УДК 311.1:519.85

Марець О.Р.
кандидат економічних наук, доцент
Львівського національного університету
імені Івана Франка

Дуда О.П.
студентка
Львівського національного університету
імені Івана Франка

ПІДБІР ТА ДІАГНОСТИКА БАГАТОФАКТОРНИХ РЕГРЕСІЙНИХ МОДЕЛЕЙ У ПАКЕТІ АНАЛІЗУ ДАНИХ R

Представлено процеси підбору та діагностики регресійної моделі на основі даних макроекономічної статистики. Використано засоби пакету статистичного аналізу R. Основну увагу приділено графічному методу діагностики регресійної моделі.

Ключові слова: регресійний метод, нормальність, незалежність залишків, лінійність, гомоскедастичність, мультиколінеарність, фактор інфляції дисперсії, Q-Q-діаграма, діаграма компонентів і залишків, діаграма впливу, діаграма Кука.

Марець О.Р., Дуда О.П. ПОДБОР И ДИАГНОСТИКА МНОГОФАКТОРНОЙ РЕГРЕССИОННОЙ МОДЕЛИ В ПАКЕТЕ АНАЛИЗА ДАННЫХ R

Представлены процессы подбора и диагностики регрессионной модели на основе данных макроэкономической статистики. Использованы средства пакета статистического анализа R. Основное внимание уделено графическому методу диагностики регрессионной модели.

Ключевые слова: регрессионный метод, нормальность, независимость остатков, линейность, гомоскедастичность, мультиколинеарность, фактор инфляции дисперсии, Q-Q-диаграмма, диаграмма компонентов и остатков, диаграмма влияния, диаграмма Кука.

Marets O.R., Duda O.P. FITTING AND CHECKING THE ASSUMPTIONS OF MULTIVARIATE REGRESSION MODELS IN R

The process of variable selection and checking the assumptions of multivariate regression model based on macroeconomic statistics is presented. Statistical analysis package R is used. The central point is a graphical method of regression model diagnostics.

Keywords: regression method, normality, independence of residuals, linearity, homoscedasticity, multicollinearity, inflation variance factor, scatterplot matrix, Q-Q-plot, component+residual plots, hat plot, Cook's distance, influence plot, R, car package, gvlma package.

Постановка проблеми. Регресійний аналіз – це загальна назва низки методів виявлення незалежних змінних, що впливають на залежну змінну. Крім того, ці методи описують тип взаємозв'язку і дають змогу скласти рівняння, що може передбачити залежну змінну на основі значень незалежних.

Аналіз останніх досліджень та публікацій. Регресійний аналіз використовували чимало вчених у дослідженнях сучасних тенденцій функціонування домогосподарств в умовах ринкової економіки. Зокрема:

– М. Жук та В. Здрок вивчали вплив індексу споживчих цін, доходу на особу та внутрішніх кредитів приватному сектору на споживчі видатки домогосподарств [1];

– З.В. Приймак з'ясовувала параметри залежності доходів населення на одну особу від ВВП на одну особу і частки середніх інвестицій у ВВП [2];

– О. Піскунова та О. Осипова будували двофакторні моделі регресії, в яких розглянуто одночасний вплив на обсяги споживання базових продуктів харчування обсягів виробництва відповідного продукту харчування та наявного доходу населення [3];

– В. Бесчастна вивчала кореляційний зв'язок між факторами соціального впливу на продуктивність праці у сільському господарстві [4];

– Ю.В. Лесик моделювала взаємозв'язок між доходами населення та іншими одержаними трансфертами і доходами від власності [5];

– О. Кузик моделював споживчу поведінку домогосподарств в Україні за період 1992–2008 рр. і дійшов висновку, що споживча поведінка вітчизняних домогосподарств добре описується кейнсіанською функцією споживання [6];

– В.О. Гнєушева вивчала чинники впливу на фінансові ресурси домогосподарств у ринкових умовах [7].

У вітчизняних наукових публікаціях спостерігаємо багато праць, які насичені теоретичними викладками та результатами розрахунків. Водночас сучасні засоби статистичного аналізу дають змогу дуже швидко виконувати необхідні розрахунки та створювати візуалізації. Завдання дослідника нині полягає у правильній інтерпретації результатів. Тож відчуваємо брак наукових робіт, які би робили акцент на процесі підбирання та діагностики регресійних моделей за допомогою сучасних засобів статистичного аналізу.

Мета статті – згенерувати багатофакторну регресійну модель, де залежна змінна – макроекономічний показник – витрати населення у розрахунку на одну особу, та здійснити діагностику цієї моделі за допомогою засобів пакету аналізу R.

Виклад основного матеріалу. Ми використали дані зі статистичного збірника «Регіони України» (2016) [8, с. 13–16], де взяли інформацію про основні соціально-економічні показники у 2015 році за областями України. Насамперед ми видалили з аналізу дані м. Київ як найяскравіше нетипове значення. Крім того, ми забезпечили порівнянність даних шляхом ділення показників на чисельність населення (кодове позначення показника закінчується на no) або на чисельність зайнятих (nz). Ми навмисне не проводили детального логічного аналізу змінних, щоби простежити, який результат видадуть засоби R.

Зауважимо, що мова програмування R складається з базового набору функцій, які вбудовані безпосередньо у програму. Крім того, доступні спеціальні пакети, які містять функції для розв'язання вузько-

спеціалізованих завдань. Ми, зокрема, скористаємося пакетами car (Companion to Applied Regression) та gvlma (Global Validation of Linear Model Assumptions).

Два найпоширеніші способи обирати незалежні змінні, які найкраще змоделюють значення залежної, – це покроковий метод і регресія за підмножинами. Ми скористаємось першим із них.

Ми почали побудову регресійної моделі з використання функцій `lm()` та `step()`, зменшили кількість чинників із 13 до 4 та отримали рівняння регресії.

Функція `step()` порівнює моделі з різними чинниками та на основі кри-терію AIC. Зауважимо, що цей критерій не бере до уваги мультиколінеарність.

На рис. 1 показано параметри моделі та показники її якості.

У першому рядку на рис. 1 – команда, яка рахує параметри моделі. Її першим аргументом є залежна змінна, далі через значок «тільда» – перелік незалежних змінних, розділених знаком «плюс».

Результуюче рівняння має такий вигляд:

$$Expenses = 6602 + 0{,}69 \times Income + 1571 \times Turnoverno - 5678 \times Impno - 31390 \times Lossnz,$$

де *Expenses* – витрати населення у розрахунку на одну особу у грн; *Income* – наявний дохід населення у розрахунку на одну особу у грн, *Turnoverno* – роздрібний товарооборот підприємств (у фактичних цінах) на одну особу у тис. грн; *Impno* – імпорт товарів та послуг на одну особу у тис. дол. США; *Lossnz* – фінансовий результат до оподаткування (збиток) на одного зайнятого у млн грн.

Відповідно до рівняння збільшення доходів домогосподарств на 1 грн збільшує витрати на 0,69 грн; збільшення роздрібного товарообороту на 1 тис. грн збільшує витрати на 1571 тис. грн; збільшення імпорту на 1 тис. грн зменшує витрати на 5678 тис. грн; збільшення фінансового результату збиткових підприємств зменшує витрати на 31 390 млн грн.

Вільний член та параметр *Lossnz* істотні зі ймовірністю 95%, параметр *Impno* – істотний зі ймовірністю 99%, параметри *Income* та *Turnoverno* істотні зі ймовірністю 99,9%. Варіація ознак, що входять у модель, пояснює 95,46% варіації залежної ознаки.

Отже, ми маємо рівняння регресії з досить хорошими показниками якості параметрів. Проте це лише початок регресійного аналізу. Необхідно ще перевірити умови застосування цих параметрів, тобто здійснити діагностику цього рівняння регресії.

Діагностика регресійної моделі включає в себе:

1) перевірку умов застосування регресійного методу (нормальність, незалежність залишків, лінійність та гомоскедастичність);

2) перевірку мультиколінеарності;

3) виявлення нетипових значень, точок високої напруги та впливових одиниць сукупності.

Важливий перший крок у діагностиці множинної регресії – це вивчення парних взаємозв'язків між змінними. Для цього визначимо коефіцієнти кореляції (рис. 2) та застосуємо графічний метод (рис. 3).

Коефіцієнти кореляції ми досліджуємо для того, щоб оцінити мультиколінеарність – наявність щільного зв'язку між незалежними змінними в моделі. Мультиколінеарність – небажане явище для регресії, оскільки вона, крім всього іншого, може давати змінним біля параметрів моделі неінтуітивний знак. Вважається, що значення коефіцієнта кореляції понад 0,7 – це привід задуматися про мультиколінеарність. На рис. 2 бачимо три таких значення, що дає підстави переглянути чинники, включені в модель.

На рис. 3 представлена матриця діаграм розсіювання значень залежної і незалежної змінних нашої моделі за допомогою функції `scatterplotMatrix()`.

Ця функція створює діаграми розсіювання для всіх пар змінних зі згладженою кривою та регресій-

```
lm(formula = Expenses ~ Income + Turnoverno + Impno + Lossnz,
    data = df)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.602e+03  2.959e+03   2.231  0.03790 *
Income       6.867e-01  1.305e-01   5.261 4.45e-05 ***
Turnoverno   1.571e+03  2.297e+02   6.837 1.59e-06 ***
Impno       -5.678e+03  1.918e+03  -2.960  0.00804 **
Lossnz      -3.139e+04  1.365e+04  -2.301  0.03292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1918 on 19 degrees of freedom
Multiple R-squared:  0.9546,    Adjusted R-squared:  0.945
F-statistic: 99.86 on 4 and 19 DF,  p-value: 1.761e-12
```

Рис. 1. Параметри регресійного рівняння залежності витрат населення від характеристик областей (виконано в R)

```
cor(df[,c("Expenses", "Income", "Turnoverno", "Impno", "Lossnz")])
             Expenses     Income Turnoverno     Impno     Lossnz
Expenses    1.0000000  0.8777226  0.8938121 0.4679988 -0.5722758
Income      0.8777226  1.0000000  0.7731074 0.5373882 -0.3877349
Turnoverno  0.8938121  0.7731074  1.0000000 0.7175109 -0.3572937
Impno       0.4679988  0.5373882  0.7175109 1.0000000  0.0795329
Lossnz     -0.5722758 -0.3877349 -0.3572937 0.0795329  1.0000000
```

Рис. 2. Коефіцієнти кореляції між змінними регресійної моделі залежності витрат населення від характеристик областей (виконано в R)

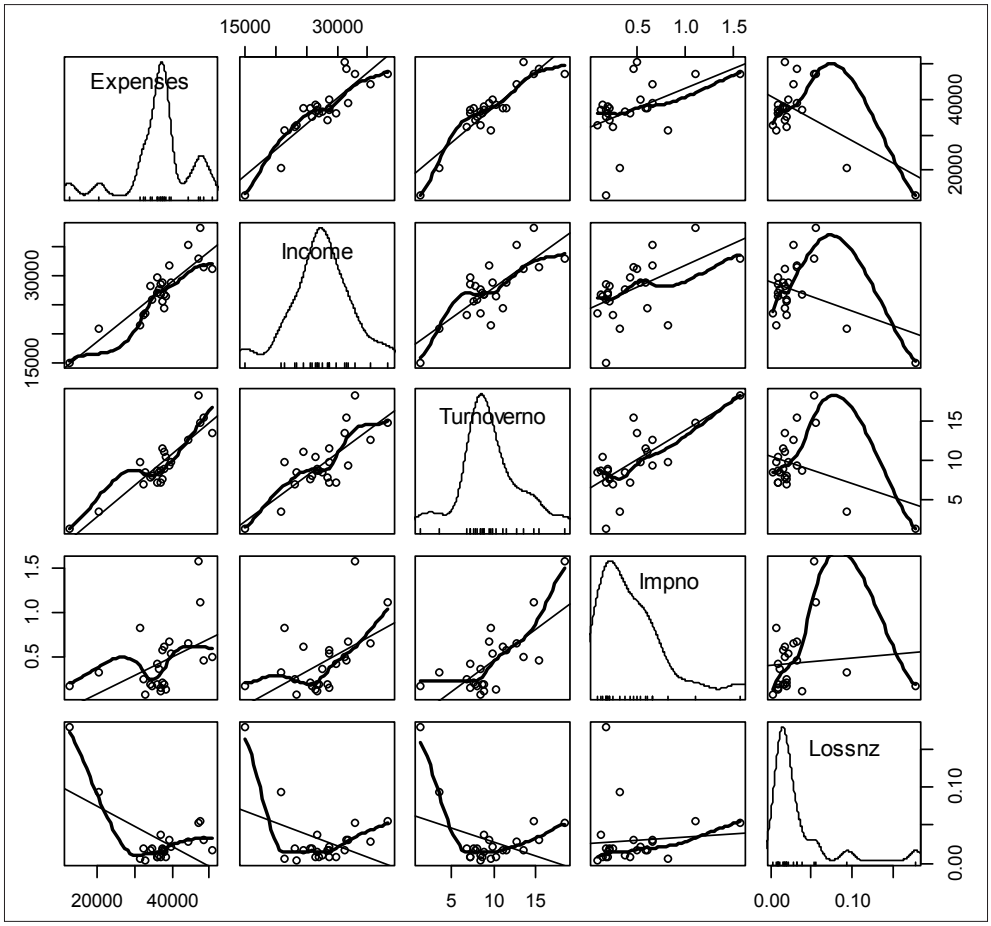

Рис. 3. Діаграма змінних регресійної моделі залежності витрат населення від характеристик областей (виконано в R)

```
vif(gvmodel)
    Income Turnoverno       Impno     Lossnz
  2.589279   4.349227   2.823652   1.625638

sqrt(vif(gvmodel)) < 2 #if false there is multicolinearity
    Income Turnoverno       Impno     Lossnz
      TRUE      FALSE       TRUE       TRUE
```

Рис. 4. Результат перевірки моделі на мультиколінеарність в R за допомогою функції `vif()`

ною прямою. На головній діагоналі представлені полігони частот та графіки-щітки для кожної змінної.

Ми бачимо, що розподіл витрат, доходів та роздрібного товарообороту нагадує нормальний, а збитки можна охарактеризувати асиметрією. Між витратами та доходами, роздрібним товарооборотом та імпортом бачимо прямий лінійний зв'язок, між витратами населення та збитками підприємств – нелінійний зв'язок.

Мультиколінеарність можна виявити за допомогою показника, який має назву «фактор інфляції дисперсії». Квадратний корінь із цього показника для кожної незалежної змінної вказує на ступінь збільшення довірчого інтервалу для параметра регресії цієї змінної порівняно з моделлю без незалежних змінних, між якими є сильна кореляція.

Фактор інфляції дисперсії можна обчислити за допомогою функції `vif()` (рис. 4).

Якщо значення квадратного кореня з цього показника перевищують 2, то це вказує на наявність мультиколінеарності. У нашому прикладі цей показник перевищив 2 для змінної *Turnoverno*.

Для правильної інтерпретації коефіцієнтів регресійної моделі необхідно, щоб дані відповідали низці умов, таких як:

– *нормальність* – значення залежної змінної розподілені нормально за фіксованих значень незалежних ознак;

– *незалежність* – значення чинникових ознак незалежні один від одного;

– *лінійність* – залежна змінна лінійно пов'язана з незалежними;

– *гомоскедастичність* – дисперсія залежної ознаки постійна за різних значень незалежних змінних. Іншими словами це явище можна назвати «однорідність дисперсії».

Якщо ці вимоги не виконані, то значення тестів істотності і довірчих інтервалів можуть бути помилковими.

Для перевірки умови нормальності використовується Q-Q-діаграма. Вона показує зв'язок між стандартизованим залишками (*Studenized Residuals*) та квантилями розподілу Стьюдента (*t Quantiles*) (рис. 5).

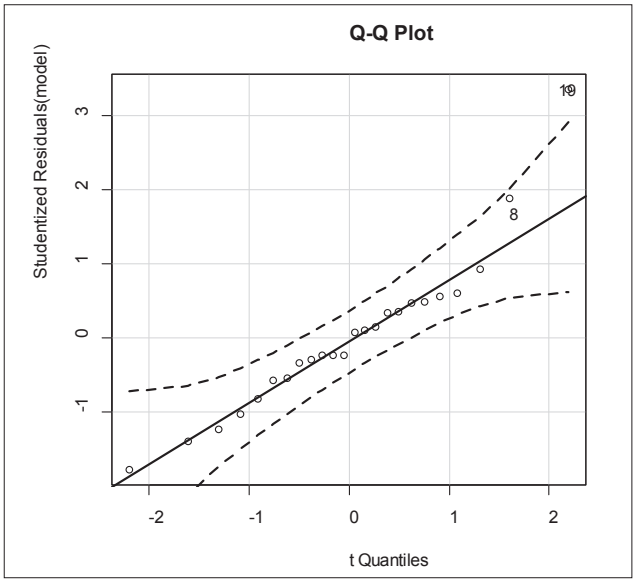

Рис. 5. Графічна перевірка нормальності розподілу (Q-Q-діаграма)

На графіку бачимо пряму лінію і 95%-ві довірчі інтервали. Якщо точки, які відповідають значенням із моделі, знаходяться щільно біля прямої та не виходять за довірчі інтервали, то є підстави вважати, що умова нормальності виконується. У нашому прикладі одиниця сукупності, яка відповідає номеру 19 (Харківська область), виходить за межі довірчого інтервалу. Фактичне значення витрат на одну особу для цієї області становить 50 662 грн, теоретичне – 45 860,84, залишок становить 4801,162, а t Стюдента – 3,3. Це означає, що ця одиниця сукупності на 3,3 середніх квадратичних відхилення відхиляється від свого середнього арифметичного.

Перевірка умови незалежності залишків передбачає перевірку автокореляції. Це явище притаманне рядам динаміки.

Наявність лінійного зв'язку між залежною та незалежними змінними можна перевірити за допомогою діаграми компонентів та залишків (рис. 6). Вони допомагають визначити систематичні відхилення від заданої лінійної моделі.

На рис. 6 маємо 4 графіки, кожен із них відповідає незалежній ознаці, яка включена в модель. По осі X відкладено фактичні значення цих ознак, по осі У – залишки. Пряма лінія на кожній моделі вказує на залежність, закладену в теоретичну модель, згладжена лінія – відповідає фактичним даним. Ми спостерігаємо дотримання лінійності на 3 графіках і частково на 4-му. Нелінійність на 4-му може свідчити про те, що некоректно сформульована функціональна форма цієї незалежної змінної у рівнянні. Можливо, треба додати нелінійні компоненти до моделі, такі як поліноміальні члени, перетворення одної чи більше змінних, або відмовитись від лінійної форми на користь будь-якого іншого її різновиду.

Для перевірки гомоскедастичності необхідно перевірити гіпотезу про сталість дисперсії залишків

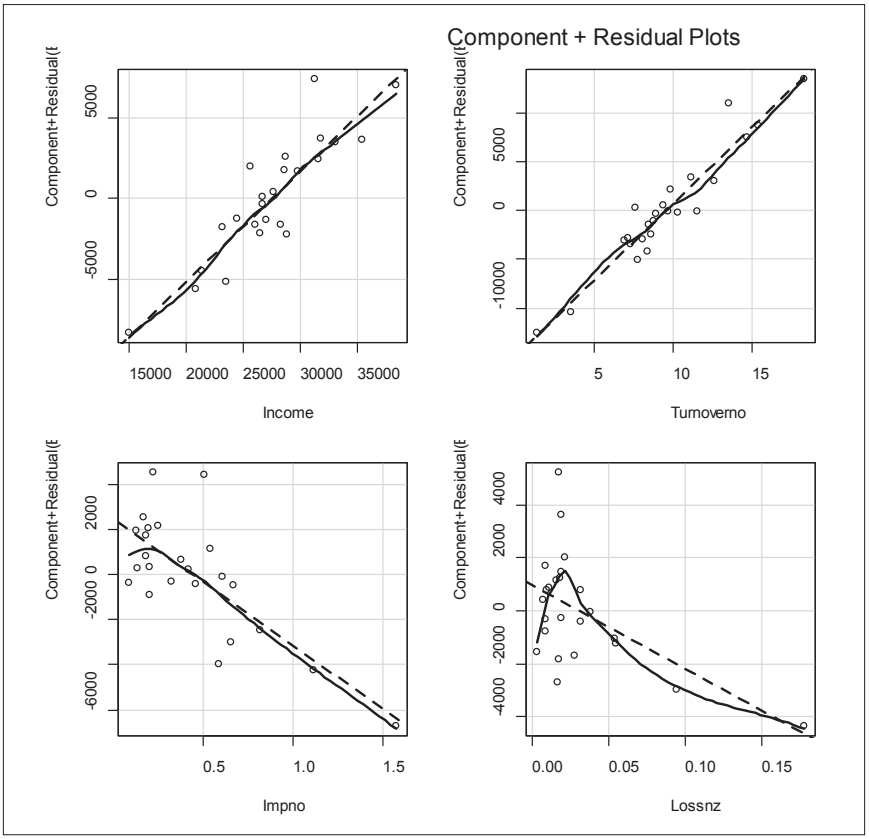

Рис. 6. Діаграма компонентів і залишків для регресійної моделі витрат домогосподарства залежно від характеристик областей (перевірка лінійності)

як альтернативу тому, що дисперсія залишків змінюється залежно від підібраних значень. Статистично істотний результат ($p < 0,05$) свідчить про гетероскедастичність (неоднорідність дисперсії залишків).

На рис. 7 показано результати перевірки моделі на гомоскедастичність (функції `ncvTest()` та `spreadLevelPlot()`). Результати тесту ($p = 0,17277 > 0,05$) свідчать про задоволення умови однорідності дисперсії.

Степеневе перетворення, яке пропонує програма (Suggested power transformation) – 0,74 – близьке до 1, отже, ніякого перетворення робити не треба.

Загальну перевірку умов застосування регресійного методу можливо здійснити за допомогою функції `gvlma()` в R (рис. 8). Програма видає свій висновок за 5 пунктами: загальна статистика, асиметрія, ексцес, форма функції, гетероскедастичність.

Отже, за рівня істотності 0,05 (тобто з імовірністю 95%) можна вважати, що умови застосування регресійної моделі виконуються.

Регресійний наліз передбачає також аналіз незвичних спостережень, а саме – нетипових значень, точок з високою напругою та впливових одиниць сукупності.

Нетипові значення – це значення, які модель погано прогнозує. Їх характеризують великі залишки. Ми вже застосували один метод виявлення нетипових значень (точка «19» на рис. 5). Функція outlierTest() рахує значення ймовірності статистичної помилки першого виду з поправкою Бонферроні для найбільшого залишку Стюдента (рис. 9).

Програма не виявила залишків, які би вказували на статистично істотні нетипові значення сукупності.

Точки високої напруги – це нетипові значення щодо інших одиниць сукупності. Вони характеризуються незвичним поєднанням незалежних змінних. Для їх ідентифікації визначають показники впливу (*hat statistic*). Результати перевірки за допомогою цього показника візуалізовано на рис. 10. Горизонтальні лінії відділяють одиниці сукупності, які в 2 та 3 рази перевищують середнє значення.

Відповідно до результатів перевірки точками з високою напругою є Закарпатська, Київська та Луганська області. Точки з високою напругою можуть бути впливовими, а можуть і не бути. Це залежить від того, чи є вони одночасно нетиповими значеннями.

Для оцінки впливових одиниць сукупності використовують відстань Кука (рис. 11), за допомогою якої ми з'ясували, що одиниця сукупності «19» виділяється на фоні інших.

Інформацію про незвичні спостереження (нетипові значення, точки високої напруги та впливові одиниці сукупності) можна звести в один графік за допомогою функції `influencePlot()` (рис. 12).

Області зі значеннями стандартизованих залишків понад 2 і менше за –2 – це нетипові значення. Області зі значенням (*hat values*) понад 0,2 і менше за 0,3 характеризуються високою напруженістю. Розмір кола пропорційний ступеню впливу одиниці сукупності.

Отже, незважаючи на те, що загальні тести за цією моделлю дали позитивні результати, ми побачили що їй притаманні певні проблеми. Ми з'ясували, що у моделі є мультиколінеарність та присутнє нетипове значення. Наші подальші дії можуть включати вилу-

```
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.858692    Df = 1     p = 0.1727759
> spreadLevelPlot(model)
Suggested power transformation:  0.7409738
```

Рис. 7. Перевірка однорідності дисперсії залишків

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model)

                     Value p-value                Decision
Global Stat        4.92316  0.2953 Assumptions acceptable.
Skewness           2.16513  0.1412 Assumptions acceptable.
Kurtosis           1.15840  0.2818 Assumptions acceptable.
Link Function      0.01186  0.9133 Assumptions acceptable.
Heteroscedasticity 1.58778  0.2076 Assumptions acceptable.
```

Рис. 8. Загальна перевірка умов застосування регресійної моделі в R

```
> outlierTest(model)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
19 3.347918          0.0035814     0.085953
```

Рис. 9. Перевірка моделі на нетипові значення в R

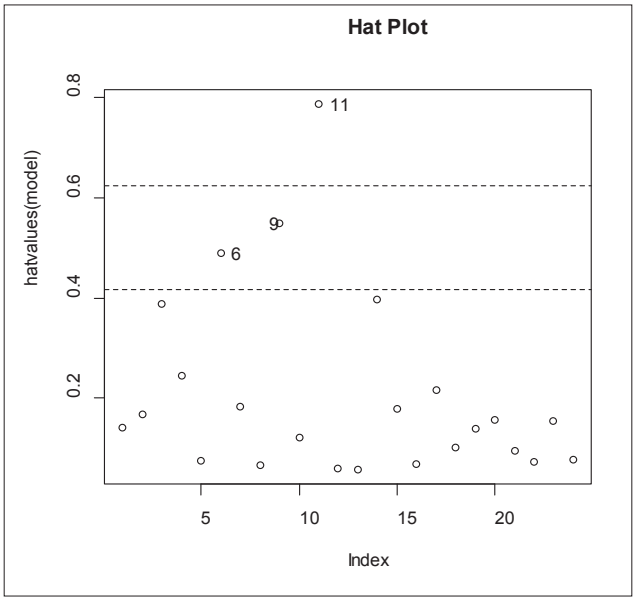

Рис. 10. Перевірка точок із високою напругою

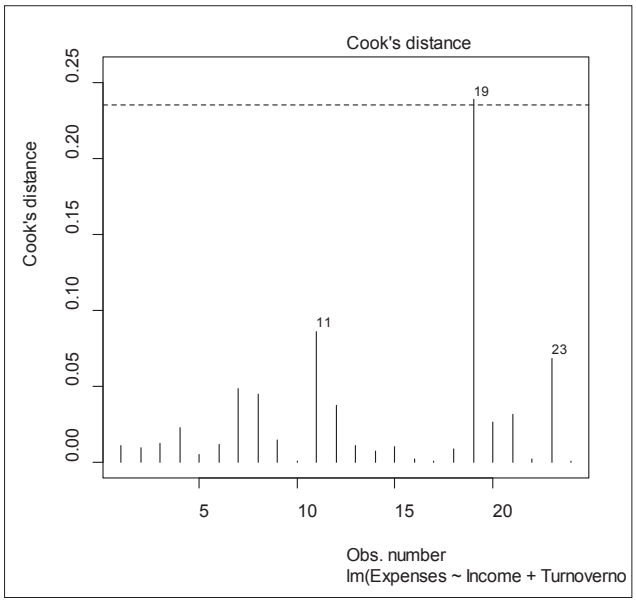

Рис. 11. Діаграма Кука для виявлення впливових одиниць сукупності

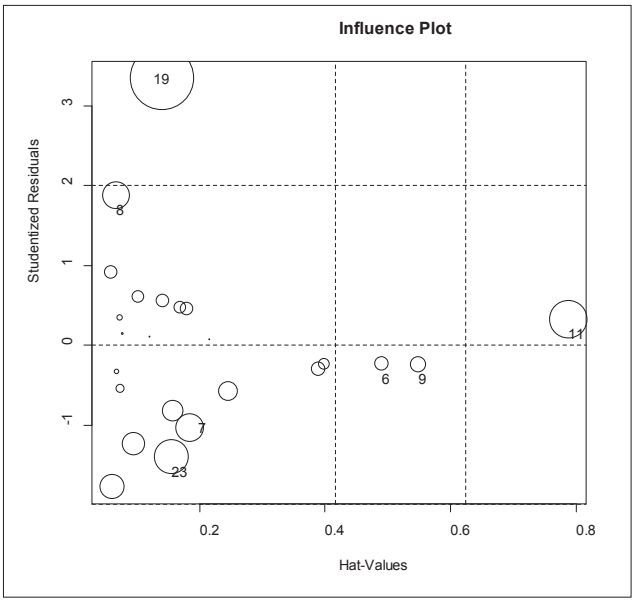

Рис. 12. Діаграма впливу одиниць сукупності на регресію

чення з моделі одиниць сукупності, перетворення, додавання чи вилучення змінних, застосування інших форм регресії або навіть інших методів дослідження. Далі йде вибір найкращої моделі шляхом порівняння найкращої з них. Продовження аналізу передбачає крос-валідацію та визначення ваги змінних, включених в модель.

Висновки. Отже, ми показали процес визначення параметрів регресійного рівняння та здійснили його діагностику. Ми переконались, що регресійний метод не можна виконувати бездумно. Необхідно насамперед вдумливо підходити до підбору чинників, які будуть формувати модель. Також ми побачили, що, незважаючи на хороші параметри істотності знайденої нами моделі, її все ж треба покращувати, а саме – усунути мультиколінеарність та видалити нетипові значення.

БІБЛІОГРАФІЧНИЙ СПИСОК:

1. Жук М. Економетричне дослідження діяльності домогосподарств в Україні / Микола Жук, Валентин Здрок // Вісник Львівського університету. Серія економічна. – 2012 – № 47. – С. 182–191.
2. Приймак З.В. Макроекономічний аналіз доходів домогосподарств країн Східної Європи / З. Приймак // Вісник Львівського університету. Серія економічна. – 2009. – № 42. – С. 529–538.
3. Піскунова О.В. Регресійний аналіз факторів, які визначають споживання продуктів харчування в регіонах України / О.В. Піскунова, О.І. Осипова // Економічний аналіз : зб. наук. праць / Тернопільський національний економічний університет; редкол.: В. А. Дерій (голов. ред.) та ін. – Тернопіль: Видавничо-поліграфічний центр Тернопільського національного економічного університету «Економічна думка», 2015. – Том 19. – № 1. – С. 230–239.

4. Бесчастна Б.В. Кореляційно-регресійний аналіз впливу соціальних факторів на продуктивність праці у сільському господарстві / Б.В. Бесчастна // Науковий вісник НУБІП України. Серія: економіка, аграрний менеджмент, бізнес. – 2015 – 221-1. – С. 48–52.
5. Лесик Ю.В. Моделювання доходів населення України / Ю.В. Лесик // Математичні методи та моделі в оподаткуванні, бізнесі, економіці : збірник тез за матеріалами XI Всеукраїнської науково-практичної інтернет-конференції. – Ірпінь: Національний університет ДФС України, 2016. – С. 61–67.
6. Кузик О.В. Макроекономічний аналіз поведінки домогосподарств в економіці України / О.В. Кузик // Дисертація на здобуття наукового ступеня кандидата економічних наук за спеціальністю 08.00.01 – економічна теорія та історія економічної думки. – Львівський національний університет імені Івана Франка, Львів, 2011.
7. Гнєушева В.О. Фінансові ресурси домогосподарств в ринкових умовах : дис. канд. екон. наук : 08.00.08 / Гнєушева Вікторія Олександрівна. – Чернігів, 2014. – С. 247.
8. Регіони України . 2016: стат. зб./ Державна служба статистики України. Київ, 2017. С. 13 – 16 – URL: http://www.ukrstat.gov.ua/
9. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / Р.И. Кабаков ; [пер. с англ. П. Волковой]. – М. : ДМК Пресс, 2014. – 580 с.