Tehnici neparametrice

Bianca Păuna*

Abstract: *Această lucrare prezintă doi estimatori mai puțin cunoscuți pentru vizualizarea funcțiilor de densitate şi a relațiilor dintre variabile. Aceşti estimatori sunt neparametrici in sensul că nu sunt necesare ipoteze a priori privind forma funcțională a dependențelor. In prezentare s-a pus accent pe doi estimatori, estimatorul Kernel al densității şi estimatorul Nadaraya – Watson al relațiilor dintre variabile. Aceste tehnici au fost folosite în continuare pentru a evidenția caracteristicile consumului în România.*

Keywords: estimatori neparametrici, estimatorul Kernel, estimatorul Nadaraya-Watson.

Clasificare JEL: C14

1 Introducere

Graficele sunt instrumente potrivite atunci când se doreşte vizualizarea anumitor caracteristici ale datelor. Dar, datorită numărului de dimensiuni care pot fi vizualizate odată, folosirea graficelor este restricționată. Grafice/Proiecții cu mai mult de trei dimensiuni sunt dificil de reprezentat şi de înțeles. Deşi graficele nu pot surprinde modele complicate ele sunt un punct de plecare pentru orice tip de investigație. Tablele sunt un alt mod de a prezenta informațiile, dar devin foarte mari şi greu de urmărit atunci când se doreşte prezentarea completă a variabilelor şi a relațiilor dintre ele.

Estimarea densităților este în general punctul de plecare al oricărei analize. Există mai multe metode care pot fi folosite pentru obținerea unui estimator al funcțiilor de densitate (histograme, estimatorul neparametric al densității, etc.). O discuție a avantajelor şi dezavantajelor privind cele două metode va fi prezentată în următoarea secțiune.

În afara densităților, relațiile dintre variabile sunt un alt punct de interes. Reprezentarea punctelor pe un grafic poate fi folositoare pentru descoperirea relațiilor dintre variabile numai în cazul în care observațiile nu sunt foarte disperse sau numeroase. Odată cu creşterea numărului de observații, sau a dispersiei punctelor, ochiului uman îi vine din ce în ce mai greu să identifice relațiile dintre variabile, şi de aceea este nevoie de un alt instrument. Alegerea este în acest caz între metodele parametrice şi cele neparametrice.

* Drd., cercetător la Institutul Național de Cercetare Economică, Academia Română.

Proprietățile estimatrului parametric OLS, nedeplasarea, consistența şi eficiența îl fac foarte popular în analizarea datelor. Prețul care trebuie plătit este în materie de ipoteze legate de distribuții şi specificarea modelului, dar aceasta este o problemă numai în cazul în care forma funcțională nu este cunoscută. Alternativ pot fi folosite metode neparamtrice. Estimatorii sunt polarizați, dar sunt consistenți, şi cel mai important, nu este nevoie să se facă nici un fel de ipoteze *apriori* privind forma dependenței, deci sunt un instrument foarte adecvat în analizele datelor.

Estimarea parametrică este instrumentul adecvat în cazurile în care se ştie că ipotezele pe care se bazează construcția modelului sunt satisfăcute. În cazul în care modelul nu este bine specificat, estimatorii coeficienților vor fi deplasați şi neconsistenți, în timp ce estimatorii neparametrici vor fi consistenți şi asimptotic nedeplasați.

Un exemplu este cazul estimării dependenței dintre variabila dependentă (***Y***) şi cea explicativă (***X***). Estimatorul OLS presupune existența unei dependențe liniare şi constă într-o line trasată printre puncte, astfel încât suma pătratului erorilor să fie minimă. Estimând coeficienții prin metoda OLS se obțin cei mai buni estimatori al unei dependențe liniare, sau liniarizate.

În cazul în care nu se cunoaşte cu exactitate forma relației dintre variabile, este necesară o alternativă mai puțin rigidă, cum este cazul metodelor neparametrice. Metodele neparametrice au apărut datorită nevoii de a evita aplicarea unei dependențe funcționale rigide modelului, deci termenul de neparametric se referă în acest caz la forma funcțională flexibilă. Exemple tipice de folosire a tehnicilor neparametrice este în construirea modelului, verificarea, inferențe şi predicție.

Echivalentul neparametric al OLS-ului este estimatorul neliniar, şi ideea este ca în loc de a se impune forme funcționale rigide, datele sunt lăsate să indice forma funcțională a dependenței dintre cele două variabile. Scopul tehnicilor neparametrice este de a contribui la reprezentarea relației adevărate dintre două variabile. Metoda cea mai la îndemâna este unirea punctelor, dar graficul care ar rezulta ar fi mult prea eratic pentru a fi de folos, de aceea orice tehnică neparametrică implică şi un proces de netezire. Procesul de netezire implică un compromis, graficul trebuie să fie suficient de neted pentru a putea descrie relația dintre cele două variabile, fără a renunța la toată variația locală. Alegerea nivelului de netezime a graficului se poate face subiectiv prin comparație între mai multe grafice cu diverse grade de netezire, sau cu ajutorul metodelor statistice.

Ideea care stă la baza estimatorul neliniar este de a calcula variabila dependentă în fiecare punct folosind informația disponibilă în vecinătatea punctului. Astfel, variabila dependentă este obținută ca o medie a valorilor funcției în vecinătate. Această procedură va produce estimatori polarizați, şi ca exemplu se consideră cazul estimării valorii funcției la maxim sau minim. Toate punctele din jurul maximului (minimului) au valori mai mici (mari) deci valoarea rezultată va fi mai mică (mare) decât valoarea reală. Este posibil să se reducă valoarea polarizăii prin reducerea dimensiunii intervalului, dar odată cu scăderea intervalului, erorile din date vor avea o influență mai mare asupra estimatorului, deci variația acestuia va creşte. Deci, procedura de alegere a intervalului optim este şi ea rezultatul unui compromis între valoarea polarizării şi a variației estimatorului.

2 Estimatorul densității

Cea mai simpla analiză neparametrică este estimarea unei funcţii de densitate a unei variabile $\boldsymbol{X_i}$. Aceasta constă în găsirea probabilităţilor asociate fiecărei valori din domeniul de definiţie a variabilei $\boldsymbol{X_i}$. Pentru obţinerea estimatorului trebuie în primul rând împărţit domeniul funcţiei în intervale egale, după care se numără datele din fiecare interval.

Pentru un interval de dimensiune *h* estimatorul histogramei este de forma următoare:

$$\hat{f}(x) = \frac{1}{nh}(\textit{numărul de puncte din interval})$$

Mai formal, relaţia de calcul a histogramei poate fi scrisă astfel:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j} I(X_i \in B_j) I(x \in B_j)$$

Factorii care influenţează forma histogramei sunt atât punctul de origine a intervalului, cât şi dimensiunea intervalului. Controlul netezimii curbei de histogramă este dat de dimensiunea intervalului. Această dependenţă a formei histogramei de cei doi factori este una din dezavantajele folosirii lor. Pentru corectarea acestei dependinţe se poate folosi metoda medierii histogramelor cu origine diferită. Aceasta constă în estimarea mai multor histograme (care au originea intervalului diferită) pentru dimenisiunea optimă a intervalui, iar histograma estimată este obţinută prim medierea tuturor histogramelor.

Histogramele prezintă unele dezavantaje care nu o fac foarte răspândită în aplicaţii. Prin definiţie, estimatorul este discontinuu: are valori constante în fiecare interval, dar are salturi la capetele intervalului, deci este un instrument nepotrivit mai ales în cazurile în care se doreşte obţinerea derivatelor de ordinul întâi.

Un alt estimator al densităţilor este estimatorul naiv. Acesta se obţine pornind de la definiţia funcţiei de densitate: probabilitatea ca $x \in [x-\varepsilon,\ x+\varepsilon]$ atunci când $\varepsilon \to 0$. Pentru estimarea probabiltăţii se foloseşte proporţia de puncte care se regăsesc în intervalul $[x-\varepsilon,\ x+\varepsilon]$.

$$\hat{f}(x) = \frac{1}{nh}(\textit{număr de } X_i \textit{ din } [x-h/2,\ x+h/2]$$

Atât histograma cât şi estimatorul naiv au definiţii similare, dar diferă din punct de vedere al calculelor făcute. Histograma are valori constante pe fiecare interval, şi salturi la capetele acestuia. Parametrul *h* controlează numărul de intervale. Estimatorul naiv este calculat pentru fiecare valoare a lui *x*. Gradul de netezire al graficului este controlat prin dimensiunea lui *h*. Din punct de vedere computaţional, histograma necesită mai puţine operaţii la estimare, în timp ce estimatorul naiv calculează pentru fiecare observaţie o valoare a densităţii, fiind mai intensiv din punct de vedere computaţional.

Estimatorul desității poate fi scris mai compact dacă se defineşte o funcție de pondere: $w(u)=\frac{1}{2}\cdot I(|u|\le 1)$ unde $I(\bullet)$ este funcția indicator care ia valoarea 1 atunci când este expresia este adevărată şi zero când este falsă:

$$\hat{f}_h(x)=\frac{1}{n}\cdot\sum_{i=1}^{n}\frac{1}{h}w(\frac{x-X_i}{h})$$

Funcția de pondere dă ponderi egale fiecărei observații din interval, deci estimatorul densității este o funcție discontinuă care are salturi în punctele $X_i\pm h$ şi are derivata egală cu zero în rest.

2.1 Estimatorul Kernel al densității

Discontinuitățile estimatorului naiv pot fi corectate cu uşurință prin alegerea unei forme funcționale diferite pentru ponderi. O funcție care să corecteze discontiunitățile trebuie să satisfacă două proprietăți:

$K(u)\to 0$ *as* $|u|\to 1$ and

$$\int_{-\infty}^{\infty}K(x)dx=1$$

Prima condiție asigură continuitatea funcției de densitate estimată, prin forțarea ponderilor spre zero la sfârşitul intervalului. Aceasta garantează continuitatea funcției de densitate pe întreg domeniul, o proprietate importantă atunci c□nd se doreşte estimarea derivatei de ordin întâi.

Cea de-a doua condiție garantează că suma ponderilro este 1, şi în consecință satisfacerea condiției pe care toate funcțiile de densitate trebuie să o satisfacă şi anume integrala funcției pe întreg domeniul de definiție să fie 1. Cum cea de a doua condiție este similară cu definiția funcției de densitate, de multe ori pentru funcția de ponderi este aleasă o denisitate cunoscută.

Ordinul funcției kernel este definit ca primul moment diferit de zero. Deci un kernel este de ordinul p dacă $k_p=\int u^p K(u)du\neq 0$. Un kernel pozitiv poate fi maximum de ordinul 2. Deoarece în estimarea densităților funcția este restricționată la valori pozitive, kernelul de ordinul 2 este cel mai popular în aplicații. Kernelii de ordin mai mare sunt folosiți şi ei în anumite cazuri, care vor fi prezentate în această secțiune.

Tin□nd cont de toate observațiile, un estimator kernel al funcției de densitate este de forma următoare:

$$\hat{f}_h(x)=\frac{1}{nh}\cdot\sum_{i=1}^{n}K(\frac{x-X_i}{h})$$

unde h este dimensiunea intervalului.

O alegere posibilă pentru funcția de ponderi este funcția de densitate normală, care fiind o funcție de densitate, satisface condițiile enumerate mai sus. Totuşi distribuția normală nu are suport finit, ceea ce înseamnă că pentru fiecare la calculul densității în fiecare punct se foloseşte informația de la toate punctele pentru care există din valori ale funcției. Acest lucru creşte costurilor din punct de vedere computațional, şi din acest motiv funcțiile kernel ale căror valori nu sunt într-un interval finit, nu sunt foarte răspândite în practică.

Una din cele mai răspândite funcții de ponderi este kernelul Epanechnikov care ia forma următoare:

$$K(u)=\frac{3}{4}\cdot(1-u^2)\cdot I(|u|\leq 1)$$

Folosirea funcțiilor de pondere care iau valori pe un domeniu infinit (cum este cazul densității normale) rezultă în creşterea semnificativă număului de operații necesare la estimare. La fiecare valoare, trebuie calculate ponderi pentru toate observațiile, deci un total de n^2 operații. Când se folosesc funcții de pondere finite numărul de operații necesare pentru estimare scade semnificativ la nh. Chiar în cazul folosirii ponderilor finite, atunci când numărul de date este foarte mare, aspectele computaționale pot fi prohibitive.

2.2 Proprietățile statistice ale estimatorului kernel al densității

Proprietătile statistice sunt criteriul cel mai răspânit pentru clasificarea diverşilor estimatori. Estimatorul neparametric este funcție de date, dar forma estimatului este sensibilă la alegerea kernelului şi a dimensiunii intervalului. Din acest motiv, proprietățile statistice (deplasarea şi variația) sunt folosite ca instrument de decizie pentru alegerea ponderilor şi a gradului de netezire.

Proprietățile estimatorului într-un punct sunt judecate prin folosirea MSE[†] (media pătratelor erorilor) care este definită în felul următor:

$$MSE_x(\hat{f})=E\{\hat{f}(x)-f(x)\}^2$$

Dacă se scrie MSE în funcție de deplasare şi de variație se obține expresia:

$$MSE_x(\hat{f})=\{E\hat{f}(x)-f(x)\}^2+var\hat{f}(x)$$

Integrând MSE pe întregul domeniu de definiție se obține MISE (media integrată a pătratelor erorilor). MISE este o măsură a acurateții globale a estimatorului şi are următoarea definiție:

$$MISE(\hat{f})=E\int\{\hat{f}(x)-f(x)\}^2dx$$

Prin aplicarea proprietăților integralelor şi cele ale operatorului de aşteptare, MISE poate fi rescrisă astfel:

[†] MSE – abrevierea lui Mean Squared Error.

$$\text{MISE}(\hat{f}) = \int E\{\hat{f}(x) - f(x)\}^2 dx = \int MSE(\hat{f}) dx$$

$$= \int \{E\hat{f}(x) - f(x)\}^2 dx + \int \text{var}\,\hat{f}(x) dx$$

Dimensiunea optimă a intervalului de estimare se obține ca o soluție a maximizării MISE. Pentru aceasta, MISE trebuie scrisă ca o funcție de dimensiunea intervalului. Modul de derivare a dimensiunii intervalului va fi prezentat în secțiunea următoare.

Estimatorul densității kernel are forma următoare[‡]:

$\hat{f}(x) = n^{-1} \sum K_h((x-y)/h)$ where $K_h = n^{-1} K((x-y)/h)$

$E(\hat{f}) = n^{-1} E\{\sum K_h((x-y)/h)\} = n^{-1} \sum E\, K_h((x-y)/h) = E\, K_h((x-y)/h)$

$var\,\hat{f}(x) = var\{n^{-1} \sum K_h((x-y)/h)\} = n^{-2}\, var \sum K_h((x-y)/h) = n^{-1}\, var\, K_h((x-y)/h).$

Pentru calculul MISE este necesar să se obțină expresii ale deplasării şi ale variației.

$deplasarea(x) = E\,\hat{f}(x) - f(x) = \int h^{-1} K((x-y)/h)\, f(y) dy - f(x)$

Se face următoare schimbare de variabile: $y = x - ht;\ dy = hdt$ şi pentru că $\int K(u) du = 1$, deplasarea se poate rescrie:

$deplasarea(x) = \int K_t(t) \{f(x-ht) - f(x)\} dt$

Prin folosirea descompunerii Taylor:

$$f(x-ht) = f(x) - ht\, f'(x) + \frac{1}{2}(ht)^2 f''(x) + \ldots$$

şi după simplificări deplasarea poate fi rescrisă astfel:

$$deplasarea(x) = \int K_t(t) \frac{1}{2} (ht)^2 f''(x)\, dt + O(h^3) \cong \frac{1}{2} h^2\, f''(x) \int t^2 K_t(t)\, dt = \frac{1}{2} h^2\, f''(x)\, k_2$$

Prin definiție variația poate fi scrisă astfel:

$$var\,\hat{f}(x) = n^{-1}\, var\, K_h((x-y)/h) = E\{K_h((x-y)/h) - E\, K_h((x-y)/h)\}^2 =$$

$$= n^{-1} E\{K_h^2((x-y)/h) + [E\, K_h((x-y)/h)]^2 - 2 K_h((x-y)/h)\, E\, K_h((x-y)/h)\} =$$

$$= n^{-1} \{E\, K_h^2((x-y)/h) + [E\, K_h((x-y)/h)]^2 - 2E\, K_h((x-y)/h)\, E\, K_h((x-y)/h)\} =$$

$$= n^{-1} \{E\, K_h^2((x-y)/h) - [E\, K_h((x-y)/h)]^2\}$$

$$var\,\hat{f}(x) =) = n^{-1} \int h^{-2} K^2((x-y)/h)\, f(y) dy - n^{-1} \{\int h^{-1} K((x-y)/h)\, f(y) dy\}^2$$

Se efectuează aceaşi schimbare de variabilă şi se foloseşte descompunerea Taylor la fel ca mai sus:

[‡] Demonstrațiile urmăresc Silverman, 1986, pg38.

$$var\,\hat{f}(x)=n^{-1}\int h^{-1}K^2(t)\,f(x-ht)dt - n^{-1}\{f(x)+bias(x)\}^2 =$$

$$=n^{-1}h^{-1}\int K^2(t)\{f(x)-ht\,f'(x)+\frac{1}{2}h^2t^2\,f''(x)\;\ldots\}dt-n^{-1}\{f(x)+bias(x)\}^2$$

La descompunerea Taylor, cel de-al doilea termen este zero, iar pentru *n* mari şi intervale *h* mici, cel de-al treilea este aproximativ egal cu zero. Cea de a doua expresie are ordin de mărime n^{-1} şi deci poate fi aproximată şi ea cu zero:

$$var\,\hat{f}(x) \cong n^{-1}h^{-1}\int K^2(t)f(x)dt - O(n^{-1}) \cong n^{-1}h^{-1}f(x)\int K^2(t)\,dt$$

Expresia obţinută pentru deplasare nu conţine *n* deci deplasarea este influenţată numai de dimensiunea intervalului şi de forma funcţională a kernelului. Variaţia estimatorului este funcţie de dimensiunea intervalului, forma funcţională a kernelului şi numărul de observaţii. Cu cât numărul de observaţii este mai mare, cetteris paribus, estimatorul are variaţia mai mică. Dimensiunea intervalului influenţează deplasare şi variaţia în mod opus. Cu cât dimensiunea intervalului este mai mare, deplasarea creşte, dar variaţia scade, ceea ce sugerează prezenţa unui compromis între dimensiunea deplasării şi a variaţiei la alegerea dimensiunii intervalului.

Proprietăţile asimtotice ale estimatorului sunt cele care indică evoluţia deplasării şi a variaţiei cu creşterea numărului de observaţi. Consistenţa estimatorului este definită astfel:

Dacă o funcţie kernel satisface următoarele condiţii[§]:

- $\int |K(u)|\,du < \infty$ and $\int K(u)du = 1$
- $lim_{|u|\to\infty} uK(u)=0$
- $EY^2 < \infty$
- $n\to\infty,\ h_n \to \infty,\ nh_n \to \infty$

dacă f este continuu în x, $\hat{f}(x) \xrightarrow{p} f(x)$ atunci când n $\to \infty$.

Majoritatea funcţiilor Kernel satisfac condiţiile de mai sus, ceea ce înseamnă că estimatorul densităţii kernel este un estimator consistent.

Deplasarea nu depinde de numărul de observaţii. Consistenţa este proprietatea care ne oferă garanţia că odată cu creşterea numărului de observaţii, estimatorii sunt din ce în ce mai apropiaţi de valoarea reală. Această proprietate nu spune însă nimic de viteza de convergenţă.

[§] Această proprietate a fost demonstrată de Parzen în 1962.

2.3 Selectarea dimensiunii optime a intervalului şi a formei funcției kernel

Dimensiunea optimă a intervalului este, aşa cum am arătat deja ca un compromis între deplasare şi variație. Pentru a se ajunge la o dimensiune optimă trebuie aleasă o funcție obiectiv de deplasare care să fie minimizată. Un candidat pentru această funcție obiectiv este MISE. Aceasta depinde atât de alegerea intervalului maxim cât şi de forma funcțională a funcției de ponderi. Prin introducerea expresiei deplasării şi variației în formula lui MISE se obține următoarea expresie:

$$MISE(x)= \{\frac{1}{2} h^2 k_2\}^2 \int f''(x)^2\, dx + n^{-1} h^{-1} \int f(x)\, dx \int K^2(t)\, dt =$$

$$= \frac{1}{4} h^4 k_2^2 \int f''(x)^2\, dx + n^{-1} h^{-1} \int K^2(t)\, dt$$

Minimizarea MISE conduce la următoarea expresie pentru dimensiunea optimă a intervalului:

$$h_{opt} = k^{-2/5} n^{-1/5} \{\int K^2(t)\, dt\}^{1/5} \{\int f''(x)^2\, dx\}^{-1/5}$$

Intervalul optim este dependent de forma funcțională a kernelului precum şi de densitatea care trebuie estimată prin expresia: $\int f''(x)^2\, dx$. Ca să se obțină o formulă analitică pentru calculul intervalului optim este necesar să se facă nişte ipoteze legate de distribuția densității.

Derivarea intevalului optim se va face pornind de la ipoteza că densitatea de estimat este normală. Alternativ, există şi metode iterative, care nu necesită specificarea formei densității, pentru obținerea dimensiunii optime a intervalului.

Introducând expresia dimensiunii optime a intervalului în formula MISE se ajunge la următoarea expresie:

$$MISE = \frac{5}{4} n^{-4/5} k_2^{2/5} \{\int f''(x)^2\, dx\}^{1/5} \{\int K^2(t)\, dt\}^{4/5}$$

Pentru un interval de dimensiune optimă, singura modalitate de a scădea MISE este prin alegerea unei forme funcționale a kernelului care minimizează MISE. În formula MISE singura parte care depinde de alegerea kernelului este următoarea:

$$C(K) = k_2^{2/5} \{\int K^2(t)\, dt\}^{4/5}$$

Kernelul optim este cel pentru care se obține valoarea MISE cea mai scăzută, condiție care se reduce la kernelul cu valoarea *C(K)* cea mai mică. Kernelul Epanenchnikov este cel care are valoarea *C(K)* cea mai redusă.

Comparația dintre diverse forme funcționale de kerneli se face prin calcularea raportului *C(K)* față de *C(K)* al kernelului Epanechnikov. Acest raport este destul de aproape de 1, în cazul densității normale este 0.95. Concluzia care se impune este că posibilitățile de îmbunătățire a MISE prin alegerea unei anumite forme de kernel sunt limitate. Din acest

motiv considerentele care predomină la alegerea formei funcționale a ponderilor sunt legate de aspectele computaționale.

Vom deriva în continuare formula pentru dimensiunea intervalului optim în cazul în care densitatea ce se doreşte estimată este normală. Densitatea normală are următoarea formă funcțională:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$

$$\int f''(x)^2\,dx=\int \frac{1}{2\pi\sigma^2}\frac{(x-\mu)^4}{\sigma^8}e^{-(x-\mu)^2/\sigma^2}dx$$

Se efectuează schimbarea de variabilă $z=(x-\mu)/\sigma$, şi $dz=dx/\sigma$ şi se obține următoarea expresie pentru integrală:

$$\int f''(x)^2\,dx=\int \frac{1}{2\pi\sigma^5}z^4e^{-z^2}dz=\frac{1}{\sigma^5}\int\frac{1}{2\pi}z^4e^{-z^2}dz=\frac{1}{\sigma^5}\int z^4\phi''(z)dz$$

unde $\phi''(z)$ este funcția de densitate standard normală.

Rezolvarea integralei se face prin părți, ținând cont şi de proprietatea că: $\int\phi''(z)dz=1$.

$$\int f''(x)^2\,dx=\frac{3}{8\pi\sigma^5}$$

Prin înlocuirea valorii integralei în formula dimensiunii optime a intervalului se obține expresia intervalului optim a lui Silverman, care are următoarea formă:

$$h_{opt}=1.364\,\sigma\cdot n^{-1/5}\cdot\left(\frac{\int K^2(t)dt}{k_2^2}\right)^{1/5}$$

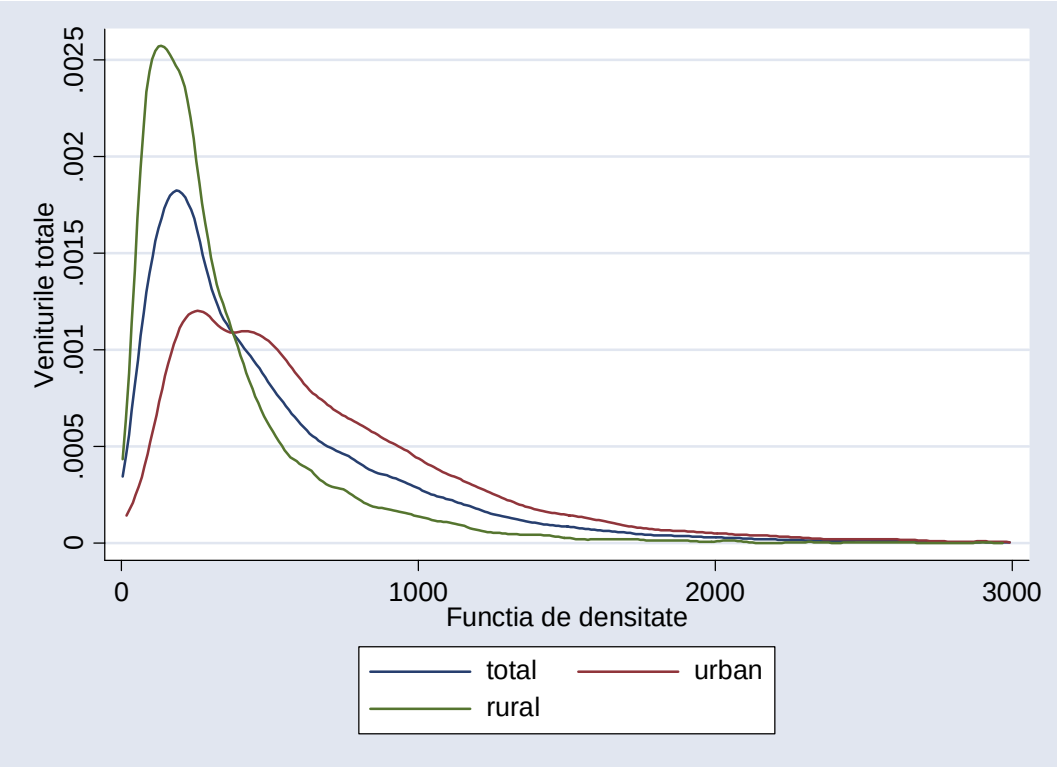
Derivarea expresiei intervalului optim Silverman este făcută în cazurile în care distribuția care se doreşte estimată este normală, dar se poate folosi fără erori majore în toate cazurile în care distribuția este uni-modală. Se poate spune deci că într-un număr mare de cazuri aproximarea Silverman este suficientă şi dă un punct de plecare bun pentru estimarea densităților.

2.4 Reprezentarea distribuțiilor variabilelor de interes

Estimatorul densității neparametrice a fost folosit pentru a studia consumul gospodăriilor din România. Datele analizate provin din Ancheta Integrată în Gospodării 2003, un sondaj pe care Institutul de Statistică îl face în fiecare an, începând cu 1995.

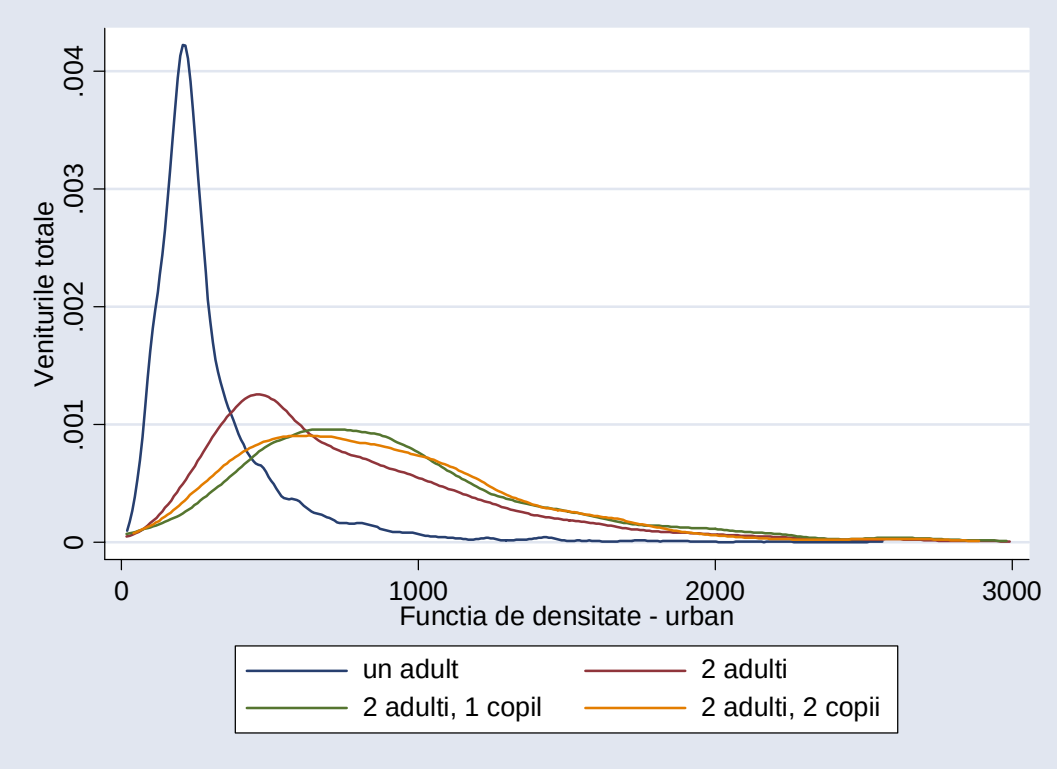
Graficul 2-1 prezintă funcția de densitate a veniturilor totale pentru toate gospodăriile şi separat pentru cele din rural şi urban. Se poate observa o diferență împortantă între cele două distribuții, veniturile din zona urbană sunt, sensibil mai mari decât cele din zona rurală. Diferența poate fi cauzată de mai mulți factori, şi nu avem suficiente informații

pentru a identifica cu exactitate care factor este responsabil cu discrepanțele privind veniturile între mediu urban şi rural. Dintre factorii responsabili menționăm, caracteristici demografice diferite ale gospodăriilor din cele două medii, care ar duce la o structură diferită a venitului total (mai mulți membri în vârstă de muncă, deci mai multe salarii), caracteristici ale forței de muncă diferite în cele două regiuni, ca de exemplu nivel de educație, calificări, etc. (salarii mai mari pentru salariații din zona urbană datorită prestării unei activități mai productive), dar şi din punct de vedere al ocupării şi nu în ultimul rând posibile diferențe salariale între cele două zone, chiar pentru munci similare, datorită condițiilor de pe piața muncii.

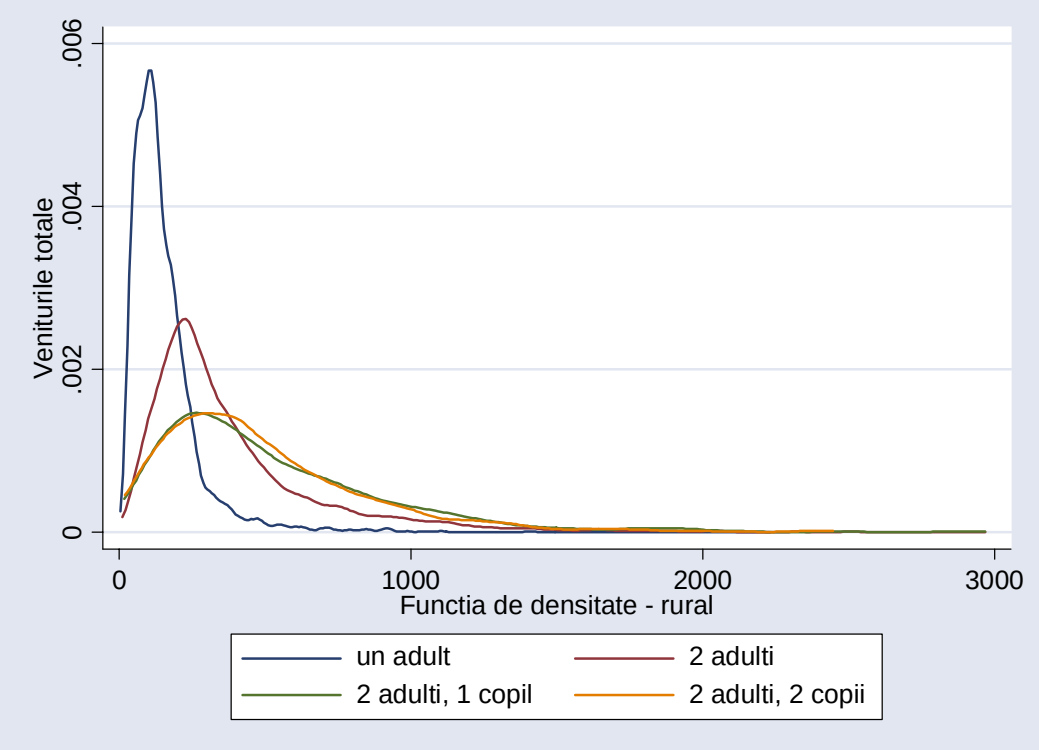


Graficul 2-1 Funcția de densitate a veniturilor totale (RON)

Următoarele două grafice investighează ipoteza că diferențele între veniturile totale ale familiilor din cele două regiuni ar putea fi explicate de diferențe în numărul de persoane care compun familia. Şi în acest sens a fost reprezentată grafic funcția de densitate separat pentru zona urbană şi rurală şi separat pentru gospodăriile formate dintr-un adult, doi adulți, doi adulți şi un copil, şi doi adulți şi doi copii, aceste tipuri de familii fiind cele mai reprezentate în eşantion.

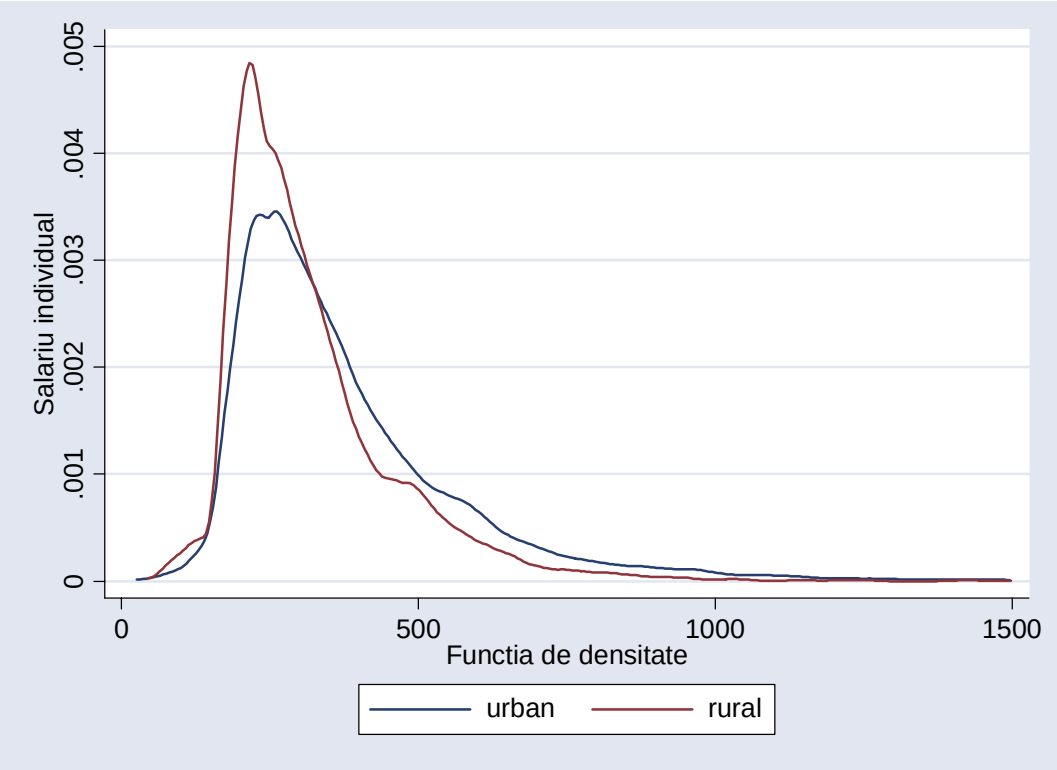


Graficul 2-2 Veniturile totale în funcție de componența familiei, în zona urbană (RON)



Graficul 2-3 Veniturile totale în funcție de componența familiei, în zona rurală (RON)

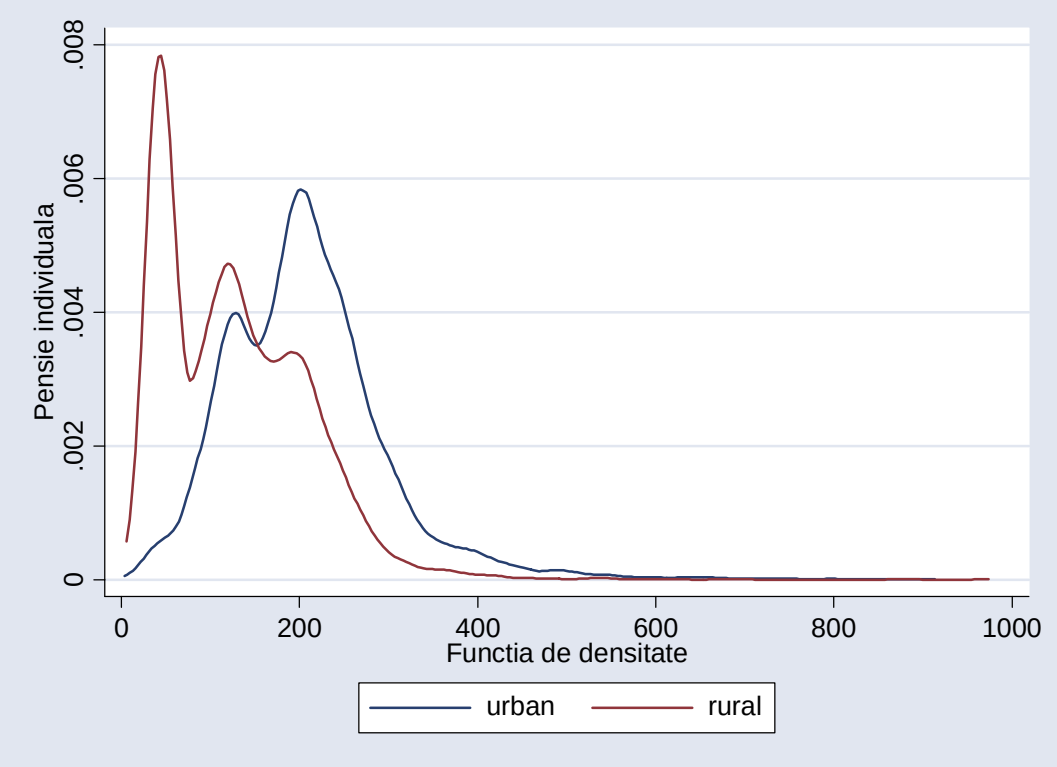
Din Graficul 2-2 şi Graficul 2-3 se poate observa că atunci c□nd se controlează pentru structura familiei diferenţele privind veniturile totale se menţin, cu alte cuvinte caracteristicile demografice ale populaţiei din zona urbană şi rurală din punct de vedere al numărului de membri ai familiei nu au o infulenţă decisivă asupra diferenţelor de venituri între cele două zone.



Graficul 2-4 Funcţia de densitate a salariilor pentru urban şi rural (RON)

Salariile individuale sunt prezentate în Graficul 2-4. Din compararea funcţiei de distribuţie pentru salariilor urbane şi rurale se pot observa anumite diferenţe □ntre cele două funcţii: o dispersie mai mare a salariilor urbane faţă de cele urbane (distribuţia salariilor rurale este mai înaltă ceea ce indică un număr mai mare de indivizi cu salarii apropiate de medie), şi de asemenea o medie marginal mai mare a acestora. Totuşi distribuţiile salariale sunt relativ apropiate, şi nu pot fi la originea diferenţelor care au fost observate la veniturile totale.

Graficul 2-5 prezintă distribuţia pensiilor urbane şi rurale.



Graficul 2-5 Funcția de densitate a pensiilor pentru urban şi rural (RON)

În cazul pensiilor situația este alta în sensul că funcțiile de distribuție ale pensiilor pentru zona urbană şi cea rurală sunt foarte diferite. În cazul pensiilor rurale distribuția este tri-modală, cu vârful cel mai mare la valori ale pensiilor foarte mici, care corespund pensiilor lucrătorilor în agricultură. distribuția pensiilor urbane are numai două vârfuri, datorită absenței pensiilor agricole în zona urbană. Al doilea vârf al distribuției pentru zona rurală corespunde primului vârful al distribuției pensiilor urbane din punct de vedere al valorii pensiei. La pensia urbană se poate observa că distribuția atinge maximul la valori ale pensiilor superioare pensiilor rurale.

Două concluzii pot fi deduse din cele două distribuții, în primul rând **diferențele dintre venituri pot fi explicate măcar parțial prin diferențele de pensie**, dar motivația existenței diferențelor între pensiile urbane şi rurale stă în statutul ocupațional al persoanelor care locuiesc în cele două regiuni. O proporție mai mare de persoane din zona rurală au lucrat şi încă lucrează în agricultura de subzistență, activitate care generează pensii şi venituri mici. Interesant este că în rest **diferențele între salarii şi pensii între cele două zone sunt marginale**. De aceea se pare că în zona urbană o proporție mai mare de persoane active este responsabilă pentru discrepanțele dintre veniturile totale ale gospodăriilor.

3 Regresia neparamterică

La regresia neparametrică se doreşte să se modeleze relaţia dintre variabila dependentă şi un set de variabile explicative. Vom începe cu cazul unei singure variabile dependente şi explicative. Pentru exemplificare să considerăm modelul:

$Y_i=m(X_i)+\varepsilon_i$

Se doreşte derivarea curbei de regresie care pote fi scrisă astfel:

$m(x)=E(Y|X=x)$

Cel mai simplu estimator poate fi obţinut prin calculul mediei variabilei dependente pe diferite intervale. În acest caz, funcţia estimată este constantă pe intervalul respectiv, dar are discontinuităţi la capetele intervalului. Acesta nu este un estimator potrivit, decât în cazul funcţiilor care manifestă asemenea salturi. Acest estimator este o generalizare a histogramei, si este denumit, prin analogie, regresogramă.

Dacă la fiecare valoare a variabilei explicative ar exista mai multe valori pentru variabila dependentă, estimatorul poate fi calculat ca medie a variabilei răspuns în fiecare punct. În cele mai frecvente cazuri, există numai o observaţie pentru fiecare valoare, în acest caz, valoarea funcţiei este estimată prin calculul mediei ponderate a variabilei dependente într-un interval mic în jurul punctului. Acest estimator specific este denumit estimatorul mediei mobile.

Deşi regresograma şi estimatorul mediei mobile sunt calculaţi similar prin medierea variabilei de răspuns pe un interval, există o diferenţă majoră între cei doi estimatori. Regresograma are valoare constantă pe un interval, numărul de intervale considerate dă număul de valori diferite pe care le poate lua funcţia. Estimatorul mediei mobile calculează valoarea funcţiei în fiecare punct al variabilei independente prin medierea unui anumit număr de valori ale variabilei dependente, deci în timp ce regresograma este o funcţie care are salturi pe intervale, estiamatorul mediei mobile este o funcţie continuă. .

3.1 Estimatorul Kernel

Estimatorul kernel pentru dependenţele dintre două variabile aplică principiile estimatorului kernel al densităţii pentru estimarea dependenţelor dintre două sau mai multe variabile. Ca şi în cazul estimatorului kernel al densităţii, estimatorul calculează o medie ponderată a variabilei dependente în fiecare punct.

$$\hat{m}(x)=\frac{1}{n}\cdot\sum_{i=1}^{n}W_{ni}(x)Y_i$$

iar ponderile $\{W_{ni}(x)\}_{i=1}^{n}$ sunt calculate cu ajutorul funcţiei kernel astfel:

$$W_{ni}(x) = \frac{K_{hn}(x - X_i)}{\hat{f}_{hn}(x)}$$

unde $\hat{f}_h(x)$ este estimatorul kernel al funcției de densitate şi *K(•)* este funcția kernel:

$$\hat{f}_h(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} K_h(x - X_i)$$

$K_u(\bullet)=h^{-1} \cdot K(\bullet/h)$.

Estimarea constă în calculul unor ponderi pentru fiecare observație din intervalul *[x-h,x+h]* la fiecare valorare a variabilei explicative. Cu excepția cazurilor în care kernelul este o funcție constantă, adică ponderile sunt egale pentru fiecare observație din intervalul *[x-h,x+h],* ponderile depind de distanța punctelor față de *x*. Deoarece suma ponderile trebuie să fie unu, acestea sunt împărțite cu densitatea variabilei explicative în acel punct.

După efectuarea tuturor subtituțiilor estimatorul kernel ia forma următoare:

$$\hat{m}_h(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}$$

Acest estimator este cunoscut şi sub numele estimatorul Nadaraya – Watson, botezat cu numele primelor persoane care au recomandat folosirea acestuia.

Proprietățile funcțiilor kernel au fost descrise în secțiunea anterioară, dar reamintim aici pe cele importante: funcțiile kernel sunt simetrice, pozitive (pentru kenelii de ordin 2, cei mai uzuali în practică) şi integrala lor trebuie să fie unu. Datorită aspectelor computaționale, cei mai folosiți sunt kernelii cu suport finit.

3.2 Proprietătile statistice ale estimatorului Nadaraya – Watson

În această secțiune se vor investiga proprietătile statistice şi asimtotice ale estimatorului Nadaraya – Watson. La derivarea formulei deplasării şi a variației, se va urmări raționamentul prezentat în Scott (1992) pg. 223. Derivarea formulei deplasării şi a variației este mai dificilă în acest caz datorită formei estimatorului definit ca raport a două variabile aleatoare corelate.

În derivații s-a folosit următoarea proprietate:

Dacă numărătorul şi numitorul unei fracţii converg către o constantă pozitivă, atunci expectaţiile asimptotice ale raportului este egal cu raportul expectaţiilor astimtotice ale numărătorului şi numitorului[**].

$$\hat{m}_h(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)}$$

Deja se cunoaşte valoarea aşteptaţa a numitorului, pentru că acesta este estimatorul kernel a densităţii, iar forumla a fost obţinută în secţiunea anterioară.

$E\hat{f}(x) = f(x) + \frac{1}{2}h^2 f''(x) k_2$

$var\hat{f}(x) = n^{-1} h^{-1} f(x) \int K^2(t)\, dt$

În continuare vom calcula valoarea aşteptată pentru numărător:

$$E\left(\frac{1}{n}\sum_{i=1}^{n} y_i \frac{1}{h} K\left(\frac{x - x_i}{h}\right)\right) = \frac{1}{n}\sum_{i=1}^{n} y_i E\left(\frac{1}{h} K\left(\frac{x - x_i}{h}\right)\right) = \sum_{i=1}^{n} y_i E\left(\frac{1}{h} K\left(\frac{x - x_i}{h}\right)\right)$$

$$= \iint y \frac{1}{h} K\left(\frac{x - z}{h}\right) f(y, z)\, dy dz$$

unde *f(y,z)* este funcţia dublă de densitate a lui *x* şi *y*.

Se efectuează următoarea schimbare de variabilă *s=(x-z)/h,* cu *ds=dz/h* şi funcţia poate fi rescrisă:

$E(\bullet) = \iint yK(s) f(x - hs, y)\, dy ds$

Densitatea dublă poate fi scrisă cu ajutorul densităţii condiţionate astfel:

f(x-hs,y)=f(y|x-hs)f(x-hs)

Cu această substituţie variabilele pot fi separate în felul următor:

E(•)=∫K(s)f(x-hs)(∫yf(y|x-hs)dy) ds=∫K(s)f(x-hs)m(x-hs)ds

Folosind descompunerea Taylor, integrala poate fi rescrisă:

E(•)=∫K(s)[f(x)-f'(x)hs+f''(x)(hs)²/2][m(x)-m'(x)hs+m''(x)(hs)²/2]ds=

=f m∫K(s)ds-h[f m'+f' m]∫sK(s)ds+h²/2[f'' m+f m''+2f' m']∫s²K(s)ds

=f(x)m(x)+ h²/2[f''(x)m(x)+f(x)m''(x)+f'(x)m'(x)]k₂

Valoare aşteptată a estimatorului va fi:

$$E\hat{m}(x) = \frac{f(x)\left[m(x) + h^2 k_2 (f''(x)m(x)/2f(x) + m''(x)/2 + f'(x)m'(x)/f(x))\right]}{f(x)[1 + h^2 k_2 f''(x)/2f(x)]}$$

[**] Scott pg. 222.

Se ştie că pentru *a* suficient de mic *1/(1+a)≅1-a*, folosind această aproximare expresia de mai sus poate fi simplificată:

$E\hat{m}(x)=m(x)+h^2k_2/2[m''(x)+2m'(x)f'(x)/f(x)]$

deci deplasarea poate fi scrisă:

deplasarea $\hat{m}(x)=h^2k_2/2[m''(x)+2m'(x)f'(x)/f(x)]$

În cazurile când punctele sunt uniform distribuite pe domeniul funcției, funcția de densitate este o constantă, derivata de ordinul unu este zero, şi termenul al doilea din paranteze este zero, iar formula deplasării poate fi simplificată. În realitate, în afara experimentelor din fizică, este foarte dificil de a avea control asupra distribuției punctelor *x,* mai ales când variabilele sunt obținute în urma unor anchete.

Forumula variației poate fi obținută în mod similar cu cea a deplasării:

$$var\ \hat{m}(x)=\frac{\sigma_\varepsilon^2\int K^2(u)du}{nhf(x)}$$

În acest moment se poate calcula MSE pentru estimatorul Nadaraya – Watson:

$$MSE\hat{m}(x)=\frac{\sigma_\varepsilon^2\int K^2(u)du}{nhf(x)}+h^4k_2{}^2[m''(x)+2m'(x)f'(x)/f(x)]^2$$

Pe lângă proprietățile statistice ne interesează şi proprietățilea asimtotice ale estimatorului. Estimatorul este deplasat, şi în formula deplasării influența creşterii numărului de observații nu este vizibilă. De aici ar părea că creşterea numărul de observații nu are nici o influență asupra scăderii deplasării. Acesta este motivul pentru care proprietatea de consistență a estimatorului este atât de importantă.

Condițiile[††] care asigură consistența estimatorului sunt enunțate în următoarea proproziție:

Dacă condițiile următoare sunt satisfăcute:

- $\int |K(u)|\,du<\infty$
- $lim_{|u|\to\infty}uK(u)=0$
- $EY^2<\infty$
- $n\to\infty,\ h_n\to\infty,\ nh_n\to\infty$

atunci în fiecare punct de continuitate a lui *m(x), f(x)* şi $\sigma^2(x)$ unde *f(x)>0*:

$$n^{-1}\sum_{i=1}^{n}W_{hi}(x)Y_i\xrightarrow{p}m(x)$$

[††] Vezi Hardle pag. 29.

Această propoziție stipulează că odată cu creşterea număului de observații, estimatorul Nadaraya – Watson converge în probabilitate către estimatorul real. Această proprietate este foarte importantă în special datorită deplasării estimatorului, recomandând folosirea acestuia atunci când există un număr suficient de observații. Demonstrația[‡‡] propoziței este complicată datorită exprimării estimatorului ca un raport a două variabile aleatoare corelate.

3.3 Alegerea gradului de netezire – metoda validării

Pentru a putea face o alegere cât mai obiectivă, este nevoie de un procedeu de selecție care să poată ordona după anumite criterii variantele posibile. Instrumentul folosit în general este minimizarea unei funcții obiectiv. În contextul stabilirii dimensiunii intervalului optim la estimarea densității, funcția de minimizat era integrala mediei pătratelor erorilor (MISE). Datorită dificultății obținerii unei expresii explicite pentru intervalul optim, una care să nu depindă de densitatea estimată, şi datorită faptului că expresia lui MISE este mult prea complicată este necesară găsirea altor criterii pentru selectarea intervalului optim.

Procedura alternativă pentru alegerea intervalului optim, care va fi descrisă în continuare, poartă numele de metoda validării (cross validation). Aceasta constă în minimizarea mediei *MSE*.

$$MSE(\lambda)=1/n\sum E\{\hat{m}(x)-m(x)\}^2$$

Deoarece funcția care se doreşte estimată nu este cunoscută, în locul lui *MSE* se foloseşte o funcție care o aproximează. Cel mai des folosit estimator a lui *MSE* este media pătratelor residuurilor *ASR* definit sub forma următoare:

$$ASR(\lambda)=1/n\sum\{y_i-\hat{m}(x_i)\}^2$$

În mod intuitiv, *ASR* nu este un estimator bun pentru *MSE,* pentru că minimizând *ASR* se obține un interval optim care duce la estimarea unei funcții care trece prin toate punctele observate, în care caz *ASR* este zero. Metoda de validare constă în calcularea valorii funcției în punctul x_i eliminând observația *i,* folosind restul de puncte. După care, pentru diferite intervale optime λ se calculează funcția *CV*. Se alege ca interval optim, valoarea λ care minimizează funcția *CV*.

Expresia pentru *CV* este următoarea:

$$CV(\lambda)=1/n\sum\{y_i-\hat{m}^{-i}(x_i)\}^2$$

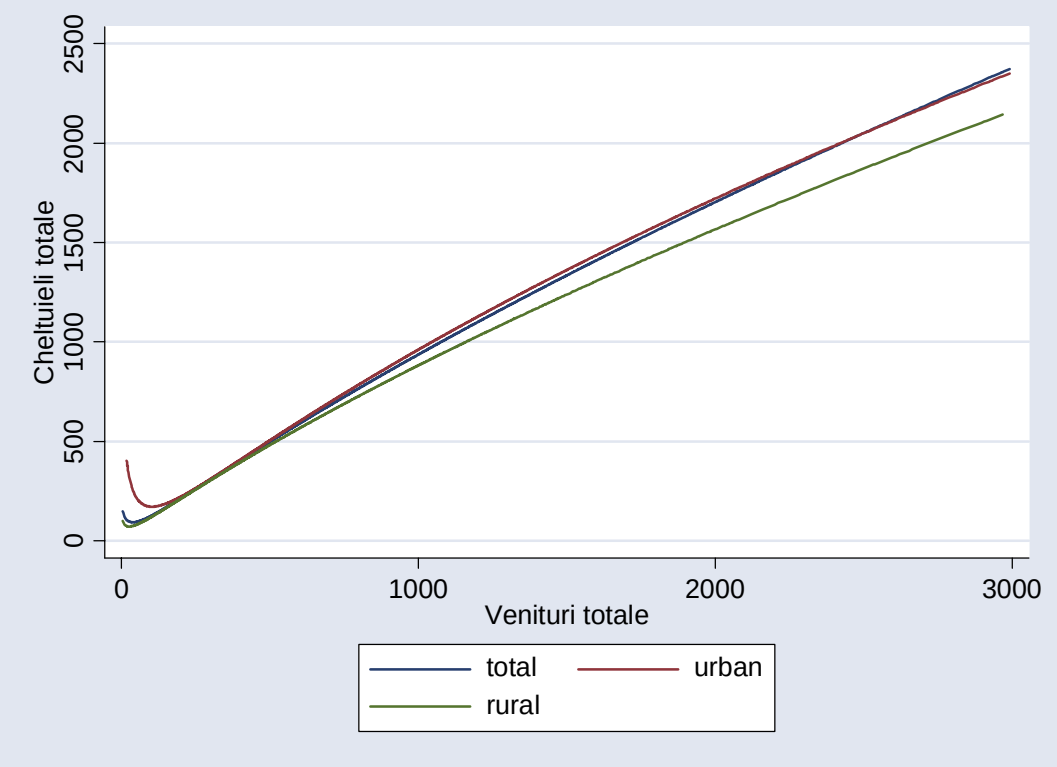
Există şi alte metode pentru alegerea dimensiunii intervalului optim (statistica C_p, etc.) dar din compararea celor două curbe folosind metodele diferite s-a constatat că rezultatele obținute sunt similare.

[‡‡] Pentru demonstrația propozitiei vezi Hardle pag. 39.

3.4 Dependen□a cheltuielilor de veniturile gospodăriilor

In această secțiune, estimatorii neparametrici sunt aplicați pentru a studia dependențele dintre variabile, ne interesează □n principal modul de variație al diverselor cheltuieli cu veniturile gospodăriilor.

Înainte de a începe prezentarea graficelor, trebuie men□ionat că metodele neparametrice func□ionează cel mai bine acolo unde există un număr suficient de mare de observa□ii. În zonele în care numărul de observa□ii este mai redus, a□a cum este la extremită□ile domeniului de referin□ă a func□iei (cozile distribu□iei de reparti□ie) valoarea estimatorului este mai instabilă, □i mai sensibilă la prezen□a valorilor extreme. Din acest motiv, atunci când analizăm graficele trebuie să fim con□tien□i de limitările estimatorului neparametric.

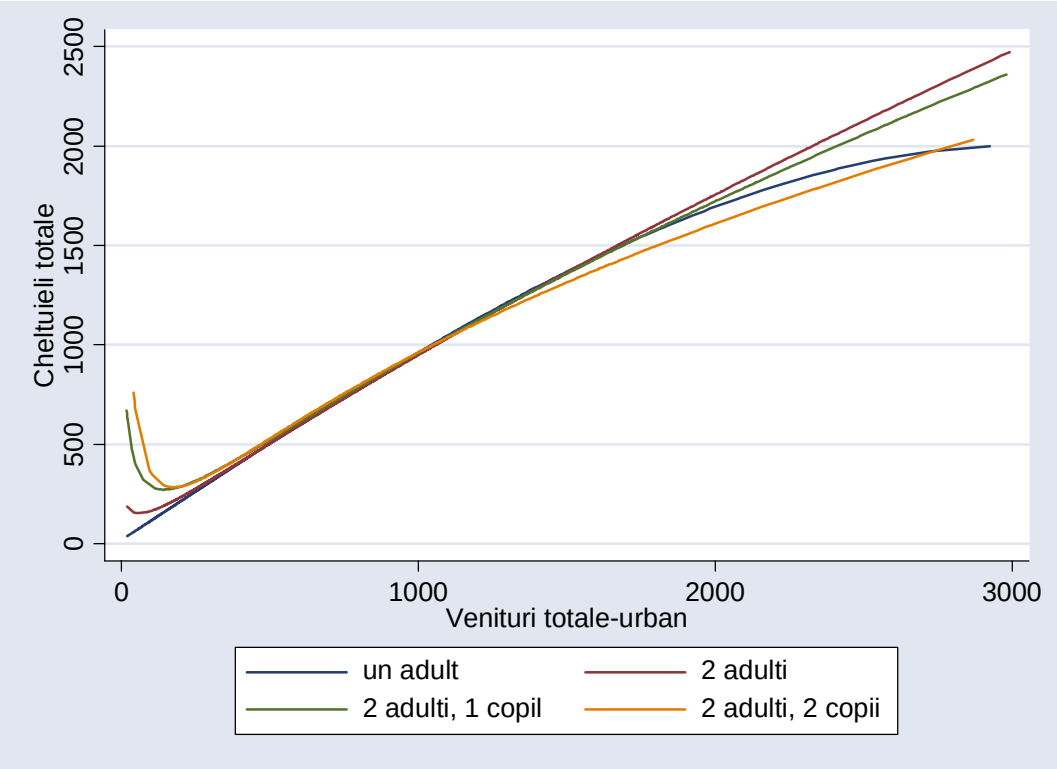


Graficul 3-1 Dependența cheltuieli totale- venituri totale pentru total populație, pentru populația urbană şi populația rurală

Prima dependență ce va fi analizată este relația dintre cheltuielile totale şi veniturile totale pentru toată populația, şi separat pentru populația urbană şi rurală. Ca o primă observație se poate remarca o dependență aproape liniară a cheltuielilor de venituri, cel puțin pentru venituri mai mici, cu propensitatea de consum mare, □n jurul unității. Aceasta □nseamnă că o proporție constantă din venituri este folosită pentru cheltuieli de către cea mai mare parte a gospodăriilor cu venituri mai mici. Insă odată cu creşterea veniturilor, cheltuielile dau semne că se aplatizează, suger□nd că odată cu creşterea veniturilor proporția care se

cheltuie scade, deci propensitatea de consum a gospodăriilor se mic□orează , un obicei de consum care este conform teoriilor economice.

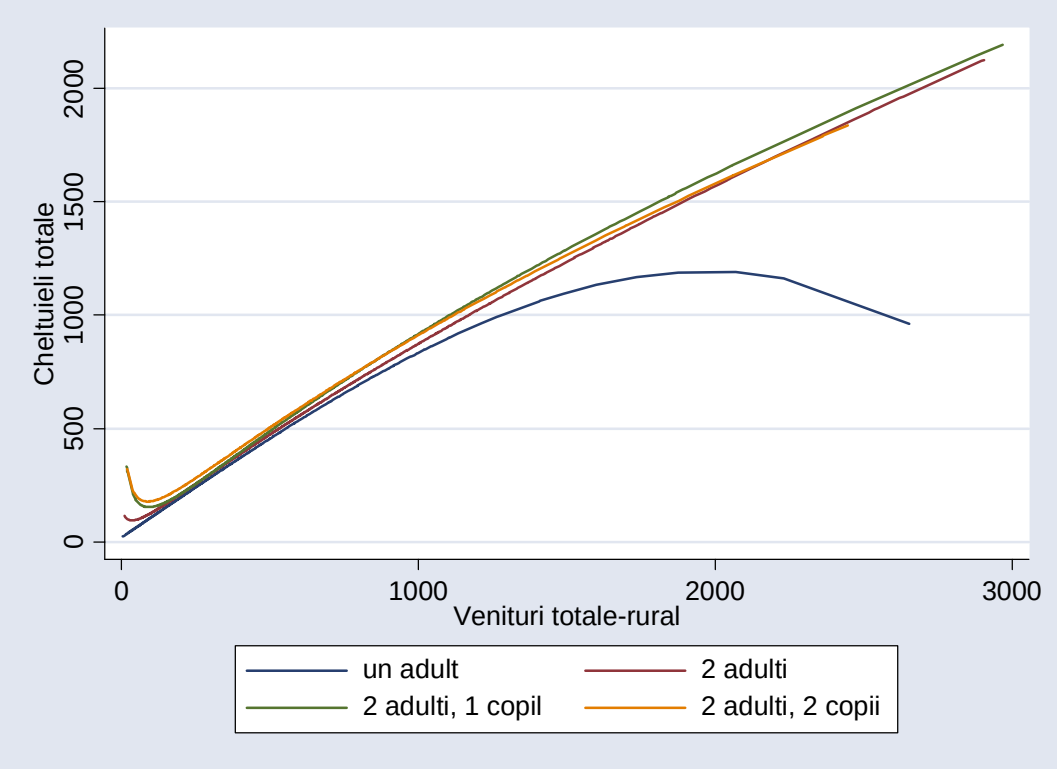
La analiza relației cheltuieli totale – venituri totale pentru cele două medii de rezidență se poate observa la venituri mici aproape o suprapunere a celor trei grafice. Odată cu cre□terea veniturilor, cheltuielile totale din zona rurală nu □in pasul cu cele din zona urbană, ecartul dintre cele două grafice mărindu-se cu cre□terea veniturilor. Explicațiile privind discrepanțele dintre consumul gospodăriilor urbane şi al celor rurale pot fi multiple, pornind de la un număr de membri mai redu□i, prezența auto-consumului care contribuie la reducerea cheltuielilor alimentare în zona rurală, dar □i o preponderen□ă mai mare în mediu rural a grupelor de v□rstă care au o propensitate de economisire mai mare.



Graficul 3-2 Rela□ia cheltuieli totale – venituri totale în func□ie de componen□a familiei în urban

Graficul 3-2 □i Graficul 3-3 prezintă rela□ia dintre cheltuieli totale □i venituri totale în func□ie de componen□a familiei. Mai ales în mediu urban se poate observa o suprapunere a celor patru grafice pentru venituri totale până la 1.500 RON. După acest prag, panta graficelor se modifică, propensitatea de consum scăzând masiv mai ales pentru familiile formate din doi adul□i □i doi copii. Gospodăriile formate din doi adul□i î□i men□in constantă proporția din venituri destinată consumului, care este totuşi subunitară. Gospodăriile formate din doi adul□i sunt fie tinere cupluri sau pensionari,

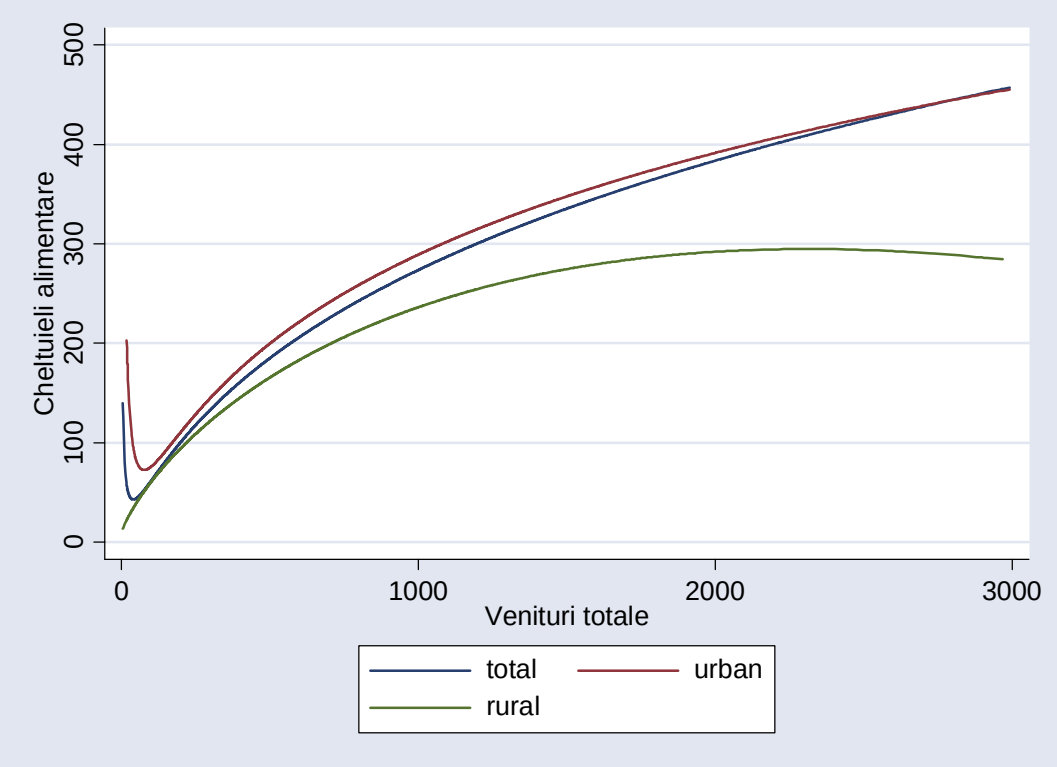
două categorii de populație care nu au motivația de economisire foarte puternică. Prezen□a copiilor pare să modifice comportamentul adul□ilor, inducând dorin□a de economisire. Acest lucru este evident însă numai atunci când veniturile sunt suficient de mari să permită atingerea unui standard minim de via□ă.



Graficul 3-3 Rela□ia cheltuieli totale – venituri totale în func□ie de componen□a familiei în rural

În zona rurală, lucrurile sunt un pic diferite. Se poate remarca o apropiere destul de mare a celor patru grafice în zona cu venituri mici □i medii, grupare care se men□ine □i la venituri mari (cu excep□ia gospodăriilor formate dintr-un adult[§§]) suger□nd că la venituri egale se cheltuie sume similare indiferent de componența familiei. Trebuie remarcat însă o propensitate de consum mai mică decât în mediu urban, adică un comportament similar cu gospodăriile urbane formate din adul□i □i copii pentru toate tipologiile de gospodării din mediu rural.

[§§]Dependen□ei în cazul gospodăriilor formate dintr-un adult la venituri mari nu este relevantă, pentru că în acea zonă graficul suferă de influen□e de la ni□te observa□ii extreme, ceea ce influen□ează dispropor□ionat forma rela□iei.

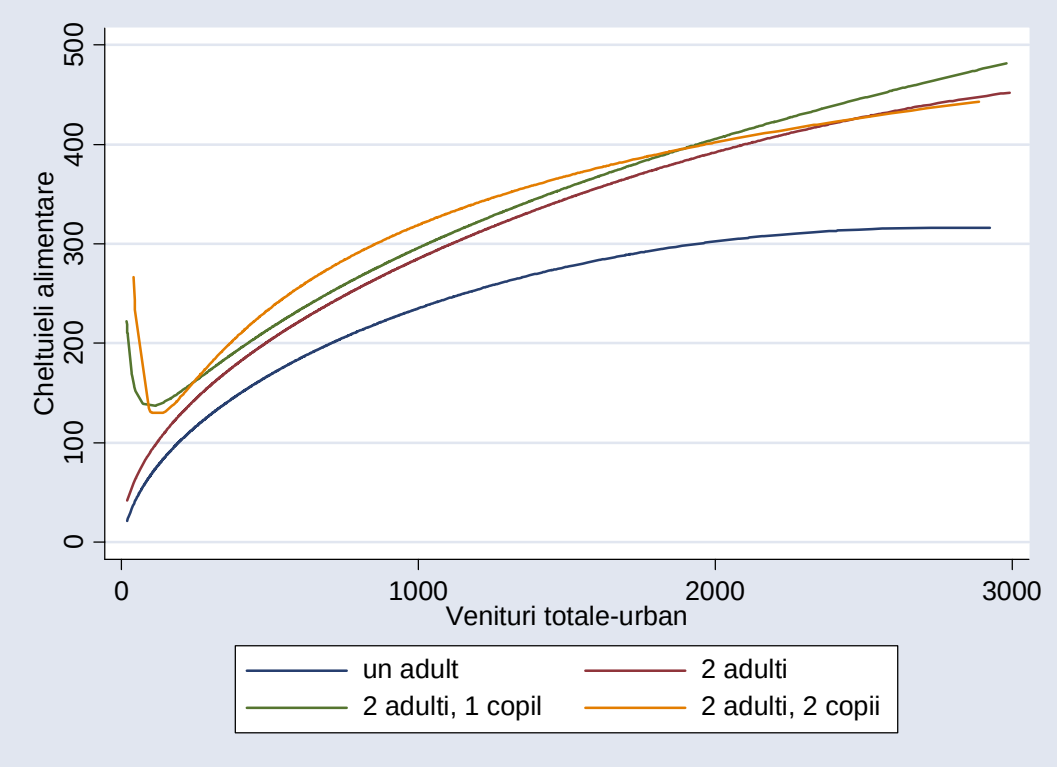


Graficul 3-4 Dependența cheltuieli alimentare – venituri totale pentru total populație, pentru populația urbană şi populația rurală

O primă observație este relația neliniară dintre cheltuielile alimentare şi veniturile totale. Deci ipoteza de liniaritate a dependențelor dintre variabile necesară □n cazul aplicării estimatorului metodei celei mai mici pătrate nu este îndeplinită. In acest caz este necesar să se liniarizeze dependențele dintre variabile prin aplicarea unor transformări – ca de exemplu logaritmarea – sau prin introducerea de variabile pătratice □n venituri.

În cazul cheltuielilor alimentare se pot observa deosebiri importante între comportamentele gospodăriilor din mediu urban fa□ă de cele din mediu rural. Deşi cheltuielile alimentare dau semne de saturație at□t □n mediu urban c□t şi □n mediu rural, cele din mediu rural se stabilizează □n jurul valorii de 300 RON, □n timp ce cele din mediu urban continuă să crească dar cu o pantă mult mai redusă. Prezența □n mediu rural a auto-consumului este o parte parțial responsabilă pentru saturația consumului de alimente □n mediu rural.

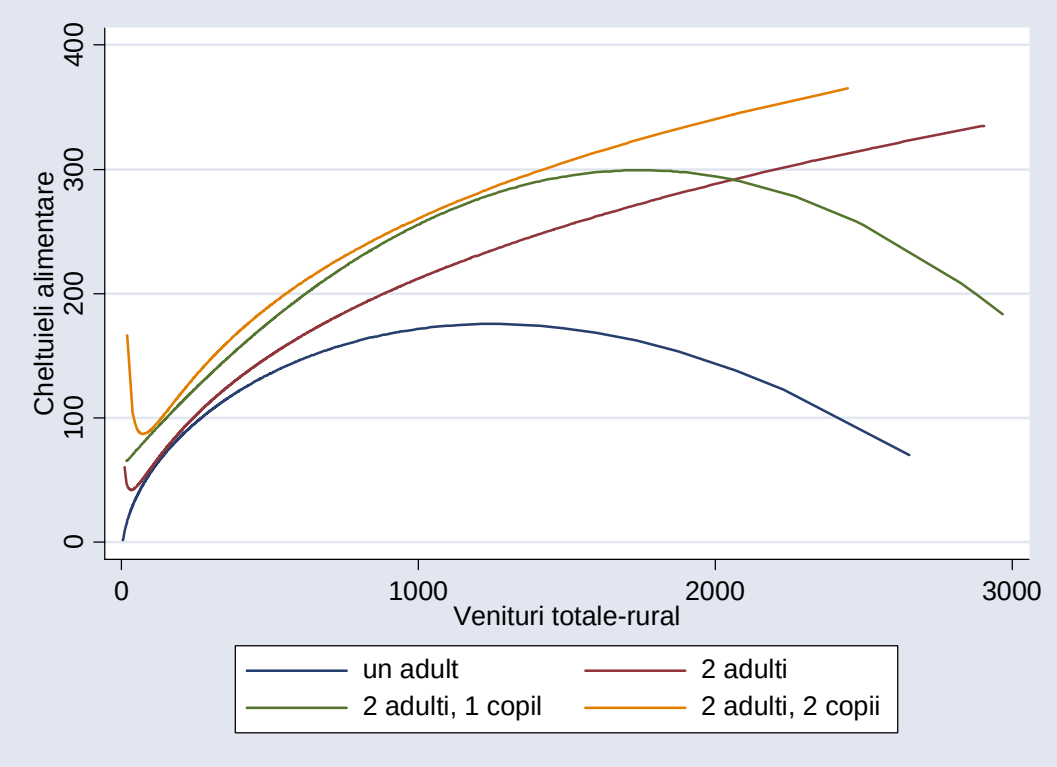
In Graficul 3-5 sunt prezentate cheltuielile alimentare In funcție de numărul de membri pentru zona urbană. Caracterisitica de neliniaritate a consumului alimentar se menține şi la gospodăriile urbane. Cum este de aşteptat, atingerea pragului de satura□ie se face la venituri mai mici □i la cheltuieli mai mici în cazul familiilor formate dintr-un singur adult. Proporția cea mai mare a cheltuielilor alimentare □n cazul veniturilor mici şi medii o au gospodăriile formate din doi adulți şi doi copii, □nsă acestea se saturează mai rapid dec□t cheltuielile celorlate tipuri de familii.



Graficul 3-5 Dependen□a cheltuieli alimentare – venituri totale în func□ie de componen□a familiei în mediu urban

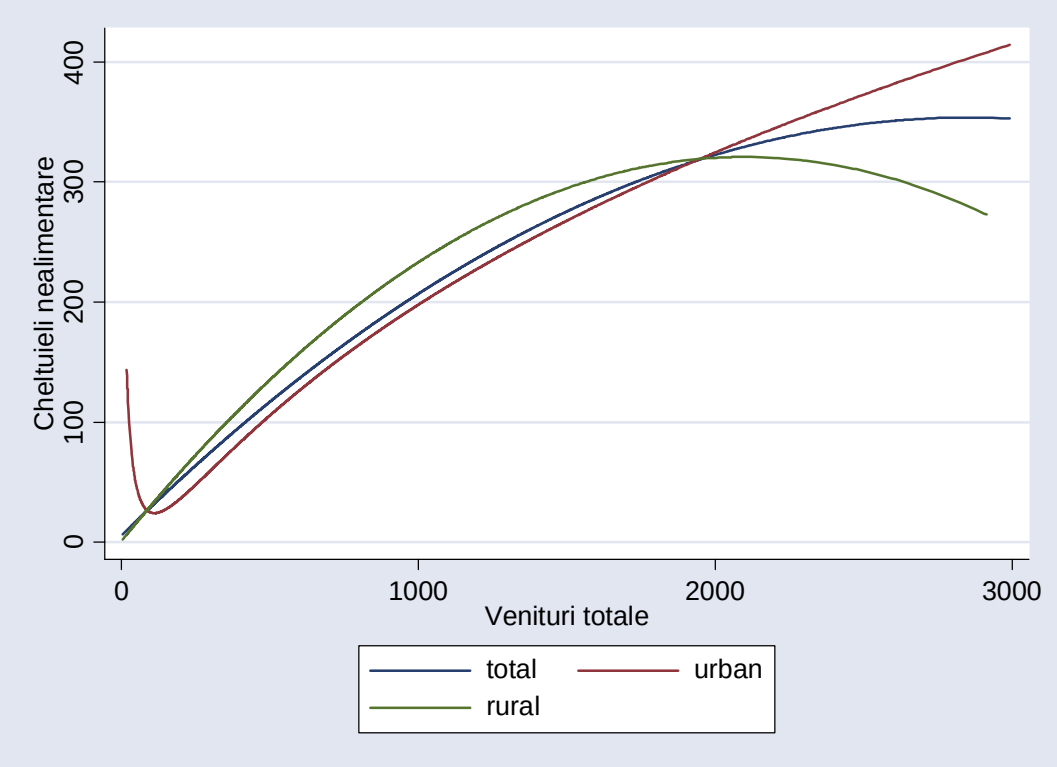
Dacă comparăm rela□ia dintre cheltuieli alimentare – venituri pentru gospodăriile din mediu urban fa□ă de cele din mediu rural se poate observa diferen□a de cheltuieli alimentare la venituri similare pentru acela□i tip de familie, lucru evident încă din Graficul 3-4.

Dacă în zona urbană exista o apropiere destul de mare a rela□iei dintre cheltuieli alimentare □i venituri la gospodăriile formate din mai mulţi adulţi, acest lucru nu mai este valabil în cazul familiilor din zona rurală. Cheltuielile cele mai mari alimentare le au, a□a cum este de a□teptat familiile formate din doi adul□i □i doi copii, iar pentru veniturile mici □i medii, familiile formate din doi adul□i □i un copil au cheltuieli foarte apropiate, aprope identice. Cheltuielile adulţilor care trăiesc singuri sunt foarte scăzute, probabil că o mare proporţie din aceştia participă □n agricultura de subzistenţă. Forma parabolică a rela□iei pentru un adult □i poate şi cea pentru doi adul□i cu un copil poate fi dată de absen□a suficientelor observa□ii în zona de venituri mari.



Graficul 3-6 Dependen□a cheltuieli alimentare – venituri totale în func□ie de componen□a familiei în mediu rural

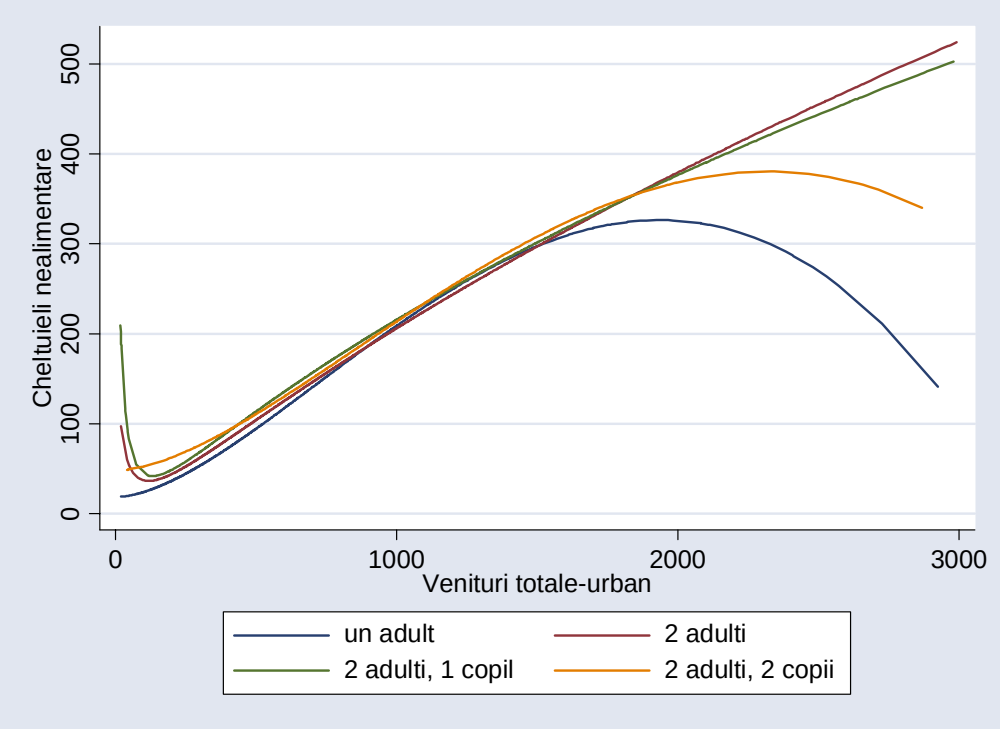
Graficul 3-7 descrie dependen□a cheltuielilor nealimentare de venituri. O primă observație este forma aproape liniară a dependenței la venituri mici şi medii. De asemenea, **nu se mai observă diferențe foarte importante □ntre consumul gospodăriilor urbane față de cele rurale**. La venituri mici cheltuielile nealimentare sunt chiar mai mari pentru gospodăriile rurale, dar la venituri peste 2000 RON situația se inversează.



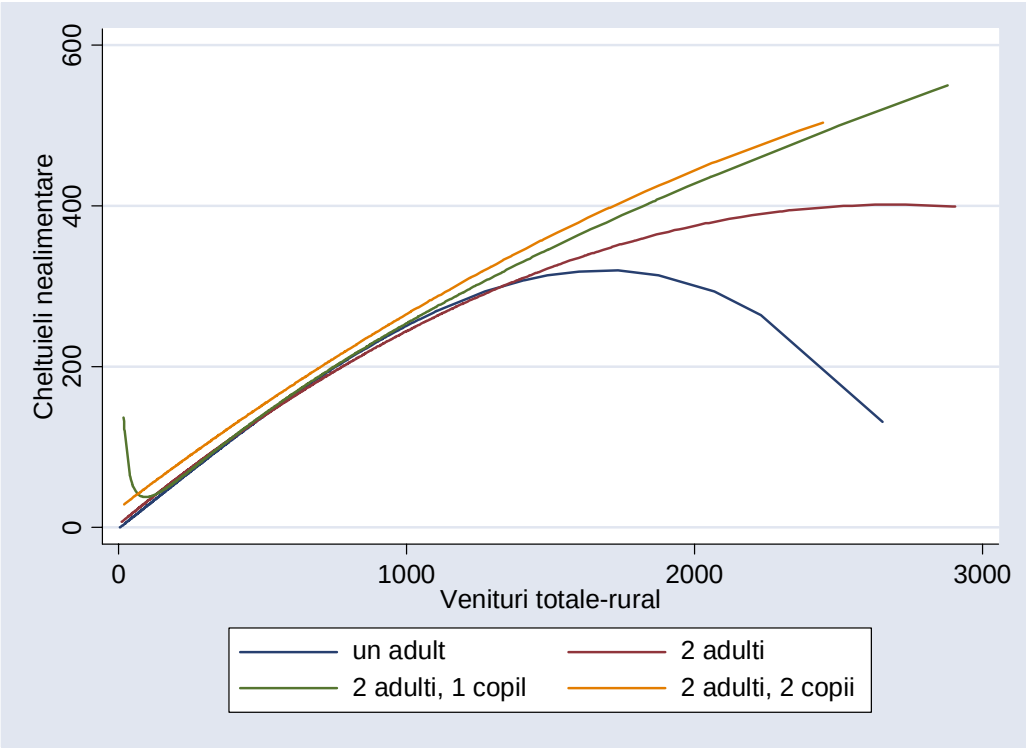
Graficul 3-7 Rela□ia cheltuieli nealimentare – venituri totale pentru total popula□ie, popula□ia urbană □i popula□ia rurală

Dependen□a **cheltuieli nealimentare – venituri pentru diverse tipuri de familii din zona urbană este similară pentru venituri mici şi medii, independent de componența familiei**, graficele practic se suprapun unul peste altul. În zona de varia□ie comună se poate observa **o dependen□ă liniară a cheltuielilor de venituri**, după care cheltuielile pentru o familie formată dintr-un adult □i cele ale familiei formate din doi adul□i □i doi copii dau semne că suferă de efectele date de influen□a valorilor extreme în absen□a datelor suficiente.

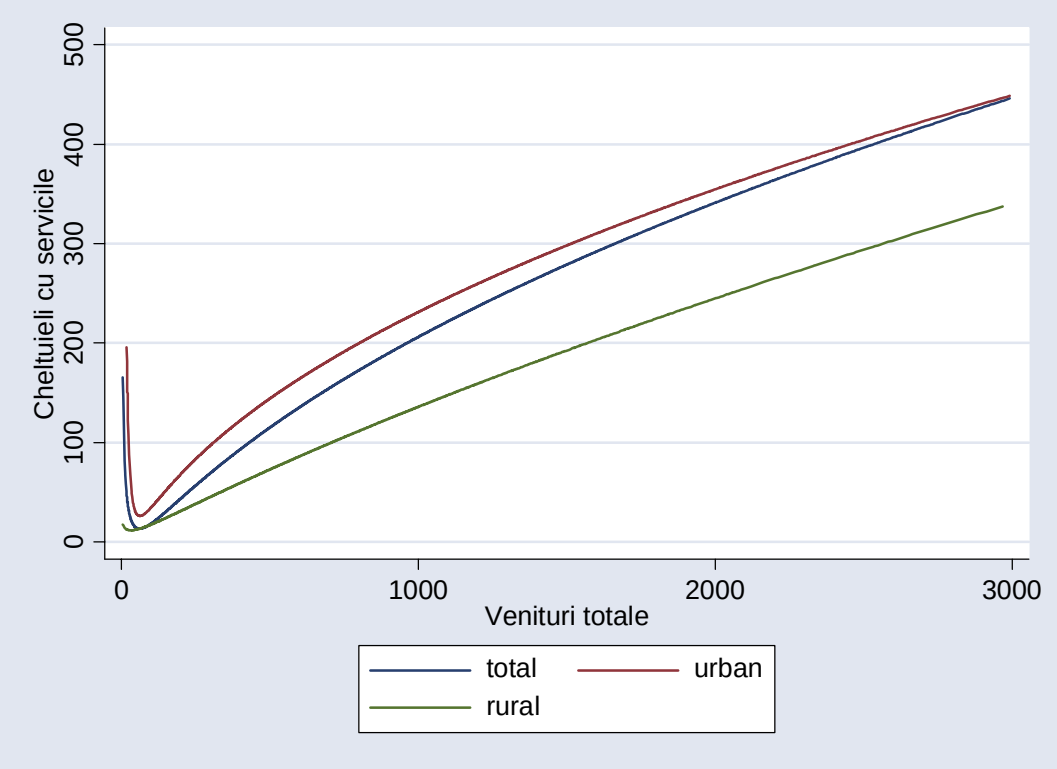
Aceleaşi caracteristici ale dependențelor cheltuielilor nealimentare – venituri sunt valabile şi pentru diversele tipuri de familii din zona rurală. Graficele sunt aproape suprapuse unul peste celălalt, la venituri mici □i medii, dependen□a fiind aproape linară pe sec□iunea respectivă, după care cheltuielile familiilor formate din doi adul□i dau semne de satura□ie în timp ce pentru doi adul□i cu unu sau doi copii ele continuă să crească în continuare.



Graficul 3-8 Dependen□a cheltuieli nealimentare – venituri totale în func□ie de componen□a familiei în mediul urban



Graficul 3-9 Dependen□a cheltuieli nealimentare – venituri totale în func□ie de componen□a familiei în mediul rural.

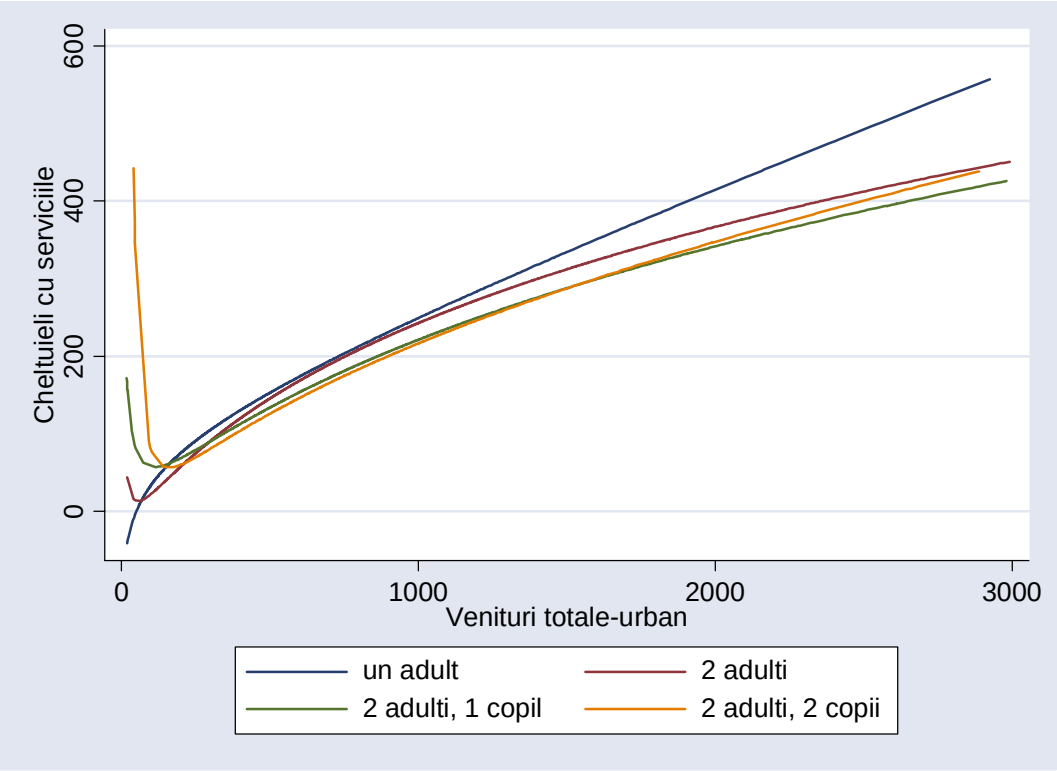


Graficul 3-10 Rela□ia dintre cheltuielile cu serviciile – veniturile totale pentru pentru total popula□ie, popula□ia urbană □i popula□ia rurală

O ultimă dependen□a pe care o vom analiza este rela□ia dintre cheltuielile cu serviciile □i venituri totale. Este interesant de observat **forma dependen□ei**, care în cazul **consumatorilor rurali este linară, iar în cazul consumatorilor urbani are o formă logaritmică**.Cheltuielile cu serviciile sunt mai mari în mediu urban, în mare parte datorită existen□ei unui pachet mai divers de servicii in mediu urban. Nu discutam aici numai de utilită□i, despre care am mai discutat, dar □i de servicii către popula□ie pentru petrecerea timpului liber □i nu numai.

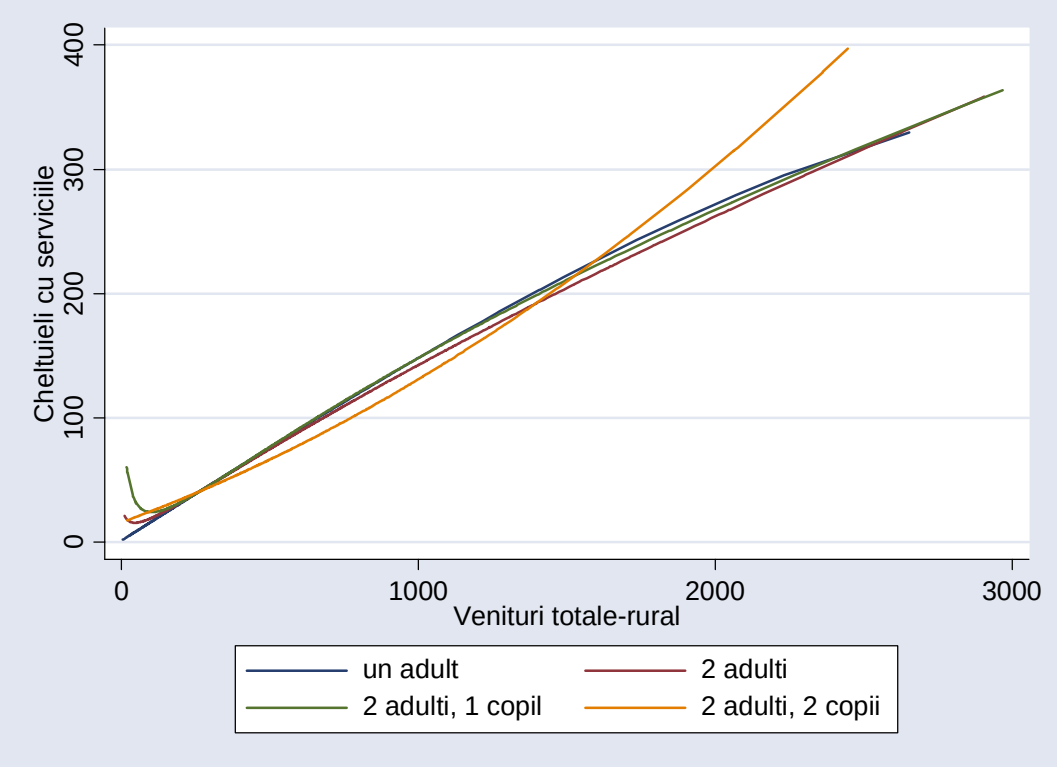
Graficul 3-11 prezintă rela□ia dintre cheltuielile cu serviciile fa□ă de venituri pentru tipuri de gospodării, □i se poate observa o **varia□ie foarte apropiată a cheltuielilor cu serviciile pentru toate gospodăriile care con□in doi adul□i**. S-ar părea că aceste cheltuieli nu depind de numărul de copii, □i reprezintă mai mult ni□te cheltuielie ale adul□ilor. Binein□eles că în servicii intră □i utilită□ile, dar există o corela□ie puternică între venituri □i consumul de utilită□i, □i mai pu□in puternică între consumul de utilită□i □i numărul de persoane din gospodărie. Spre expemplu, considera□i cheltuielile cu încălzirea, acestea depind de volumul de încălzit, de tipul de locuin□ă;de tipul de încălzire folosit, temperatura exterioară, etc. □i mai pu□in de numărul de persoane care locuiesc în camera respectivă. Veniturile, pe de altă parte, pot fi decisive în

stabilirea tipului de încălzire folosit[***], i a numărului de camere încălzite. În cazul unor venituri mici, o familie poate decide să încălzească o cameră în timp ce la venituri mari aceai familie poate decide să încălzească mai multe camere.



Graficul 3-11 Dependena cheltuieli cu serviciile – venituri totale în funcie de componena familiei în mediul urban

[***] În cazul unor venituri mai mari, o gospodărie poate decide să-i instaleze o centrală de apartament în loc să se încălzească cu sobe



Graficul 3-12 Dependen□a cheltuielicu serviciile – venituri totale în func□ie de componen□a familiei în mediul rural

Interesant este însă dependen□a cheltuieli cu serviciile – veniturile totale pentru gospodăriile urbane formate dintr-un singur adult. Acestea depă□esc, la venituri similare, cheltuielile cu serviciile pentru gospodăriile formate din mai multe persoane. Cum am observat deja, acest tip de cheltuieli par să fie speific adul□ilor □i adul□ii singuri cheltuie mai mult la venituri similare, probabil datorită cheltuielilor mai mari pe care aceste persoane le au □n petrecerea timpului liber. Adul□ii singuri au cheltuieli mai reduse at□t la bunurile alimentare c□t şi la cele nealimentare, ce le rămâne din venituri cheltuiesc pentru servicii.

În Graficul 3-12 se prezintă dependen□a cheltuieli cu serviciile – venituri totale pentru gospodăriile din mediu rural. În acest caz, cheltuielile cu serviciile sunt destul de apropiate independent de tipologia familiei, ele depinzând numai de venituri. Fa□ă de dependen□ele urbane cheltuielile adul□ilor singuri nu mai ies în eviden□ă, în sensul ca □in pasul cu restul cheltuielilor la venituri similare, însă dacă le raportăm la numărul de adul□i sunt practic duble în compara□ie cu restul familiilor. În schimb, familiile cu doi adul□i □i doi copii la venituri mari au cheltuieli cu serviciile mult mai mari decât restul.

4 Concluzii

In această lucrare au fost folosite metodele neparametrice pentru a studia caracteristicile consumului gospodăriilor. Au fost construite funcții de densitate pentru venituri, dar şi estimatori ai relațiilor dintre cheltuieli şi venituri.

Din punctul de vedere al veniturilor, s-a constatat că veniturile urbane sunt sensibil superioare veniturilor rurale. Iar acest lucru nu a fost dat de tipologia diferită a familiilor □n cele două medii, pentru că diferențele □ntre venituri s-au menținut şi atunci c□nd s-a studiat veniturile medii □n funcție de tipul familiei, şi nici de nivele de salarizare sensibil diferite □ntre cele două medii, pentru că deşi distribuțiile salariilor nu sunt identice, nici nu sunt at□t de diferite □nc□t să fie responsabile de diferențele □nregistrate la venituri.

Cele mai mari discrepanțe s-au □nregistrat □n pensii, distribuția pensiilor rurale este tri-modală cu cel mai mare procent de populație la pensiile mici – corespunzătoare pensiilor agricultorilor. In zona urbană pensiile agricole lipsesc, distribuția av□nd doar două v□rfuri, iar cel mai mare procent de pensionari sunt la nivelul mare de pensie. Astfel discrepanțele veniturilor pot fi explicate măcar parțial de diferențele privind statutul ocupațional al persoanelor din urban şi rural, □n zona rurală se află mai puțini salariați şi mai mulți lucrători/pensionari din agricultură.

La analiza dependențelor diverselor cheltuieli față de venituri rezultatele au fost mai variate. Cheltuielile alimentare sunt foarte diferite □ntre rural şi urban, iar la venituri similare sunt mai mari □n urban, □n timp ce cheltuielile nealimentare sunt mai mari □n rural pentru venituri medii şi mici. Cheltuielile cu serviciile □n zonele rurale sunt mult mai scăzute dec□t □n zonele urbane, iar cel puțin parțial responsabil pentru asta este lipsa se ofertă □n rural.

Cheltuielile cu alimentele sunt tipul de cheltuială cel mai sensibil la tipologia familiei, □n timp ce cheltuielile cu serviciile sunt aproape identice la venituri similare indiferent de caractristicile gospodăriei, chiar şi □n cazul familiilor formate dintr-un singur adult □n mediu urban, ceea ce practic □nseamnă că un adult singur cheltuie de două ori mai mult cu serviciile dec□t un adult căsătorit.

Din punctul de vedere al relației dintre variabile s-a observat că există cazuri c□nd dependența veniturilor de cheltuieli nu este liniară, mai ales □n cazul cheltuielilor alimentare. Acest lucru nu este surprinzător, dar poate ridica probleme privind folosirea estimatorului celor mai mici pătrate. In cazul □n care neliniaritatea dependențelor nu poate fi corectată prin introducerea unor variabile suplimentare, atunci probabil că ar trebui să se liniarizeze relația prin logaritmare, sau poate este necesară folosirea unor metode neparametrice.

Bibliografie

1. Ahmad, I.A., Lin, P. E. (1984): "*Fitting a multiple regression*", Journal of Statistics Planning and Inference, , vol. 2, pp. 163-176.

2. Benedetti, J. K. (1977): "*On the Nonparametric Estimates of Regression Function*" *Journal of the Royal Statistical Society,* Series B, vol. B, vol 39, pp. 248-253.

3. Bierns, H. J., Pott-Bitter, H. A. (1990): "*Specification of Household Engel Curves by Nonparametric Regression*" Econometric Reviews, vol.9, pp.123-184.

4. Chu, C. K., Marron, J. S. (1991): "*Choosing a Kernel Regression Estimator*" Statistical Science, vol. 6, pp. 404-436.

5. Devroye, L. P., Gyorfi, L. (1985): *Nonparametric Density Estimation,* New York, Wiley.

6. Dodge, Y. (1986): "*Some Difficulties Involving Nonparametric Estimation of a Density Function*" Journal of Official Statistics, vol. 2, pp. 193-202

7. Epanechnikov (1969) *Nonparametric estimation of a multidimensional probability density function* Theory Probability Appl., vol. 14, pp. 153-158.

8. Greene, W.H. (1993): *Econometric Analysis,* New Jersey: Prentice Hall International Editions.

9. Greblick, W., Krzyak (1980): "*Asymptotic Properties of Kernel Estimates of a Regression Function*" Journal of Statistical Planning and Inference, vol. 4, pp. 81-90.

10. Hardle, W. (1990): *Applied Nonparametric Regression,* New Zork, Cambridge University Press.

11. Hardle, W. (1991): *Smoothing techniques with implementation in S, ser. Springer Series in Statistics. New York: Springer-Verlag*

12. Hardle, W., Oliver, L. (1994): *Applied Nonparametric Methods,* in Handbook of Econometrics vol. 4.

13. Iordan, M., Regep, M., Chilian, M. N. (2001): *"Household Consumption in the Central and East - European Countries Aspiring to Joint the EU"* Romanian Journal of Economic, vol. 1-2.

14. Maddala, G. S. (1992): *Introduction to econometrics* Maxwell MacMillan International Editions.

15. Pagan, A., Ullah, A. (1999):*Nonparametric Econometrics,* Cambridge, Cambridge University Press.

16. Pindyck, R. S., Rubinfeld, D.L. (1991): *Econometric Models and Economic Forecasts,* New York: McGraw – Hill International Edition.

17. Scott, D. W. (1992): *Multivariate Density Estimation: theory practice and visualization, Wiley.*

18. Silverman, B.W. (1986): *Denisty Estimation for Statistics and Data Analysis* London: Chapmand and Hall.

19. Stanciu, Mariana. 2006. Metode de cercetare a modelelor de consum. Bucureşti: CIDE

20. Stanciu, Mariana. 2001. Structuri moderne ale consumului european. Bucureşti: Editura Genicod