

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



TESTY V LINEÁRNEJ REGRESII PO SELEKCII
PREMENNÝCH

DIPLOMOVÁ PRÁCA

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

**TESTY V LINEÁRNEJ REGRESII PO SELEKCII
PREMENNÝCH**

DIPLOMOVÁ PRÁCA

Študijný program: Ekonomicko - finančná matematika a modelovanie
Študijný odbor: Matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: Mgr. Samuel Rosa, PhD.



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Jakub Poljovka
Študijný program: ekonomicko-finančná matematika a modelovanie
(Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Testy v lineárnej regresii po selekcii premenných
Tests in linear regression after model selection

Anotácia: Bežnými súčasťami práce s lineárnou regresiou sú výber vhodného modelu a testovanie hypotéz o parametroch. Problém je však, že tieto dva postupy spolu nefungujú ideálne: ak najprv vykonáme výber vhodného modelu (selekciu premenných), tak je zasiahnutá korektnosť testov o parametroch. V bakalárskej práci Mariána Badára (2021) bolo simulačne ukázané, že chyba prvého druhu testov signifikancie môže byť skutočne výrazne zmenená. V diplomovej práci sa pokúsime buď analyticky alebo pomocou veľmi dôkladných simulácií zistiť, za akých podmienok dochádza k ako veľkému poškodeniu chyby prvého druhu a ako je možné tento problém riešiť. Dôkladnejšie tiež naštudujeme a odprezentujeme literatúru venujúcu sa kvalite testov po vykonaní selekcie premenných.

Vedúci: Mgr. Samuel Rosa, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.
Dátum zadania: 11.01.2022

Dátum schválenia: 14.01.2022

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

PodĎakovanie Touto cestou by som sa chcel poĎakovať svojmu školiteľovi Mgr. Samuelovi Rosovi, PhD. za jeho cenné rady, trpezlivosť, pravidelnú pomoc pri doladovaní obsahu práce a hlavne množstvo času, ktoré mi venoval pri pravidelných konzultáciách. Naše spoločné stretnutia boli pre mňa veľmi prínosné po odbornej stránke, ale vďačím tiež za psychickú podporu a náhradné nápady vždy, keď som mal pocit, že sa nachádzam v slepej uličke a neviem ako ďalej postupovať.

Vďaka patrí tiež všetkým mojim vyučujúcim a prednášajúcim na fakulte, za odovzdanie svojich vedomostí. Stretol som sa tu vždy s férovým a často aj veľmi ústretovým prístupom k študentom, čo mi veľmi spríjemnilo a uľahčilo moje štúdium.

Abstrakt

V našej práci sa zaoberáme testovaním hypotéz o signifikantnosti parametrov nášho záujmu po selekcii ostatných parametrov v modeli pomocou Akaikeho informačného kritéria. Nadväzujeme tak na empirické výsledky získané pomocou Monte-Carlo simulácií. Naším cieľom je tieto výsledky čo najlepšie uchopiť a popísať pomocou analytických metód. Práca obsahuje jednak prehľad dostupnej literatúry na túto tému, ale aj vlastné výsledky. Odvádzame pravdepodobnosť, že pomocou Akaikeho kritéria vyselektujeme neúplný model oproti úplnému. Následne odlišujeme dva rôzne prístupy pri selekcii. V prvom z nich umožníme selekciu parametra ktorého signifikantnosť chceme testovať a následne ukážeme, že tento prístup je v závislosti od nastavenej hladiny testu buď totožný s postupom, v ktorom iba testujeme hypotézu bez prevedenia selekcie, alebo s postupom, v ktorom prevedieme iba selekciu bez následného testovania hypotézy. V druhom prístupe trváme na tom, aby bol testovaný parameter obsiahnutý v modeli. V tomto prípade odhadujeme chybu prvého druhu testu prevedeného po selekcii premenných pomocou simulácií. Ukazuje sa, že táto chyba závisí od skutočných hodnôt parametrov modelu, ktoré reálne nepoznáme, čo vyriešime bayesovským prístupom k týmto premenným. V záverečnej kapitole nakoniec poskytujeme analýzy, ktoré hovoria o súvisi medzi výberovou koreláciou vysvetľujúcich premenných a potrebnou korekciou kvantilu príslušného Fisher-Snedecorovho rozdelenia daného vyselektovaným modelom.

Kľúčové slová: Lineárny regresný model, Selekcia premenných, Akaikeho informačné kritérium, Testovanie signifikantnosti parametra , Chyba prvého druhu

Abstract

In our thesis, we focus on testing hypotheses about the significance of parameters of our interest after selecting other parameters in the model using the Akaike information criterion. We build on empirical results obtained through Monte Carlo simulations and aim to describe them using analytical methods. The thesis includes an overview of available literature on this topic, as well as our own results. We derive the probability that the Akaike criterion will select an incomplete model over a complete one. Then, we distinguish between two different approaches to selection. In the first approach, we allow the selection of the parameter whose significance we want to test, and we show that this approach is either identical to the procedure where we only test the hypothesis without conducting selection or to the procedure where we only perform selection without subsequent hypothesis testing, depending on the chosen test level. In the second approach, we insist on the tested parameter being included in the model. In this case, we estimate the type I error of the test conducted after variable selection using simulations. We show that this error depends on the actual parameter values of the model, which are unknown, and we solve this problem using a Bayesian approach to these variables. Finally, in the concluding chapter, we provide analyses that reveal the relationship between selection correlation of explanatory variables and the necessary correction of the quantile of the relevant Fisher-Snedecor distribution given by the selected model.

Keywords: Linear regression model, Selection of variables, Akaike information criterion, Test of parameters significance, Type I error

Obsah

Úvod	8
1 Lineárny regresný model	11
1.1 Odhad parametrov	11
1.2 Formulácia základných tvrdení využívaných v práci	12
2 Prehľad literatúry	18
2.1 Testovanie v modeli získanom selekciou premenných	19
2.2 Podreportovanie chyby testu v modeli získanom selekciou premenných	28
3 Selekcia nulovej premennej	33
3.1 Výber modelu na základe Akaikeho kritéria	34
3.2 Pravdepodobnosť selekcie nulovej premennej	40
4 Testovanie hypotéz o parametri po aplikovaní selekcie	44
4.1 Testovanie po aplikovaní selekcie na reálnych dátach	48
4.2 Analýza nepodmienennej pravdepodobnosti chyby prvého druhu	50
5 Analýza výberovej korelácie vysvetľujúcich premenných	52
Záver	56
Zoznam použitej literatúry	58

Úvod

Ako ľudstvo od nepamäti radi skúmame rôzne súvislosti medzi všakovakými javmi. Uspokojujeme tým jednak našu zvedavosť, ale najmä zvyšujeme našu schopnosť prežiť a vyvíjať sa. Všimli sme si, že dostatok vlahy a slnka má priaznivý efekt na našu úrodu, že dlhší pobyt v hygienicky nepriaznivom prostredí spôsobuje choroby a podobne. Vďaka tomu vieme vhodne reagovať na jednotlivé podnety a usmerňovať ich.

Na hľadanie všakovakých závislostí a súvislostí medzi rôznymi javmi a stavmi máme dnes už vcelku veľmi silné nástroje, vďaka mnohým výsledkom zo štatistiky. Práve štatistika nám vie často veľmi dobre naznačiť, či súvislosti ktoré nachádzame pozorovaním okolitého sveta sú skôr dielom náhody, alebo ide naopak o nejakú zákonitosť.

Jedným zo základných z týchto nástrojov je lineárna regresia. Pri nej predpokladáme, že premenná ktorú sme sa rozhodli skúmať (zväčša preto, lebo je pre nás v praxi z nejakého dôvodu dôležitá), je lineárne závislá od rôznych iných premenných. Táto závislosť však nie je deterministická, preto do hry vstupuje aj náhoda, ktorá môže túto lineárnu závislosť trochu (a niekedy možno aj viac) vychýliť jedným, alebo druhým smerom. Lineárna regresia nám potom mimo iného pomáha posudzovať odpovede na jednotlivé otázky (ktoré odborne nazývame hypotézami) o významnosti jednotlivých parametrov. Ak by napríklad niekto sformuloval hypotézu: “Množstvo zrážok nemá vplyv na množstvo úrody,” vedeli by sme sa pozrieť na dáta, aplikovať na ne lineárnu regresiu (teda predpoklad, že množstvo úrody závisí od množstva zrážok lineárne) a na základe toho úsudiť, že pravdivosť tejto hypotézy je veľmi nepravdepodobná.

V tejto práci sa budeme konkrétne zaoberať situáciou, keď nás bude zaujímať, či nejaké konkrétne premenné majú vplyv na premenú ktorú pozorujeme, pričom ale predpokladáme, že môže existovať množstvo iných vysvetľujúcich premenných, ktoré môžu pôsobiť na premennú ktorú pozorujeme. Ilustrujme si túto situáciu bližšie na nejakom príklade.

Predpokladajme, že máme firmu na výrobu hnojív, ktorá chce otestovať nový typ hnojiva. Predpokladajme, že existuje lineárna závislosť (aspoň na nejakom intervale) medzi množstvom použitého hnojiva a celkovou úrodou. Je ale jasné, že na množstvo úrody bude určite pôsobiť ešte veľa iných faktorov (zrážky, slnko, teplota, kvalita pôdy, množstvo škodcov, ...). S tým by však ešte nebol problém, stačilo by nám totiž iba vytvoriť

lineárny regresný model so všetkými týmito parametrami a testovať hypotézu: “Množstvo použitého hnojiva nemá vplyv na množstvo úrody.” Problém však je, že my v skutočnosti nevieme, ktoré z týchto “ostatných faktorov” sú naozaj dôležité, a ktoré sú také, že na výsledné množstvo úrody nemajú žiaden vplyv. Ak totiž takéto parametre zahrnieme do modelu, budú nás zbytočne mýliť a vychyľovať od skutočnosti.

Vyradenie rušivých parametrov z modelu vieme riešiť pomocou tzv. informačných kritérií (v tejto práci budeme používať najmä Akaikeho informačné kritérium, ale existujú aj iné, napr. Bayesovo informačné kritérium). Náš problém s hnojivami teda môžeme riešiť tak, že si vyberieme vhodný lineárny regresný model na základe Akaikeho kritéria a následne budeme testovať hypotézu: “Množstvo použitého hnojiva nemá vplyv na množstvo úrody.”

Tento prístup má však žiaľ tiež jeden závažný problém. Simulácie totiž naznačujú, že pri testovaní hypotézy takýmto postupom môžeme hypotézu nesprávne zamietnuť s väčšou pravdepodobnosťou, ako keby sme testovali hypotézu štandardným spôsobom (teda keby sme mali správny model rovno k dispozícii a nemuseli by sme si ho vyberať informačnými kritériami). Tento jav napr. ilustruje pomocou simulácií Marián Badár vo svojej bakalárskej práci [1]. Cieľom tejto diplomovej práce bude preto zaoberať sa týmto problémom po matematickej stránke a skúsiť ho čo najviac popísať a vysvetliť.

V prvej kapitole formulujeme základnú teóriu ohľadom lineárneho regresného modelu a formulujeme základné definície, vety a tvrdenia, ktoré budeme neskôr v práci využívať.

V druhej kapitole poskytneme čitateľovi základný prehľad dostupnej literatúry k danej téme a zároveň sa pokúsime čitateľovi ozrejmiť niektoré z hlavných výsledkov článku [3].

Kapitola 3 poskytuje hlavné teoretické výsledky tejto práce. Jej obsahom je výpočet pravdepodobnosti, že sa selekciou prostredníctvom Akaikeho kritéria vyberie neúplný model oproti úplnému a taktiež výpočet chyby prvého druhu testu o nulovosti sledovaného parametra, v prípade, ak umožníme aby bol tento parameter vyselektovaný z modelu.

Prípadom keď vyžadujeme aby bola testovaná premenná obsiahnutá v modeli sa zaoberáme v kapitole 4. Budeme sa snažiť odhadnúť, aká je skutočná pravdepodobnosť chyby prvého druhu daného testu vykonaného po selekcii. Skutočnosť že táto pravdepodobnosť závisí od parametrov modelu, ktoré v realite nepoznáme budeme riešiť pomocou bayesovského prístupu k daným parametrom a uvedené prístupy demonštrujeme na príkladoch.

Nakoniec v kapitole 5 uvedieme niektoré zo základných pozorovaní, týkajúcich sa vzájomného prepojenia medzi výberovou koreláciou vysvetľujúcich premenných a veľkosťou korekcie kvantilu Fisher-Snedecorovho rozdelenia (potrebnej na úpravu chybovosti testu), s ktorým porovnávame príslušnú testovú štatistiku.

1 Lineárny regresný model

Lineárna regresia je základným regresným modelom, často používaným v ekonometrii, všetkých prírodných vedách, ale aj mnohých iných odvetviach. Vieme ju charakterizovať nasledujúcim predpisom:

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \dots + x_{ik}\beta_k + \varepsilon_i \quad i = 1, 2, \dots, n \quad (1)$$

Premenná y_i je tzv. vysvetľovaná premenná a premenné $x_{i2}, x_{i3}, \dots, x_{ik}$ sú vysvetľujúce premenné (čiže ako tieto názvy naznačujú, pomocou x-ových premenných sa snažíme vysvetliť správanie premennej y). Zároveň y-ové aj x-ové premenné sú nám štandardne známe z dát. Hodnoty $\beta_1, \beta_2, \dots, \beta_k$ sú parametrami lineárneho regresného modelu, ktorých skutočnú hodnotu štandardne nepoznáme a snažíme sa ich odhadnúť. Premenná ε_i je náhodná premenná vnášajúca do modelu neistotu, často označovaná aj ako “náhodný šum”. Štandardne sa predpokladá (a my tak budeme tiež predpokladať naprieč celou diplomovou prácou), že premenné ε_i sú premenné typu iid (“independent, identically distributed”, čiže nezávislé a rovnako rozdelené) a platí $\varepsilon_i \sim N(0, \sigma^2)$, kde σ^2 je vo všeobecnosti neznáma disperzia. Nakoniec k je počet parametrov v modeli a n je počet pozorovaní.

Model (1) vieme zapísať aj maticovo v tvare:

$$Y = X\beta + \varepsilon. \quad (2)$$

V tomto zápise predstavuje Y n -rozmerný vektor premenných y_i , X označuje maticu rozmerov $n \times k$, pričom $X_{ij} = x_{ij}$ a ε je vektor premenných ε_i , čiže $\varepsilon \sim N(0, \sigma^2 I_n)$.

1.1 Odhad parametrov

Keď už máme zostavený lineárny regresný model, ostáva otázkou, ako odhadnúť neznáme parametre β , ktoré nepoznáme. Dá sa ukázať, že najlepší nevychýlený odhad $\hat{\beta}$ je argument minima funkcie $f(b) = (Y - Xb)^T(Y - Xb)$ (dôkaz možno nájsť napr. v knihe [5]).

Tvrdenie 1.1. *Za predpokladu plnej hodnosti matice X je argument minima funkcie $f(b) = (Y - Xb)^T(Y - Xb)$ vektor $\hat{\beta}$ tvaru:*

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

Ide o jedno z najznámejších tvrdení týkajúcich sa lineárnej regresie. Dôkaz možno taktiež nájsť napríklad v knihe [5].

1.2 Formulácia základných tvrdení využívaných v práci

V tejto podkapitole uvidíme sériu základných, všeobecne známych tvrdení, ktoré budeme potrebovať v tejto práci využiť.

Ako sme si už povedali, vektor $\hat{\beta}$ je najlepším nevychýleným odhadom neznámeho vektora β . O jeho presnom rozdelení nám hovorí nasledujúce tvrdenie:

Tvrdenie 1.2. *Náhodný vektor $\hat{\beta}$ má normálne rozdelenie so strednou hodnotou β a variančnou maticou $\sigma^2(X^T X)^{-1}$.*

Dôkaz. Vieme, že $\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1} X^T \varepsilon$.

Vidíme, že jediná náhodnosť v predpise náhodného vektora $\hat{\beta}$ je spôsobená náhodným vektorom ε . Vektor $\hat{\beta}$ je teda lineárnou transformáciou normálneho vektora, a teda $\hat{\beta}$ je tiež normálny vektor.

Vieme, že náhodný vektor $\hat{\beta}$ sa riadi normálnym rozdelením. Pre úplnosť dôkazu nám teda ostáva už iba určiť jeho strednú hodnotu a variančnú maticu:

$$E(\hat{\beta}) = E(\beta + (X^T X)^{-1} X^T \varepsilon) = E(\beta) + E((X^T X)^{-1} X^T \varepsilon) = \beta + (X^T X)^{-1} X^T E(\varepsilon) = \beta,$$

pričom sme využili, že β a X sú nenáhodné a $E(\varepsilon) = 0$.

$$\begin{aligned} Var(\hat{\beta}) &= Var(\beta + (X^T X)^{-1} X^T \varepsilon) = (X^T X)^{-1} X^T Var(\varepsilon) X (X^T X)^{-1} = \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}, \end{aligned}$$

pričom sme využili, že β a X sú nenáhodné a $Var(\varepsilon) = \sigma^2 I$.

□

V lineárnej regresii veľmi často využívame maticu H definovanú ako $H = X(X^T X)^{-1} X^T$ (ide o takzvaný projektor na priestor generovaný vysvetľujúcimi premennými) a maticu M definovanú ako $M = I - H$ (ide o takzvaný projektor na vektor rezíduí). V nasledujúcom tvrdení si dokážeme, že ide skutočne o projekčné matice.

Tvrdenie 1.3. *Matice H a M definované ako $H = X(X^T X)^{-1} X^T$ a $M = I - H$ sú maticami projekcie.*

Dôkaz. Vieme, že ak matica A je maticou projekcie, musí spĺňať dve vlastnosti, a to $A^T = A$ a $A^2 = A$. Cieľom tohto dôkazu je teda ukázať, že matice H a M spĺňajú obe tieto vlastnosti.

Očividne platí $H = X(X^T X)^{-1} X^T = H^T$. Zároveň $H^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H$. Potom pre maticu $M = I - H$ dostávame $M^T = I - H^T = I - H = M$ a $M^2 = (I - H)^2 = I - 2H + H^2 = I - 2H + H = I - H = M$ □

Pre projekčné matice budeme neskôr využívať nasledujúce tvrdenie:

Tvrdenie 1.4. *Ak A je maticou projekcie, tak pre jej vlastné hodnoty platí, že sú buď rovné 1, alebo rovné 0.*

Dôkaz. Nech x je vlastný vektor matice A pre vlastnú hodnotu λ . Potom platí $Ax = \lambda x$. Keďže A je projekčná matica, platí $A = A^2$, a teda:

$$\lambda x = Ax = A^2 x = A(Ax) = A(\lambda x) = \lambda Ax = \lambda^2 x.$$

Keďže máme $\lambda x = \lambda^2 x$ a vlastný vektor x isto nie je nulový, musí platiť $\lambda = \lambda^2$, a teda buď $\lambda = 0$, alebo $\lambda = 1$ □

Pred ďalším formulovaním tvrdení sa budeme zaoberať nasledujúcimi dvomi modelmi:

- $\mathcal{M}_1 : Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon_1,$
- $\mathcal{M}_2 : Y = X_1 \beta_1 + \varepsilon_2,$

pričom predpokladáme, že máme n pozorovaní a $\beta_1 \in \mathbb{R}^{k_1}$ a $\beta_2 \in \mathbb{R}^{k_2}$. Vidíme, že model $\mathcal{M}_2$ je v skutočnosti iba submodelom modelu $\mathcal{M}_1$. Označme blokovú maticu $[X_1 \ X_2]$ ako Z_1 , maticu X_1 ako Z_2 , vektor $(\beta_1 \ \beta_2)^T$ ako θ_1 a vektor β_2 ako θ_2 . Modely $\mathcal{M}_1$ a $\mathcal{M}_2$ potom môžeme maticovo zapísať nasledujúcim spôsobom:

- $\mathcal{M}_1 : Y = Z_1\theta_1 + \varepsilon_1$
- $\mathcal{M}_2 : Y = Z_2\theta_2 + \varepsilon_2$

Nakoniec zdefinujme matice $H_1 = Z_1(Z_1^T Z_1)^{-1} Z_1^T$, $H_2 = Z_2(Z_2^T Z_2)^{-1} Z_2^T$, $M_1 = I - H_1$ a $M_2 = I - H_2$, o ktorých budeme formulovať nasledujúce dve tvrdenia.

Tvrdenie 1.5. *Matice H_1 a H_2 formulované vyššie komutujú a platí: $H_2 H_1 = H_1 H_2 = H_2$.*

Dôkaz. Nech matica X_1 je matica rozmerov $n \times k_1$ a matica X_2 je matica rozmerov $n \times k_2$.

Definujme maticu A tvaru: $A := \begin{bmatrix} I_{k_1} \\ 0_{k_2 \times k_1} \end{bmatrix}$.

Potom platí:

$$Z_2 = X_1 = [X_1 \ X_2] \begin{bmatrix} I_{k_1} \\ 0_{k_2 \times k_1} \end{bmatrix} = Z_1 A.$$

Pre maticu H_2 teda dostávame vyjadrenie:

$$H_2 = Z_2(Z_2^T Z_2)^{-1} Z_2^T = Z_1 A(A^T Z_1^T Z_1 A)^{-1} A^T Z_1^T.$$

Potom pre súčin $H_1 H_2$ dostávame:

$$H_1 H_2 = Z_1(Z_1^T Z_1)^{-1} Z_1^T Z_1 A(A^T Z_1^T Z_1 A)^{-1} A^T Z_1^T = Z_1 A(A^T Z_1^T Z_1 A)^{-1} A^T Z_1^T = H_2$$

a pre súčin $H_2 H_1$ máme:

$$H_2 H_1 = Z_1 A(A^T Z_1^T Z_1 A)^{-1} A^T Z_1^T Z_1 (Z_1^T Z_1)^{-1} Z_1 = Z_1 A(A^T Z_1^T Z_1 A)^{-1} A^T Z_1^T = H_2.$$

□

Tvrdenie 1.6. Matice M_1 a M_2 formulované vyššie komutujú a platí: $M_2M_1 = M_1M_2 = M_1$.

Dôkaz. Platí:

$$M_1M_2 = (I - H_1)(I - H_2) = I - H_1 - H_2 + H_1H_2.$$

S využitím tvrdenia 1.5 dostávame $H_1H_2 = H_2$, a teda $M_1M_2 = I - H_1 = M_1$. Analogicky:

$$M_2M_1 = (I - H_2)(I - H_1) = I - H_1 - H_2 + H_2H_1 = I - H_1 = M_1.$$

□

Zavedme náhodné premenné vyjadrujúce súčet štvorcov rezíduí (residual sum od squares) $RSS_1 = (Y - Z_1\hat{\theta}_1)^T(Y - Z_1\hat{\theta}_1)$, $RSS_2 = (Y - Z_2\hat{\theta}_2)^T(Y - Z_1\hat{\theta}_2)$. Pre tieto náhodné premenné uvádzame nasledujúce, v ekonometrii všeobecne známe tvrdenie.

Tvrdenie 1.7. Definujme nasledujúcu testovú štatistiku:

$$F = \frac{n - k_1 - k_2}{k_2} \frac{RSS_2 - RSS_1}{RSS_1}$$

Za predpokladu nulovosti parametra β_2 a úplnej hodnosti matice Z_1 platí:

$$F \sim F_{k_2, n - k_1 - k_2}.$$

Nasleduje tvrdenie, ktoré hovorí o ortogonálnej transformácii normálneho náhodného vektora s nulovou strednou hodnotou a identickou variančnou maticou.

Tvrdenie 1.8. Nech pre náhodný vektor N platí $N \sim N(0_n, I_n)$ a Q je matica rozmerov $n \times q$, pričom jej stĺpce sú navzájom kolmé a majú dĺžku 1 (v euklidovskej norme). Potom pre rozdelenie náhodného vektora Q^TN platí $Q^TN \sim N(0_q, I_q)$.

Dôkaz. Keďže Q^TN je lineárna transformácia normálneho vektora, vektor Q^TN je tiež normálnym vektorom. Na dokončenie dôkazu stačí teda už len vyjadriť jeho strednú hodnotu a variančnú maticu:

$$E(Q^TN) = Q^TE(N) = Q^T0_n = 0_q,$$

pričom sme mohli Q^T vyňať pred strednú hodnotu, keďže ide o nenáhodnú maticu.

$$\text{Var}(Q^T N) = Q^T \text{Var}(N) Q = Q^T I_n Q = Q^T Q = I_q,$$

čím sme dokázali $Q^T N \sim N(0_q, I_q)$. □

V ekonometrii často využívame súčet štvorcov rezíduí, označovaný ako RSS (residual sum of squares). Bude preto potrebné vysloviť nasledujúce tvrdenie o jeho pravdepodobnostnom rozdelení.

Tvrdenie 1.9. *Majme lineárny regresný model $Y = X\beta + \varepsilon$ obsahujúci n pozorovaní a $k < n$ parametrov, pričom $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$. Definujme $RSS := (Y - X\hat{\beta})^T(Y - X\hat{\beta})$. Potom za predpokladu úplnej hodnosti matice X pre pravdepodobnostné rozdelenie náhodnej premennej RSS platí:*

$$RSS \sim \sigma_\varepsilon^2 \chi_{n-k}^2$$

Neskôr v práci sa budeme potrebovať odvolať na nasledujúce dve tvrdenia, umožňujúce pracovať s podmienenou strednou hodnotou a podmienenou varianciou.

Tvrdenie 1.10. *Nech X, Y sú ľubovoľné náhodné vektory. Potom pre strednú hodnotu $E(X)$ platí nasledujúca rovnosť:*

$$E(X) = E(E(X | Y)).$$

Dôkaz. Dôkaz urobíme iba pre spojité náhodné vektory. Nech náhodné vektory X, Y majú rozmery k a m a združenú hustotu $f(x, y)$. Potom pre podmienenú strednú hodnotu $E(X | Y)$ máme vzorec:

$$E(X | Y) = \frac{\int_{\mathbb{R}^k} x f(x, Y) dx}{\int_{\mathbb{R}^k} f(x, Y) dx}.$$

Opätovným aplikovaním strednej hodnoty dostávame:

$$E(E(X | Y)) = \int_{\mathbb{R}^m} \frac{\int_{\mathbb{R}^k} x f(x, y) dx}{\int_{\mathbb{R}^k} f(x, y) dx} f_Y(y) dy, \quad (3)$$

kde $f_Y(y) = \int_{\mathbb{R}^k} f(x, y) dx$ je marginálna hustota náhodnej premennej Y . Rovnosť (3) sa nám tým pádom upraví na tvar:

$$E(E(X | Y)) = \int_{\mathbb{R}^m} \int_{\mathbb{R}^k} x f(x, y) dx dy.$$

Zámenou poradia integrálov dostávame:

$$E(E(X | Y)) = \int_{\mathbb{R}^k} \int_{\mathbb{R}^m} x f(x, y) dy dx.$$

Vektor x je konštantou vzhľadom na integračnú premennú y , a teda ho môžeme vyňať pred zátvorku, čiže vieme napísať nasledujúcu sériu rovností:

$$E(E(X | Y)) = \int_{\mathbb{R}^k} x \int_{\mathbb{R}^m} f(x, y) dy dx = \int_{\mathbb{R}^k} x f_X(x) dx = E(X),$$

čím sme dokázali vyslovené tvrdenie. □

Tvrdenie 1.11. *Nech X, Y sú ľubovoľné náhodné vektory. Potom pre varianciu $Var(X)$ platí:*

$$Var(X) = E(Var(X | Y)) + Var(E(X | Y)).$$

Dôkaz. Z definície variance máme pre $Var(X | Y)$ nasledujúci predpis:

$$Var(X | Y) = E[(X | Y - E(X | Y))(X | Y - E(X | Y))^T],$$

čo vieme upraviť do tvaru:

$$Var(X | Y) = E(XX^T | Y) - E(X | Y)[E(X | Y)]^T. \quad (4)$$

Aplikovaním strednej hodnoty a využitím tvrdenia 1.10 dostávame:

$$\begin{aligned} E(Var(X | Y)) &= E(E(XX^T | Y)) - E[E(X | Y)[E(X | Y)]^T] = \\ &= E(XX^T) - E[E(X | Y)[E(X | Y)]^T] - E(E(X | Y))[E(E(X | Y))]^T + E(E(X | Y))[E(E(X | Y))]^T \\ &= E(XX^T) - Var(E(X | Y)) + E(X)[E(X)]^T = \\ &= Var(X) - Var(E(X | Y)). \end{aligned}$$

Z uvedenej série rovností sme dostali:

$$E(Var(X | Y)) = Var(X) - Var(E(X | Y)),$$

z čoho už priamo vyplýva:

$$Var(X) = E(Var(X | Y)) + Var(E(X | Y)).$$

□

2 Prehľad literatúry

V tejto kapitole sa pokúsime čitateľovi poskytnúť prehľad základnej literatúry zaoberajúcej sa podobnou problematikou. Jedným z prvých článkov zaoberajúcich sa problémom testovania hypotéz v modeli, ktorý vznikol selekciou bol článok [2]. V tomto článku sa autor zaoberá lineárnym regresným modelom obsahujúcim dve vysvetľujúce premenné, pričom jednu z nich sa snaží vyselektovať a pre druhú z nich sa snaží nájsť odhad (ktorý berie v úvahu uvedenú selekciu), pre ktorý v závere odvádza prvý moment a výchylku (bias).

V článku [4] sa autor zaoberá situáciou, keď máme na výber väčšie množstvo regresorov, ktoré je možné zahrnúť do modelu, ale máme vopred fixne stanovený počet, koľko z týchto regresorov chceme pomocou selekcie do tohto modelu vybrať. Za istých zjednodušujúcich predpokladov (známa variancia rezíduí a nulová výberová korelácia vysvetľujúcich premenných) potom zavedie takzvané pravidlo palca (rule of thumb) upravujúce hodnotu hladiny klasického Studentovho t-testu vo vyselektovanom modeli.

Sumár najnovších poznatkov ohľadom testovania hypotéz v modeloch, ktoré prešli selekciou je obsiahnutý v článku [6], v ktorom sa autor zaoberá otázkou, ako najlepšie odhadnúť parametre modelu, ktorý bol získaný selekciou, v zmysle minimalizovania strednej kvadratickej chyby odhadu (mean squared error).

V nasledujúcej podkapitole sa pokúsime čitateľovi bližšie priblížiť článok [3], ktorý sa zaoberá otázkou, akej veľkej chyby (v zmysle odlišnosti skutočnej hodnoty parametra od odhadu pre daný parameter) sa vieme pri odhadovaní parametrov v modeli, ak ignorujeme fakt, že daný model bol vybraný selekciou z viacerých prípustných modelov. Keďže je tento článok zameraný pre odbornú verejnosť, sú v ňom viaceré myšlienky a tvrdenia uvedené bez bližšieho kontextu, čo môže spôsobovať nezainteresovanému čitateľovi značné ťažkosti. Naším cieľom preto bude daný text sprístupniť aj pre čitateľa, ktorý síce má základné matematické a štatistické vzdelanie, ale nie je v danej téme odborníkom.

2.1 Testovanie v modeli získanom selekciou premenných

Skúmame lineárny regresný model $\mathcal{M}$ nasledujúceho tvaru:

$$\mathcal{M} : y = X\beta + Z\gamma + \varepsilon,$$

pričom β je k - rozmerný vektor parametrov, ktorých signifikantnosť chceme testovať (teda chceme testovať hypotézu $H_0 : \beta = 0$) a γ je m -rozmerný vektor takzvaných rušivých parametrov, ktorých skutočná hodnota nás nezaujíma (a môžeme sa preto pokúsiť ich z modelu vyselektovať). Vektor y je potom n - rozmerný vektor vysvetľovaných premenných, X, Z sú $n \times k$ a $n \times m$ - rozmerné matice vysvetľujúcich premenných a ε je n - rozmerný vektor chýb.

Pre parametre β, γ urobíme metódou najmenších štvorcov ich odhady $\hat{\beta}$ a $\hat{\gamma}$:

$$\begin{aligned} \min_{\beta, \gamma} F(\beta, \gamma) &= (y - X\beta - Z\gamma)^T (y - X\beta - Z\gamma) \\ &= y^T y + \beta^T X^T X \beta + \gamma^T Z^T Z \gamma - 2y^T X \beta - 2y^T Z \gamma + 2\beta^T X^T Z \gamma. \end{aligned}$$

Z nutnej podmienky optima potom dostávame:

$$\frac{\partial F(\hat{\beta}, \hat{\gamma})}{\partial \beta} = 2X^T X \hat{\beta} - 2X^T y + 2X^T Z \hat{\gamma} = 0, \quad (5)$$

$$\frac{\partial F(\hat{\beta}, \hat{\gamma})}{\partial \gamma} = 2Z^T Z \hat{\gamma} - 2Z^T y + 2Z^T X \hat{\beta} = 0. \quad (6)$$

Úpravou rovnice (5) dostávame:

$$\begin{aligned} X^T X \hat{\beta} &= X^T y - X^T Z \hat{\gamma}, \\ \hat{\beta} &= (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T Z \hat{\gamma}. \end{aligned} \quad (7)$$

Analogicky úpravou rovnice (6) dostávame:

$$\hat{\gamma} = (Z^T Z)^{-1} Z^T y - (Z^T Z)^{-1} Z^T X \hat{\beta}.$$

Dosadením (7) dostávame:

$$\hat{\gamma} = (Z^T Z)^{-1} Z^T y - (Z^T Z)^{-1} Z^T X [(X^T X)^{-1} X^T y - (X^T X)^{-1} X^T Z \hat{\gamma}]. \quad (8)$$

Zavedme značenie:

$$H_X = X(X^T X)^{-1} X^T, \quad M_X = I_n - H_X. \quad (9)$$

Úpravou rovnice (8) potom dostávame:

$$\begin{aligned}(Z^T Z - Z^T H_X Z) \hat{\gamma} &= Z^T y - Z^T H_X y, \\ Z^T M_X Z \hat{\gamma} &= Z^T (I_n - H_X) y = Z^T M_X y, \\ \hat{\gamma} &= (Z^T M_X Z)^{-1} Z^T M_X y.\end{aligned}\tag{10}$$

Dosadením do (5) potom dostávame:

$$\hat{\beta} = (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T Z (Z^T M_X Z)^{-1} Z^T M_X y.\tag{11}$$

Ľahko teraz vyriešime aj postačujúcu podmienku minima, pretože rovno dostávame:

$$\begin{bmatrix} \frac{\partial^2 F(\hat{\beta}, \hat{\gamma})}{\partial \beta^2} & \frac{\partial^2 F(\hat{\beta}, \hat{\gamma})}{\partial \beta \partial \gamma} \\ \frac{\partial^2 F(\hat{\beta}, \hat{\gamma})}{\partial \beta \partial \gamma} & \frac{\partial^2 F(\hat{\beta}, \hat{\gamma})}{\partial \gamma^2} \end{bmatrix} = \begin{bmatrix} 2X^T X & 0 \\ 0 & 2Z^T Z \end{bmatrix}.$$

Ľahko vidíme, že Hessova matica je kladne definitná, keďže matice $X^T X$ a $Z^T Z$ sú kladne definitné.

Zavedme nasledujúce značenie:

$$b_r = (X^T X)^{-1} X^T y, \quad \hat{\theta} = (Z^T M_X Z)^{\frac{1}{2}} \hat{\gamma}, \quad Q = (X^T X)^{-1} X^T Z (Z^T M_X Z)^{-\frac{1}{2}}.\tag{12}$$

V zmysle značenia (12) vieme teraz rovnosť (11) zapísať nasledujúcim spôsobom:

$$\hat{\beta} = b_r - Q \hat{\theta}.\tag{13}$$

Označme odhad (13) parametra β v úplnom modeli $\mathcal{M}$ ako b_u (unrestricted estimator of β). Všimnime si, že ak by sme parameter β chceli metódou najmenších štvorcov odhadnúť v submodeli $\mathcal{M}_s$:

$$\mathcal{M}_s : y = X\beta + \varepsilon,$$

dostávame podľa tvrdenia 1.1 odhad metódou najmenších štvorcov pre parameter β rovný b_r (restricted estimator of β).

Autori Danilov a Magnus potom navrhujú vo svojom článku [3] zaviesť takzvaný WALS (weighted average least - squares estimator) odhad b parametra β , definovaný predpisom:

$$b = \lambda b_u + (1 - \lambda) b_r,$$

pričom $\lambda \in [0, 1]$. Ide teda o vážený priemer odhadu parametra β v úplnom modeli $\mathcal{M}$ a v submodeli $\mathcal{M}_s$, pričom voľba parametra λ je voliteľná ľubovoľne.

Doteraz sme sa zaoberali situáciou, keď sme brali do úvahy iba dva možné modely, úplný model $\mathcal{M}$ a submodel $\mathcal{M}_s$. Keďže vektor γ je ale m - rozmerný vektor, máme teoreticky až 2^m možných submodelov modelu $\mathcal{M}$. Označme teda $\mathcal{M}_i$ ako i -ty submodel modelu $\mathcal{M}$, pričom tento model obsahuje niektoré parametre vektora γ . Nech S_i^T je ľubovoľná permutácia matice $[I_{r_i} \ 0]$ (pričom $r_i \in 1, 2, \dots, m$ je hodnosť identickej matice I_{r_i}). Pre model $\mathcal{M}_i$ teda dostávame nasledujúci predpis:

$$\mathcal{M}_i : y = X\beta + ZS_i\gamma_i + \varepsilon,$$

pričom γ_i je príslušný podvektor vektora γ a matica S_i je zvolená tak, aby stĺpce matice ZS_i zodpovedali príslušným parametrom. Pomocou metódy najmenších štvorcov teraz vieme nájsť odhady parametrov β a γ_i modelu $\mathcal{M}_i$ analogicky ako tomu bolo v prípade úplného modelu $\mathcal{M}$, akurát v roli matice Z bude tentokrát vystupovať matica ZS_i . Pre odhady b_i, c_i parametrov β a γ_i modelu $\mathcal{M}_i$ dostávame:

$$b_i = b_r - (X^T X)^{-1} X^T ZS_i c_i, \quad (14)$$

$$c_i = (S_i^T Z^T M_X ZS_i)^{-1} S_i^T Z^T M_X y. \quad (15)$$

Rovnicu (14) vieme ďalej upraviť. Dosadením vzťahu (15) za c_i dostávame:

$$b_i = b_r - (X^T X)^{-1} X^T ZS_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T Z^T M_X y,$$

$$b_i = b_r - (X^T X)^{-1} X^T Z (Z^T M_X Z)^{-\frac{1}{2}} (Z^T M_X Z)^{\frac{1}{2}} S_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T Z^T M_X y.$$

Použitím značenia (12) dostávame:

$$b_i = b_r - Q(Z^T M_X Z)^{\frac{1}{2}} S_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T Z^T M_X y,$$

$$b_i = b_r - Q(Z^T M_X Z)^{\frac{1}{2}} S_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T (Z^T M_X Z) (Z^T M_X Z)^{-1} Z^T M_X y,$$

čo sa nám po využití vzťahov (10) a (12) upraví do tvaru:

$$b_i = b_r - Q(Z^T M_X Z)^{\frac{1}{2}} S_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T (Z^T M_X Z) \hat{\gamma},$$

$$b_i = b_r - Q(Z^T M_X Z)^{\frac{1}{2}} S_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T (Z^T M_X Z)^{\frac{1}{2}} \hat{\theta}.$$

Zavedením značenia:

$$W_i = (Z^T M_X Z)^{\frac{1}{2}} S_i (S_i^T Z^T M_X ZS_i)^{-1} S_i^T (Z^T M_X Z)^{\frac{1}{2}}, \quad (16)$$

dostávame:

$$b_i = b_r - QW_i \hat{\theta}. \quad (17)$$

Poznámka: Dá sa ľahko overiť, že matica W_i , závislá iba od známych matíc Z, X a S_i , je idempotentná.

V nasledujúcich krokoch si tiež ukážeme, že odhad c_i je závislý iba od náhodného vektora $\hat{\theta}$:

$$\begin{aligned} c_i &= (S_i^T Z^T M_X Z S_i)^{-1} S_i^T Z^T M_X y, \\ c_i &= (S_i^T Z^T M_X Z S_i)^{-1} S_i^T (Z^T M_X Z)^{\frac{1}{2}} (Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X y, \end{aligned}$$

a teda s využitím vzťahu (10) a značenia (12) dostávame:

$$c_i = (S_i^T Z^T M_X Z S_i)^{-1} S_i^T (Z^T M_X Z)^{\frac{1}{2}} \hat{\theta}.$$

Podobne ako tomu bolo v prípade keď sme mali k dispozícii iba modely $\mathcal{M}$ a $\mathcal{M}_s$, môžeme definovať takzvaný WALS odhad b parametra β nasledujúcim spôsobom:

$$b := \sum_{i=1}^{2^m} \lambda_i b_i. \quad (18)$$

Znova ide o vážený priemer odhadov b_i , a teda požadujeme, aby váhy λ_i spĺňali podmienky $\lambda_i \geq 0$ a $\sum_{i=1}^{2^m} \lambda_i = 1$. Navyše budeme ďalej požadovať, aby váhy λ_i boli závislé iba od rezíduí, a teda λ_i sú náhodné premenné závislé iba od náhodného vektora $M_X y$ ($\lambda_i = \lambda_i(M_X y)$). Potom vieme vyjadrenie (18) s využitím vzťahu (17) upraviť nasledujúcim spôsobom:

$$b = \sum_{i=1}^{2^m} \lambda_i b_i = \sum_{i=1}^{2^m} \lambda_i (b_r - Q W_i \hat{\theta}) = b_r \sum_{i=1}^{2^m} \lambda_i - Q \sum_{i=1}^{2^m} \lambda_i W_i \hat{\theta} = b_r - Q \sum_{i=1}^{2^m} \lambda_i W_i \hat{\theta}.$$

Označme vážený priemer idempotentných matíc $\sum_{i=1}^{2^m} \lambda_i W_i$ ako W . Matica W je potom tiež symetrická, a keďže matice W_i sú idempotentné, a teda kladne semidefinitné (ich vlastné hodnoty sú iba 0 a 1), aj matica W bude kladne semidefinitná. Pre WALS odhad b parametra β tak dostávame finálne vyjadrenie:

$$b = b_r - Q W \hat{\theta} \quad (19)$$

Keďže sa nám podarilo vyjadriť odhad b vo vhodnom tvare, bude nás teraz samozrejme zaujímať stredná hodnota, variancia a stredná kvadratická chyba tohto odhadu. Predtým ako ale vyslovíme a dokážeme tvrdenie o ich tvare, uvedieme nasledujúce pomocné tvrdenie.

Tvrdenie 2.1. Náhodné vektory b_r a $\hat{\theta}$ sú nezávislé.

Dôkaz. Dosadením značenia (12) a vzťahu (10) dostávame:

$$b_r = (X^T X)^{-1} X^T y,$$

$$\hat{\theta} = (Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X y.$$

Všimnime si, že náhodné vektory b_r a $\hat{\theta}$ sú iba lineárnou transformáciou normálneho vektora y , a teda sú to tiež normálne vektory. Aby sme teda ukázali ich nezávislosť, stačí nám dokázať, že ich kovariančná matica je nulová (je totiž známe, že ak je kovariancia dvoch normálnych náhodných vektorov nulová, tak sú nezávislé). Poďme teda túto kovarianciu spočítať:

$$\begin{aligned} \text{cov}(b_r, \hat{\theta}) &= \text{cov}((X^T X)^{-1} X^T y, (Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X y) = (X^T X)^{-1} X^T \text{cov}(y, y) M_X Z (Z^T M_X Z)^{-\frac{1}{2}} = \\ &= \sigma_\varepsilon^2 (X^T X)^{-1} X^T M_X Z (Z^T M_X Z)^{-\frac{1}{2}}. \end{aligned}$$

Použitím vzťahu (9) nakoniec dostávame:

$$\begin{aligned} \text{cov}(b_r, \hat{\theta}) &= \sigma_\varepsilon^2 (X^T X)^{-1} X^T M_X Z (Z^T M_X Z)^{-\frac{1}{2}} = \sigma_\varepsilon^2 (X^T X)^{-1} X^T (I - X(X^T X)^{-1} X^T) Z (Z^T M_X Z)^{-\frac{1}{2}} \\ &= \sigma_\varepsilon^2 (X^T X)^{-1} (X^T - X^T) Z (Z^T M_X Z)^{-\frac{1}{2}} = 0, \end{aligned}$$

a teda b_r a $\hat{\theta}$ sú nezávislé. □

Nakoniec uveďme pred vyslovením ďalšieho tvrdenia nasledujúce značenie:

$$\eta = \frac{1}{\sigma_\varepsilon} \theta, \quad \hat{\eta} = \frac{1}{\sigma_\varepsilon} \hat{\theta}. \quad (20)$$

Tvrdenie 2.2. Za predpokladu, že váhy λ_i sú náhodnými premennými závislými iba od $M_X y$, spĺňajúce $\lambda_i \geq 0$ a $\sum_{i=1}^{2^m} \lambda_i = 1$, platí:

- $E(b) = \beta - \sigma_\varepsilon Q E(W \hat{\eta} - \eta),$
- $\text{Var}(b) = \sigma_\varepsilon^2 ((X^T X)^{-1} + Q \text{Var}(W \hat{\eta}) Q^T),$
- $\text{MSE}(b) = \sigma_\varepsilon^2 ((X^T X)^{-1} + Q \text{MSE}(W \hat{\eta}) Q^T),$

pričom $\text{MSE}(b)$ (mean squared error) definujeme ako $E((b - \beta)(b - \beta)^T)$ a $\text{MSE}(W \hat{\eta})$ definujeme ako $E((W \hat{\eta} - \eta)(W \hat{\eta} - \eta)^T)$.

Dôkaz. V zmysle tvrdenia 1.10 môžeme strednú hodnotu $E(b)$ vyjadriť ako $E(E(b | M_X y))$.

Vyjadríme si teda podmienenú strednú hodnotu $E(b | M_X y)$:

$$E(b | M_X y) = E(b_r - QW\hat{\theta} | M_X y) = E(b_r | M_X y) - E(QW\hat{\theta} | M_X y). \quad (21)$$

Všimnime si, že v zmysle značenia (12) je $\hat{\theta}$ závislé iba od $\hat{\gamma}$, pričom z vyjadrenia (10) si môžeme všimnúť, že $\hat{\gamma}$ je determinovaná hodnotou $M_X y$. Matica $W = \sum_{i=1}^{2m} \lambda_i W_i$ je tiež determinovaná hodnotou $M_X y$, pretože hodnoty λ_i sú z predpokladu determinované touto hodnotou a W_i sú nenáhodné. Matica Q je tiež nenáhodná, a teda celý súčin $QW\hat{\theta}$ je determinovaný hodnotou $M_X y$, z čoho už priamo vyplýva $E(QW\hat{\theta} | M_X y) = QW\hat{\theta}$.

Náhodné vektory b_r a $M_X y$ sú oba lineárnou transformáciou normálneho vektora y , a teda sú tiež normálne. Ľahko sa ukáže, že ich kovariancia je nulová, čiže náhodné vektory b_r a $M_X y$ sú nezávislé:

$$\begin{aligned} \text{cov}(b_r, M_X y) &= \text{cov}((X^T X)^{-1} X^T y, (I - X(X^T X)^{-1} X^T) y) = \\ &= (X^T X)^{-1} X^T \text{cov}(y, y) (I - X(X^T X)^{-1} X^T) = \sigma_\varepsilon^2 (X^T X)^{-1} X^T (I - X(X^T X)^{-1} X^T) = \\ &= \sigma_\varepsilon^2 ((X^T X)^{-1} X^T - (X^T X)^{-1} X^T) = 0. \end{aligned}$$

Z nezávislosti vektorov b_r a $M_X y$ vyplýva, že $E(b_r | M_X y) = E(b_r)$. Vzťah (21) tak môžeme upraviť do tvaru:

$$\begin{aligned} E(b | M_X y) &= E(b_r) - QW\hat{\theta} = (X^T X)^{-1} X^T E(y) - QW\hat{\theta} = \\ &= (X^T X)^{-1} X^T E(X\beta + Z\gamma + \varepsilon) - QW\hat{\theta} = \beta + (X^T X)^{-1} X^T Z\gamma - QW\hat{\theta}, \end{aligned}$$

a teda:

$$E(b | M_X y) = \beta + (X^T X)^{-1} X^T Z\gamma - QW\hat{\theta}. \quad (22)$$

Potom:

$$E(b) = E(E(b | M_X y)) = E(\beta + (X^T X)^{-1} X^T Z\gamma - QW\hat{\theta}) = \beta + (X^T X)^{-1} X^T Z\gamma - E(QW\hat{\theta}).$$

Použitím značenia (20) dostávame:

$$E(b) = \beta + (X^T X)^{-1} X^T Z\gamma - \sigma_\varepsilon Q E(W\hat{\eta}) = \beta + (X^T X)^{-1} X^T Z (Z^T M_X Z)^{-\frac{1}{2}} (Z^T M_X Z)^{\frac{1}{2}} \gamma - \sigma_\varepsilon Q E(W\hat{\eta}),$$

čo aplikovaním značenia (12) vieme upraviť do nasledujúceho taru:

$$E(b) = \beta + Q\theta - \sigma_\varepsilon Q E(W\hat{\eta}) = \beta + \sigma_\varepsilon Q\eta - \sigma_\varepsilon Q E(W\hat{\eta}),$$

z čoho už priamo dostávame prvú časť vysloveného tvrdenia:

$$E(b) = \beta - \sigma_\varepsilon Q(W\hat{\eta} - \eta). \quad (23)$$

Využitím tvrdenia 1.11 dostávame:

$$Var(b) = E(Var(b | M_X y)) + Var(E(b | M_X y)). \quad (24)$$

Hodnotu $Var(b | M_X y)$ z rovnosti (24) vieme pomocou (19) upraviť nasledujúcim spôsobom:

$$\begin{aligned} Var(b | M_X y) &= Var(b_r - QW\hat{\theta} | M_X y) = Var(b_r | M_X y) = Var(b_r) = \\ &= Var((X^T X)^{-1} X^T y) = (X^T X)^{-1} X^T Var(y) X (X^T X)^{-1} = \sigma_\varepsilon^2 (X^T X)^{-1}, \end{aligned}$$

a teda

$$Var(b | M_X y) = Var(b_r) = \sigma_\varepsilon^2 (X^T X)^{-1}. \quad (25)$$

pričom sme využili už spomenuté fakty, že matica W a odhad $\hat{\theta}$ sú deteminované vektorom $M_X y$ a vektor b_r je nezávislý od $M_X y$. Keďže matica $\sigma_\varepsilon^2 (X^T X)^{-1}$ je nenáhodná, tak aj $E(\sigma_\varepsilon^2 (X^T X)^{-1}) = \sigma_\varepsilon^2 (X^T X)^{-1}$.

Pozrime sa teraz na druhý sčítanec rovnosti (24). Využitím (22) dostávame:

$$Var(E(b | M_X y)) = Var(\beta + (X^T X)^{-1} X^T Z \gamma - QW\hat{\theta}).$$

Teraz už len stačí využiť, že β a $(X^T X)^{-1} X^T Z \gamma$ sú nenáhodné a vzťah (20), čím dostávame:

$$Var(E(b | M_X y)) = \sigma_\varepsilon^2 Q Var(W\hat{\eta}) Q^T.$$

Z toho nám spojením oboch sčítancov v rovnosti (24) už priamo vyplýva druhá časť vysloveného tvrdenia:

$$Var(b) = E(Var(b | M_X y)) + Var(E(b | M_X y)) = \sigma_\varepsilon^2 (X^T X)^{-1} + \sigma_\varepsilon^2 Q Var(W\hat{\eta}) Q^T,$$

a teda:

$$Var(b) = \sigma_\varepsilon^2 ((X^T X)^{-1} + Q Var(W\hat{\eta}) Q^T).$$

Nakoniec hodnotu $MSE(b)$ vieme vyjadriť nasledujúcim spôsobom:

$$MSE(b) = E((b - \beta)(b - \beta)^T) = Var(b - \beta) + E(b - \beta)(E(b - \beta))^T =$$

$$\begin{aligned}
&= \text{Var}(b) + \sigma_\varepsilon Q E(W\hat{\eta} - \eta)(E(W\hat{\eta} - \eta))^T Q^T \sigma_\varepsilon = \\
&= \sigma_\varepsilon^2 ((X^T X)^{-1} + Q \text{Var}(W\hat{\eta}) Q^T) + \sigma_\varepsilon^2 Q E(W\hat{\eta} - \eta)(E(W\hat{\eta} - \eta))^T Q^T = \\
&= \sigma_\varepsilon^2 ((X^T X)^{-1} + Q(\text{Var}(W\hat{\eta}) + E(W\hat{\eta} - \eta)(E(W\hat{\eta} - \eta))^T) Q^T) = \\
&\sigma_\varepsilon^2 ((X^T X)^{-1} + Q(\text{Var}(W\hat{\eta} - \eta) + E(W\hat{\eta} - \eta)(E(W\hat{\eta} - \eta))^T) Q^T) = \\
&= \sigma_\varepsilon^2 ((X^T X)^{-1} + Q \text{MSE}(W\hat{\eta}) Q^T),
\end{aligned}$$

čím sme dokázali aj poslednú časť vysloveného tvrdenia. $\square$

Poznámka: Všimnime si, že $\text{MSE}(b)$, ani $\text{Var}(b)$ nijak nezávisia od hodnoty β (závisia iba od hodnoty γ , respektíve η).

Podme sa ďalej zaoberať strednou kvadratickou chybou odhadu b_u , ktorý má podľa (13) nasledujúci predpis $b_u = b_r - Q\hat{\theta}$ (odhad parametra β v úplnom modeli). Potom:

$$\text{MSE}(b_u) = \text{Var}(b_u) + E(b_u - \beta)(E(b_u - \beta))^T. \quad (26)$$

Zamerajme sa teda na $E(b_u)$. V zmysle tvrdenia 1.10 môžeme strednú hodnotu odhadu b_u vyjadriť nasledujúcim spôsobom:

$$E(b_u) = E(E(b_u \mid M_X y)).$$

Potom:

$$E(b_u \mid M_X y) = E(b_r - Q\hat{\theta} \mid M_X y) = E(b_r \mid M_X y) - E(Q\hat{\theta} \mid M_X y).$$

V dôkaze tvrdenia 2.2 sme tiež dokázali, že vektory b_r a $M_X y$ sú nezávislé, z čoho vyplýva $E(b_r \mid M_X y) = E(b_r)$. Taktiež vieme, že hodnota $\hat{\theta}$ je determinovaná vektorom $M_X y$, a teda $E(Q\hat{\theta} \mid M_X y) = Q\hat{\theta}$, a teda dostávame:

$$\begin{aligned}
E(b_u \mid M_X y) &= E(b_r) - Q\hat{\theta} = E((X^T X)^{-1} X^T y) - Q\hat{\theta} = (X^T X)^{-1} X^T E(X\beta + Z\gamma + \varepsilon) - Q\hat{\theta} = \\
&= \beta + (X^T X)^{-1} X^T Z\gamma - Q(Z^T M_X Z)^{\frac{1}{2}} \hat{\gamma} = \beta + (X^T X)^{-1} X^T Z\gamma - (X^T X)^{-1} X^T Z\hat{\gamma},
\end{aligned}$$

pričom sa využilo značenie (12). Pre strednú hodnotu odhadu b_u tak dostávame:

$$E(b_u) = E(E(b_u \mid M_X y)) = E(\beta + (X^T X)^{-1} X^T Z\gamma - (X^T X)^{-1} X^T Z\hat{\gamma}),$$

a teda:

$$E(b_u) = \beta. \quad (27)$$

Z rovnosti (27) nám zároveň priamo vyplýva $E(b_u - \beta) = 0$. Výraz (26) sa nám tým pádom zjednoduší do tvaru:

$$MSE(b_u) = Var(b_u). \quad (28)$$

Varianciu odhadu b_u vieme opäť v zmysle tvrdenia (1.11) vyjadriť nasledujúcim spôsobom:

$$Var(b_u) = E(Var(b_u | M_X y)) + Var(E(b_u | M_X y)).$$

Pre prvý zo sčítancov platí:

$$E(Var(b_u | M_X y)) = E(Var(b_r - Q\hat{\theta} | M_X y)) = E(Var(b_r | M_X y)) = E(Var(b_r)) = \sigma_\varepsilon^2 (X^T X)^{-1},$$

pričom sa opäť využilo, že hodnota $Q\hat{\theta}$ je determinovaná vektorom $M_X y$, hodnota b_r je nezávislá od $M_X y$ a vzťah (25). Analyzujeme druhý sčítanec:

$$Var(E(b_u | M_X y)) = Var(E(b_r | M_X y) + E(Q\hat{\theta} | M_X y)) = Var(E(b_r) + Q\hat{\theta}) = Var(Q\hat{\theta}).$$

Spojením oboch sčítancov dostávame:

$$Var(b_u) = E(Var(b_u | M_X y)) + Var(E(b_u | M_X y)) = \sigma_\varepsilon^2 (X^T X)^{-1} + Q Var(\hat{\theta}) Q^T. \quad (29)$$

Využitím značenia (12) a výsledku (10) dostávame:

$$\begin{aligned} Var(\hat{\theta}) &= Var((Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X y) = (Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X Var(y) M_X Z (Z^T M_X Z)^{-\frac{1}{2}} = \\ &= \sigma_\varepsilon^2 (Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X Z (Z^T M_X Z)^{-\frac{1}{2}} = \sigma_\varepsilon^2 I_m, \end{aligned}$$

pričom sa využila idempotentnosť matice M_X . Dosadením $Var(\hat{\theta}) = \sigma_\varepsilon^2 I_m$ do vzťahu (29) dostávame:

$$Var(b_u) = \sigma_\varepsilon^2 ((X^T X)^{-1} + Q Q^T). \quad (30)$$

Vďaka vzťahu (28) dostávame tiež:

$$MSE(b_u) = \sigma_\varepsilon^2 ((X^T X)^{-1} + Q Q^T). \quad (31)$$

Teraz môžeme provnať, aký je rozdiel strednej kvadratickej chyby odhadu b_u parametra β (teda ak by sme na odhad parametra β použili iba úplný model) a strednej kvadratickej chyby odhadu b parametra β (teda ak by sme na odhad parametra β použili aj modely, ktoré vznikli selekciou premenných). Vďaka tvrdeniu 2.2 a výsledku (31) dostávame:

$$MSE(b_u) - MSE(b) = \sigma_\varepsilon^2 ((X^T X)^{-1} + Q Q^T) - \sigma_\varepsilon^2 ((X^T X)^{-1} + Q MSE(W\hat{\eta}) Q^T),$$

a teda:

$$MSE(b_u) - MSE(b) = \sigma_\varepsilon^2 Q (I_m - MSE(W\hat{\eta})) Q^T. \quad (32)$$

2.2 Podreportovanie chyby testu v modeli získanom selekciou premenných

Teraz sa konečne budeme zaoberať otázkou, k akej miere podreportovania chyby testu dochádza, ak odignorujeme fakt, že model v ktorom odhadneme parameter β sme dostali selekciou. K dispozícii teda máme model

$$\mathcal{M} : y = X\beta + Z\gamma + \varepsilon,$$

z ktorého selekciou vyberieme niektorý zo submodelov

$$\mathcal{M}_i : y = X\beta + ZS_i\gamma_i + \varepsilon.$$

V našom prípade tak dostávame pre vyselektovaný model hodnotu $\lambda_i = 1$ a ostatné parametre λ sú nulové. Budeme tiež predpokladať, že to ktorú z hodnôt λ určíme ako 1 bude determinované hodnotou M_{Xy} (túto podmienku spĺňa napríklad selekcia pomocou Akaikého kritéria).

Z tvrdenia 2.2 máme, že pre skutočnú strednú kvadratickú chybu odhadu $b = b_i$ parametra β platí:

$$MSE(b) = \sigma_\varepsilon^2[(X^T X)^{-1} + QMSE(W\hat{\eta})Q^T].$$

Pokiaľ by sme ale selekciu neuvažili a model $\mathcal{M}_i$ by sme tak považovali za jediný prípustný, strednú kvadratickú chybu odhadu $b = b_i$ parametra β by sme vypočítali nasledujúcim spôsobom:

$$\widetilde{MSE}(b) = E((b_i - \beta)(b_i - \beta)^T) = Var(b_i) + E(b_i - \beta)[E(b_i) - \beta]^T,$$

pričom ak neuvažujeme selekciu a model $\mathcal{M}_i$ považujeme za jednoznačne správny, tak strednú hodnotu $E(b_i - \beta)$ považujeme za nulovú a dostávame:

$$\widetilde{MSE}(b) = Var(b_i),$$

čo vieme využitím (19) ďalej upraviť:

$$\widetilde{MSE}(b) = Var(b_r - QW\hat{\theta}).$$

Uvedomme si, že $W = \sum_{j=1}^{2^m} \lambda_j W_j = W_i$, a teda

$$\widetilde{MSE}(b) = Var(b_r - QW_i\hat{\theta}).$$

Matice Q a W_i sú nenáhodné a z tvrdenia 2.1 vieme, že vektory b_r a $\hat{\theta}$ sú nezávislé, čiže platí:

$$\widetilde{MSE}(b) = Var(b_r) + QW_i Var(\hat{\theta})W_i Q^T.$$

To vieme ďalej upraviť využitím značenia (12) a vyjadrenia (10):

$$\widetilde{MSE}(b) = Var(b_r) + QW_i Var((Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X y) W_i Q^T,$$

$$\widetilde{MSE}(b) = Var(b_r) + QW_i (Z^T M_X Z)^{-\frac{1}{2}} Z^T M_X Var(y) M_X Z (Z^T M_X Z)^{-\frac{1}{2}} W_i Q^T,$$

$$\widetilde{MSE}(b) = Var(b_r) + \sigma_\varepsilon^2 QW_i Q^T,$$

pričom sme využili $Var(y) = \sigma_\varepsilon^2 I_n$ a idempotentnosť matice W_i . Nakoniec využijeme:

$$Var(b_r) = Var((X^T X)^{-1} X^T y) = \sigma_\varepsilon^2 (X^T X)^{-1},$$

čím pre $\widetilde{MSE}(b)$ dostávame finálne vyjadrenie:

$$\widetilde{MSE}(b) = \sigma_\varepsilon^2 [(X^T X)^{-1} + QW_i Q^T].$$

Nech $\omega \in \mathbb{R}^k$ je ľubovoľný vektor. Chceli by sme porovnať odhady:

$$MSE(\omega^T b) = \sigma_\varepsilon^2 [\omega^T (X^T X)^{-1} \omega + \omega^T Q MSE(W \hat{\eta}) Q^T \omega],$$

$$\widetilde{MSE}(\omega^T b) = \sigma_\varepsilon^2 [\omega^T (X^T X)^{-1} \omega + \omega^T QW_i Q^T \omega].$$

Definujme premennú UR (underreporting ratio) ako:

$$UR = 1 - \frac{\widetilde{MSE}(\omega^T b)}{MSE(\omega^T b)}. \quad (33)$$

Takto definovaná premenná UR nám potom hovorí, že pokiaľ sa stredné kvadratické chyby $MSE(\omega^T b)$ a $\widetilde{MSE}(\omega^T b)$ veľmi nelíšia, hodnota premennej UR bude blízka 0 (nedochádza k podreportovaniu skutočnej chyby). Ak je naopak stredná kvadratická chyba $\widetilde{MSE}(\omega^T b)$ veľmi nízka oproti kvadratickej chybe $MSE(\omega^T b)$, hodnota premennej UR je blízka 1 (dochádza k veľkej miere podreportovania skutočnej chyby).

Zavedme značenie:

$$R := R(\hat{\eta}) = MSE(W \hat{\eta}) \quad q := \frac{Q^T \omega}{\sqrt{\omega^T Q Q^T \omega}} \quad q_0^2 := \frac{\omega^T Q Q^T \omega}{\omega^T (X^T X)^{-1} \omega}. \quad (34)$$

Pomocou značenia (34) vieme teraz hodnotu (33) upraviť nasledujúcim spôsobom:

$$\begin{aligned}
UR &= 1 - \frac{\widetilde{MSE}(\omega^T b)}{MSE(\omega^T b)} = 1 - \frac{\sigma_\varepsilon^2[\omega^T(X^T X)^{-1}\omega + \omega^T Q W_i Q^T \omega]}{\sigma_\varepsilon^2[\omega^T(X^T X)^{-1}\omega + \omega^T Q MSE(W\hat{\eta})Q^T \omega]} = \\
&= 1 - \frac{\omega^T[(X^T X)^{-1} + Q W Q^T]\omega}{\omega^T[(X^T X)^{-1} + Q R Q^T]\omega} = \frac{\omega^T Q(R - W)Q^T \omega}{\omega^T Q R Q^T \omega + \omega^T(X^T X)^{-1}\omega} = \\
&= \frac{\frac{\omega^T Q}{\sqrt{\omega^T Q Q^T \omega}}(R - W)\frac{Q^T \omega}{\sqrt{\omega^T Q Q^T \omega}}}{\frac{\omega^T Q}{\sqrt{\omega^T Q Q^T \omega}}R\frac{Q^T \omega}{\sqrt{\omega^T Q Q^T \omega}} + \frac{\omega^T(X^T X)^{-1}\omega}{\omega^T Q Q^T \omega}} = \frac{q^T(R - W)q}{q^T R q + \frac{1}{q_0^2}}.
\end{aligned}$$

V zmysle značenie (34) teda dostávame:

$$UR = \frac{q^T(R - W)q}{q^T R q + \frac{1}{q_0^2}}. \quad (35)$$

Bude nás ďalej zaujímať stredná hodnota premennej UR . Môžeme vyjadriť:

$$E(UR) = E\left(\frac{q^T(R - W)q}{q^T R q + \frac{1}{q_0^2}}\right) = q_0^2 E\left(\frac{q^T(R - W)q}{q_0^2 q^T R q + 1}\right).$$

Všimnime si, že vektor q je zvolený tak, že platí $q^T q = 1$, a teda môžeme napísať:

$$E(UR) = q_0^2 E\left(\frac{q^T(R - W)q}{q_0^2 q^T R q + q^T q}\right).$$

Podľa autorov článku [3] potom platí:

$$E(UR) = q_0^2 E\left(\frac{q^T(R - W)q}{q_0^2 q^T R q + q^T q}\right) = q_0^2 E(q^T(q_0^2 R + I_m)^{-\frac{1}{2}}(R - W)(q_0^2 R + I_m)^{-\frac{1}{2}}q).$$

Keďže náhodnosť sa vyskytuje iba v matici W , ďalej dostávame:

$$E(UR) = q_0^2 q^T (q_0^2 R + I_m)^{-\frac{1}{2}} (R - E(W)) (q_0^2 R + I_m)^{-\frac{1}{2}} q.$$

Nášim ďalším cieľom bude nájsť nejaké horné ohraňenie pre $E(UR)$. Predpokladajme, že $q_0^2 \rightarrow \infty$. Označme:

$$E^*(UR) = \max_q q_0^2 q^T (q_0^2 R + I_m)^{-\frac{1}{2}} (R - E(W)) (q_0^2 R + I_m)^{-\frac{1}{2}} q.$$

To vieme ďalej upraviť nasledujúcim spôsobom:

$$E^*(UR) = \max_q q_0^2 \frac{1}{q_0^2} q^T (R + \frac{1}{q_0^2} I_m)^{-\frac{1}{2}} (R - E(W)) (R + \frac{1}{q_0^2} I_m)^{-\frac{1}{2}} q.$$

Keďže $q_0^2 \rightarrow \infty$, tak $\frac{1}{q_0^2} \rightarrow 0$ a dostávame:

$$E^*(UR) = \max_q q^T (R^{-\frac{1}{2}} (R - E(W)) R^{-\frac{1}{2}}) q.$$

Všimnime si, že matica $R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}}$ je symetrická, keďže matice $R = MSE(W\hat{\eta})$ a $E(W)$ sú symetrické. Potom existuje spektrálny rozklad $U\Lambda U^T$ tejto matice, pričom matica U je ortogonálna. Využitím tohto spektrálneho rozkladu dostávame:

$$E^*(UR) = \max_q q^T (R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}})q = \max_q q^T U\Lambda U^T q.$$

Všimnime si, že platí $q^T U U^T q = q^T q = 1$ a matica Λ je diagonálna matica, pričom jej diagonála obsahuje vlastné hodnoty matice $R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}}$. Súčin $q^T U\Lambda U^T q$ potom nie je nič iné, ako vážený priemer vlastných hodnôt matice $R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}}$. Vážený priemer vlastných hodnôt potom isto nie je väčší ako najväčšia vlastná hodnota, a teda platí:

$$E^*(UR) = \max_q q^T (R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}})q \leq \max_{1 \leq j \leq m} \xi_j(R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}}),$$

pričom $\xi_j(R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}})$ je j -ta vlastná hodnota matice $R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}}$. To vieme ďalej upraviť do tvaru:

$$\begin{aligned} E^*(UR) &\leq \max_{1 \leq j \leq m} \xi_j(R^{-\frac{1}{2}}(R - E(W))R^{-\frac{1}{2}}) = \max_{1 \leq j \leq m} \xi_j(I_m - R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}}) = \\ &= 1 - \min_{1 \leq j \leq m} \xi_j(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}}), \\ E^*(UR) &\leq 1 - \min_{1 \leq j \leq m} \xi_j(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}}). \end{aligned} \quad (36)$$

Pre strednú hodnotu $E(W)$ platí $E(W) = E(\sum_{i=1}^{2^m} \lambda_i W_i)$, a keďže matice W_i sú nenáhodné, tak $E(W) = \sum_{i=1}^{2^m} E(\lambda_i) W_i$. Pre maticu $R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}}$ tým pádom dostávame:

$$R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}} = \sum_{i=1}^{2^m} E(\lambda_i) R^{-\frac{1}{2}} W_i R^{-\frac{1}{2}}.$$

Skúmame maticový rozdiel $R^{-\frac{1}{2}}W_i R^{-\frac{1}{2}} - R^{-1}$. Platí:

$$R^{-\frac{1}{2}}W_i R^{-\frac{1}{2}} - R^{-1} = R^{-\frac{1}{2}}(W_i - I_m)R^{-\frac{1}{2}}.$$

Uvedomme si, že matica W_i je idempotentná, a teda jej vlastné hodnoty sú iba nuly a jednotky, z čoho vyplýva, že matica $W_i - I_m$ je záporne semidefinitná. Potom pre ľubovoľný vektor $a \in \mathbb{R}^m$ platí $a^T R^{-\frac{1}{2}}(W_i - I_m)R^{-\frac{1}{2}}a \leq 0$, z čoho vyplýva, že aj matica $R^{-\frac{1}{2}}W_i R^{-\frac{1}{2}} - R^{-1}$ je záporne semidefinitná. Potom pre vlastné hodnoty matíc $R^{-\frac{1}{2}}W_i R^{-\frac{1}{2}}$ a R^{-1} platí nerovnosť:

$$\xi_j(R^{-\frac{1}{2}}W_i R^{-\frac{1}{2}}) \leq \xi_j(R^{-1}).$$

Potom s využitím nezápornosti hodnôt λ_i platia aj nerovnosti:

$$\xi_j(E(\lambda_i)R^{-\frac{1}{2}}W_iR^{-\frac{1}{2}}) \leq \xi_j(E(\lambda_i)R^{-1}),$$

ktoré môžeme sčítať cez všetky hodnoty i a dostávame:

$$\begin{aligned} \xi_j \left(\sum_{i=1}^{2^m} E(\lambda_i)R^{-\frac{1}{2}}W_iR^{-\frac{1}{2}} \right) &\leq \xi_j \left(\sum_{i=1}^{2^m} E(\lambda_i)R^{-1} \right), \\ \xi_j \left(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}} \right) &\leq \xi_j \left(R^{-1} \sum_{i=1}^{2^m} E(\lambda_i) \right) = \xi_j(R^{-1}) \sum_{i=1}^{2^m} E(\lambda_i), \end{aligned}$$

a teda dostávame vzťah:

$$\xi_j \left(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}} \right) \leq \xi_j(R^{-1}) \sum_{i=1}^{2^m} E(\lambda_i).$$

Uvedomme si, že $\lambda_i = 1$ v prípade, že selekciou vyberieme i -ty model a $\lambda_i = 0$ inak. Potom $E(\lambda_i) = p_i$, pričom hodnotou p_i značíme pravdepodobnosť, že selekciou vyberieme i -ty model. Súčet $\sum_{i=1}^{2^m} E(\lambda_i) = \sum_{i=1}^{2^m} p_i$ je potom z vety o úplnej pravdepodobnosti rovný 1, z čoho vyplýva:

$$\begin{aligned} \xi_j \left(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}} \right) &\leq \xi_j(R^{-1}), \\ \min_{1 \leq j \leq m} \xi_j \left(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}} \right) &\leq \min_{1 \leq j \leq m} \xi_j(R^{-1}) = \frac{1}{\max_j \xi_j(R)}. \end{aligned} \quad (37)$$

Využitím vzťahu (37) v nerovnosti (36) nakoniec dostávame horné ohraničenie pre hodnotu $E(UR)$:

$$\begin{aligned} E(UR) \leq E^*(UR) &\leq 1 - \min_{1 \leq j \leq m} \xi_j(R^{-\frac{1}{2}}E(W)R^{-\frac{1}{2}}) \leq 1 - \frac{1}{\max_j \xi_j(R)}, \\ E(UR) &\leq 1 - \frac{1}{\max_j \xi_j(R)}. \end{aligned}$$

3 Selekcia nulovej premennej

V tejto kapitole sa budeme zaoberať situáciou, keď máme na výber z dvoch lineárnych regresných modelov, pričom jeden z modelov bude obsahovať parametre, ktoré sú v skutočnosti nulové (ide teda o rušivé parametre, ktoré sú v modeli navyše). Bude nás teda zaujímať pravdepodobnosť, že sa nám tieto rušivé parametre podarí odstrániť buď prostredníctvom selekcie, alebo prostredníctvom testovania hypotézy o ich signifikantnosti. Situáciu môžeme vyjadriť nasledovne:

- $\mathcal{M}_1 : Y = X_1\beta_1 + X_2\beta_2 + \varepsilon_1$
- $\mathcal{M}_2 : Y = X_1\beta_1 + \varepsilon_2$

Vektor Y je v tomto prípade vektor n vysvetľovaných premenných, X_1 a X_2 sú matice rozmerov $n \times k_1$ a $n \times k_2$, ktorých stĺpce tvoria vektory vysvetľujúcich premenných, β_1 je vektor k_1 parametrov, ktoré sú pre oba modely spoločné, β_2 je vektor k_2 parametrov zahrnutých iba v modeli $\mathcal{M}_1$ a $\varepsilon_1, \varepsilon_2$ sú vektory chýb, pre ktoré platí $\varepsilon_1, \varepsilon_2 \sim N(0, \sigma_\varepsilon^2 I_n)$. V tomto prípade budeme tiež predpokladať, že vektor parametrov β_2 je v skutočnosti nulový. Z danej podmienky potom automaticky vyplýva: $\varepsilon_1 = \varepsilon_2$ (keďže modely $\mathcal{M}_1$ a $\mathcal{M}_2$ sú v teoreticky rovnaké, aj keď v praxi dávajú iné odhady).

Pre väčšiu úspornosť našich zápisov budeme používať maticu Z_1 ako blokovú maticu $[X_1 \ X_2]$ a maticu $Z_2 = X_1$. Tiež zavedieme vektory $\theta_1 = (\beta_1 \ \beta_2)^T$ a $\theta_2 = \beta_1$. Modely $\mathcal{M}_1$ a $\mathcal{M}_2$ môžeme potom pomocou tohto značenia kompaktne zapísať nasledovným spôsobom:

$$\mathcal{M}_1 : Y = Z_1\theta_1 + \varepsilon$$

$$\mathcal{M}_2 : Y = Z_2\theta_2 + \varepsilon$$

Budeme sa konkrétne zaoberať otázkou: “Aká je pravdepodobnosť, že o parametri β_2 rozhodneme, že je nulový?” Takéto rozhodnutie môžeme urobiť dvoma rôznymi spôsobmi. Buď si podľa Akaikeho kritéria (*AIC*) vyberieme model $\mathcal{M}_2$, ako model vysvetľujúci dáta Y (keď pomocou *AIC* rozhodneme, že parameter β_2 možno vyradiť), alebo si síce vyberieme $\mathcal{M}_1$, ale následne nezamietneme hypotézu (na hladine α) o nulovosti parametra β_2 . Pravdepodobnosť, že o parametri β_2 rozhodneme že je nulový môžeme potom schematicky

napísať nasledujúcim spôsobom:

$$P(\text{nez.}H_0 : \beta_2 = 0) = P(AIC_2 < AIC_1) + P(\text{nez.}H_0 : \beta_2 = 0 \wedge AIC_1 < AIC_2).$$

V nasledujúcej podkapitole sa teda budeme zaoberať pomocnou otázkou: “Aká je pravdepodobnosť, že ako model vysvetľujúci dáta Y si podľa AIC vyberieme model $\mathcal{M}_2$ oproti $\mathcal{M}_1$?”

3.1 Výber modelu na základe Akaikeho kritéria

Hoci sme v úvode kapitoly zaviedli predpoklad, že parameter $\beta_2 = 0$, bude pre nás výhodnejšie keď naše úvahy viac zovšeobecníme, a teda od tohto predpokladu na chvíľu upustíme (znova ho využijeme keď ho budeme potrebovať). Chceme sa totiž vyhnúť potrebe robiť veľmi podobné úvahy na dvakrát.

Pomocou AIC vyberieme model $\mathcal{M}_2$ ako model vysvetľujúci dáta Y práve vtedy, keď AIC modelu $\mathcal{M}_2$ (ozn. AIC_2) bude menšie ako AIC modelu $\mathcal{M}_1$ (ozn. AIC_1). Pravdepodobnosť, že pomocou AIC uprednostníme $\mathcal{M}_2$ oproti $\mathcal{M}_1$, potom môžeme vyjadriť nasledovne:

$$\begin{aligned} P(AIC_2 < AIC_1) &= \\ &= P\left(\ln(Z_2\hat{\theta}_2 - Y)^T(Z_2\hat{\theta}_2 - Y) - \ln n + \frac{2k_2}{n} < \ln(Z_1\hat{\theta}_1 - Y)^T(Z_1\hat{\theta}_1 - Y) - \ln n + \frac{2k_1}{n}\right) = \\ &= P\left(\ln \frac{(Z_2\hat{\theta}_2 - Y)^T(Z_2\hat{\theta}_2 - Y)}{(Z_1\hat{\theta}_1 - Y)^T(Z_1\hat{\theta}_1 - Y)} < \frac{2(k_1 - k_2)}{n}\right) \end{aligned} \quad (38)$$

Poznámka: Za predpokladu nulovej hypotézy ($H_0 : \beta_2 = 0$) sa dá ukázať, že čitateľ aj menovateľ zlomku vystupujúceho ako argument logaritmu majú χ^2 rozdelenie. Naším cieľom preto bude daný zlomok vyjadriť pomocou náhodnej premennej, ktorá sa riadi Fisher-Snedecorovým rozdelením. Toto by sme vedeli urobiť priamo, keby dané χ^2 - rozdelenia boli nezávislé, čo však neplatí. S touto závislosťou sa preto budeme musieť nejakým spôsobom vysporiadať.

Pred ďalším pokračovaním zavedme značenie:

$$H_1 = Z_1(Z_1^T Z_1)^{-1} Z_1^T, \quad H_2 = Z_2(Z_2^T Z_2)^{-1} Z_2^T, \quad M_1 = I - H_1, \quad M_2 = I - H_2,$$

pričom budeme predpokladať, že matice Z_1 a Z_2 majú plnú hodnotu. Teraz môžeme vyjadriť závislosť náhodných premenných vyskytujúcich sa v (38) pomocou kovariancie:

$$\begin{aligned} \text{cov}(Z_2 \hat{\theta}_2 - Y, Z_1 \hat{\theta}_1 - Y) &= \text{cov}(Z_2(Z_2^T Z_2)^{-1} Z_2^T Y - Y, Z_1(Z_1^T Z_1)^{-1} Z_1^T Y - Y) = \\ &= \text{cov}(H_2 Y - Y, H_1 Y - Y) = \text{cov}((H_2 - I_n)Y, (H_1 - I_n)Y) = (H_2 - I_n) \text{cov}(Y, Y) (H_1 - I_n) = \\ &= (H_2 - I_n) \text{cov}(Z_1 \theta_1 + \varepsilon, Z_1 \theta_1 + \varepsilon) (H_1 - I_n) = (H_2 - I_n) \text{cov}(\varepsilon, \varepsilon) (H_1 - I_n) = \\ &= M_2 \sigma_\varepsilon^2 I_n M_1 = \sigma_\varepsilon^2 M_2 M_1. \end{aligned}$$

Z tvrdenia 1.6 vyplýva, že $M_2 M_1 = M_1$, vďaka čomu dostávame:

$$\text{cov}(Z_2 \hat{\theta}_2 - Y, Z_1 \hat{\theta}_1 - Y) = \sigma_\varepsilon^2 M_1$$

O vektoroch $Z_2 \hat{\theta}_2 - Y$ a $Z_1 \hat{\theta}_1 - Y$ vieme taktiež povedať, že majú normálne rozdelenie. Vektory Y , $\hat{\theta}_1$ a $\hat{\theta}_2$ sú totiž lineárne kombinácie vektora Y , čo implikuje, že aj ich lineárne kombinácie sú lineárnou kombináciou vektora Y (čo je normálny vektor).

Aby sme mali kompletnú informáciu o vektoroch $Z_2 \hat{\theta}_2 - Y$ a $Z_1 \hat{\theta}_1 - Y$, ostáva nám už iba určiť ich strednú hodnotu a varianciu:

$$E(Z_1 \hat{\theta}_1 - Y) = Z_1 E(\hat{\theta}_1) - E(Y) = Z_1 \theta_1 - Z_1 \theta_1 = 0_n$$

$$E(Z_2 \hat{\theta}_2 - Y) = E(H_2 Y - Y) = (H_2 - I) E(Y) = -M_2 Z_1 \theta_1$$

$$\text{Var}(Z_1 \hat{\theta}_1 - Y) = \text{Var}(H_1 Y - Y) = \text{Var}(-M_1 Y) = M_1 \text{Var}(Y) M_1 = \sigma_\varepsilon^2 M_1$$

$$\text{Var}(Z_2\hat{\theta}_2 - Y) = \text{Var}(H_2Y - Y) = \text{Var}(-M_2Y) = M_2\text{Var}(Y)M_2 = \sigma_\varepsilon^2 M_2$$

Zavedme značenie:

$$P_1 := Z_1\hat{\theta}_1 - Y, \quad P_2 := Z_2\hat{\theta}_2 - Y. \quad (39)$$

Potom pre náhodný vektor $(P_1 \ P_2)^T$ dostávame, že sa riadi nasledujúcim rozdelením:

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} \sim N \left(\begin{pmatrix} 0_n \\ -M_2Z_1\theta_1 \end{pmatrix}, \begin{bmatrix} \sigma_\varepsilon^2 M_1 & \sigma_\varepsilon^2 M_1 \\ \sigma_\varepsilon^2 M_1 & \sigma_\varepsilon^2 M_2 \end{bmatrix} \right). \quad (40)$$

Potom náhodný vektor P_2 môžeme jednoducho definovať pomocou náhodného vektora P_1 a nezávislého náhodného vektora $N \sim N(0_n, I_n)$ nasledujúcim spôsobom:

$$P_2 := P_1 - M_2Z_1\theta_1 + \sigma_\varepsilon(M_2 - M_1)N. \quad (41)$$

Dá sa totiž ukázať, že práve pre takto zvolený vektor P_2 sa náhodný vektor $(P_1 \ P_2)^T$ riadi rozdelením (40) (stačí jednoducho overiť, že sa zachováva stredná hodnota aj variančná matica).

Podme sa teda ďalej zaoberať skalárnym súčinom $P_2^T P_2$:

$$P_2^T P_2 = (P_1 - M_2Z_1\theta_1 + \sigma_\varepsilon(M_2 - M_1)N)^T (P_1 - M_2Z_1\theta_1 + \sigma_\varepsilon(M_2 - M_1)N) =$$

$$= P_1^T P_1 + \theta_1^T Z_1^T M_2 Z_1 \theta_1 + \sigma_\varepsilon^2 N^T (M_2 - M_1) N - 2P_1^T M_2 Z_1 \theta_1 +$$

$$+ 2\sigma_\varepsilon P_1^T (M_2 - M_1) N - 2\sigma_\varepsilon \theta_1^T Z_1^T M_2 (M_2 - M_1) N,$$

pričom sme opäť využili tvrdenie 1.3 a tvrdenie 1.6, z ktorých priamo vyplýva $(M_2 - M_1)^2 = (M_2 - M_1)$.

Podme teraz bližšie preskúmať výraz $2\sigma_\varepsilon P_1^T (M_2 - M_1) N$. Spätným preznačením P_1 podľa (39) dostávame:

$$2\sigma_\varepsilon P_1^T (M_2 - M_1) N = 2\sigma_\varepsilon (Z_1\hat{\theta}_1 - Y)^T (M_2 - M_1) N = 2\sigma_\varepsilon Y^T (Z_1(Z_1^T Z_1)^{-1} Z_1^T - I) (M_2 - M_1) N =$$

$$= 2\sigma_\varepsilon Y^T (H_1 - I)(M_2 - M_1)N = -2\sigma_\varepsilon Y^T M_1(M_2 - M_1)N = 0,$$

pričom sme opäť využili tvrdenie 1.6. Skúmame ďalej výraz $2P_1^T M_2 Z_1 \theta_1$. Dostávame:

$$\begin{aligned} 2P_1^T M_2 Z_1 \theta_1 &= 2Y^T (Z_1 (Z_1^T Z_1)^{-1} Z_1^T - I) M_2 Z_1 \theta_1 = 2Y^T (H_1 - I) M_2 Z_1 \theta_1 = -2Y^T M_1 M_2 Z_1 \theta_1 = \\ &= -2Y^T M_1 Z_1 \theta_1 = -2Y^T (I - H_1) Z_1 \theta_1 = -2Y^T Z_1 \theta_1 + 2Y^T Z_1 \theta_1 = 0. \end{aligned}$$

Nakoniec využijeme $M_2(M_2 - M_1) = M_2 - M_1$ a teda člen $-2\sigma_\varepsilon \theta_1^T Z_1^T M_2(M_2 - M_1)N$ môžeme upraviť na výraz $-2\sigma_\varepsilon \theta_1^T Z_1^T (M_2 - M_1)N$. Pre skalárny súčin $P_2^T P_2$ potom dostávame nasledujúce vyjadrenie:

$$P_2^T P_2 = P_1^T P_1 + \theta_1^T Z_1^T M_2 Z_1 \theta_1 + \sigma_\varepsilon^2 N^T (M_2 - M_1)N - 2\sigma_\varepsilon \theta_1^T Z_1^T (M_2 - M_1)N. \quad (42)$$

Aby sme sa vedeli lepšie vysporiadať s touto rovnosťou, vyslovíme a dokážeme nasledujúce tvrdenie:

Tvrdenie 3.1. *Existujú spektrálne rozklady matíc $M_1 = U_1 \Lambda_1 U_1^T$ a $M_2 = U_2 \Lambda_2 U_2^T$ také, že platí:*

1. *Ak je nejaká zložka na diagonále matice Λ_1 rovná 1, tak rovnaká zložka na diagonále matice Λ_2 je rovná 1*
2. *Matice vlastných vektorov sa rovnajú: $U_1 = U_2$*

Dôkaz. Keďže M_1 a M_2 sú maticami projekcie (tvrdenie 1.3), ich vlastné hodnoty sú buď rovné 1, alebo 0 (tvrdenie 1.4). Predpokladajme, že vektor u je vlastným vektorom matice M_1 pre vl. hodnotu 1. Potom $M_1 u = u$, čo po prenasobení maticou M_2 zľava upravíme na tvar $M_2 M_1 u = M_2 u$. Vieme, že $M_2 M_1 = M_1$ (tvrdenie 1.6), a teda máme $M_1 u = M_2 u$. Zároveň ale $M_1 u = u$, takže dostávame $M_2 u = M_1 u = u$, čiže u je zároveň vlastným vektorom matice M_2 pre vl. hodnotu 1. Tým sme dokázali prvú časť tvrdenia.

Ukázali sme, že ak je u vlastný vektor matice M_1 pre vlastnú hodnotu 1, tak je zároveň vlastným vektorom matice M_2 pre vlastnú hodnotu 1.

Taktiež platí, že ak je u vlastný vektor matice M_2 pre vlastnú hodnotu 0, tak je zároveň vlastným vektorom matice M_1 pre vlastnú hodnotu 0 (stačí prenasobiť rovnicu $M_2 u = 0$ maticou M_1 zľava a využiť tvrdenie 1.6).

Nech v je ľubovoľný vektor pre ktorý platí $v \in C(Z_2)$ (značenie $C(Z_2)$ vyjadruje stĺpcový priestor (column space) matice Z_2). Ako sa už ukázalo, z uvedeného vyplýva,

že v je vlastný vektor matice M_2 (pre vlastnú hodnotu 0). Keďže matica U_2 obsahuje ortonormálnu bázu všetkých vlastných vektorov, tak zjavne niektoré z týchto vlastných vektorov musia tvoriť ortonormálnu bázu priestoru $C(Z_2)$.

Zjavne platí, že $C(Z_2) \subset C(Z_1)$. Preto ak $v \in C(Z_2)$, tak $v \in C(Z_1)$, a teda v je vlastný vektor matice M_1 pre vlastnú hodnotu 0. Potom matica U_1 musí obsahovať ortonormálnu bázu priestoru $C(Z_2)$.

Ukázali sme, že matica U_1 , aj matica U_2 obsahujú ortonormálnu bázu priestoru $C(Z_2)$. Bez ujmy na všeobecnosti potom môžeme zvoliť matice U_1 a U_2 tak, aby išlo o rovnakú bázu.

Teraz nech u je ľubovoľný stĺpec matice U_1 . Potom ak $u \in C(Z_2)$, tak už vieme, že u je zároveň stĺpcom matice U_2 . Predpokladajme teda, že $u \notin C(Z_2)$. To ale znamená, že u je kolmý na všetky bázoové vektory priestoru $C(Z_2)$, keďže stĺpce matice U_1 sú na seba kolmé, a teda u je z ortogonálneho doplnku priestoru $C(Z_2)$. Z poznatkov z lineárnej algebry vieme, že ortogonálnym doplnkom stĺpcového priestoru je ľavý nulový priestor, a teda $u \in N(Z_2^T)$. Potom platí:

$$M_2 u = (I - Z_2(Z_2^T Z_2)^{-1} Z_2^T) u = u - 0 = u,$$

čiže u je zároveň vlastným vektorom matice M_2 (pre vlastnú hodnotu 1), a teda ho môžeme vložiť aj do matice U_2 . □

S využitím tohto tvrdenia teraz pre vyjadrenie rozdielu $M_2 - M_1$ dostávame:

$$M_2 - M_1 = U_1 \Lambda_2 U_1^T - U_1 \Lambda_1 U_1^T = U_1 (\Lambda_2 - \Lambda_1) U_1^T.$$

Teraz si stačí uvedomiť, že matica $(\Lambda_2 - \Lambda_1)$ je tiež diagonálnou maticou, ktorej diagonála je tvorená nulami a jednotkami. Označme maticu vlastných vektorov matice U_1 zodpovedajúcich jednotkám na diagonále matice $(\Lambda_2 - \Lambda_1)$ ako V . Keďže vieme, že matica Λ_1 má hodnotu $n - k_1$ a matica Λ_2 má hodnotu $n - k_2$, matica $(\Lambda_2 - \Lambda_1)$ bude mať (s využitím tvrdenia 3.1) hodnotu $(n - k_2) - (n - k_1) = k_1 - k_2$. Matica V je teda tvorená $k_1 - k_2$ navzájom kolmými n -rozmernými stĺpcami veľkosti 1. Pre rozdiel $M_2 - M_1$ tak dostávame $M_2 - M_1 = V V^T$. Dosadením do (42) potom dostávame:

$$P_2^T P_2 = P_1^T P_1 + \theta_1^T Z_1^T M_2 Z_1 \theta_1 + \sigma_\varepsilon^2 N^T V V^T N - 2\sigma_\varepsilon \theta_1^T Z_1^T V V^T N.$$

S využitím vzťahu $M_1 Z_1 = (I - H_1) Z_1 = Z_1 - Z_1 = 0$ môžeme danú rovnosť prepísať do tvaru:

$$P_2^T P_2 = P_1^T P_1 + \theta_1^T Z_1^T (M_2 - M_1) Z_1 \theta_1 + \sigma_\varepsilon^2 N^T V V^T N - 2\sigma_\varepsilon \theta_1^T Z_1^T V V^T N,$$

pričom po dosadení $M_2 - M_1 = V V^T$ dostávame:

$$P_2^T P_2 = P_1^T P_1 + \theta_1^T Z_1^T V V^T Z_1 \theta_1 + \sigma_\varepsilon^2 N^T V V^T N - 2\sigma_\varepsilon \theta_1^T Z_1^T V V^T N,$$

čo vieme upraviť na tvar:

$$P_2^T P_2 = P_1^T P_1 + \sigma_\varepsilon^2 \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)^T \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right). \quad (43)$$

Teraz už máme všetko potrebné na to, aby sme vedeli vyjadriť pravdepodobnosť $P(AIC_2 < AIC_1)$. Z rovnosti (38) totiž máme:

$$P(AIC_2 < AIC_1) = P \left(\ln \frac{(Z_2 \hat{\theta}_2 - Y)^T (Z_2 \hat{\theta}_2 - Y)}{(Z_1 \hat{\theta}_1 - Y)^T (Z_1 \hat{\theta}_1 - Y)} < \frac{2(k_1 - k_2)}{n} \right),$$

čo vieme s využitím značenia (39) prepísať na tvar:

$$P(AIC_2 < AIC_1) = P \left(\ln \frac{P_2^T P_2}{P_1^T P_1} < \frac{2(k_1 - k_2)}{n} \right).$$

Použitím rovnosti (43) ďalej dostávame:

$$P(AIC_2 < AIC_1) = P \left(\ln \frac{P_1^T P_1 + \sigma_\varepsilon^2 \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)^T \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)}{P_1^T P_1} < \frac{2(k_1 - k_2)}{n} \right),$$

čo vieme upraviť do tvaru:

$$P(AIC_2 < AIC_1) = P \left(\frac{\sigma_\varepsilon^2 \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)^T \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)}{P_1^T P_1} < \exp \left(\frac{2(k_1 - k_2)}{n} \right) - 1 \right)$$

Všimnime si, že jediné náhodné vektory, ktoré v danej pravdepodobnosti vystupujú sú N a P_1 , ktoré sú navyše nezávislé. Z tvrdenia 1.8 vyplýva, že ak $N \sim N(0_n, I_n)$, tak $V^T N \sim N(0_{(k_1-k_2)}, I_{(k_1-k_2)})$. Môžeme si teda všimnúť, že výraz $\left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)^T \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)$ má v skutočnosti necentrálne χ^2 rozdelenie so stupňom voľnosti $k_1 - k_2$ a necentrálnym parametrom

$\frac{1}{\sigma_\varepsilon^2} \theta_1^T Z_1^T V V^T Z_1 \theta_1 = \frac{1}{\sigma_\varepsilon^2} \theta_1^T Z_1^T (M_2 - M_1) Z_1 \theta_1$. Zároveň vieme, že $P_1^T P_1 = (Z_1 \hat{\theta}_1 - Y)^T (Z_1 \hat{\theta}_1 - Y) = Y^T M_1 Y$, a teda $P_1^T P_1 \sim \sigma_\varepsilon^2 \chi_{n-k_1}^2$ (tvrdenie 1.9). Výraz

$$\frac{n - k_1}{k_1 - k_2} \frac{\sigma_\varepsilon^2 \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)^T \left(V^T N + \frac{-V^T Z_1 \theta_1}{\sigma_\varepsilon} \right)}{P_1^T P_1}$$

má teda necentrálne Fisher-Snedecorovo rozdelenie s necentrálnymi parametrami

$\frac{1}{\sigma_\varepsilon^2} \theta_1^T Z_1^T (M_2 - M_1) Z_1 \theta_1$ a 0, a stupňami voľnosti $k_1 - k_2$ a $n - k_1$. Z toho potom už priamo vyplýva, že pravdepodobnosť $P(AIC_2 < AIC_1)$ vieme vyjadriť ako distribučnú funkciu tohto rozdelenia v bode $\frac{n-k_1}{k_1-k_2} \left(\exp \left(\frac{2(k_1-k_2)}{n} \right) - 1 \right)$.

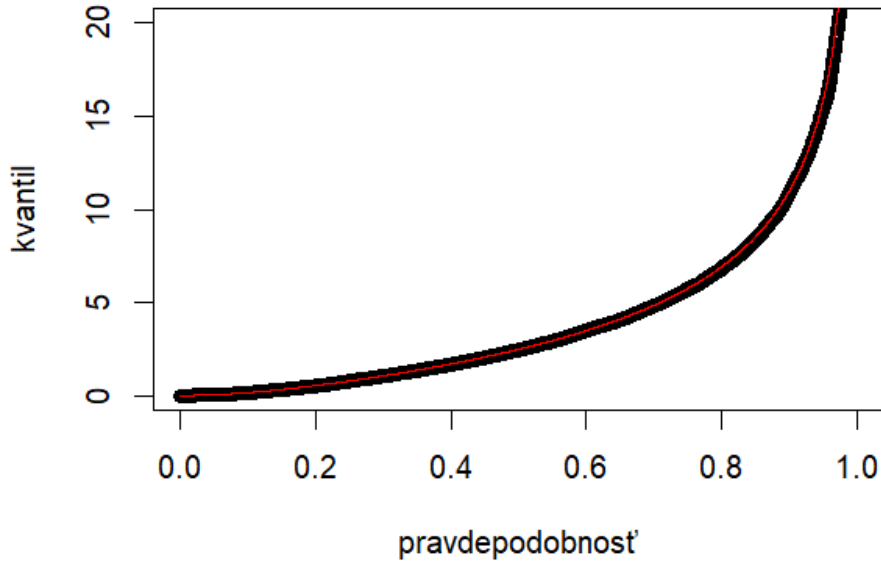
Všimnime si, že kľúčom v odvodení pravdepodobnosti $P(AIC_2 < AIC_1)$ bolo zistenie, že náhodná premenná $F := \left(\frac{P_2^T P_2}{P_1^T P_1} - 1 \right) \frac{n-k_1}{k_1-k_2}$ sa riadi necentrálnym Fisher-Snedecorovým rozdelením so stupňami voľnosti $k_1 - k_2$, $n - k_1$ a necentrálnym parametrom $\frac{1}{\sigma_\varepsilon^2} \theta_1^T Z_1^T (M_2 - M_1) Z_1 \theta_1$. Podiel $\frac{P_2^T P_2}{P_1^T P_1}$ nie je pritom nič iné, ako podiel súčtov štvorcov rezíduí v jednotlivých modeloch, teda $\frac{RSS_2}{RSS_1}$. Potom zistenie, že sa náhodná premenná F naozaj riadi týmto rozdelením vieme jednoduchým spôsobom overiť aj empiricky, pomocou simulácií.

Overenie sme urobili v softvéri R. Zostavili sme si lineárny regresný model s desiatimi pozorovaniami a troma parametrami $\beta_0, \beta_1, \beta_2$, nastavenými na hodnoty 5, 10 a 2. Zvolili sme stĺpce matice X tak, aby boli navzájom lineárne nezávislé a varianciu σ_ε^2 sme nastavili na hodnotu 100. Následne sme urobili 10000 simulácií, pričom v každej sme spočítali RSS_1 pre úplný model a RSS_2 pre model neobsahujúci parameter β_2 . V každej simulácii sme spočítali testovú štatistiku F a nakoniec sme dané testové štatistiky usporiadali a nechali vykresliť, čím sme dostali empirickú kvantilovú funkciu. V rovnakom grafe sme následne vykreslili reálnu kvantilovú funkciu Fisher-Snedecorovho rozdelenia so stupňami voľnosti $k_1 - k_2$, $n - k_1$ a necentrálnym parametrom $\frac{1}{\sigma_\varepsilon^2} \theta_1^T Z_1^T (M_2 - M_1) Z_1 \theta_1$. Výsledok uvádzame na obrázku 1.

Keďže vidíme, že empirická kvantilová funkcia (čiernou farbou vykreslené dáta) sa zhoduje s nami odvodenou teoretickou kvantilovou funkciou (červená krivka), dá sa očakávať, že teoretické výsledky odvodené v tejto podkapitole sú správne.

3.2 Pravdepodobnosť selekcie nulovej premennej

V tejto podkapitole sa vrátime k otázke: “Aká je pravdepodobnosť, že o parametri β_2 rozhodneme, že je naozaj nulový?”. Zopakujme, že naše “rozhodnutie” spočíva v tom, že



Obr. 1: Empirický kvantil testovej štatistiky F

najprv si podľa AIC vyberieme model a pokiaľ vyberieme model $\mathcal{M}_1$, tak následne ešte otestujeme hypotézu o signifikantnosti parametra β_2 . Matematicky teda túto pravdepodobnosť vieme vyjadriť ako $P(AIC_2 < AIC_1) + P(nez.H_0 : \beta_2 = 0 \wedge AIC_1 < AIC_2)$. Hodnotu pravdepodobnosti $P(AIC_2 < AIC_1)$ už poznáme. Poďme sa teda bližšie zaoberať pravdepodobnosťou $P(nez.H_0 : \beta_2 = 0 \wedge AIC_1 < AIC_2)$.

Z tvrdenia 1.7 dostávame za predpokladu nulovosti parametra β_2 :

$$T := \frac{n - k_1 - k_2}{k_2} \frac{(Z_2 \hat{\theta}_2 - Y)^T (Z_2 \hat{\theta}_2 - Y) - (Z_1 \hat{\theta}_1 - Y)^T (Z_1 \hat{\theta}_1 - Y)}{(Z_1 \hat{\theta}_1 - Y)^T (Z_1 \hat{\theta}_1 - Y)} \sim F_{k_2, n - k_1 - k_2},$$

respektíve

$$T := \frac{n - k_1 - k_2}{k_2} \frac{RSS_2 - RSS_1}{RSS_1} \sim F_{k_2, n - k_1 - k_2}, \quad (44)$$

kde $RSS_2 := (Z_2 \hat{\theta}_2 - Y)^T (Z_2 \hat{\theta}_2 - Y)$ a $RSS_1 := (Z_1 \hat{\theta}_1 - Y)^T (Z_1 \hat{\theta}_1 - Y)$. Hypotézu H_0 teda zamietame na hladine α , ak $T > qF(1 - \alpha, k_2, n - k_1 - k_2)$

(kde pravá strana značí $(1 - \alpha)$ -kvantil Fisher-Snedecorovho rozdelenia s parametrami k_2 a $n - k_1 - k_2$).

Nerovnica $AIC_2 > AIC_1$ je ekvivalentná s nerovnicou $\ln \frac{RSS_2}{RSS_1} > \frac{2k_1}{n}$, čo vieme upraviť na tvar:

$$T := \frac{n - k_1 - k_2}{k_2} \frac{RSS_2 - RSS_1}{RSS_1} > \frac{n - k_1 - k_2}{k_2} \left(\exp \left(\frac{2k_1}{n} \right) - 1 \right).$$

Pre pravdepodobnosť $P(\text{nez}.H_0 : \beta_2 = 0 \wedge AIC_1 < AIC_2)$ teda dostávame:

$$\begin{aligned} & P(\text{nez}.H_0 : \beta_2 = 0 \wedge AIC_1 < AIC_2) = \\ & = P \left(T \in \left(\frac{n - k_1 - k_2}{k_2} \left(\exp \left(\frac{2k_1}{n} \right) - 1 \right), qF(1 - \alpha, k_2, n - k_1 - k_2) \right) \right) \\ & = \max \left\{ 0, 1 - \alpha - pF \left(\frac{n - k_1 - k_2}{k_2} \left(\exp \left(\frac{2k_1}{n} \right) - 1 \right), k_2, n - k_1 - k_2 \right) \right\}, \end{aligned}$$

pričom zavedené značenie pF vyjadruje distribučnú funkciu Fisher-Snedecorovho rozdelenia v danom bode a s danými parametrami. K vyjadreniu pomocou maximovej funkcie sme sa museli uchýliť preto, lebo musíme zvlášť odlíšiť prípad:

$$qF \left(\frac{n - k_1 - k_2}{k_2} \left(\exp \left(\frac{2k_1}{n} \right) - 1 \right), k_2, n - k_1 - k_2 \right) > qF(1 - \alpha, k_2, n - k_1 - k_2),$$

v tomto prípade je totiž daná pravdepodobnosť zjavne nulová. Pre pravdepodobnosť, že o parametri β_2 rozhodneme, že je naozaj nulový potom platí:

$$\begin{aligned} & P(AIC_2 < AIC_1) + P(\text{nez}.H_0 : \beta_2 = 0 \wedge AIC_1 < AIC_2) = \\ & pF \left(\frac{n - k_1 - k_2}{k_2} \left(\exp \left(\frac{2k_1}{n} \right) - 1 \right), k_2, n - k_1 - k_2 \right) + \\ & + \max \left\{ 0, 1 - \alpha - pF \left(\frac{n - k_1 - k_2}{k_2} \left(\exp \left(\frac{2k_1}{n} \right) - 1 \right), k_2, n - k_1 - k_2 \right) \right\}. \end{aligned}$$

Všimnime si, že výraz na pravej strane je buď rovný $1 - \alpha$,

ak $1 - \alpha > pF\left(\frac{n-k_1-k_2}{k_2} \left(\exp\left(\frac{2k_1}{n}\right) - 1\right), k_2, n - k_1 - k_2\right)$, alebo
 $pF\left(\frac{n-k_1-k_2}{k_2} \left(\exp\left(\frac{2k_1}{n}\right) - 1\right), k_2, n - k_1 - k_2\right)$,
 ak $1 - \alpha < pF\left(\frac{n-k_1-k_2}{k_2} \left(\exp\left(\frac{2k_1}{n}\right) - 1\right), k_2, n - k_1 - k_2\right)$.

Z uvedeného vyplýva, že hoci sme pri rozhodnutí o signifikantnosti parametra β_2 kombinovali testovanie hypotézy a rozhodnutie podľa AIC , výsledok je rovnaký, ako keby sme buď iba testovali hypotézu, alebo sa rozhodli iba na základe AIC . Ak je totiž hladina testu dostatočne malá (menšia ako $1 - \frac{n-k_1-k_2}{k_2} \left(\exp\left(\frac{2k_1}{n}\right) - 1\right)$), selekcia nemá na testovanie hypotézy o parametri β_2 žiaden vplyv, pretože dostávame rovnaký výsledok, ako keby sme daný test previedli bez prevedenia selekcie. Naopak, ak je hladina testu príliš veľká (väčšia ako $1 - \frac{n-k_1-k_2}{k_2} \left(\exp\left(\frac{2k_1}{n}\right) - 1\right)$), testovanie hypotézy o nulovosti parametra β_2 po selekcii už nemá význam, pretože pre pravdepodobnosť výberu modelu M_1 oproti modelu M_2 dostávame rovnakú hodnotu, ako keby sme iba vykonali selekciu bez testovania. Inými slovami, ak sme si v tomto prípade raz už vybrali model $\mathcal{M}_1$ pomocou selekcie, test o nulovosti parametra β_2 bude vždy zamietnutý.

4 Testovanie hypotéz o parametri po aplikovaní selekcie

V časti 2.2 sa nám už podarilo ukázať, že ak pripustíme selekciu parametra, ktorého nulovosť chceme testovať, je výsledok tohoto procesu analogický, ako keby sme použili iba test, alebo iba selekciu (v závislosti od hladiny testu). Ďalej sa už teda budeme zaoberať iba takým procesom testovania, kedy sa budeme snažiť selekciou vylúčiť iba zvyšné parametre, a teda nie parameter, ktorého nulovosť chceme testovať.

V tejto kapitole sa teda budeme zaoberať otázkou, akej chyby sa v skutočnosti dopustíme, ak aplikujeme štandardné testovanie hypotézy o nulovosti parametra v lineárnom regresnom modeli, ktorý sme získali selekciou premenných. Ukazuje sa, že ide o vcelku náročný problém (minimálne pre autora tohto textu), preto si ako úplný model (na ktorý aplikujeme selekciu) zvolíme model s iba tromi parametrami (intercept, parameter ktorý testujeme a parameter ktorý selektujeme). Poďme si danú situáciu lepšie popísať.

Majme nasledujúce dva lineárne regresné modely:

$$\mathcal{M}_1 : Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

$$\mathcal{M}_2 : Y = \beta_0 + X_1\beta_1 + \varepsilon.$$

Vektor Y je vektor vysvetlovaných premenných spoločný pre oba modely. Vektor X_1 je vektorom vysvetľujúcich premenných prislúchajúci parametru β_1 , ktorého nulovosť chceme testovať. Vektor X_2 je vektorom vysvetľujúcich premenných prislúchajúcich parametru β_2 , ktorý sa budeme snažiť vyselektovať pomocou Akaikeho kritéria (teda podľa AIC sa rozhodneme, či hypotézu o nulovosti β_1 budeme testovať v modeli $\mathcal{M}_1$, alebo v modeli $\mathcal{M}_2$). Nakoniec parameter β_0 zodpovedá interceptu a ε zodpovedá vektoru rezíduí, pričom pre jeho rozdelenie platí $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$.

Poznámka: Keďže sa zaoberáme zjednodušeným prípadom, keď je úplný model $\mathcal{M}_1$ tvorený iba tromi parametrami, X_1 a X_2 už pre nás nebudú matice ako v kapitole 3, ale iba vektory. Podobne β_1 a β_2 už pre nás nebudú vektory parametrov, ale iba hodnota jedného parametra.

Pre modely $\mathcal{M}_1$ a $\mathcal{M}_2$ budeme tiež používať maticový zápis:

$$\mathcal{M}_1 : Y = Z_1\theta_1 + \varepsilon,$$

$$\mathcal{M}_2 : Y = Z_2\theta_2 + \varepsilon,$$

pričom $Z_1 = [\mathbf{1}_n, X_1, X_2]$, $Z_2 = [\mathbf{1}_n, X_1]$, $\mathbf{1}_n$ je n -rozmerný vektor jednotiek a θ_1, θ_2 sú vektory parametrov $(\beta_0, \beta_1, \beta_2)^T, (\beta_0, \beta_1)^T$. Nakoniec zaveďme nasledovné submodely $\mathcal{M}_{1s}$ a $\mathcal{M}_{2s}$ modelov $\mathcal{M}_1$ a $\mathcal{M}_2$:

$$\mathcal{M}_{1s} : Y = \beta_0 + X_2\beta_2 + \varepsilon,$$

$$\mathcal{M}_{2s} : Y = \beta_0 + \varepsilon,$$

spolu s ich maticovou verziou:

$$\mathcal{M}_{1s} : Y = Z_{1s}\theta_{1s} + \varepsilon,$$

$$\mathcal{M}_{2s} : Y = Z_{2s}\theta_{2s} + \varepsilon,$$

kde $Z_{1s} = [\mathbf{1}_n, X_2]$, $Z_{2s} = \mathbf{1}_n$, $\theta_{1s} = (\beta_0, \beta_2)^T$ a $\theta_{2s} = \beta_0$.

Poznámka: Uvedomme si, že skutočným modelom (ktorým sa generujú dáta Y) je model $\mathcal{M}_1$. Ak teda pomocou selekcie vyberieme model $\mathcal{M}_2$, môže sa stať, že sme pre dané dáta Y vybrali model, ktorým sa tieto dáta neriadia. V takom prípade skutočné parametre β_0, β_1 a ε v modeli $\mathcal{M}_2$ prakticky neexistujú. Napriek tomu pre ne ale môžeme robiť odhady.

Povedzme teda napríklad, že pomocou *AIC* sme vyseletovali parameter β_2 , a teda sme sa rozhodli aplikovať na naše dáta model $\mathcal{M}_2$. Teraz otestujeme hypotézu $H_0 : \beta_1 = 0$. Použijeme štandardnú testovú štatistiku:

$$T := \frac{RSS_{2s} - RSS_2}{S^2}, \quad (45)$$

kde RSS_{2s} a RSS_2 sú súčty štvorcov rezíduí v príslušných modeloch $\mathcal{M}_{2s}, \mathcal{M}_2$ a $S^2 = \frac{RSS_2}{n-2}$. Testovú štatistiku T následne porovnáme s kvantilovou funkciou Fisher-Snedecorovho rozdelenia so stupňami voľnosti 1, $n - 2$ a pokiaľ táto štatistika prekročí $(1 - \alpha)$ -kvantil daného rozdelenia, hypotézu H_0 zamietneme (pričom α je zvolená hladina testu). Z teórie (napr. [5]) potom vieme, že pravdepodobnosť chyby prvého druhu (teda že zamietneme H_0 , napriek tomu, že je táto hypotéza platná) pri takto vykonanom teste je štandardne práve α . Problémom však je, že sme tento test aplikovali po vykonaní selekcie, čo má na jeho chybovosť signifikantný vplyv, čiže už nebude pravdepodobnosť chyby prvého druhu rovná hladine α .

Naším cieľom teda bude určiť, aká tá chybovosť bude po aplikovaní selekcie naozaj. V ideálnom prípade by sme chceli nájsť nejaký vzorec, do ktorého vieme dosadiť jednotlivé parametre modelov a ako výstup dostaneme danú chybovosť. Hoci sme si ale pre analýzu vybrali ten najjednoduchší prípad, kde selektujeme iba jeden parameter a zároveň testujeme tiež iba jeden parameter, tento cieľ sa žiaľ ukázal ako príliš ambiciózny. Pri odhade skutočnej chybovosti testu po aplikovaní selekcie sa preto uchýlime k použitiu simulácií.

Ak by sme skutočné hodnoty parametrov β_0 , β_2 a σ_ε^2 poznali, tak by bol simulačný prístup veľmi jednoduchý. Stačí do modelu $\mathcal{M}_1$ dosadiť všetky hodnoty parametrov a dostatočne veľa krát vygenerovať ε -ny. V každej simulácii potom skúsime vyselektovať parameter β_2 a v závislosti od zvoleného modelu po selekcii budeme pomocou testovej štatistiky (45) testovať (pravdivú) hypotézu $H_0 : \beta_1 = 0$. Potom pravdepodobnosť $P(\text{zam}.H_0 \mid AIC_1 < AIC_2)$ odhadneme ako podiel počtu prípadov kedy sme selekciou vybrali model $\mathcal{M}_1$ a zamietli sme hypotézu H_0 , ku počtu všetkých prípadov. Analogicky, pravdepodobnosť $P(\text{zam}.H_0 \mid AIC_2 < AIC_1)$ odhadneme ako podiel počtu prípadov kedy sme selekciou vybrali model $\mathcal{M}_2$ a zamietli sme hypotézu H_0 , ku počtu všetkých prípadov. Intuitívne sa dá potom očakávať, že zvyšovaním počtu simulácií vieme takýmto jednoduchým postupom pravdepodobnosti $P(\text{zam}.H_0 \mid AIC_1 < AIC_2)$ a $P(\text{zam}.H_0 \mid AIC_2 < AIC_1)$ odhadnúť ľubovoľne presne.

Problémom tohto prístupu však je, že v reálnom svete parametre β_0 , β_2 a σ_ε^2 (o parametri β_1 predpokladáme, že je nulový) nepoznáme. V takom prípade potom žiaľ simulačný postup popísaný v predošlom odstavci nefunguje, pretože nevieme čo máme za tieto parametre dosadiť.

Napriek tomu že sú pre nás parametre β_0 , β_2 a σ_ε^2 neznáme, vieme si o ich hodnotách vytvoriť nejaké presvedčenie v závislosti od vysvetľujúcich premenných X_2 a vysvetľovaných premenných Y , ktoré pre nás známe sú. Dáva totiž napríklad zmysel, že v závislosti od dát máme inú mieru presvedčenia o tom, že parameter β_2 leží v intervale $(0, 1)$, oproti tomu, že leží v intervale $(1000, 1001)$. K našej situácii teda vieme zaujať Bayesovský prístup, ktorého typickou črtou je, že mieru presvedčenia o neznámych parametroch vyjadrujeme pomocou hustoty pravdepodobnosti (a teda sa ku konštantným, ale neznámym parametrom správame ako ku náhodným premenným).

Otázkou zostáva, aká hustota vhodne popisuje mieru nášho presvedčenia o paramet-

roch β_0 , β_2 a σ_ε^2 . Odpoveď na túto otázku je našťastie v teórii bayesovskej štatistiky už zodpovedaná a je bližšie popísaná v knihe [7]. Ukazuje sa, že vhodná neinformatívna apriórna hustota popisujúca mieru presvedčenia o skutočných hodnotách parametrov θ_{1s} a σ_ε^2 je charakterizovaná funkciou $\pi(\theta_{1s}, \sigma_\varepsilon^2) \propto \frac{1}{\sigma_\varepsilon^3}$, pričom zavedené značenie $\propto$ hovorí, že apriórna hustota $\pi(\theta_{1s}, \sigma_\varepsilon^2)$ je úmerná pravej strane. Dá sa ukázať, že posterior pre úplný model $\mathcal{M}_1$ bude mať v takom prípade nasledujúci tvar:

$$\pi(\theta_{1s}, \sigma_\varepsilon^2 | Y) \propto \frac{1}{\sigma_\varepsilon^{n+3}} \exp \left(-\frac{1}{2\sigma_\varepsilon^2} (Y - Z_{1s}\theta_{1s})^T (Y - Z_{1s}\theta_{1s}) \right). \quad (46)$$

Zároveň máme v knihe [7] odvodené aj nasledovné marginálne hustoty pre parametre θ_{1s} a σ_ε^2 :

$$\pi(\theta_{1s} | Y) \propto \left(\|Y - X_{1s}\hat{\theta}_{1s}\|^2 + \|X_{1s}(\hat{\theta}_{1s} - \theta_{1s})\|^2 \right)^{-\frac{n+1}{2}}, \quad (47)$$

$$\pi(\sigma_\varepsilon^2 | Y) \propto \left(\frac{\|Y - X_{1s}\hat{\theta}_{1s}\|^2}{\sigma_\varepsilon^2} \right)^{\frac{n+1}{2}} \exp \left(-\frac{1}{2\sigma_\varepsilon^2} \|Y - X_{1s}\hat{\theta}_{1s}\|^2 \right), \quad (48)$$

pričom značením $\|\cdot\|$ myslíme klasickú euklidovkú normu.

Označme pravdepodobnosti $P(zam.H_0 : \beta_1 = 0 | \theta_{1s}, \sigma_\varepsilon^2, AIC_1 < AIC_2)$ a $P(zam.H_0 : \beta_1 = 0 | \theta_{1s}, \sigma_\varepsilon^2, AIC_2 < AIC_1)$ ako $\Psi(AIC_1 < AIC_2 | \theta_{1s}, \sigma_\varepsilon^2)$, $\Psi(AIC_2 < AIC_1 | \theta_{1s}, \sigma_\varepsilon^2)$. Ďalej označme odhady pravdepodobností $P(zam.H_0 : \beta_1 = 0 | AIC_1 < AIC_2)$ a $P(zam.H_0 : \beta_1 = 0 | AIC_2 < AIC_1)$ pri neznámych (a teda kvázi “náhodných”) parametroch θ_{1s} , σ_ε^2 ako $\Psi(AIC_1 < AIC_2)$, $\Psi(AIC_2 < AIC_1)$. Potom pri neznámych parametroch θ_{1s} , σ_ε^2 dáva zmysel vyjadriť odhady pre chybovosti $\Psi(AIC_1 < AIC_2)$, $\Psi(AIC_2 < AIC_1)$ nášho modelu nasledujúcim spôsobom:

$$\Psi(AIC_1 < AIC_2) = \int_{\Theta} \int_{\Sigma} \Psi(AIC_1 < AIC_2 | \theta_{1s}, \sigma_\varepsilon^2) \pi(\theta_{1s}, \sigma_\varepsilon^2 | Y) d\sigma_\varepsilon^2 d\theta_{1s}, \quad (49)$$

$$\Psi(AIC_2 < AIC_1) = \int_{\Theta} \int_{\Sigma} \Psi(AIC_2 < AIC_1 | \theta_{1s}, \sigma_\varepsilon^2) \pi(\theta_{1s}, \sigma_\varepsilon^2 | Y) d\sigma_\varepsilon^2 d\theta_{1s}, \quad (50)$$

pričom Θ je oblasť z ktorej pochádza vektor θ_{1s} a Σ je oblasť z ktorej pochádza hodnota σ_ε^2 (ide vlastne o oblasti $\mathbb{R}^2$ a $\mathbb{R}$).

Poznámka: Integrály (49) a (50) nevieme spočítať explicitne. Na prípadné zistenie ich hodnoty treba preto buď využiť simulácie, alebo nejakú numerickú metódu.

Pre skutočnú pravdepodobnosť zamietnutia $H_0 : \beta_1 = 0$, teda $P(zam.H_0 : \beta_1 = 0 | \bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2)$, pričom pod $\bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2$ myslíme skutočné hodnoty parametrov θ_{1s} a σ_ε^2 , vieme

jednoduchým spôsobom nájsť aj intervalový odhad. Pre ľubovoľné hodnoty parametrov θ_{1s} a σ_ε^2 vieme totiž odhadnúť pravdepodobnosť $\Psi(AIC_1 < AIC_2 \mid \theta_{1s}, \sigma_\varepsilon^2)$, prípadne $\Psi(AIC_2 < AIC_1 \mid \theta_{1s}, \sigma_\varepsilon^2)$ pomocou simulácií (parametre θ_{1s} a σ_ε^2 dosadíme do modelu $\mathcal{M}_1$ a dostatočne veľa krát vygenerujeme z tohto modelu Y -ny. Potom pravdepodobnosť $\Psi(AIC_1 < AIC_2 \mid \theta_{1s}, \sigma_\varepsilon^2)$ odhadneme jednoducho ako podiel počtu tých realizácií kedy sa selekciou zvolil úplný model a hypotéza H_0 sa zamietla ku počtu všetkých realizácií kedy sa selekciou vybral úplný model. Analogicky vieme odhadnúť pravdepodobnosť $\Psi(AIC_2 < AIC_1 \mid \theta_{1s}, \sigma_\varepsilon^2)$. Potom nám stačí vygenerovať n dvojíc parametrov $\theta_{1s}, \sigma_\varepsilon^2$ z hustoty (46) (pomocou Monte Carlo accept - reject metódy), pre každú z dvojíc simulačne odhadnúť $\Psi(AIC_1 < AIC_2 \mid \theta_{1s}, \sigma_\varepsilon^2)$, respektíve $\Psi(AIC_2 < AIC_1 \mid \theta_{1s}, \sigma_\varepsilon^2)$ a následne tieto odhady chybovostí pri vygenerovaných parametroch usporiadať. Keďže predpokladáme, že parametre $\bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2$ tiež pochádzajú z hustoty (46) (Bayesovský prístup), skutočná chybovosť $\Psi(AIC_1 < AIC_2 \mid \bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2)$, respektíve $\Psi(AIC_2 < AIC_1 \mid \bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2)$ je rovnocenná (v zmysle, že pochádza z rovnakej hustoty) s nami vygenerovanými chybovosťami. Z týchto vygenerovaných usporiadaných chybovostí potom stačí zahodiť 2,5% najmenších odhadov, 2,5% najväčších odhadov a zo zostávajúcich chybovostí vziať najmenší a najväčší odhad ako ľavé a pravé ohraničenie 95% - ného intervalu spoľahlivosti.

Podobným spôsobom vieme určiť aj bodový odhad, ktorý je alternatívou k odhadom (49), (50). Z usporiadaných generovaných pravdepodobností $\Psi(AIC_1 < AIC_2 \mid \theta_{1s}, \sigma_\varepsilon^2)$, respektíve $\Psi(AIC_2 < AIC_1 \mid \theta_{1s}, \sigma_\varepsilon^2)$ určíme bodový odhad pravdepodobnosti $\Psi(AIC_1 < AIC_2 \mid \bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2)$, respektíve $\Psi(AIC_2 < AIC_1 \mid \bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2)$ jednoducho ako medián.

4.1 Testovanie po aplikovaní selekcie na reálnych dátach

Postup pri odhadovaní skutočnej chybovosti popísaný v tejto kapitole budeme demonštrovať na nasledujúcom príklade. Majme lineárny regresný model, ktorého vysvetľované a vysvetľujúce premenné sú uvedené v nasledujúcej tabuľke:

Y	X ₁	X ₂
3.734	0.58	5.79
4.182	-0.84	2.76
3.717	0.21	3.67
4.088	1.11	6.85
3.930	0.71	5.12
4.612	-0.10	7.10
3.393	1.61	7.05
3.598	-0.11	4.38

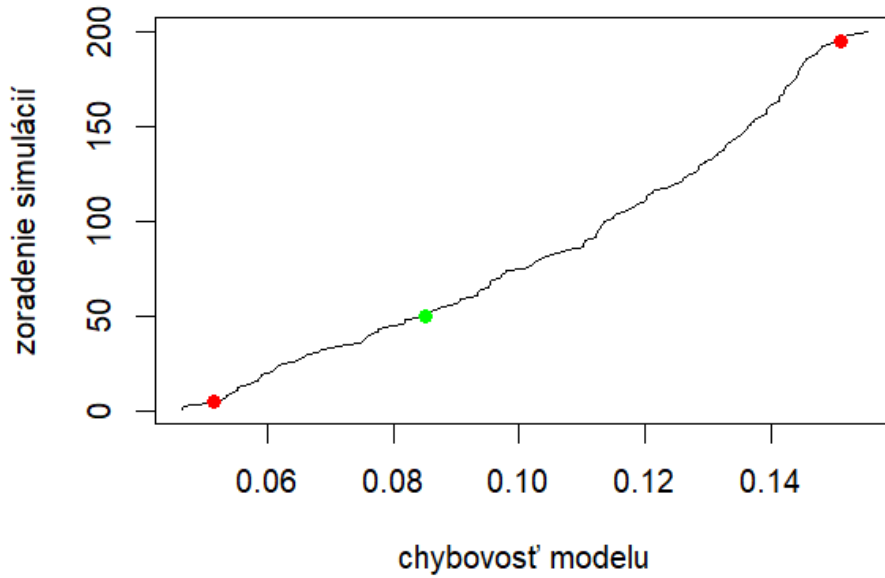
Skutočné parametre tohto modelu sú $\beta_0 = 3$, $\beta_2 = 0.2$, $\sigma_\varepsilon^2 = 0.5$ (pričom sa samozrejme budeme tváriť, že tieto parametre sú pre nás neznáme). Vysvetľujúce premenné Y vznikli teda náhodným vygenerovaním z tohto modelu. Je zároveň splnený predpoklad $H_0 : \beta_1 = 0$. Naším cieľom je odhadnúť, aká je skutočná chybovosť testu aplikovaného po selekcii v tomto modeli, teda aká je pravdepodobnosť, že sa pri skutočných parametroch tohto modelu vygeneruje Y také, že hypotézu H_0 po aplikovaní selekcie zamietneme.

Lahko vieme spočítať, že $AIC_1 = -2.33$ a $AIC_2 = -1.82$. Keďže vidíme, že $AIC_1 < AIC_2$, zvolíme si v tomto prípade selekciou úplný model, a teda sa budeme primárne zaujímať o chybovosť úplného modelu.

Podme sa teraz zamerať na zistenie vhodného 95%-ného intervalu spoľahlivosti pre chybovosť nášho modelu. O skutočných premenných θ_{1s} , σ_ε^2 máme teda aposteriórne presvedčenie (46). Pomocou Monte Carlo accept - reject metódy vygenerujeme 200 dvojíc parametrov θ_{1s} , σ_ε^2 z hustoty (46) a pre každú z dvojíc simulačne odhadneme k akej chybovosti v modeli dochádza, v prípade ak sme selekciou vybrali úplný model. Výsledky týchto simulácií uvádzame na obr. 2.

Na obrázku 2 môžeme vidieť aký odhad chybovosti sme dostali pre každú dvojicu nasimulovaných parametrov θ_{1s} , σ_ε^2 , pričom červené body vymedzujú 95% - ný interval spoľahlivosti pre skutočnú chybovosť nášho modelu a zelený bod znázorňuje skutočnú chybovosť nášho modelu (ktorú sme získali dosadením skutočných hodnôt parametrov θ_{1s} , σ_ε^2 do modelu a následnou analogickou simuláciou ako pri vygenerovaných parametroch).

V našom modeli sme teda odhadli, že jeho skutočná chybovosť pri testovaní $H_0 : \beta_1 = 0$, leží za predpokladu že sa vyseletoval úplný model na 95% v intervale (0.051; 0.151)



Obr. 2: Interval spoľahlivosti pre chybovosť testu v úplnom modeli

(pričom skutočná hodnota tejto chybovosti je približne 0.085). Nakoniec môžeme tiež určiť bodový odhad chybovosti v testovaní H_0 ako medián z vygenerovaných chybovostí, a teda ako 0.113.

Poznámka: Podobne by sme mohli teoreticky postupovať, aj keby sme mali k dispozícii model s viacerými premennými. Prakticky je však táto situácia veľmi ťažko realizovateľná, kvôli výpočtovej náročnosti. Jednak nám so zvyšujúcim sa počtom parametrov exponenciálne rastie počet modelov, ktoré sa môžu vyselektovať, ale aj generovanie z veľarozmerných hustôt je problematické (pretože dochádza k častému zamietaniu hodnôt pri Monte Carlo accept - reject metóde).

Poznámka: Je potrebné si uvedomiť, že získané odhady v našom ilustračnom príklade môžu byť do istej miery závislé od voľby prioru $\pi(\theta_{1s}, \sigma_\varepsilon^2) \propto \frac{1}{\sigma_\varepsilon^3}$, najmä preto, že pre väčšiu prehľadnosť sme v príklade použili model, ktorý obsahuje veľmi málo dát (iba 8). V modeli s dostatočným počtom dát je však tento problém zanedbateľný.

4.2 Analýza nepodmienenenej pravdepodobnosti chyby prvého druhu

Doteraz sme sa zaujímali iba o podmienenú pravdepodobnosť chyby prvého druhu za podmienky, že vieme, ktorý model sa selekciou zvolil. Rovnako sa ale môžeme zaujímať o

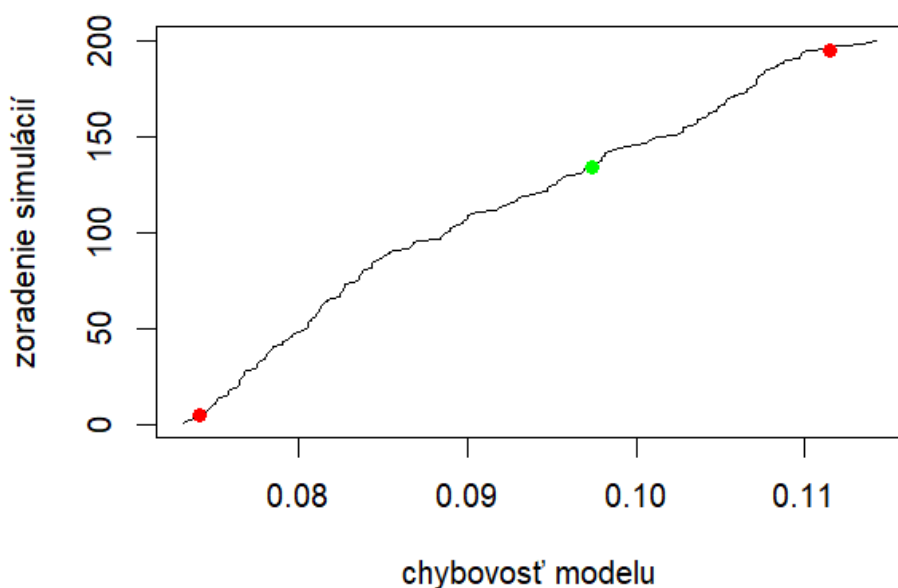
celkovú pravdepodobnosť chyby prvého druhu pri zamietaní $H_0 : \beta_1 = 0$, a to bez ohľadu na to, ktorý model sa zvolí selekciou.

V princípe môžeme postupovať veľmi podobne. Ak by sme poznali skutočné parametre modelu, stačí nám ich do modelu dosadiť, nasimulovať dostatočné množstvo vektorov Y a pri každej simulácii previesť najprv selekciu pomocou AIC a následne vo vyselektovanom modeli test pomocou klasickej Fisher-Snedecorovej testovej štatistiky. Výslednú nepodmienenú pravdepodobnosť chyby prvého druhu vieme potom veľmi jednoducho odhadnúť, ako počet prípadov kedy došlo k zamietnutiu H_0 deleno počet všetkých simulácií. Dá sa potom rovnako očakávať, že zvyšovaním počtu simulácií vieme takto dostať pre chybu prvého druhu ľubovoľne presný odhad.

Keďže ale v realite presné hodnoty parametrov nepoznáme, znova si o nich budeme potrebovať vytvoriť presvedčenie pomocou posterioru (46). Z tohto posterioru teraz môžeme generovať parametre $\theta_{1s}, \sigma_\varepsilon^2$ analogicky ako v kapitole 3. Pre každú dvojicu takto vygenerovaných parametrov vieme potom ich dosadením do modelu a následnou simuláciou odhadnúť $P(zam.H_0 : \beta_1 = 0 \mid \theta_{1s}, \sigma_\varepsilon^2)$. Z týchto vygenerovaných pravdepodobností vieme potom vytvoriť bodový odhad a interval spoľahlivosti pre skutočnú pravdepodobnosť chyby prvého druhu, teda $P(zam.H_0 : \beta_1 = 0 \mid \bar{\theta}_{1s}, \bar{\sigma}_\varepsilon^2)$, analogickým spôsobom ako sme to urobili pre pravdepodobnosti podmienené výberom modelu v kapitole 3.

Týmto prístupom sme nakoniec vypočítali odhad pre celkovú pravdepodobnosť chyby prvého druhu pri testovaní hypotézy H_0 (nepodmienenú výberom modelu) v príklade z podkapitoly 3.1. Výsledky simulácií uvádzame na obr. 3:

V modeli sme odhadli, že skutočná chybovosť pri testovaní $H_0 : \beta_1 = 0$, leží na 95% v intervale (0.074;0.112) (pričom skutočná hodnota tejto chybovosti je približne 0.097). Nakoniec môžeme tiež určiť bodový odhad chybovosti v testovaní H_0 ako medián z vygenerovaných chybovostí, a teda ako 0.089.



Obr. 3: Interval spoľahlivosti pre pravdepodobnosť chyby prvého druhu

5 Analýza výberovej korelácie vysvetľujúcich premenných

V tejto kapitole uvedieme niektoré z našich empirických pozorovaní. Rovnako ako v kapitole 4, budeme sa zaoberať situáciou, keď máme na výber nasledujúce dva modely:

$$\mathcal{M}_1 : Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

$$\mathcal{M}_2 : Y = \beta_0 + X_1\beta_1 + \varepsilon.$$

Predpokladajme, že na týchto modeloch opäť prevedieme selekciu prostredníctvom Akaikeho kritéria a na vyselektovanom modeli budeme testovať hypotézu $H_0 : \beta_1 = 0$ pomocou štandardnej Fisher-Snedecorovej testovej štatistiky F :

$$F := (n - k) \frac{RSS_{sub} - RSS}{RSS},$$

pričom RSS je reziduálny súčet štvorcov vo vyselektovanom modeli, RSS_{sub} je reziduálny súčet štvorcov v jeho submodeli (neobsahujúcom parameter β_1), n je počet meraní a k je počet parametrov vo vyselektovanom modeli. Túto štatistiku následne porovnávame s 95%-ným kvantilom rozdelenia $F_{1,n-k}$.

V tejto práci sa nám už podarilo viackrát demonštrovať, že tento postup nie je korektný (pravdepodobnosť chyby prvého druhu pri zamietaní hypotézy H_0 nie je rovná 5%, čiže

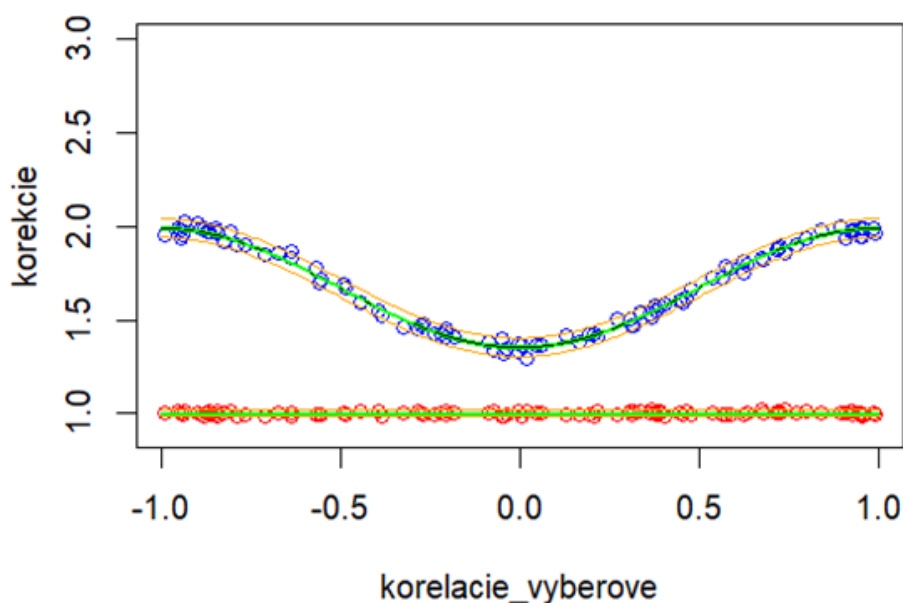
nastavenej hladine testu). Budeme sa preto snažiť nájsť vhodnú konštantu c , ktorou treba prenásobiť 95%-ný kvantil príslušného Fisher - Snedecorovho rozdelenia, aby skutočná chyba prvého druhu testu prevedeného po selekcii bola 5%.

Naším prirodzeným želaním je samozrejme nájsť nejakú jednoduchú charakteristiku modelu $\mathcal{M}_1$, ktorá by nám pomohla korekciu c určiť. Takouto charakteristikou by mohla byť napríklad výberová korelácia vysvetľujúcich premenných X_1 a X_2 . Do akej miery vplýva táto výberová korelácia na korekciu c vieme potom analyzovať simulačne.

Použijeme teda nasledujúci simulačný prístup: nastavíme disperziu σ_ε^2 , parametre β_0 , $\beta_1 = 0$, β_2 , stredné hodnoty vektorov X_1 , X_2 a disperzie vektorov X_1 , X_2 na nami zvolené hodnoty. Následne vygenerujeme koreláciu X_1 a X_2 rovnomerne náhodne z intervalu $(-1, 1)$. Potom vygenerujeme vysvetľujúce premenné X_1 a X_2 z normálneho rozdelenia tak, aby platili ich nastavené stredné hodnoty, disperzie a korelácia. Vygenerované vysvetľujúce premenné X_1 a X_2 dosadíme do modelu $\mathcal{M}_1$ a urobíme dostatočný počet simulácií, v ktorých budeme generovať chyby ε (čím zároveň generujeme dáta Y). V každej takto prevedenej simulácii potom vieme pomocou AIC vybrať model a vo vybanom modeli následne spočítať testovú štatistiku F . Pre oba modely vieme potom tieto testové štatistiky zoradiť, čím dostaneme empirickú kvantilovú funkciu. Vhodné korekcie c_1 a c_2 ktorými chceme prenásobiť 95%-ný kvantil Fisher-Snedecorovho rozdelenia daný modelom $\mathcal{M}_1$, respektíve $\mathcal{M}_2$, vypočítame ako podiel empirického 95%-ného kvantilu v danom modeli a teoretického 95%-ného kvantilu Fisher-Snedecorovho rozdelenia daného príslušným modelom.

Ukazuje sa, že pokiaľ je parameter β_2 tiež nulový, tak sú korekcie c_1 , c_2 kvantilov Fisher-Snedecorovho rozdelenia daného modelom $\mathcal{M}_1$, respektíve $\mathcal{M}_2$ jednoznačne určené výberovou koreláciou vysvetľujúcich premenných X_1 a X_2 . Túto situáciu ilustrujeme na obrázku 4, ktorý bližšie popisujeme v nasledujúcom odseku.

Počet pozorovaní v modeli je 10. Hodnotu σ_ε^2 sme nastavili na 25, stredné hodnoty vektorov X_1 , X_2 sme nastavili na 0 a 5, ich disperzie na 1 a 100. Intercept β_0 sme zvolili rovný 3 a ako sme už uviedli, parametre β_1 a β_2 sú nulové. Vykonali sme 100 realizácií vysvetľujúcich premenných X_1 , X_2 a pre každú realizáciu sme následne urobili 100000 simulácií chýb ε , čím sa nám vygenerovali vysvetľované premenné Y . Postupom popísaným v predchádzajúcom odseku sme pre každú realizáciu vektorov X_1 , X_2 určili korekcie



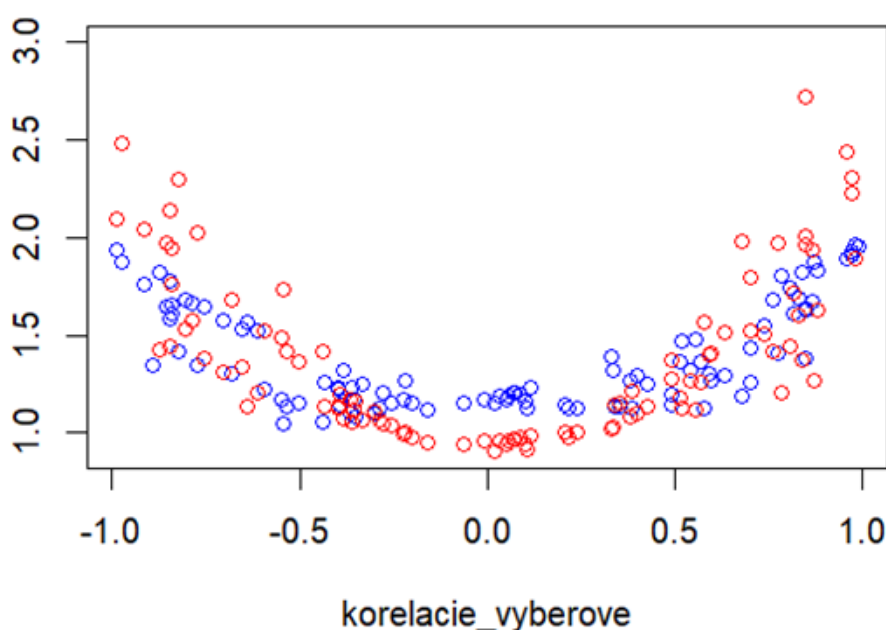
Obr. 4: Korekcia kvantilu v závislosti od výberovej korelácie a vyselektovaného modelu

c_1 , c_2 kvantilov Fisher-Snedecorovho rozdelenia určeného modelom $\mathcal{M}_1$, respektíve $\mathcal{M}_2$. Na obrázku 4 potom vykresľujeme modrou k akej korekcii dochádza pri danej výberovej korelácií vektorov X_1 a X_2 za predpokladu, že sa selekciou zvolil model $\mathcal{M}_1$ a analogicky červenou farbou vykresľujeme k akej korekcii dochádza pri danej výberovej korelácií za predpokladu, že sa zvolil model $\mathcal{M}_2$.

Podarilo sa nám teda empiricky zistiť, že pokiaľ je hodnota parametra β_2 nulová a pomocou AIC vyselektujeme model $\mathcal{M}_2$, žiadna korekcia nie je potrebná, a teda $c_2 = 1$. Za predpokladu že vyselektujeme model $\mathcal{M}_1$ už síce korekcia potrebná je, ale jej hodnota je jednoznačne (aspoň sa tak javí) určená výberovou koreláciou vektorov X_1 , X_2 . Ukazuje sa dokonca, že závislosť medzi výberovou koreláciou a korekciou c_1 je vcelku dobre popísateľná pomocou goniometrickej regresie. Konkrétne v tomto prípade sa nám podarilo hodnotu c_1 odhadnúť ako $\hat{c}_1 = 1.672 - 0.32\cos(\text{cor}(X_1, X_2))$ (tento odhad vykresľujeme na obr. 4 ako zelenú krivku spolu s príslušnými 95%-nými intervalmi spoľahlivosti vyznačenými pomocou oranžových kriviek).

Úplne analogicky môžeme postupovať, keď všetky ostatné hodnoty zachováme, akurát určíme parameter $\beta_2 = 0.2$. Túto situáciu vieme načrtnúť na obrázku 5.

Vidíme, že v tomto prípade sa už žiaľ korekcie c_1 , c_2 nejavia byť jednoznačne určené koreláciou vysvetľujúcich premenných X_1 , X_2 v ani jednom z modelov $\mathcal{M}_1$, $\mathcal{M}_2$ (opäť



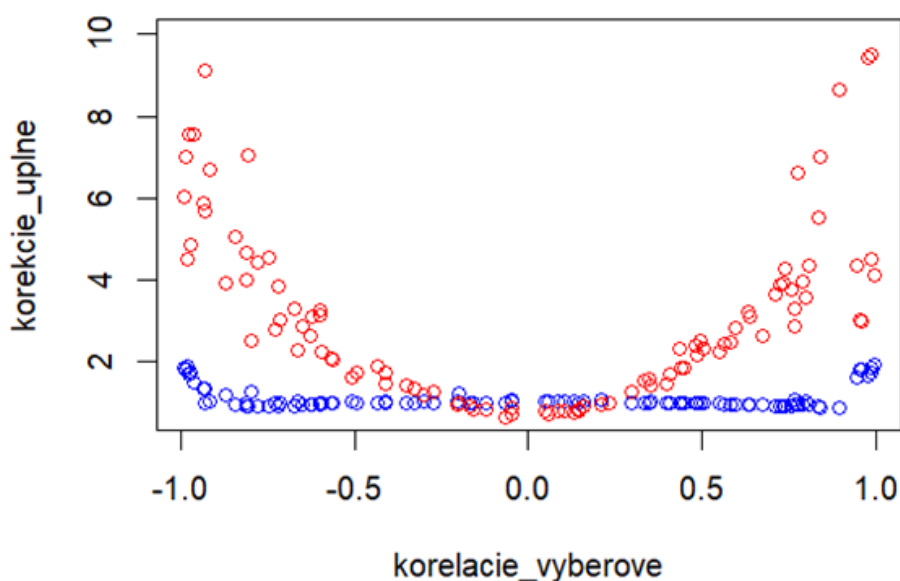
Obr. 5: Korekcia kvantilu v závislosti od výberovej korelácie a vyselektovaného modelu

vykresľujeme korekcie v úplnom modeli modrou a v neúplnom modeli červenou farbou) .

Nakoniec urobíme rovnakú analýzu pre hodnotu parametra β_2 rovnú 0.5, ktorej výsledky máme zaznamenané na obrázku 6:

Z obrázka 6 potom vidíme, že pokiaľ sme selekciou zvolili model $\mathcal{M}_1$, a výberová korelácia medzi vysvetľujúcimi premennými X_1, X_2 nie je veľmi vysoká, nie je potrebná žiadna korekcia, a teda $c_1 = 1$. Toto pozorovanie je zároveň v súlade so sedliackym rozumom. Uvedomme si totiž, že hodnota 0.5 zodpovedá už vcelku veľkej hodnote parametra β_2 a pokiaľ sa premenné X_1, X_2 dostatočne líšia (pričom výberová korelácia je vcelku dobrou mierou odlišnosti) je veľmi nepravdepodobné, že by sme pomocou *AIC* náhodne vyselektovali model $\mathcal{M}_2$. Potom ak sa takmer vždy vyselektuje model $\mathcal{M}_1$, situácia je veľmi podobná ako keby k žiadnej selekcii ani nedochádzalo a hypotézu by sme rovno testovali v modeli $\mathcal{M}_1$.

Poznámka: Najväčším problémom tejto kapitoly ostáva, že v realite nepoznáme skutočnú hodnotu parametra β_2 , a teda nemôžeme vedieť, ktorý z uvedených scenárov je pre nás relevantný.



Obr. 6: Korekcia kvantilu v závislosti od výberovej korelácie a vyselektovaného modelu

Záver

V práci sme sa postupne pozreli na problematiku testovania hypotéz o parametroch lineárneho regresného modelu po aplikovaní selekcie z viacerých uhlov pohľadu. V prvej kapitole sme popísali základné poznatky ohľadom lineárnej regresie a uviedli sme dôležité, všeobecne známe matematické tvrdenia, o ktoré sa neskôr v práci opierame.

V kapitole 2 sme čitateľa oboznámili s literatúrou, ktorá sa problematikou testovania po aplikovaní selekcie tiež zaoberá. Následne sme podrobnejšie popísali výsledky článku [3], pričom sme doplnili viaceré medzikroky, ktoré autori pôvodného textu vynechali. Zaviedli sme takzvaný *WALS* odhad parametra β , pre ktorý sme v tvrdení 2.2 odvodili všeobecný vzorec pre jeho strednú hodnotu, varianciu a strednú kvadratickú chybu. Taktiež sme odvodili všeobecný vzorec pre hodnotu rozdielu strednej kvadratickej chyby odhadu parametra β získaného v úplnom modeli a strednú kvadratickú chybu *WALS* odhadu. Nakoniec sme zaviedli premennú *UR*, ktorej cieľom je kvantifikovať mieru podreportovania skutočnej chyby testu a pre jej strednú hodnotu sme urobili horný odhad.

V kapitole 3 sme definovali Akaikeho kritérium, pomocou ktorého sme ďalej v práci vykonávali selekciu. Podarilo sa nám všeobecne vypočítať hodnotu pravdepodobnosti, že pomocou Akaikeho kritéria vyselektujeme neúplný model oproti úplnému. Následne sme sa zaoberali testovaním hypotézy o nulovosti parametra β_2 v prípade, že pripustíme možnosť

odstránenia tohto parametra prostredníctvom selekcie. Nakoniec sa ukázalo, že proces v ktorom sa najprv pokúsime parameter β_2 vyselektovať a následne testujeme hypotézu o nulovosti tohto parametra je rovnaký, ako keby sme iba previedli selekciu, respektíve iba testovali hypotézu (v závislosti od nastavenej hladiny testu).

V štvrtej kapitole sme sa snažili odhadnúť skutočnú chybu prvého druhu testu prevedeného po selekcii v prípade, že neumožníme odstránenie testovaného parametra prostredníctvom selekcie. Keďže táto chyba závisí od skutočných hodnôt parametrov modelu, zaviedli sme bayesovský prístup, ktorý nám umožnil správať sa k neznámych parametrom ako k náhodným premenným. Popísali sme, ako pomocou tohto prístupu nájsť bodový odhad a interval spoľahlivosti pre skutočnú chybovosť testu prevedeného po selekcii, a to jednak v závislosti od vyselektovaného modelu, ale aj bez ohľadu na vyselektovaný model. Následne sme tento prístup demonštrovali na konkrétnom príklade.

Nakoniec sme v kapitole 5 popísali ako súvisí výberová korelácia medzi vysvetľujúcimi premennými a veľkosť korekcie kvantilu príslušného Fisher-Snedecorovho rozdelenia daného vyselektovaným modelom. Ukázalo sa, že táto súvislosť je veľmi silná v prípade, ak je premenná ktorú sa snažíme vyselektovať naozaj nulová. V prípade že tomu tak nie je, už žiaľ nedokážeme povedať, že by hodnota potrebnej korekcie kvantilu Fisher-Snedecorovho rozdelenia bola daná výberovou koreláciou vysvetľujúcich premenných. To sa žiaľ ukazuje ako vcelku vážny problém, pretože v praxi reálne parametre modelu nepoznáme.

Zoznam použitej literatúry

- [1] Badár, M. 2021: *Selekcia premenných v lineárnej regresii a kvalita testov*, FMFI UK Bratislava.
- [2] Bancroft, T. A. 1944: *On Biases in Estimation Due to the Use of Preliminary Tests of Significance*, The Annals of Mathematical Statistics, Iowa State College, dostupné na internete (19.4.2023): <https://www.jstor.org/stable/pdf/2236199.pdf>
- [3] Danilov, D., Magnus, J. R. 2004: *On the harm that ignoring pretesting can cause*, Journal of Econometrics, Eindhoven University of Technology, dostupné na internete (19.4.2023): <https://www.janmagnus.nl/papers/JRM070.pdf>
- [4] Lovell, M. C. 1983: *Data Mining*, The Review of Economics and Statistics, The MIT Press, dostupné na internete (19.4.2023): <https://www.jstor.org/stable/pdf/1924403.pdf>
- [5] Johnston, J., Dinardo, J.: *Econometric Methods*, McGraw-Hill, 1997, 531 s., ISBN 9780071153423, 007115342X
- [6] Magnus, J. R. 1999: *The traditional pretest estimator*, All Russian mathematical portal, dostupné na internete (19.4.2023): <https://www.mathnet.ru/links/c18d36d5eb07845f6a54a46d05bb330f/tvp775.pdf>
- [7] Pázman, A. 2009: *Bayesovská štatistika*, FMFI, Bratislava, ISBN 978-80-223-2666-7