Odhady kvantilů v lékařských aplikacích - Regresní modely.

Bohumír Procházka

Státní zdravotní ústav Praha, Šrobárova 48

V lékařském výzkumu, ale i v ostatních aplikacích matematické statistiky, je často požadováno vytvoření norem na základě standardní populace. K tomuto účelu bývají velmi často používány odhady kvantilů. Problém nastává pokud je nutno odhadnout kvantily v regresním modelu. V praxi bývá používáno několika různých přístupů.

V tomto příspěvku se pokusím popsat některé z používaných metod a ukázat jejich aplikaci na příkladech dat antropometrického výzkumu. Budou zde popsány například metody na založené odhadech střední hodnoty a rozptylu sledované proměnné v závislosti na jiné, rušivé veličině, které odhadují kvantily pro jednotlivé hodnoty nezávisle proměnné na základě předpokladu normality měřené veličiny. Protože tento předpoklad nebývá často splněn, používají někteří autoři transformace (log, Box-Cox, ...), které převádějí měřenou veličinu tak aby její rozložení bylo blíže normálnímu rozdělení. Jindy bývá využito i jádrových odhadů nebo dalších metod vyhlazování dat. V poslední době se ještě nabízí další možnost která je založena na možnosti zobecnit pojem kvantilů na regresní model.

V praktických příkladech budou použity výsledky studií, závislosti výšky postavy, hmotnosti a váhovýškového indexu na věku dítěte.

Nejjednodušší přístup používaný v lékařské praxi je založen na rozdělení celého souboru podle nezávisle proměnné na menší skupiny, tak aby rozdíly středních hodnot závisle proměnné uvnitř takto vytvořených skupiny byly zanedbatelně malé. Pak je možné provést odhad kvantilů pro každou skupinu odděleně a to buď přímým výpočtem výběrových kvantilů, nebo použitím předpokladu o typu rozložení. V druhém případě pak jsou odhadnuty parametry předpokládaného rozložení a s jejich využitím pak konstruovány odhady kvantilů tohoto rozložení, případně toleranční meze.

Jako příklad takového přístupu je možné jmenovat článek M.A.Wilkoxe a spol.(1993) o závislosti porodní hmotnosti na týdnu porodu nebo D.Campell a spol.(1993) zabývající se podobným problémem. V těchto článcích autoři odhadují pro každý týden porodu 10%, 50% a 90%-ní kvantily za předpokladu normality porodní váhy jako kvantily normálního rozložení se střední hodnotou a rozptylem odhadnutým pro příslušný týden porodu. Nepříjemná vlastnost tohoto přístupu je dána již tím, že uvažuje jednotlivé podskupiny nezávisle na sobě a vzhledem k tomu, že je nutno použít dostatečně jemné dělení, narůstá rychle nepřesnost těchto odhadů. K získání dostatečně spolehlivých výsledků je pak nutný velmi vysoký počet měření. Tímto rozdrobením celkové informace se vlastně zříkáme možnosti využít vlastností regresního modelu. Dalším důsledkem toho přístupu je z praktického pohledu i skutečnost, že křivky vzniklé spojením příslušných odhadů nejsou hladké a jsou zatížené individuální chybou odhadu každé skupiny (týdne porodu).

Někteří autoři tento poslední problém řeší použitím metod vyhlazování dat a to buď číselným vyhlazováním (např. spline) nebo případně i pomocí grafických úprav.

Dalším možným přístupem je vytvoření lokálního odhadu parametrů (lokální vzhledem k nezávisle proměnné). Tento přístup je použit například ve studii L.Lhotské a spol.(1992). Jedná se o celostátní antropometrickou studii v které bylo v průběhu roku 1991 změřeno 90910 dětí ve věku od 0 do 18 let. U každého jedince byl jednak evidován věk v okamžiku měření (ve dnech) a další antropometrické údaje. Zde se dále budeme zmiňovat o obvodu hlavy. Jedním z cílů studie bylo získání percentilových růstových grafů, které se v lékařské praxi používají jako standard k porovnání sledovaného jedince a jeho vývoje s celou populací. Všechny odhady byly prováděny pro každé pohlaví zvlášť.

Při odhadu kvantilů bylo použito předpokladu normality sledované veličiny. Pro každý den x věku byly vypočteny odhady parametrů lineární regrese pro jedince jejichž věk se neliší o více než o 15 dní. Dále byla v tomto modelu odhadnuta střední hodnota M a rozptyl S^2 sledované veličiny ve dni x. Nakonec byly vypočteny příslušné kvantily normálního rozložení se střední hodnotou $M(x)$ a rozptylem $S^2(x)$. Tímto způsobem byly získány odhady kvantilů v závislosti na věku, nebyly však ještě dostatečně hladké obzvlášť pro extrémní kvantily. V dalším kroku bylo použito k vyhlazení těchto funkcí klouzavých průměrů.

Jiným používaným přístupem k tomuto problému je postup který navrhl Cole(1988). Ten navrhuje postup při kterém je uvažována širší třída distribucí. Postup je založen na Box-Coxově transformaci. Tito autoři navrhli ve svém článku z roku 1964 použít jednoparametrickou transformaci

$y^{(\lambda)} = \frac{(y^\lambda - 1)}{\lambda}$ pro $\lambda \neq 0$

$y^{(\lambda)} = \ln(y)$ pro $\lambda = 0$

nebo dvouparametrickou transformaci

$y^{(\lambda)} = \frac{((y+\delta)^\lambda - 1)}{\lambda}$ pro $\lambda \neq 0$

$y^{(\lambda)} = \ln(y + \delta)$ pro $\lambda = 0$

které se snaží aby transformovaná veličina měla přibližně normální rozložení. Dále budeme používat první z nich. Pro odhad parametrů tohoto modelu je tedy nutno spočítat jednak odhady exponentu λ ale i střední hodnoty μ a rozptylu σ^2 transformované veličiny. Označme tyto odhady L, M a S^2 .

Autor navrhuje nejprve rozdělit soubor podle nezávisle proměnné x (v našem případě podle věku) do p skupin. Takto získáme tři posloupnosti L, M a S^2 , $i = 1, ..., p$, z nich navrhuje autor vytvořit spojité funkce $L(x), M(x)$ a a $S^2(x)$ ať již lineárním dodefinováním mezi sousedními časy, nebo použitím polynomické regrese, případně i jiným způsobem vyhlazení. Z takto získaných odhadů je již možno konstruovat za předpokladu normality požadované kvantily. Dříve než přikročíme k jejich odhadu, je samozřejmě nutné ověřit normalitu transformované veličiny. α-kvantil pak definuje:

$Q(x) = M(x)(1 + L(x)S(x)z_\alpha)^{\frac{1}{L(x)}}$ pro $L(t) \neq 0$

$Q(x) = M(x)exp(S(x)z_\alpha)$ pro $L(t) = 0$

kde z_α je α-kvantil normálního rozložení.

Tento přístup použila i Chinn a spol.(1992) v antropometrické studii anglických a skotských dětí. V studie zahrnuje 8107 dětí ve věku 5-11 let. Autoři nejprve vytvořili

skupiny po jednom roku věku pro každé pohlaví zvlášť. V těchto skupinách provedli Box-Coxovu transformaci a funkce $L(x), M(x)$ a $S(x)$ získali pomocí polynomické regrese třetího stupně pro tyto hodnoty. Tyto funkce pak již přímo použili k definici kvantilů.

Jinou možností, kterou používá S.Chinn (1992) je využití znalosti o tvaru heteroscedascity výšky postavy v závislosti na věku osoby. Autorka navrhuje provést lineární transformaci stabilisující rozptyl a pak pro logaritmy transformovaných hodnot použije kubickou regresi. Za předpokladu normality pak konstruhuje pomocí kritických hodnot normálního rozdělení požadované kvantily.

Dalším možným přístupem je použít myšlenky regresních kvantilů, které poprvé definoval Koenker a Bassett (1978), a jejich zobecnění na nelineární regresní model. Kvantily definujeme jako vektor parametrů minimalizující součet:

$$\sum \rho_\alpha(Y - g(x, \beta))$$

kde $\rho_\alpha(z) = z(\alpha I_{[z>0]} + (\alpha - 1)I_{[z\leq 0]})$, g(.,.) je zvolená regresní funkce, x nezávisle proměnná a t neznámý parametr. Statistické vlastnosti těchto odhadů jsou studovány mnohými autory např. Ruppert Carroll(1981), Jurečková (1984) Procházka(1993) a.j. Výhodou tohoto přístupu je že tvar vypočtených kvantilů odpovídá použitému modelu, tedy zvolené regresní funkci. Další pozitivní vlastností je i ta skutečnost, že takto definované výběrové regresní kvantily si zachovávají tu vlastnost, že $100\alpha\%$ pozorování leží pod touto funkcí. Obzvláště u velmi rozsáhlých studií se však objevuje problém s volbou vhodného modelu. V případě simulací tento problém samozřejmě odpadá a odhady fungují bez vážnějších problémů. V reálné praxi je ale každý model do jisté míry pouze aproximací skutečného modelu. Ve skutečnosti se to pak projevuje tak, že je-li pro zvolený model dostatečně velký rozsah souboru je často snadné prokázat neshodu modelu s daty. Tento problém nabývá významu obzvláště pokud se pokoušíme nalézt model pro celý věkový rozsah. Autoři proto nejčastěji rozdělují růst na menší etapy (ranné dětství, období do puberty a puberta).

Obraťme nyní pozornost k praktickým výpočtům. Pro poslední z výše jmenovaných přístupů jsem vytvořil program. Odhady regresních kvantilů jsou počítány iteračně tak, že výpočet končí pokud proces dosáhne se zvolenou přesností minima. Pokud však není dosaženo se zvolenou přesností toho, že získaný kvantil neleží nad odpovídajícím procentem pozorování, opakuje se výpočet do té doby než je toto kritérium splněno.

Nejprve jsem se věnoval studii Chinn a spol.(1992), a to především závislosti váhovýškového indexu (BMI) na věku. Ve studii jsou vypočteny kvantily pro každé pohlaví zvlášť. Aby bylo možné výsledky porovnat, použil jsem, stejně jako autoři článku, pro výpočet regresních kvantilů kubickou regresi. Porovnání obou odhadů pro skupinu chlapců je zobrazeno v grafu 1. (tečkovanou čarou jsou nakresleny odhady z článku a plnou odhady regresních kvantilů pořízené jak na Mainframe tak i na PC) Váhovýškový index má sám o sobě poměrně velký rozptyl, a tak je zřejmé, že jednotlivé kvantily budou poměrně vzdálené a že odhady získané různými metodami se mohou obzvláště pro extrémní kvantily lišit. Neshoda odhadů na mainframe a PC je způsobena jinou požadovanou přesností a velkým rozptylem sledované veličiny a projevuje se především u 97%-ního kvantilu. Z grafu je zřejmá celkem dobrá shoda mediánu a kvantilů pro nízká α. Podstatný rozdíl mezi odhady je pro 97% a 90%-ní kvantily. Výpočet byl ukončen pokud relativní změna parametrů byla

menší než 0.00001. Dosažená procenta bodů pod křivkou jsou v tabulce 1 společně s počtem iterací a použitým procesorovým časem. Kvantily získané pomocí Colovy metody jsou systematicky nižší, což poukazuje na porušení normality transformované veličiny.

Výpočty těchto kvantilů jsem provedl na počítači Amdahl 5890-300E (42MIPs) později jsem opakoval tento výpočet na počítači PC386DX 25MHz. Hrubé porovnání rychlosti a přesnosti výpočtu je též v tabulce 1. Průměrná doba jedné iterace na mainframe je 0.0947 sec a na pc 2.74 sec. Rychlost výpočtu na mainframe je samozřejmě nesrovnatelně vyšší, ale i výpočet na PC je zřejmě použitelný.

Tabulka 1.:

	Mainframe			PC		
Kvantil	Počet iterací	Procento pod	Čas sec.	Počet iterací	Procento pod	Čas sec
.03	73	3.04	6.09	48	3.01	108
.10	87	10.01	7.46	16	9.98	133
.20	67	19.99	6.47	55	20.01	88
.50	47	49.94	5.23	150	49.99	378
.80	79	79.96	6.74	21	80.01	162
.90	69	89.97	6.34	49	90.02	47
.97	63	96.98	7.59	41	96.99	125

Druhým příkladem použití regresních kvantilů je závislost obvodu hlavy na věku dítěte v intervalu od 0 do 8 let věku. Pro tento příklad jsem použil dat ze studie L.Lhotské (1992). Pro výpočet regresních kvantilů bylo použito zobecnění logistické funkce

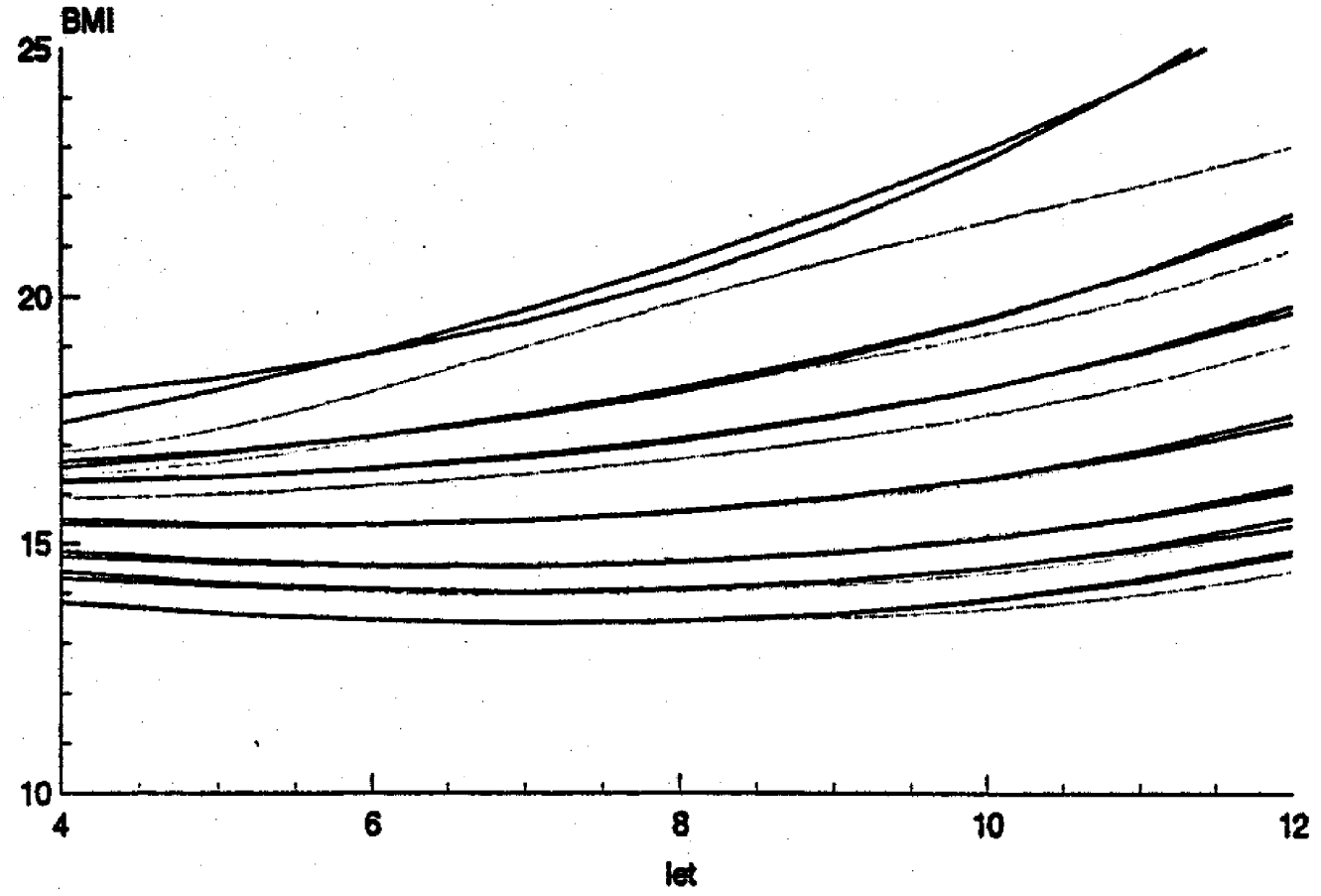
$$g(x,\beta) = \beta_1 + \frac{\beta_2}{1+\beta_3(x+\beta_4)^{\beta_5}}$$

Vypočtené regresní kvantily, opět pro chlapce (jedná se o 21192 chlapců), jsou zobrazeny v grafu 2. Pro výpočet všech sedmi kvantilů bylo na PC386DX (25MHz) potřeba asi 158 min (289 iteračních kroků), což odpovídá průměrnému času potřebnému na jeden iterační krok 32.6s. Protože z výše zmíněné studie jsem měl k disposici pouze grafy, z kterých bylo obtížné přesně odečíst hodnoty odhadů a protože odhady obvodu hlavy byly počítány pouze pro věk od 0 do 3 let, je obtížné nakreslit a porovnat v jednom grafu výsledky obou zmíněných metod.

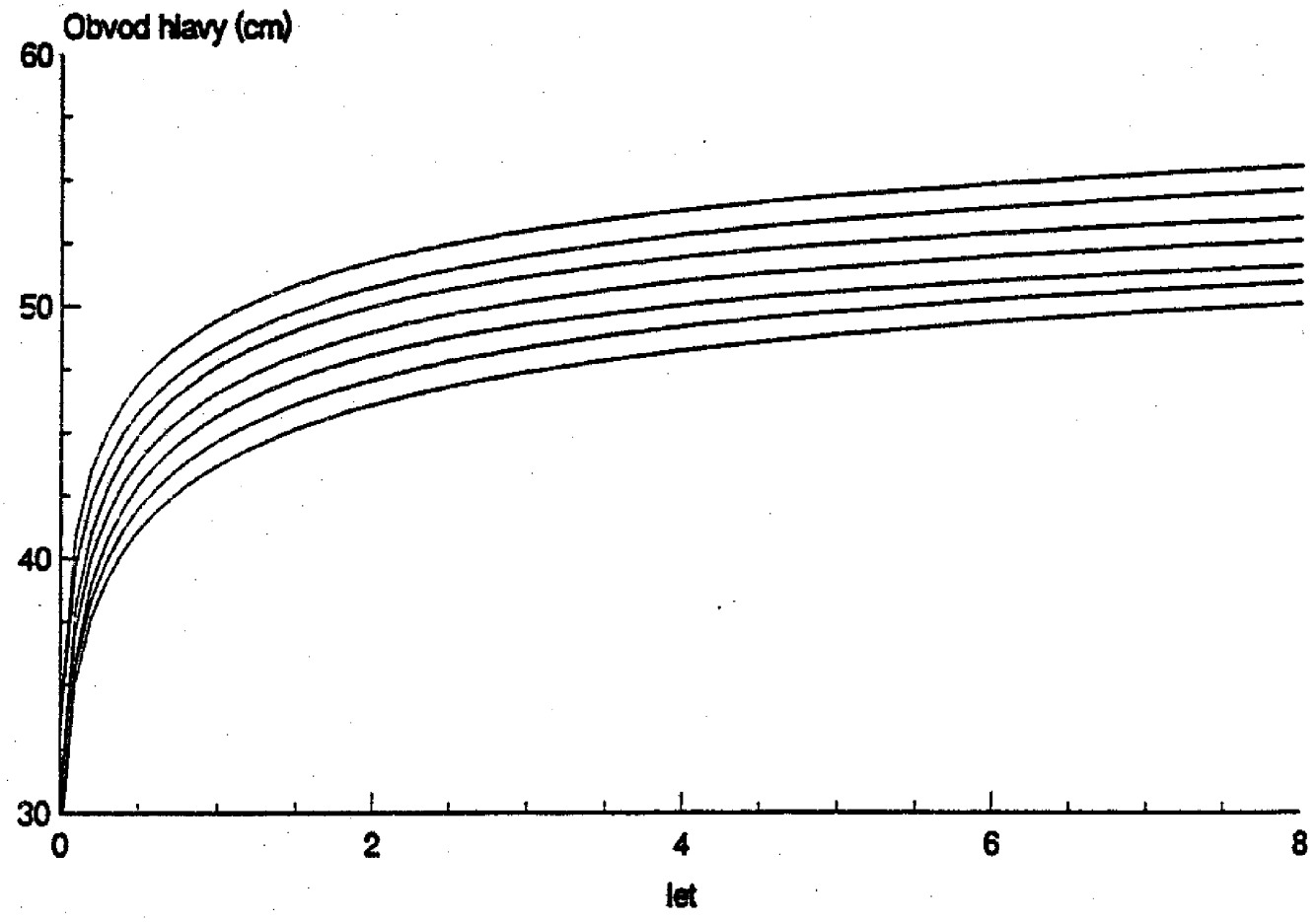
Poděkování:

Děkuji prof W.W.Hollandovi a S.Chinn z oddělení Public Health UMDS Londýn za poskytnutí reálných dat a za možnost provést výpočty na mainframe počítači. Výpočty jsem provedl v rámci studijního pobytu podporovaného komisí evropského společenství pro spolupráci ve vědě a technice se středo a východoevropskými zeměmi. Dále děkuji dr L.Lhotské za poskytnutí přístupu k datům celostátní antropometrické studie grantu MZČR číslo 0268-3 a za možnost provést výše zmíněné výpočty.

Graf 1.



Graf 2.



Literatura:

D.Campell, M.Hall, J.Lemon, R.Carr-Hill, C.Pithard, M.Samphier (1993): Clinical birthweight standards for a total population in the 1980s. British Journal of Obsterics and Gynaecology vol.100,436-445.

S.Chinn, R.J.Rona, M.C.Gulliford, J.Hammond (1992): Weight-for-hight in childern aged 4-12 yars. A new index compared to the normalized body mass index. European Journal of Clinical Nutrition, 46, 489-500.

S.Chinn (1992): A new method for calculation of height centiles for preadolescent children. Annals of Human Biology, vol.19, No.3, 221-232.

T.J.Cole (1988): Fitting Smoothed Centiles to Reference Data. JRSS A 151, parth 3,385-418.

Jurečková,J.(1984) : Regression quantiles and trimmed least squares estimator under a general design. Kybernetika 20, 345-357.

Koenker,R. and Bassett,G.(1978) : Regression quantiles. Econometrica 46, 33-50.

L.Lhotská, P.Bláha, J.Vignerová, Z.Roth, M.Prokopec (1993): V.celostátní antropometrický výzkum dětí a mládeže 1991 (České země). Státní zdravotní ústav Praha.

Procházka,B.(1993) : Useknuté odhady v modelu nelineární regrese. CSc disertační práce, Karlova Universita, Praha.

Ruppert,D. and Carroll,R.J.(1980) : Trimmed least-squares estimation in the linear model. J.Amer.Statist.Assoc. 75, 828-838.

M.A.Wilcox, I.R.Johnson, P.V.Maynard, S.J.Smith, C.E.D.Chilvers (1993): The individualised birthweight ratio: a more logical outcome measure of pregnancy than birthweight alone. British Journal of Obsterics and Gynaecology vol.100,342-347.