mgr. Petrovi Bubelínymu za výběr zajímavého tématu, četné konzultace a jeho cenné rady.

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

V Praze dne 7.5.2010 Barbora Zuzáková

Obsah

Název práce: Mnohorozměrná lineární regerse
Autor: Barbora Zuzáková
Katedra: Katedra pravděpodobnosti a metamatické statistiky
Vedoucí bakalářské práce: Mgr. Bubelíny Peter, Katedra pravděpodobnosti a metamatické statistiky
e-mail vedoucího: bubeliny@seznam.cz

Abstrakt: Předložená bakalářská práce se zabývá problematikou mnohorozměrné lineární regrese. Jsou zde vyloženy základní charakteristiky a vlastnosti tohoto modelu. V první kapitole lze nalézt stručné shrnutí jednorozměrné lineární regrese, jejíchž poznatků se využívá i v teorii mnohorozměrné. Tato práce se podrobněji zabývá odhadováním regresních parametrů a testováním hypotéz. V sekci mnohorozměrné lineární regrese je tématu testování hypotéz věnována speciální pozornost. Jsou zde rozebrány nejpoužívanější testové statistiky, které jsou porovnány pomocí jednoduchých simulací.

Klíčová slova: mnohorozměrná lineární regrese, regresní parametr, metoda nejmenších čtverců, metoda maximální věrohodnosti, testování hypotézy.

Title: Multivariate Linear Regresion
Author: Barbora Zuzáková
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Peter Bubelíny, Department of Probability and Mathematical Statistics
Supervisor's e-mail address: bubeliny@seznam.cz

Abstract: This bachelor thesis explores the topic of multivariate linear regression. The basic characteristics and properties of this method are set out in the work. The first section provides a brief summary of the univariate linear regression because it is necessary for understanding the multivariate theory. Furthermore the work deals in detail with regression estimation and testing hypotheses. The topic of testing hypothesis is given special attention in the section on the multivariate linear regression. Most frequently used test statistics are also discussed and their usage is compared in several simple simulations.

Keywords: Multivariate linear regression, regression parameter, least squares estimation, maximum likelihood estimation, testing hypothesis.

Úvod

Model mnohorozměrné lineární regrese je důležitou statistickou metodou ke zkoumání závislosti proměnných, jejichž hodnoty získáváme při realizaci experimentů. Své uplatněné regrese nachází například v lékařství, meteorologii nebo i ve finančním sektoru. Tato bakalářská práce je shrnutím teorie mnohorozměrné lineární regrese. Obsahuje základní vzorce potřebné k určení regresních parametrů a testování hypotéz.

První kapitola shrnuje teorii jednorozměrné lineární regrese, která tvoří základy k odvození teorie mnohorozměrné lineární regrese. Tato kapitola seznamuje s obecným modelem jednorozměrné lineární regrese. Je zaměřena na metody odhadů regresních parametrů a reziduálních rozptylů, které jsou potřebné při testování hypotéz. Dále se zabývá tím, jaké lze v tomto modelu testovat hypotézy a jak přesně je testovat. Na závěr této kapitoly je připojen příklad nejjednodušších modelů jednorozměrné lineární regrese. Více informací k problematice z této kapitoly lze nalézt např. v knihách [1] a [2].

Druhá kapitola se soustřeďuje na teorii mnohorozměrné lineární regrese. Taktéž popisuje základní model, vysvětluje, jak lze odhadovat regresní parametry a matici chyb, která se používá při testování hypotéz. Podrobněji se zabývá problematikou testování hypotéz. Seznamuje s nejčastějšími testovými statistikami založenými na poměru věrohodností, Pillaiově a Lawley-Hotellingově stopě. Na rozdíl od první kapitoly je zde stručně vysvětleno, jak je možné určit predikční oblast pro nové pozorování. Na závěr této kapitoly je opět připojen příklad, který lze popsat jedním z nejzákladnějších mnohorozměrných lineárních modelů. Tento příklad je ještě doprovázen simulacemi sestrojenými v programu R (tento simulační program je přiložen k práci na kompaktním disku). Tyto simulace slouží hlavně k porovnání jednotlivých testů. Více informací k pojmům této kapitoly lze nalézt např. v knize [3].

Kapitola 1

Jednorozměrná lineární regrese

1.1 Vymezení základních pojmů

Jednorozměrná regresní analýza je běžně používaná statistická metoda sloužící ke zkoumání závislosti proměnné Y na souboru nezávislých proměnných $X_1, \ldots, X_p$, kde $p \geq 1$. Takto můžeme například pozorovat závislost výšky člověka na věku a pohlaví, nebo hustotu dané látky na měnící se teplotě. Cílem regresní analýzy je otestovat, na kterých konkrétních $X_1, \ldots, X_p$ proměnná Y závisí, jak na nich závisí, popřípadě předpovídat hodnoty Y, máme - li k dispozici pozorování proměnných $X_1, \ldots, X_p$.

Proměnnou Y můžeme tedy vyjádřit jako funkci $X_1, \ldots, X_p$ takto: $Y = f(X_1, \ldots, X_p) + e$. Tento model se nazývá regresní, funkce f je regresní funkce a e je náhodná chyba. Tato náhodná chyba může vzniknout například v důsledku nepřesností měření vektoru veličiny Y. Jednotlivé chyby pozorovat nelze.

Pokud je funkce f lineární, mluvíme o lineárním regresním modelu. Model můžeme přepsat ve tvaru

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e, \tag{1.1}$$

kde koeficienty $\beta_1, \ldots, \beta_p$ jsou blíže nespecifikované neznámé regresní parametry. Tyto parametry se snažíme odhadnout.

Obecně máme data sestávající se z nezávislých pozorování, které vytvoří n-rozměrný náhodný vektor $\mathbf{Y}$ a p n-rozměrných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_p$, jejichž

hodnoty známe. Dále tedy budeme předpokládat, že $\mathbf{Y}$ je náhodný vektor se složkami $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X}$ je matice typu $(n \times p)$, jejíž i-tý sloupec je sloupec vektoru $\mathbf{X}_i = (x_{1i}, \ldots, x_{ni})^T$, $i = 1, \ldots, p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ je vektor jednotlivých regresních parametrů a $\mathbf{e} = (e_1, \ldots, e_n)^T$ je vektor chyb.

Poznámka. První sloupec matice $\mathbf{X}$ zpravidla tvoří vektor jedniček. Pokud by tomu tak nebylo, nadrovina vyjadřující závislost proměnné Y na proměnných $X_1, \ldots, X_p$ by vždy procházela počátkem. První sloupec tvořený vektorem jedniček nám umožní posouvání této nadroviny o konstantu β_1, tudíž přímka pak neprochází počátkem, ale protíná svislou ypsilonovou osu v hodnotě parametru β_1.

Model tedy můžeme přepsat maticově do tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1.2}$$

což můžeme rozepsat po složkách jako

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}. \tag{1.3}$$

Pro jednotlivé složky vektoru $\mathbf{Y}$ platí

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i, \qquad \text{pro } i = 1, 2, \ldots, n. \tag{1.4}$$

Předpokládáme, že model splňuje následující podmínky:

1. Y_i je náhodná veličina a její hodnota je funkcí vysvětlujících veličin $x_{i1}, \ldots, x_{ip}$, tato funkce je lineární a stejná pro všechna $i = 1, 2, \ldots, n$.

2. Vektor chyb je vektor náhodných veličin $\boldsymbol{e} = (e_1, \ldots, e_n)^T$, který splňuje podmínky
$$E\boldsymbol{e} = \mathbf{0}, \qquad var\,\boldsymbol{e} = \sigma^2\mathbf{I}, \tag{1.5}$$
kde parametr σ^2 taktéž není znám.

1.2 Odhady parametrů

1.2.1 Metoda nejmenších čtverců

Nejpoužívanější metodou pro bodový odhad koeficientů je metoda nejmenších čtverců. Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ a matici daných čísel $\mathbf{X}_{(n\times p)}$, tedy model tvaru (1.2), kde vektor chyb $\mathbf{e}$ splňuje podmínky (1.5). Pro střední hodnotu a rozptyl vektoru $\mathbf{Y}$ tedy platí, že $E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, $var\mathbf{Y} = \sigma^2\mathbf{I}$. Metoda spočívá v proložení dat $(x_{i1}, \ldots, x_{ip}, Y_i)$ nadrovinou tak, aby součet čtverců odchylek (tj. vzdálenost Y_i a odhadu Y_i) byl co možná nejmenší (tato nadrovina se nazývá regresní). Tedy hledáme minimum pro výraz

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.6}$$

Označme tento odhad $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$. Za předpokladu, že hodnost matice $\mathbf{X}$ je rovna p platí, že matice $\mathbf{X}^T\mathbf{X}$ je typu $p \times p$ a je regulární.

Věta 1 *Pro odhad vektoru parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců platí $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.*

Důkaz. Pro vektor $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ platí

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

$$= \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Y} = \mathbf{0}.$$

Nechť $\boldsymbol{\beta}^*$ je jiný odhad. Potom

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

$$= [(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*)]^T[(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*)].$$

Po roznásobení a využití podmínky $\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$ dostáváme

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

Výraz $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ je konstantní, tedy $S(\boldsymbol{\beta}^*)$ je minimální právě tehdy, když výraz $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ je minimální. Matice $\mathbf{X}^T\mathbf{X}$ je pozitivně definitní, neboť její hodnost $h(\mathbf{X}) = p$. Pak

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \geq 0$$

a rovnost nastává pro $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}$. □

Výraz $R = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ se nazývá reziduální součet čtverců. Označme dále $s^2 = R/(n-p)$. Této veličině se říká reziduální rozptyl. Pro rezidualní rozptyl platí $Es^2 = \sigma^2$ a tedy s^2 je nestranným odhadem parametru σ^2.

Poznámka. Reziduální součet čtverců můžeme ekvivalentně psát jako $\mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y}$, kde $\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Tvrzení 2 *Pro střední hodnotu a rozptyl parametru $\hat{\boldsymbol{\beta}}$ platí $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $var\hat{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.*

Důkaz. Lze nalézt např. v [2]. □

Pokud dále předpokládáme, že vektor chyb $\mathbf{e}$ má normální rozdělení, tedy $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, pak vektor $\mathbf{Y}$ má taktéž normální rozdělení, tedy $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, a pro odhad parametru $\hat{\boldsymbol{\beta}}$ platí následující tvrzení.

Tvrzení 3 *1. $\hat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$,*

2. $\frac{R}{\sigma^2} \sim \chi^2_{n-p}$,

3. $\hat{\boldsymbol{\beta}}$ a R jsou nezávislé.

Důkaz. Lze nalézt např. v [2]. □

Tvrzení 4 *Nechť v_{ij} je (i,j)-tý prvek matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Pak pro každé $i = 1, \ldots, p$ má náhodná veličina*

$$T_i = \frac{\widehat{\beta}_i - \beta_i}{\sqrt{s^2 v_{ii}}}$$

t-rozdělení o $n - p$ stupních volnosti.

Důkaz. Lze nalézt např. v [2]. □

Po vypočtení regresních koeficientů můžeme dále testovat hypotézu $\hat{\beta}_{q+1} = \hat{\beta}_p = \cdots = 0$ pro $q < p$. Tedy je to test toho, zda některé z vysvětlujících proměnných nemají žádný podstatný efekt na závislou proměnnou Y. Chceme porovnat úplný model $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ s redukovaným modelem $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$ (na tento model lze převést každou podmnožinu vysvětlujících proměnných X_i velikosti q pouhým přečíslováním indexů). Test se provádí pomocí porovnání reziduálních součtů

čtverců obou modelů. Označme R reziduální součet čtverců úplného modelu a R_0 reziduální součet čtverců redukovaného modelu. Testová statistika má potom tvar

$$F = \frac{R - R_0}{(p-q)s^2} = \frac{R - R_0}{p-q} \cdot \frac{n-p}{R},$$

kde s^2 je reziduální rozptyl úplného modelu.

Tvrzení 5 *Za platnosti nulové hypotézy $\hat{\beta}_{q+1} = \hat{\beta}_p = \cdots = 0$ pro $q < p$ má testová statistika F F-rozdělení o $p-q$ a $n-p$ stupních volnosti.*

Důkaz. Lze nalézt např. v [2]. □

Pokud zamítáme nulovou hypotézu, znamená to, že alespoň jedna z vysvětlujících proměnných $X_{q+1}, \ldots, X_p$ má významný vliv na proměnnou Y.

1.2.2 Metoda maximální věrohodnosti

Mějme model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ a předpokládejme, že $e_1, \ldots, e_n$ jsou nezávislé náhodné veličiny s normálním rozdělením $N(0, \sigma^2)$. Náhodný vektor $\mathbf{e}$ má potom n-rozměrné normální rozdělení $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Vektor $\mathbf{Y}$ má tedy n-rozměrné normální rozdělení $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Pro hustotu vektoru $\mathbf{Y}$ platí

$$\begin{aligned} f(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - x_{i1}\beta_1 - \cdots - x_{ip}\beta_{ip})^2\right). \end{aligned}$$

Řekneme, že odhad $\boldsymbol{\beta}^*$ je odhad parametru $\boldsymbol{\beta}$ metodou maximální věrohodnosti, platí-li

$$f(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}^*, \sigma) \geq f(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma) \qquad \text{pro každé } \boldsymbol{\beta} \in \mathbb{R}^p.$$

Pro libovolné $\sigma > 0$ je tato nerovnost splněna právě tehdy, když je výraz

$$\begin{aligned} S(\boldsymbol{\beta}) = S(\beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n}(Y_i - x_{i1}\beta_1 - \cdots - x_{ip}\beta_{ip})^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

minimální. Podle věty 1, dokázané v předešlé části, je tento výraz minimální pro $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \hat{\boldsymbol{\beta}}$, kde $\hat{\boldsymbol{\beta}}$ je odhad vektoru parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců. Tedy odhad metodou maximální věrohodnosti je roven odhadu metodou nejmenších čtverců. Označme $R = S(\hat{\boldsymbol{\beta}})$ (tedy R je reziduální součet čtverců). Dále mějme funkci

$$f(\mathbf{Y}, \mathbf{X}\hat{\boldsymbol{\beta}}, \sigma) = (2\pi)^{-n/2}\sigma^{-n}\ \exp\left(-\frac{R}{2\sigma^2}\right),$$

kterou nyní chápeme jako funkci proměnné σ. Tato funkce $f(\mathbf{Y}, \mathbf{X}\hat{\boldsymbol{\beta}}, \sigma)$ nabývá svého maxima v bodě $\sigma = \sqrt{R/n}$. Proto odhady parametrů $\beta_1, \ldots, \beta_p, \sigma$ metodou maximální věrohodnosti jsou $\hat{\beta}_1, \ldots, \hat{\beta}_p, \sqrt{R/n}$.

1.3 Speciální případy jednorozměrné lineární regrese

Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ a vektor daných čísel $\mathbf{X} = (x_1, \ldots, x_n)^T$. Mezi nejzákladnější lineární modely patří přímka procházející počátkem. Pro tuto situaci máme model

$$Y_i = \beta x_i + e_i \qquad \text{pro } i = 1, 2, \ldots, n,$$

kde $e_1, \ldots, e_n$ jsou nezávislé náhodné veličiny s rozdělením $N(0, \sigma^2)$. Pro použití metody nejmenších čtverců odvozené výše si stačí uvědomit, že vektor $\boldsymbol{\beta}$ je jednorozměrný, tudíž má jediný prvek β a matice $\mathbf{X}$ je typu $n \times 1$.

Pak podle věty 1 pro odhad parametru $\hat{\beta}$ platí $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. V tomto případě je $(\mathbf{X}^T\mathbf{X})^{-1} = (\sum_{i=1}^n x_i^2)^{-1}$ a $\mathbf{X}^T\mathbf{Y} = \sum_{i=1}^n x_i Y_i$. Tedy

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Dále užitím vztahů $R = \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}$ a $s^2 = R/(n-p)$ dostáváme vztah pro reziduální rozptyl ve tvaru

$$s^2 = \frac{R}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - \hat{\beta}\sum_{i=1}^n x_i Y_i}{n-1} = \frac{\sum_{i=1}^n {x_i}^2 \cdot \sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n x_i Y_i)^2}{(n-1)\sum_{i=1}^n {x_i}^2}.$$

U tohoto modelu nejčastěji testujeme hypotézu $H_0 : \beta = 0$, tedy hypotézu, že Y_i na x_i vůbec nezávisí. K tomu použijeme tvrzení 4. Veličina

$$T = \frac{\hat{\beta}}{s}\sqrt{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i Y_i}{s\sqrt{\sum_{i=1}^{n} x_i^2}}$$

má rozdělení $t_{(n-1)}$ a pokud $\mid T \mid \geq t_{(n-1)}(1-\alpha/2)$, pak hypotézu H_0 zamítáme.

Dalším používaným regresním modelem je obecná přímka, tj. model ve tvaru

$$Y_i = \beta_1 + \beta_2 x_i + e_i \qquad \text{pro } i = 1, 2, \ldots, n.$$

Pro ten platí

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix},$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}.$$

Označme

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Podle věty 1 pro odhady parametrů β_1 a β_2 metodou nejmenších čtverců platí následující vztahy

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2\overline{x}. \tag{1.7}$$

Pro reziduální rozptyl platí

$$s^2 = \frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} Y_i - \hat{\beta}_2 \sum_{i=1}^{n} x_i Y_i}{n-2}.$$

U tohoto modelu nejčastěji testujeme hypotézu $H_0 : \beta_2 = 0$ (tedy hypotézu, že Y_i na x_i vůbec nezávisí). Pro T_2 platí

$$T_2 = \frac{\hat{\beta}_2}{s} \sqrt{\sum_{j=i}^{n} x_i^2 - n\overline{x}^2}.$$

Pokud $\mid T_2 \mid \geq t_{(n-2)}(1 - \alpha/2)$, pak hypotézu H_0 na hladině α zamítáme.

Kapitola 2

Mnohorozměrná lineární regrese

2.1 Popis modelu

Na rozdíl od jednorozměrného lineárního regresního modelu, který popisuje závislost jedné proměnné Y na souboru proměnných $X_1, \ldots, X_p$, mnohorozměrný model popisuje závislost souboru proměnných $Y_1, \ldots, Y_q$ na souboru vysvětlujících proměnných $X_1, \ldots, X_p$.

Mějme tedy soubor závislých proměnných $Y_1, \ldots, Y_q$ a ke každé z nich mějme k dispozici n pozorování, tedy hodnoty $Y_{i1}, \ldots, Y_{iq}$ pro $i = 1, \ldots, n$. Označme $\mathbf{Y}_{.h}$ n-rozměrný vektor odpovídající závislé proměnné Y_h, $h = 1, \ldots, q$. Tedy $\mathbf{Y}_{.h} = (Y_{1h}, \ldots, Y_{nh})^T$. Na každý vektor $\mathbf{Y}_{.h}$ lze aplikovat jednorozměrný lineární regresní model tvaru

$$\mathbf{Y}_{.h} = \mathbf{X}\boldsymbol{\beta}_{.h} + \mathbf{e}_{.h}, \qquad h = 1, \ldots, q, \tag{2.1}$$

$$E\mathbf{e}_{.h} = \mathbf{0}, \qquad var\mathbf{e}_{.h} = \sigma_{hh}\mathbf{I}_n,$$

kde $\mathbf{X} = (\mathbf{X}_{.1}^T, \ldots, \mathbf{X}_{.p}^T)$ je matice známých čísel velikosti $(n \times p)$ a vektory $\mathbf{e}_{.h}$ jsou vektory chyb. Stejně tak jako v jednorozměrném modelu první sloupec matice $\mathbf{X}$ obvykle tvoří vektor jedniček. Tato matice je stejná pro každý vektor $\mathbf{Y}_{.h}$, $h = 1, \ldots, q$. Vektor regresních parametrů $\boldsymbol{\beta}_{.h} = (\beta_{1h}, \ldots, \beta_{ph})^T$ a vektor chyb $\mathbf{e}_{.h} = (e_{1h}, \ldots, e_{nh})^T$ mohou být různé pro různé $h = 1, \ldots, q$.

Zapišme vektory $\mathbf{Y}_{.h}$, $\boldsymbol{\beta}_{.h}$ a $\mathbf{e}_{.h}$ do matic následovně

$$\mathbf{Y}_{n\times q} = (\mathbf{Y}_{.1}, \ldots, \mathbf{Y}_{.q}) \qquad \boldsymbol{\beta}_{p\times q} = (\boldsymbol{\beta}_{.1}, \ldots, \boldsymbol{\beta}_{.q}) \qquad \mathbf{e}_{n\times q} = (\mathbf{e}_{.1}, \ldots, \mathbf{e}_{.q}).$$

Mnohorozměrný lineární model pak můžeme napsat ve tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.2}$$

což lze po složkách rozepsat jako

$$\begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1q} \\ Y_{21} & Y_{22} & \cdots & Y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nq} \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1q} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq} \end{pmatrix}$$
$$+ \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1q} \\ e_{21} & e_{22} & \cdots & e_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nq} \end{pmatrix}. \tag{2.3}$$

Pro matici $\mathbf{e}$ náhodných chyb předpokládáme $E\mathbf{e} = \mathbf{0}$ a dále

$$Cov(e_{ih}, e_{i'h'}) = \begin{cases} \sigma_{hh'} & \text{pokud } i = i' \\ 0 & \text{pokud } i \neq i' \end{cases},$$

což lze zjednodušit jako

$$Cov(e_{ih}, e_{i'h'}) = \sigma_{hh'}\delta_{ii'}, \qquad \text{kde } \delta_{ii'} = \begin{cases} 1 & \text{pokud } i = i' \\ 0 & \text{pokud } i \neq i' \end{cases}.$$

Poznámka. Při konstruování statistických testů a oblastí spolehlivost předpokládáme, že náhodná veličina e_{ih} má normální rozdělení a pro její střední hodnotu a rozptyl platí: $Ee_{ih} = 0$, $vare_{ih} = \sigma_{hh}$. Tedy náhodný vektor $\mathbf{e}_{\cdot h}$ má mnohorozměrné normální rozdělení se střední hodnotou $E\mathbf{e}_{\cdot h} = \mathbf{0}$ a rozptylem $var\mathbf{e}_{\cdot h} = \sigma_{hh}\mathbf{I}_n$.

Mnohorozměrný lineární model můžeme zapsat i v jiném tvaru. Nyní budeme nahlížet na řádky matic $\mathbf{Y}$, $\mathbf{X}$ a $\mathbf{e}$ jako na vektory, tedy

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{1\cdot}^T \\ \vdots \\ \mathbf{Y}_{n\cdot}^T \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_{1\cdot}^T \\ \vdots \\ \mathbf{X}_{n\cdot}^T \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} \boldsymbol{\epsilon}_{1\cdot}^T \\ \vdots \\ \boldsymbol{\epsilon}_{n\cdot}^T \end{pmatrix}. \tag{2.4}$$

Pro tento model platí

$$\mathbf{Y}_{i\cdot}^T = \mathbf{X}_{i\cdot}^T\boldsymbol{\beta} + \boldsymbol{\epsilon}_{i\cdot}^T, \qquad \text{pro } i = 1, \ldots, n. \tag{2.5}$$

Pro vektor náhodných chyb $\boldsymbol{\epsilon}_{i\cdot}$ platí

$$E\boldsymbol{\epsilon}_{i\cdot} = \mathbf{0}, \qquad Var\boldsymbol{\epsilon}_{i\cdot} = \Sigma_{q\times q} = (\sigma_{hh'})_{h,h'=1}^{q},$$

$$Cov(\boldsymbol{\epsilon}_{i\cdot}, \boldsymbol{\epsilon}_{j\cdot}) = 0 \qquad \text{pro } i \neq j.$$

2.2 Odhady parametrů

Klíčovým krokem k určení regresních koeficientů mnohorozměrného modelu je přepsat si tento model jako jednorozměrný. Model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \qquad E\mathbf{e}_{\cdot h} = \mathbf{0}, \qquad Cov(e_{ih}, e_{i'h'}) = \sigma_{hh'}\delta_{ii'} \tag{2.6}$$

přepíšeme jako

$$\begin{pmatrix} \mathbf{Y}_{\cdot 1} \\ \mathbf{Y}_{\cdot 2} \\ \vdots \\ \mathbf{Y}_{\cdot q} \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{\cdot 1} \\ \boldsymbol{\beta}_{\cdot 2} \\ \vdots \\ \boldsymbol{\beta}_{\cdot q} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{\cdot 1} \\ \mathbf{e}_{\cdot 2} \\ \vdots \\ \mathbf{e}_{\cdot q} \end{pmatrix}, \tag{2.7}$$

kde vektor chyb má střední hodnotu $\mathbf{0}$ a variační matici tvaru

$$\begin{pmatrix} \sigma_{11}\mathbf{I}_n & \sigma_{12}\mathbf{I}_n & \cdots & \sigma_{1q}\mathbf{I}_n \\ \sigma_{12}\mathbf{I}_n & \sigma_{22}\mathbf{I}_n & \cdots & \sigma_{2q}\mathbf{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1q}\mathbf{I}_n & \sigma_{2q}\mathbf{I}_n & \cdots & \sigma_{qq}\mathbf{I}_n \end{pmatrix}. \tag{2.8}$$

Na tento model nahlížíme jako na jednorozměrný, neboť levá strana rovnice je nq-rozměrný vektor a pravá strana je rovna součinu matice typu $nq \times pq$ a nq-rozměrného vektoru, ke kterému je přičten ještě nq-rozměrný vektor chyb. Následující definice nám umožňují zapsat tento model v přehledném tvaru.

Definice 6 *(Operátor Vec) Nechť $\boldsymbol{Y} = (Y_{ij})$ je matice typu $n \times q$. Pak definujeme Vec($\boldsymbol{Y}$) jako nq-rozměrný sloupcový vektor, pro který platí Vec($\boldsymbol{Y}$) $= (Y_{11}, \ldots, Y_{n1}, Y_{12}, \ldots, Y_{n2}, \ldots, Y_{1q}, \ldots, Y_{nq})^T$.*

Definice 7 *(Kroneckerův součin matic) Nechť* $\boldsymbol{A}$ *je matice typu* $k \times q$ *a nechť* $\boldsymbol{B}$ *je matice typu* $n \times p$*. Pak Kroneckerův součin matic* $\boldsymbol{A}$ *a* $\boldsymbol{B}$ *je definován jako*

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1q}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{k1}\boldsymbol{B} & \cdots & a_{kq}\boldsymbol{B} \end{pmatrix}.$$

Lemma 8 *Pro Kroneckerův součin matic platí*
$[\boldsymbol{A} \otimes \boldsymbol{B}][\boldsymbol{C} \otimes \boldsymbol{D}] = [\boldsymbol{AB} \otimes \boldsymbol{CD}]$*, pokud je násobení matic definováno.*

Důkaz. Důkaz plyne přímo z definice, stačí si rozepsat matice po složkách a porovnat jednotlivé součiny. □

Rovnici (2.7) můžeme přepsat také do tvaru

$$\mathrm{Vec}(\mathbf{Y}) = [\mathbf{I}_q \otimes \mathbf{X}]\mathrm{Vec}(\mathbf{B}) + \mathrm{Vec}(\mathbf{e}). \tag{2.9}$$

Pro vektor $\mathrm{Vec}(\mathbf{e})$ platí

$$E(\mathrm{Vec}(\mathbf{e})) = \mathbf{0}, \qquad var(\mathrm{Vec}(\mathbf{e})) = \Sigma \otimes \mathbf{I}_n. \tag{2.10}$$

2.2.1 Metoda nejmenších čtverců

Nejpoužívanější metoda pro odhad koeficientů modelu (2.2) je metoda nejmenších čtverců. Dále budeme předpokládat, že matice $\mathbf{X}$ je regulární.

Věta 9 *Pro odhad matice parametrů* $\boldsymbol{\beta}$ *mnohorozměrného lineárního modelu* $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ *platí* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$*.*

Důkaz. Označme matici $\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Ekvivalentní zápis tvrzení věty s použitím matice $\mathbf{M}$ je $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{MY}$. Model (2.6) nyní berme jako jednorozměrný lineární a k odhadu parametru $Vec(\hat{\boldsymbol{\beta}})$ můžeme aplikovat větu 1, podle níž platí

$$Vec(\hat{\boldsymbol{\beta}}) = ([\mathbf{I}_q \otimes \mathbf{X}]^T[\mathbf{I}_q \otimes \mathbf{X}])^{-1}[\mathbf{I}_q \otimes \mathbf{X}]^T\, Vec(\mathbf{Y}).$$

Potom platí

$$[\mathbf{I}_q \otimes \mathbf{X}]\, Vec(\hat{\boldsymbol{\beta}}) = [\mathbf{I}_q \otimes \mathbf{X}]([\mathbf{I}_q \otimes \mathbf{X}]^T[\mathbf{I}_q \otimes \mathbf{X}])^{-1}[\mathbf{I}_q \otimes \mathbf{X}]^T\, Vec(\mathbf{Y}).$$

Z lemmatu 8 a vlastností operací matic plyne rovnost

$$[\mathbf{I}_q \otimes \mathbf{X}]([\mathbf{I}_q \otimes \mathbf{X}]^T[\mathbf{I}_q \otimes \mathbf{X}])^{-1}[\mathbf{I}_q \otimes \mathbf{X}]^T = [\mathbf{I}_q \otimes \mathbf{M}].$$

Tedy pro uvažovaný jednorozměrný model platí

$$[\mathbf{I}_q \otimes \mathbf{X}]\, Vec(\hat{\boldsymbol{\beta}}) = [\mathbf{I}_q \otimes \mathbf{M}]\, Vec(\mathbf{Y}),$$

což můžeme přepsat do tvaru

$$\begin{pmatrix} \mathbf{X}\hat{\boldsymbol{\beta}}_{\cdot 1} \\ \vdots \\ \mathbf{X}\hat{\boldsymbol{\beta}}_{\cdot q} \end{pmatrix} = \begin{pmatrix} \mathbf{M}\mathbf{Y}_{\cdot 1} \\ \vdots \\ \mathbf{M}\mathbf{Y}_{\cdot q} \end{pmatrix}$$

a to je ekvivalentní zápisu $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{M}\mathbf{Y}$. □

2.2.2 Metoda maximální věrohodnosti

Pro odvození odhadů metodou maximální věrohodnosti použijeme model (2.4) a na řádky matic $\mathbf{Y}, \mathbf{X}$ a $\mathbf{e}$ budeme nahlížet jako na vektory. Dále budeme předpokládat, že matice Σ je regulární, jednotlivé řádky matice $\mathbf{Y}$ jsou nezávislé a mají mnohorozměrné normální rozdělení, tj. $\mathbf{Y}_{i\cdot} \sim N_q(\boldsymbol{\beta}^T\mathbf{X}_{i\cdot}, \Sigma)$ pro $i = 1, \ldots, n$. Pro hustotu vektoru $\mathbf{Y}_{i\cdot}$ tedy platí

$$f(\mathbf{Y}_{i\cdot}, \boldsymbol{\beta}^T\mathbf{X}_{i\cdot}, \Sigma) = \frac{1}{\sqrt{(2\pi)^q \mid \Sigma \mid}} \exp\left(-\frac{1}{2}(\mathbf{Y}_{i\cdot} - \boldsymbol{\beta}^T\mathbf{X}_{i\cdot})^T\Sigma^{-1}(\mathbf{Y}_{i\cdot} - \boldsymbol{\beta}^T\mathbf{X}_{i\cdot})\right)$$

a věrohodnostní funkce pro $\mathbf{Y}$ je tvaru

$$L(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \Sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^q \mid \Sigma \mid}} \exp\left(-\frac{1}{2}(\mathbf{Y}_{i\cdot} - \boldsymbol{\beta}^T\mathbf{X}_{i\cdot})^T\Sigma^{-1}(\mathbf{Y}_{i\cdot} - \boldsymbol{\beta}^T\mathbf{X}_{i\cdot})\right).$$

Zlogaritmujeme-li ji, dostaneme funkci tvaru

$$\ell(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \Sigma) = -\frac{nq}{2}\ \log(2\pi) - \frac{n}{2}\ \log(\mid \Sigma \mid)$$

$$-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_{i\cdot} - \boldsymbol{\beta}^T\mathbf{X}_{i\cdot})^T\Sigma^{-1}(\mathbf{Y}_{i\cdot} - \boldsymbol{\beta}^T\mathbf{X}_{i\cdot}).$$

Pro libovolný jednorozměrný model, kde variační matice vektoru chyb je brána jako konstanta, nabývá věrohodnostní rovnice svého maxima pro odhad vektoru $\hat{\boldsymbol{\beta}}$ (metodou nejmenších čtverců). Z předchozí kapitoly dále víme, že mnohorozměrný lineární model lze pomocí operátoru Vec a Kroneckerova součinu převést na jednorozměrný. Odhad matice $\hat{\boldsymbol{\beta}}$ metodou nejmenších čtverců nezávisí na matici Σ, a proto je věrohodnostní rovnice maximalizována pro libovolné Σ nahrazením matice $\boldsymbol{\beta}$ maticí odhadů $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Tedy maximálně věrohodný odhad je roven odhadu metodou nejmenších čtverců.

Odhad Σ metodou maximální věrohodnosti nám dává následující tvrzení.

Tvrzení 10 *Nechť* $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ *a* $r(\boldsymbol{X})$ *je hodnost matice* $\boldsymbol{X}$. *Pak platí*

1. *Odhad matice* Σ *metodou maximální věrohodnosti je* $\hat{\Sigma} = \frac{1}{n}\ \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{M})\,\boldsymbol{Y}$.
2. $\frac{n}{n-r(\boldsymbol{X})}\ \hat{\Sigma} = \frac{1}{n-r(\boldsymbol{X})}\ \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{M})\,\boldsymbol{Y}$ *je nestranný odhad matice* Σ.

Důkaz. Lze nalézt např. v [3]. □

2.3 Testování hypotéz

Jak přesně testovat hypotézy si ukážeme až v následující podkapitole. Zde pouze zformulujeme úvahy, které jsou společné pro všechny budoucí testové statistiky. Uvažujme mnohorozměrný lineární model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{2.11}$$

Tento model chceme porovnat s redukovaným modelem

$$\mathbf{Y} = \mathbf{X}_0\boldsymbol{\Gamma} + \mathbf{e}, \tag{2.12}$$

kde $\boldsymbol{\Gamma}$ je matice regresních koeficientů typu $s \times q$, kde $s \leq p$ a pro matici $\mathbf{X}_0$ platí $C(\mathbf{X}_0) \subset C(\mathbf{X})$ (kde symbol $C(\mathbf{X})$, resp. $C(\mathbf{X}_0)$ značí vektorový prostor generovaný sloupcovými vektory matice $\mathbf{X}$, resp. $C(\mathbf{X}_0)$). Pro matici $\mathbf{e}$ předpokládáme platnost rovnosti (2.10), přičemž matice $\Sigma \otimes \mathbf{I}_n$ není známá.

Definujme matici $\mathbf{M}_0 = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T$. Pak test hypotézy platnosti modelu (2.12) je založen na testové statistice, která využívá matice

$$\mathbf{H} \equiv \mathbf{Y}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} \qquad \text{a} \qquad \mathbf{E} \equiv \mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y}.$$

Testová statistika je pak často funkcí matice

$$\mathbf{H}\mathbf{E}^{-1} = \mathbf{Y}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}(\mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y})^{-1}.$$

Testová statistika hypotéz jednorozměrného modelu je funkcí stejného tvaru, ale oproti mnohorozměrnému modelu je tato statistika skalární veličina. Na rozdíl od jednorozměrných modelů pro mnohorozměrné neexistuje jedna univerzální testová statistika vhodná ke zkoumání platností různých hypotéz.

Dále uvažujme testování hypotézy

$$H_0 : \boldsymbol{\Lambda}^T\boldsymbol{\beta} = \mathbf{0} \qquad \text{proti} \qquad H_1 : \boldsymbol{\Lambda}^T\boldsymbol{\beta} \neq \mathbf{0}, \tag{2.13}$$

kde $\boldsymbol{\Lambda}^T = \mathbf{P}^T\mathbf{X}$.

Označme $\mathbf{M}_{MP} = \mathbf{MP}((\mathbf{MP})^T(\mathbf{MP}))^{-1}(\mathbf{MP})^T = \mathbf{MP}(\mathbf{P}^T\mathbf{MP})^{-1}\mathbf{P}^T\mathbf{M}$. Pak lze ukázat (celé odvození lze nalézt v [3]), že redukovaný model (2.12) můžeme napsat ve tvaru

$$\mathbf{Y} = (\mathbf{M} - \mathbf{M}_{MP})\boldsymbol{\Gamma} + \mathbf{e}.$$

A tedy matice $\mathbf{H}$ má v tomto případě tvar

$$\begin{aligned}
\mathbf{H} &\equiv \mathbf{Y}^T[\mathbf{M} - (\mathbf{M} - \mathbf{M}_{MP})]\mathbf{Y} = \mathbf{Y}^T\mathbf{M}_{MP}\mathbf{Y} \\
&= \mathbf{Y}^T\mathbf{MP}(\mathbf{P}^T\mathbf{MP})^{-1}\mathbf{P}^T\mathbf{MY} \\
&= \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}(\mathbf{PX}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\boldsymbol{\Lambda}^T\hat{\boldsymbol{\beta}})^T[\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\Lambda}]^{-1}(\boldsymbol{\Lambda}^T\hat{\boldsymbol{\beta}}).
\end{aligned}$$

Poznámka. Při výpočtu matice $\mathbf{H}$ je třeba určit inverzní k matici, která nemusí být regulární. V tomto případě použijeme tzv. pseudoinverzní matici.

Podobně můžeme testovat hypotézu

$$H_0 : \boldsymbol{\Lambda}^T\boldsymbol{\beta} = \mathbf{W} \qquad \text{proti} \qquad H_1 : \boldsymbol{\Lambda}^T\boldsymbol{\beta} \neq \mathbf{W}, \tag{2.14}$$

kde $\mathbf{W}$ je známá matice čísel. Nechť $\mathbf{G}$ je známé řešení rovnice $\mathbf{\Lambda}^T\mathbf{G} = \mathbf{W}$, pak matice $\mathbf{H}$ a $\mathbf{E}$ jsou tvaru

$$\begin{aligned}\mathbf{H} &\equiv (\mathbf{Y} - \mathbf{XG})^T\mathbf{M}_{MP}(\mathbf{Y} - \mathbf{XG}) \\ &= (\mathbf{Y} - \mathbf{XG})^T\mathbf{MP}(\mathbf{P}^T\mathbf{MP})^{-1}\mathbf{P}^T\mathbf{M}(\mathbf{Y} - \mathbf{XG}) \\ &= (\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{W})^T[\mathbf{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}]^{-1}(\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{W}) \\ \mathbf{E} &= \mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y}.\end{aligned}$$

Zvláštním případem hypotézy $H_0 : \mathbf{\Lambda}^T\boldsymbol{\beta} = \mathbf{0}$ je hypotéza

$$H_0 : \mathbf{\Lambda}^T\boldsymbol{\beta}\boldsymbol{\xi} = \mathbf{0} \qquad \text{proti} \qquad H_1 : \mathbf{\Lambda}^T\boldsymbol{\beta}\boldsymbol{\xi} \neq \mathbf{0}, \tag{2.15}$$

kde ξ je libovolný q-rozměrný vektor a $\Lambda^T = \mathbf{P}^T\mathbf{X}$. Nyní můžeme model (2.11) přepsat do tvaru

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

Toto je test hypotézy jednorozměrného lineárního modelu, kde $\mathbf{e}\boldsymbol{\xi}$ je vektor chyb, pro jehož rozdělení platí $\mathbf{e}\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{\xi}^T\Sigma\boldsymbol{\xi}\mathbf{I}_n)$, a $\boldsymbol{\beta}\boldsymbol{\xi}$ je vektor regresních parametrů, jehož odhad metodou nejmenších čtverců je $\hat{\boldsymbol{\beta}}\boldsymbol{\xi}$. Pak z teorie jednorozměrné lineární regrese plyne

$$\frac{(\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}}\boldsymbol{\xi})^T(\mathbf{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}}\boldsymbol{\xi}/r(\mathbf{\Lambda})}{(\mathbf{Y}\boldsymbol{\xi})^T(\mathbf{I} - \mathbf{M})(\mathbf{Y}\boldsymbol{\xi})/n - r(\mathbf{X})} \sim F(r(\mathbf{\Lambda}), n - r(\mathbf{X})).$$

Předchozí test hypotézy můžeme zobecnit následovně: uvažujme matici $\mathbf{Z}$ velikosti $q \times r$ známých čísel s hodností $r(\mathbf{Z}) = r < q$. Zkoumat platnost hypotézy

$$H_0 : \mathbf{\Lambda}^T\boldsymbol{\beta}\mathbf{Z} = \mathbf{0} \qquad \text{proti} \qquad H_1 : \mathbf{\Lambda}^T\boldsymbol{\beta}\mathbf{Z} \neq \mathbf{0} \tag{2.16}$$

je to samé, jako zkoumat platnost hypotézy tvaru (2.13) pro mnohorozměrný model tvaru

$$\mathbf{YZ} = \mathbf{X}\boldsymbol{\beta}\mathbf{Z} + \mathbf{eZ}.$$

2.3.1 Testové statistiky

Předpokládejme, že chceme otestovat platnost redukovaného modelu tvaru (2.12) nebo hypotézu tvaru (2.13). Potom matice E je tvaru

$$\mathbf{E} = \mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y}$$

a matice H je tvaru

$$\mathbf{H} = \mathbf{Y}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} \qquad \text{nebo} \qquad \mathbf{H} = \mathbf{Y}^T\mathbf{M}_{MP}\mathbf{Y}.$$

Definice 11 *: Nechť* $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ *jsou nezávislé a mají mnohorozměrné normální rozdělaní* $N(\boldsymbol{\mu}_i, \Sigma)$*,* $i = 1, \ldots, n$*. Pak*

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{W}_i \boldsymbol{W}_i^T$$

má Wishartovo rozdělení o n *stupních volnosti, variační maticí* Σ *a maticí parametrů* $\boldsymbol{Q}$*, kde*

$$\boldsymbol{Q} = \frac{1}{2}\, \Sigma^{-1} \sum_{i=1}^{n} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T,$$

tedy $\boldsymbol{S} \sim W(n, \Sigma, Q)$*. Jestliže navíc* $\boldsymbol{Q} = \boldsymbol{0}$*, pak říkáme, že* $\boldsymbol{W}$ *má centrální Wishartovo rozdělení.*

Poznámka.

1. Jestliže označíme $\mathbf{W}$ jako matici, jejíž řádky tvoří vektory $\mathbf{W}_1^T, \ldots, \mathbf{W}_n^T$, pak můžeme psát, že $W \sim N(\boldsymbol{\mu}, \Sigma \otimes \mathbf{I}_n)$, kde řádky matice $\boldsymbol{\mu}$ tvoří jednotlivé vektory $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$. Dále platí $\mathbf{S} = \mathbf{W}^T\mathbf{W}$ má Wishartovo rozdělení $W(n, \Sigma, \mathbf{Q})$ popsané výše.

2. Pokud má $\mathbf{W}$ rozdělení $N(\boldsymbol{\mu}, \Sigma \otimes \mathbf{I}_n)$ a definujeme-li $\mathbf{W}^* = \mathbf{AW}$, kde $\mathbf{A}$ je čtvercová matice velikosti n, pak $\mathbf{W}^*$ má rozdělení $N(\mathbf{A}\boldsymbol{\mu}, \Sigma \otimes \mathbf{A}\mathbf{I}_n\mathbf{A}^T)$.

Věta 12 *: Uvažujme mnohorozměrný lineární model tvaru* $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ *a matice* $\boldsymbol{H} = \boldsymbol{Y}^T(\boldsymbol{M} - \boldsymbol{M}_0)\,\boldsymbol{Y}$ *a* $\boldsymbol{E} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{M})\,\boldsymbol{Y}$*. Pak za předpokladu normality (tj.* $\boldsymbol{Y}_{i\cdot} \sim N_q(\boldsymbol{\beta}^T\boldsymbol{X}_{i\cdot}, \Sigma)$ *pro* $i = 1, \ldots, n$*) a nezávislosti* $\boldsymbol{Y}_{i\cdot}$ *platí:*

1. *matice* $\boldsymbol{H}$ *a* $\boldsymbol{E}$ *mají nezávislé Wishartovo rozdělení, konkrétně*

$$\boldsymbol{E} \sim W\left(n - r(\boldsymbol{X}), \Sigma, \boldsymbol{0}\right),$$

$$\boldsymbol{H} \sim W\left(r(\boldsymbol{X}) - r(\boldsymbol{X}_0), \Sigma, \frac{1}{2}\, \Sigma^{-1}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{X}\boldsymbol{\beta}\right).$$

   Platí-li redukovaný model, pak $\boldsymbol{H} \sim W\left(r(\boldsymbol{X}) - r(\boldsymbol{X}_0), \Sigma, \boldsymbol{0}\right)$*,*

2. *matice* $\boldsymbol{H}$ *a* $\boldsymbol{E}$ *jsou nezávislé,*

3. *matice* $\boldsymbol{MY}$ *a* $\boldsymbol{E}$ *jsou nezávislé.*

Důkaz. Nejprve dokážeme část *1.* $\mathbf{Y}$ je matice, jejíž řádky tvoří nezávislé vektory $\mathbf{Y}_{1\cdot}, \ldots, \mathbf{Y}_{n\cdot}$ a platí $\mathbf{Y}_{i\cdot} \sim N_q(\boldsymbol{\beta}^T\mathbf{X}_{i\cdot}, \Sigma)$ pro $i = 1, \ldots, n$. Tedy matice $\mathbf{Y}$ má rozdělení $N(\mathbf{X}\boldsymbol{\beta}, \Sigma \otimes \mathbf{I}_n)$. Nejprve dokážeme tvrzení věty pro matici $\mathbf{E}$.

Matice $(\mathbf{I}_n - \mathbf{M})$ je idempotentní, neboť platí $(\mathbf{I}_n - \mathbf{M})(\mathbf{I}_n - \mathbf{M}) = \mathbf{I}_n - \mathbf{M} - \mathbf{M} + \mathbf{M}\mathbf{M} = (\mathbf{I}_n - \mathbf{M})$. Dále označme hodnost této matice jako k. Platí $k = r(\mathbf{I}_n - \mathbf{M}) = tr(\mathbf{I}_n - \mathbf{M}) = tr(\mathbf{I}_n) - tr(\mathbf{M}) = n - r(\mathbf{M}) = n - r(\mathbf{X})$. Označíme-li $\mathbf{S} = (\mathbf{I}_n - \mathbf{M})\mathbf{Y}$, pak matici $\mathbf{E}$ můžeme spát ve tvaru

$$\mathbf{E} = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{M})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{M})(\mathbf{I}_n - \mathbf{M})\mathbf{Y} = \mathbf{S}^T\mathbf{S},$$

kde $\mathbf{S} \sim N((\mathbf{I}_n - \mathbf{M})\mathbf{X}\boldsymbol{\beta}, \Sigma \otimes (\mathbf{I}_n - \mathbf{M})\mathbf{I}_n(\mathbf{I}_n - \mathbf{M})) = N(\mathbf{0}, \Sigma \otimes (\mathbf{I}_n - \mathbf{M}))$, neboť $(\mathbf{I}_n - \mathbf{M})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. Dále rozepíšeme $\mathbf{E}$ následovně

$$\mathbf{E} = \mathbf{S}^T\mathbf{S} = \mathbf{S}^T(\mathbf{I}_n - \mathbf{M})\mathbf{S} + \mathbf{S}^T(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{M}))\mathbf{S} = \mathbf{S}^T(\mathbf{I}_n - \mathbf{M})\mathbf{S} + \mathbf{S}^T\mathbf{M}\mathbf{S}.$$

Matice $\mathbf{M}\mathbf{S}$ má rozdělení $N(\mathbf{0}, \Sigma \otimes \mathbf{0})$, neboť $\mathbf{M}(\mathbf{I}_n - \mathbf{M})\mathbf{M} = (\mathbf{M} - \mathbf{M}\mathbf{M})\mathbf{M} = \mathbf{0}$. A tedy $\mathbf{S}^T\mathbf{M}\mathbf{S}$ je rovno $\mathbf{0}$ s pravděpodobností jedna. Zbývá dokázat, že pokud $\mathbf{S} \sim N(\mathbf{0}, \Sigma \otimes (\mathbf{I}_n - \mathbf{M}))$, pak $\mathbf{E} = \mathbf{S}^T(\mathbf{I}_n - \mathbf{M})\mathbf{S}$ má rozdělení $W(n - r(\mathbf{X}), \Sigma, \mathbf{0}) = W(k, \Sigma, \mathbf{0})$. Matice $(\mathbf{I}_n - \mathbf{M})$ má hodnost $k \leq n$, tudíž konečným počtem lineárních kombinací na řádky této matice ji lze převést na matici, jejíž prvních k řádků budou tvořit vektory lineárně nezávislé a $(k+1)$-ní až n-tý řádek budou tvořit nulové vektory. Provádět lineární kombinace na řádky matice $(\mathbf{I}_n - \mathbf{M})$ znamená násobit ji nějakou maticí, označme ji $\mathbf{B}$, zleva (tedy $\mathbf{B}(\mathbf{I}_n - \mathbf{M})$). Matice $(\mathbf{I}_n - \mathbf{M})$ je však idempotentní a symetrická. Provedeme-li stejné lineární kombinace, ale tentokrát na sloupce (tedy $(\mathbf{I}_n - \mathbf{M})\mathbf{B}^T$), převedeme ji na matici, jejíž prvních k sloupců budou tvořit vektory lineárně nezávislé a $(k+1)$-ní až n-tý sloupec vektory nulové. Odtud plyne, že existuje nějaká regulární matice $\mathbf{C}$ taková, že

$$\mathbf{C}(\mathbf{I}_n - \mathbf{M})\mathbf{C}^T = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{n \times n},$$

označme matici na pravé straně předchozí rovnosti symbolem $\mathbf{I}_{k,n}$. Tedy matici $(\mathbf{I}_n - \mathbf{M})$ můžeme přepsat do tvaru $(\mathbf{I}_n - \mathbf{M}) = \mathbf{C}^{-1}\mathbf{I}_{k,n}(\mathbf{C}^{-1})^T$. Z toho, že matice $(\mathbf{I}_n - \mathbf{M})$ je idempotentní plyne, že

$$\mathbf{I}_{k,n}(\mathbf{C}^{-1})^T(\mathbf{C}^{-1})\mathbf{I}_{k,n} = \mathbf{I}_{k,n}$$

a tedy

$$(\mathbf{C}^{-1})^T(\mathbf{C}^{-1})\mathbf{I}_{k,n}(\mathbf{C}^{-1})^T(\mathbf{C}^{-1}) = \mathbf{I}_{k,n}. \tag{2.17}$$

Nyní provedeme substituci $\mathbf{S}^* = \mathbf{CS}$. Pak $\mathbf{S}^*$ má rozdělení $N(\mathbf{0}, \Sigma \otimes \mathbf{I}_{k,n})$. A tedy

$$\mathbf{E} = \mathbf{S}^T(\mathbf{I}_n - \mathbf{M})\mathbf{S} = \mathbf{S}^T\mathbf{C}^{-1}\mathbf{I}_{k,n}(\mathbf{C}^{-1})^T\mathbf{S}$$
$$= (\mathbf{S}^*)^T(\mathbf{C}^{-1})^T\mathbf{C}^{-1}\mathbf{I}_{k,n}(\mathbf{C}^{-1})^T\mathbf{C}^{-1}\mathbf{S}^* = (\mathbf{S}^*)^T\mathbf{I}_{k,n}\mathbf{S}^*$$

s využitím vztahu (2.17). Rozdělíme matici $\mathbf{S}^*$ velikosti $n \times q$ na matice $\mathbf{S}_1^*$ velikosti $k \times q$ a $\mathbf{S}_2^*$ velikosti $(n-k) \times q$. Pak $\mathbf{E} = (\mathbf{S}_1^*)^T(\mathbf{S}_1^*)$ a $\mathbf{S}_1^* \sim N(\mathbf{0}, \Sigma \otimes \mathbf{I}_k)$. Podle definice Wishartova rozdělení platí $\mathbf{E} \sim W(k, \Sigma, \mathbf{0})$.

Tvrzení o rozdělení matice $\mathbf{H}$ se dokáže podobně. Nejprve ukážeme, že matice $(\mathbf{M} - \mathbf{M}_0)$ je idempotentní.

$$(\mathbf{M} - \mathbf{M}_0)(\mathbf{M} - \mathbf{M}_0) = \mathbf{MM} - \mathbf{M}_0\mathbf{M} - \mathbf{MM}_0 + \mathbf{M}_0\mathbf{M}_0 =$$

$$\mathbf{M} - \mathbf{M}_0\mathbf{M} - \mathbf{MM}_0 + \mathbf{M}_0 = \mathbf{M} - \mathbf{M}_0 \Leftrightarrow \mathbf{M}_0\mathbf{M} + \mathbf{MM}_0 - \mathbf{M}_0 = \mathbf{M}_0.$$

Pronásobením poslední rovnice maticí $\mathbf{M}$ zprava a následně zleva dostaneme požadovanou rovnost. Stejně tak jako v předchozím důkazu označme hodnost matice $(\mathbf{M} - \mathbf{M}_0)$ jako k. Pak platí $k = r(\mathbf{M} - \mathbf{M}_0) = tr(\mathbf{M} - \mathbf{M}_0)$ $= tr(\mathbf{M}) - tr(\mathbf{M}_0) = r(\mathbf{M}) - r(\mathbf{M}_0) = r(\mathbf{X}) - r(\mathbf{X}_0)$. Označíme-li $\mathbf{S}$ jako matici $(\mathbf{M} - \mathbf{M}_0)\mathbf{Y}$ pak zapíšeme $\mathbf{H}$ ve tvaru

$$\mathbf{H} = \mathbf{Y}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} = \mathbf{Y}^T(\mathbf{M} - \mathbf{M}_0)(\mathbf{M} - \mathbf{M}_0)\mathbf{Y} = \mathbf{S}^T\mathbf{S}$$

a platí $\mathbf{S} \sim N((\mathbf{M} - \mathbf{M}_0)\mathbf{X}\boldsymbol{\beta}, \Sigma \otimes (\mathbf{M} - \mathbf{M}_0))$. Pak

$$\mathbf{H} = \mathbf{S}^T\mathbf{S} = \mathbf{S}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{S} + \mathbf{S}^T(I - (\mathbf{M} - \mathbf{M}_0))\mathbf{S}$$

a stejně jako v předchozím důkazu je $\mathbf{S}^T(I - (\mathbf{M} - \mathbf{M}_0))\mathbf{S}$ rovno $\mathbf{0}$ s pravděpodobností jedna, neboť $(I - (\mathbf{M} - \mathbf{M}_0))(\mathbf{M} - \mathbf{M}_0)\mathbf{X}\boldsymbol{\beta} = (\mathbf{M} - \mathbf{M}_0)\mathbf{X}\boldsymbol{\beta}$ $-(\mathbf{M}-\mathbf{M}_0)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ a $(I-(\mathbf{M}-\mathbf{M}_0))(\mathbf{M}-\mathbf{M}_0)(I-(\mathbf{M}-\mathbf{M}_0)) = ((\mathbf{M}-\mathbf{M}_0)$ $- (\mathbf{M} - \mathbf{M}_0))(\mathbf{M} - \mathbf{M}_0) = \mathbf{0}$ a tedy $(I - (\mathbf{M} - \mathbf{M}_0))\mathbf{S} \sim N(\mathbf{0}, \Sigma \otimes \mathbf{0})$. Zbývá dokázat, že pokud $\mathbf{S} \sim N((\mathbf{M} - \mathbf{M}_0)\mathbf{X}\boldsymbol{\beta}, \Sigma \otimes (\mathbf{M} - \mathbf{M}_0))$, pak $\mathbf{H} = \mathbf{S}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{S}$ má Wishartovo rozdělení $W(k, \Sigma, \mathbf{Q})$, kde $\mathbf{Q} = \frac{1}{2}\,\Sigma^{-1}\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{M} - \mathbf{M}_0)\mathbf{X}\boldsymbol{\beta}$. Stejně jako v předchozím důkazu existuje regulární matice $\mathbf{C}$ taková, že

$$\mathbf{C}(\mathbf{M} - \mathbf{M}_0)\mathbf{C}^T = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{n \times n}.$$

A opět provedeme substituci $\mathbf{S}^* = \mathbf{CS}$ a rozložíme matici $\mathbf{S}^*$ velikosti $n \times q$ na matice $\mathbf{S}_1^*$ velikosti $k \times q$ a $\mathbf{S}_2^*$ velikosti $(n - k) \times q$. Pak $\mathbf{H} = (\mathbf{S}_1^*)^T\mathbf{S}_1^*$

a $\mathbf{S}_1^* \sim N((\mathbf{M}-\mathbf{M}_0)\mathbf{X}\boldsymbol{\beta}, \Sigma \otimes \mathbf{I}_k)$. Podle definice Wishartova rozdělení platí $\mathbf{H} \sim W(k, \Sigma, \mathbf{D})$. Zbývá tedy už jen dopočítat centralitu Wishartova rozdělení. Podle definice tohoto rozdělení platí

$$\mathbf{D} = \frac{1}{2}\,\Sigma^{-1/2}((\mathbf{M}-\mathbf{M}_0)\mathbf{X}\boldsymbol{\beta})^T(\mathbf{M}-\mathbf{M}_0)\mathbf{X}\boldsymbol{\beta} = \mathbf{Q}.$$

Za platnosti nulové hypotézy je matice $\mathbf{M}-\mathbf{M}_0$ rovna $\mathbf{0}$ a zřejmě matice $\mathbf{H}$ má Wishartovo rozdělení $W\left(r(\mathbf{X}) - r(\mathbf{X}_0), \Sigma, 0\right)$.

K důkazu bodu *2.* stačí ukázat, že $(\mathbf{I}-\mathbf{M})\mathbf{Y}$ a $(\mathbf{M}-\mathbf{M}_0)\mathbf{Y}$ jsou nezávislé. Rozdělení vektorů $\mathbf{Y}_{i\cdot}$ je normální pro každé $i = 1, \ldots, n$, takže k nezávislosti $(\mathbf{I}-\mathbf{M})\mathbf{Y}$ a $(\mathbf{M}-\mathbf{M}_0)\mathbf{Y}$ stačí ukázat, že $cov((\mathbf{I}-\mathbf{M})\mathbf{Y}, (\mathbf{M}-\mathbf{M}_0)\mathbf{Y}) = \mathbf{0}$.

$$\begin{aligned} cov((\mathbf{I}-\mathbf{M})\mathbf{Y}_{i\cdot}, (\mathbf{M}-\mathbf{M}_0)\mathbf{Y}_{i\cdot}) &= (\mathbf{I}-\mathbf{M})var(\mathbf{Y}_{i\cdot})(\mathbf{M}-\mathbf{M}_0)^T \\ &= (\mathbf{I}-\mathbf{M})\Sigma(\mathbf{M}-\mathbf{M}_0) \\ &= \Sigma\mathbf{M} - \Sigma\mathbf{M}_0 - \mathbf{M}\Sigma\mathbf{M} + \mathbf{M}\Sigma\mathbf{M}_0. \end{aligned}$$

Pronásobíme-li poslední rovnost maticí $\mathbf{M}$ zleva dostáváme

$$\mathbf{M}\Sigma\mathbf{M} - \mathbf{M}\Sigma\mathbf{M}_0 - \mathbf{M}\mathbf{M}\Sigma\mathbf{M} + \mathbf{M}\mathbf{M}\Sigma\mathbf{M}_0 = \mathbf{0}.$$

A dále $cov((\mathbf{I}-\mathbf{M})\mathbf{Y}_{i\cdot}, (\mathbf{M}-\mathbf{M}_0)\mathbf{Y}_{j\cdot}) = \mathbf{0}$ pro $i \neq j$ plyne z nezávislosti vektorů $\mathbf{Y}_{1\cdot}, \ldots, \mathbf{Y}_{n\cdot}$.

Důkaz bodu *3.* provedeme podobně jako v bodě *2.* K nezávislosti matic $\mathbf{M}\mathbf{Y}$ a $\mathbf{E}$ stačí ukázat nezávislost matic $\mathbf{M}\mathbf{Y}$ a $(\mathbf{I}-\mathbf{M})\mathbf{Y}$. Rozdělení vektorů $\mathbf{Y}_{i\cdot}$ je normální pro každé $i = 1, \ldots, n$, tedy stačí ukázat, že $cov(\mathbf{M}\mathbf{Y}, (\mathbf{I}-\mathbf{M})\mathbf{Y}) = \mathbf{0}$.

$$\begin{aligned} cov(\mathbf{M}\mathbf{Y}_{i\cdot}, (\mathbf{I}-\mathbf{M})\mathbf{Y}_{i\cdot}) &= \mathbf{M}var(\mathbf{Y}_{i\cdot})(\mathbf{I}-\mathbf{M})^T \\ &= \mathbf{M}\Sigma - \mathbf{M}\Sigma\mathbf{M}. \end{aligned}$$

Pronásobíme-li poslední rovnost maticí $\mathbf{M}$ zleva, dostaneme

$$\mathbf{M}\Sigma\mathbf{M} - \mathbf{M}\mathbf{M}\Sigma\mathbf{M} = \mathbf{0}.$$

A stejně jako v předchozím důkazu $cov(\mathbf{M}\mathbf{Y}_{i\cdot}, (\mathbf{I}-\mathbf{M})\mathbf{Y}_{j\cdot}) = \mathbf{0}$ pro $i \neq j$. Odtud již nezávislost matic $\mathbf{M}\mathbf{Y}$ a $\mathbf{E}$ plyne. □

Platnost hypotézy (2.15) můžeme otestovat pomocí testové statistiky založené na poměru věrohodností, přesněji poměru maxima věrohodnostní funkce redukovaného modelu a globálního maxima věrohodnostní funkce. Maximální hodnota věrohodnostní funkce je

$$L(\mathbf{Y}, \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\Sigma}) = 2\pi^{-nq/2}|\hat{\Sigma}|^{-n/2}e^{-nq/2},$$

kde $\hat{\boldsymbol{\beta}}$ je odhad matice parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců a $\hat{\Sigma}$ je matice popsaná v tvrzení 10. Předpokládáme-li platnost redukovaného modelu, pak pro odhady matic $\boldsymbol{\Gamma}$ a Σ_H metodou maximální věrohodnosti platí $\hat{\boldsymbol{\Gamma}} = (\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{Y}$ a $\hat{\Sigma}_H = \mathbf{Y}^T(\mathbf{I}-\mathbf{M}_0)\mathbf{Y}/n$. Maximální hodnota věrohodnostní funkce za předpokladu nulové hypotézy je

$$L(\mathbf{Y}, \mathbf{X}_0\hat{\boldsymbol{\Gamma}}, \hat{\Sigma}_H) = 2\pi^{-nq/2}|\hat{\Sigma}_H|^{-n/2}e^{-nq/2}.$$

Testová statistika založená na poměru věrohodností má tvar

$$\frac{L(\mathbf{Y}, \mathbf{X}_0\hat{\boldsymbol{\Gamma}}, \hat{\Sigma}_H)}{L(\mathbf{Y}, \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_H|}\right)^{n/2}.$$

Uvědomíme-li si, že $\hat{\Sigma} = \mathbf{E}/n$, $\hat{\Sigma}_H = (\mathbf{E}+\mathbf{H})/n$ a funkce $f(x) = x^{2/n}$ je ryze rostoucí, pak můžeme tuto testovou statistiku psát ekvivalentně ve tvaru

$$U = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = |\mathbf{I}+\mathbf{H}\mathbf{E}^{-1}|^{-1}.$$

(Podrobnější odvození lze nalézt např. v [3]).

Za platnosti nulové hypotézy má U nějaké rozdělení, označme ho $U(q, d, n-r(\mathbf{X}))$, kde d je buď $r(\mathbf{X})-r(\mathbf{X}_0)$, je-li $\mathbf{H}$ tvaru $\mathbf{Y}^T(\mathbf{M}-\mathbf{M}_0)\mathbf{Y}$, nebo $r(\boldsymbol{\Lambda})$, je-li $\mathbf{H}$ tvaru $\mathbf{Y}^T\mathbf{M}_{MP}\mathbf{Y}$. Test hypotézy na hladině α je zamítnut, jestliže napozorovaná hodnota U je menší než α-kvantil rozdělení $U(q, d, n-r(\mathbf{X}))$.

V [4] je stanovena následující přibližná aproximace pro distribuční funkci U

za platnosti nulové hypotézy. Nechť

$$\begin{aligned} r &= r(\mathbf{X}), \\ d &= r(\mathbf{X}) - r(\mathbf{X}_0) = r(\mathbf{\Lambda}), \\ s &= \frac{qd}{2} - 1, \\ f &= (n-r) + d - \frac{1}{2}(d+q+1), \\ t &= \begin{cases} [(q^2d^2-4)/(q^2+d^2-5)]^{1/2} & \text{pokud } \min(q,d) \geq 2 \\ 1 & \text{pokud } \min(q,d) = 1 \end{cases}, \end{aligned}$$

pak platí přibližně

$$\frac{1-U^{1/t}}{U^{1/t}} \cdot \frac{ft-s}{qd} \sim F(qd, ft-s)$$

a hypotézu zamítáme, jestliže

$$\frac{1-U^{1/t}}{U^{1/t}} \cdot \frac{ft-s}{qd} \geq F_{(qd,ft-s)}(1-\alpha).$$

Poznámka. Je-li $\min(q,d)$ rovno 1 nebo 2, pak je toto rozdělení přesné.

Mezi další dvě známé testové statistiky patří Lawley-Hotellingova stopa

$$T^2 = (n - r(\mathbf{X}))\text{tr}\left(\mathbf{H}\mathbf{E}^{-1}\right) = \text{tr}\left(\mathbf{H}\mathbf{S}^{-1}\right)$$

a Pillaiova stopa

$$V = \text{tr}\left(\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}\right).$$

Přesné hodnoty jejich rozdělení lze nalézt v tabulkách, existují ale i jejich aproximace. Jedna taková je navržena v [5]. Nechť $d = r(\mathbf{X}) - r(\mathbf{X}_0) = r(\mathbf{\Lambda})$ a $n - r(\mathbf{X}) = n - r$, pak za platnosti nulové hypotézy H_0 má přibližně

$$GT^2 \sim F(qd, D),$$

kde

$$\begin{aligned} D &= 4 + \frac{qd+2}{B-1}, \\ B &= \frac{(n-r+d-q-1)(n-r-1)}{(n-r-q-3)(n-r-q)}, \\ G &= (qd)^{-1}\left(\frac{D}{D-2}\right)\left(\frac{n-r-q-1}{n-r}\right). \end{aligned}$$

Pro velké n má veličina T^2 přiblížné asymptotické rozdělní $T^2 \sim \chi^2(qd)$.

Poznámka. je-li $\min(q, d) = 1$, pak je toto rozdělení přesné.

Za platnosti nulové hypotézy H_0 má Pillaiova stopa přibližné rozdělení

$$\frac{n-r-q+s}{|q-d|+s} \cdot \frac{V}{s-V} \sim F(s(|q-d|+s), s(n-r-q+s)),$$

kde

$$s = \min(q, d).$$

Pro asymptotickou aproximaci platí $(n-r)V \sim \chi^2(qd)$.

Poznámka. Stupně volnosti v jednotlivých rozděleních musí být vždy větší než 0. Proto musíme volit dostatečně velký počet pozorování n tak, aby tato podmínka byla splněna.

Všechny výše uvedené testové statistiky platí za předpokladu normality. Pokud tento předpoklad není splněn, pak tyto testové statistiky můžeme taktéž použít jako hrubé aproximace, neboť i bez předpokladu normality platí

$$\begin{aligned} \mathrm{E}\left(\mathbf{E}/(n-r(\mathbf{X}))\right) &= \Sigma, \\ \mathrm{E}\left(\mathbf{H}/r(\mathbf{M}-\mathbf{M}_0)\right) &= \Sigma \end{aligned}$$

a můžeme je tedy použít při výpočtech hodnot testových statistik.

2.3.2 Predikční oblasti

Předpokládejme, že chceme nějakým způsobem odhadovat hodnoty nového pozorování vektoru $\mathbf{Y}_0^T$ na základě námi předem zvolených hodnot vektoru $\mathbf{X}_0 = (X_{01}, \ldots, X_{0q})$. Je zřejmé, že vektor $\mathbf{Y}_0$ je generován stejným procesem jako matice $\mathbf{Y}$, proto $var(\mathbf{Y}_0) = \Sigma$ a $\mathbf{Y}_0$ je nezávislý na $\mathbf{Y}$. Z předchozí teorie plyne, že nejlepší (lineární) nestranný odhad vektoru $\mathbf{Y}_0$ je $\hat{\mathbf{Y}}_0 = \hat{\boldsymbol{\beta}}^T\mathbf{X}_0$, kde složky vektoru $\mathbf{X}_0$ jsou nějaké, námi předem zvolené, hodnoty. Oblast spolehlivosti pro $\mathbf{Y}_0$ může pak být založena na distribuční funkci $\mathbf{Y}_0 - \hat{\mathbf{Y}}_0$. Lze totiž ukázat, že $E(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0) = \mathbf{0}$ a $var(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0) = (1 + \mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0)\Sigma$ (celé odvození lze nalézt např. v [3]). Pokud $\mathbf{Y}_0$ a $\mathbf{Y}$ mají mnohorozměrné

normální rozdělení, pak i $\mathbf{Y}_0 - \hat{\mathbf{Y}}_0$ má mnohorozměrné normální rozdělení $N(\mathbf{0}, (1 + \mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0)\Sigma)$. Aplikací definice 11 (za n v této definice dosazujeme hodnotu 1) dostáváme

$$(1 + \mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0)^{-1}(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)^T \sim W(1, \Sigma, \mathbf{0}).$$

Z toho plyne, že náhodná veličina

$$\text{tr}\left(\frac{(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)^T}{(1 + \mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0)}\,\mathbf{S}^{-1}\right)$$

má stejnou distribuční funkci jako Lawley - Hotellingova testová statistika T^2, kde parametr $d = 1$. Z McKeonovy aproximace plyne přesné rozdělní této náhodné veličiny, tedy

$$\frac{(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)^T\mathbf{S}^{-1}(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)}{(1 + \mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0)}\frac{n - r(\mathbf{X}) - q + 1}{q(n - r(\mathbf{X}))} \sim F(q, n - r(\mathbf{X}) - q + 1).$$

Pak tedy $(1 - \alpha)\%$ oblast spolehlivosti je tvořena všemi vektory $\mathbf{Y}_0$ splňujícími nerovnost

$$(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0)^T\mathbf{S}^{-1}(\mathbf{Y}_0 - \hat{\mathbf{Y}}_0) \leq F_{q,n-r-q+1}(1-\alpha)\frac{q(n-r)}{n-r-q+1}(1+\mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0).$$

2.4 Příklad

Mějme speciální mnohorozměrný lineární model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{e}$, kde $q = p = 2$. Maticový zápis tohoto modelu bude tvaru

$$\begin{pmatrix} Y_{11} & Y_{12} \\ \vdots & \vdots \\ Y_{n1} & Y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} + \begin{pmatrix} e_{11} & e_{12} \\ \vdots & \vdots \\ e_{n1} & e_{n2} \end{pmatrix}.$$

Za předpokladu normality pro sloupcové vektory $\mathbf{e}_{\cdot 1}, \mathbf{e}_{\cdot 2}$ matice $\mathbf{e}$ platí $E\mathbf{e}_{\cdot 1}$ $= E\mathbf{e}_{\cdot 2} = \mathbf{0}$ a $var\mathbf{e}_{\cdot 1} = \sigma_{11}\mathbf{I}_n$, $var\mathbf{e}_{\cdot 2} = \sigma_{22}\mathbf{I}_n$, $(cov(\mathbf{e}_{\cdot 1}, \mathbf{e}_{\cdot 2}) = cov(\mathbf{e}_{\cdot 2}, \mathbf{e}_{\cdot 1})$ $= \sigma_{12}\mathbf{I}_n)$. Pro odhad matice parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců platí podle věty 9 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Pro matice $\mathbf{X}^T\mathbf{X}$ a $\mathbf{X}^T\mathbf{Y}$ platí

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_{i1} & \sum_{i=1}^n Y_{i2} \\ \sum_{i=1}^n x_i Y_{i1} & \sum_{i=1}^n x_i Y_{i2} \end{pmatrix}.$$

Označme

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij} \qquad \text{pro } j = 1, 2.$$

Pro jednotlivé sloupce matice $\mathbf{Y}$ platí

$$\mathbf{Y}_{\cdot j} = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \end{pmatrix} \mathbf{X} + \mathbf{e}_{\cdot j}, \qquad \text{pro } j = 1, 2,$$

což můžeme ještě po složkách rozepsat jako

$$Y_{ij} = \beta_{1j} + \beta_{2j} x_i + e_{ij}, \qquad \text{pro } i = 1, \ldots, n, \qquad \text{pro } j = 1, 2.$$

Nyní si stačí uvědomit, že pro vektory $\mathbf{Y}_{\cdot 1}, \mathbf{Y}_{\cdot 2}$ platí jednorozměrný lineární model, speciálně ve formě obecné přímky. Proto na jednotlivé složky matice parametrů $\boldsymbol{\beta}$ můžeme aplikovat vzorce (1.7) odvozené v kapitole 1.3, tedy bude platit

$$\hat{\beta}_{21} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_{i1} - \overline{Y}_1)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_{22} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_{i2} - \overline{Y}_2)}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

$$\hat{\beta}_{11} = \overline{Y}_1 - \hat{\beta}_{21}\overline{x}, \qquad \hat{\beta}_{12} = \overline{Y}_2 - \hat{\beta}_{22}\overline{x}.$$

Pro nestranný odhad $\mathbf{S}$ matice Σ podle tvrzení 10 platí

$$\mathbf{S} = \frac{1}{n - r(\mathbf{X})}\ \mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y} = \frac{1}{n - r(\mathbf{X})}\ \mathbf{Y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y} =$$

$$= \frac{1}{n - r(\mathbf{X})}\ \mathbf{Y}^T\mathbf{Y} - \frac{1}{n - r(\mathbf{X})}\ \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

a pro jednotlivé složky této matice platí

$$\sigma_{ij} = \frac{1}{n-2}\left(\sum_{k=1}^{n} Y_{ki}Y_{kj} - \hat{\beta_{1j}}\sum_{k=1}^{n} Y_{ki} - \hat{\beta_{2j}}\sum_{k=1}^{n} Y_{ki}x_k\right), \qquad \text{pro } i, j = 1, 2.$$

Poznámka.

1. Matice $\mathbf{S}$ je symetrická, platí totiž $\sigma_{12} = \sigma_{21}$.

2. Složky σ_{11} a σ_{22} matice $\mathbf{S}$ mají stejný tvar jako reziduální rozptyl příslušných jednorozměrných lineárních modelů.

Uvažujme testování hypotézy toho, že veličina $\mathbf{Y}$ vůbec na hodnotách $\mathbf{X}_{.2}$ nezávisí, tj. $\beta_{21} = \beta_{22} = 0$. Zvolíme-li $\mathbf{\Lambda}$ jako

$$\mathbf{\Lambda} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

pak test hypotézy $H_0 : \mathbf{\Lambda}^T\boldsymbol{\beta} = 0$ proti hypotéze $H_1 : \mathbf{\Lambda}^T\boldsymbol{\beta} \neq 0$ odpovídá přesně hypotéze $\beta_{21} = \beta_{22} = 0$, tedy že model na $x_1, \ldots, x_n$ nezávisí.

Spočtěme nyní matice $\mathbf{H}$ a $\mathbf{E}$.

$$\begin{aligned} \mathbf{H} &= (\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}})^T[\mathbf{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}]^{-1}(\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}}) \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \begin{pmatrix} \hat{\beta}_{21}^2 & \hat{\beta}_{21}\hat{\beta}_{22} \\ \hat{\beta}_{21}\hat{\beta}_{22} & \hat{\beta}_{22}^2 \end{pmatrix} \\ \mathbf{E} &= \mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y} = (n-2)\hat{\Sigma}^* = (n-2)\mathbf{S}. \end{aligned}$$

Testová statistika založená na poměru věrohodností má tvar

$$\begin{aligned} U &= |\mathbf{I} + \mathbf{H}\mathbf{E}^{-1}|^{-1} \\ &= \left|\mathbf{I} + \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-2)} \begin{pmatrix} \hat{\beta}_{21}^2 & \hat{\beta}_{21}\hat{\beta}_{22} \\ \hat{\beta}_{21}\hat{\beta}_{22} & \hat{\beta}_{22}^2 \end{pmatrix} \mathbf{S}^{-1}\right|^{-1} \end{aligned}$$

Dále v tomto modelu platí $p = q = r = 2$ a $d = r(\mathbf{\Lambda}) = 1$. Můžeme tedy vypočítat parametry s a f, které určují rozdělení náhodné veličiny U. Platí

$$\begin{aligned} s &= qd/2 - 1 = 0 \\ f &= (n-r) + d - (d+q+1)/2 = n - 3. \end{aligned}$$

Dále je vidět, že parametr $t = 1$, neboť $\min(q, d) = 1$ a rozdělení náhodné veličiny U bude přesné. Za platnosti nulové hypotézy platí

$$\frac{1-U}{U} \cdot \frac{n-3}{2} \sim F(2, n-3)$$

a hypotézu zamítáme, je-li

$$\frac{1-U}{U} \cdot \frac{n-3}{2} \geq F_{(2,n-3)}(1-\alpha).$$

Poznámka. Z rozdělení výše uvedené náhodné veličiny je vidět, že počet pozorování n musí být větší než 3.

Stejnou hypotézu můžeme testovat také pomocí Lawley-Hotellingovy a Pillaiovy stopy. Ty jsou tvaru

$$\begin{aligned} T^2 &= (n - r(\mathbf{X})))\mathrm{tr}\left(\mathbf{H}\mathbf{E}^{-1}\right) = \mathrm{tr}\left(\mathbf{H}\mathbf{S}^{-1}\right) \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \ \mathrm{tr}\left(\begin{pmatrix} \hat{\beta}_{21}^2 & \hat{\beta}_{21}\hat{\beta}_{22} \\ \hat{\beta}_{21}\hat{\beta}_{22} & \hat{\beta}_{22}^2 \end{pmatrix}\mathbf{S}^{-1}\right), \\ V &= \mathrm{tr}\left(\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}\right) \\ &= \mathrm{tr}\left(\mathbf{H}((n-2)\mathbf{S}+\mathbf{H})^{-1}\right). \end{aligned}$$

Dále spočteme parametry B, D, G

$$\begin{aligned} B &= \frac{(n-r+d-q-1)(n-r-1)}{(n-r-q-3)(n-r-q)} = \frac{n-3}{n-7}, \\ D &= 4 + \frac{qd+2}{B-1} = n-3, \\ G &= (qd)^{-1}\left(\frac{D}{D-2}\right)\left(\frac{n-r-q-1}{n-r}\right) = \frac{n-3}{2(n-2)}. \end{aligned}$$

Rozdělení náhodné veličiny T^2 bude opět přesné (neboť $\min(q,d) = 1$) a bude tvaru

$$\frac{n-3}{2(n-2)}\ T^2 \sim F(2, n-3).$$

Za platnosti nulové hypotézy má Pillaiova stopa přibližné rozdělení

$$\frac{n-3}{2} \cdot \frac{V}{1-V} \sim F(2, n-3).$$

V obou dvou těchto případech nulovou hypotézu zamítáme pro velké hodnoty testových statistik.

2.4.1 Simulace

Všechny výše uvedené testové statistiky mohou danou hypotézu otestovat různě. Abychom je porovnaly mezi sebou, sestrojili jsme v programu R simulace (tento program je přiložen k bakalářské práci na kompaktním disku).

Pro zjednodušení jsme uvažovali pouze dvourozměrný lineární regresní model a k simulacím jsme využili předchozího příkladu. Pro daný dvourozměrný lineární model jsme si spočetli všechny výše uvedené testové statistiky, tj. založené na poměru věrohodností, založené na Lawley-Hottelingově stopě a Pillaiově stopě, a přidali jsme ještě jednu testovou statistiku využívající jednorozměrnou lineární regresi. Zvolili jsme běžně používanou hladinu $\alpha = 5\%$ a na této hladině jsme testovali nulovou hypotézu $H_0 : \boldsymbol{\Lambda^T \beta} = \mathbf{0}$, kde $\boldsymbol{\Lambda^T} = (0, 1)^T$. Celý tento proces jsme opakovali deset tisíckrát a sledovali jsme, kolikrát je nulová hypotéza zamítnuta. Podíl zamítnutých nulových hypotéz a počtu opakování nám pak dal síly a odhady významnosti jednotlivých testů a umožnil nám tak jejich porovnání.

Při sestrojování simulace jsme postupovali následovně. Nejdříve jsme si zvolili parametr n, tedy počet pozorování, parametry σ a $\beta = \beta_{21} = \beta_{22}$ tak, že matice Σ a $\boldsymbol{\beta}$ byly tvaru

$$\Sigma = \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} 1 & 1 \\ \beta & \beta \end{pmatrix}.$$

Dále jsme si vygenerovali matici $\mathbf{X}$ velikosti $n \times 2$, jejíž první sloupec tvoří vektor jedniček a druhý sloupec náhodné hodnoty z rovnoměrného rozdělení na předem určeném intervalu $[0, a]$, kde parametr a jsme vhodně měnili v závislosti na parametru n. Také jsme si vygenerovali matici chyb $\mathbf{e}$ velikosti $n \times 2$, pro jejíž sloupcové vektory $\mathbf{e}_{\cdot 1}, \mathbf{e}_{\cdot 2}$ platí $E\mathbf{e}_{\cdot 1} = E\mathbf{e}_{\cdot 2} = \mathbf{0}$ a $var\mathbf{e}_{\cdot 1} = var\mathbf{e}_{\cdot 2} = \mathbf{I}_n$, $cov(\mathbf{e}_{\cdot 1}, \mathbf{e}_{\cdot 2}) = \sigma \mathbf{I}_n$. Na základě těchto dat jsme spočetli hodnoty matice $\mathbf{Y}$ pomocí vztahu $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

Matici $\mathbf{Y}$ jsme prohlásili za matici napozorovaných hodnot vektorů $\mathbf{Y}_{\cdot 1}$ a $\mathbf{Y}_{\cdot 2}$ a nadále jsme pracovali pouze s ní a maticí $\mathbf{X}$. Testovali jsme hypotézu $H_0 : \beta = 0$ (tj. hypotézu $\beta_{21} = \beta_{22} = 0$, což znamená, že hodnoty matice $\mathbf{Y}$ nezávisí na hodnotách $\mathbf{X}_{\cdot 2}$) proti alternativě $H_1 : \beta \neq 0$ pomocí všech výše uvedených testů. Tyto testy dále budeme značit U, T^2, V, kde U je test založený na poměru věrohodností, T^2 test založený na Lawley-Hotellingově stopě a V test založený na Pillaiově stopě.

K těmto třem testům jsme přidali ještě jeden, využívající jednorozměrný lineární model, který nám taktéž umožnil testovat hypotézu $H_0 : \beta = 0$. Uvažujme jednorozměrný lineární model tvaru

$$Y_{ij} = \beta_{1j} + \beta_{2j} x_i + e_{ij}, \qquad \text{pro } i = 1, \ldots, n, \qquad \text{pro } j = 1, 2.$$

Pro $j = 1$ a $j = 2$ jsme testovali hypotézu $H_0^* : \beta_{2j} = 0$ proti alternativě $H_1^* : \beta_{2j} \neq 0$. Zamítáme-li pak aspoň jednu nulovou hypotézu z těchto dvou, je tento test ekvivalentní testu $H_0 : \beta = 0$ popsanému v předchozím odstavci. K testování těchto nulových hypotéz využijeme vzorce z tvrzení 4. Testová statistika pro jednotlivé jednorozměrné modely je pak tvaru

$$J_j = \frac{\widehat{\beta}_{2j}}{\sqrt{\hat{\sigma}_{jj} v_{22}}}, \qquad \text{pro } j = 1, 2,$$

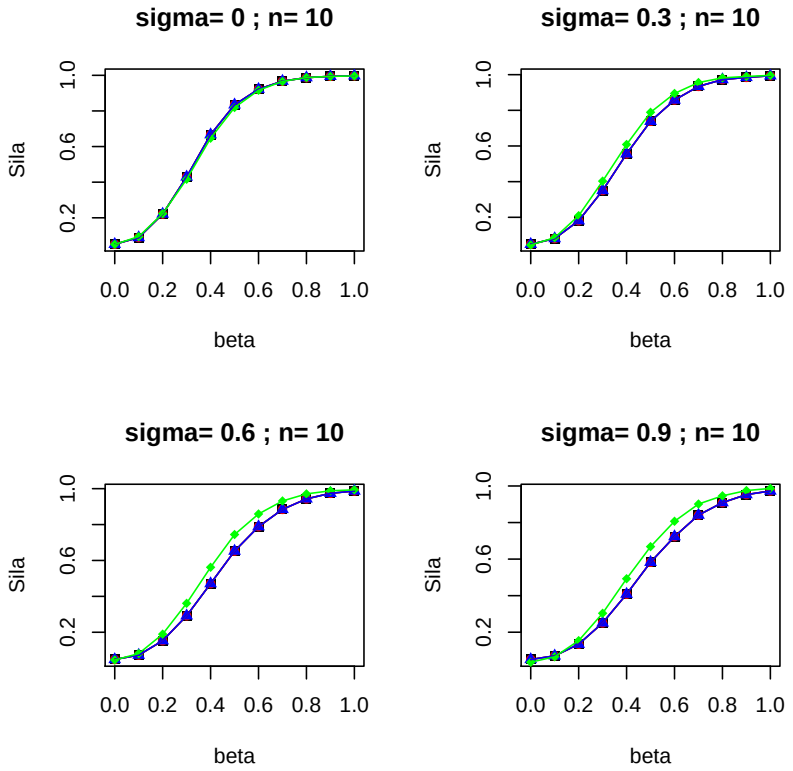
kde prvek v_{22} je $(2,2)$-tý prvek matice $(\mathbf{X}^T\mathbf{X})^{-1}$. Pak hypotézu $H_0 : \beta = 0$ zamítáme na hladině α, je-li $|J_1| \geq t_{(n-2)}(1-\alpha/4)$ nebo $|J_2| \geq t_{(n-2)}(1-\alpha/4)$. (V argumentu funkce t je výraz $1 - \alpha/4$, což plyne z Bonferriho nerovnosti pro testování platnosti dvou hypotéz zároveň). Test založený na jednorozměrné lineární regresi budeme značit J.

Prováděli jsme několik různých simulací, v nichž jsme měnili jak parametr n, tak parametr σ. Každý test jsme prováděli pro všechny hodnoty $\beta = 0, 0.1, 0.2, 0.3, \ldots, 1$. Na následujících stranách lze nalézt grafy závislosti jednotlivých testů na parametru β pro různé hodnoty parametru σ a pozorování n. Testu založeném na poměru věrohodností odpovídá v grafech černá barva, testu založeném na Lawley-Hotellingově stopě červená, testu založeném na Pillaiově stopě modrá a konečně testu založeném na jednorozměrné lineární regresi zelená. Křivky závislosti síly testu na parametru β testů založených na mnohorozměrné lineární regresi se v grafech překrývají a reprezentuje je tedy pouze jedna (příslušná testu založenému na Pillaiově stopě). Dále jsou pod grafy zobrazeny tabulky odhadů hladiny významnosti jednotlivých testů pro různé hodnoty σ a různé hodnoty pozorování n.

Nejprve jsme zvolili $n = 10$ a pro toto n jsme zvolili $a = 7$ a prováděli jsme testy pro $\sigma = 0, 0.3, 0.6, 0.9$. Na obrázku (2.1) jsou grafy závislosti síly testů na parametru β pro jednotlivé σ. V tabulce (2.1) jsou odhady hladin významnosti testů pro parametr $\beta = 0$, tyto hodnoty odpovídají průsečíkům jednotlivých křivek s ypsilonovou osou v grafech na obrázku (2.1). Jak je vidět z této tabulky, všechny testové statistiky dodržují hladinu významnosti. Hodnoty všech testů se pohybují přibližně kolem hodnoty 0.05. Dále z grafů pro $\sigma = 0.3, 0.6, 0.9$ z obrázku (2.1) můžeme vypozorovat, že největší sílu má test založený na otestování jednotlivých jednorozměrných modelů, neboť při pevně zvoleném β z intervalu $[0, 1]$ má nejvyšší hodnotu. Křivky závislosti síly testu na parametru β založené na poměru věrohodností, Pillaiově stopě

a Lawley-Hotellingově stopě se v grafech překrývají. Můžeme tedy usoudit, že tyto testy jsou ekvivalentní. Pro $\sigma = 0$ je s těmito třemi testy dokonce ekvivalentní i test založený na jednorozměrné regresi, což můžeme vypozorovat z prvního grafu obrázku (2.1). S rostoucím σ roste i rozdíl mezi sílou testu založeného na jednorozměrné lineární regresi a sílou ostatních třech testů založených na mnohorozměrné lineární regresi.

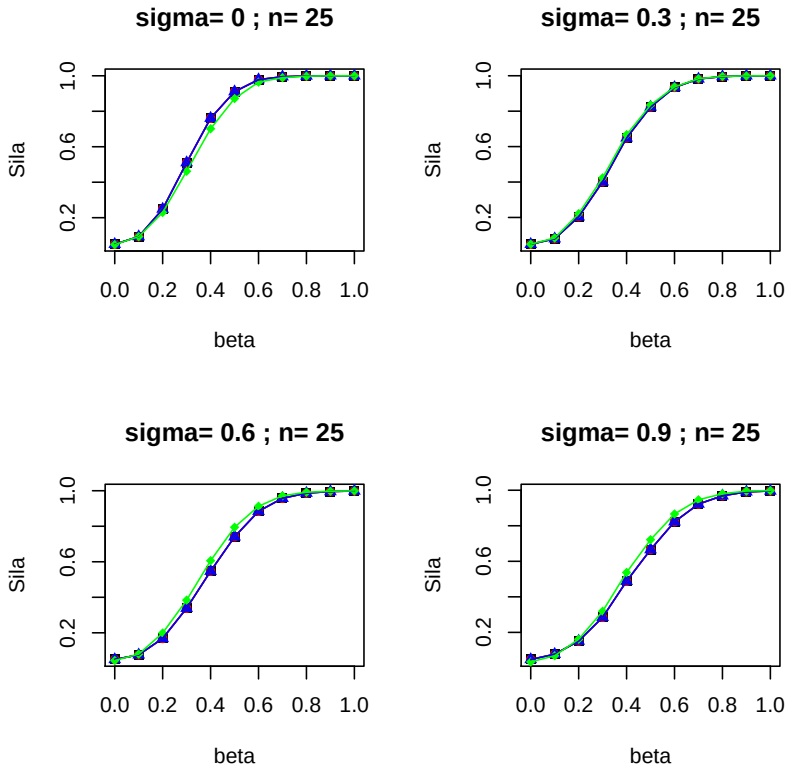
V dalších simulacích jsme volili $n = 25$ ($a = 4$), $n = 50$ ($a = 2$) a $\sigma = 0.3k$ pro $k = 0, 1, \ldots, 3$. Z tabulek (2.2) a (2.3) můžeme vyčíst, že všechny testy opět dodržují hladinu významnosti. Jak je vidět z obrázků (2.2) a (2.3) všechny testy založené na mnohorozměrné lineární regresi (tedy testy U, T^2 a V) jsou opět téměř ekvivalentní. Pro nezávislé vektory chyb (tedy $\sigma = 0$) má tentokrát testová statistika založená na jednorozměrné regresi nejmenší sílu. Pro velké hodnoty kovariance vektoru chyb je ale ze všech zkoumaných testů opět nejsilnější.



Obrázek 2.1: Grafy závislosti síly testu na β pro $n = 10$ - U (černá), T^2 (červená), V (modrá), J (zelená).

Tabulka 2.1: Odhady hladiny významnosti testů (tj. pro $\beta = 0$) pro $n = 10$ a pro parametry $\sigma = 0, 0.3, 0.6, 0.9$.

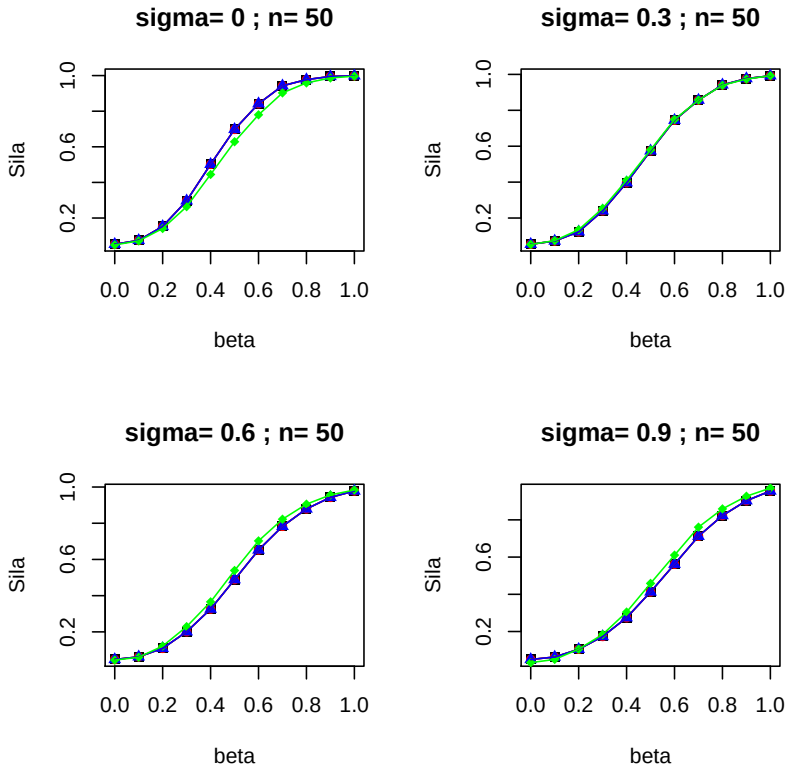
σ	U	T^2	V	J
0	0.0513	0.0513	0.0513	0.0509
0.3	0.0488	0.0488	0.0488	0.0448
0.6	0.0484	0.0484	0.0484	0.0457
0.9	0.0509	0.0509	0.0509	0.0360



Obrázek 2.2: Grafy závislosti síly testu na β pro $n = 25$ - U (černá), T^2 (červená), V (modrá), J (zelená).

Tabulka 2.2: Odhady hladiny významnosti testů (tj. pro $\beta = 0$) pro $n = 25$ a pro parametry $\sigma = 0, 0.3, 0.6, 0.9$.

σ	U	T^2	V	J
0	0.0501	0.0501	0.0501	0.0483
0.3	0.0486	0.0486	0.0486	0.0492
0.6	0.0498	0.0498	0.0498	0.0441
0.9	0.0475	0.0475	0.0475	0.0338



Obrázek 2.3: Grafy závislosti síly testu na β pro $n = 50$ - U (černá), T^2 (červená), V (modrá), J (zelená).

Tabulka 2.3: Odhady hladiny významnosti testů (tj. pro $\beta = 0$) pro $n = 50$ a pro parametry $\sigma = 0, 0.3, 0.6, 0.9$.

σ	U	T^2	V	J
0	0.0527	0.0527	0.0527	0.0492
0.3	0.0538	0.0538	0.0538	0.0529
0.6	0.0476	0.0476	0.0476	0.0441
0.9	0.0498	0.0498	0.0498	0.0329

Závěr

Cílem této bakalářské práce bylo charakterizovat model mnohorozměrné lineární regrese a popsat jeho vlastnosti. Ukázali jsme, jak lze odhadovat regresní parametry jak metodou nejmenších čtverců, tak metodou maximální věrohodnosti. Dále jsme se v této práci zabývali testováním hypotéz. V teorii mnohorozměrné lineární regrese jsme se věnovali podrobněji testovým statistikám založeným na poměru věrohodností, Lawley-Hottelingově stopě a Pillaiově stopě. Rozdíly mezi jednotlivými testovými statistikami jsme se snažili zjistit pomocí simulace sestrojené v programu R. Tato simulace zkoumala zda všechny testy dodržují stanovenou hladinu významnosti a dále porovnávala jednotlivé testy z hlediska jejich sil. Došli jsme k závěrům, že chování jednotlivých testů založených na mnohorozměrné lineární regresi, tj. založené na poměru věrohodností, Lawley-Hottelingově stopě a Pillaiově stopě, je v naší simulaci stejné, tyto testy jsou tedy ekvivalentní. Odlišuje se od nich pouze test založený na jednorozměrné lineární regresi. Pro malé počty pozorování a malé hodnoty kovariance vektorů chyb je tento test ekvivalentní s ostatními. Roste-li hodnota kovariance vektorů chyb, je výhodnější volit test založený na jednorozměrné lineární regresi, neboť má ze všech zkoumaných testů největší sílu. Pro malé hodnoty kovariance vektorů chyb se s rostoucím počtem pozorování však síla tohoto testu zhoršuje, a proto je výhodnější použít libovolný ze tří testů založených na mnohorozměrné lineární regresi.

Literatura

[1] David Birkes, Yadolah Dodge: *Alternative Methods of Regression*, John Wiley & Sons, Inc. , 1993.

[2] Jiří Anděl: *Statistické metody*, MATFYZPRESS, 1998.

[3] Ronald Christensen: *Advanced Linear Modeling*, Springer - Verlag New York, Inc., 2001.

[4] C. R. Rao: *An asymptotic expansion of the distribution of Wilks' criterion*, Bulletin of the International Statistical Institute, 1951.

[5] James J. McKeon: *F approximations to the distribution of Hotteling's* T_0^2, Biometrika, 1974.