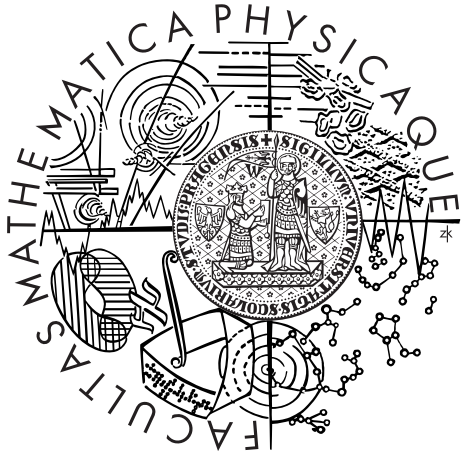
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Hana Jelínková

Porovnání prediktorů

Katedra pravděpodobnosti a matematické statistiky

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

Praha 2011

Srdečně děkuji vedoucímu diplomové práce, doc. RNDr. Karlu Zvárovi, Csc., za jeho cenné rady a kritické připomínky, které mi výrazně pomohly při psaní této práce, a rovněž za jeho trpělivost a ochotu při udílení konzultací. Dále děkuji RNDr. Janě Rubešové, Ph.D., za poskytnutá data.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 1. srpna 2011 Hana Jelínková

Název práce: Porovnání prediktorů
Autor: Hana Jelínková
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: doc. RNDr. Karel Zvára, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V předložené práci se zaměřujeme na popis a odvození testů zkoumajících vliv prediktorů na vysvětlení závisle proměnné.

Nejprve se soustředíme na porovnání vlivu dvou prediktorů. Za tímto účelem uvažujeme celkem tři různé testy rovnosti dvou korelačních koeficientů, pomocí nichž vliv prediktorů porovnáváme. Součástí práce je, kromě odvození těchto testů, také jejich porovnání provedené na základě simulací, při nichž je zkoumáno, jak dodržují hladinu významnosti a jakou mají sílu.

V další části práce uvažovaný postup pro porovnávání dvou prediktorů zobecníme a přejdeme k porovnávání relativní důležitosti dvou skupin prediktorů v modelu lineární regrese. Závěr práce je pak věnován numerické demonstraci teorie.

Klíčová slova: prediktor, korelační koeficient, lineární regresní model.

Title: Comparing of predictors
Author: Hana Jelínková
Department: Department of Probability and Mathematical Statistics
Supervisor: doc. RNDr. Karel Zvára, CSc., Department of Probability and Mathematical Statistics

Abstract: In the present work we target at the description and derivation of tests which examine the effect of predictors on explaining a dependent variable.

First we focus on the comparison of contribution of two predictors. For this purpose we consider three different tests of equality of two correlation coefficients, which we use to compare the contribution of predictors. The thesis contains apart from the derivation of these tests also their comparison made on the basis of simulations in which we test if they hold the nominal level and how powerful these tests are.

In another part of the work we generalize the considered procedure for comparing two predictors and move on to the comparison of the relative importance of two groups of predictors in a model of linear regression. The end of the work is then devoted to a numerical demonstration of the theory.

Keywords: predictor, correlation coefficient, linear regression model

Obsah

Úvod

Tématem této diplomové práce jsou postupy navržené za účelem porovnání prediktorů. Práce je rozdělena do dvou částí. V první, teoretické, části se nejprve soustředíme na porovnání vlivu dvou prediktorů na vysvětlení závisle proměnné. Za tímto účelem jsou podrobně popsány a odvozeny některé testy testující rovnost dvou korelačních koeficientů, jmenovitě test Hotellingův, Williamsův a test poměrem věrohodností. Zatímco odvození Hotellingova testu a testu poměrem věrohodností je dohledatelné v literatuře, o Williamsově testu se veškerá nám dostupná literatura zmiňuje vždy pouze jako o modifikaci Hotellingova testu, avšak jakékoli podrobnější vysvětlení chybí. Proto odvození Williamsova testu bylo jedním z hlavních cílů této práce.

Dále zobecníme postup použitý při porovnávání vlivu dvou prediktorů a přejdeme k porovnávání relativní důležitosti dvou skupin prediktorů v modelu lineární regrese.

První polovina praktické části práce si klade za cíl porovnání testů rovnosti dvou korelačních koeficientů odvozených v teoretické části. Pro každý uvažovaný test je pomocí simulací zkoumáno, jak test v různých nastaveních dodržuje hladinu významnosti a jakou má sílu. Výsledky provedených simulací prezentujeme v tabulkách, které pro lepší názornost doplňujeme také grafy. Druhá polovina praktické části je pak věnována numerické demonstraci teorie obsažené v první části práce. K tomu využíváme data obsahující údaje o 102 studentech bakalářského studijního oboru Geografie–kartografie Přírodovědecké fakulty Univerzity Karlovy v Praze, kteří se zapsali ke studiu v akademickém roce 2003/04. Veškeré výpočty provádíme ve volně šiřitelném statistickém programu **R** verze **2.13.0**.

Příloha práce obsahuje některé skripty, které jsme použili při výpočtech a simulacích v programu **R**. Ty jsou pak k dispozici spolu s textem diplomové práce také na přiloženém CD.

Kapitola 1

Značení a definice některých pojmů z teorie testování hypotéz

V této kapitole zavedeme značení, které budeme používat v průběhu celé práce, a dále uvedeme některé pojmy z teorie testování hypotéz.

Tučný písmem budeme značit vektory a matice, normálním písmem skaláry. Transpozici matice $\mathbf{A}$ označíme $\mathbf{A}'$. Označením pro d-dimenzionální sloupcový vektor jedniček je $\mathbf{1}_d$. Jednotkovou d-dimenzionální matici označíme $\mathbf{I}_d$. Označení pro další symboly zavedeme v průběhu práce.

Protože se celá tato práce týká testování hypotéz, zopakujeme si zde ještě některé pojmy z jeho teorie. Nechť náhodný vektor $\mathbf{X} = (X_1, X_2, ..., X_n)'$ má rozdělení, které závisí na parametru $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)'$. O tomto parametru víme, že patří do nějaké množiny Θ, kterou nazýváme *parametrický prostor*. Dále předpokládejme, že o parametru $\boldsymbol{\theta}$ existují dvě navzájem si konkurující hypotézy. První z nich je hypotéza H_0: $\boldsymbol{\theta} \in \Theta_0 (\subset \Theta)$, tu nazýváme *nulová hypotéza*. Druhou je tzv. *alternativní hypotéza*: H_1: $\boldsymbol{\theta} \in \Theta_1 (= \Theta - \Theta_0)$. Naším úkolem je na základě náhodného vektoru $\mathbf{X}$ rozhodnout, zda hypotézu H_0 zamítnout ve prospěch alternativy H_1, nebo nezamítnout. Postup, při kterém zjišťujeme, zda hypotézu zamítnout, či nezamítnout, se nazývá *testování hypotéz*. Naše rozhodnutí ale nemusí být vždy správné. Pokud zamítneme hypotézu H_0, ačkoli je správná, pak se dopustíme *chyby prvního druhu*. Když naopak tuto hypotézu nezamítneme, ačkoli není správná, dopustíme se *chyby druhého druhu*. Nástroj, pomocí něhož rozhodujeme o zamítnutí, či nezamítnutí nulové hypotézy, se nazývá *testová statistika*. Množina hodnot, kterých tato statistika nabývá, se rozpadá na dvě disjunktní množiny. Množinu $W \subset \mathbf{R}^n$ takovou, že je-li $\mathbf{X} \in W$, potom hypotézu H_0 zamítneme ve prospěch alternativy H_1, nazýváme *kritický obor testu*. Jestliže $\mathbf{X} \notin W$ hypotézu H_0 nezamítneme. Kritický obor volíme tak, aby pravděpodobnost chyby prvního druhu byla menší nebo rovna zvolenému malému kladnému číslu α. Obvykle volíme α rovno 0,05. Hodnotu $sup_{\boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}(\mathbf{X} \in W)$ nazýváme *hladinou významnosti testu*. Pokud má testová statistika spojité rozdělení, můžeme vždy zvolit test, jehož hladina významnosti je právě rovna α. *Silofunkcí testu* budeme rozumět funkci proměnné $\boldsymbol{\theta}$, která udává pravděpodobnost, že zamítneme nulovou hypotézu za podmínky, že neplatí, tj. $\beta(\boldsymbol{\theta}) = P(\boldsymbol{X} \in W | \boldsymbol{\theta} \in \Theta_1)$. Řekneme, že test T_1 má větší sílu než test T_2, pro nějaké pevné $\tilde{\boldsymbol{\theta}} \in \Theta_1$ na hladině významnosti α, jestliže oba testy mají hladinu významnosti α a silofunkce testu T_1 je v bodě $\tilde{\boldsymbol{\theta}}$ větší než silofunkce testu T_2 v tomtéž bodě.

Kapitola 2

Testy rovnosti dvou korelačních koeficientů

Uvažujme náhodný vektor $(Y, X_1, X_2)'$ mající trojrozměrné normální rozdělení se střední hodnotou $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2)'$, $\mu_i \in \mathbf{R}$, $i = 0, 1, 2$, a regulární varianční maticí

$$\mathbf{V} = \begin{pmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & \rho_{02}\sigma_0\sigma_2 \\ \rho_{01}\sigma_0\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{02}\sigma_0\sigma_2 & \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

kde $\sigma_i > 0$, $i = 0, 1, 2$, jsou po řadě směrodatné odchylky náhodných veličin Y, X_1 a X_2, $\rho_{01}, \rho_{02} \in (-1, 1)$ korelační koeficienty náhodných veličin Y a X_1, resp. Y a X_2, a $\rho_{12} \in (-1, 1)$ korelační koeficient veličin X_1 a X_2.

Naším úkolem v této kapitole bude odvození testů testujících hypotézu o rovnosti korelačních koeficientů ρ_{01} a ρ_{02} (tj. hypotézu H_0: $\rho_{01} = \rho_{02}$). Motivací pro testování této hypotézy je úloha, v níž máme k dispozici tři náhodné veličiny; Y jako vysvětlovanou proměnnou (resp. závisle proměnnou) a X_1 a X_2 jako proměnné vysvětlující (resp. nezávislé). K vysvětlení proměnné Y chceme však použít jen jednu z uvažovaných vysvětlujících proměnných. To, kterou z nich vybereme, rozhodneme právě na základě korelačních koeficientů mezi vysvětlovanou a vysvětlujícími proměnnými, přičemž vezmeme tu, která má s vysvětlovanou proměnnou větší korelaci.

Existuje řada testů testujících námi uvažovanou hypotézu, my zde uvedeme tři z nich, a to test Hotellingův, Williamsův a test poměrem věrohodností.

Hypotézu H_0 budeme testovat proti oboustranné alternativě H_1: $\rho_{01} \neq \rho_{02}$, a to na základě náhodného výběru (Y_i, X_{1i}, X_{2i}) o rozsahu n $(1 \leq i \leq n)$ z trojrozměrného rozdělení $N_3(\boldsymbol{\mu}, \boldsymbol{V})$.

Používat budeme následujícího značení:

- $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$, $\mathbf{X}_j = (X_{j1}, X_{j2}, ..., X_{jn})'$, kde $j = 1, 2$.
- Výběrové průměry $\mathbf{Y}$, $\mathbf{X_1}$, resp. $\mathbf{X_2}$:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ji}, \quad j = 1, 2.$$

- Výběrové rozptyly $\mathbf{Y}$, $\mathbf{X_1}$, resp. $\mathbf{X_2}$:

$$s_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad \text{a} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2, \quad j = 1, 2.$$

- Výběrové korelační koeficienty:

$$r_{01} = r_{10} = \frac{s_{01}}{s_0 s_1}, \qquad r_{02} = r_{20} = \frac{s_{02}}{s_0 s_2} \qquad \text{a} \qquad r_{12} = r_{21} = \frac{s_{12}}{s_1 s_2},$$

kde

$$s_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$$

a

$$s_{0j} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{ji} - \bar{X}_j), \quad j = 1, 2.$$

- Korelační matice a výběrová korelační matice:

$$\mathbf{P} = \begin{pmatrix} 1 & \rho_{01} & \rho_{02} \\ \rho_{10} & 1 & \rho_{12} \\ \rho_{20} & \rho_{21} & 1 \end{pmatrix} \qquad \text{a} \qquad \mathbf{R} = \begin{pmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{pmatrix}.$$

Poznámka 1. V rámci zjednodušení zápisu budeme v textu někdy namísto symbolů s_j^2, $j = 0, 1, 2$, označujících výběrové rozptyly používat symboly s_{jj}, $j = 0, 1, 2$. $\Diamond$

Poznámka 2. Při odvozování Hotellingovy a Williamsovy testové statistiky budeme pracovat s rozdělením náhodného vektoru $\mathbf{Y}$ za podmínek $\mathbf{X_1} = \mathbf{x_1}$ a $\mathbf{X_2} = \mathbf{x_2}$. Potom symboly s_{ij} a r_{ij} $i, j = 0, 1, 2$, definované výše jako funkce náhodných vektorů $\mathbf{Y}$ a $\mathbf{X}_j$, $j = 1, 2$, budeme používat i v případě, kdy tyto funkce budou namísto na těchto náhodných vektorech záviset na pevných hodnotách vektorů $\mathbf{x_1}$ a $\mathbf{x_2}$. Z textu bude nicméně zřejmé, co kdy daný symbol označuje. $\Diamond$

2.1 Zavedení testových statistik a testů

2.1.1 Hotellingův test

Hotellingův test (viz [5]) používá k testování nulové hypotézy H_0: $\rho_{01} = \rho_{02}$ testovou statistiku

$$T_1 = (r_{01} - r_{02})\sqrt{\frac{(n-3)(1+r_{12})}{2\,|\mathbf{R}|}}, \tag{2.1}$$

kde $|\mathbf{R}| = 1 - r_{12}^2 - r_{01}^2 - r_{02}^2 + 2r_{01}r_{02}r_{12}$ je determinant matice $\mathbf{R}$. Uvedená testová statistika má za platnosti nulové hypotézy t-rozdělení s $n-3$ stupni volnosti.

Hypotézu H_0 tedy zamítneme na hladině α ve prospěch alternativy H_1, jestliže absolutní hodnota testové statistiky T_1 je větší nebo rovna $(1-\alpha/2)$-kvantilu t-rozdělení s $n-3$ stupni volnosti (tj. $|T_1| \geq t_{n-3}(1-\alpha/2)$).

Odvození testové statistiky Hotellingova testu

Uvažujme náhodný vektor $(Y, X_1, X_2)'$, který má trojrozměrné normální rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\mathbf{V}$. Uvažujme dále náhodný výběr (Y_i, X_{1i}, X_{2i}), $1 \leq i \leq n$, z tohoto rozdělení. Testová statistika Hotellingova testu je odvozena za předpokladu, že hodnoty náhodných veličin X_{1i}, X_{2i}, $1 \leq i \leq n$, jsou pevné a rovné hodnotám jedné dané realizace náhodného výběru. Podmíníme-li tedy náhodné veličiny Y_i realizacemi x_{1i} a x_{2i} náhodných veličin X_{1i} a X_{2i}, $1 \leq i \leq n$, získáme vzhledem k tomu, že podmíněné rozdělení náhodné veličiny Y_i za podmínky $X_{1i} = x_{1i}$ a $X_{2i} = x_{2i}$ je normální se střední hodnotou $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ a rozptylem σ^2 (viz [1, str. 67]), regresní model:

$$M_1: \quad Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad 1 \leq i \leq n,$$

kde

$$\beta_0 = \mu_0 - \beta_1 \mu_1 - \beta_2 \mu_2$$

a

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \frac{\sigma_0}{1-\rho_{12}^2} \begin{pmatrix} \frac{1}{\sigma_1}(\rho_{01} - \rho_{02}\rho_{12}) \\ \frac{1}{\sigma_2}(\rho_{02} - \rho_{01}\rho_{12}) \end{pmatrix} \tag{2.2}$$

jsou regresní koeficienty modelu a kde

$$\sigma^2 = \frac{\sigma_0^2}{1-\rho_{12}} |\mathbf{P}|$$

je reziduální rozptyl.

Hodnoty x_{1i} a x_{2i}, $1 \leq i \leq n$, jsou rovny hodnotám jedné dané realizace náhodného výběru. Symboly s_{ij} a r_{ij}, $i, j = 0, 1, 2$, používané dále, budeme proto uvažovat jako funkce těchto pevných hodnot x_{1i} a x_{2i}, $1 \leq i \leq n$, nikoli jako funkce náhodných vektorů $\mathbf{X_1}$ a $\mathbf{X_2}$ (viz Poznámka 2). Označme

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

matici modelu M_1 a předpokládejme, že tato matice má lineárně nezávislé sloupce ($h(\mathbf{X}) = 3$), tedy že model M_1 je regulární. Dále, abychom si v následujícím zjednodušili výpočty, přejdeme nyní od modelu

$$\begin{aligned} M_1: \quad Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \qquad 1 \leq i \leq n, \\ &= (\beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2) + \beta_1 s_1 \left(\frac{x_{1i} - \bar{x}_1}{s_1} \right) + \beta_2 s_2 \left(\frac{x_{2i} - \bar{x}_2}{s_2} \right) + \epsilon_i \end{aligned}$$

k modelu s centrovanými a škálovanými regresory, tedy k modelu

$$M_1^*: \quad Y_i = \beta_0^* + \beta_1^* \left(\frac{x_{1i} - \bar{x}_1}{s_1} \right) + \beta_2^* \left(\frac{x_{2i} - \bar{x}_2}{s_2} \right) + \epsilon_i, \quad 1 \leq i \leq n,$$

kde $\beta_0^* = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2$, $\beta_1^* = \beta_1 s_1$ a $\beta_2^* = \beta_2 s_2$.

Z lineární nezávislosti sloupců matice $\mathbf{X}$ plyne, že druhý ani třetí sloupec této matice není násobkem sloupce prvního, tedy že $\mathbf{x_{1\bullet}} \neq c\mathbf{1}$ a $\mathbf{x_{2\bullet}} \neq d\mathbf{1}$, kde $c, d \in \mathbf{R}$, což nám umožňuje vydělit vektory

$$\begin{pmatrix} x_{11} - \bar{x}_1 \\ x_{12} - \bar{x}_1 \\ \vdots \\ x_{1n} - \bar{x}_1 \end{pmatrix}, \qquad \begin{pmatrix} x_{21} - \bar{x}_2 \\ x_{22} - \bar{x}_2 \\ \vdots \\ x_{2n} - \bar{x}_2 \end{pmatrix}$$

výběrovými směrodatnými odchylkami a přechod k modelu M_1^*. Označme

$$\mathbf{X}^* = \begin{pmatrix} 1 & \frac{x_{11} - \bar{x}_1}{s_1} & \frac{x_{21} - \bar{x}_2}{s_2} \\ 1 & \frac{x_{12} - \bar{x}_1}{s_1} & \frac{x_{22} - \bar{x}_2}{s_2} \\ \vdots & \vdots & \vdots \\ 1 & \frac{x_{1n} - \bar{x}_1}{s_1} & \frac{x_{2n} - \bar{x}_2}{s_2} \end{pmatrix} \tag{2.3}$$

matici modelu M_1^* a $\boldsymbol{\beta^*} = (\beta_0^*, \beta_1^*, \beta_2^*)$ vektor regresních koeficientů. Z konstrukce modelu M_1^* je zřejmé, že $E\boldsymbol{Y} = \boldsymbol{X\beta} = \boldsymbol{X^*\beta^*}$, tedy že modely M_1 a M_1^* jsou totožné. Matici $\mathbf{X}^*$ dostaneme z matice $\mathbf{X}$ postupným násobením zprava dvěma trojúhelníkovými regulárními maticemi. Nejprve maticí

$$\mathbf{A_1} = \begin{pmatrix} 1 & -\bar{x}_1 & -\bar{x}_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{a poté maticí} \quad \mathbf{A_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{s_1} & 0 \\ 0 & 0 & \frac{1}{s_2} \end{pmatrix}.$$

Z věty 8.16 na str. 63 v [2] pak vyplývá, že

$$h(\mathbf{X}^*) = h\left[(\mathbf{X}\mathbf{A_1})\mathbf{A_2}\right] = h(\mathbf{X}) = 3.$$

Tedy model M_1^* je také, stejně jako model M_1, regulární.

V regresním modelu porovnáváme vliv regresorů na závisle proměnnou (při nezměněných hodnotách ostatních regresorů) pomocí regresních koeficientů. Požadujeme však, aby regresory, jejichž přínos porovnáváme, měly nulový průměr a stejný rozptyl (zpravidla jednotkový). Takto upravené regresory jsou potom ve stejném měřítku a příslušné regresní koeficienty je možno stejně interpretovat. Čím je absolutní hodnota regresního koeficientu větší, tím větší má příslušný regresor přínos k vysvětlení závisle proměnné. V našem případě jsou takto standardizovanými koeficienty regresní koeficienty β_1^* a β_2^* modelu M_1^*. Od testu hypotézy H_0: $\rho_{01} = \rho_{02}$ proti alternativě H_1: $\rho_{01} \neq \rho_{02}$ tedy přecházíme k testu hypotézy

$$H_0^*: \ \beta_1^* = \beta_2^* \quad \text{proti alternativě} \quad H_1^*: \ \beta_1^* \neq \beta_2^*.$$

Tuto hypotézu lze testovat pomocí testové statistiky

$$T(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}) = \frac{V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})}{\sqrt{\widehat{\text{var}}\,[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})]}}, \tag{2.4}$$

kde

$$V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}) = b_1^* - b_2^* = b_1 s_1 - b_2 s_2$$

a kde

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \frac{s_0}{1 - r_{12}^2} \begin{pmatrix} \frac{1}{s_1}(r_{01} - r_{02} r_{12}) \\ \frac{1}{s_2}(r_{02} - r_{01} r_{12}) \end{pmatrix} \tag{2.5}$$

jsou odhady regresních koeficientů β_1 a β_2 metodou nejmenších čtverců. Rozptyl $\text{var}\,[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})]$ ve jmenovateli (2.4) je nahrazen odhadem, protože, jak později uvidíme, závisí na rozptylu modelu σ^2, který neznáme.

Upravme nyní čitatele a jmenovatele (2.4) a vyjádřeme tuto statistiku pomocí výběrových korelačních koeficientů. Soustřeďme se nejprve na výpočet rozptylu $V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})$.

Označme $\mathbf{b}^* = (b_0^*, b_1^*, b_2^*)'$ odhad regresních koeficientů $\boldsymbol{\beta^*} = (\beta_0^*, \beta_1^*, \beta_2^*)'$ modelu M_1^*. Potom vzhledem k tomu, že (viz [1, str. 82])

$$\begin{aligned} \text{var}\ \mathbf{b}^* &= \sigma^2 (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \\ &= \frac{\sigma^2}{(n-1)(1-r_{12}^2)} \begin{pmatrix} \frac{(n-1)(1-r_{12}^2)}{n} & 0 & 0 \\ 0 & 1 & -r_{12} \\ 0 & -r_{12} & 1 \end{pmatrix}, \end{aligned}$$

neboť

$$\mathbf{X}^{*\prime} \mathbf{X}^* = \begin{pmatrix} n & 0 & 0 \\ 0 & (n-1) & (n-1) r_{12} \\ 0 & (n-1) r_{12} & (n-1) \end{pmatrix},$$

platí

$$\begin{aligned} \text{var}\,[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}))] &= \text{var}\,(b_1^* - b_2^*) = \text{var}\ b_1^* + \text{var}\ b_2^* - 2\text{cov}\,(b_1^*, b_2^*) \\ &= \frac{2\sigma^2}{(n-1)(1-r_{12})}. \end{aligned} \tag{2.6}$$

Jak už jsme se zmínili dříve, rozptyl $\text{var}\,[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}))]$ závisí na neznámém parametru σ^2. Nahradíme ho proto jeho nestranným odhadem (viz [1, str. 82]), kterým je náhodná veličina

$$s^2 = \frac{RSS}{n-3},$$

kde RSS je reziduální součet čtverců modelu M_1^*. Máme

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i, \qquad \mathbf{X}^{*\prime}\mathbf{Y} = \begin{pmatrix} n\bar{Y} \\ (n-1) s_0 r_{01} \\ (n-1) s_0 r_{02} \end{pmatrix}$$

a

$$\mathbf{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y} = \begin{pmatrix} \bar{Y} \\ \frac{s_0(r_{01}-r_{02}r_{12})}{1-r_{12}^2} \\ \frac{s_0(r_{02}-r_{01}r_{12})}{1-r_{12}^2} \end{pmatrix}.$$

Pak pro reziduální součet čtverců platí (viz [1, str. 82])

$$\begin{aligned} RSS &= \mathbf{Y}'\mathbf{Y} - \mathbf{b}^{*\prime}\mathbf{X}^{*\prime}\mathbf{Y} \\ &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \left(1 - \frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{(1-r_{12})(1+r_{12})}\right) \\ &= \frac{(n-1)s_0^2\,|\mathbf{R}|}{1-r_{12}^2} \end{aligned}$$

a odhad rozptylu σ^2 je

$$s^2 = \frac{(n-1)s_0^2\,|\mathbf{R}|}{(1-r_{12}^2)(n-3)}.$$

Potom dosadíme-li do (2.6), dostaneme

$$\widehat{\text{var}}[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})] = \frac{2s_0^2\,|\mathbf{R}|}{(n-3)(1-r_{12})(1-r_{12}^2)}. \tag{2.7}$$

Jmenovatele statistiky (2.4) tedy již známe, zbývá upravit čitatele. Dosadíme-li ve vzorci $V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}) = b_1 s_1 - b_2 s_2$ za b_1 a b_2 (viz (2.5)) dostaneme po zjednodušení

$$V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}) = \frac{r_{01} - r_{02}}{1 - r_{12}} s_0. \tag{2.8}$$

Odtud a z (2.7) pak po dosazení do (2.4) získáme

$$\begin{aligned} T_1 &= T(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}) = \frac{V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})}{\sqrt{\widehat{\text{var}}\,[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})]}} \\ &= \frac{\frac{r_{01}-r_{02}}{1-r_{12}} s_0^2}{\sqrt{\frac{2s_0^2|\mathbf{R}|}{(n-3)(1+r_{12})(1-r_{12})^2}}} = \frac{r_{01} - r_{02}}{\sqrt{\frac{2|\mathbf{R}|}{(1+r_{12})(n-3)}}}, \end{aligned}$$

což je však přesně testová statistika uvedená na začátku (viz (2.1)).

Rozdělení Hotellingovy testové statistiky

Zbývá určit rozdělení testové statistky T_1. Máme (viz [1, str. 83])

$$\boldsymbol{b^*} \sim N(\boldsymbol{\beta^*}, \sigma^2(\boldsymbol{X^{*\prime}X^*})^{-1}).$$

Potom

$$b_1^* - b_2^* \sim N\left(\beta_1^* - \beta_2^*, \frac{2\sigma^2}{(n-1)(1-r_{12})}\right)$$

a za platnosti nulové hypotézy H_0 platí

$$\frac{b_1^* - b_2^*}{\sqrt{\frac{2\sigma^2}{(n-1)(1-r_{12})}}} \sim N(0,1).$$

Dále máme (viz opět [1, str. 83])

$$\frac{RSS}{\sigma^2} = \frac{(n-1)s_0^2\,|\mathbf{R}|}{(1-r_{12}^2)\sigma^2} \sim \chi^2_{n-3}$$

a protože b^* a s^2 jsou nezávislé, má veličina

$$T_1 = \frac{\frac{b_1^* - b_2^*}{\sqrt{\frac{2\sigma^2}{(n-1)(1-r_{12})}}}}{\sqrt{\frac{(n-1)s_0^2|\mathbf{R}|}{(1-r_{12}^2)\sigma^2}}}\sqrt{n-3} = \frac{r_{01}-r_{02}}{\sqrt{\frac{2|\mathbf{R}|}{(1+r_{12})(n-3)}}}$$

za platnosti nulové hypotézy H_0 Studentovo rozdělení t_{n-3}.

Jak již jsme uvedli v úvodu této části, testová statistika Hotellingova testu je odvozena za předpokladu, že hodnoty náhodných veličin X_{1i}, X_{2i}, $1 \le i \le n$, jsou pevné a rovné hodnotám jedné dané realizace náhodného výběru. Uvažována je tedy pouze variabilita náhodného vektoru $\mathbf{Y}$, nikoli však již variabilita vektorů $\mathbf{X_1}$ a $\mathbf{X_2}$.

2.1.2 Williamsův test

Dalším testem používaným k testování hypotézy $H_0 : \rho_{01} = \rho_{02}$ je Williamsův test (viz [9]). Tento test je velmi podobný Hotellingovu, avšak na rozdíl od něj zohledňuje variabilitu nezávisle proměnných X_1 a X_2. Williamsův test k testování hypotézy H_0 používá statistiku

$$T_2 = (r_{01} - r_{02})\sqrt{\frac{(n-1)(1+r_{12})}{2\left(\frac{n-1}{n-3}\right)|\mathbf{R}| + \bar{r}^2(1-r_{12})^3}},$$

kde $\bar{r}$ označuje aritmetický průměr korelačních koeficientů r_{01} a r_{02}, tedy $\bar{r} = (r_{01} + r_{02})/2$, a kde $|\mathbf{R}|$ opět označuje determinant matice $\mathbf{R}$. Vzhledem k tomu, že testová statistika T_2 má za platnosti hypotézy H_0 přibližně t-rozdělení s $n-3$ stupni volnosti, zamítneme hypotézu H_0 na hladině α ve prospěch alternativy H_1, jestliže absolutní hodnota testové statistiky T_2 je větší nebo rovna $(1-\alpha/2)$-kvantilu t-rozdělení s $n-3$ stupni volnosti (tj. $|T_2| \ge t_{n-3}(1-\alpha/2)$).

Odvození testové statistiky Williamsova testu

Uveďme nyní odvození tohoto testu. Uvažujme opět nejprve náhodný vektor $(Y, X_1, X_2)'$ mající normální rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\mathbf{V}$. Nechť je dále $(Y_i, X_{1i}, X_{i2})'$, $1 \le i \le n$, náhodný výběr z tohoto rozdělení. Potom, podmíníme-li náhodné veličiny Y_i realizacemi x_{1i} a x_{2i} náhodných veličin X_{1i} a X_{2i}, $1 \le i \le n$, získáme, stejně jako při odvozování Hotellingova testu, regresní model

$$M_1 : Y_i = (\beta_0 + \beta_1\bar{x}_1 + \beta_2\bar{x}_2) + \beta_1 s_1 \left(\frac{x_{1i} - \bar{x}_1}{s_1}\right) + \beta_2 s_2 \left(\frac{x_{2i} - \bar{x}_2}{s_2}\right) + \epsilon_i,$$

kde $1 \leq i \leq n$, a test hypotézy H_0: $\rho_{01} = \rho_{02}$ potom odpovídá testu hypotézy

$$H_0^* : \quad \beta_1 s_1 = \beta_2 s_2.$$

Tuto hypotézu jsme v předchozí kapitole testovali pomocí testové statistiky

$$T_1 = T(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2}) = \frac{V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})}{\sqrt{\widehat{\text{var}}\left[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})\right]}}. \tag{2.9}$$

Dívali jsme se tedy na statistiku V, resp. T, jako na funkci náhodného vektoru $\mathbf{Y}$ a vektorů konstant $\mathbf{x_1}$ a $\mathbf{x_2}$. Tento přístup však nepočítá s variabilitou náhodných vektorů $\mathbf{X_1}$ a $\mathbf{X_2}$. Abychom tuto variabilitu zohlednili, uvažujme nyní statistiku V, resp. T, jako funkci náhodných vektorů $\mathbf{Y}$, $\mathbf{X_1}$ a $\mathbf{X_2}$. Tedy

$$T(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2}) = \frac{V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})}{\sqrt{\widehat{\text{var}}\left[V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})\right]}}. \tag{2.10}$$

Připomeňme, že

$$V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2}) = b_1 s_1 - b_2 s_2,$$

přičemž symboly s_1 a s_2 zde označují výběrové rozptyly náhodných vektorů $\mathbf{X_1}$ a $\mathbf{X_2}$.

Abychom vyjádřili testovou statistiku $T(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})$ pomocí výběrových korelačních koeficientů, potřebujeme spočítat rozptyl statistiky $V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})$. Ten získáme pomocí podmiňování, a to tak, že rozptyl této statistiky rozložíme na součet rozptylu podmíněné střední hodnoty a střední hodnoty podmíněného rozptylu, přičemž podmiňovat budeme opět náhodný vektor $\mathbf{Y}$ realizacemi $\mathbf{x_1}$ a $\mathbf{x_2}$ náhodných vektorů $\mathbf{X_1}$ a $\mathbf{X_2}$. Následující vztah lze nalézt např. v [1, str. 59];

$$\begin{aligned} \text{var}\left[V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})\right] &= \text{var}_{\mathbf{X_1}, \mathbf{X_2}}\left[E(V \,|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2})\right] \\ &\quad + E_{\mathbf{X_1}, \mathbf{X_2}}\left[\text{var}(V \,|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2})\right]. \end{aligned} \tag{2.11}$$

Soustřeďme se nejprve na výpočet druhého členu pravé strany této rovnosti. Podmíněnou střední hodnotu $E_{\mathbf{X_1}, \mathbf{X_2}}\left[\text{var}(V \,|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2})\right]$ nahradíme odhadem podmíněného rozptylu $\text{var}(V \,|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2})$. Ten je však roven odhadu rozptylu $\text{var}\left[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})\right]$ počítaném při odvozování Hotellingova testu v předchozí podkapitole (viz (2.7)), tedy

$$\widehat{\text{var}}[V|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}] = \frac{2s_0^2\,|\mathbf{R}|}{(n-3)(1-r_{12})(1-r_{12}^2)}. \tag{2.12}$$

Přejděme dále k výpočtu prvního členu pravé strany (2.11). Vzhledem k tomu, že

$$E(V|\mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}) = \beta_1 s_1 - \beta_2 s_2,$$

je první člen (2.11) roven $\text{var}_{\mathbf{X_1}, \mathbf{X_2}}\left[\beta_1 s_1 - \beta_2 s_2\right]$. Tento rozptyl aproximujeme pomocí Taylorova rozvoje funkce.

Nechť $f(s_1^2, s_2^2) = \beta_1\sqrt{s_1^2} - \beta_2\sqrt{s_2^2}$ je funkce náhodných veličin s_1^2 a s_2^2. Potom vzhledem k tomu, že tato funkce má v nějakém okolí bodu (Es_1^2, Es_2^2) derivace všech řádů a náhodné veličiny s_1^2 a s_2^2 mají konečné momenty, můžeme funkci f v bodě (Es_1^2, Es_2^2) aproximovat Taylorovým polynomem prvního řádu, tedy

$$f(s_1^2, s_2^2) \approx f(Es_1^2, Es_2^2) + \left(\begin{array}{c} \frac{\partial f(u_1,u_2)}{\partial u_1} \\ \frac{\partial f(u_1,u_2)}{\partial u_2} \end{array} \right)' \Bigg|_{u_i = Es_i^2} \left(\begin{array}{c} s_1^2 - Es_1^2 \\ s_2^2 - Es_2^2 \end{array} \right),$$

kde $u_i \in \mathbf{R}$ a $i = 1, 2$. Z toho poté vyplývá následující vztah pro aproximaci rozptylu funkce náhodných veličin:

$$\operatorname{var}\left[f(s_1^2, s_2^2)\right] \approx \left(\begin{array}{c} \frac{\partial f(u_1,u_2)}{\partial u_1} \\ \frac{\partial f(u_1,u_2)}{\partial u_2} \end{array} \right)' \Bigg|_{u_i = Es_i^2} \operatorname{var} \left(\begin{array}{c} s_1^2 \\ s_2^2 \end{array} \right) \left(\begin{array}{c} \frac{\partial f(u_1,u_2)}{\partial u_1} \\ \frac{\partial f(u_1,u_2)}{\partial u_2} \end{array} \right) \Bigg|_{u_i = ES_i^2}, \quad (2.13)$$

kde opět $u_i \in \mathbf{R}$ a $i = 1, 2$.

Abychom mohli tento vztah použít, potřebujeme nejprve vyjádřit střední hodnotu a varianční matici vektoru výběrových rozptylů $(s_1^2, s_2^2)'$.

Lemma 1. *Nechť s_1^2, resp. s_2^2, je výběrový rozptyl náhodného vektoru $\mathbf{X_1}$, resp. $\mathbf{X_2}$. Potom*

$$E \left(\begin{array}{c} s_1^2 \\ s_2^2 \end{array} \right) = \left(\begin{array}{c} \sigma_1^2 \\ \sigma_2^2 \end{array} \right) \quad a \quad var \left(\begin{array}{c} s_1^2 \\ s_2^2 \end{array} \right) = \left(\begin{array}{cc} \frac{2\sigma_1^4}{n-1}, & \frac{2\rho_{12}^2\sigma_1^2\sigma_2^2}{n-1} \\ \frac{2\rho_{12}^2\sigma_1^2\sigma_2^2}{n-1}, & \frac{2\sigma_2^4}{n-1} \end{array} \right).$$

Důkaz. Vzhledem k tomu, že náhodná veličina $\sum_{i=1}^{n} \left(\frac{X_{1i} - \bar{X}_1}{\sigma_1}\right)^2$ má χ^2-rozdělení s $(n-1)$ stupni volnosti, pro střední hodnotu a rozptyl náhodné veličiny s_1^2 platí

$$Es_1^2 = \sigma_1^2 \qquad \text{a} \qquad \operatorname{var}\, s_1^2 = \frac{\sigma_1^4}{(n-1)^2} 2(n-1) = \frac{2\sigma_1^4}{(n-1)}.$$

Analogické vztahy dostaneme pro střední hodnotu a rozptyl s_2^2.

Při výpočtu kovariance $\operatorname{cov}(s_1^2, s_2^2)$ použijeme podobného postupu jako při výpočtu rozptylu statistiky $V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})$ (viz (2.11)). Vyjádříme-li výběrové rozptyly s_1^2 a s_2^2 maticově:

$$s_1^2 = \frac{1}{n-1} \mathbf{X_1'MX_1} \quad \text{a} \quad s_2^2 = \frac{1}{n-1} \mathbf{X_2'MX_2},$$

kde $\mathbf{M} = \mathbf{I} - \frac{1}{n}\mathbf{1'1}$ je matice, pro niž platí $\mathbf{M'} = \mathbf{M}$ a $\mathbf{MM} = \mathbf{M}$, dostaneme

$$\operatorname{cov}(s_1^2, s_2^2) = \frac{1}{(n-1)^2} \operatorname{cov}(\mathbf{X_1'MX_1}, \mathbf{X_2'MX_2}).$$

Kovarianci $\operatorname{cov}(\mathbf{X_1'MX_1}, \mathbf{X_2'MX_2})$ pak (viz [1, str. 61]) můžeme rozložit na součet střední hodnoty podmíněné kovariance a kovariance dvou podmíněných středních hodnot. Podmiňovat budeme tentokrát náhodný vektor $\mathbf{X_1}$ realizacemi $\mathbf{x_2}$ náhodného vektoru $\mathbf{X_2}$. Platí

$$\begin{aligned}\operatorname{cov}(\mathbf{X_1'MX_1}, \mathbf{X_2'MX_2}) &= E_{\mathbf{X_2}}\left[\operatorname{cov}(\mathbf{X_1'MX_1}, \mathbf{X_2'MX_2}\,|\mathbf{X_2} = \mathbf{x_2})\right] \\ &+ \operatorname{cov}_{\mathbf{X_2}}\left[E(\mathbf{X_1'MX_1}|\mathbf{X_2} = \mathbf{x_2}), E(\mathbf{X_2'MX_2}|\mathbf{X_2} = \mathbf{x_2})\right].\end{aligned}$$

Protože první člen výrazu pravé strany této rovnosti je roven nule, neboť se jedná o kovarianci mezi náhodnou veličinou a konstantou, je hodnota kovariance náhodných veličin $\mathbf{X_1'MX_1}$ a $\mathbf{X_2'MX_2}$ rovna pouze druhému členu. Vypočtěme nejprve podmíněné střední hodnoty

$$E(\mathbf{X_1'MX_1}|\mathbf{X_2} = \mathbf{x_2}) \quad \text{a} \quad E(\mathbf{X_2'MX_2}|\mathbf{X_2} = \mathbf{x_2}).$$

Výpočet druhé z nich je snadný, neboť jde o střední hodnotu konstanty:

$$E(\mathbf{X_2'MX_2}|\mathbf{X_2} = \mathbf{x_2}) = \mathbf{x_2'Mx_2}.$$

K hodnotě první z nich dojdeme takto. Uvažujme náhodný vektor $\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$, $1 \le i \le n$. Potom vzhledem k tomu, že podmíněné rozdělení náhodné veličiny X_{1i} za podmínky $X_{2i} = x_{2i}$, $1 \le i \le n$, je normální se střední hodnotou $\mu_1 + \rho_{12}\frac{\sigma_1}{\sigma_2}(x_{2i} - \mu_2)$ a rozptylem $\sigma_1^2(1-\rho_{12}^2)$ (viz [1, str. 67]), je možné vyjádřit náhodný vektor $\mathbf{X_1}$ pomocí náhodného vektoru $\mathbf{Z}$, jehož složky Z_i mají normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem, následovně:

$$\mathbf{X_1} = \sqrt{\sigma_1^2(1-\rho_{12})}\mathbf{Z} + \boldsymbol{\mu_1} + \tfrac{\rho_{12}\sigma_1}{\sigma_2}(\mathbf{x_2} - \boldsymbol{\mu_2}).$$

Dosadíme-li toto vyjádření $\mathbf{X_1}$ do $E(\mathbf{X_1'MX_1}|\mathbf{X_2} = \mathbf{x_2})$ a uvědomíme-li si opět, že $\mathbf{Z'MZ}$ má χ^2-rozdělení s $n-1$ stupni volnosti, po několika úpravách obdržíme

$$E(\mathbf{X_1'MX_1}|\mathbf{X_2} = \mathbf{x_2}) = \sigma_1^2(1-\rho_{12}^2)(n-1) + \left(\frac{\rho_{12}\sigma_1}{\sigma_2}\right)^2 \mathbf{x_2'Mx_2}.$$

Potom tedy

$$\begin{aligned}\operatorname{cov}(\mathbf{X_1'MX_1}, \mathbf{X_2'MX_2}) &= \operatorname{cov}_{\mathbf{X_2}}\left[E(\mathbf{X_1'MX_1}|\mathbf{X_2} = \mathbf{x_2}), E(\mathbf{X_2'MX_2}|\mathbf{X_2} = \mathbf{x_2})\right] \\ &= \left(\frac{\rho_{12}\sigma_1}{\sigma_2}\right)^2 \operatorname{var}(\mathbf{X_2'MX_2}) = \left(\frac{\rho_{12}\sigma_1}{\sigma_2}\right)^2 2(n-1)\sigma_2^4 \\ &= 2\rho_{12}^2\sigma_1^2\sigma_2^2(n-1)\end{aligned}$$

a

$$\operatorname{cov}(s_1^2, s_2^2) = \frac{2\rho_{12}^2\sigma_1^2\sigma_2^2}{n-1}.$$

□

Dosadíme-li nyní do (2.13) za střední hodnotu a varianční matici vektoru výběrových rozptylů $(s_1^2, s_2^2)'$ (viz Lemma 1) dostaneme

$$\begin{aligned}\mathrm{var}_{\mathbf{X_1},\mathbf{X_2}}[f(s_1^2,s_2^2)] &\approx \begin{pmatrix} \frac{1}{2}\beta_1\sigma_1^{-\frac{1}{2}} \\ -\frac{1}{2}\beta_2\sigma_2^{-\frac{1}{2}} \end{pmatrix}' \begin{pmatrix} \frac{2\sigma_1^4}{n-1}, & \frac{2\rho_{12}^2\sigma_1^2\sigma_2^2}{n-1} \\ \frac{2\rho_{12}^2\sigma_1^2\sigma_2^2}{n-1}, & \frac{2\sigma_2^4}{n-1} \end{pmatrix} \begin{pmatrix} \frac{1}{2}\beta_1\sigma_1^{-\frac{1}{2}} \\ -\frac{1}{2}\beta_2\sigma_2^{-\frac{1}{2}} \end{pmatrix} \\ &\approx \frac{1}{2}\beta_1^2\frac{\sigma_1^2}{n-1}+\frac{1}{2}\beta_2^2\frac{\sigma_2^2}{n-1}-\beta_1\beta_2\frac{\rho_{12}^2\sigma_1\sigma_2}{n-1}.\end{aligned}$$

Dále v předchozím výrazu dosadíme také za β_1 a β_2 (viz (2.2)) a vzhledem k tomu, že nám jde o rozdělení statistiky V za platnosti nulové hypotézy, položíme $\rho = \rho_{01} = \rho_{02}$. Po úpravě obdržíme

$$\mathrm{var}_{\mathbf{X_1},\mathbf{X_2}}[f(s_1^2,s_2^2)] \approx \frac{\rho^2\sigma_0^2(1-\rho_{12})}{(n-1)(1+\rho_{12})}.$$

Tedy

$$\mathrm{var}_{\mathbf{X_1},\mathbf{X_2}}[E(V\,|\mathbf{X_1}=\mathbf{x_1},\mathbf{X_2}=\mathbf{x_2})] \approx \frac{\rho^2\sigma_0^2(1-\rho_{12})}{(n-1)(1+\rho_{12})}.$$

Protože však hodnoty parametrů ρ, ρ_{12} a σ_0^2 v předchozím výrazu neznáme, použijeme ve výpočtu rozptylu podmíněné střední hodnoty jejich odhady. Jako odhad korelačního koeficientu ρ_{12} vezmeme výběrový korelační koeficient r_{12}, jako odhad rozptylu σ_0^2 náhodné veličiny Y výběrový rozptyl s_0^2 a jako odhad parametru ρ použijeme aritmetický průměr výběrových korelačních koeficientů r_{01} a r_{02} (tj. $\bar{r} = (r_{01}+r_{02})/2$). Pak tedy

$$\widehat{\mathrm{var}}[E(V\,|\mathbf{X_1}=\mathbf{x_1},\mathbf{X_2}=\mathbf{x_2})] = \frac{s_0^2\bar{r}(1-r_{12})}{(n-1)(1+r_{12})}. \tag{2.14}$$

Tím jsme tedy získali i druhý ze sčítanců výrazu na pravé straně (2.11). Dosadíme-li nyní (2.14) a (2.12) do (2.11) dostaneme

$$\begin{aligned}\widehat{\mathrm{var}}\,[V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})] &= \frac{s_0^2\bar{r}(1-r_{12})}{(n-1)(1+r_{12})}+\frac{2s_0^2\,|\mathbf{R}|}{(n-3)(1+r_{12})(1-r_{12})^2} \\ &= \frac{2s_0^2\,|\mathbf{R}|\left(\frac{n-1}{n-3}\right)+\bar{r}^2s_0^2(1-r_{12})^3}{(1+r_{12})(1-r_{12})^2(n-1)}.\end{aligned} \tag{2.15}$$

Spočetli jsme tedy rozptyl statistiky $V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})$, podívejme se ještě na její střední hodnotu. Tu získáme opět pomocí podmiňování. Platí (viz [1, str. 58])

$$E\left[V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})\right] = E_{\mathbf{X_1},\mathbf{X_2}}\left[E(V\,|\mathbf{X_1}=\mathbf{x_1},\mathbf{X_2}=\mathbf{x_2})\right].$$

Vzhledem k tomu, že

$$E(V|\mathbf{X_1}=\mathbf{x_1},\mathbf{X_2}=\mathbf{x_2}) = \beta_1 s_1-\beta_2 s_2,$$

dostáváme

$$E_{\mathbf{X_1},\mathbf{X_2}}\left[\beta_1 s_1-\beta_2 s_2\right] = \beta_1 E_{\mathbf{X_1}}\left[s_1\right]-\beta_2 E_{\mathbf{X_2}}\left[s_2\right]. \tag{2.16}$$

K dokončení výpočtu nám stačí určit $E_{\mathbf{X_1}}[s_1]$. Protože, jak již bylo řečeno, náhodná veličina $\sum_{i=1}^{n}\left(\frac{X_{1i}-\bar{X}_1}{\sigma_1}\right)^2$ má χ^2-rozdělení s $(n-1)$ stupni volnosti a protože střední hodnota náhodné veličiny mající χ-rozdělení s $n-1$ stupni volnosti je $\sqrt{2}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$ (viz [11]), platí

$$E_{\mathbf{X_1}}[s_1] = E\left[\frac{\sigma_1}{\sqrt{n-1}}\sqrt{\sum_{i=1}^{n}\left(\frac{X_{1i}-\bar{X}_1}{\sigma_1}\right)^2}\right] = \sigma_1\frac{\sqrt{2}}{\sqrt{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}.$$

Analogicky dostaneme $E_{\mathbf{X_2}}[s_2]$. Potom, dosadíme-li do (2.16) za vypočtené střední hodnoty a za β_1 a β_2 z (2.2), získáme po úpravě

$$E(V|\mathbf{X_1}=\mathbf{x_1},\mathbf{X_2}=\mathbf{x_2}) = E_{\mathbf{X_1},\mathbf{X_2}}(\beta_1 s_1 - \beta_2 s_2) = \frac{c\sigma_0}{1-\rho_{12}}(\rho_{01}-\rho_{02}),$$

kde $c = \frac{\sqrt{2}}{\sqrt{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$. Z tohoto vyjádření vyplývá, že střední hodnota statistiky $V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})$ je rovna nule právě tehdy, když platí nulová hypotéza, což nám potvrzuje, že navržená testová statistika

$$T(\mathbf{Y},\mathbf{X_1},\mathbf{X_2}) = \frac{V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})}{\sqrt{\widehat{\text{var}}\ [V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})]}} \tag{2.17}$$

správně reaguje na splnění nulové hypotézy.

Dokončeme nyní výpočet testové statistiky T_2. Jmenovatele T_2 známe a z předchozí podkapitoly známe i čitatele (viz (2.8)):

$$V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2}) = \frac{r_{01}-r_{02}}{1-r_{12}}s_0.$$

Odtud a z (2.15) pak po dosazení dostaneme

$$\begin{aligned} T_2 &= T(\mathbf{Y},\mathbf{X_1},\mathbf{X_2}) = \frac{V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})}{\sqrt{\widehat{\text{var}}\,[V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})]}} \\ &= \frac{\frac{r_{01}-r_{02}}{1-r_{12}}s_0}{\sqrt{\frac{2s_0^2|\mathbf{R}|\left(\frac{n-1}{n-3}\right)+\bar{r}^2 s_0^2(1-r_{12})^3}{(1+r_{12})(1-r_{12})^2(n-1)}}} = \frac{r_{01}-r_{02}}{\sqrt{\frac{2|\mathbf{R}|\left(\frac{n-1}{n-3}\right)+\bar{r}^2(1-r_{12})^3}{(1+r_{12})(n-1)}}}, \end{aligned}$$

což je přesně Williamsova statistika.

Rozdělení testové statistiky Williamsova testu

V předchozí podkapitole jsme odvodili, že Hotellingova testová statistika má t-rozdělení s $n-3$ stupni volnosti. Hotellingovu testovou statistiku jsme odvodili za předpokladu, že hodnoty náhodných vektorů $\mathbf{X_1}$ a $\mathbf{X_1}$ jsou pevné. Při odvozování Williamsovy testové statistiky jsme vzali v úvahu i variabilitu těchto náhodných vektorů a rozptyl $V(\mathbf{Y},\mathbf{X_1},\mathbf{X_2})$ dostali pouze aproximativně. Z tohoto důvodu Williamsova statistika nemá již přesně t-rozdělení, jako měla statistika Hotellingova. Nicméně budeme předpokládat, že statistika T_2 t-rozdělení s $n-3$

stupni rovnosti má přibližně. Potvrdí nám to také simulační experiment uvedený v kapitole 4. Počet stupňů volnosti je dán tím, že jsme museli opět odhadnout neznámý parametr σ^2, na kterém závisí rozptyl statistiky $V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})$. Při testování nulové hypotézy H_0 proti oboustranné alternativě H_1 budeme tedy absolutní hodnotu Williamsovy testové statistiky porovnávat s $(1-\alpha/2)$-kvantilem t-rozdělení s $n-3$ stupni volnosti.

2.1.3 Test poměrem věrohodností

Další možností testování hypotézy H_0: $\rho_{01} = \rho_{02}$ je test založený na věrohodnostním poměru. Uvažujme opět náhodný výběr $(Y_i, X_{1i}, X_{2i})'$, $1 \leq i \leq n$, z trojrozměrného normální rozdělení se střední hodnotou $\boldsymbol{\mu}$ a regulární varianční maticí $\mathbf{V}$, jehož hustota je

$$f(y, x_1, x_2) = (2\pi)^{-\frac{3}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} y-\mu_0 \\ x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}' \mathbf{V}^{-1} \begin{pmatrix} y-\mu_0 \\ x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \right\}.$$

Test poměrem věrohodností s rušivými parametry (viz [1, str. 183–184]) testuje hypotézu $H_0 : \rho_{01} = \rho_{02}$ pomocí testové statistiky

$$T_3 = 2\,[L(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) - L(\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}})],$$

kde

$$\begin{aligned} L(\boldsymbol{\theta_n}) = -\tfrac{3}{2} n \log(2\pi) \quad &- \quad \tfrac{1}{2} n \log |\mathbf{V}| \\ &- \quad \tfrac{1}{2} \textstyle\sum_{i=1}^{n} \left\{ \begin{pmatrix} Y_i-\mu_0 \\ X_{1i}-\mu_1 \\ X_{2i}-\mu_2 \end{pmatrix}' \mathbf{V}^{-1} \begin{pmatrix} Y_i-\mu_0 \\ X_{1i}-\mu_1 \\ X_{2i}-\mu_2 \end{pmatrix} \right\} \end{aligned}$$

je logaritmická věrohodnostní funkce. Dále kde

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}} = (\boldsymbol{\tau_0'}, \tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}')' = ((\rho, \rho)', (\tilde{\rho}_{12}, \tilde{\mu}_0, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma_0^2}, \tilde{\sigma_1^2}, \tilde{\sigma_2^2})')'$$

je odhad parametru $\boldsymbol{\theta}_{\boldsymbol{n}} = (\boldsymbol{\tau_n'}, \boldsymbol{\psi_n'})' = ((\rho_{01}, \rho_{02})', (\rho_{12}, \mu_0, \mu_1, \mu_2, \sigma_0^2, \sigma_1^2, \sigma_2^2)')'$ metodou maximální věrohodnosti za podmínky, že $\rho_{01} = \rho_{02} = \rho$, a kde

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} = (\hat{\boldsymbol{\tau}}_{\boldsymbol{n}}', \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}')' = ((\hat{\rho}_{01}, \hat{\rho}_{02})', (\hat{\rho}_{12}, \hat{\mu}_0, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma_0^2}, \hat{\sigma_1^2}, \hat{\sigma_2^2})')'$$

je odhad parametru $\boldsymbol{\theta}_{\boldsymbol{n}} = (\boldsymbol{\tau_n'}, \boldsymbol{\psi_n'})'$ metodou maximální věrohodnosti, který není vázaný žádnými podmínkami.

Hypotéza, kterou testujeme, se zajímá pouze o $\boldsymbol{\tau_n}$, parametry obsažené v $\boldsymbol{\psi_n}$ jsou sice potřebné k popisu pravděpodobnostního modelu, nicméně nulová hypotéza se jich vůbec netýká. Z toho důvodu se $\boldsymbol{\tau_n}$ nazývá *cílový parametr* a $\boldsymbol{\psi_n}$ *rušivý parametr*.

Maximálně věrohodný odhad $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ parametru $\boldsymbol{\theta}_{\boldsymbol{n}}$ získáme snadno. Odhady středních hodnot μ_i, $i = 0, 1, 2$, jsou po řadě rovny výběrovým průměrům $\bar{Y}$, $\bar{X}_1$ a $\bar{X}_2$, tedy

$$\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ji}, \quad j = 1, 2.$$

Odhady rozptylů σ_j^2 jsou po řadě rovny $\frac{n-1}{n}$-násobkům výběrových rozptylů s_j^2, $j = 0, 1, 2$, tedy

$$\hat{\sigma_0^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad \text{a} \quad \hat{\sigma_j^2} = \frac{1}{n}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2, \quad j = 1, 2,$$

a konečně odhady korelačních koeficientů ρ_{01}, ρ_{02} a ρ_{12} jsou po řadě rovny výběrovým korelačním koeficientům r_{01}, r_{02} a r_{12} , tedy

$$\hat{\rho}_{01} = \frac{s_{01}}{s_0 s_1}, \qquad \hat{\rho}_{02} = \frac{s_{02}}{s_0 s_2} \qquad \text{a} \qquad \hat{\rho}_{12} = \frac{s_{12}}{s_1 s_2}.$$

Maximálně věrohodný odhad $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ parametru $\boldsymbol{\theta}_{\boldsymbol{n}}$ za platnosti nulové hypotézy H_0: $\rho_{01} = \rho_{02}$ lze získat nalezením maxima Lagrangeovy funkce

$$F(\boldsymbol{\theta}_{\boldsymbol{n}}; \eta) = L(\boldsymbol{\theta}_{\boldsymbol{n}}) + \eta(\rho_{01} - \rho_{02}).$$

Tato úloha nicméně vede k soustavě rovnic (viz [6]), jejichž analytickým řešením se zde věnovat nebudeme. Maximálně věrohodný odhad $\tilde{\boldsymbol{\theta}}_{\boldsymbol{n}}$ získáme iteračně, přičemž jako první iteraci vezmeme vektor $(\bar{r}, \bar{r}, \hat{\rho}_{12}, \hat{\mu}_j, \hat{\sigma_j^2})'$, $j = 0, 1, 2$, kde $\bar{r} = (r_{01} + r_{02})/2$.

Rozdělení testové statistiky T_3, pokud jsou splněny všechny potřebné předpoklady (viz [1, str. 184]), konverguje za platnosti hypotézy H_0 v distribuci k rozdělení χ^2 s jedním stupněm volnosti. V našem případě vzhledem k tomu, že pracujeme s trojrozměrným normálním rozdělením, není se splněním těchto podmínek problém. Potom tedy, jestliže hodnota testové statistiky T_3 je větší nebo rovna $(1-\alpha)$-kvantilu χ_1^2-rozdělení (tj. $T_3 \geq \chi_1^2(1-\alpha)$)), hypotézu H_0 zamítneme.

2.1.4 Nabídka prostředí R

V prostředí **R** lze najít naprogramovaný Williamsův test, a to v knihovně `psych` pod názvem `r.test`. Tuto funkci lze kromě testování hypotézy o rovnosti dvou korelačních koeficientů použít například také k testování nulovosti jediného korelačního koeficientu, záleží pouze na zadaných vstupních parametrech. Vstupními parametry pro Williamsův test jsou proměnné `n`, `r12`, `r13`, `r23` a `twotailed`, kde `n` je počet pozorování, `r12`, `r13`, resp. `r23` výběrové korelační koeficienty námi označované r_{01}, r_{02}, resp. r_{12}. Pomocí parametru `twotailed` nastavíme, zda má být proveden test jednostranné nebo oboustranné alternativy. Výsledkem je testová statistika Williamsova testu a příslušná p-hodnota.

Kapitola 3

Porovnání důležitosti dvou skupin prediktorů v modelu lineární regrese

V předchozí kapitole jsme se věnovali úloze, při níž jsme měli k dispozici dvě vysvětlující proměnné, a naším cílem bylo mezi nimi vybrat tu, která by lépe vysvětlila závisle proměnnou. Nyní bude naším úkolem uvedený postup zobecnit a přejít od porovnávání dvou prediktorů k porovnání relativní důležitosti dvou skupin prediktorů. Uvažovat nicméně budeme pouze případ, kdy prediktory jsou pevné hodnoty, nikoli náhodné veličiny, jak tomu bylo v předchozí kapitole. Čerpat budeme především z článku [7], kde je také možno nalézt jednu z ukázek použití této úlohy. Tento článek se nicméně zabývá logistickou regresí, zatímco my zde budeme uvažovat lineární regresní model.

Mějme náhodný vektor $\mathbf{Y} = (Y_1, ..., Y_n)'$ a matici daných čísel $\mathbf{H} = (h_{ij})$ typu $n \times J$, kde $J < n$. Předpokládejme, že pro náhodný vektor $\mathbf{Y}$ platí lineární regresní model

$$M: \quad \boldsymbol{Y} = \boldsymbol{H\theta} + \boldsymbol{\epsilon},$$

kde $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)'$ je vektor neznámých regresních parametrů a $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)'$ je náhodný vektor mající normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2\mathbf{I}_n$, přičemž $\sigma^2 > 0$ je rovněž neznámý parametr.

Předpokládejme dále, že sloupce matice $\mathbf{H}$ jsou lineárně nezávislé, tedy že model M je regulární, a že tyto sloupce jsou uspořádány v matici $\mathbf{H}$ v následujícím pořadí: vektor jedniček $\mathbf{1}$, K vektorů matice $\mathbf{X}$, M vektorů matice $\mathbf{Z}$ a na závěr L vektorů matice, kterou označíme $\mathbf{W}$. Tedy

$$\mathbf{H} = (\mathbf{1}, \mathbf{X}, \mathbf{Z}, \mathbf{W}).$$

Matice $\mathbf{X}$ a matice $\mathbf{Z}$ obsahují skupiny prediktorů, jejichž relativní důležitost budeme porovnávat, matice $\mathbf{W}$ pak ty regresory, které se sice v modelu vyskytují, ale o které se blíže nezajímáme. Model M lze tedy rozepsat takto

$$M: \quad \boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{W\delta} + \boldsymbol{\epsilon},$$

resp. po složkách

$$M: \quad Y_i = \alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{m=1}^{M} \gamma_m z_{im} + \sum_{l=1}^{L} \delta_l w_{il} + \epsilon_i, \quad 1 \le i \le n.$$

Zřejmě $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta'}, \boldsymbol{\gamma'}, \boldsymbol{\delta'})'$ a $J = K + M + L + 1$.

Abychom si zjednodušili další zápis, předpokládejme nyní navíc, že součty složek sloupců matic $\mathbf{X}$ a $\mathbf{Z}$ jsou rovné nule. Odečtení průměrů sloupců matic $\mathbf{X}$ a $\mathbf{Z}$ od složek těchto sloupců způsobí v regresním modelu pouze změnu konstantního členu α, o který se v tomto případě nezajímáme. Proto můžeme tento předpoklad učinit bez újmy na obecnosti.

V předchozí kapitole jsme rozhodovali o větší důležitosti prediktoru na základě korelačních koeficientů. K vysvětlení závisle proměnné jsme vybrali ten prediktor, který měl s vysvětlovanou proměnnou větší korelaci. Testovali jsme tedy hypotézu H_0: $\rho_{01} = \rho_{02}$. Nicméně při odvozování testových statistik Hotellingova a Williamsova testu jsme v obou případech přešli od testu hypotézy o rovnosti korelačních koeficientů k testu hypotézy o rovnosti standardizovaných regresních koeficientů. Testovali jsme tedy hypotézu

$$H_0^*: \quad \beta_1 s_1 = \beta_2 s_2.$$

Platí-li však tato rovnost, platí také rovnost druhých mocnin $\beta_1^2 s_1^2$ a $\beta_2^2 s_2^2$ a po úpravě také rovnost

$$\beta_1^2 s_1^2 \sum_{i=1}^{n} \left(\frac{x_{1i} - \bar{x}_1}{s_1} \right)^2 = \beta_2^2 s_2^2 \sum_{i=1}^{n} \left(\frac{x_{2i} - \bar{x}_2}{s_2} \right)^2, \tag{3.1}$$

jež je, vybavíme-li si značení z předchozí kapitoly, ekvivalentní s

$$\beta_1^* \mathbf{x}_{1\bullet}^{*\prime} \mathbf{x}_{1\bullet}^* \beta_1^* = \beta_2^* \mathbf{x}_{2\bullet}^{*\prime} \mathbf{x}_{2\bullet}^* \beta_2^*,$$

resp.

$$\left\| \mathbf{x}_{1\bullet}^* \beta_1^* \right\|^2 = \left\| \mathbf{x}_{2\bullet}^* \beta_2^* \right\|^2.$$

Symbolem $\mathbf{x}_{1\bullet}^*$, resp. $\mathbf{x}_{2\bullet}^*$, značíme druhý, resp. třetí, sloupec matice $\mathbf{X}^*$ modelu M_1^* (viz (2.3)). Od porovnávání standardizovaných regresních koeficientů jsme tedy přešli k porovnávání čtverců délek vektorů $\mathbf{x}_1^* \beta_1^*$ a $\mathbf{x}_2^* \beta_2^*$. Je-li délka jednoho z vektorů větší, potom je také přínos příslušného prediktoru v modelu větší. A to nám dává nyní návod k porovnání relativní důležitosti dvou skupin prediktorů.

Definujme ω jako poměr čtverců délek vektorů $\boldsymbol{X\beta}$ a $\boldsymbol{Z\gamma}$, tedy

$$\omega = \frac{\left\| \boldsymbol{X\beta} \right\|^2}{\left\| \boldsymbol{Z\gamma} \right\|^2} = \frac{\boldsymbol{\beta' X' X \beta}}{\boldsymbol{\gamma' Z' Z \gamma}}. \tag{3.2}$$

Potom, pokud obě skupiny prediktorů přispívají k vysvětlení závisle proměnné stejným dílem, je $\omega = 1$. Pokud je podíl různý od jedné, pak jedna ze skupin prediktorů přispívá k vysvětlení závisle proměnné více než druhá. Abychom zjistili, jak tomu je, potřebujeme testovat hypotézu

$$H_0: \quad \omega = 1.$$

Poznámka 3. Vektory $\mathbf{x}^*_{1\bullet}$ a $\mathbf{x}^*_{2\bullet}$ matice $\mathbf{X}^*$ jsou nejen centrované, ale také škálované, zatímco sloupce matic $\mathbf{X}$ a $\mathbf{Z}$ uvažované dále v definici ω jsou pouze centrované. To však nevadí, neboť výběrové rozptyly s_1^2 a s_2^2 se na obou stranách rovnice vykrátí (viz (3.1)). $\Diamond$

Odvození testové statistiky pro test hypotézy H_0: $\omega = 1$ a jejího rozdělení

K odvození testové statistiky potřebujeme nejprve odhadnout regresní koeficienty $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$, které se vyskytují ve vyjádření ω. Odhad $\hat{\boldsymbol{\theta}} = (\hat{\alpha},\, \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\delta}}')'$ parametru $\boldsymbol{\theta} = (\alpha,\, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\delta}')'$ metodou nejmenších čtverců je roven (viz [1, str. 81])

$$\hat{\boldsymbol{\theta}} = (\hat{\alpha},\, \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\delta}}')' = (\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{Y}.$$

Složky odhadu $\hat{\boldsymbol{\beta}}$, resp. $\hat{\boldsymbol{\gamma}}$, parametru $\boldsymbol{\beta}$, resp. $\boldsymbol{\gamma}$, pak zřejmě získáme vybráním příslušných složek vektoru $\hat{\boldsymbol{\theta}}$. Poznamenejme, že v tomto případě je odhad parametru $\boldsymbol{\theta}$ metodou nejmenších čtverců roven odhadu metodou maximální věrohodnosti, neboť úloha minimalizace výrazu

$$(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{\theta})'(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{\theta}),$$

jakožto funkce $\boldsymbol{\theta}$, je ekvivalentní úloze maximalizace logaritmické věrohodnostní funkce

$$L(\boldsymbol{\theta}) = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^2) - \tfrac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{\theta})'(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{\theta}).$$

Dosadíme-li maximálně věrohodné odhady $\hat{\boldsymbol{\beta}}$ a $\hat{\boldsymbol{\gamma}}$ do (3.2), získáme maximálně věrohodný odhad $\hat{\omega}$ (viz [1, str. 148] Zehnaova věta - princip invariance pro maximálně věrohodné odhady), tedy

$$\hat{\omega} = \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\gamma}}'\boldsymbol{Z}'\boldsymbol{Z}\hat{\boldsymbol{\gamma}}}. \tag{3.3}$$

Poznámka 4. Poznamenejme na tomto místě další možnou interpretaci uvažovaného poměru. Čitatel $\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^2$ je až na multiplikační konstantu roven výběrovému rozptylu složek $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, analogicky jmenovatel $\|\boldsymbol{Z}\hat{\boldsymbol{\gamma}}\|^2$. To vyplývá z předpokladu $\mathbf{1}'\mathbf{X} = \mathbf{0}$, resp. $\mathbf{1}'\mathbf{Z} = \mathbf{0}$. Potom totiž zřejmě

$$\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \overline{X\beta}\mathbf{1}\right\|^2 = \left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^2, \quad \text{resp.} \quad \left\|\boldsymbol{Z}\hat{\boldsymbol{\gamma}} - \overline{Z\gamma}\mathbf{1}\right\|^2 = \|\boldsymbol{Z}\hat{\boldsymbol{\gamma}}\|^2.$$

Poměr čtverců délek vektorů $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ a $\boldsymbol{Z}\hat{\boldsymbol{\gamma}}$, zde označený $\hat{\omega}$, lze tedy interpretovat také jako poměr výběrových rozptylů složek těchto vektorů. $\Diamond$

Dále k testu hypotézy H_0: $\omega = 1$ potřebujeme znát rozdělení statistiky $\hat{\omega}$. Nicméně my, namísto přímo rozdělení $\hat{\omega}$, odvodíme rozdělení přirozeného logaritmu tohoto odhadu, což vzhledem k tomu, že $\omega > 0$ a $P(\hat{\omega} > 0) = 1$, můžeme, a hypotézu H_0: $\omega = 1$ budeme testovat pomocí testové statistiky

$$Z = \frac{\log\hat{\omega}}{\sqrt{\widehat{\operatorname{var}}(\log\hat{\omega})}}.$$

Rozptyl odhadu $\log\hat{\omega}$ ve jmenovateli nahrazujeme jeho odhadem, protože zaprvé závisí na neznámém parametru σ^2 a zadruhé protože místo funkce log(.) budeme při výpočtu rozptylu $\log\hat{\omega}$ používat její lineární aproximaci.

Poznámka 5. Odvoďme rozdělení $\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^2$. Z [10, str. 226] víme, že pro matici $\mathbf{X}'\mathbf{X}$, jakožto symetrickou matici, existuje spektrální rozklad, tedy existují ortonormální matice $\mathbf{Q}$ a diagonální matice $\boldsymbol{\Lambda}$ s diagonálními prvky $\lambda_1 \geq \lambda_2 \geq \lambda_3$ tak, že platí

$$\mathbf{X}'\mathbf{X} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'.$$

Z [1, str. 83] dále máme

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}),$$

kde $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ je varianční matice var$(\hat{\boldsymbol{\beta}})$. Potom

$$\boldsymbol{V} = \boldsymbol{Q}'\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{Q}'\boldsymbol{\beta},\, \boldsymbol{Q}'\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{Q})$$

a tedy

$$\mathbf{V}_j \sim N(\boldsymbol{q}_j{}'\boldsymbol{\beta},\, \boldsymbol{q}_j{}'\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{q}_j).$$

Z toho vyplývá

$$\begin{aligned}\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}'\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'\hat{\boldsymbol{\beta}} = \boldsymbol{V}'\boldsymbol{\Lambda}\boldsymbol{V} \\ &= \textstyle\sum_{j=1}^{k}\lambda_j{\mathbf{V}_j}^2 = \sum_{j=1}^{k}\lambda_j\left(\sqrt{\boldsymbol{q}_j{}'\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{q}_j}\right)^2\left(\frac{\mathbf{V}_j}{\sqrt{\boldsymbol{q}_j{}'\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{q}_j}}\right)^2.\end{aligned}$$

Potom vzhledem k tomu, že $\left(\frac{\mathbf{V}_j}{\sqrt{\boldsymbol{q}_j{}'\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{q}_j}}\right)^2$ má necentrální χ^2_1-rozdělení s parametrem necentrality $\lambda = (\boldsymbol{q}_j{}'\boldsymbol{\beta})^2$, je $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ lineární kombinací obecně závislých necentrálních χ^2_1. Vidíme tedy, že odvodit přesně rozdělení $\hat{\omega}$ by bylo komplikované. Z tohoto důvodu jsme přešli k logaritmu $\hat{\omega}$ a odvodíme rozdělení $\log\hat{\omega}$ asymptoticky. $\lozenge$

Zaveďme dále funkci τ, a to následovně:

$$\tau(\boldsymbol{u}'_{\boldsymbol{\beta}},\boldsymbol{u}'_{\boldsymbol{\gamma}})' = \log\left(\frac{\boldsymbol{u}'_{\boldsymbol{\beta}}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u}_{\boldsymbol{\beta}}}{\boldsymbol{u}'_{\boldsymbol{\gamma}}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{u}_{\boldsymbol{\gamma}}}\right),$$

kde $\boldsymbol{u}_{\boldsymbol{\beta}}\in\boldsymbol{R}^{\boldsymbol{K}}$, $\boldsymbol{u}_{\boldsymbol{\gamma}}\in\boldsymbol{R}^{\boldsymbol{M}}$, a označme

$$\tau = \tau(\boldsymbol{\beta}',\boldsymbol{\gamma}') = \log\omega, \qquad \hat{\tau} = \tau(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\gamma}}')' = \log\hat{\omega}.$$

Potom

$$Z = \frac{\hat{\tau}}{\sqrt{\widehat{\text{var}}\ \hat{\tau}}}. \tag{3.4}$$

Ukažme nejprve, že $\hat{\tau}$ je konzistentním odhadem τ. Z [3, str. 42] víme, že odhad $\hat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ metodou nejmenších čtverců je konzistentní, jestliže existuje taková matice $\mathbf{Q}$ typu $J \times J$, která je regulární ($h(\mathbf{Q}) = J$) a pro kterou platí

$$\lim_{n\to\infty} \frac{\mathbf{H}'\mathbf{H}}{n} = \mathbf{Q}. \tag{3.5}$$

Předpokládáme-li tedy, že vlastnost (3.5) je splněna, potom je odhad $\hat{\tau}$ parametru τ, jakožto spojitá funkce konzistentních odhadů, také konzistentní.

Dále vzhledem k tomu, že $\hat{\omega}$ je maximálně věrohodný odhad parametru ω, z již zmíněné Zehnaovy věty vyplývá, že $\hat{\tau}$ je, jakožto funkce $\hat{\omega}$, také maximálně věrohodný odhad. Potom vzhledem k tomu, že tento odhad je konzistentní, má odhad $\hat{\tau}$ za podmínky, že jsou splněny podmínky regularity, asymptoticky normální rozdělení (viz [1, str. 159–160] věta 7.100).

Střední hodnotu a rozptyl odhadu $\hat{\tau}$ získáme aproximací pomocí Taylorova rozvoje funkce analogicky jako při odvozování Williamsovy testové statistiky v předchozí kapitole. Vzhledem k tomu, že funkce τ má v nějakém okolí bodu

$$E(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\gamma}}')' = (\boldsymbol{\beta}',\boldsymbol{\gamma}')'$$

derivace všech řádů, můžeme ji v tomto bodě aproximovat Taylorovým polynomem prvního řádu, tedy

$$\tau(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\gamma}}')' \approx \tau(\boldsymbol{\beta}',\boldsymbol{\gamma}')' + \left(\begin{array}{c} \frac{\partial \tau(\boldsymbol{u}_\beta',\boldsymbol{u}_\gamma')'}{\partial \boldsymbol{u}_\beta} \\ \frac{\partial \tau(\boldsymbol{u}_\beta',\boldsymbol{u}_\gamma')'}{\partial \boldsymbol{u}_\gamma} \end{array} \right)' \Bigg|_{\boldsymbol{u}_\beta = \boldsymbol{\beta}, \boldsymbol{u}_\gamma = \boldsymbol{\gamma}} \left(\begin{array}{c} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{array} \right).$$

Odtud potom dostaneme následující dva vztahy pro aproximaci střední hodnoty a rozptylu odhadu $\hat{\tau}$. Pro střední hodnotu platí

$$E\hat{\tau} = E\left[\tau(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\gamma}}')'\right] \approx E\left[\tau(\boldsymbol{\beta}',\boldsymbol{\gamma}')'\right] = \log\left(\frac{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}}\right) = \tau$$

a pro rozptyl

$$\text{var } \hat{\tau} = \text{var } \left[\tau(\hat{\boldsymbol{\beta}}',\hat{\boldsymbol{\gamma}}')'\right] \approx \boldsymbol{w}'\text{var}\left(\begin{array}{c} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{array} \right)\boldsymbol{w},$$

kde

$$\boldsymbol{w} = 2\left(\begin{array}{c} \frac{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}} \\ \frac{-\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}} \end{array} \right),$$

neboť

$$\frac{\partial \tau}{\partial \boldsymbol{\beta}} = 2\frac{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}} \qquad \text{a} \qquad \frac{\partial \tau}{\partial \boldsymbol{\gamma}} = -2\frac{\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}}.$$

Z [1, str. 82] víme, že

$$\text{var}\left(\begin{array}{c} \hat{\alpha} \\ \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \\ \hat{\boldsymbol{\delta}} \end{array} \right) = \sigma^2(\boldsymbol{H}'\boldsymbol{H})^{-1}.$$

Varianční matici $\mathrm{var}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix}$ pak získáme z matice $\sigma^2(\mathbf{H}'\mathbf{H})^{-1}$ vynecháním prvního řádku a prvního sloupce a vynechání posledních L řádků a L sloupců.

Statistika $\hat{\tau} = \log \hat{\omega}$ má tedy asymptoticky normální rozdělení se střední hodnotou $\log(\omega)$ a rozptylem $\boldsymbol{w'V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}}\boldsymbol{w}$, kde

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}} = \mathrm{var}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix}.$$

Všimněme si však, že rozptyl $\hat{\tau} = \log \hat{\omega}$ závisí na parametru σ^2, který neznáme. Proto ho nahradíme jeho nestranným a zároveň konzistentním odhadem. Tím je, jak víme z předchozí kapitoly, náhodná veličina

$$\hat{\sigma}^2 = \frac{RSS}{n-J},$$

kde

$$RSS = (\boldsymbol{Y} - \boldsymbol{H}\hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \boldsymbol{H}\hat{\boldsymbol{\theta}}),$$

je reziduální součet čtverců modelu M.

Dosadíme-li na závěr do (3.4), dostaneme vyjádření testové statistiky pro test hypotézy H_0: $\omega = 1$, tedy

$$Z \doteq \frac{\log(\hat{\omega})}{\sqrt{\mathbf{w'}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}}\mathbf{w}}}.$$

Testová statistka Z má vzhledem k rozdělení $\hat{\tau}$ za platnosti nulové hypotézy H_0 asymptoticky normované normální rozdělení. Testujeme-li tedy hypotézu H_0: $\omega = 1$ např. proti alternativě H_1: $\omega \neq 1$, zamítneme H_0 na hladině α, když platí

$$|Z| \doteq \frac{|\log(\hat{\omega})|}{\sqrt{\mathbf{w'}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}}\mathbf{w}}} \geq u(1-\alpha/2),$$

kde $u(1-\alpha/2)$ značí $(1-\alpha/2)$-kvantil normovaného normálního rozdělení.

Poznámka 6. Všimněme si, že statistiku Z jsme odvodili obdobně jako Williamsovu statistiku T_2 v kapitole 2. Navrhli jsme statistiku, pomocí Taylorova rozvoje aproximovali její rozptyl a neznámý parametr σ^2, na kterém výsledný výraz závisel, nahradili jeho nestranným odhadem $\hat{\sigma}^2$. Z [1, str. 83] víme, že

$$\frac{\hat{\sigma}^2(n-J)}{\sigma^2} \sim \chi^2_{n-J}.$$

Proto o rozdělení statistiky Z můžeme přibližně mluvit jako o t-rozdělení s $n-J$ stupni volnosti. Asymptoticky pak samozřejmě dostaneme uvažované rozdělení $N(0,1)$. $\Diamond$

3.1 Nabídka prostředí R

V prostředí **R** je pro porovnání relativní důležitosti dvou skupin prediktorů v modelu lineární regrese k dispozici funkce `relimp` z knihovny `relimp`. Tuto funkci lze použít na objekt třídy `lm`, ale také na šest dalších tříd. Jmenujme zejména třídu `glm`, kam patří také logistická regrese. Argumenty funkce `relimp` jsou mimo jiné proměnné `set1` a `set2`, které obsahují vektory indexů prediktorů obsažených v první a druhé porovnávané skupině. Tuto funkci lze použít jak k porovnání dvou skupin prediktorů, tak také k porovnání dvojice prediktorů. Výsledkem je potom objekt třídy `relimp`, který je složen ze čtyř prvků: `model`, `sets`, `log.ratio` a `se.log.ratio`. Hodnota `log.ratio` je rovna polovině z námi počítané veličiny $\hat{\tau} = \log \hat{\omega}$, hodnota `se.log.ratio` je pak odhadem směrodatné odchylky odhadu `log.ratio` a je opět poloviční v porovnání s hodnotou odhadu směrodatné odchylky odhadu $\hat{\tau} = \log \hat{\omega}$.

Poznámka 7. Funkce `relimp` pracuje s poměrem výběrových směrodatných odchylek složek vektorů $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ a $\boldsymbol{Z}\hat{\boldsymbol{\gamma}}$, zatímco námi uvažovaný poměr $\hat{\omega}$ je poměrem výběrových rozptylů těchto vektorů. Označíme-li poměr výběrových směrodatných odchylek $\hat{\omega}^*$, potom zřejmě platí $\hat{\omega}^* = \sqrt{\hat{\omega}}$ a následně po zlogaritmování $\log \hat{\omega}^* = \frac{1}{2} \log \hat{\omega}$ a $\sqrt{\mathrm{var}(\log \hat{\omega}^*)} = \frac{1}{2}\sqrt{\mathrm{var}(\log \hat{\omega})}$. Proto jsou výsledky funkce `relimp` o polovinu menší než naše. $\Diamond$

Kapitola 4

Porovnání testů rovnosti dvou korelačních koeficientů

Naším úkolem v této kapitole bude porovnat testy rovnosti dvou korelačních koeficientů odvozené v teoretické části textu. Uvažovat budeme tedy test Hotellingův, Williamsův a test poměrem věrohodností. Testy porovnáme na základě simulací, při nichž budeme zkoumat, jak dodržují hladinu významnosti α a jakou mají sílu.

4.1 Nastavení simulací

Popišme, jak budeme při porovnávání testů postupovat. Hladinu významnosti α testů i jejich sílu budeme odhadovat tak, že 1000krát vygenerujeme náhodný výběr (Y_i, X_{1i}, X_{2i}), $1 \leq i \leq n$, z trojrozměrného normálního rozdělení se střední hodnotou $\boldsymbol{\mu}$ a varianční maticí $\mathbf{V}$. Pro každý takový vygenerovaný náhodný výběr provedeme uvažované testy nulové hypotézy. Za odhady hladiny významnosti, resp. za odhady síly testů, potom vezmeme relativní četnosti zamítnutí hypotéz jednotlivými testy. Ke generování náhodných veličin použijeme statistický software **R 2.13.0**.

V celé simulaci je možné volit následující parametry: hladinu testů α, počet pozorování n, střední hodnoty a rozptyly náhodných veličin Y, X_1 a X_2 a korelační koeficienty ρ_{01}, ρ_{02} a ρ_{12}. Hladinu testů α zvolíme rovnu 0,05. Střední hodnotu $(Y, X_1, X_2)'$ položíme bez újmy na obecnosti rovnu nulovému vektoru. Vektor směrodatných odchylek $\boldsymbol{\sigma}$= $(\sigma_0, \sigma_1, \sigma_2)$ položíme roven $(1, 2, 3)$. Při zkoumání dodržování hladiny testů a jejich síly nás bude zajímat, zda a jak jejich hodnoty závisí na hodnotách korelačních koeficientů ρ_{01}, ρ_{02} a ρ_{12}, případně zda jsou ovlivněny počtem pozorování n. Zkoumáme-li dodržování hladiny významnosti, předpokládáme, že platí hypotéza H_0: $\rho_{01} = \rho_{02}$, volíme tedy korelační koeficienty ρ_{01} a ρ_{02} shodné. Zkoumáme-li, jakou mají testy sílu, předpokládáme, že platí alternativní hypotéza H_1: $\rho_{01} \neq \rho_{02}$, volíme tedy korelační koeficienty ρ_{01} a ρ_{02} rozdílné. Hodnoty korelačních koeficientů ρ_{01}, ρ_{02} a ρ_{12} budeme volit v rozmezí intervalu $(-1, 1)$, a to tak, aby matice $\mathbf{P}$ byla pozitivně definitní, tedy aby hodnota determinantu $|\mathbf{P}| = 1 - \rho_{12}^2 - \rho_{01}^2 - \rho_{02}^2 + 2\rho_{01}\rho_{02}\rho_{12}$ matice $\mathbf{P}$ byla kladná. Počet pozorování n budeme volit v rozmezí mezi 10 a 150.

Poznámka 8. Podle Sylvestrova kritéria je matice $\mathbf{A}_{n\times n}$ pozitivně definitní, pokud jsou všechny determinanty matic, které vznikly vynecháním posledních $n-i$,

$1 \leq i \leq n$, řádků a sloupců, kladné. V našem případě je $|\mathbf{P}_0| = 1$, $|\mathbf{P}_1| = 1$ a $|\mathbf{P}_2| = 1 - \rho_{01}^2 > 0$, neboť $\rho_{01} \in (-1, 1)$. K pozitivní definitnosti matice $\mathbf{P}$ tedy opravdu stačí ověřit pouze, že $|\mathbf{P}_3| = |\mathbf{P}| > 0$. $\Diamond$

Abychom zjistili, zda daný test dodržuje požadovanou hladinu $\alpha_0 = 0{,}05$, budeme testovat hypotézu

$$H_0^\alpha : \alpha = \alpha_0 \quad \text{proti oboustranné alternativě} \quad H_1^\alpha : \alpha \neq \alpha_0,$$

kde α označuje skutečnou hladinu významnosti daného testu. Odhad této skutečné hladiny jsme získali tak, že jsme 1000krát vygenerovali náhodný výběr (Y_i, X_{1i}, X_{2i}), $1 \leq i \leq n$, a pro každý takový vygenerovaný náhodný výběr provedli uvažované testy nulové hypotézy H_0: $\rho_{01} = \rho_{02}$. Ty testovanou hypotézu buď zamítly, nebo nezamítly. K dispozici tedy máme nyní náhodný výběr z alternativního rozdělení a k testu hypotézy H_0^α můžeme použít testovou statistiku

$$T_\alpha = \frac{|\hat{\alpha} - \alpha_0|}{\sqrt{\alpha_0(1 - \alpha_0)}}\sqrt{1000},$$

která má za platnosti nulové hypotézy H_0^α asymptoticky normované normální rozdělení (viz [4, str. 131]). Pokud tento test hypotézu H_0^α zamítne, budeme říkat, že test nedodržuje hladinu významnosti. Jestliže je hodnota odhadu $\hat{\alpha}$ menší než $\alpha_0 - u(1 - \alpha_0/2)\sqrt{\frac{\alpha_0(1-\alpha_0)}{n}} \doteq 0{,}0365$, potom test hladinu významnosti podhodnocuje. Takovýto test nazveme konzervativní. Naopak jestliže je hodnota $\hat{\alpha}$ větší než $\alpha_0 + u(1 - \alpha_0/2)\sqrt{\frac{\alpha_0(1-\alpha_0)}{n}} \doteq 0{,}0635$, potom test hladinu významnosti nadhodnocuje. Takovýto test nazveme antikonzervativní.

Odhady hladiny významnosti a síly testů budeme ukládat do tabulek a v grafech znázorníme případnou závislost těchto odhadů na hodnotách korelačních koeficientů a na počtech pozorování. Tak lépe uvidíme rozdíly mezi jednotlivými testy a budeme je moci lépe porovnat. V případě, že některý z testů bude nadhodnocovat hladinu významnosti, vyznačíme v tabulce příslušnou hodnotu odhadu červeně. Pokud některý z testů bude naopak hladinu významnosti podhodnocovat, vyznačíme příslušnou hodnotu odhadu v tabulce modře. V tabulkách obsahujících odhady síly testů označíme tučně ten test, který bude pro danou volbu parametrů nejsilnější. Pokud bude mít tuto vlastnost více testů, označíme tučně všechny tyto testy. V případě, že některá volba parametrů nebude přípustná, tj. korelační matice $\mathbf{P}$ nebude pozitivně definitní, nahradíme v tabulce příslušný odhad pomlčkou. Na závěr poznamenejme, že v tabulkách používáme zkratku ML pro označení testu poměrem věrohodností, H pro test Hotellingův a W pro Williamsův test.

4.2 Výsledky simulací

4.2.1 Dodržování hladiny významnosti

Nyní se tedy pokusíme uvažované testy porovnat podle toho, jak dodržují hladinu významnosti. Nejprve se podívejme, zda je hladina testů ovlivněna hodnotou korelačního koeficientu ρ_{12}. Výsledky simulací pro různé hodnoty korelačních

koeficientů ρ_{01}, ρ_{02}, ρ_{12} a pro $n = 100$ nalezneme v tabulce 4.1. Z té vyplývá, že hladina Williamsova testu a testu poměrem věrohodností se s hodnotou korelačního koeficientu ρ_{12} nemění. Oba testy hladinu významnosti dodržují. Nicméně hodnota odhadu hladiny testu poměrem věrohodností je vždy nepatrně větší než hodnota odhadu hladiny Williamsova testu. To je patrné i z grafu na obrázku 4.1, kde je znázorněna závislost hladiny testů na korelačním koeficientu ρ_{12} pro $\rho_{01} = \rho_{02} = 0{,}4$ a $n = 100$. Z tohoto obrázku, stejně tak jako z dříve uvedené tabulky, také zřetelně vidíme, že hladina významnosti Hotellingova testu se s hodnotou ρ_{12} mění. Zhruba můžeme říci, že čím je hodnota tohoto koeficientu menší, tím je počet zamítnutí nulové hypotézy větší. S rostoucí hodnotou korelačního koeficientu ρ_{12} se počet zamítnutí nulové hypotézy Hotellingovým testem blíží počtu zamítnutí testem Williamsovým.

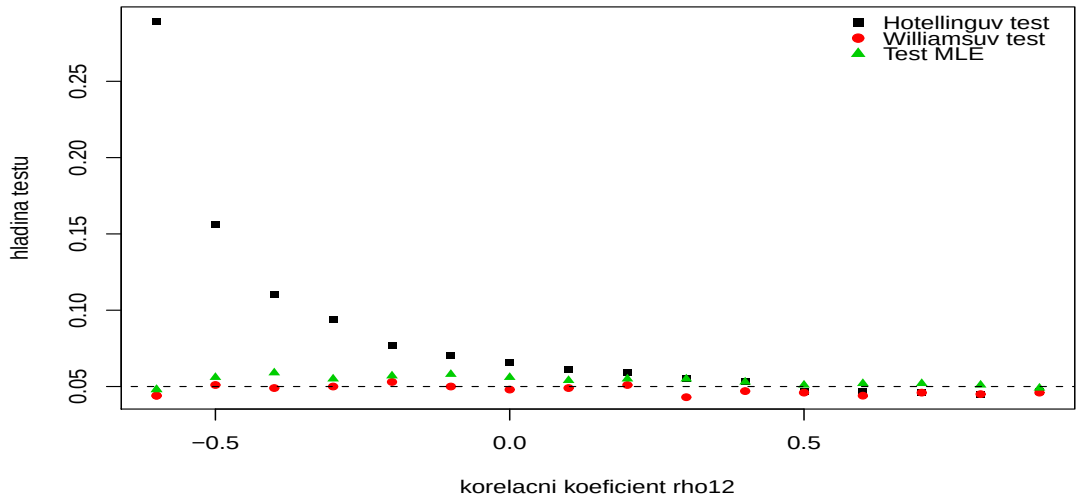
Tabulka 4.1: Výsledky simulací zkoumajících dodržování hladiny významnosti $\alpha = 0{,}05$ pro $n = 100$ a různé hodnoty ρ_{01}, ρ_{02} (zkoumáme závislost na korelačním koeficientu ρ_{12}).

	$\rho_{01} = \rho_{02} = 0{,}1$			$\rho_{01} = \rho_{02} = 0{,}4$			$\rho_{01} = \rho_{02} = 0{,}7$		
ρ_{12}	ML	H	W	ML	H	W	ML	H	W
-0,9	0,053	**0,075**	0,051	-	-	-	-	-	-
-0,8	0,051	0,058	0,049	-	-	-	-	-	-
-0,7	0,051	0,054	0,045	-	-	-	-	-	-
-0,6	0,052	0,052	0,045	0,048	**0,289**	0,044	-	-	-
-0,5	0,052	0,051	0,046	0,056	**0,156**	0,051	-	-	-
-0,4	0,052	0,051	0,047	0,059	**0,110**	0,049	-	-	-
-0,3	0,052	0,051	0,048	0,055	**0,094**	0,050	-	-	-
-0,2	0,052	0,051	0,049	0,057	**0,077**	0,053	-	-	-
-0,1	0,052	0,051	0,049	0,058	**0,070**	0,050	-	-	-
0,0	0,051	0,050	0,048	0,056	**0,066**	0,048	0,054	**0,601**	0,049
0,1	0,052	0,049	0,048	0,054	0,061	0,049	0,052	**0,221**	0,047
0,2	0,053	0,049	0,049	0,055	0,059	0,051	0,056	**0,138**	0,050
0,3	0,052	0,048	0,048	0,055	0,055	0,043	0,057	**0,097**	0,054
0,4	0,053	0,048	0,048	0,053	0,053	0,047	0,058	**0,071**	0,054
0,5	0,052	0,047	0,046	0,051	0,047	0,046	0,055	0,061	0,052
0,6	0,053	0,045	0,045	0,052	0,047	0,044	0,053	0,056	0,046
0,7	0,053	0,045	0,045	0,052	0,046	0,046	0,053	0,053	0,047
0,8	0,052	0,046	0,046	0,051	0,045	0,045	0,049	0,046	0,046
0,9	0,048	0,047	0,047	0,049	0,047	0,046	0,052	0,048	0,047

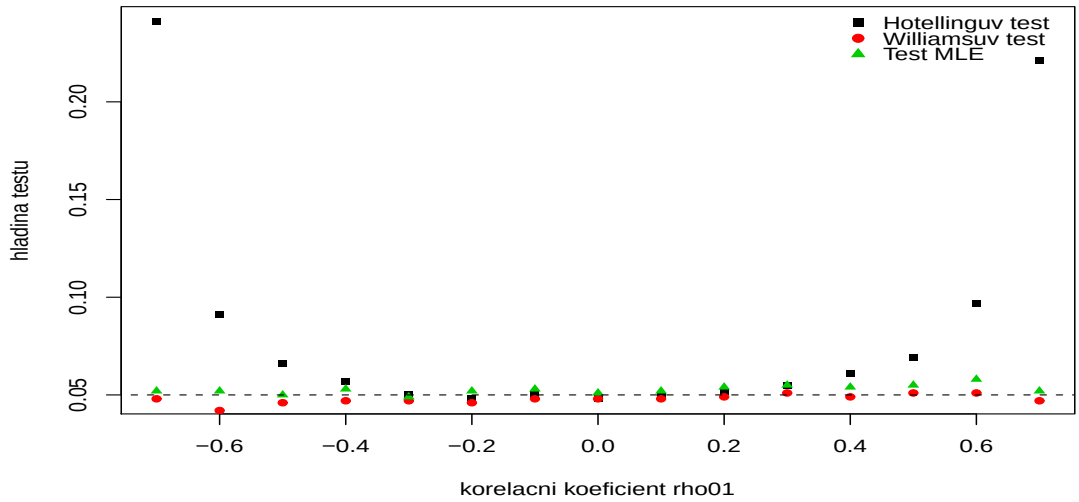
Tabulka 4.2: Výsledky simulací zkoumajících dodržování hladiny významnosti $\alpha = 0{,}05$ pro $\rho_{12} = 0{,}1$ a $n = 100$ (zkoumáme závislost na korelačním koeficientu ρ_{01}, resp. ρ_{02}).

$\rho_{01} = \rho_{02}$	ML	H	W	$\rho_{01} = \rho_{02}$	ML	H	W
-0,9	-	-	-	0,1	0,052	0,049	0,048
-0,8	-	-	-	0,2	0,054	0,052	0,049
-0,7	0,052	**0,241**	0,048	0,3	0,055	0,055	0,051
-0,6	0,052	**0,091**	0,042	0,4	0,054	0,061	0,049
-0,5	0,050	**0,066**	0,046	0,5	0,055	**0,069**	0,051
-0,4	0,053	0,057	0,047	0,6	0,058	**0,097**	0,051
-0,3	0,049	0,050	0,047	0,7	0,052	**0,221**	0,047
-0,2	0,052	0,048	0,046	0,8	-	-	-
-0,1	0,053	0,050	0,048	0,9	-	-	-

Dále si v tabulce 4.1, případně v tabulce 4.2, můžeme všimnout, že hladina Hotellingova testu se mění také s hodnotou korelačních koeficientů ρ_{01} a ρ_{02}, tedy je závislá také na těchto korelačních koeficientech. Jak ze zmíněných tabulek, tak z grafu na obrázku 4.2, vyplývá, že hladina významnosti Hotellingova testu roste s rostoucí absolutní hodnotou korelačního koeficientu ρ_{01}, resp. korelačního koeficientu ρ_{02}. Čím více se hodnota těchto korelačních koeficientů přibližuje nule, tím lépe Hotellingův test dodržuje hladinu významnosti. Z grafu na obrázku 4.2 dále vidíme, že ani hladina Williamsova testu ani testu poměrem věrohodností na hodnotách korelačních koeficientů ρ_{01} a ρ_{02} nezávisí. Oba tyto testy ve všech testovaných případech hladinu významnosti dodržují.



Obrázek 4.1: Grafické znázornění závislosti hladiny testů na korelačním koeficientu ρ_{12} pro $\rho_{01} = \rho_{02} = 0{,}4$ a $n = 100$.



Obrázek 4.2: Grafické znázornění závislosti hladiny testů na korelačním koeficientu ρ_{01}, resp. ρ_{02}, pro $\rho_{12} = 0{,}1$ a $n = 100$.

Před tím než přejdeme k výsledkům dalších simulací, pokusme se vysvětlit závislost Hotellingova testu na korelačních koeficientech ρ_{01}, ρ_{02} a ρ_{12}. Z teoretické části práce víme, že Hotellingův test testuje nulovou hypotézu H_0: $\rho_{01} = \rho_{02}$ pomocí testové statistiky

$$T_1 = \frac{V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})}{\sqrt{\widehat{\text{var}}\left[V(\mathbf{Y}, \mathbf{x_1}, \mathbf{x_2})\right]}} = \frac{r_{01} - r_{02}}{\sqrt{\frac{2|\mathbf{R}|}{(1+r_{12})(n-3)}}},$$

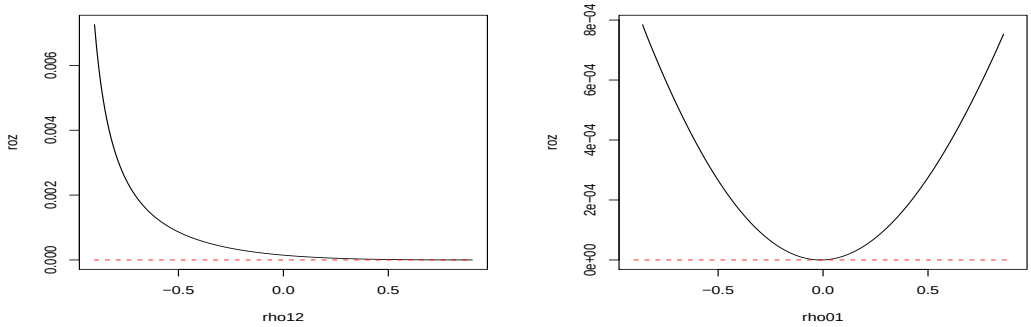
Williamsův test pomocí testové statistiky

$$T_2 = \frac{V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})}{\sqrt{\widehat{\text{var}}\left[V(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2})\right]}} = \frac{r_{01} - r_{02}}{\sqrt{\frac{2|\mathbf{R}|}{(1+r_{12})(n-3)} + \frac{\bar{r}^2(1-r_{12})^3}{(1+r_{12})(n-1)}}}.$$

Obě testové statistiky porovnáváme při testování nulové hypotézy s $(1-\alpha/2)$-kvantilem t-rozdělení s $n-3$ stupni volnosti. Na první pohled je zřejmé, že jmenovatel testové statistiky T_1 je menší než jmenovatel testové statistiky T_2, tedy absolutní hodnota testové statistiky T_1 je větší než absolutní hodnota testové statistiky T_2. Jak moc se absolutní hodnoty těchto dvou statistik liší, závisí na hodnotě výrazu

$$\frac{\bar{r}^2(1-r_{12})^3}{(1+r_{12})(n-1)}. \tag{4.1}$$

Hodnota tohoto výrazu klesá s rostoucí hodnotou korelačního koeficientu ρ_{12} k nule (viz první z grafů na obrázku 4.3) a roste s rostoucí absolutní hodnotou korelačních koeficientů ρ_{01} a ρ_{02} (viz druhý z grafů na obrázku 4.3). S rostoucí hodnotou výrazu (4.1) roste hodnota jmenovatele Williamsovy testové statistiky a zvyšuje se rozdíl mezi absolutní hodnotou Wiliamsovy a Hotellingovy testové statistiky. Obě testové statistiky nicméně porovnáváme se stejnou hodnotou $(1-\alpha/2)$-kvantilu. Proto se stává, že Williamsův test nulovou hypotézu nezamítne a Hotellingův test ano, což způsobuje, že Hotellingův test v popsaných případech nedodržuje hladinu významnosti.



Obrázek 4.3: Grafické znázornění závislosti hodnoty výrazu (4.1) na hodnotě korelačního koeficientu ρ_{12} (graf VLEVO) a na hodnotě korelačního koeficientu ρ_{01}, resp. ρ_{02} (graf VPRAVO).

Vidíme tedy, že zahrnutí variability náhodných vektorů $\mathbf{X_1}$ a $\mathbf{X_2}$ do úvah při odvozování testové statistiky testující nulovou hypotézu H_0: $\rho_{01} = \rho_{02}$ je pro

spolehlivost testu nezbytné.

Podívejme se dále, zda je dodržování hladiny významnosti testů ovlivněno počtem pozorování. Výsledky provedených simulací shrnuje tabulka 4.3. Na první pohled je zřejmé, že na počtu pozorování závisí test poměrem věrohodností. Pokud je počet pozorování dostatečně velký, nemá test poměrem věrohodností s dodržením hladiny významnosti problém, avšak pokud počet pozorování zmenšíme, řekněme na třicet, začne tento test hladinu významnosti výrazně nadhodnocovat. Čím je potom počet pozorování menší, tím je hladina tohoto testu více nadhodnocena. Hladina Williamsova a Hotellingova testu se zdá býti ovlivněna počtem pozorování zřetelně méně. Hotellingův test dodržuje, resp. nedodržuje, hladinu významnosti v případech popsaných dříve v závislosti na hodnotách korelačních koeficientů. Williamsův test dodržuje hladinu významnosti téměř ve všech testovaných případech, výjimkou jsou ty, kdy je počet pozorování velmi malý. Pokud je n rovno deseti, Williamsův test hladinu významnosti podhodnocuje.

Tabulka 4.3: Výsledky simulací zkoumajících dodržování hladiny významnosti $\alpha = 0{,}05$ pro různé hodnoty korelačních koeficientů ρ_{01}, ρ_{02} a ρ_{12} (zkoumáme závislost na počtu pozorování n).

	$\rho_{01} = \rho_{02} = 0{,}1$ $\rho_{12} = 0{,}1$			$\rho_{01} = \rho_{02} = 0{,}4$ $\rho_{12} = 0{,}5$			$\rho_{01} = \rho_{02} = 0{,}7$ $\rho_{12} = 0{,}3$		
n	ML	H	W	ML	H	W	ML	H	W
150	0,057	0,052	0,052	0,055	0,055	0,053	0,062	**0,097**	0,057
140	0,059	0,058	0,058	0,057	0,056	0,056	0,059	**0,100**	0,055
130	0,046	0,045	0,045	0,049	0,047	0,045	0,053	**0,088**	0,050
120	0,057	0,053	0,052	0,056	0,054	0,052	0,052	**0,098**	0,052
110	0,055	0,050	0,050	0,055	0,053	0,049	0,047	**0,087**	0,045
100	0,052	0,049	0,048	0,051	0,047	0,046	0,057	**0,097**	0,054
90	0,056	0,054	0,051	0,056	0,051	0,050	0,048	**0,096**	0,045
80	0,046	0,040	0,039	0,045	0,043	0,043	0,041	**0,077**	0,038
70	0,056	0,047	0,047	0,054	0,051	0,049	0,046	**0,082**	0,042
60	0,052	0,048	0,048	0,051	0,047	0,045	0,063	**0,099**	0,051
50	0,062	0,051	0,050	**0,064**	0,055	0,052	0,062	**0,102**	0,056
40	0,047	0,043	0,040	0,051	0,047	0,044	0,043	**0,079**	**0,032**
30	**0,069**	0,056	0,054	**0,065**	0,055	0,054	**0,069**	**0,093**	0,058
20	**0,070**	0,048	0,044	**0,070**	0,053	0,051	**0,066**	**0,085**	0,043
10	**0,103**	0,042	0,037	**0,103**	0,044	0,040	**0,101**	**0,086**	**0,031**

Jak už jsme poznamenali dříve, hodnota odhadu hladiny testu poměrem věrohodností je vždy nepatrně větší než hodnota odhadu hladiny Williamsova testu. Pokusme se nyní vysvětlit, proč tomu tak je. Porovnejme za tímto účelem hodnoty statistiky T_2^2 a T_3. V tabulce 4.4 nalezneme pro $n = 100$ a pro různé hodnoty korelačních koeficientů ρ_{01}, ρ_{02} a ρ_{12} poměry počtu simulací, pro něž je $T_3 < T_2^2$, a celkového počtu provedených simulací. Z této tabulky vyplývá, že v drtivé většině případů je hodnota testové statistiky T_3 větší než druhá mocnina testové statistiky T_2. Jak testová statistika T_3, tak T_2^2, mají asymptoticky χ_1^2-rozdělení. Porovnávali-li bychom hodnoty T_3 a T_2^2 s $(1-\alpha)$-kvantilem χ_1^2-rozdělení, mohlo by se stát, že by test poměrem věrohodností nulovou hypotézu zamítl víckrát než Williamsův test. Z tohoto důvodu hodnota odhadu hladiny testu poměrem věrohodností je nepatrně větší než hodnota odhadu hladiny Williamsova testu,

a to i přesto, že my porovnáváme testovou statistiku T_2 s kvantilem $t_{n-3}(1-\alpha/2)$.

Tabulka 4.4: Poměry počtu simulací, pro něž je $T_3 < T_2^2$, a celkového počtu provedených simulací pro $n = 100$ a pro různé hodnoty korelačních koeficientů ρ_{01}, ρ_{02} a ρ_{12}.

ρ_{12}	$\rho_{01} = 0{,}1$	$\rho_{01} = 0{,}4$	ρ_{12}	$\rho_{01} = 0{,}1$	$\rho_{01} = 0{,}4$	$\rho_{01} = 0{,}7$
-0,9	0,019	-	0,1	0,011	0,011	0,000
-0,8	0,014	-	0,2	0,011	0,010	0,001
-0,7	0,013	-	0,3	0,013	0,010	0,002
-0,6	0,013	0,019	0,4	0,013	0,011	0,003
-0,5	0,012	0,007	0,5	0,013	0,012	0,005
-0,4	0,011	0,006	0,6	0,014	0,013	0,009
-0,3	0,011	0,007	0,7	0,014	0,012	0,011
-0,2	0,010	0,006	0,8	0,015	0,012	0,011
-0,1	0,011	0,007	0,9	0,013	0,013	0,012

Dodržování hladiny významnosti - shrnutí

Provedené simulace ukázaly, že bez ohledu na hodnotu korelačních koeficientů dodržuje hladinu významnosti Williamsův test a test poměrem věrohodností. Na oba tyto testy se můžeme spolehnout, pokud máme dostatečně velký počet pozorování. Pro test poměrem věrohodností je potřeba míti k dispozici alespoň 40 pozorování, pro Williamsův test je to okolo 20. Pokud je počet pozorování příliš malý, test poměrem věrohodností hladinu významnosti nadhodnocuje, Williamsův test podhodnocuje. Poznamenejme dále, že ve všech zkoumaných případech byla hladina Williamsova testu vždy nepatrně nižší než hladina testu poměrem věrohodností.

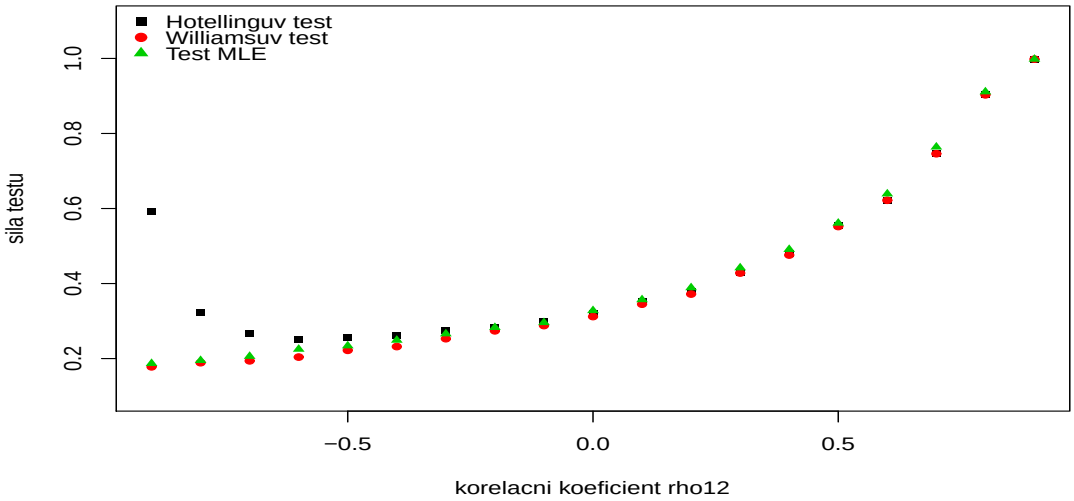
Hotellingův test se ukázal použitelný jen v omezeném počtu případů (viz popis výše), což je způsobeno již několikrát zmíněnou podmíněností tohoto testu.

4.2.2 Porovnání síly testů

V této části se soustředíme na porovnávání testů z hlediska jejich síly. Věnujme se nejprve opět závislosti síly testů na hodnotě korelačního koeficientu ρ_{12}. Z výsledků simulací, které nalezneme v tabulce 4.5, vyplývá, že na hodnotě korelačního koeficientu ρ_{12} je závislá síla všech uvažovaných testů. Tato závislost je zřetelně viditelná i z grafického znázornění na obrázku 4.4. Pokud je hodnota koeficientu ρ_{12} záporná, je nejsilnějším z uvažovaných testů Hotellingův test, avšak, jak jsme zjistili v předchozí části textu, Hotellingův test právě v těchto případech nedodržuje hladinu významnosti, proto ho není možné s ostatními testy porovnávat. Druhým nejsilnějším testem, pokud je ρ_{12} záporné, a nejsilnějším testem, pokud je ρ_{12} kladné, je test poměrem věrohodností. Jak z tabulky 4.5, tak z grafu na obrázku 4.4, vyplývá, že síla testu poměrem věrohodností je vždy o něco větší než síla Williamsova testu, přičemž síla obou těchto testů roste s rostoucí hodnotou koeficientu ρ_{12}.

Tabulka 4.5: Výsledky simulací zkoumajících sílu testů pro $n = 100$ a různé hodnoty korelačních koeficientů ρ_{01} a ρ_{02} (zkoumáme závislost na korelačním koeficientu ρ_{12}).

	$\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}3$			$\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}5$			$\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}7$		
ρ_{12}	ML	H	W	ML	H	W	ML	H	W
-0,9	0,187	**0,593**	0,178	-	-	-	-	-	-
-0,8	0,195	**0,323**	0,189	0,572	**0,965**	0,657	-	-	-
-0,7	0,206	**0,266**	0,194	0,683	**0,866**	0,673	-	-	-
-0,6	0,225	**0,252**	0,204	0,707	**0,820**	0,696	0,860	**1,000**	0,989
-0,5	0,234	**0,255**	0,222	0,734	**0,802**	0,718	0,994	**0,999**	0,992
-0,4	0,249	**0,262**	0,232	0,759	**0,803**	0,746	0,996	**0,998**	0,996
-0,3	0,266	**0,275**	0,253	0,778	**0,814**	0,768	0,997	**0,998**	0,997
-0,2	0,283	**0,284**	0,274	0,817	**0,834**	0,806	0,997	**0,998**	0,997
-0,1	0,296	**0,300**	0,288	0,848	**0,856**	0,839	**0,999**	**0,999**	**0,999**
0,0	**0,328**	0,321	0,312	0,883	**0,888**	0,874	**0,999**	**0,999**	**0,999**
0,1	**0,357**	0,352	0,345	**0,914**	0,913	0,907	**1,000**	**1,000**	0,999
0,2	**0,389**	0,379	0,372	**0,938**	**0,938**	0,934	**1,000**	**1,000**	**1,000**
0,3	**0,442**	0,430	0,428	**0,965**	0,965	0,961	**1,000**	**1,000**	**1,000**
0,4	**0,491**	0,480	0,476	**0,980**	0,979	0,978	**1,000**	**1,000**	**1,000**
0,5	**0,561**	0,554	0,552	**0,998**	0,996	0,996	**1,000**	**1,000**	**1,000**
0,6	**0,639**	0,622	0,622	**0,999**	**0,999**	**0,999**	**1,000**	**1,000**	**1,000**
0,7	**0,764**	0,747	0,746	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**
0,8	**0,911**	0,903	0,903	**1,000**	**1,000**	**1,000**	-	-	-
0,9	**0,998**	**0,998**	**0,998**	**1,000**	**1,000**	**1,000**	-	-	-

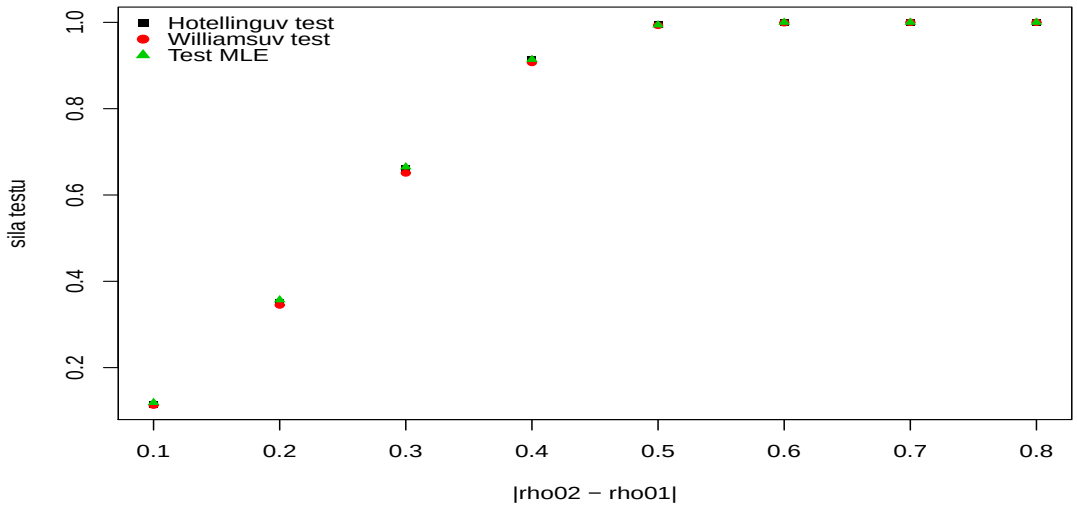


Obrázek 4.4: Grafické znázornění závislosti síly testů na korelačním koeficientu ρ_{12} pro $\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}3$ a $n = 100$.

V tabulce 4.5, případně v tabulce 4.6, si dále můžeme všimnout, že síla testů roste také s rostoucí absolutní hodnotou rozdílu korelačních koeficientů ρ_{01} a ρ_{02}. Tato závislost by pro nás však neměla být nijak překvapivá. Je zřejmá již z vyjádření čitatele Hotellingovy, resp. Williamsovy statistiky. Grafické znázornění závislosti síly testů na absolutní hodnotě rozdílu korelačních koeficientů ρ_{01} a ρ_{02} nalezneme na obrázku 4.5.

Tabulka 4.6: Výsledky simulací zkoumajících sílu testů pro $n = 100$ a různé hodnoty korelačního koeficientu ρ_{12} (zkoumáme závislost na absolutní hodnotě rozdílu korelačních koeficientů ρ_{01} a ρ_{02}).

		$\rho_{12} = 0{,}1$			$\rho_{12} = 0{,}4$			$\rho_{12} = 0{,}7$		
ρ_{01}	ρ_{02}	ML	H	W	ML	H	W	ML	H	W
0,1	0,2	**0,119**	0,115	0,113	**0,156**	0,145	0,145	**0,274**	0,264	0,264
0,1	0,3	**0,357**	0,352	0,345	**0,491**	0,480	0,476	**0,764**	0,747	0,746
0,1	0,4	**0,665**	0,661	0,651	**0,831**	0,819	0,818	**0,989**	0,988	0,988
0,1	0,5	**0,914**	0,913	0,907	**0,980**	0,979	0,978	**1,000**	**1,000**	**1,000**
0,1	0,6	0,994	**0,995**	0,993	**0,999**	**0,999**	**0,999**	**1,000**	**1,000**	**1,000**
0,1	0,7	**1,000**	**1,000**	0,999	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**
0,1	0,8	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	-	-	-
0,1	0,9	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	-	-	-

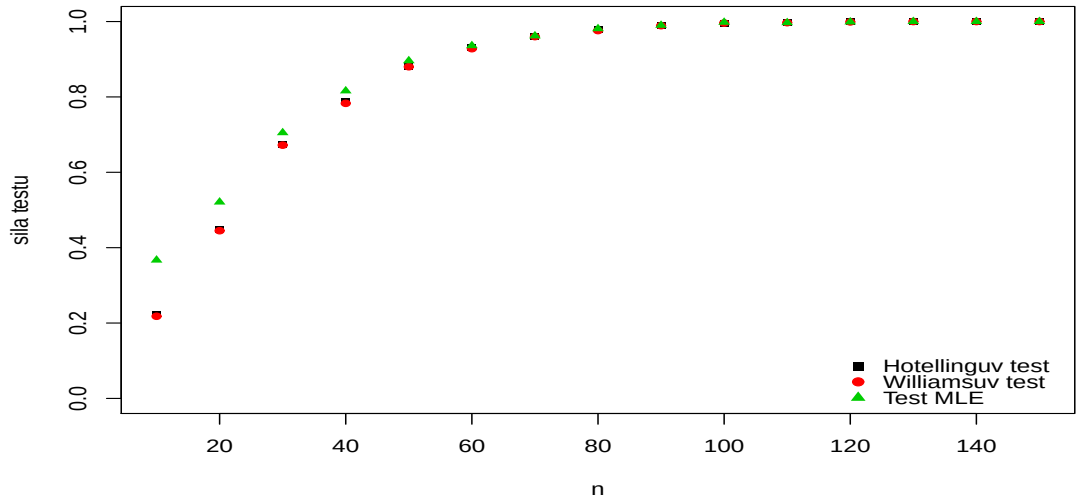


Obrázek 4.5: Grafické znázornění závislosti síly testů na absolutní hodnotě rozdílu korelačních koeficientů ρ_{01} a ρ_{02} pro $\rho_{12} = 0{,}1$ a $n = 100$.

Tabulka 4.7: Výsledky simulací zkoumajících sílu testů pro $\rho_{12} = 0{,}5$ a různé hodnoty korelačních koeficientů ρ_{01} a ρ_{02} (zkoumáme závislost na počtu pozorování n).

	$\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}3$			$\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}5$			$\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}7$		
n	ML	H	W	ML	H	W	ML	H	W
150	**0,714**	0,711	0,711	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**
140	**0,697**	0,687	0,683	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**
130	**0,671**	0,659	0,658	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**
120	**0,632**	0,617	0,616	**0,999**	**0,999**	**0,999**	**1,000**	**1,000**	**1,000**
110	**0,605**	0,592	0,589	**0,997**	**0,997**	**0,997**	**1,000**	**1,000**	**1,000**
100	**0,561**	0,554	0,552	**0,998**	**0,996**	**0,996**	**1,000**	**1,000**	**1,000**
90	**0,561**	0,554	0,552	**0,990**	**0,990**	0,989	**1,000**	**1,000**	**1,000**
80	**0,518**	0,505	0,502	**0,981**	0,980	0,976	**1,000**	**1,000**	**1,000**
70	**0,458**	0,441	0,438	**0,962**	0,960	0,960	**1,000**	0,999	0,999
60	**0,422**	0,406	0,405	**0,936**	0,931	0,928	**1,000**	**1,000**	**1,000**
50	**0,311**	0,292	0,292	**0,896**	0,881	0,880	**1,000**	**1,000**	**1,000**
40	**0,273**	0,245	0,244	**0,816**	0,787	0,783	**1,000**	**1,000**	0,999
30	**0,232**	0,195	0,192	**0,705**	0,674	0,672	**0,993**	0,990	0,990
20	**0,176**	0,113	0,112	**0,521**	0,449	0,445	**0,946**	0,928	0,924
10	**0,171**	0,088	0,082	**0,367**	0,224	0,218	**0,758**	0,612	0,599

Podívejme se ještě na závěr, zda síla testů závisí na počtu pozorování. Z tabulky 4.7, resp. z obrázku 4.6, vyplývá, že síla všech uvažovaných testů roste s rostoucím počtem pozorování. Pro jakýkoli počet pozorování je nejsilnějším testem test poměrem věrohodností. Pokud je počet pozorování velký, je rozdíl mezi silami testů poměrně malý. Pro menší počet pozorování je rozdíl mezi silou testu poměrem věrohodností a silou Williamsova testu, resp. Hotellingova testu, o něco zřetelnější. Připomeňme však, že pro malý počet pozorování test poměrem věrohodností výrazně nadhodnocoval hladinu významnosti. Proto porovnávání sil testů, pokud je počet pozorování malý, není korektní.



Obrázek 4.6: Grafické znázornění závislosti síly testů na počtu pozorování n pro $\rho_{01} = 0{,}1$, $\rho_{02} = 0{,}5$ a $\rho_{12} = 0{,}5$.

Porovnání síly testů - shrnutí

Jelikož má smysl porovnávat pouze testy, které dodržují hladinu významnosti, vynecháme ze souhrnného porovnání Hotellingův test, který v mnoha testovaných případech hladinu nadhodnocoval. Ze zbylých dvou uvažovaných testů se ukázal silnější test poměrem věrohodností. Ten však, pokud byl počet pozorování malý, nadhodnocoval hladinu významnosti. Proto, pokud máme k dispozici pouze malý počet pozorování, měli bychom dát raději přednost Williamsovu testu, který hladinu významnosti dodržuje v těchto případech o něco lépe.

Z provedených simulací dále vyplývá, že síla všech testů roste s rostoucí hodnotou korelačního koeficientu ρ_{12}, rostoucí absolutní hodnotou rozdílu korelačních koeficientů ρ_{01} a ρ_{02} a také s rostoucím počtem pozorování n.

Kapitola 5

Porovnání relativní důležitosti prediktorů - numerická demonstrace

K numerické demonstraci teorie obsažené v kapitolách 2 a 3 použijeme data pocházející z [8]. Jedná se o údaje o 102 studentech bakalářského studijního oboru Geografie–kartografie Přírodovědecké fakulty Univerzity Karlovy v Praze, kteří se zapsali ke studiu v akademickém roce 2003/04. Datový soubor, který máme k dispozici, obsahuje následující informace:

název veličiny	vysvětlení
`pr1 - pr3`	průměry známek za 1. až 3. rok bakalářského studia
`prumer`	celkový průměr známek za bakalářské studium
`matem`	body za písemku z matematiky u přijímacích zkoušek
`zem`	body za písemku ze zeměpisu u přijímacích zkoušek
`ss1 - ss4`	průměrné známky posledních čtyř let střední školy
`Pohlaví`	označení pohlaví studenta (1 - žena, 2 - muž)
`zcel`	celkový počet bodů u přijímacích zkoušek

5.1 Testy rovnosti dvou korelačních koeficientů

Naším prvním úkolem bude rozhodnout, zda průměr známek za bakalářské studium je lépe vysvětlen průměrem z průměrných známek posledních čtyř let střední školy nebo výsledky přijímacích zkoušek. Za tímto účelem budeme testovat hypotézu

$$H_0 : \quad \rho_{01} = \rho_{02},$$

kde ρ_{01} je korelační koeficient náhodných veličin *prumer* a *zcel* a ρ_{02} je korelační koeficient náhodných veličin *prumer* a *ssprumer*. Náhodná veličina *ssprumer* vyjadřuje průměr z průměrných známek posledních čtyř let střední školy. Předpokládejme, že výběr, který máme k dispozici, pochází z trojrozměrného normálního rozdělení. Z dat vypočítáme výběrové korelační koeficienty. Vyjde $r_{01} = -0{,}614$, $r_{02} = 0{,}662$ a $r_{12} = -0{,}381$. Abychom mohli korelace mezi náhodnými veličinami porovnat, změníme znaménka korelačních koeficientů r_{01} a r_{12}, která vyšla

záporná, na kladná. Tuto změnu znaménka můžeme provést bez újmy na obecnosti. Nulovou hypotézu H_0 budeme testovat proti oboustranné alternativě H_1: $\rho_{01} \neq \rho_{02}$. K tomu použijeme testy odvozené v kapitole 2, tedy test Hotellingův, Williamsův a test poměrem věrohodností. Testové statistiky těchto testů spolu s příslušnými kvantily a p-hodnotami nalezneme v tabulce 5.1. Z té vyplývá, že $|T_i| < t_{99}(0{,}975)$, $i = 1, 2$, a $T_3 < \chi^2_1(0{,}95)$. Ani jeden z testů nulovou hypotézu na hladině 5 % nezamítnul. Nepodařilo se nám tedy prokázat, že by jeden z uvažovaných prediktorů měl na vysvětlovanou proměnnou větší vliv než druhý.

Tabulka 5.1: Výsledky Hotellingova a Williamsova testu a testu poměrem věrohodností - test hypotézy H_0: $\rho_{01} = \rho_{02}$.

test	test. statistika	kvantil	p-hodnota
Hotellingův test	$T_1 = -0{,}669$	$t_{99}(0{,}975) = 1{,}984$	0,505
Williamsův test	$T_2 = -0{,}628$	$t_{99}(0{,}975) = 1{,}984$	0,532
Test poměrem věrohodností	$T_3 = 0{,}404$	$\chi^2_1(0{,}95) = 3{,}841$	0,525

5.2 Porovnání důležitosti dvou prediktorů v modelu lineární regrese

Uvažujme následující regresní model vyjadřující závislost prospěchu za bakalářské studium na prospěchu na střední škole a na výsledcích přijímacích zkoušek:

$$M_1: \quad prumer_i = \beta_0 + \beta_1 zcel_i + \beta_2 ssprum_i + \epsilon_i, \quad i = 1, ..., 102, \tag{5.1}$$

Tabulka 5.2: Výstup programu **R** pro odhad modelu M_1.

```
lm(formula = prumer ~ zcel + ssprumer)
Residuals:

      Min         1Q     Median         3Q        Max
-0.811749  -0.214233  -0.007698   0.218281   0.860930

Coefficients:
```

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	2.599772	0.247759	10.493	< 2e-16 ***
zcel	-0.008440	0.001386	-6.088	2.18e-08 ***
ssprumer	0.527605	0.073288	7.199	1.19e-10 ***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3169 on 99 degrees of freedom

Multiple R-squared: 0.5909,      Adjusted R-squared: 0.5826
F-statistic: 71.49 on 2 and 99 DF,  p-value: < 2.2e-16
```

Parametry β_0, β_1 a β_2 tohoto modelu odhadneme metodou nejmenších čtverců. Odhady b_0, b_1 a b_2 těchto parametrů nalezneme v tabulce 5.2 spolu s p-hodnotami

t-testů testujících nulovost jednotlivých regresních koeficientů. Z této tabulky je zřejmé, že všechny parametry jsou v modelu významné. Dosazením odhadů regresních koeficientů do (5.1) dostaneme

$$\widehat{prumer}_i = 2{,}600 - 0{,}008 \,.\, zcel_i + 0{,}528 \,.\, ssprumer_i, \quad i = 1, ..., 102.$$

Koeficient determinace modelu M_1 je roven 0,591, což znamená, že prospěch na střední škole a výsledky přijímacích zkoušek vysvětlují 59,1 % variability prospěchu na vysoké škole.

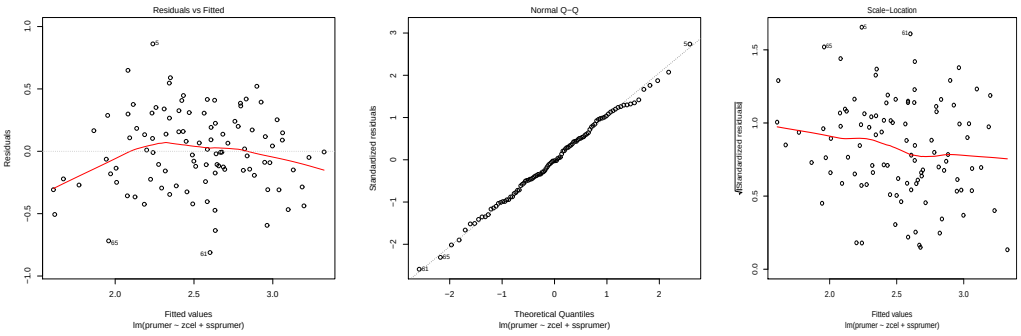
Než začneme provádět další analýzy, je třeba ověřit, zda náš regresní model splňuje následující předpoklady:

- normalita vektoru chyb $\boldsymbol{\epsilon}$,
- shoda rozptylů jednotlivých složek reziduální složky (homoskedasticita),
- nezávislost reziduí (nezávislost jednotlivých pozorování).

Normalitu reziduí budeme testovat pomocí Shapirova-Wilkova testu (viz [10, str. 129]). Častým případem porušení předpokladu homoskedasticity je monotonní závislost rozptylu na střední hodnotě závisle proměnné. Hypotézu, že rozptyl je konstantní proti alternativní hypotéze, že závisí na střední hodnotě, otestujeme pomocí Breuschova-Paganova testu (viz [10, str. 125]). Normalitu i homoskedasticitu reziduí budeme testovat na hladině 5 %. Nezávislost jednotlivých pozorování budeme předpokládat, neboť nemáme důvod si myslet, že při sběru dat došlo k jejímu porušení.

Poznámka 9. V **R** je Shapiro-Wilkův test normality možno nalézt pod názvem `shapiro.test()` ve standardní knihovně `stat`, Breuschův-Paganův test pak v knihovně `lmtest` jako `bptest()`. Jedním z parametrů funkce `bptest()` je parametr `studentize`. Pokud je jeho hodnota `TRUE`, potom se provede modifikace Breuschova-Paganova testu, kterou navrhl Koenker (viz [10, str. 125]). Tato modifikace je v **R** nastavena defaultně a pracovat s ní budeme i my. ◊

Zjistěme tedy, zda model M_1 splňuje všechny tyto předpoklady. Nejprve testujme normalitu reziduí. Protože je p-hodnota Shapirova-Wilkova testu rovna 0,998, normalitu reziduí na hladině 5 % nezamítáme. Výsledek tohoto testu potvrzuje i druhý z grafů na obrázku 5.1. Jedná se o tzv. normální diagram (viz [10, str. 113]). Stejně tak na hladině 5 % nezamítáme hypotézu, že rozptyl je konstantní (p-hodnota Breuschova-Paganova testu je 0,189). To, že s dodržením předpokladu homoskedasticity bychom neměli mít problém, naznačuje i první a třetí graf na obrázku 5.1 (rezidua jsou rovnoměrně rozprostřena po ploše grafu). Můžeme tedy předpokládat, že tento model splňuje všechny předpoklady normálního modelu lineární regrese.



Obrázek 5.1: Grafická analýza reziduí modelu M_1.

Jak již jsme uvedli v teoretické části, v regresním modelu porovnáváme vliv regresorů na závisle proměnnou pomocí regresních koeficientů. Požadujeme však, aby regresory, jejichž přínos porovnáváme, měly nulový průměr a jednotkový rozptyl. Převeďme tedy model M_1 do požadovaného tvaru, tedy:

$$\begin{aligned} M_1: \quad prumer_i &= (\beta_0 + \beta_1\overline{zcel} + \beta_2\overline{ssprum}) + \\ &+ \beta_1 s_1\left(\frac{zcel_i - \overline{zcel}}{s_1}\right) + \beta_2 s_2\left(\frac{ssprum_i - \overline{ssprum}}{s_2}\right) + \epsilon_i, \end{aligned}$$

kde $i = 1, ..., 102$ a

$$s_1 = \frac{1}{101}\sum_{i=1}^{102}\left(zcel_i - \overline{zcel}\right)^2 \quad \text{a} \quad s_2 = \frac{1}{101}\sum_{i=1}^{102}\left(ssprum_i - \overline{ssprum}\right)^2.$$

Označíme-li

$$\beta_0^* = \beta_0 + \beta_1\overline{zcel} + \beta_2\overline{ssprum}, \qquad \beta_1^* = \beta_1 s_1, \qquad \beta_2^* = \beta_2 s_2$$

a dále

$$zcel_i^* = \left(\frac{zcel_i - \overline{zcel}}{s_1}\right), \qquad ssprumer_i^* = \left(\frac{ssprumer_i - \overline{ssprumer}}{s_2}\right),$$

můžeme přejít k modelu

$$M_1^*: \quad prumer_i = \beta_0^* + \beta_1^* zcel_i^* + \beta_2^* ssprum_i^* + \epsilon_i, \quad i = 1, ..., 102.$$

Odhadneme-li koeficienty β_0^*, β_1^* a β_2^*, dostaneme po dosazení rovnici regresní funkce

$$\widehat{prumer}_i = 2{,}519 - 0{,}208\ .\ zcel_i^* + 0{,}246\ .\ ssprum_i^*, \quad i = 1, ...102.$$

Nyní vidíme, že absolutní hodnota odhadu b_2^* regresního koeficientu β_2^* je větší než absolutní hodnota odhadu b_1^* regresního koeficientu β_1^*. Prospěch ze střední školy by tedy měl mít větší vliv na prospěch na vysoké škole než přijímací zkoušky. V předchozí části jsme také zjistili, že $|r_{01}| < |r_{02}|$. Nicméně testy testující hypotézu o rovnosti příslušných korelačních koeficientů tuto hypotézu nezamítly. Větší

vliv prospěchu ze střední školy na prospěch na vysoké škole se nám tedy prokázat nepodařilo.

Spočtěme ještě na závěr této části poměr

$$\hat{\omega} = \frac{b_1^* \mathbf{zcel}' \mathbf{zcel} b_1^*}{b_2^* \mathbf{ssprumer}' \mathbf{ssprumer} b_2^*},$$

kde $\mathbf{zcel} = (zcel_1, ..., zcel_{102})'$ a $\mathbf{ssprumer} = (ssprumer_1, ..., ssprumer_{102})'$, a otestujme hypotézu

$$H_0 : \omega = \frac{\beta_1^* \mathbf{zcel}' \mathbf{zcel} \beta_1^*}{\beta_2^* \mathbf{ssprumer}' \mathbf{ssprumer} \beta_2^*} = 1 \quad \text{proti alternativě} \quad H_1 : \omega \neq 1.$$

Test této hypotézy je sice navržený za účelem porovnání relativní důležitosti dvou skupin prediktorů, je možné ho však samozřejmě použít i k porovnání vlivu dvou prediktorů.

Dostáváme $\hat{\omega} = 0{,}715$, $\hat{\tau} = \log \hat{\omega} = -0{,}335$ a $\sqrt{\widehat{\text{var}}\ \hat{\tau}} = 0{,}505$. Potom testová statistika testující hypotézu $H_0 : \omega = 1$ je rovna

$$Z = \frac{\hat{\tau}}{\sqrt{\widehat{\text{var}}\ \hat{\tau}}} = -\frac{0{,}335}{0{,}505} = -0{,}665$$

a vzhledem k tomu, že $|Z| < 1{,}96$, hypotézu H_0 na hladině 5 % nezamítáme. To je ve shodě s výsledky předchozích testů. Všimněme si také, že absolutní hodnoty statistik T_1, T_2, $\sqrt{T_3}$ (viz tabulka 5.1) a Z se příliš neliší.

Nebylo by při vysvětlování prospěchu na vysoké škole správnější přihlédnout také k pohlaví studentů?

Data, která máme k dispozici, dále obsahují informaci o pohlaví studentů. Přidejme tedy do uvažovaného modelu regresor *Pohlavi* jako faktor a otestujme, zda se liší závislost pro dívky a chlapce a zda by nebylo správnější použít při vysvětlování prospěchu na vysoké škole pro chlapce jiný regresní model než pro dívky. Uvažujme následující model s interakcemi:

$$\begin{aligned} M_2^{int}: \quad prumer_i &= \beta_0 + \beta_1 zcel_i + \beta_2 ssprum_i + \beta_3 Pohlavi_i, \\ &+ \lambda_{13} zcel_i \cdot Pohlavi_i + \lambda_{23} zcel_i \cdot Pohlavi_i + \epsilon_i, \end{aligned}$$

kde $i = 1, ..., 102$. Odhady regresních koeficientů nalezneme v tabulce 5.3. Pomocí příslušných testů jsme ověřili, že můžeme předpokládat splnění podmínek normálního modelu lineární regrese. Abychom zjistili, zda se závislost pro chlapce a dívky liší, ověřme významnost interakcí v modelu. Naším úkolem je testovat hypotézu

$$H_0 : \lambda_{13} = 0 \qquad \text{a} \qquad \lambda_{23} = 0$$

proti alternativě

$$H_1 : \lambda_{13} \neq 0 \qquad \text{nebo} \qquad \lambda_{23} \neq 0.$$

Zajímá nás tedy, zda je od modelu M_2^{int} možné přejít k podmodelu

$$M_2: \quad prumer_i = \beta_0 + \beta_1 zcel_i + \beta_2 ssprum_i + \beta_3 Pohlavi_i + \epsilon_i, \quad i = 1, ..., 102.$$

Tabulka 5.3: Výstup programu **R** pro odhad modelu M_2^{int}.

```
lm(formula = prumer ~ zcel * Pohlavi + ssprumer * Pohlavi)
Residuals:

      Min        1Q    Median        3Q       Max
 -0.73408  -0.19418   0.00011   0.20182   0.92036

Coefficients:
```

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
(Intercept)	2.5628247	0.3639934	7.041	2.86e-10	***
zcel	-0.0080680	0.0020241	-3.986	0.000131	***
Pohlavi2	-0.3360868	0.4874979	-0.689	0.492227	
ssprumer	0.4787332	0.0962497	4.974	2.87e-06	***
zcel:Pohlavi2	-0.0002973	0.0026626	-0.112	0.911324	
Pohlavi2:ssprumer	0.3242590	0.1512823	2.143	0.034608	*

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3 on 96 degrees of freedom

Multiple R-squared: 0.6444,     Adjusted R-squared: 0.6258
F-statistic: 34.79 on 5 and 96 DF,  p-value: < 2.2e-16
```

O podmodelu rozhodneme pomocí F-testu (viz [10, str. 30]). Příslušná testová statistika má F-rozdělení s 2 a 96 stupni volnosti. Protože je testová statistika tohoto testu rovna 2,774 a příslušná p-hodnota se rovná 0,067, na hladině 5 % nezamítáme hypotézu H_0 o nevýznamnosti interakcí ve prospěch uvedené alternativy. To tedy znamená, že interakce v tomto modelu jsou nevýznamné a od modelu M_2^{int} lze přejít k modelu M_2. V modelu M_2 jsou všechny regresory významné. Koeficient determinace tohoto modelu je roven 0,624. Dosadíme-li odhadnuté parametry, dostaneme následující dvě rovnice:

$$\begin{aligned} \widehat{prumer}_i &= 2{,}349 - 0{,}008 \,.\, zcel_i + 0{,}599 \,.\, ssprum_i,\ i = 1, ..., 102, \quad \text{(Dívky)} \\ \widehat{prumer}_i &= 2{,}349 + 0{,}188 - 0{,}008 \,.\, zcel_i + 0{,}599 \,.\, ssprum_i. \qquad \text{(Chlapci)} \end{aligned}$$

Zjistili jsme tedy, že vliv regresorů na vysvětlovanou proměnnou je pouze aditivní. Regresní rovnice pro chlapce a dívky se liší jen hodnotou konstantního členu. Proto při testování relativní důležitosti prediktorů *zcel* a *ssprumer* není nutné přihlížet k pohlaví studentů.

5.3 Porovnání důležitosti dvou skupin prediktorů v modelu lineární regrese

V této části se opět pokusíme vysvětlit prospěch studentů na vysoké škole tentokrát však pomocí jiných vysvětlujících proměnných. Ty můžeme rozdělit do dvou skupin. První skupina uvažovaných regresorů obsahuje prediktory týkající se prospěchu studentů na střední škole, jsou to regresory $ss1$, $ss2$, $ss3$ a $ss4$.

Druhá skupina obsahuje prediktory *matem* a *zem*, tedy proměnné s výsledky přijímacích zkoušek. Uvažujme následující model:

$$M_3: \; prumer_i = \alpha + \beta_1 ss1_i + \beta_2 ss2_i + \beta_3 ss3_i + \beta_4 ss4_i + \gamma_1 matem_i + \gamma_2 zem_i + \epsilon_i, \quad i = 1, ..., 102. \tag{5.2}$$

Tabulka 5.4: Výstup programu **R** pro odhad modelu M_3.

```
lm(formula = prumer ~ ss1 + ss2 + ss3 + ss4 + matem + zem)
Residuals:

      Min        1Q    Median        3Q       Max
-0.83706  -0.22945   0.00247   0.20896   0.88350

Coefficients:
```

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	2.740436	0.254292	10.777	< 2e-16 ***
ss1	-0.099071	0.141253	-0.701	0.48478
ss2	0.183820	0.181560	1.012	0.31390
ss3	0.192616	0.174488	1.104	0.27243
ss4	0.260392	0.135959	1.915	0.05847 .
matem	-0.005737	0.001701	-3.372	0.00108 **
zem	-0.014552	0.002601	-5.594	2.13e-07 ***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.' 0.1 ' ' 1

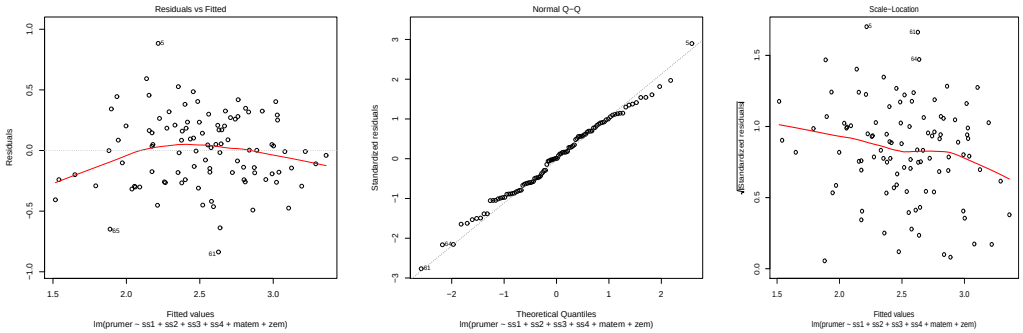
```
Residual standard error: 0.308 on 95 degrees of freedom
```

Multiple R-squared: 0.6292,	Adjusted R-squared: 0.6058
F-statistic: 26.87 on 6 and 95 DF,	p-value: < 2.2e-16

Regresní koeficienty tohoto modelu odhadneme opět metodou nejmenších čtverců. Tyto odhady nalezneme v tabulce 5.4. Po jejich dosazení do (5.2) dostaneme

$$\widehat{prumer}_i = 2{,}740 - 0{,}099 \; . \; ss1_i + 0{,}183 \; . \; ss2_i + 0{,}193 \; . \; ss3_i + 0{,}260 \; . \; ss4_i - 0{,}006 \; . \; matem_i - 0{,}014 \; . \; zem_i, \quad i = 1, ..., 102.$$

Můžeme předpokládat, že model M_3 splňuje všechny podmínky normálního modelu lineární regrese (p-hodnota Shapirova-Wilkova testu je rovna 0,839 a p-hodnota Breuschova-Paganova testu je 0,144). Splnění podmínek nevyvracejí ani grafy na obrázku 5.2.



Obrázek 5.2: Grafická analýza reziduí modelu M_3.

Koeficient determinace R^2 tohoto modelu je roven 0,629. Tabulka 5.4 dále obsahuje testové statistiky a p-hodnoty testů testujících významnost jednotlivých regresorů v modelu. Z těchto testů vyplývá, že ne všechny regresory jsou v modelu významné. Některé regresory by tedy bylo možno z modelu vyloučit a model zjednodušit. To však není naším záměrem. Naším úkolem je porovnat vliv uvažovaných skupin prediktorů na vysvětlení proměnné *prumer*. Označme

$$\mathbf{X} = \begin{pmatrix} \mathbf{ss1} - \mathbf{1}\overline{ss1} \\ \mathbf{ss2} - \mathbf{1}\overline{ss2} \\ \mathbf{ss3} - \mathbf{1}\overline{ss3} \\ \mathbf{ss4} - \mathbf{1}\overline{ss4} \end{pmatrix}' \quad \text{a} \quad \mathbf{Z} = \begin{pmatrix} \mathbf{matem} - \mathbf{1}\overline{matem} \\ \mathbf{zem} - \mathbf{1}\overline{zem} \end{pmatrix}'$$

a dále

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' \quad \text{a} \quad \boldsymbol{\gamma} = (\gamma_1, \gamma_2)'.$$

Model centrujeme za účelem zjednodušení dalšího zápisu (viz kapitola 3). Testujme hypotézu

$$H_0: \quad \omega = \frac{\boldsymbol{\beta' X' X \beta}}{\boldsymbol{\gamma' Z' Z \gamma}} = 1 \quad \text{proti} \quad H_1: \quad \omega \neq 1.$$

Nechť $\hat{\boldsymbol{\beta}}$ a $\hat{\boldsymbol{\gamma}}$ jsou odhady parametrů $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ metodou nejmenších čtverců. Potom

$$\hat{\omega} = \frac{\hat{\boldsymbol{\beta}}' \boldsymbol{X}' \boldsymbol{X} \hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\gamma}}' \boldsymbol{Z}' \boldsymbol{Z} \hat{\boldsymbol{\gamma}}} = 1{,}443, \qquad \hat{\tau} = \log \hat{\omega} = 0{,}366$$

a testová statistika pro test hypotézy H_0 je

$$Z = \frac{\hat{\tau}}{\sqrt{\widehat{\text{var}}\ \hat{\tau}}} = 0{,}814.$$

Vzhledem k tomu, že $|Z| < u(1 - \alpha/2) = 1{,}96$, hypotézu H_0 na hladině 5 % nezamítáme. Ani o jedné skupině prediktorů tedy nemůže říct, že by měla na vysvětlení závisle proměnné větší vliv.

Soubor dat, který máme k dispozici, obsahuje kromě celkového průměru za první tři roky bakalářského studia také jednotlivé průměry za první až třetí rok. Uvažujme následující tři modely pro závislost těchto průměrů ($pr1$, $pr2$, resp. $pr3$) na prospěchu ze střední školy a na výsledcích přijímacích zkoušek:

$$\begin{aligned}
M_{3PR1}: \ pr1_i &= \alpha_1 + \beta_{11} ss1_i + \beta_{12} ss2_i + \beta_{13} ss3_i + \beta_{14} ss4_i + \gamma_{11} matem_i + \gamma_{12} zem_i + \epsilon_{1i}, \\
M_{3PR2}: \ pr2_i &= \alpha_2 + \beta_{21} ss1_i + \beta_{22} ss2_i + \beta_{23} ss3_i + \beta_{24} ss4_i + \gamma_{21} matem_i + \gamma_{22} zem_i + \epsilon_{2i}, \\
M_{3PR3}: \ pr3_i &= \alpha_3 + \beta_{31} ss1_i + \beta_{32} ss2_i + \beta_{33} ss3_i + \beta_{34} ss4_i + \gamma_{31} matem_i + \gamma_{32} zem_i + \epsilon_{3i},
\end{aligned}$$

kde $i = 1, ..., 102$. Naším úkolem bude opět otestovat, zda jedna z uvažovaných skupin prediktorů vysvětluje závisle proměnnou lépe. Bodové odhady regresních koeficientů a koeficient determinace modelů nalezneme v tabulce 5.5, výsledky testů relativní důležitosti dvou skupin prediktorů pak v tabulce 5.6. Testovali jsme opět hypotézu H_0: $\omega = 1$ proti oboustranné alternativě H_1: $\omega \neq 1$ a jak vyplývá z tabulky 5.6, tuto hypotézu jsme na hladině 5 % nezamítli ani pro jeden z uvažovaných modelů. Tedy ani tentokrát se nám nepodařilo prokázat větší

důležitost jedné ze dvou skupin prediktorů.

Tabulka 5.5: Odhady regresních koeficientů a koeficienty determinace modelů M_{3PR1}, M_{3PR2} a M_{3PR3}.

Model	Intercept	ss1	ss2	ss3	ss4	matem	zem	R^2
M_{3PR1}	3,103	-0,206	0,178	0,236	0,352	-0,009	-0,015	0,596
M_{3PR2}	3,249	-0,223	0,423	0,059	0,263	-0,006	-0,020	0,572
M_{3PR3}	2,050	0,028	0,085	0,226	0,190	-0,003	-0,012	0,470

Tabulka 5.6: Výsledky testů relativní důležitosti dvou skupin prediktorů v modelech M_{3PR1}, M_{3PR2} a M_{3PR3}.

Model	$\hat{\omega}$	$\hat{\tau}$	Z	p-hodnota
M_{3PR1}	1,212	0,192	0,388	0,698
M_{3PR2}	0,883	-0,125	-0,254	0,800
M_{3PR3}	2,468	0,903	1,425	0,154

Zkusme znovu na závěr přidat do uvažovaných modelů M_3, M_{3PR1}, M_{3PR2} a M_{3PR3} regresor $Pohlavi$ a otestujme ještě, zda by nebylo správnější použít při vysvětlování prospěchu na vysoké škole pro chlapce jiný regresní model než pro dívky. Uvažujme model s interakcemi:

$$\begin{aligned} M_4^{int}: \quad prumer_i &= \alpha + \beta_1 ss1_i + \beta_2 ss2_i + \beta_3 ss3_i + \beta_4 ss4_i + \gamma_1 matem_i + \gamma_2 zem_i \\ &+ \delta_1 Pohlavi_i + \lambda_1 ss1_i \cdot Pohlavi_i + \lambda_2 ss2_i \cdot Pohlavi_i + \lambda_3 ss3_i \cdot Pohlavi_i \\ &+ \lambda_4 ss4_i \cdot Pohlavi_i + \lambda_5 matem_i \cdot Pohlavi_i + \lambda_6 zem_i \cdot Pohlavi_i + \epsilon_i, \end{aligned}$$

kde $i = 1, ..., 102$. Analogicky by vypadaly modely M_{4PR1}^{int}, M_{4PR2}^{int}, M_{4PR3}^{int} pro závisle proměnné $pr1$, $pr2$ a $pr3$. Ověřme významnost interakcí v uvažovaných modelech. Testujme hypotézu

$$H_0 : \lambda_k = 0, \qquad k = 1, ..., 6$$

proti alternativě

$$H_1 : \lambda_k \neq 0 \text{ pro alespoň jedno } \lambda_k, \ k = 1, ..., 6.$$

Zajímá nás tedy, zda je od modelů M_4^{int}, M_{4PR1}^{int}, M_{4PR2}^{int}, resp. M_{4PR3}^{int} možné přejít k podmodelům bez interakcí. O podmodelech rozhodneme pomocí F-testu. Příslušné testové statistiky mají F-rozdělení s 6 a 88 stupni volnosti. Nalezneme je v tabulce 5.7 spolu s p-hodnotami. Z této tabulky vyplývá, že na hladině 5 % nezamítáme hypotézu H_0 o nevýznamnosti interakcí ve prospěch uvedené alternativy pro všechny uvažované modely. Od modelů s interakcemi tedy opět můžeme přejít k modelům bez interakcí. Regresní rovnice pro chlapce se od rovnice pro dívky liší jen hodnotou konstantního členu. Proto ani v tomto případě při testování relativní důležitosti dvou skupin prediktorů v uvažovaných modelech není nutné přihlížet k pohlaví studentů.

Tabulka 7: Výsledky F-testů testujících významnost interakcí v modelech M_4^{int}, M_{4PR1}^{int}, M_{4PR2}^{int}, resp. M_{4PR3}^{int}.

Model	test.stat.	p-hodnota
M_4^{int}	1,339	0,249
M_{4PR1}^{int}	1,521	0,181
M_{4PR2}^{int}	0,668	0,676
M_{4PR3}^{int}	0,836	0,546

Závěr

Cílem této diplomové práce bylo popsat, případně odvodit, některé postupy používané při zkoumání vlivu prediktorů na vysvětlení závisle proměnné. Nejprve jsme se soustředili na porovnání vlivu dvou prediktorů. Ten jsme porovnávali na základě příslušných korelačních koeficientů. Testů rovnosti dvou korelačních koeficientů existuje celá řada, my jsme se zaměřili na tři z nich, a to na test Hotellingův, test Williamsův a test poměrem věrohodností. Tyto testy jsme popsali a odvodili. Podařilo se nám také vysvětlit odvození Williamsova testu, které není dohledatelné v nám dostupné literatuře.

Uvažované testy rovnosti korelačních koeficientů jsme v praktické části práce podrobili simulační studii, při níž jsme zjišťovali, jak dodržují hladinu významnosti a jakou mají sílu. Z provedených simulací vyplynulo, že téměř pro jakoukoli volbu parametrů je použitelný Williamsův test. Tento test se sice ukázal nepatrně slabší než test poměrem věrohodností, dodržuje však hladinu významnosti i při nižších počtech pozorování. Můžeme se na něj tedy spolehnout i v případech, kdy již test poměrem věrohodností selhává a hladinu významnosti nadhodnocuje. Williamsův test podhodnocuje hladinu významnosti pouze, pokud je počet pozorování již velmi malý. V případě, že máme dostatečně velký počet pozorování, bychom nicméně doporučili dát přednost testu poměrem věrohodností, neboť, jak již jsme uvedli, je silnější než test Williamsům. Dále simulace ukázaly jasnou nespolehlivost Hotellingova testu. Tento test dodržuje hladinu významnosti jen pro některé volby parametrů. Proto bychom jeho používání nedoporučovali.

Dále bylo naším úkolem přejít od porovnávání dvou prediktorů k porovnávání dvou skupin prediktorů. Omezili jsme se na případ lineárního regresního modelu a odvodili příslušný test.

Poslední kapitolu práce jsme věnovali numerické demonstraci teorie. K dispozici jsme měli údaje o 102 studentech bakalářského studijního oboru Geografie–kartografie Přírodovědecké fakulty Univerzity Karlovy v Praze, kteří se zapsali ke studiu v akademickém roce 2003/04. Naším cílem bylo na základě těchto dat rozhodnout, zda na vysvětlení výsledků na vysoké škole mají větší vliv výsledky studentů na střední škole nebo výsledky přijímacích zkoušek na vysokou školu. K tomu jsme využili veškeré testy odvozené v teoretické části práce. Porovnávali jsme jak vliv dvou prediktorů na závisle proměnnou, tak vliv dvou skupin prediktorů. Žádný z provedených testů nicméně rozdíl ve vlivech neprokázal.

Seznam použité literatury

[1] Anděl, J. (2005). *Základy matematické statistiky*, Matfyzpress, Praha.

[2] Bican, L. (2002). *Lineární algebra a geometrie*, Academia, Praha.

[3] Cipra, T. (1984). *Ekonometrie*, SPN Praha.

[4] Dupač, V., Hušková, M. (2005). *Pravděpodobnost a matematická statistika*, Nakladatelství Karolinum, Praha.

[5] Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters, *The Annals of Mathematical Statistics*, 11, 271–283.

[6] Neill, J. J., Dunn, O. J. (1975). Equality of dependent correlation coefficients, *Biometrics*, 31, 531–543.

[7] Silber, J. H., Rosenbaum, P. R., Ross, R. N. (1995). Comparing the contributions of groups of predictors: Which outcomes vary with hospital rather than patient characteristics?, *Journal of the American Statistical Association*, 90, 7–18.

[8] Rubešová, J. (2009). *Statistické metody pro hodnocení predikční validity*, Disertační práce PřF UK, Praha.

[9] Williams, E. J. (1959). The comparison of regression variables, *Journal of the Royal Statistical Society, Series B*, 21, 396–399.

[10] Zvára, K. (2008). *Regrese*, Matfyzpress, Praha.

[11] `http://mathworld.wolfram.com/ChiDistribution.html` (19.6.2011).

Seznam tabulek

Seznam obrázků

Příloha A

Skripty použité při výpočtech a simulacích v programu R

Zde uvádíme pouze některé z používaných skriptů. Zbylé je možno nalézt na přiloženém CD.

A.1 Simulační experiment - porovnání testů rovnosti dvou korelačních koeficientů

```
rm(list=ls()); library(maxLik); library(mvtnorm)

# vypocet gradientu a hessianu logaritmicke verohodnostni funkce
# trojrozmerne normalni rozdeleni

logLH0 = deriv3(~-n*log(s0*s1*s2)-
n/2*log(1+2*r*r*r12-2*r*r-r12*r12)-
(S11*(1-r12*r12)/s0/s0+S12*(r*r12-r)/s0/s1+S13*(r*r12-r)/s0/s2+
S12*(r*r12-r)/s0/s1+S22*(1-r*r)/s1/s1+S23*(r*r-r12)/s1/s2+
S13*(r*r12-r)/s0/s2+S23*(r*r-r12)/s1/s2+S33*(1-r*r)/s2/s2)/
(1 + 2*r*r*r12-2*r*r-r12*r12)/2,
namevec=c("s0","s1","s2","r","r12"),
function.arg=function(s0,s1,s2,r,r12,S11,S12,S13,S22,S23,S33,n))

# logaritmicka verohodnostni funkce - predpokladame platnost H0
# trojrozmerne normalni rozdeleni
# atributy funkce jsou jeji gradient a hessian

logLikH0 = function(theta0,S,n)
{
 s0=theta0[1];s1=theta0[2];s2=theta0[3]
 S11=S[1,1];S12=S[1,2];S13=S[1,3];S22=S[2,2];S23=S[2,3];S33=S[3,3]
 r = theta0[4]; r12=theta0[5]
 lL = logLH0(s0,s1,s2,r,r12,S11,S12,S13,S22,S23,S33,n)
 val = c(lL)
 attr(val,"gradient") = unclass(attr(lL,"gradient"))
 attr(val,"hessian") = attr(lL,"hessian")[1,,]
 val
}
```

```
# logaritmicka verohodnostni funkce
# trojrozmerne normalni rozdeleni

logLik = function(theta0,S,n)
{
 s0=theta0[1];s1=theta0[2];s2=theta0[3]
 S11=S[1,1];S12=S[1,2];S13=S[1,3];S22=S[2,2];S23=S[2,3];S33=S[3,3]
 r01=theta0[4];r02=theta0[5];r12=theta0[6]
 -n*log(s0*s1*s2)-n/2*log(1+2*r01*r02*r12-r01*r01-r02*r02-r12*r12)-
 (S11*(1-r12*r12)/s0/s0+S12*(r02*r12-r01)/s0/s1+
 S13*(r01*r12-r02)/s0/s2+S12*(r02*r12-r01)/s0/s1+
 S22*(1-r02*r02)/s1/s1+ S23*(r01*r02-r12)/s1/s2+
 S13*(r01*r12-r02)/s0/s2+S23*(r01*r02-r12)/s1/s2+
 S33*(1-r01*r01)/s2/s2)/(1+2*r01*r02*r12-r01*r01-r02*r02-r12*r12)/2
}

# test pomerem verohodnosti

testML = function(xMce)
{
 sd0 = apply(xData,2,sd)*sqrt(1-1/nrow(xMce))
 rho0 = cor(xData)
 S = var(xData)*(nrow(xData)-1)
 theta0 = c(s0=sd0[1],s1=sd0[2],s2=sd0[3],
          r01=rho0[1,2],r02=rho0[1,3],r12=rho0[2,3])
 theta0H0 = c(s0=sd0[1],s1=sd0[2],s2=sd0[3],
             r=(rho0[1,2]+rho0[1,3])/2,r12=rho0[2,3])
 S11=S[1,1]; S12=S[1,2];S13=S[1,3];S22=S[2,2];S23=S[2,3];S33=S[3,3]
 mL = sum(logLik(theta0=theta0,S=S,n=nrow(xData)))
 mL0 = maxLik(logLikH0,start=theta0H0,S=S,n=nrow(xData))
 ML=2*(mL-mL0$maximum)                    # testova statistika
 p=pchisq(ML,1,lower.tail=FALSE)          # p-hodnota
 return(list("chisq"=ML,"p.value"=p,"theta"=mL,"theta0"=mL0$estimate))
}

# Hotellinguv a Williamsuv test

testHW = function(xMce)
{
 n = nrow(xMce)
 r = cor(xMce)
 # testova statistika a p-hodnota Hotellingova testu
 t1 = (r[1,2]-r[1,3])/sqrt(2*det(r)/(n-3)/(1+r[2,3]))
 p1 = 2* pt(-abs(t1),n-3)
 # testova statistika a p-hodnota Williamsova testu
 rBar = (r[1,2]+r[1,3])/2
 t2 = (r[1,2]-r[1,3])/sqrt((2*det(r)/(n-3)/(1+r[2,3]))+
      (rBar^2*(1-r[2,3])^3/(n-1)/(1+r[2,3])))
```

```
 p2 = 2* pt(-abs(t2),n-3)
 list(T=c(t1,t2),p.value=c(p1,p2))
}

 # generovani dat: trojrozmerne normalni rozdeleni

xSimul = function(rho=c(0.4,0.6,0.5),mu=c(0,0,0),
          sigma=c(1,2,3),n=100,seed=1)
{
 rhoMce = matrix(c(1,rho[1],rho[2],rho[1],1,rho[3],
          rho[2],rho[3],1),3,3)
 sigmaMce = diag(sigma)%*%rhoMce%*%diag(sigma)
 set.seed(seed)
 rmvnorm(n,mean=mu,sigma=sigmaMce)
}

# test pozitivni definitnosti matice P
# testujeme, zda je determinant matice P kladny


pdR = function(rho)
{
 (1+2*rho[1]*rho[2]*rho[3]>rho[1]^2+rho[2]^2+rho[3]^2)
}

# simulace - zjistujeme hladinu a silu testu
# vysledky ukladame do souboru

DATUM = function()
{
 d = Sys.Date()
 paste(substring(d,3,4),substring(d,6,7),substring(d,9,10),sep="")
}

# vypis do souboru
OUT = paste("C:/simulace",DATUM(),".txt",sep="")

# nastaveni parametru
# zde priklad pro zavislost na korelacnim koeficientu rho12

n = 100                    # pocet pozorovani
B = 1000                   # pocet opakovani
rho12 = c(0.1,0.3)            # korelacni koeficienty rho01 a rho02
rho3seq = seq(-0.9,0.9,by=0.1)         # korelacni koeficient rho12
capture.output(cat("(",as.character(Sys.time()),")"),
file=OUT,append=FALSE)
capture.output(cat("rho12:"),print(round(rho12,2)),
file=OUT,append=TRUE)
capture.output(cat("rho3:"),print(round(rho3seq,2)),
file=OUT,append=TRUE)
```

```
capture.output(cat("B:"),print(B), file=OUT,append=TRUE)
capture.output(cat("n:"),print(n), file=OUT,append=TRUE)
stat = NULL

for (rho3 in rho3seq)
{
 pp = matrix(NA,B,17)
 colnames(pp) = c("ML","Hotell","Will","pML","pHotell","pWill",
 "s0","s1","s2","r01","r02","r12","s00","s01","s02","r","r12")
 rho = c(rho12,rho3)
 names(rho) = c("rho01","rho02","rho12")
 if(!pdR(rho)) stop("'korelacni' matice neni pozitivne definitni")
 for (b in 1:B)
  {
   xData = xSimul(rho=rho,n=n,sigma=c(1,2,3),seed=b)
   tML = testML(xData)
   tHW = testHW(xData)
   pp[b,1] = tML$chisq
   pp[b,4] = tML$p.value
   pp[b,2:3] = tHW$T
   pp[b,5:6] = tHW$p.value
   pp[b,7:12] = tML$theta
   pp[b,13:17] = tML$theta0
  }
 radek = apply(pp[,4:6],2,function(x) mean(x<0.05))
 stat = rbind(stat,radek)
 capture.output(print(c(round(rho,2),round(radek,3))),
 file=OUT,append=TRUE)
}
```

A.2 Porovnání důležitosti dvou skupin prediktorů v modelu lineární regrese

```
# Funkce, kterou pouzivame pri porovnavani dulezitosti
# dvou skupin prediktoru v modelu linearni regrese

 testreldul<-function(object,set1,set2,alternative="two.sided")
# argumenty funkce:
 # object: linearni regresni model
 # set1, set2: vektory indexu prediktoru obsazenych v prvni,
 # resp. druhe skupine
 # alternative: "two.sided" pro test oboustranne alternativy,  # "less", resp.
"greater", pro test jednostranne alternativy
{
 if(length(intersect(set1, set2))!=0)
   stop("Zadane mnoziny indexu nejsou disjunktni!")
 if( (length(set1)==0) | (length(set2)==0))
   stop("Nektera ze zadanych mnozin vektoru je prazdna.")
```

```
b<-coef(object)[set1]     # odhady regresnich koeficientu prediktoru
c<-coef(object)[set2]     # obsazenych v porovnavanych skupinach

modelmatrix<-model.matrix(object)       # matice modelu

# matice obsahujici skupiny prediktoru, jejichz relativni dulezitost
# porovnavame
X<-matrix(modelmatrix[,set1],ncol=length(b))
Z<-matrix(modelmatrix[,set2],ncol=length(c))

newX<-sweep(X,2,apply(X,2,mean))               # centrovani
newZ<-sweep(Z,2,apply(Z,2,mean))
omega<-(t(b) %*% t(newX) %*% newX %*% b)/
       (t(c) %*% t(newZ) %*% newZ %*% c)       # vzorec (3.3)
tau<-log(omega)

w<-2*c((t(newX) %*% newX %*% b/sum(t(b) %*% t(newX) %*% newX %*% b)),
  (- t(newZ) %*% newZ %*% c)/sum(t(c) %*% t(newZ) %*% newZ %*% c ))
indices<-c(set1,set2)
# rozptylova matice vektoru odhadu regresnich koeficientu
Sigma<-vcov(object)[indices,indices]

var.tau<-t(w)%*%Sigma%*%w       # rozptyl odhadu tau
sd.tau<-sqrt(var.tau)           # smerodatna odchylka odhadu tau

Z<-tau/sd.tau       # testova statistika - vzorec (3.4)
alpha<-0.05

if (alternative == "less")
 {
  p.hodnota <- pnorm(Z)       # p-hodnota, kvantil a interval spolehlivosti
  kvantil<-qnorm(1-alpha)     # pro alternativy "less", "greater"
  int<-c(-Inf,tau+sd.tau*kvantil)                 # a "two.sided"
 }
if (alternative == "greater")
 {
  p.hodnota <- pnorm(Z,lower.tail = FALSE)
  kvantil<-qnorm(1-alpha)
  int<-c(tau-sd.tau*kvantil,Inf)
 }
if (alternative == "two.sided")
 {
  kvantil<-qnorm(1-alpha/2)
  p.hodnota <- 2 * pnorm(-abs(Z))
  int<-c(tau-sd.tau*kvantil,tau+sd.tau*kvantil)
 }

# vystup obsahuje odhad omega, odhad tau, smerodatnou odchylku tau
# testovou statistiku Z, kvantil normovaneho normalniho rozdeleni
# p-hodnotu testu a 95% intervalovy odhad tau
```

```
 V<-matrix(c("omega:",round(omega,3),"tau:" ,round(tau,3), "smerodatna
    odchylka tau:", round(sd.tau,3), "testova statistika:", round(Z,3),
    "kvantil normovaneho normalniho rozdelení:", round(kvantil,3),
    "p-hodnota:", round(p.hodnota,3), "95% intervalovy odhad tau (D): ",
    round(int[1],3), "95% intervalovy odhad tau (H): ", round(int[2],3)),
    ncol=2,byrow=T)
 return(V)
}

# ukazka pouziti

# porovnani dvou prediktoru
summary(m1<-lm(prumer ~ zcel + ssprumer))
testreldul(m1,set1<-c(2),set2<-c(3),alternative="two.sided")

# porovnani dvou skupin prediktoru
summary(mPR<-lm(prumer ~ ss1 + ss2 + ss3 +ss4 + matem + zem))
testreldul(mPR,set1<-c(2,3,4,5),set2<-c(6,7),alternative="two.sided")
```

Příloha B

Obsah přiloženého CD

Součástí této diplomové práce je i CD, které kromě textu diplomové práce obsahuje i další soubory. Složka *Skripty* obsahuje soubory se skripty, které jsme použili při výpočtech a simulacích v programu **R**. Ve složce *Výsledky simulací* nalezneme soubory s výstupy programu pro porovnání testů rovnosti dvou korelačních koeficientů.