# ROBUSTNÍ IDENTIFIKACE ŽIVOTNOSTNÍHO MODELU

P. Kovanic, P. Volf, ÚTIA ČSAV, Praha

## 1 Úvod

Práce je věnována nové metodě analýzy nelineárních regresních modelů pro přežívání. Problém přežívání je již dlouhou dobu předmětem zájmu jak teoretiků, tak i praktiků z různých oborů. Je to dáno jednak složitostí řešení této úlohy v reálných podmínkách a jednak jejím významem.

Standardní programové systémy pro statistické zpracování dat (např. obecně známé SPSS či BMDP) jako nezbytnou součást zahrnují i programy pro zpracování životnostních dat statistickými metodami. Používání těchto programů v praxi však naráží na značné potíže:

Pokud se tyto programy neomezují jen na prosté uspořádávání dat a zahrnují i odhadování parametrů modelů životnosti, jsou závislé na apriorní volbě pravděpodobnostního modelu náhodné složky dat. U reálných procesů lze však volbu tohoto modelu podpořit přesvědčivými argumenty jen výjimečně a ani dodatečné ověření správnosti volby není snadnou záležitostí. Výsledky zkoumání dat pak mohou vykazovat značnou citlivost k volbě statistického modelu, nejsou dostatečně robustní k předpokladům. Apriorní volba statistického modelu je dále komplikována tím, že v některých životnostních úlohách nejsou apriori určeny ani meze definičního intervalu distribuční funkce, přičemž předpoklad o nekonečném nosiči by byl absurdní. Tak např. únavové lomy strojních součástí jistě nenastanou před prvním zatěžovacím cyklem a lze vyloučit, že by součást vydržela nekonečně mnoho zatěžovacích cyklů. Pacient sice může zemřít ihned po podání léku nebo po operačním zásahu, ale jistě existuje konečná doba života, kterou nepřežije žádný pacient. Apriorně neznámé meze nosiče dat je pak často nutné rovněž odhadovat z dat.

Další potíží může být závislost rozptylu náhodné doby života na nějaké jiné veličině (kovariátě regresoru). Tento přirozený jev vylučuje možnost použití konstantního modelu neurčitosti a zvyšuje počet neznámých parametrů.

Programy založené na klasických postupech matematické statistiky dávají často výsledky silně ovlivněné jednotlivými daty, která jsou oproti ostatním netypická, odpovídají bodům odlehlým od hlavního shluku. Nejsou tedy dostatečně robustní k odlehlým datům.

Programy založené na postupech robustní statistiky mají sice zlepšenou robustnost k předpokladům, avšak nejsou dostatečně univerzální při používání. O tom svědčí současná existence velkého množství takových metod řešících tentýž problém s týmiž daty a dávajících různé výsledky.

Používání těchto metod obecně nezajišťuje maximální využití informace obsažené v datech. Proto bývá uspokojivá kvalita výsledků podmíněna dostupností dostatečného množství dat. To však často nelze v praxi zajistit, neboť životnostní data bývají vzácná či drahá (např. jeden výsledek = jeden lidský život nebo jeden automobil zničený dlouhou sérií testů). Požadavek informační optimality programu je proto v životnostních úlohách obzvláště naléhavý.

Při sledování životnosti se vyskytují i nedokončené testy (a jim odpovídající tzv. cenzorovaná data). Výsledek měření ($T$) pak poskytuje informaci 'doba života převyšuje $T$' místo 'doba života se rovná $T$'. I takto omezená informace však může – zejména při malém počtu dat – podstatně ovlivnit výsledek analýzy a musí proto být adekvátním způsobem vzata v úvahu.

Doba života může záviset na mnoha vlivech. Některé z těchto vlivů mohou být rovněž měřeny.

Doba života se pak určuje jako funkce několika proměnných. Tato funkce bývá často nelineární a protože kriteriální funkce (hodnotící shodu dat s odhady) nemusí být kvadratická, je k nalezení nejlepších odhadů hledaných parametrů třeba řešit nelineární rovnice. Navíc mohou pro některé parametry nebo jejich funkce platit některá podstatná omezení, vyplývající z reálného smyslu úlohy. Řešení životnostního problému pak zahrnuje podmíněnou extremalizaci funkce v mnohorozměrném prostoru, což vede k nutnosti používat složité algoritmy numerické matematiky. U takových algoritmů však vyvstává naléhavá otázka algoritmické spolehlivosti, tj. schopnosti řádného fungování 'za pokud možno všech' okolností.

Zmíněná nelinearita životnostních modelů má další závažný důsledek. Všechny lineární modely jsou si podobné, mají tutéž strukturu, a to i ve vícerozměrném případě. To dovoluje používat jediný program pro identifikaci kteréhokoliv lineárního modelu bez ohledu na specifičnost oblastí, z nichž pocházejí data. Tyto modely tedy tvoří jedinou třídu. Nelineárních modelů je však nekonečně mnoho. To vede k nutnosti specializace identifikačních programů, k takovému apriornímu vymezení třídy řešených úloh, které by udrželo celkovou složitost programu na rozumné úrovni. Protože však uživatel zpravidla požaduje co nejobecnější použitelnost programů, je třeba najít přijatelný kompromis mezi jednoduchostí (a tudíž i spolehlivostí) a univerzální aplikovatelností.

Požadavky kladené na metodu pro identifikaci životnostních modelů lze tedy shrnout takto: Požaduje se robustnost k apriorním statistickým předpokladům a dostatečně univerzální použitelnost k datům rušeným neurčitostmi různé povahy (a pocházejícím z omezeného intervalu majícího neznámé meze), robustnost k odlehlým datům, informační optimalita, zajišťující co nejlepší využití všech (necenzorovaných i cenzorovaných) dat, algoritmická spolehlivost a rozumný kompromis mezi složitostí a univerzální použitelností. Dostupné programové systémy těmto náročným požadavkům kladeným praxí plně nevyhovují. To je důvodem k hledání nových řešení.

## 2 Formulace úlohy a stručný popis řešení

Úlohu zformulujeme pro názornost jako úlohu zkoumání životnosti strojních součástí, i když lze předpokládat, že obdobná úloha se vyskytuje i v dalších oborech, jako např. ve zdravotnictví, v pojišťovnictví aj., a že popisovaný přístup je tedy použitelný obecněji.

Byla provedena série $N$ nezávislých pokusů s cyklickým zatěžováním strojních součástí a měřením jejich životnosti. V různých pokusech byla amplituda vibrací $Z$ (určující maximální namáhání) nastavena na různé hodnoty, ale v průběhu pokusu se už neměnila.

Předpokládá se, že změřená 'doba života' $T_k$ (počet period vibrací do únavového lomu nebo do jiného ukončení zkoušky) $k-$té zkušebně zatěžované součásti (např. pružiny) závisí na nezáporné zátěži $Z_k$ (na amplitudě vibrací) dle vztahu

$$T_k = t_k * R_k, \tag{1}$$

kde

$$t_k = B * (Z_k + C)^{-A} + D \tag{2}$$

je doba života dle teoretického modelu s neznámými parametry a $R_k$ je hodnota $k-$tého multiplikativního rezidua.[1] Spojitá a diferencovatelná distribuční funkce neznámého typu $P_Z(R)$ reziduí $R$.

[1] Obvykle se jako reziduum označuje rozdíl mezi skutečnou hodnotou závislé proměnné a hodnotou

mající nekonstantní parametr měřítka

$$S(Z) = S_0 \exp(S_1 Z), \tag{3}$$

je definována na omezeném intervalu $(R_L, R_U)$, který při dané zátěži $Z$ odpovídá intervalu dob života od $T_L(Z)$ (horní hranice dob života, během nichž téměř jistě lom nenastane) do $T_U(Z)$ (doba života do téměř jistého lomu).

Je dána $N-$tice datových trojic $(Z_k, T_k, I_k)$, kde $I_k$ je indikátor cenzorovanosti dat rovný 1 (ukončení zkoušky lomem) nebo 0 (jiné ukončení).

Požaduje se odhad vektoru neznámých parametrů

$$\Theta = (S_0, S_1, R_L, R_U, A, B, C, D) \tag{4}$$

a odhad distribuční funkce $P_Z(R)$ ve tvaru použitelném v procedurách umožňujících výpočet pravděpodobnosti $P\ (T \leq t|Z)$ i kvantilů $q(\alpha) = P_Z^{-1}(\alpha)$ pro všechna $Z$ z definičního oboru.

Stojíme tedy před úlohou nelineární regresní analýzy, která má určité nestandardní prvky a na jejíž řešení klademe přirozené požadavky, shrnuté v úvodní části. Právě snaha vyrovnat se s těmito požadavky co nejlépe vedla k hledání nové, nestandardní metody, založené na gnostickém popisu neurčitosti [1]. Stěžejním pojmem je přitom (globální) gnostický odhad distribuční funkce P($R$) rozdělení reziduí $R$. Tato d. f. je určena daty (tj. v našem případě rezidui $R_k$, případně indikátory $I_k$) a vektorem parametrů $\Theta$. Dosadí-li se v nějakém přiblížení odhady parametrů do modelu (2), lze vypočítat hodnoty reziduí $R_k$ pro všech $N$ datových trojic. Globální odhad distribuční funkce v kterémkoliv bodě $R$ je pak určen těmito rezidui a aktuálními odhady parametrů $S_0, S_1, R_L$ a $R_U$ určenými z podmínky nejlepší shody modelu distribuční funkce s bezprostřední reprezentací dat – reziduí, s tzv. bodovou empirickou distribuční funkcí (BEDF). Pro $R_{[i]}$ značící $i-$té multiplikativní (uspořádané) reziduum má tato funkce tvar

$$F_p(R_{[i]}) := \frac{i - 0.5}{N}, \quad (i = 1, .., N) \tag{5}$$

Stručně může tedy být postup řešení shrnut do těchto kroků:

1. Nastavit nebo jednoduchou procedurou zhruba odhadnout $\Theta_0$ jako první přiblížení k $\Theta$.
2. Vypočítat rezidua užitím (1) a (2), uspořádat je a dle (5) určit bodové odhady hodnot jejich empirické distribuční funkce ($BEDF$).
3. Korigovat $BEDF$ pro zahrnutí cenzorovanosti některých dat s použitím Kaplanova–Meierova odhadu.
4. Pomocí optimalizační procedury najít další iteraci $\Theta_i$ z podmínky minimalizace kritéria dobré shody gnostického odhadu distribuční funkce $P(R)$ reziduí s $BEDF$.
5. Pokračovat od bodu 2 pro $i + 1$ do ustálení odhadu $\Theta_i$.
6. Výsledný odhad parametrů dosadit do procedur pro odhadování pravděpodobné doby života nebo přípustné zátěže.

---

vypočtenou modelem. V modelu (1) je $R_k$ exponenciální funkcí takového "aditivního" rezidua, proto hovoříme o "multiplikativním reziduu".

Optimalizační procedura v 4. je klíčovým místem celého výpočtu. Kritérium dobré shody může mít několik variant, my jsme zvolili kvadratické kritérium

$$Q = \sum_{n=1}^{N} \{BEDF(R_{[n]}) - P(R_{[n]})\}^2 \tag{6}$$

. Kritickou hodnotou testu dobré shody Kolmogorova-Smirnova je maximální absolutní rozdíl hodnot teoretické distribuční funkce od obvyklé (stupňovité) empirické distribuční funkce (EDF). K minimalizaci této vzdálenosti však dochází současně s minimalizací největší vzdálenosti teoretické distribuční funkce od bodu dle (5), který půlí příslušný stupeň EDF. Najdeme-li proto volné parametry hledané spojité distribuční funkce tak, aby minimalizovaly vhodný funkcionál jejích vzdáleností od BEDF, přiblížíme se k průběhu, který by zajistil nejlepší shodu ve smyslu zmíněného testu.

Vzniká přirozená otázka, proč by takováto procedura měla konvergovat k rozumnému řešení. Vysvětlení souvisí s jednou základní vlastností globální gnostické distribuční funkce [4]. Je totiž robustní (necitlivá) k outlierům, dokonce i k periferním shlukům dat. V kritériu shody má proto hlavní váhu "většina" reziduí, rezidua s hodnotami blízkými jedné. Celá procedura je tak daleko méně náchylná ke kmitání daleko od optima. Musíme si ovšem i uvědomit, že optimalizační procedury jsou ve složitém případě schopny najít pouze lokální extrémy (procedury pro vyhledávání globálních extrémů nezvládají prohledat celou oblast v rozumném čase). Proto je důležité dobře nastavit první přiblížení, případně analýzu opakovat několikrát z různých výchozích řešení.

# 3 Podrobnější popis metody

## 3.1 Počáteční odhady parametrů modelu

Při prvním přiblížení může být využita apriorní informace, zkušenost s obdobnými daty. Pokud nemáme dostatečnou informaci tohoto typu, použijeme k prvnímu odhadu standardní metodu analýzy nelineární regrese. Pro tento účel použijeme model v aditivním, logaritmizovaném tvaru a parametry $A, \ldots, D$ odhadneme metodou nejmenších čtverců. Musíme se však vypořádat s přítomností cenzorovaných dat. To nám umožní tzv. $EM$ algoritmus. Tento algoritmus a podmínky pro jeho konvergenci jsou zkoumány v [2]. V [3] autoři popisují modifikaci $EM$ algoritmu pro řešení úlohy analýzy lineární regrese s cenzorovanými daty. Při tomto postupu se opakovaně střídají dva kroky: $E$-krok "rekonstruuje" cenzorovaná data (tj. pro $I_k = 0$) v tom smyslu, že odhadne $\hat{Y}_k = E\{Y|Y > Y_k, Z_k, \hat{A}, \ldots, \hat{D}\}$, kde $\hat{A}, \ldots, \hat{D}$ jsou odhady parametrů získané v předchozím $M$-kroku. Další $M$-krok pak prostě najde metodou nejmenších čtverců nový odhad parametrů $A, \ldots, D$ z právě zrekonstruovaných dat $Y_k(I_k = 1), \hat{Y}_k(I_k = 0), Z_k$.

## 3.2 Problém nosiče dat

Gnostické distribuční funkce jsou definovány nad neomezeným intervalem $R_+ := (0, \infty)$. Jak jsme se již zmínili v úvodu, reálná data mohou pocházet z omezeného intervalu $\mathcal{D} := (T_L, T_U)$. Meze tohoto intervalu mohou být neznámé a často představují nejvýznamnější výsledek zkoumání modelu. Tak např. ve zmíněném případě únavových lomů je fyzikální interpretace mezní hodnoty $T_L(Z)$ známá, je to doba (měřená počtem zatěžovacích cyklů) od začátku experimentu do okamžiku, kdy při zátěži $Z$ dojde k tzv. inicializaci trhliny. Od tohoto bodu se totiž bude trhlina rychle zvětšovat až - zpravidla

k nezadržitelnému - lomu součásti. Jde tedy o mez, která by při řádném užívání součásti neměla být překročena a znalost této meze úzce souvisí se zárukami, které může výrobce uživateli poskytnout. V úloze o přežívání pacienta s transplantovaným orgánem mají ovšem velký význam odhady obou mezí. K odhadování mezních hodnot reziduí a tudíž i dob života (a mezních zátěží) je třeba provést vhodnou transformaci $\tau := \mathcal{D} \to \mathcal{R}_+$. Pro transformaci reziduí se osvědčila funkce

$$\rho_k = \frac{R_k - R_L}{1 - R_k/R_U}, \tag{7}$$

kde $R_k$ je skutečné $k-té$ reziduum vypočtené z (1) a (2) a $\rho_k$ je jeho ekvivalent dosazovaný do vzorce pro distribuční funkci definovanou nad nekonečným nosičem.

### 3.3 Optimalizační algoritmus

Kdybychom chtěli zajistit nejlepší shodu distribučních funkcí EDF a P($R$) přímo ve smyslu Kolmogorov-Smirnovova testu, museli bychom řešit nehladkou minimaxní podmíněnou vícerozměrovou úlohu. Nehladkosti úlohy jsme se však vyhnuli použitím bodové empirické distribuční funkce a hladkého kritéria dobré shody (6). Podmíněná extrémalizace tohoto funkcionálu je sice značně jednodušší, nikoliv však jednoduchá. K výběru vhodné metody účinně pomohl dialogový programový systém OPTIA [5] umožňující připojit uživatelský program a odzkoušet v něm kterýkoliv ze širokého výběru minimalizačních algoritmů. Jako nejvhodnější k řešení dané úlohy se ukázala metoda Schittkowského [6]. Tato metoda je rychlá a spolehlivě řeší úlohu lokální vícerozměrové podmíněné minimalizace s omezeními běžných typů. Globální minimalizační metody se ukázaly jako prakticky nepoužitelné pro nepřijatelné časové nároky. Proto se opakovaně použije lokální metoda Schittkowského s různými počátečními podmínkami k ověření, zda nalezené řešení opravdu představuje globální extrém.

## 4 Využití metody

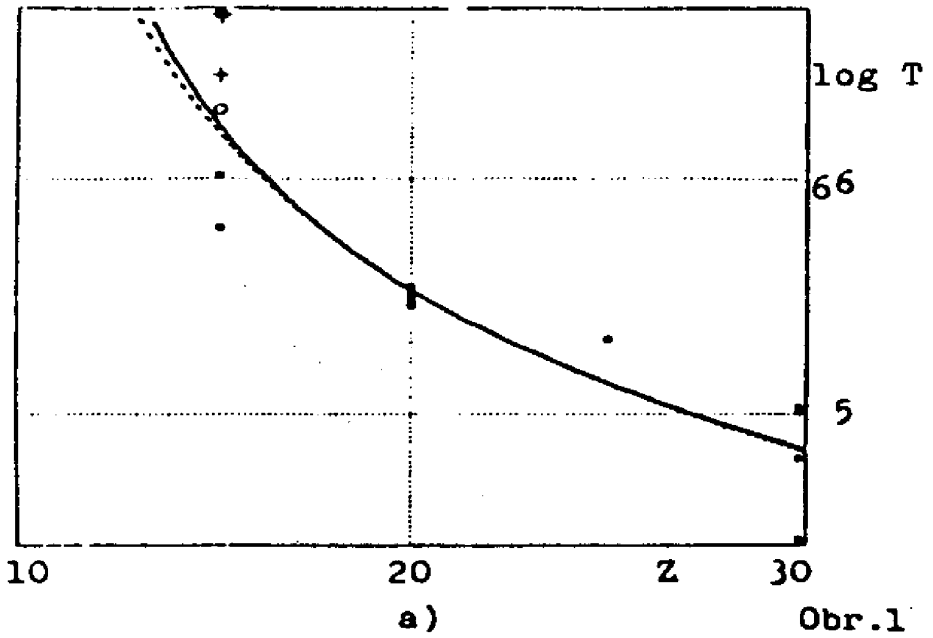
Výzkum a vývoj popisovaných procedur byl motivován potřebami výzkumného oddělení podniku TATRA. Cílem metody bylo zpracování testů životnosti součástek automobilů. Postup byl realizován v ÚTIA ve formě programu GI 5 pro počítače třídy PC/AT. V dohledné době bude k dispozici i verze pro pracovní stanice s operačním systémem UNIX. Formulace úlohy je místy poplatná této konkrétní aplikaci, využitelnost procedur je však značně širší.

### Příklad

Při ověřování metody byla používána data ze životnostních testů na únavové lomy odpovídajících výšeuvedenému popisu. Opakování $EM$ iterací přivedlo k těmto počátečním odhadům parametrů modelu: $\hat{A} = 2.146$, $\hat{B} = 7.622$, $\hat{C} = -10.575$, $\hat{D} = 1881$. Obr. 1a ukazuje regresní křivku s těmito parametry, čárkovaně pak vůbec první odhad, neuvažující cenzorování (vlastně tedy první $M$-krok $EM$ procedury). Cenzorovaná data jsou označena o, jejich rekonstruované hodnoty pak znakem +.

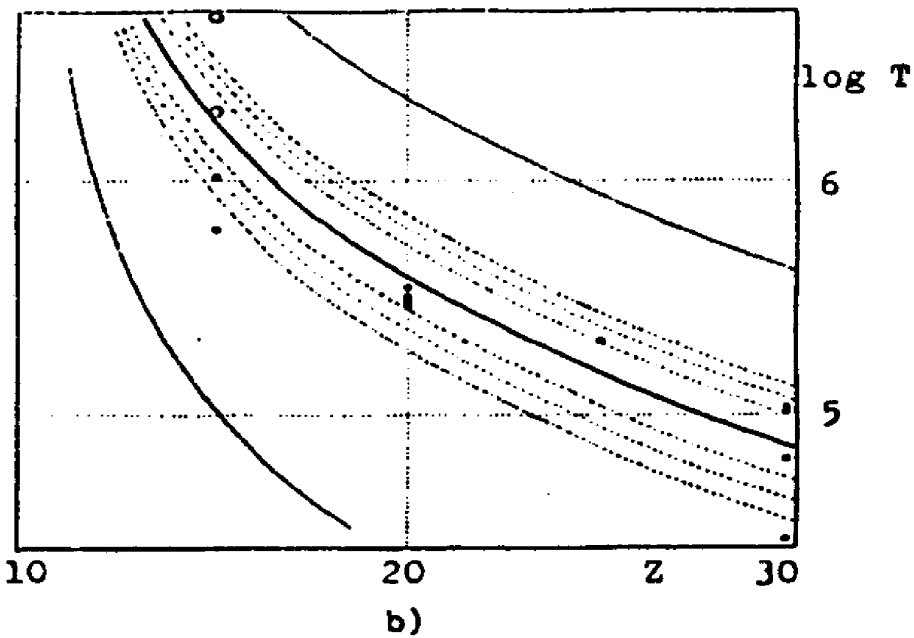
Další optimalizací s použitím popsané metody jsme došli ke konečným odhadům:
$\hat{S}_0 = 1.415$, $\hat{S}_1 = -0.0015$, $\hat{R}_L = 0.194$, $\hat{R}_U = 0.708$, $\hat{A} = 1.815$, $\hat{B} = 7.291$, $\hat{C} = -10,493$, $\hat{D} = 877$.

Obr. 1b zobrazuje výslednou regresní funkci, je to zároveň křivka mediánů globálního gnostického odhadu distribuce $T(Z)$. Na tomto obrázku jsou vyznačeny i 5 %, 10 % a 20 % dolní a horní kvantily

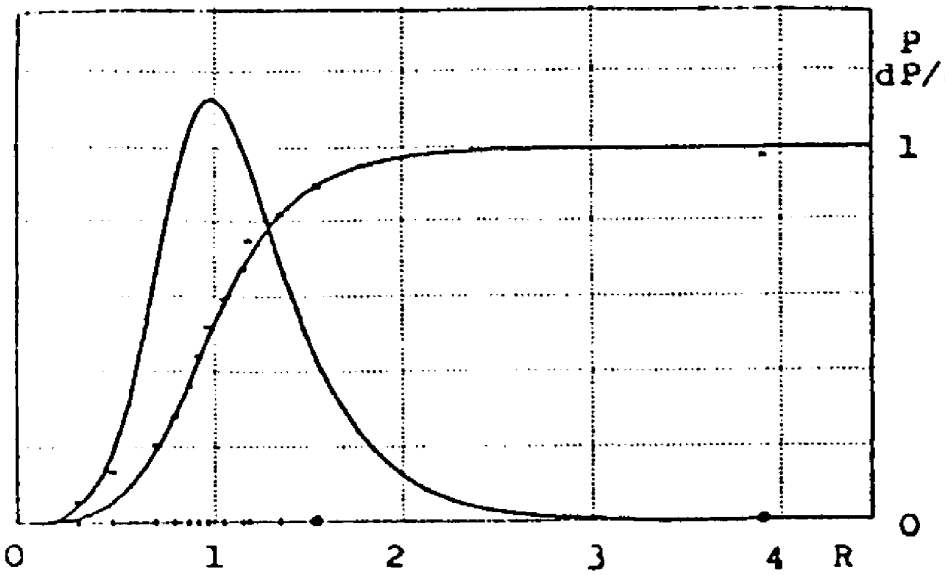



a)

Obr.1

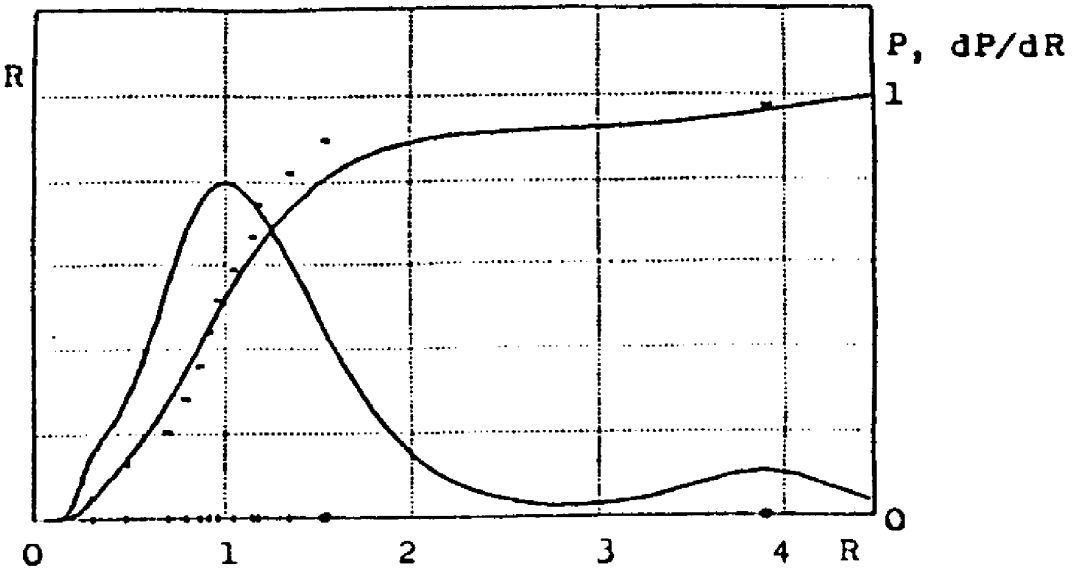



b)




a)

Obr.2




b)

této distribuce, i křivky $T_L(Z)$ a $T_U(Z)$. Obrázek 2 ukazuje tvar gnostického odhadu distribuční funkce a hustoty rozdělení multiplikativních reziduí, a to při hodnotě parametru měřítka odpovídající $Z = 25$. Je dobře vidět, jak lokální (b) varianta reaguje na odlehlé reziduum, zatímco globální odhad distribuce (a) potlačuje jeho vliv. Na ose reziduí jsou cenzorovaná rezidua označena o. Odhadnuté hodnoty bodové empirické distribuční funkce $BEDF$ jsou v obrázku označeny ×.

# Literatura


[1] Kovanic P., A New Theoretical and Algorithmical Basis for Estimation, Identification and Control, Automatica 22, (1986), 6, 657-674

[2] Dempster A. P., Laird N. M., Rubin D. B. , Maximum likelihood estimation from incomplete data via the EM algorithm, J. Roy. Statist. Soc., ser. B 39 (1977), 1–22.

[3] James I. R., Smith P. J., Consistency results for linear regression with censored data, Ann. Statist. 12, (1984), 590–600.

[4] Kovanic P., Robustní odhady distribučních funkcí, Robust '92.

[5] Fidler J., Doležal J., Pacovský J., Dialogue System OPTIA for Minimization of Functions of Several Variables – User's Guide, Inst. of Inf. Theory and Automation, Prague, (1991).

[6] Schittkowski, K., NLPQL: A Fortran Subroutine Solving Constrained Nonlinear Programming Problems, Annals of Operation Research, 5, (1985/6), 485-500 .