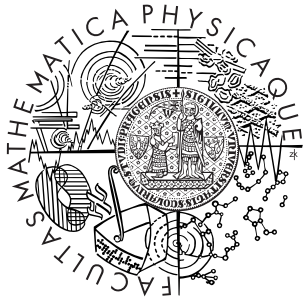
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Katarína Mordinová

Errors in variables

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Zdeněk Hlávka, Ph.D.
Studijní program: matematika, finanční a pojistná matematika

2011

Na tomto mieste by som chcela predovšetkým poďakovať vedúcemu diplomovej práce Mgr. Zdeňkovi Hlávkovi, Ph.D. za rady a pripomienky pri vypracovaní tejto práce. Moje poďakovanie patrí aj mojim rodičom za podporu počas celého štúdia.

Prehlasujem, že som svoju diplomovú prácu napísala samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V Prahe dňa 10. apríla 2011

Obsah

4

Názov práce: Errors in variables
Autor: Katarína Mordinová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci bakalárskej práce: Mgr. Zdeněk Hlávka, Ph.D.
e-mail vedúceho: Zdenek.Hlavka@mff.cuni.cz

Abstrakt: Práca pojednáva o regresných modeloch známych ako Errors in variables. V úvodnej kapitole tejto práce definujeme základné pojmy, ktoré v práci používame a základné vzťahy, ktoré s regresnou analýzou súvisia. V druhej kapitole sa venujeme lineárnemu regresnému modelu a jeho vlastnostiam. V tretej kapitole sa zaoberáme jednotlivými typmi modelov errors in variables. V záverečnej kapitole tejto práce si ukážeme možnú aplikáciu modelov v medicíne.
Kľúčové slová: regresná analýza, errors in varibles, lineárny regresný model

Title: Errors in variables
Author: Katarína Mordinová
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Zdeněk Hlávka, Ph.D.
Supervisor's e-mail address: Zdenek.Hlavka@mff.cuni.cz

Abstract: The topic of the diploma thesis is Errors in variables. In the opening chapter, we define basic terms used in the thesis and we introduce the regression analysis and basic relations related to this term. In the second chapter, we attend to linear regression model and its characteristics. In the third chapter, we attend to the errors in variables models. In the last chapter of this thesis we present a possible application in medicine.
Keywords: regresion analysis, errors in variables, linear regression model

Úvod

V úvodnej kapitole tejto práce sú zadefinované základné pojmy, ktoré v práci používame. Keďže sa s pojmom regresia stretávame v rôznych oboroch a oblastiach, je zrejmé, že regresná analýza má početné možnosti aplikácie. Štandardne sa s jej aplikáciou stretávame v oboroch financí, ekonómie, obchodu, práva, štatistiky, medicíny a dalo by sa pokračovať ďalšími vednými obormi. V podstate, sa táto jednoduchá metóda dá využiť pri riešení širokého spektra problémov. S niektorými konkrétnymi aplikáciami sa neskôr budeme zaoberať aj v tejto práci.

Po zavedení základných pojmov prejdeme k základným vzťahom medzi skúmanými náhodnými veličinami. Predovšetkým si tieto vzťahy ukážeme na lineárnom regresnom modeli a normálnom regresnom modeli. Tretia kapitola je následne venovaná modelom errors in variables a ich podrobnejším popisom. Finálna, štvrtá kapitola, sa zaoberá aplikáciou modelov errors in variables v praxi. Pre potreby tejto práce sme zvolili aplikáciu errors in variables v medicíne, vo výskume rakoviny pŕs.

Kapitola 1

Základné pojmy

Regresia respektívne regres je výraz pochádzajúci z latinského výrazu *regressus*, čo obecne znamená návrat, pohyb späť. S pojmom regresia sa stretávame v rôznych oboroch a oblastiach, keďže regresná analýza má početné možnosti aplikácie. Čiastočný, no určite nie úplný zoznam by zahrnoval financie, ekonómiu, obchod, právo, štatistiku, medicínu a dalo by sa pokračovať ďalšimi vednými obormi. V podstate, sa táto jednoduchá metóda dá využiť pri riešení širokého spektra problémov. S niektorými konkrétnymi aplikáciami sa neskôr budeme zaoberať aj v tejto práci.

Samotný pojem regresie zaviedol do štatistiky britský vedec Francis Galton okolo roku 1880. Názov regresnej analýzy, ktorou sa budeme zaoberať, sa odvodzuje práve od jeho článku: *Regression towards mediocrity in hereditary stature*, z roku 1886 [Gal86].

V ktorom sa zaoberal vyšetrovaním závislosti priemernej výšky potomkov na výške rodičov. Avšak so snahou hľadať vzájomný vzťah náhodných veličín sa stretávame už oveľa skôr, napríklad v prácach Galilea Galileiho, R.J.Boscowitcha, Laplaca či Gaussa.

Regresná analýza sa postupom času stala jednou z najviac rozšírených a používaných metód pre analyzovanie dát. Hlavnou príčinou tejto skutočnosti je najmä to, že sa jedná o relatívne jednoduchú a zrozumiteľnú metódu pre skúmanie funkčného vzťahu medzi premennými. Tento vzťah je

vyjadrený vo forme rovnice alebo modelu, ktoré spájajú takzvané závislé premenné a jednu či viac vysvetľujúcich premenných.

Vzhľadom k tomu, že k tejto téme existuje obrovské množstvo literatúry, je občas dosť ťažké sa vyznať v označení jednotlivých premenných vyskytujúcich sa v teórii regresnej analýzy. Najjednoduchší spôsob ako začať je uvažovať na začiatok vzťah medzi jednou závislou a jednou vysvetľujúcou premennou. Použitie výrazu závislá premenná v tomto prípade znamená, že táto premenná je niakym spôsobom naviazaná na hodnotu vysvetľujúcej premennej.

Teda, že napríklad ak vysvetľujúcu premennú vynásobím vhodným koeficientom, dostanem hodnotu závislej premennej. A práve z tejto úvahy vychádzal už spomenutí Francis Galton pri svojom experimente s meraní synov a otcov. Jeho experiment môžeme zjednodušene popísať nasledovne: uvažujme n rodín, v ktorých máme nameranú výšku otca a syna. Pre prehladnosť začnime s prvou rodinou. Výška otca je pre nám vysvetľujúca premenná, je to náhodná veličina a označíme si ju x_1. Výška syna je pre nám závislá premenná, ktorá je tiež náhodnou veličinou a označme si ju y_1. Vzťah medzi x_1a y_1 potom môžeme popísať rovnicou v tvare:

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1.$$

V tejto rovnici β_0 a β_1 predstavujú koeficienty, prostredníctvom ktorých vyjadrujeme vájomnú závislosť premenných x_1 a y_1. Pri meraní, môže dôjsť k chybe a tento fakt je teda nutné reflektovať zahrnutím náhodnej zložky ε_1 do rovnice.

Rovnakou rovnicou možno vyjadriť vzťah medzi výškou otca a syna v každej zo skúmaných n rodín. Tým sa dostávame k ďalšiemu značeniu. Symbolom Y si označíme náhodný vektor, ktorého zložky sú tvorené náhodnými veličinami $y_1, .., y_n$. Teda je to vektor, ktorý nám udáva namerané výšky synov v jednotlivých rodinách. Obdobne potom označme X náhodný vektor, ktorého zložky sú tvorené náhodnými veličinami $x_1, ..., x_n$ a udáva nám namerané výšky otcov v jednotlivých rodinách. Rovnica závislosti Y a X má potom

tvar:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Tu už ale parametere β_0 a β_1 predstavujú náhodný vektor. Tak isto ε predstavuje už náhodný vektor chýb merania a je tvorený zložkami $\varepsilon_1, \varepsilon_2, .., \varepsilon_n$.

V literatúre zaoberajúcej sa touto problematikou potom môžme nájsť rôzne pomenovania pre premenné. Závislé premenné y_i nájdeme označenú ako regresandy a odozvy. Vysvetľujúce premenné x_i ako prediktory, regresory alebo nezávislé premenné. Avšak názov nezávislé premenné je najmenej preferovaný a to hlavne z dôvodu, že samotné vysvetľujúce premenné sú len málokedy vzájomne nezávislé. Teda tento názov može byť mierne zavádzajúci, a v tejto práci sa mu radšej vyhneme.

1.1 Podklady

Pozrime sa teraz obecne na podklady, s ktorými budeme pracovať v tejto práci. Základom pre nás sú niake dostupné data. Pri skúmaní a vyhodnocovaní týchto dát je vhodné postupovať na základe overenej šablóny, ktorú by sme mohli nazvať krokmi regresnej analýzy. V stručnosti nám totiž poskytuje návod ako postupovať pri výbere dát i modelu, na základe ktorého budeme tieto data vyhodnocovať, a tiež nám sprostredkuje určitý obraz o tom, k akým výsledkom sa na základe zvoleneho modelu môžeme no finále dopracovať.

Kroky regresnej analýzy môžeme rozdeliť do nasledujúcich častí:

1. Určenie problému
2. Selekcia potenciálne dôležitých premenných
3. Zber dát
4. Špecifikácia modelu
5. Potvrdenie modelu a posudok
6. Použitie vybraného modelu pri riešení daného problému

1.1.1 Určenie problému

Základným kameňom regresnej analýzy je formulácia daného problému. To zahŕňa definovanie problému a vymedzenie otázok, ktoré majú byť zodpovedané analýzou. Tento krok je prvý a snád najdôležitejší krok v regresnej analýze. Je tak dôležitý, pretože ním sú silne ovplyvnené aj nasledujúce kroky, a chyby v ňom môžu viesť k výberu irelevantnej množiny premenných alebo k nevhodnej voľbe štatistického modelu analýzy.

1.1.2 Selekcia potenciálne dôležitých premenných

Ďalším krokom po vymedzení problému je výber množiny premenných, ktoré sú potrebné pre vysvetlenie alebo predikciu odozvy y. Je nutné sa zamerať na tie data, ktoré nám skutočne vplývajú na odozvu. To nie je vždy tak jednoduché ako to na prvy pohľad vyzerá.

1.1.3 Zber dát

Zber dát prebieha z prostredia, ktoré je podrobené nášmu štúdiu. Niekedy sú data zozbierané na základe kontrolovaného nastavenia, takže faktory, ktoré nie sú naším primárnym záujmom možno udržiavať na konštantnej úrovni. Menej častý je prístup kedy sú data zozbierané na základe neexperimentálnych podmienok, kedy je len veľmi málo možné zasahovať do ich kontroly. V oboch prípadoch ale zozbierané data pozostávajú z pozorovaní o n zložkách. Každé z týchto n pozorovaní je zaznamenané v tabuľke 1, ktorá má štandardne nasledujúci tvar:

číslo pozorovania	Y	X_1	...	X_p	ε
1	y_1	x_{11}	...	x_{1p}	ε_1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	y_n	x_{n1}	...	x_{np}	ε_n

Tabuľka 1: Štandardný zápis dát

Sťĺpce v tabuľke reprezentujú premennú, zatiaľ čo riadky prislúchajú pozorovaniam, ktoré sú tvorené z množiny $(1 + p)$ hodnôt pre každú zložku. Teda dostávame z toho jednu hodnotu pre odozvu a jednu hodnotu pre každú hodnotu p prediktorov. Pre náš predošlý príklad merania výšky synov by sme teda mali tabuľku zloženú iba zo stľpcov pre Y a X_1.

Zápis prediktora x_{ij} prislúcha napozorovanej i-tej hodnote j-tej vysvetľujúcej premennej resp. prediktoru. Prvý index nám hovorí o poradí pozorovania a druhý o poradí vysvetľujúcej premennej. Na základe tohto môžeme pripísať lineárny regresný model pre i-té pozorovanie v tvare:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i,$$

kde $i = 1, ..., n$ a $j = 1, ..., p$.

1.1.4 Špecifikácia modelu

Podoba modelu, ktorá súvisí s odozvou a jej vzťahom k množine prediktorov, môže byť špecifikovaná na začiatok na základe už v praxi využívaných modeloch, ktoré vznikli na podklade skúseností a znalostí expertov v danom obore nášho štúdia. Predpokladaný model pravdaže môžeme následne modifikovať alebo úplne zmeniť na základe nazberaných dát. Je nutné myslieť na to, že model je v tomto kroku načrtnutý len v obrysoch, pretože stále môže závisieť na neznámych parametroch.

V tomto kroku ďalej potrebujeme určiť podobu funkcie $f(X_1, X_2, ..., X_p)$.

Táto funkcia môže byť klasifikovaná v dvoch typoch, a to ako lineárna alebo nelineárna.
Všetky nelineárne funkcie, ktoré možno transformovať do lineárných funkcíi sa nazývajú linearizovateľné funkcie. Avšak nie všetky nelineárne funkcie sú linearizovateľné. V literatúre sa dokonca niekedy stretávame s názorom, že iba nelineárne funkcie, ktoré sú nelinearizovateľné sú skutočne nelineárne funkcie.

1.1.5 Potvrdenie modelu a posudok

Potom, ako sme definovali model a zozbierali data, je naším ďalším krokom odhad regresných parametrov na základe nazbieraných dát. Metóda, ktorá sa najčastejšie využíva pre odhad parametrov sa nazýva metóda najmenších štvorcov *(least-squares method)*. Za určitých predpokladov dáva táto metóda odhady s požadovanými vlastnosťami. V niektorých príkladoch, ak je jeden alebo viac predpokladov tejto metódy porušených, sa využívajú aj iné metódy ako napríklad metóda maximálnej vierohodnosti *(maximum likelihood method)*, hrebeňová metóda *(ridge method)* alebo metóda hlavných komponent *(principal components method)*.

Následne potom aplikujeme vybranú metódu odhadu parametrov na regresné parametre. Odhady regresných parametrov $\beta_0, \beta_1, ..., \beta_p$ sa označujú $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$. Odhadnutá regresná rovnica má potom tvar:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}. \tag{1.1}$$

Strieška nad parametrom označuje, že sa jedná o odhad parametru a toto označenie odhadu budeme využívať aj v celej práci. Čo sa týka hodnoty $\hat{y}_i$, tú nazývame vyrovnaná hodnota *(fitted value)* i-teho pozorovania.

Správnosť štatistickej metódy, akou je i regresná analýza, závisi na určitých predpokladoch. Predpoklady sa zvyčajne vyslovujú o dátach a o modely. Presnosť rozboru a záverov je odvodená od analýzy zásadne závislej na správnosti týchto predpokladov. Predtým, ako použijeme rovnicu (1.1) za akýmkoľvek

účelom, potrebujeme si najprv overiť či budú špecifické predpoklady platiť. Snažíme sa teda odpovedať na nasledujúce otázky:

1.Aké sú požadované predpoklady?

2.Je každý z týchto predpokladov odôvodnený?

3.Čo budeme robiť ak jeden či viacero predpokladov nebude platiť?

V tomto kroku sa kladie dôraz na to, aby kontrola platnosti predpokladov bola vykonaná predtým ako sú vyvodené akékoľvek závery na základe nášho rozboru. Na regresnú analýzu sa totiž pohliada ako na cyklický proces, ktorého výstupy sa používajú pre diagnostiku, potvrdenie, kritiku a možnú modifikáciu vstupov.

1.1.6 Použitie vybraného modelu pri riešení daného problému

Explicitné určenie regresnej rovnice je najdôležitejší produkt analýzy. Je to zhrnutie vzťahu medzi Y a $X_1, X_2, ..., X_p$. Rovnicu možno použiť pre rôzne účely. Možno ju využiť pri zisťovaní dôležitosti individuálneho prediktora pre analyzovanie efektov metódy, čo zahŕňa meniace sa hodnoty prediktorov, alebo pre predpoveď hodnôt odozvy pre danú množinu prediktorov. Aj keď je regresná rovnica finálny produkt, je tu aj množstvo ďalších vedľajších produktov. Vidíme regresnú analýzu ako množinu analytických techník, ktoré sa používajú pre lepšie pochopenie vnútorného vzťahu medzi premennými a daným prostredím. Úlohou regresnej analýzy je naučiť sa čo najviac je možné o prostredí na základe nahromadených dát. Zdôrazdňuje sa, že to čo je odkryté počas procesu formulácie rovnicu, môže byť niekedy tak cenné a informatívne ako samotná finálna rovnica.

Kapitola 2

Lineárny regresný model

Uvažujme náhodný vektor $Y = (y_1, .., y_n)$, ktorého zložky tvoria nezávislé náhodné veličiny $y_1, .., y_n$ so spojitým rozdelením. Stredné hodnoty týchto veličín možno vyjadriť v tvare:

$$E\ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, \tag{2.1}$$

kde $i = 1, .., n$. Vektor $\beta = (\beta_0, \beta_1.., \beta_k)$ je tvorený parametrami $\beta_0, \beta_1.., \beta_k$, ktoré nazývame regresnými koeficientmi. Jedná sa o neznáme konštanty.

Veličiny x_{ij} sú známe konštanty a možno ich usporiadať do regresnej matice $\mathbf{X}$ typu $(n, k+1)$:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{nk} \end{pmatrix} \tag{2.2}$$

Matica $\mathbf{X}$ je veľmi dôležitá pri odhade vektoru β , kedy konkrétne lineárne kombinácie jej stĺpcov predstavujú strednú hodnotu náhodného vektoru Y, a teda odhad vektoru β.

Ďalej predpokladajme, že platí:

$$Var\ y_i = \sigma^2, \tag{2.3}$$

pre $i = 1, .., n$, kde $\sigma^2 > 0$ je neznámy parameter, ktorý označujeme ako rozptyl náhodných veličín y_i. Model (2.1) teda môžeme prepísať v tvare:

$$Y \sim (\mathbf{X}\beta, \sigma^{\mathbf{2}}\mathbf{I}), \tag{2.4}$$

kde $\mathbf{I}$ je jednotková matica a výraz $\sigma^2\mathbf{I}$ dáva variančnú maticu vektora Y.

$$\mathbf{Var\ Y} = \begin{pmatrix} var\ y_1 & cov\ (y_1, y_2) & ... & cov\ (y_1, y_n) \\ cov\ (y_2, y_1) & var\ y_2 & ... & cov\ (y_2, y_n) \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ cov\ (y_n, y_1) & cov\ (y_n, y_2) & ... & var\ y_n \end{pmatrix} \tag{2.5}$$

Variančná matica v tomto tvare (2.5) potom udáva rovnaký rozptyl a nekorelovanosť jednotlivých zložiek $y_1, .., y_n$ náhodného vektoru Y.

Hodnosť matice $\mathbf{X}$ označme ako $h(\mathbf{X}) = p$. V prípade, že $h(\mathbf{X}) = (k + 1)$ hovoríme, že model (2.4) je model s úplnom hodnosťou. Ak $k = 1$ hovoríme o jednoduchej lineárnej regresíi, kedže uvažujeme len jednu nezávislú premennú x. Ak $k > 1$ hovoríme o mnohonásobnej lineárnej regresíi. Ak $h(\mathbf{X}) < (k + 1)$, potom sa jedná o model s neúplnou hodnosťou. Obecne predpokladáme, že $h(\mathbf{X}) > 0$.

Model (2.4) možno ekvivalentne zapísať aj pomocou náhodnej zložky $\varepsilon \sim (0, \sigma^2\mathbf{I})$ v tvare:

$$Y = \mathbf{X}\beta + \varepsilon, \tag{2.6}$$

kde $\varepsilon = (\varepsilon_1, .., \varepsilon_n)$ je náhodný vektor chýb.

2.1 Odhady parametrov v lineárnom regresnom modeli

Z toho ako sme si zaviedli lineárny regresný model v predošlej podkapitole je jasné, že nás pri skúmaní dát zaújmajú hlavne odhady neznámych parametrov. Teda regresných parametrov $\beta_0, \beta_1.., \beta_k$ odhad rozptylu σ^2 a odhad vektoru

stredných hodnôt $\mathbf{X}\beta$ v modeli (2.4).

Aby sme mohli bližšie popísať akým spôsobom odhadujeme požadované parametre, je vhodné si najprv uviesť niekoľko tvrdení, z ktorých budeme vychádzať.

Pri práci s maticami s neúplnou hodnosťou sa nám určite zíde tzv. *Pravidlo piatich matíc*, ktorého dôkaz môžeme nájsť napríklad v publikácii [And05]. Toto pravidlo môžeme napísať v nasledujúcom tvare:

$$\mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T X} = \mathbf{X} \tag{2.7}$$

Úpravou potom dostávame tvar:

$$(\mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T})\mathbf{X} = \mathbf{0} \tag{2.8}$$

kde označme:

$$\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T} \tag{2.9}$$

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T} \tag{2.10}$$

Odkiaľ máme:

$$\mathbf{HX} = \mathbf{X}$$

$$\mathbf{MX} = \mathbf{0}.$$

Definícia 1 (Idempotentná matica)**.** *Štvorcovú maticu* $\mathbf{A}$ *nazývame idempotentnou, ak platí:* $\mathbf{AA} = \mathbf{A}$. *Hodnosť idempotentnej matica je rovná jej stope, tj.* $h(\mathbf{A}) = tr\ \mathbf{A}$.

Definícia 2 (Lineárny priestor)**.** *Uvažujme pevne dané vektory* $x_1, .., x_k \in R^n$ *tj. do* $n-$*rozmerného euklidovského priestoru. Podpriestor* R^n *obsahujúci práve všetky lineárne kombinácie vektorov* $x_1, .., x_k$ *označení ako:*

$$M(x_1, .., x_k) = \{x \in R^n : x = \sum_{i=1}^{k} \alpha_i x_i,\ \alpha_i \in R^1; i = 1, .., k\}$$

nazývame lineárnym priestorom generovaným vektormi $x_1, .., x_k$. *Ak sú vektory* $x_1, .., x_k$ *stĺpce matice* $\mathbf{X}$ *potom píšeme* $M(\mathbf{X})$.

Definícia 3 (Ortogonálny doplnok)**.** *Nech je niaky lineárny podpriestor priestoru $\Omega \subset R^n$, tj. s každými dvoma prvkami tam patrí tiež ich lineárna kombinácia. Množina:*

$$\Omega^{\perp} = \{x \in R^n : (x, y) = 0,\ \forall y \in \Omega\}$$

je tiež lineárny podpriestor R^n a nazýva sa ortogonálnym doplňkom podpriestoru $\Omega \subset R^n$.

Ďalej platí, že dva lineárne podpriestory Ω_1 a Ω_2 z R^n majú rovnaký ortogonálny doplnok, práve keď sú totožné. Pretože platí:

$$(\Omega^{\perp})^{\perp} = \Omega.$$

Platí, teda že $M(\mathbf{X}\mathbf{X}^{\mathbf{T}}) = M(\mathbf{X})$.

Definícia 4 (Projekcia do podpriestoru)**.** *Nech je Ω podpriestor euklidovského priestoru R^n. Ortogonálnym priemetom vektoru $x \in R^n$ do podpriestoru Ω nazveme taký vektor $r_{\Omega}(x)$, pre ktorý platí že:*

$$\begin{aligned} r_{\Omega}(x) &\in \Omega \\ x - r_{\Omega}(x) &\in \Omega^{\perp}. \end{aligned}$$

Potom platia nasledujúce tvrdenia, ktorých dôkazy nájdeme v [Zvá89].

Tvrdenie 2.1. *:*

Priemet $r_{\Omega}(x)$ je daný jednoznačne.

Tvrdenie 2.2. *:*

Vektor $x - r_{\Omega}(x)$je ortogonálnym priemetom vektora x do podpriestoru $\Omega^{\perp}$.

Tvrdenie 2.3. *:*

Platí:

$$\begin{aligned} x \in \Omega &\leftrightarrow r_\Omega(x) = x \\ x \in \Omega^\perp &\leftrightarrow r_\Omega(x) = 0. \end{aligned} \tag{2.11}$$

Tvrdenie 2.4. *:*

Pre $\forall x \in R^n$ *platí:*

$$||x||^2 = ||r_\Omega(x)||^2 + ||x - r_\Omega(x)||^2. \tag{2.12}$$

Tvrdenie 2.5. *:*

Nech $x \in R^n$*. Ak* $z \in \Omega$ *také, že pre* $\forall y \in \Omega$ *platí:*

$$||x - z||^2 \leq ||x - y||^2 \tag{2.13}$$

potom tiež $z = r_\Omega(x)$ *.*

2.2 Odhad vektora stredných hodnôt

Pozrime sa najprv na odhad vektoru stredných hodnôt $E\ Y$ v modeli $Y \sim (\mathbf{X}\beta, \sigma^\mathbf{2}\mathbf{I})$. Označme si $\mu = \mathbf{X}\beta$. Z definície *1* je zrejmé, že $\mu \in M(\mathbf{X})$. Teda možno vektor odhadnúť μ vektorom $\hat{Y}$ z priestoru $M(\mathbf{X})$, ktorý je k vektoru Y najbližší. Keďže platí tvrdenie (2.13), je $\hat{Y}$ priemetom vektora do $M(\mathbf{X})$. Projekčná maticu si označme $\mathbf{H}$ a platí:

$$\hat{Y} = \mathbf{H}\,Y.$$

Z (2.9) máme, že:

$$\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T}$$

.

Veta 2.1. *:*

V modeli $Y \sim (\mathbf{X}\beta, \sigma^2\mathbf{I})$ *je vektor* $\hat{Y}$ *najlepším nestranným lineárným odhadom vektoru stredných hodnôt* $\mu = \mathbf{X}\beta$*. Je daný jednoznačne a platí:*

$$\mathbf{Var}\ \hat{\mathbf{Y}} = \sigma^2\mathbf{H}.$$

Dôkaz. Keďže projekčná matica $\mathbf{H}$ je idempotentná,nestrannosť $\hat{Y}$ plynie z toho, že $\mathbf{H}$ je projekčná matica na $M(\mathbf{X})$.Potom: $\mathbf{Var}\ \hat{\mathbf{Y}} = \mathbf{Var}\ \mathbf{H}Y = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} = \sigma^2\mathbf{H}$.
Z vlastností idempotentnej matice máme, že: $\mathbf{HX} = \mathbf{X}$. Teda:

$$E\ \hat{Y} = E\ \mathbf{H}Y = \mathbf{HX}\beta = \mathbf{X}\beta = \mu.$$

Overme ešte jednoznačnosť odhadu $\hat{Y}$. Uvažujme niaky lineárny nestranný odhad vektoru μ. Označme ho $\tilde{Y} = \mathbf{A}Y + a$. Pre $\forall\beta$ spĺňa matica $\mathbf{A}$ typu (n, n) rovnosť:

$$E\ \tilde{Y} = \mathbf{AX}\beta + a = \mathbf{X}\beta.$$

Keďže je možné aby vektor $\beta = 0$, musí byť teda $a = 0$ a zároveň musí platiť $\mathbf{AX} = \mathbf{X}$. Teda aj $\mathbf{AH} = \mathbf{H}$.

K porovnaniu nestranných odhadov vektorového parametru používame ich variančné matice. Platí, že ak sú $\hat{Y}$ a $\tilde{Y}$ dva nestranné odhady vektoru μ , potom je odhad $\hat{Y}$ lepší ak je matica $(\mathbf{Var}\ \tilde{Y} - \mathbf{Var}\ \hat{Y})$ pozitívne semidefinitná. Teda, že pre $\forall$ vektor $q \in R^n$ je:$\mathbf{Var}\ (q^T\hat{Y}) \leq \mathbf{Var}\ (q^T\tilde{Y})$.

$$\begin{aligned}
\mathbf{Var}\ \tilde{\mathbf{Y}} &= \mathbf{Var}\ \mathbf{A}Y \\
&= \sigma^2\mathbf{A}\mathbf{A}^\mathbf{T} \\
&= \sigma^2[\mathbf{H} + (\mathbf{A} - \mathbf{H})][\mathbf{H} + (\mathbf{A} - \mathbf{H})] \\
&= \sigma^2\mathbf{H} + \sigma^2(\mathbf{A} - \mathbf{H})(\mathbf{A} - \mathbf{H})^\mathbf{T}, \\
\mathbf{Var}\ \hat{\mathbf{Y}} &= \sigma^2\mathbf{H}.
\end{aligned}$$

□

2.3 Reziduum

Vektor $v = Y - \hat{Y}$ sa nazýva vektor reziduíi resp. reziduum. Porovnáva nám napozorované hodnoty závislej premennej s odhadom jej stredných hodnôt. Z tvrdenia 2 plynie, že je priemetom do priestoru $M(\mathbf{X})^{\perp}$, ktorý nazývame reziduálnym priestorom.

Príslušná projekčná matica má tvar:

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T}.$$

Reziduálny súčet štvorcov označme:

$$\mathbf{RSS} = ||v||^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (y_i - \hat{y_i})^{\mathbf{2}}.$$

Je to štvorec vzdialenosti vektorov Y a $\hat{Y}$, ktorý meria ich nezhodu.

Reziduálny rozptyl potom dostávame ako:

$$\mathbf{S^2} = \frac{\mathbf{RSS}}{n-p}\ .$$

Veta 2.2 (Veta o reziduách)**.** *:*

V lineárnom modeli $Y \sim (\mathbf{X}\beta, \sigma^{\mathbf{2}}\mathbf{I})$ *platí:*

(i) $v = \mathbf{M}\,Y = \mathbf{M}\varepsilon$

(ii) $v \sim (\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{M})$

(iii) $\mathbf{RSS} = \varepsilon^{\mathbf{T}}\mathbf{M}\varepsilon$

(iv) $E\ \mathbf{RSS} = (n-p)\sigma^{\mathbf{2}}$

(v) $E\ \mathbf{S^2} = \sigma^2$

(vi) $\mathbf{X^T}v = \mathbf{0}$.

Dôkaz. Tvrdenie *1* vyplýva z toho, že $v = \mathbf{M}Y = \mathbf{M}(\mathbf{X}\beta + \varepsilon) = \mathbf{M}$, pretože platí: $\mathbf{MX} = \mathbf{0}$.

Keďže $\varepsilon \sim (0, \sigma^2\mathbf{I})$ a $v = Y - \hat{Y}$, teda $v = \mathbf{M}Y \sim (\mathbf{0}, \sigma^2\mathbf{M})$, čím sme dostali tvrdenie *2*.

Pre tvrdenie *3* vychádzame z toho, že $\mathbf{RSS} = \varepsilon^\mathbf{T}\mathbf{M}\varepsilon$,kde $\mathbf{M} = \mathbf{M^T}\mathbf{M}$ je idempotentná matica.
Potom $\mathbf{RSS} = ||v||^\mathbf{2} = ||\mathbf{M}\varepsilon||^\mathbf{2} = (\mathbf{M}\varepsilon)^\mathbf{T}(\mathbf{M}\varepsilon) = \varepsilon^\mathbf{T}\mathbf{M^T}\mathbf{M}\varepsilon = \varepsilon^\mathbf{T}\mathbf{M}\varepsilon$.

Tvrdenie *4* dostaneme úpravou: $E\ \mathbf{RSS} = E\ ||v||^\mathbf{2} = E\ v^T v = tr\ E\ v^T v = trVar\ v = \sigma^\mathbf{2} tr\mathbf{M} = \sigma^\mathbf{2} tr(\mathbf{I} - \mathbf{H}) = \sigma^\mathbf{2}(n - p)$.
Keďže platí, že: $h(\mathbf{H}) = h(\mathbf{X})$ a stopa tr idempotentnej matice je rovná jej hodnosti.

Pre tvrdenie *5* potom máme: $E\ \mathbf{S^2} = E\ (\frac{\mathbf{RSS}}{n-p}) = \frac{E\ \mathbf{RSS}}{(n-p)} = \sigma^\mathbf{2}\frac{n-p}{n-p} = \sigma^\mathbf{2}$.

Tvrdenie *6* plynie z toho, že: $\mathbf{MX} = \mathbf{0}$ a $v = \mathbf{M}\varepsilon$.

□

2.3.1 Odhad vektora regresných koeficientov β

Pri odhade vektora regresných parametrov $\beta = (\beta_0, \beta_1.., \beta_k)$ vychádzame z hodnosti matice $\mathbf{X}$. Odhad vektoru β totiž vyjadruje strednú hodnotu náhodného vektora Y ako konkrétnu lineárnu kombináciu stĺpcov matice $\mathbf{X}$.

Ak má matica $\mathbf{X}$ plnú hodnosť, potom možno každý vektor z $M(\mathbf{X})$ vyjadriť pomocou lineárnej kombinácie stĺpcov matice $\mathbf{X}$ nekonečne mnoho spôsobmi.

Jedna z obecne najviac užívaných metód pri odhade regresných parametrov je metóda najmenších štvorcov. Za jej vznikom stáli nezávisle na sebe Carl Friedrich Gauss(1795) a Adrien Marie Legendre(1805).

Prvotné aplikácie tejto metódy boli v astronómii a geodetike. S prvým verejným výtiskom sa stretávame v roku 1805 v Legendreho knihe o určovaní obežných dráh komét.

Pre zjednodušenie si uveďme príklad z praxe, ktorý môžme nájsť v knihe Alternative methods of regression od D. Birkes,Y.Dodge [Dod93].

Pri experimente sa skúmal vzťah medzi dvoma procedúrami merania obsahu kyseliny v chemikálii. Prvá metóda, založená na extrakcii a vážení bola drahšia. Zatiaľ čo druhá, založená na titrácii, je oveľa lacnejšia.

Základom pokusu teda bolo zistiť vzájomný vzťah týchto dvoch metód použitím regresie. Pričom ak by vzťah existoval, znamenalo by to v budúcnosti nahradenie drahšej procedúry tou lacnejšou.

Obe procedúry boli použité na 20-tich vzorkách chemikálie, čím sa dospelo k nameraným hodnotám kyseliny v tabuľke 2.

Hodnota y_i predstavuje výsledok i-teho merania drahej metódy a hodnota x_i predstavuje výsledok i-teho merania lacnej metódy merania obsahu kyseliny. Použitím týchto dát sa snažíme získať rovnicu, ktorá by vyjadrovala vzájomný vzťah oboch procedúr.
Pozrime sa najprv na lineárny regresný model, daný rovnicou:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Pre tento prípad uvažujme len jednu vysvetľujúcu premennú. Model s týmto nastavením potom označujeme ako jednoduchý lineárny regresný model.

číslo vzorky	Drahá procedura (Y)	Lacná procedúra (X)
1	76	123
2	70	109
3	55	62
4	71	104
5	55	57
6	48	37
7	50	44
8	66	100
9	41	16
10	43	28
11	82	138
12	68	105
13	88	159
14	58	75
15	64	88
16	88	164
17	89	169
18	88	167
19	84	149
20	88	167

Tabuľka 2: obsah kyseliny

Rovnica je v tvare:

$$Y = \beta^T X + \varepsilon.$$

Ďalším krokom je určenie odhadov pre regresné koeficienty β_0 a β_1. Keď následne použijeme lacnejšiu procedúru X na zmeranie množstva kyseliny vo vzorku, potom môžeme pre drahšiu procedúru Y písať odhad i-teho merania v tvare:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i.$$

Pre odhad paremetrov β_0 a β_1 použijeme metódu najmenších štvorcov.

Uvažujeme nameranú závislosť ako súbor dvojíc $(x_1, y_1); (x_2, y_2); ...; (x_n, y_n)$.

Týmito bodmi prekladáme funkciu v tvare $y = f(x, \beta_0, \beta_1, ..., \beta_p)$, pričom hľadáme štatistické odhady $\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_p}$ jej parametrov $\beta_0, \beta_1, ..., \beta_p$.

Základné kritérium je, aby bol súčet štvorcov odchýliek empirických hodnôt y_i, označujeme ho S, minimálny od vyrovnaných hodnôt $y = f(x_i, \beta_0, \beta_1, ..., \beta_p)$, tj.od teoretickej regresnej funkcie. Teda mala by platiť rovnosť:

$$S(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} (y_i - f(x_i, \beta_0, ..., \beta_p))^2$$

$$\widehat{\beta} = agrmin_\beta \sum_{i=1}^{n} (y_i - f(x_i, \beta_0, ..., \beta_p))^2$$

Kedže uvažujeme lineárný regresný model, obmedzíme si tvar funkcie f na lineárnu regresnú funkciu v tvare:

$$f(x_i, \beta_0, ..., \beta_p) = \beta_0 f_0(x) + \beta_1 f_1(x) + \beta_2 f_2(x) + ... + \beta_p f_p(x)$$

Pre existenciu minima funkcie S je nutná podmienka:

$$\frac{\partial S}{\partial \beta_j} = 0$$

pre $j = 0, 1, ..., p$.

Pre parameter β_j dostávame:

$$\sum_{i=1}^{n} f_j(x_i) f_0(x_i) \beta_0 + ... + \sum_{i=1}^{n} f_j(x_i) f_p(x_i) \beta_p = \sum_{i=1}^{n} y_i f_j(x_i)$$

kde označíme:

$$\sum_{i=1}^{n} f_j(x_i) f_h(x_i) = a_{jh}$$

$$\sum_{i=1}^{n} y_i f_j(x_i) = a_j$$

Dostávame sústavu rovníc:

$$\begin{aligned} a_{00}\beta_0 + a_{01}\beta_1 + ... + a_{0p}\beta_p &= a_0 \\ a_{10}\beta_0 + a_{11}\beta_1 + ... + a_{1p}\beta_p &= a_1 \\ &\vdots \\ a_{p0}\beta_0 + a_{p1}\beta_1 + ... + a_{pp}\beta_p &= a_p \end{aligned}$$

Jej vyriešením dostávame hľadané odhady $\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_p}$.

Vráťme sa teraz späť k nášmu meraniu hodnôt kyseliny. Odhady $\widehat{\beta_0}$ a $\widehat{\beta_1}$ majú potom tvar:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

a

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}.$$

Kde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ predstavujú priemery.

Pre nami namerané data potom dostávame hodnoty odhadov $\widehat{\beta_0} = 35.46$ a $\widehat{\beta_1} = 0.3216$. Vhodná regresná priamka pre má experiment má teda tvar:

$$\widehat{y_i} = 35.46 + 0.3216x_i.$$

Označme si $\hat{\beta}$ odhad vektora regresných parametrov β. Z definície priemetu

platí, že $\mathbf{X}\hat{\beta} \in M(\mathbf{X})$. V zhľadom k predošlému, nám $\hat{\beta}$ predstavuje ľubovoľné riešenie sústavy $\mathbf{X}\hat{\beta} = \hat{Y}$. Čo možno prepísať ako:

$$\mathbf{X}\hat{\beta} + v = Y, \tag{2.14}$$

keďže $v = Y - \hat{Y}$.

Z vety o reziduách máme, že $\mathbf{X^T}v = \mathbf{0}$, teda (2.14) možno prepísať v tvare $\mathbf{X^T}(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$. To je ekvivalentné s normálnou rovnicou :

$$\mathbf{X^T}\mathbf{X}\hat{\beta} = \mathbf{X^T}\,Y. \tag{2.15}$$

Odkiaľ dostávame vyjadrenie pre $\hat{\beta}$ v tvare:

$$\hat{\beta} = (\mathbf{X^T}\mathbf{X})^{-}\mathbf{X^T}\,Y. \tag{2.16}$$

Riešenie ale nemusí byť jednoznačné, záleží na hodnosti matice $\mathbf{X}$.

Veta 2.3. *: V modeli $Y \sim (\mathbf{X}\beta, \sigma^2\mathbf{I})$ s úplnou hodnosťou matice $\mathbf{X}$ je vektor $\hat{\beta} = (\mathbf{X^T}\mathbf{X})^{-}\mathbf{X^T}\,Y$ jediným riešením normálnej rovnice $\mathbf{X^T}\mathbf{X}\hat{\beta} = \mathbf{X^T}\,Y$. Vektor $\hat{\beta}$ je najlepším nestranným lineárným odhadom vektoru regresných koeficientov β.*

Dôkaz. Z vlastností otrogonálneho doplnku platí, že $M(\mathbf{X}\mathbf{X^T}) = M(\mathbf{X})$. To nám dáva $h(\mathbf{X}) = \mathbf{h}(\mathbf{X}\mathbf{X^T})$, teda je $\mathbf{X^T}\mathbf{X}$ regulárna matica v modeli s úplnou hodnosťou.
Normálna rovnica má jediné riešenie v tvare $\hat{\beta} = (\mathbf{X^T}\mathbf{X})^{-}\mathbf{X^T}\,Y$ a vektor $\hat{\beta}$ je jednoznačnou lineárnou funkciou $\hat{Y}$, pretože platí: $\hat{\beta} = (\mathbf{X^T}\mathbf{X})^{-}\mathbf{X^T}\,\hat{Y}$.
Ďalej podľa vety 1 je najlepším nestranným lineárnym odhadom svojej strednej hodnoty.

□

Veta 2.4. *Pre momenty $\hat{\beta}$ platí $E\ \hat{\beta} = \beta$ a $Var\ \hat{\beta} = \sigma^2(\mathbf{X^T}\mathbf{X})^{-}$.*

Dôkaz. Keďže predpokladáme, že $Y \sim (\mathbf{X}\beta, \sigma^{\mathbf{2}}\mathbf{I})$ a $h(\mathbf{X}) = p$, z predošlej vety 2.3 teda možno písať: $\hat{\beta} = (\mathbf{X^T X})^{-}\mathbf{X^T}\, Y$.
Teda:

$$E\ \hat{\beta} = E\ (\mathbf{X^T X})^{-}\mathbf{X^T}\, Y = (\mathbf{X^T X})^{-}\mathbf{X^T X}\beta = \beta,$$

a $Var\ \hat{\beta} = Var\ (\mathbf{X^T X})^{-}\mathbf{X^T}\, Y = \sigma^{\mathbf{2}}(\mathbf{X^T X})^{-}$. □

Veta 2.5. *Nech* $Y \sim (\mu, \mathbf{W}^{-})$ *, kde* $\mathbf{W}$ *typu* (n, n) *je pozitívne definitná matica. Nech matica* $\mathbf{A}$ *typu* (n, n) *je pozitívne semidefinitná matica, taká že platí* $\mu\mathbf{A} = \mathbf{0}$ *a* $h(\mathbf{A}) = r > 0$*. Potom platí,že ak je matica* $\mathbf{AW}^{-}$ *idempotentná:*

$$Y^T\mathbf{A}\, Y \sim \chi^{\mathbf{2}}_r.$$

Veta 2.6. *Nech* $Y \sim (\mu, \mathbf{W}^{-})$ *, kde* $\mathbf{W}$ *typu* (n, n) *je pozitívne definitná matica. Nech* $\Omega \subset R_n$ *a* $\mu \in \Omega^{\perp}_{\mathbf{W}}$*. Ak dim* $\Omega = p > 0$ *potom platí:*

$$||\mathbf{P_\Omega}\, Y||^{\mathbf{2}}_{\mathbf{W}} \sim \chi^{\mathbf{2}}_p.$$

Dôkaz. Z definície idempotentnej matice vieme, že projekčná matica $\mathbf{P}$ je idempotentná a matica $\mathbf{WP_\Omega}$ je symetrická. Teda možno písať:

$$||\mathbf{P_\Omega}\, Y||^{\mathbf{2}} = Y^T\mathbf{P^T_\Omega W P_\Omega}\, Y = Y^T\mathbf{WP_\Omega}\, Y$$

Matica $\mathbf{WP_\Omega}$ spĺňa predpoklady vety 2.5 , teda štvorec dĺžky priemetu $\mathbf{P_\Omega}\, Y$ pri skalárnom súčine $(.,.)_{\mathbf{W}}$ má rozdelenie χ^2_p. □

Veta 2.7. *Nech* $Y \sim (\mu, \mathbf{W}^{-})$ *,* Ω_1, Ω_2 *sú podpriestory* R^n *, pričom platí že:*

$$\Omega_1 \subset \Omega^{\perp}_{2\mathbf{W}}.$$

Ak sú $\mathbf{P_1}, \mathbf{P_2}$ *projekčné matice na podpriestory* Ω_1, Ω_2 *pri skalárnom súčine* $(.,.)_{\mathbf{W}}$ *, potom platia tvrdenia:*

(i) $\mathbf{P_1}\, Y, \mathbf{P_2}\, Y$ *sú nezávislé náhodné vektory*

(ii) $||\mathbf{P_1}Y||^{\mathbf{2}}_{\mathbf{W}}$ *a* $||\mathbf{P_2}Y||^{\mathbf{2}}_{\mathbf{W}}$ *sú nezávislé náhodné veličiny.*

Dôkaz. Tvrdenie *1* znamená, že pre ľubovoľné $x, z \in R^n$ platí nasledujúce:

$$\mathbf{0} = (\mathbf{P_1}x, \mathbf{P_2}z)_{\mathbf{W}} = x^T\mathbf{P_1^T W P_2}z.$$

Teda je $\mathbf{P_1^T W P_2} = \mathbf{0}$. S využitím symetrie projekčných matíc máme symetriu matíc $\mathbf{WP_1}$, $\mathbf{WP_2}$ a:

$$\mathbf{0} = \mathbf{P_1^T W P_2} = \mathbf{P_1^T W W^- W P_2} = \mathbf{W P_1 W^- P_2^T W}.$$

Keďže $\mathbf{W}$ je regulárna, platí aj $\mathbf{P_1 W^- P_2^T} = \mathbf{0}$. Teda náhodné vektory $\mathbf{P_1}Y, \mathbf{P_2}Y$ sú nezávislé.

Tvrdenie *2* je priamym dôsledkom tvrdenia *1*.

□

Definícia 5 (Overiteľná lineárna hypotéza)**.** *:*

Nech pre model $Y \sim (\mathbf{X}\beta, \sigma^{\mathbf{2}}\mathbf{I})$ *je* $\mu^0 \in M(\mathbf{X})$ *pevne zvolený vektor,* $\omega \subset M(\mathbf{X})$ *je pevne zvolený podpriestor. Potom požiadavok*

$$E\ Y - \mu^0 \in \omega \tag{2.17}$$

sa nazýva overiteľná lineárna hypotéza.

Veta 2.8. *Najlepší nestranný lineárny odhad vektora stredných hodnôt* $E\ Y$ *v modeli* $Y \sim (\mathbf{X}\beta, \sigma^{\mathbf{2}}\mathbf{I})$ *za platnosti overiteľnej lineárnej hypotézy (2.17) je rovný:*

$$\hat{Y_h} = \mu^0 + P_\omega(Y - \mu^0).$$

Dôkaz. Pretože je $\mu^0 \in M(\mathbf{X})$, teda existuje $\beta \in R_k$ také, že $\mu^0 = \mathbf{X}\beta^{\mathbf{0}}$. Označme si $z = Y - \mu^0$ a $\gamma = \beta - \beta^0$. Potom je $z \sim (\mathbf{X}\gamma, \sigma^{\mathbf{2}}\mathbf{I})$. Ak platí overiteľná lineárna hypotéza (2.17) , je $E\ z \in \omega \subset M(\mathbf{X})$. Teda najlepší nestranný lineárny odhad pre $E\ z$ je:

$$\hat{z_h} = P_\omega z = P_\omega(Y - \mu^0).$$

Keďže $\hat{Y_h} = \hat{z_h} + \mu^0$, potom je $E\ Y$ najlepší nestranný lineárny odhad za platnosti hypotézy. □

Veta 2.9. *Ak je* $E\ Y - \mu^0 \in \omega$ *overiteľná lineárna hypotéza v modeli* $Y \sim (\mathbf{X}\beta, \sigma^2\mathbf{I})$ *potom platia nasledujúce tvrdenia:*

(i) $\hat{Y_h} = \hat{Y} - d$

(ii) $v_h = v + d$

(iii) $d^T v = 0$

(iv) $\mathbf{RSS_h} = \mathbf{RSS} + ||d||^2$,

kde vektor d *je priemetom vektora* $Y - \mu^0$ *do* $M(\mathbf{X}) \cap \omega^\perp$, *a* $\mathbf{d} = dim\ M(\mathbf{X}) - dim\ \omega = dim\ M(\mathbf{X}) \cap \omega^\perp$.

Dôkaz. Uvažujme znovu $z = Y - \mu^0$ a $\gamma = \beta - \beta^0$, $\beta \in R_k$ také, že $\mu^0 = \mathbf{X}\beta^\mathbf{0}$ a $z \sim (\mathbf{X}\gamma, \sigma^2\mathbf{I})$. Označme si $\Omega = M(\mathbf{X})$. Keďže P je projekčná matica, z jej vlastností platí, že:

$$\hat{z} - \hat{z_h} = (P_\Omega - P_\omega)z = (P_{\Omega \cap \omega^\perp})z.$$

Keďže $z = Y - \mu^0$, potom dostávame:

$$\hat{Y} - \mu^0 - (\hat{Y_h} - \mu^0) = (P_{\Omega \cap \omega^\perp})(Y - \mu^0) = d.$$

tj. tvrdenie *1*. Keď ho odčítame od identity $Y = Y$, dostávame tvrdenie *2*. Tvdenie *3* dostaneme ako dôsledok toho, že $d \in \Omega \cap \omega^\perp$ a $v \in \Omega^\perp$. Potom ale možme písať:

$$\mathbf{RSS_h} = ||v_h||^\mathbf{2} = ||v + d||^\mathbf{2} = ||v||^\mathbf{2} + ||d||^\mathbf{2} = \mathbf{RSS} + ||d||^\mathbf{2}.$$

čo je tvrdenie *4*. □

Veta 2.10. *Uvažujme konzistentnú sústavu lineárnych rovníc v tvare:*

$$\mathbf{A}\beta = a, \tag{2.18}$$

kde $\mathbf{A}$ *je daná matica typu* (t,k) *a* $a \in R_t$ *je daný vektor.*
Potom omedzenie (2.18) je overiteľná lineárna hypotéza v modeli $Y \sim (\mathbf{X}\beta, \sigma^2\mathbf{I})$ *práve keď platí:*

$$M(\mathbf{A^T}) \subset M(\mathbf{X^T}). \tag{2.19}$$

Potom pre ľubovolné riešenie $\hat{\beta}$ *normálnej rovnice platí:*

$$||d||^2 = (\mathbf{A}\hat{\beta} - a)^T[\mathbf{A}(\mathbf{X^T X})^-\mathbf{A^T}]^-(\mathbf{A}\hat{\beta} - a)$$

a

$$\mathbf{d} = h(\mathbf{A}).$$

Veta 2.11. *Ak* $Y \sim (\mathbf{X}\beta, \sigma^2\mathbf{I})$ *je model s úplnou hodnosťou, potom je každá lineárna hypotéza (2.18) overiteľná. Ak má naviac matica* $\mathbf{A}$ *lineárne nezávislé riadky, platí:*

(i) $\hat{\beta}_h = \hat{\beta} - (\mathbf{X^T X})^-\mathbf{A^T}[\mathbf{A}(\mathbf{X^T X})^-\mathbf{A^T}]^-(\mathbf{A}\hat{\beta} - a)$

(ii) $\mathbf{RSS_h} = \mathbf{RSS} + (\hat{\beta} - \hat{\beta_\mathbf{h}})^\mathbf{T}\mathbf{X^T X}(\hat{\beta} - \hat{\beta_\mathbf{h}})$.

Dôkaz. Ak $p = k$, potom platí $M(\mathbf{X^T}) = R_k$ a triviálne platí $M(\mathbf{A^T}) \subset M(\mathbf{X^T})$. Keďže predpokladáme hodnosť matice $h(\mathbf{A}) = t$, má sústava $\mathbf{A}\beta = a$ vždy riešenie. Vzhľadom k lineárnej nezávislosti riadkov matice $\mathbf{A}$ a pozitívnej definitnosti matice $(\mathbf{X^T X})^-$ dostávame aj pozitívnu definitnosť a regularitu matice $\mathbf{A}(\mathbf{X^T X})^-\mathbf{A^T}$. Dosadením do $\hat{Y_h} = \hat{Y} - d$ a dostávame:

$$\hat{Y_h} = \mathbf{X}\hat{\beta} - \mathbf{X}(\mathbf{X^T X})^-\mathbf{A^T}[\mathbf{A}(\mathbf{X^T X})^-\mathbf{A^T}]^-(\mathbf{A}\hat{\beta} - a) = \mathbf{X}\hat{\beta_h}.$$

tj. tvrdenie *1*.
A ďalej využitím predošlej vety môžeme písať:

$$\begin{aligned} \mathbf{RSS_h} - \mathbf{RSS} &= ||d||^\mathbf{2} \\ &= (\mathbf{A}\hat{\beta} - a)^-[\mathbf{A}(\mathbf{X^T X})^-\mathbf{A^T}]^-(\mathbf{A}\hat{\beta} - a) \\ &= (\hat{\beta} - \hat{\beta_h})^- X^T X(\hat{\beta} - \hat{\beta_h}). \end{aligned}$$

□

2.4 Normálný lineárny model

Predpokladajme, že Y má mnohorozmerné normálne rozdelenie, tj. $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Pre rozdelenie vektor chýb ε, potom dostávame, že $\varepsilon \sim N(0, \sigma^2\mathbf{I})$.

Veta 2.12. *Nech* $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$*, potom platia nasledujúce tvrdenia:*

(i) $\hat{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{H})$

(ii) $v \sim N(\mathbf{0}, \sigma^2\mathbf{M})$

(iii) $\hat{Y}$ *a* v *sú nezávislé náhodné vektory*

(iv) **RSS** *nezávisí na* $\hat{Y}$

(v) $\frac{\mathbf{RSS}}{\sigma^2} \sim \chi^2_{n-p}$

Dôkaz. Tvrdenie *1* je trviálnym dôsledkom rozdelenia $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, ktoré predpokladáme. Keďže $\hat{Y} = \mathbf{H}Y$, platí $\hat{Y} \sim N(\mathbf{HX}\beta, \sigma^2\mathbf{HH})$. Z faktu, že $\mathbf{H}$ je projekčná matica na $M(\mathbf{X})$ dostávame $\mathbf{HH} = \mathbf{H}$ a $\mathbf{HX} = \mathbf{X}$.

Pri tvrdení *2* stačí využiť faktu, že matica $\mathbf{M}$ sa premieta na $M(\mathbf{X})^\perp$, a teda platí $\mathbf{MX} = \mathbf{0}$.

Pre tvrdenie *3* a tvrdenie *4* využijeme vetu 2.7 , kedy tvrdenie *3* z nej priamo plynie a tvrdenie *4* je jej bezprostredný dôsledok.

tvrdenie *5* je vlastne upravou vety 2.6 , kedy miesto $\mathbf{W}^-$ použijeme $\sigma^2\mathbf{I}$.

□

Veta 2.13. *V normálnom lineárnom modele* $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ *s úplnom hodnosťou platí pre odhad vektora regresných koeficientov:*

(i) $\hat{\beta} \sim N(\beta, \sigma^2\mathbf{C})$

(ii) $t_i = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim t_{n-p}$,

pre $i = 1, .., k$ *a* $\mathbf{C} = (\mathbf{X^T X})^-$.

Veta 2.14. *Nech normálny lineárny model* $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ *má ľubovolné riešenie normálnej rovnice* $\hat{\beta}$. *Potom štatistika:*

$$F = \frac{||\hat{Y} - \mathbf{X}\beta||^2}{\mathbf{RSS}} \frac{n-p}{p} \tag{2.20}$$

má Fischer-Snedecorovo rozdelenie $F_{p,n-p}$.

Na základe poslednej vety možno v normálnom modeli s úplnou hodnosťou testovať hypotézy o hodnote celého vektoru β.

Uvažujme náhodný výber $(X_1, .., X_n)$, tj. postupnosť nezávislých náhodných veličín, resp. vektorov, z rozsahu n s rovnakým rozdelením charakterizovaným distribučnou funkciou.

Definícia 6 (Intervalový odhad)**.** *:*

Interval $(D(X_1, .., X_n), H(X_1, .., X_n))$ *nazveme intervalovým odhadom* θ *o spoľahlivosti* $1 - \alpha$ *ak:*

$$P_\theta[D(X_1, .., X_n) < \theta < H(X_1, .., X_n)] = 1 - \alpha,$$

tj. intervalový odhad pokrýva skutočnú hodnotu θ *s pravdepodobnosťou* $1 - \alpha$.

Krajné body intervalu spoľahlivosti sú funkciami náhodného výberu $(X_1, .., X_n)$, sú to teda náhodné veličiny.

Nech β^0 je vopred daný vektor. Hypotézu $\beta = \beta^0$ zamietame na hladine α v prospech alternatívy $\beta \neq \beta^0$, práve keď pre F z vety 2.14 platí:

$$F > F_{k,n-p}(1-\alpha),$$

kde $F_{k,n-k}(1-\alpha)$ je $(1-\alpha)-$kvantil rozdelenia $F_{k,n-k}$. Ak hypotézu $\beta = \beta^0$ nezamietame, potom množina všetkých takýchto vektorov β^0 vytvorí konfidenčnú množinu pre β s koeficientom spoľahlivosti $1-\alpha$ v tvare:

$$K_2 = \beta \in R_k : (\beta - \hat{\beta})^T \mathbf{X^T X} (\beta - \hat{\beta}) \leqq k s^2 F_{k,n-p}(1-\alpha). \tag{2.21}$$

Veta 2.15. *Uvažujme overiteľnú lineárnu hypotézu:*

$$E\ Y - \mu^0 \in \omega \tag{2.22}$$

v modeli $Y \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ *a nech platí* $\mathbf{d} = p - dim\ \omega > 0$. *Potom za platnosti (2.22) má náhodná veličina*

$$F = \frac{\mathbf{RSS_h} - \mathbf{RSS}}{\mathbf{RSS}} \frac{n-p}{\mathbf{d}}$$

rozdelenie $F_{d,n-p}$, *náhodná veličina*

$$B = \frac{\mathbf{RSS_h} - \mathbf{RSS}}{\mathbf{RSS_h}}$$

má rozdelenie $B_{d/2,(n-p)/2}$ *a náhodná veličina*

$$B^* = \frac{\mathbf{RSS}}{\mathbf{RSS_h}}$$

má rozdelenie $B_{(n-p)/2,d/2}$.

Kapitola 3

Errors in variables

V jednoduchosti možno povedať, že EIV modely sú regresné modely u ktorých sú regresory pozorované s chybami. Navzdory tejto pomerne zjednodušenej definícii, je možné za EIV modelmi vidieť oveľa širšiu škálu. Táto problematika je teda prirodzene podrobne skúmaná viacerými autormi. Avšak medzi najznámejších autorov patria napríklad Wayne A. Fuller [Ful87], Charles E. McCulloch [CE01] alebo Annette J. Dobson [Dob02].

Modely delíme priamo na linearne EIV modely *(lienar EIV models)*, nelineárne EIV modely *(nonlinear EIV models)* a čiastočne lineárne modely *(partially linear EIV models)* respektívne v niektorej literatúre ich nájdeme pod anglickým názvom *(partly linear EIV models)*. Jednotlivé modely následne možno rozčleniť na základe informácii o správaní sa jednotlivých vstupov. Čím dostávame relatívne zložitú štruktúru tohoto celku regresnej analýzy.

3.1 Lineárne EIV modely

Pri skúmaní jednotlivých častí, začneme najprv modelom, ktorý intuitívne naväzuje na doposiaľ uvedené informácie. Prirodzene sa teda dostávame k lineárnemu EIV modelu, ktorý možno chápať ako priamu nadstavbu predošlých kapitol.

Lineárny EIV model pre jednu závislú premennú je definovaný pre i-te pozorovanie ako:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$$w_i = x_i + u_i.$$

Kedže zápis skúmaných dát si možno predstaviť v tvare, ktorý udáva nasledujúca tabuľka:

číslo pozorovania	Y	X	ε	U	W
1	y_1	x_1	ε_1	u_1	w_1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	y_n	x_n	ε_n	u_n	w_n

Tabuľka 2

Prirodzene teda dostávame:

$$Y = \beta^T X + \varepsilon,$$

$$W = X + U,$$

kde Y je vektor závislej premennej, vektor U je náhodný člen respektívne chyba merania a X je vektor regresorov (kovariantov). V tomto modele sú regresori x_i pozorované s chybou, teda je priamo pozorovaná iba premenná w_i nazývaná tiež *prokázateľná premenná (manifest variable)*. Regresory x_i sa označujú aj ako *latentná premenná (latent variable)*.

Obecne možno lineárne EIV modely rozdeliť aj podľa toho ako sa prejavujú regresori x_i. V prípade, že x_i sú pevné, tj. konštanty, hovoríme o funkcionálnych modeloch *(functional models)*. Ak uvažujeme x_i náhodné,

dostávame takzvané štrukturálne modely *(structual models)*.

Maticovo možno model prepísať:

$$Y = \mathbf{X}\beta + \varepsilon,$$

$$\mathbf{W} = \mathbf{X} + \mathbf{U},$$

kde $\mathbf{X}, \mathbf{U}, \mathbf{W}$ sú matice v tvare:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix} \tag{3.1}$$

$$\mathbf{U} = \begin{pmatrix} 0 & u_{11} & ... & u_{1p} \\ 0 & u_{21} & ... & u_{2p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 0 & u_{n1} & ... & u_{np} \end{pmatrix} \tag{3.2}$$

$$\mathbf{W} = \begin{pmatrix} 1 & w_{11} & ... & w_{1p} \\ 1 & w_{21} & ... & w_{2p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & w_{n1} & ... & w_{np} \end{pmatrix} \tag{3.3}$$

3.2 Nelineárne EIV modely

Nelineárny EIV model je definovaný ako:

$$E(Y|X_t) = g(X_t),$$

$$W_t = X_t + U_t.$$

V tomto modely predpokladáme, že vzťah medzi Y a X_t je nelineárny a regresory X_t sú namerané s chybami. Závislú premennú označujeme Y, X_t je premenná pozorovaná s chybou, Z_t je premenná nameraná bez chyby a W_t je *prokazateľná premenná(manifest variable).*

3.3 Čiastočne lineárne EIV modely

Čiastočne lineárny EIV model je definovaný ako:

$$Y = X_i^T \beta + g(T_i) + \varepsilon_i, \tag{3.4}$$

kde X_i predstavujú vektor kovariátov, T_i sú skalárne kovariáty, funkcia $g(.)$ je neznáma a chyby modelu ε_i sú nezávislé s podmienenou hustotou 0 danou kovariátmi. Počet pozorovaní pre daný model nech je n. Čiastočne lineárny model je predstavený napríklad v článku od autorov Engle, Granger, Rice a Weiss [RE86]. Kde sa jeho tvorcovia zaoberajú súvislosťou aký má vplyv počasie na poptávku po elektrine. Ďalšie zaújmavé stúdie možno nájsť v publikáciách od Heckmanna [E.N86], Chena [Che88], Speckmanna [Spe88], Lianga a Härdla [HL97] a Severiniho a Staniswalissa [A.T94].

Pri skúmaní modelu nás primárne zaújma odhad neznámeho parametru β a neznámej funkcie $g(.)$ v modely (3.4), kde kovariáty X_i sú merané s chybami merania. Namiesto priameho pozorovania hodnôt X_i teda pozorujeme hodnoty:

$$W_i = X_i + U_i. \tag{3.5}$$

Pozorované hodnoty W_i pozostávajú teda z dvoch častí, a to z chýb merania, ktoré označime ako U_i a z kovariátov X_i. Pre uvedené hodnoty U_i platí, že sú nezávislé a stejne rozdelené. Teda nezávislé na (Y_i, X_i, T_i), s priemerom 0 a kovariančnou maticou Σ_{uu}. Predpokladajme, že kovariančná matica Σ_{uu} je pre nás známa. Ak sú hodnoty X_i pozorovateľné, potom by sme štandardne dostali hodnoty odhadu parametru β. Pre každé pevné β, nech je $\hat{g}(T, \beta)$ odhadom pre $g(T)$. Potom napríklad v implamentácii použitej v práci Severiniho a Staniswalissa [A.T94], $\hat{g}(T, \beta)$ maximalizuje vážený pravdepodobnostný odhad a chyby modelu ε_i majú normálne rozdelenie s váhami danými pomocou jádrových váh zo symetrického jádrového odhadu hustoty $K(.)$ a šírkou pásma h. Keďže máme $\hat{g}(T, \beta)$ je možné odhadnúť β pomocou metódy najmenších štvorcov ako:

$$min \sum_{i=1}^{n} [Y_i - X_i^T \beta + g(T_i, \beta)]^2. \tag{3.6}$$

V tomto konkrétnom prípade, odhad pre β je teda určený explicitne. Nech $\hat{g}_{y,h}(.)$ a $\hat{g}_{x,h}(.)$ sú jádra so šírkou pásma h pre Y, X a T. Potom :

$$\hat{\beta}_n = [\sum_{i=1}^{n} (X_i - \hat{g}_{x,h}(T_i))(X_i - \hat{g}_{x,h}(T_i))^T]^{-1} \times \sum_{i=1}^{n} (X_i - \hat{g}_{x,h}(T_i))(Y_i - \hat{g}_{y,h}(T_i)) \tag{3.7}$$

Jednou z dôležitých vlastností odhadu (3.7) je, že nevyžaduje vyhľadzovanie, takže šírka pásma v rozsahu $h \sim n^{-1/5}$ vedie k výsledku:

$$n^{1/2}(\hat{\beta}_n - \beta) \to Normal(0, B^{-1} C B^{-1}). \tag{3.8}$$

Kde B je kovariančná matica pre $X - E(X|T)$ a C je kovariančná matica pre

$\varepsilon[X - E(X|T)]$. Forma zápisu (3.7) môže byť použitá ak neberieme v úvahu chybu merania a nahradíme X pomocou W, potom výsledný odhad je nekonzistentný s β. Tvar, v ako máme odhad vyjadrený, ale ukazuje ešte niečo ďalšie. Je obecne známe, že v lineárnej regresíi, nekonzistentnosť spôsobenú chybou merania je možné obísť použitým "korekcie útlumu" *(correction for attenuation)*. V kontexte semiparametrických modelov, táto úvaha nás privádza k tomu, aby sme použili odhad v tvare:

$$\hat{\beta}_n = [\sum_{i=1}^{n}(W_i - \hat{g}_{w,h}(T_i))(W_i - \hat{g}_{w,h}(T_i))^T - n\Sigma_{uu}]^{-1} \times \sum_{i=1}^{n}(W_i - \hat{g}_{w,h}(T_i))(Y_i - \hat{g}_{y,h}(T_i)). \tag{3.9}$$

Odhad (3.9) je teda možné odvodiť rovnakým postupom ako je odvodený odhad od Severiniho a Staniswalisa.

Kapitola 4

Aplikácia EIV modelov v praxi

4.1 Výskum rakoviny prsu

Využitie EIV modelov praxi je možné názorne predviesť na nasledujúcej štúdii, pochádzajúcej od Gail Gong, Alice S. Whittemore a Stelly Grosser [GG90], ktorú s aplikáciou na errors in variables modely bližšie rozobrala Alice S. Whittemore [Whi89]. Štúdia sa zaoberala meraním úmrtnosti spojenej s výskytom rakoviny prs u žien žijúcich v oblasti San Francisca. Zrojom dát bol register nádorových ochorení v oblasti San Francisca. Skúmala sa úmrtnosť v skupine $N = 2495$ žien, ktorým bola diagnostikovaná rakoviná prs. Vek žien sa pohyboval v rozmedzí 55 - 64 rokov.
Pri štúdii sa autori zamerali hlavne na to, ako miera úmrtnosti závisí na závažnosti ochorenia pri diagnóze a na čase, ktorý uplynul od určenia diagnózy. Každá pacientka bola zaradená do jednej z 5-tich skupín, ktoré reflektovali jednotlivé štádiá ochorenia. Pričom väčšie číslo skupiny predstavovalo rozvinutejšiu formu ochorenia, viz tabuľka 3.

Pri určovaní štádia ochorenia sa vychádzalo z informácii uvedených v zdravotných záznamoch jednotlivých pacientiek. Keďže záznamy boli často nekompletné, je prevdepodobné, že niektoré pacientky boli zaradené do nesprávnej skupiny. Pacientky boli sledované po dobu 10 rokov, pričom toto obdobie bolo rozdelené na päť dvojročných cyklov. V rámci jedného cyklu malo každé štádium konštantnú mieru úmrtnosti, tj. ku zmene miery

úmrtnosti daného štádia mohlo dôjsť iba na počiatku nového dvojročného cyklu.

4.1.1 Rozbor podkladových dát

Nech $\lambda = \lambda_{jk}$ je matica typu (J, K), ktorá nám udáva pravdepodobnosti úmrtnosti. Hodnoty $J = 1, .., 5$ predstavujú jednotlivé cykly, a hodnoty $K = 1, .., 5$ predstavujú jednotlivé štádiá ochorenia, do ktorých sme pacientky rozdelili. Rozloženie smrtí (D) a počtu uplynulých mesiacov od určenie diagnózy (DM) pre sledovaný počet 2 495 pacientiek zachycuje tabuľka 3.

	1. N = 637		2. N = 1045		3. N = 441		4. N = 167		5. N = 205	
Mesiace	**D**	**PM**	**D**	**PM**	**D**	**PM**	**D**	**PM**	**D**	**PM**
< 24	12	12 773	42	20 456	45	8 359	38	3 072	123	2 788
24 - 48	24	9 048	70	13 096	66	4 485	28	1 508	30	962
48 - 72	21	5 469	33	6 767	14	1 539	8	664	10	267
72 - 96	6	2 639	18	2 651	4	451	1	274	2	131
96 - 120	0	596	2	477	0	82	0	52	1	51

Tabuľka 3: Štádium choroby

Data jednotlivých pacientiek pozostávajú z nasledujúcich údajov: t predstavuje čas do smrti alebo do ukončenia štúdie, premenná δ predstavuje indikátor úmrtia, nadobúda hodnoty 1 pri úmrtí a hodnoty 0 v prípade života pacientky. Posledným údajom je kovariančný vektor indikátorov štádia ochorenia, v ktorom sa pacietka nachádza. Označili sme ho ako $x = (x_1, .., x_K)$. Pre pacientku v štádiu k je $x = e_k$, kde e_k je $K-$dimenzionálny stĺpcový vektor s $k-$tou komponentou rovnou 1 a ostatnými nulovými komponentami.

Predpokladáme, že štádium, čas do smrti a čas do kontroly sú nezávislé náhodné veličiny. Potom hustota pre (t, δ) daná $x = e_k$ je úmerná k:

$$p(t, \delta, e_k, \lambda) = \exp\{\sum_{j=1}^{J} [\delta m_j(t) \log \lambda_{jk} - M_j(t)\lambda_{jk}]\}. \tag{4.1}$$

kde, $m_j(t)$ je indikátor rovný 1 ak t patrí do daného dvojročného cyklu, a rovný 0 inak. $M_j(t)$ je doba trávená v periode j, a konštanta úmrtnosti je nezávislá na λ. Pre hypotetického pacienta, ktorý zomrel v čase $t = 33$ mesiacov je potom $m_2(t) = 1$ a $m_j(t) = 0, j \neq 2$. Okrem toho je $M_1(t) = 24$, $M_2(t) = 9$ a $M_j(t) = 0$ pre $j > 2$. Nasčítaním (4.1) cez N sledovaných subjektov možno napísať log-likekihood kernel dát (t_i, δ_i) pre $i = 1, ..., N$ ako:

$$S_1 = \sum_{jk} [\triangle_{jk} \log \lambda_{jk} - \mu_{jk}\lambda_{jk}], \tag{4.2}$$

kde $\triangle_{jk}$ a μ_{jk} sú počty úmrtí (D) a uplynulé mesiace od určenie diagnózy (MD) v období j pre štádium k pacientov. Teda (4.2) predstavuje log-likekihood kernel pre $JxK = 25$ poissonových premenných $\triangle_{jk}$, ktoré majú priemere $\mu_{jk}\lambda_{jk}$. Teda regresné modely pre neznáme úmrtnostné miery λ_{jk} možno analyzovať prostredníctvom Poissonnových regresných metód.

Najobecnejší model *Model1*, umožňuje aby sa λ_{jk} voľne menili. Graf maximálne vierohodných odhadov hazard rates pre tento model ukazuje tendencie zvyšovania spolu so štádiom, avšak bez jasne viditeľných znakov paralelizmu v rámci daného skúmaného štádia.
The constant hazard model je označení ako *Model2*. Tento model predpokladá, že λ_{jk} sú konštantné vrámci všetkých 5-tich periód pre jednotlivé štádiá. Tento mdel predpokladá nezávislosť hazard rate na čase od určenia diagnózy, čo je značne reštriktívne. Vskutku, likehood ratio test porovnania *Model1* vs *Model2* vychádza silne rozporne, kedy chi-square= 97.1 s 20-timi stupňami voľnosti a $p < 0.01$.

Nie tak príliš reštriktívny je potom proportional hazard model *Model3*, ktorý predpokladá, že log hazard rates pre prvé štyri štádia sú zhodné. Do výpočtu sa nezahŕňa piate štádium, pretože prežitie pacientov v piatom štádiu, ktorí sú diagnostikovaný s metastázami sa správa inak ako u pacientov zaradených v ostatných štádiach onkologického ochrenia. Likehood ratio test *Model1* vs *Model3* potom odmieta *Model3* , keďže chi-square= 23.2 s 12-timi stupňami voľnosti a $p < 0.025$.

V tomto bode by sme mohli usudzovať, že obecný hazard model, ktorý sebou síce nesie riziko komplikovaných interpretácií, je vhodným modelom. Avšak autori skúmali aj možnosť, že aj jeden z viac reštriktívnych modelov je vhodný. A jeho nevhodnosť navonok, je na vrub chybám v priraďovaní pacientov do jednotlivých štádií ochorenia.

4.1.2 Analýza Errors-in-variables

Nech $\gamma = (\gamma_{kl})$ je matica pravdepodobností typu $(5, 5)$. Potom γ_{kl} je pravdepodobnosť, že pacientka pozorovaná v štádiu l je v skutočnosti v štádiu k, kde $k, l = 1, ..., 5$. Táto informácia nám obmedzila maticu γ na takú, že v nej nepozorujeme prechodné obdobia, medzi jednotlivými štádiami ochorenia. Má teda tvar:

$$\gamma = \begin{pmatrix} \gamma_{11} & 0 & 0 & 0 & 0 \\ 1-\gamma_{11} & \gamma_{22} & 0 & 0 & 0 \\ 0 & 1-\gamma_{22} & \gamma_{33} & 0 & 0 \\ 0 & 0 & 1-\gamma_{33} & \gamma_{44} & 0 \\ 0 & 0 & 0 & 1-\gamma_{44} & 1 \end{pmatrix} \tag{4.3}$$

Teda pre pacientku pozorovanú v štádiu 1 predpokladáme, že je v štádiu 1 alebo v štádiu 2 s pravdepodobnosťami γ_{11} a $1 - \gamma_{11}$. To isté potom platí aj pre ostatné štádia ochorenia $2, 3$ a 4. Štádium 5, v ktorom pacientky mali metastatický nález, bolo určené jednoznačne.

Parametre (λ, γ) boli odhadnuté maximalizáciou vierohodnosti pozorovaných dát, pomocou pozorovaných štádiových premenných $X_1, ..., X_N$:

$$\prod_{i=1}^{N} \sum_{k=1}^{K} f_{t,\delta,x|X}(t_i, \delta_i, e_k | X_i; \lambda, \gamma). \tag{4.4}$$

Kde $f_{t,\delta,x|X}$ je podmienená pravdepodobnstná hustota, ktorá je funkciou (t, δ, x) a je daná hodnotou vektora štádíi X. Za určitých podmienok je potom úmerná k:

$$\prod_{i=1}^{N} \sum_{k=1}^{K} p(t_i, \delta_i, e_k, \lambda) f_{x|X}(e_k|X_i; \gamma). \tag{4.5}$$

kde $p(.)$ je dané (4.1), $f_{x|X}$ je mnohočlenná hustota pre vektor x skutočného štádia, podmienením pozorovaným štádiom a parametrizovaný pomocou γ z (4.3). Konstanta úmernosti pritom nezávisí na (λ, γ). Úmernosť (4.4) a (4.5) je splnená, za predpokladu, že platia nasledujúce tri predpoklady: $(i.)$ časy do úmrtia a kontroly sú stochasticky nezávislé, $(ii.)$ rozdelenie času do smrti nezávisí na X a $(iii.)$ rozdelenie času do kontroly nezávisí na x. V takom prípade potom možno predpokladať, že (4.4) a (4.5) dávajú oba správne pozorované stavy indikátorov (x, X).

Prvý predpoklad zabezpečí, že časy do úmrtia majú rozdelenie u skúmaných pacientov v danom pravdivom štádiu, rovnajúcom sa neskúmaným pacientom. Tento postup je bežne používaný v analýze prežitia. Druhý predpoklad ukotvuje nezávislosť prognózy prežitia na presnosti zaradenia množstva skúmaných pacientov do jednotlivých skupín podľa ich skutočného stavu choroby. Tento predpoklad vychádza z analýzy chyby merania. Posledný, tretí predpoklad nakoniec zabezpečí, že skúmaná prognóza je nezávislá na skutočnom stave množstva pacientov pozorovaných v danom štádiu. Predpoklad teda platí približne v prípade, že pravdepodobnosti zaredenia do nesprávneho štádia choroby sú nízke, alebo keď pravdepodobnosť výberu má malý vzťah k rozsahu choroby a stavu štádia.

Odhady $(\hat{\lambda}, \hat{\gamma})$ maximalizace (4.5) nie sú dostupné v priamej forme a možno ich získať použitím EM algoritmu popísaného napríklad v práci od A.P. Dempster, N.M.Laird a D.B.Rubin [A.P77]. $E - krok$ $(E - step)$ z EM algoritmu napočítava $Q(\lambda^c, \gamma^c)$, predpoklad úplnosti dát logaritmickej vierohodnosti, závisí na pozorovaných datach a aktuálnom odhade (λ^c, γ^c). Kompletnosť dát v tomto prípade predstavuje hodnoty $(t_i, \delta_i, X_i, x_i)$ pre

$i = 1, ..., N$. $M - krok$ $(M - step)$ následne maximalizuje Q, tak aby sme dosiahli nový odhad. Z $E - kroku$ dostávame:

$$Q(\lambda, \gamma) = \Xi_1(\lambda) + \Xi_2(\gamma), \tag{4.6}$$

kde:

$$\Xi_1(\lambda) = \sum_{jk} [\varepsilon(\triangle_{jk}) \log \lambda_{jk} - \varepsilon(\mu_{jk})\lambda_{jk}], \tag{4.7}$$

a ďalej:

$$\Xi_2(\gamma) = \sum_{kl} [\varepsilon(\nu_{kl}) \log \gamma_{kl}]. \tag{4.8}$$

Tu ν_{kl} predstavuje počet $k - teho$ štádia pacientov s pozorovaným stavom diagnózy l a $\varepsilon(.)$ udáva podmienené očakávané hodnoty dané (t_i, δ_i, X_i) pre $i = 1, ..., N$ parametrizáciou (λ, γ). Tieto očakávania sú potom splnené jednoduchým nahradením nepozorovaného vektoru x pre každého pacienta, vektorom $\chi = (\chi_1, ..., \chi_K)$ pravdepodobností, ktoré vyjadrujú skutočný vzťah k jednotlivým stavom choroby. Jednotlivé komponenty χ reprezentujú potom podmienené pravdepodobnosti zavedenia v jednotlivých stavoch, daných pozorovaným stavom pacienta, datami prežitia a odhadmi parametrov [GG90].
Vo výsledku teda možno zhrnúť, že prevedené simulácie v [GG90] odhaľujú nasledujúce dve zistenie: (i) v prípade, že ignorujeme možnosť nesprávneho zaradenia do skupiny, môžeme tým spôsobiť odmietnutie proportional hazard modelu, i keď je tento model správny. Na druhú stranu zistenie (ii) v prípade, že nijakým spôsobom nevalidujeme data, likelihood ratio test je skoro nepoužiteľný pri detekcíi nesprávneho zaradenia do skupiny, jedine, že by sme skúmali skutočne veľké množstvo dát.

45 Záver

V úvodnej kapitole tejto práce sme zadefinovali základné pojmy a pojem regresie a na príklade merania synov a otcov sme si ukázali jednu z možností praktického využitia regresie. V druhej kapitole sme rozobrali lineárny regresný model a jeho vlastnosti, pričom sme sa zaoberali i normálnym lineárnym modelom, jeho charakteristikami a možnosťami odhadu jeho parametrov. Tieto poznatky sme následne v tretej kapitole využili pri zavedení jedtnotlivých typov modelov errors in variables.

V poslednej kapitole tejto práce sme uviedli aplikáciu modelov error in variables v medicíne, kde sme sa zaoberali výskumom rakoviny pŕs. Na základe uvedených poznatkov sme dospeli k rozhodnutiu, že použitie analyzy errors in variables nás doviedlo k niekoľkých predpokladom a vlastnostiam, ktoré sú požadované po skúmaných dátach. Obdobne je teda možné nájsť množstvo zaujímavých aplikácií v rôznych oblastiach vedy.

Literatúra

[And05] J. Anděl. Základy matematické statistiky. MATFYZPRESS, 2005.

[A.P77] D.B.Rubin A.P.Dempster, N.M.Laird. Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the royal statistical society. Series B, Vol.39, pp. 1-39, 1977.

[A.T94] G.J. Staniswalis A.T.Severini. Quasilikelihood estimation in semiparametric models. Journal of the American statistical association, Vol. 89, pp. 501-511, 1994.

[CE01] Shayle R. Searle Charles E.McCulloch. Generalized,linear and mixed models. John Wiley Sons, Inc, 2001.

[Che88] H. Chen. Convergence rates for parametric components in a partly linear model. The annals of statistics, Vol.16, pp. 136-146, 1988.

[Dob02] Annette J. Dobson. An introduction to generalized linear models. Chapman Hall/CRC, 2002.

[Dod93] David Birkes;Yadolah Dodge. Alternative Methods of regression. A Wiley-Interscience Publication, 1993.

[E.N86] E.N.Heckman. Spline smoothing in partly linear models. Journal of the royal statistical society. Series B, Vol. 48, pp. 244-248, 1986.

[Ful87] Wayne A. Fuller. Measurement error models. John Wiley Sons, Inc, 1987.

[Gal86] Francis Galton. Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute, 1886.

[GG90] Stella Grosser Gail Gong, Alice S. Whittemore. Censored survival data with misclassified covariates: A case study of breast-cancer mortality. Journal of the American statistical association, Vol. 85, No. 409, pp.20-28, 1990.

[HL97] W. Härdle H. Liang. Asymptotic normality of parametric part in partially linear heteroscedastic regression models. Humboldt university Berlin, DP 33, SFB 373, 1997.

[RE86] J. Rice A. Weiss R.F. Engle, C.W.J. Granger. Semiparametric estimates of the relation between weather and electricity sales. Journal of the American statistical association, Vol. 81, pp.310-320, 1986.

[Spe88] P. Speckman. Kernel smoothing in partial linear models. Journal of the royal statistical society. Series B, Vol.50, pp. 413-436, 1988.

[Whi89] Alice S. Whittemore. Errors-in-variables regression problems in epidemiology. American mathematical society,Vol.112, pp.17-32, 1989.

[Zvá89] Karel Zvára. Regresní analýza. Academia, 1989.