МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЧЕРКАСЬКИЙ ДЕРЖАВНИЙ ТЕХНОЛОГІЧНИЙ УНІВЕРСИТЕТ

Кваліфікаційна наукова
праця на правах рукопису

Ткаченко Олександр Миколайович

УДК 519.25:621.391

ДИСЕРТАЦІЯ

ПОЛІНОМІАЛЬНІ МЕТОДИ ТА ЗАСОБИ ОЦІНЮВАННЯ ПАРАМЕТРІВ РЕГРЕСІЇ З ВИКОРИСТАННЯМ МОДЕЛЕЙ НЕГАУСОВИХ ПОМИЛОК

122 – комп'ютерні науки

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_________________О.М. Ткаченко

Науковий керівник
Гончаров Артем Володимирович,
кандидат технічних наук, доцент

Черкаси – 2021

АНОТАЦІЯ

Ткаченко О.М. Поліноміальні методи та засоби оцінювання параметрів регресії з використанням моделей негаусових помилок. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 – комп'ютерні науки (12 Інформаційні технології). – Черкаський державний технологічний університет, Черкаси, 2021.

Дисертаційна робота спрямована на вирішення актуальної науково-технічної задачі створенні методів і засобів математичного і комп’ютерного моделювання процесів оцінювання параметрів регресійних залежностей шляхом адаптації ймовірнісних моделей досліджуваних даних для покращення точності отримуваних оцінок на основі урахування відмінностей статистичних властивостей регресійних помилок від гаусової ідеалізації.

У першому розділі роботи проведено огляд інформаційних джерел за тематикою дисертаційного дослідження, який свідчить, що не зважаючи на більш ніж двохсотлітню історію регресійного аналізу задача підвищення точності оцінювання параметрів на основі урахування специфіки регресійних моделей залишається актуальною. На сучасному етапі розвитку науково-технічного прогресу вирішення подібних задач неможливо уявити без використання комп’ютерних технологій та спеціалізованих програмних засобів основу яких, складають відповідні математичні моделі та обчислювальні методи.

Основна гіпотеза про потенційне підвищення ефективності опрацювання (критерієм якого є величина дисперсії оцінок інформативних параметрів) базується на тому, що використання класичного методу найменших квадратів забезпечує отримання оптимальних оцінок лише при виконанні ряду теоретичних припущень, зокрема при гаусовому (нормальному) розподілі помилок. В ситуаціях коли ефект нормалізації відсутній для підвищення ефективності (мінімізації дисперсії) оцінок регресійних параметрів застосовується адаптивний підхід, що базується на

максимізації правдоподібності. Проте практичне використання параметричних методів адаптивного оцінювання, ускладнюється суттєвими вимогами до наявності апріорної інформації про тип та значення параметрів розподілу регресійних помилок.

Відомо, що одним із компромісних підходів до вирішення задач, пов’язаних із опрацюванням негаусових даних, є напрямок, що базується на використанні нелінійних перетворень у вигляді стохастичних поліномів Кунченка, який надає додаткові важелі для підвищення точності порівняно із лінійними методами, які оптимізовані для гаусових моделей.

У другому розділі роботи запропоновано новий підхід до адаптивного знаходження оцінок параметрів на основі використання для опису випадкової складової регресійних моделей статистик вищих порядків, що надало можливість реалізаційно просто враховувати відхилення від гаусової ідеалізації в процесі синтезу результуючих обчислювальних методів та алгоритмів.

На основі апарату стохастичних поліномів та отриманих із використанням моментно-кумулянтного опису модифікацій регресійних моделей здійснено синтез обчислювальних методів адаптивного оцінювання параметрів регресійних моделей лінійного, поліноміального і нелінійного типу. Показано, що загальна задача, алгоритмічно може бути зведена до розв’язання системи нелінійних стохастичних рівнянь із застосуванням чисельної ітераційної процедури Ньютона-Рафсона.

Проведений у третьому розділі роботи теоретичний аналіз показав, що застосування запропонованого способу поліноміального оцінювання загалом забезпечує зменшення дисперсії отримуваних оцінок, порівняно із відомими оцінками методу найменших квадратів. При цьому ступінь ефективності не залежить від типу регресійних моделей та є однаковим для всіх складових параметрів регресії. Зростання точності досягається завдяки врахуванню негаусовості регресійних помилок. Кількісно величина ефективності поліноміальних оцінок (відносно лінійних) залежить від ступеня негаусовості, яка описується значеннями кумулянтних коефіцієнтів вищих порядків.

У четвертому розділі роботи представлено опис стуктури розробленого програмного комплексу, набір модулів якого забезпечують як безпосереднє вирішення задачі знаходження адаптивних оцінок параметрів регресійних залежностей так і реалізацією комп’ютерного статистичного моделювання на основі методу Монте-Карло і бутстреп-аналізу. Сукупність отриманих результатів статистичного моделювання у цілому підтверджують теоретично доведену ефективність поліноміальних оцінок відносно лінійних оцінок методу найменших квадратів. На прикладі моделі помилок із експоненціальним степеневим розподілом з також показано, що при відсутності апріорної інформації про значення параметрів регресійних помилок адаптивні поліноміальні оцінки можуть бути більш точними порівняно із класичними оцінкам (зростання точності щодо методу найменших квадратів складає до 60%, а відносно методу максимальної правдоподібності до 10%).

Наукова новизна дисертаційної роботи полягає у створенні методів математичного моделювання процесів оцінювання параметрів регресійних залежностей при негаусовому розподілі їх випадкових помилок на основі використання статистик вищих порядків, методу максимізації полінома (методу Кунченка), що дозволяє зменшити дисперсію інформативних оцінок параметрів і забезпечує високу якість статистичного опрацювання в комп’ютерних системах. Вперше розроблено методи адаптивного поліноміального оцінювання параметрів регресії, які за рахунок використання удосконалених на основі застосування для опису регресійних моделей статистик вищих порядків дозволяють підвищити точність оцінок інформативних параметрів в умовах апріорної невизначеності щодо ймовірнісного характеру негаусових регресійних помилок. Крім того, отримали подальший розвиток елементи теорії оцінювання параметрів регресії з використанням методу максимізації полінома, що дозволяє отримувати субефективні, адаптивні і компромісні з точки зору практичної реалізації рішення з оцінювання інформативних параметрів регресії за негаусового характеру випадкових помилок.

Практична цінність одержаних результатів визначається тим, що отримані методи та засоби математичного і комп'ютерного моделювання дали змогу: розробити методику отримання адаптивних оцінок інформативних параметрів регресійних залежностей при асиметричному і симетричному характері їх випадкових помилок, яка може бути використана для побудови нових обчислювальних алгоритмів аналізу даних; синтезувати нові алгоритми адаптивного оцінювання параметрів однофакторних і багатофакторних лінійних, поліноміальних та нелінійних регресійних залежностей, які дають змогу враховувати негаусовість регресійних залишків; провести аналіз можливості та ефективності застосування чисельного методу Ньютона-Рафсона та його модифікацій для розв'язку систем нелінійних рівнянь методу максимізації полінома для знаходження оцінок інформативних параметрів лінійних та нелінійних регресійних залежностей; розробити методику та отримати аналітичні вирази для визначення теоретичної точності поліноміальних оцінок, з використанням яких можна здійснювати порівняльний аналіз відносно оцінок класичними методами найменших квадратів та максимальної правдоподібності. На основі використання програмного пакету MATLAB і спеціалізованої для аналізу даних мови високого рівня R розроблено програмні засоби, призначені для вирішення задач статистичного моделювання знаходження оцінок параметрів регресійних залежностей.

Розроблені програмні засоби були успішно застосовані при вирішенні прикладної задачі в рамках дослідження механічних характеристик системи подачі філаменту та реологічних характеристик екструдера 3D принтера. Результати застосування запропонованих моделей та методів оцінювання дозволили забезпечити зменшення на 10-25 % величини дисперсії оцінок інформативних параметрів нелінійних регресійних моделей.

Ключові слова: регресійні моделі, негаусові помилки, моменти, кумулянти, оцінювання параметрів, стохастичні поліноми

ABSTRACT

Tkachenko O.M. Polynomial methods and tools for estimating regression parameters using Non-Gaussian error models.

Thesis for a Doctor of Philosophy degree in specialty 122 – Computer Science. - Cherkasy State Technological University. Cherkasy, 2021.

The work was solved the scientific and technical problem of development and application of mathematical and computer modeling methods for processes of estimating regression parameters under the condition of non-Gaussian character of their errors.

A new approach to the adaptive finding of parameter estimates based on the use of higher-order regression models for describing the random component is proposed, which made it possible to simply take into account deviations from Gaussian idealization in the synthesis and analysis of the effectiveness of the resulting methods.

Based on the apparatus of stochastic Kunchenko polynomials and modifications of regression models obtained using instantaneous-cumulative description, the synthesis of computational methods of adaptive parameters estimation of regression models for linear, polynomial, and nonlinear type is carried out. It is shown that the general problem can be algorithmically reduced to solving a system of nonlinear stochastic equations using a numerical Newton-Rafson iterative procedure.

The properties of polynomial estimates under the condition of the asymmetric and symmetric character of non-Gaussian errors are analyzed and their efficiency is compared with classical estimates of least squares and maximum likelihood. It is shown that the application of the proposed approach reduces the variance of polynomial estimates compared to the known least squares estimates, and the increase in accuracy is achieved by taking into account the non-Gaussian regression errors.

The developed software package, its structure, and set of modules provide both a direct solution to the problem of finding adaptive estimates of the parameters of regression dependencies and the implementation of computer statistical modeling based on the Monte Carlo method and bootstrap analysis. The set of obtained results of statistical modeling confirms the theoretically proven efficiency of polynomial estimates.

The example of the error model with exponential power distribution shows that in the absence of a priori information about the values of regression error parameters, adaptive polynomial estimates can be more accurate even compared to adaptive estimates of maximum plausibility.

Keywords: regression models, non-Gaussian errors, moments, cumulants, parameter estimation, stochastic polynomials

Список публікацій в яких опубліковані основні наукові результати дисертації:

1. Zabolotnii, S., Warsza, Z. L., & Tkachenko, O. (2018). Polynomial Estimation of Linear Regression Parameters for the Asymmetric PDF of Errors. Automation 2018. Advances in Intelligent Systems and Computing, 743, 758–772. doi:10.1007/978-3-319-77179-3_75 (**Scopus**)
2. Zabolotnii, S. W., Warsza, Z. L., & Tkachenko, O. (2020). Estimation of Linear Regression Parameters of Symmetric Non-Gaussian Errors by Polynomial Maximization Method. Automation 2019. Advances in Intelligent Systems and Computing, 920, 636–649. doi:10.1007/978-3-030-13273-6_59 (**Scopus**)
3. Zabolotnii, S., Khotunov, V., Chepynoha, A., & Tkachenko, O. (2021). Estimating parameters of linear regression with an exponential power distribution of errors by using a polynomial maximization method. *Eastern-European Journal of Enterprise Technologies*, *1*(4 (109), 64–73. https://doi.org/10.15587/1729-4061.2021.225525 (**Scopus**)

Список публікацій які засвідчують апробацію матеріалів дисертації:

4. Заболотній С.В. Поліноміальні адаптивні процедури регресійного аналізу із використанням моделей негаусових помилок на основі статистик вищих порядків/ С.В. Заболотній, О.М. Ткаченко // Тези доповідей IV Міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи) – 2017» (ComInt – 2017): Київ, 16-18 травня 2017 р. – К: КНУ ім. Т. Шевченка, – 2017. С. 113-114

5. Заболотній С.В. Застосування методу максимізації полінома для оцінювання параметрів однофакторної лінійної регресії при негаусовому розподілі помилок / С.В. Заболотній, О.М. Ткаченко // Тези доповідей VI Міжнародної науково-практичної конференції «Обробка сигналів і негаусівських процесів – 2017» (ОСНП-2017): Черкаси, 24-26 травня 2017 р. – Черкаси: ЧДТУ, – 2017. С. 74-76.
6. Заболотній С.В. Аналіз ефективності поліноміальних оцінок параметрів лінійної регресії при симетричному розподілі негаусових помилок / С.В Заболотній, М.П. Рудь, О.М. Ткаченко // Сучасні прилади, матеріали і технології для неруйнівного контролю і технічної діагностики машинобудівного і нафтогазопромислового обладнання, 14-16 листопада 2017: Тези доповідей 8-ма міжнародна н/т конф. – Івано-Франківськ, 2017. – С.130-131
7. Заболотній С.В. Особливості поліноміального оцінювання параметрів регресії при негаусовому симетричному розподілі помилок / С.В Заболотній, М.П. Рудь, О.М. Ткаченко // Автоматика та комп'ютерно-інтегровані технології у промисловості, телекомунікаціях, енергетиці та транспорті, 16-17 листопада 2017: Матеріали Всеукраїнської науково-практичної інтернет-конференції. – Кропивницький, 2017. – С.177-179
8. Заболотній С.В., Рудь М.П., Ткаченко О.М. Застосування методу максимізації полінома для оцінювання параметрів нелінійних регресійних моделей // Праці VII Міжнародної науково-практичної конференції «Обробка сигналів і негаусівських процесів», присвяченої пам'яті професора Ю.П. Кунченка: Тези доповідей. [Електронний ресурс]. – Черкаси: ЧДТУ, 2019, С. 76-79.

ЗМІСТ

ВСТУП

Актуальність теми. Регресійний аналіз є однією із найбільш поширених груп задач, що вирішуються в рамках сучасного прикладного напрямку розвитку комп’ютерних наук та інформаційних технологій відомого як «Наука про Дані» (Data Science). Бурхливий розвиток цього напрямку пов'язаний із неперервним і стрімким ростом даних, які генерує людство. Опрацювання таких обсягів даних не можливо вже уявити без використання обчислювальної техніки і відповідних програмних засобів, орієнтованих на комп’ютерне моделювання і статистичний аналіз. Серед них можна виділити як спеціалізовані комерційні продукти типу MATLAB, SAS, STATISTICA, так і високорівневі мови програмування, зокрема, R та Python, які орієнтовані на задачі аналізу даних. Математичним фундаментом цих програмних інструментальних засобів є різноманітні моделі та обчислювальні методи. Одним із основних критеріїв їх розробки та застосування є підвищення точності опрацювання на основі урахування специфіки реальних даних.

Оцінювання інформативних параметрів регресійних залежностей є однією із самих ранніх задач статистичного аналізу. За більш ніж два століття різноманітні регресійні моделі набули широкого поширення при вирішенні багатьох прикладних задач технічного, геофізичного, біомедичного, економічного та інших аспектів людської діяльності. У більшості з них для знаходження оцінок параметрів застосовуються різні варіації методу найменших квадратів (МНК), розробником якого вважають Гауса. Широке поширення МНК обумовлено тим факторам, що він дозволяє отримувати рішення у замкнутій формі, не потребує додаткової апріорної інформації про ймовірнісні властивості моделі похибок та при виконанні ряду умов теореми Гауса-Маркова є найкращим лінійним незміщеним оцінювачем. Однією із таких умов є нормалізація розподілу регресійних помилок, що забезпечує мінімізацію дисперсії оцінок параметрів. Проте в реальних ситуаціях часто може спостерігатися суттєве відхилення розподілу даних від нормального (гаусового) закону, що суттєво погіршує точність МНК-оцінок, зокрема, при наявності окремих віддалених спостережень.

Існують різні концептуальні підходи для боротьби з «негаусовістю» регресійних помилок. Один з них базується на застосуванні робастних або непараметричних процедур, зокрема, робастних версій МНК. Їх застосування спрямовано в першу чергу на забезпечення стійкості щодо впливу екстремальних відхилень. Однак існує велика кількість ситуацій, коли відхилення від гаусового розподілу породжене не «шкідливими» поодинокими викидами, а специфікою даних. У цьому випадку може бути використано підхід, заснований на нормалізуючих перетвореннях. Цей підхід дозволяє шляхом трансформації вихідних даних здійснити перехід до гаусової моделі помилок, що навіть при невеликих обсягах вибірок забезпечує можливість перевірки гіпотез і розрахунку довірчих інтервалів з використанням традиційних статистик.

Проте у випадках, коли одним із основних критеріїв є мінімізація невизначеності (дисперсії) оцінок параметрів регресійних моделей застосовується параметричний підхід, який базується на методі максимальної правдоподібності (ММП). З обчислювальної точки зору такий підхід характеризується значно більшою (порівняно з МНК) складністю, а також істотним збільшенням обсягу апріорної інформації. Це пов'язано з необхідністю попередньої специфікації (вибору) закону розподілу ймовірностей для моделі помилок та оцінюванням їх загалом неінформативних параметрів.

Дослідження останніх років показують, що компромісним з точки зору складності і повноти імовірнісного опису є підхід, який базується на застосуванні апарату стохастичних поліномів Кунченка та статистик вищих порядків, зокрема, моментів та кумулянтів. В даній роботі застосовується відносно новий підхід до статистичного оцінювання, заснований на використанні методу максимізації полінома (ММПл). Відзначимо функціональність ММПл, який вже був успішно застосований для знаходження оцінок параметрів різноманітних типів сигналів та завад, для знаходження моментів зміни (розладки) властивостей випадкових послідовностей, а також при сумісному (адаптивному) виявленні сигналів на фоні негаусових шумів. Сукупність отриманих в цих роботах результатів свідчить, що при негаусовому характері статистичних даних за рахунок додаткових нелінійних перетворень ММПл-оцінки параметрів можуть бути більш ефективними (мати меншу дисперсією) порівняно з лінійними оцінками, які оптимізовані для гаусових моделей.

Проте наведені задачі за своєю математичною постановкою суттєво відрізняються від задач регресійного аналізу. Тому, актуальність даної дисертаційної роботи обумовлена необхідністю модифікації регресійних моделей, які містять адитивну суміш детермінованої та негаусової випадкової складової, в також розробки на основі апарату стохастичних поліномів Кунченка нових обчислювальних методів статистичного оцінювання інформативних параметрів регресійних залежностей. Додатковою умовою, яка накладається на результуючі алгоритми, є їх адаптивність. Вона полягає у забезпеченні працездатності в умовах апріорної невизначеності щодо властивостей випадкової складової регресійних моделей, що є характерним для реальних ситуацій їх використання.

Зв'язок роботи з науковими програмами, планами, темами. Напрям дисертаційного дослідження відповідає планам науково-дослідних робіт Черкаського державного технологічного університету. Дисертаційна робота виконувалась на кафедрі радіотехніки, телекомунікаційних і робототехнічних систем відповідно до основних наукових напрямів та найважливіших проблем фундаментальних досліджень у галузі природничих, технічних, суспільних і гуманітарних наук НАНУ на 2019–2023 роки, зокрема «Дослідження математичних моделей, проблем комп’ютерної математики, оптимізації, оцінювання, ідентифікації». Дисертаційна робота проводилась у відповідності до планів держбюджетної НДР «Розробка технології та пристроїв адитивного виробництва індивідуальних хірургічних імплантатів та протезів з біосумісних полімерних матеріалів» (№ держ. реєстрації 0117U000937).

Мета і завдання дослідження. Мета дисертаційної роботи полягає у створенні методів і засобів математичного і комп’ютерного моделювання процесів оцінювання параметрів регресійних залежностей шляхом адаптації ймовірнісних моделей досліджуваних даних для покращення точності отримуваних оцінок на основі урахування відмінностей статистичних властивостей регресійних помилок від гаусової ідеалізації.

Для досягнення мети дослідження необхідно розв’язати такі задачі:

- здійснити аналіз існуючих підходів до оцінювання інформативних параметрів регресійних моделей за умови відмінності розподілу їх випадкової компоненти від гаусового закону;
- здійснити модифікацію опису випадкової складової регресійних моделі на основі використання опису у вигляді статистик вищих порядків для застосування апарату стохастичних поліномів Кунченка;
- здійснити синтез обчислювальних методів та алгоритмів адаптивного оцінювання параметрів регресії із застосуванням методу максимізації поліномів;
- провести аналіз властивостей поліноміальних оцінок параметрів регресії та здійснити порівняння їх ефективності із оцінками методу найменших квадратів та максимальної правдоподібності;
- розробити програмні засоби статистичного моделювання адаптивного оцінювання параметрів регресійних моделей із використанням методу Монте-Карло і бутстреп-аналізу.

Об'єкт дослідження – процеси оцінювання параметрів регресійних залежностей за умови негаусовості випадкових помилок.

Предметом дослідження є математичні моделі адитивної взаємодії детермінованої складової регресивних залежностей і негаусових випадкових регресійних помилок, що засновані на застосуванні статистик вищих порядків (моментів та кумулянтів), методи і засоби моделювання процесів оцінювання інформативних параметрів регресії, що орієнтовані на створення засобів їх комп’ютерної реалізації.

Методи дослідження. Теоретичні дослідження ґрунтуються на використанні апарату теорії ймовірності та математичної статистики, регресійного аналізу, теорії адаптивних систем, методів статистичного оцінювання і обчислювальної математики, а також методів побудови комп'ютерних засобів моделювання. Достовірність отриманих результатів і висновків перевірена порівнянням теоретичних положень з експериментальними даними, отриманими за допомогою комп'ютерного статистичного моделювання.

Наукова новизна одержаних результатів полягає у створенні методів математичного моделювання процесів оцінювання параметрів регресійних залежностей при негаусовому розподілі їх випадкових помилок на основі використання статистик вищих порядків, методу максимізації полінома (методу Кунченка), що дозволяє зменшити дисперсію інформативних оцінок параметрів і забезпечує високу якість статистичного опрацювання в комп'ютерних системах.

Вперше запропоновано:

- методи адаптивного поліноміального оцінювання параметрів регресії, які за рахунок використання удосконалених моделей регресійних залежностей в умовах апріорної невизначеності щодо ймовірнісного характеру негаусових регресійних помилок дозволяють підвищити точність оцінок параметрів регресії.

Удосконалено:

- моделі регресійних залежностей на основі використання статистик вищих порядків для опису їх випадкової складової, що дозволяє реалізаційно просто враховувати негаусовість статистичних даних.

Отримали подальший розвиток:

- елементи теорії оцінювання параметрів регресії з використанням методу максимізації полінома, що дозволяє отримувати субефективні, адаптивні і компромісні з точки зору практичної реалізації рішення з оцінювання інформативних параметрів регресії за негаусового характеру випадкових помилок.

Практичне значення одержаних результатів полягає в тому, що:

- розроблено методику отримання адаптивних оцінок інформативних параметрів регресійних залежностей при асиметричному і симетричному характері їх випадкових помилок, яка може бути використана для побудови нових обчислювальних алгоритмів аналізу даних;

- синтезовано нові алгоритми адаптивного оцінювання параметрів однофакторних і багатофакторних лінійних, поліноміальних та нелінійних регресійних залежностей, які дають змогу враховувати негаусовість регресійних залишків;
- проведено аналіз можливості та ефективності застосування чисельного методу Ньютона-Рафсона та його модифікацій для розв’язку систем нелінійних рівнянь методу максимізації поліномів для знаходження оцінок інформативних параметрів лінійних та нелінійних регресійних залежностей;
- розроблено методику та отримано аналітичні вирази для визначення теоретичної точності поліноміальних оцінок, з використанням яких можна здійснювати порівняльний аналіз відносно оцінок класичними методами найменших квадратів та максимальної правдоподібності;
- на основі використання програмного пакету MATLAB і спеціалізованої для аналізу даних мови високого рівня R розроблено програмні засоби, які можуть бути використані для вирішення задач статистичного моделювання знаходження оцінок параметрів регресійних залежностей.

Основні результати дисертаційної роботи використовуються для учбового процесу в спецкурсі «Теорія нелінійної статистичної радіотехніки» (Додаток Б), який викладається в Черкаському державному технологічному університеті.

Особистий внесок здобувача. Наукові та практичні положення дослідження, представлені в дисертаційній роботі, отримані автором особисто або при його особистій участі та підтверджені у ряді публікацій де він є співавтором. Зокрема, у роботі [1] здійснено синтез обчислювальних алгоритмів знаходження адаптивних поліноміальних оцінок параметрів лінійної регресії із використанням поліномів степені *S=2* для випадку асиметричних регресійних помилок, а роботах [2, 6, 7] із використанням поліномів степені *S=3* для випадку симетричних регресійних помилок та проведено порівняльний аналіз їх ефективності. Узагальнено застосування знаходження адаптивних поліноміальних оцінок на випадок багатофакторної лінійної регресійної моделі (у роботі [3]) та проведено порівняльний аналіз ефективності із класичними методами. У роботі [4] проведено аналіз існуючих підходів до знаходження оцінок параметрів регресійних залежностей при негаусовому розподілі помилок. У роботі [5] запропоновані модифікацію регресійних моделей із використанням часткового опису статистиками вищих порядків для застосування методу максимізації полінома. У роботі [8] здійснено узагальнення методу пошуку адаптивних поліноміальних оцінок параметрів на випадок нелінійних регресій та застосовано синтезовані моделі та методи для вирішення реальної прикладної задачі.

Апробація результатів дисертації. Основні положення дисертаційної роботи доповідалися й обговорювалися на 8 (у тому числі 2-х закордонних) науково-технічних конференціях: «Обчислювальний інтелект (результати, проблеми, перспективи) – 2017» (ComInt-2017, Київ); «Обробка сигналів і негаусівських процесів» (ОСНП-2017, ОСНП-2019 Черкаси); «Сучасні прилади, матеріали і технології для неруйнівного контролю і технічної діагностики машинобудівного і нафтогазопромислового обладнання» (Івано-Франківськ, 2017), «Автоматика та комп’ютерно-інтегровані технології у промисловості, телекомунікаціях, енергетиці та транспорті» (Кропивницький, 2017); «Advances in Automation, Robotics and Measurement Techniques» (Automation 2018, Warszawa, Poland); Progress in Automation, Robotics and Measurement Techniques (Automation 2019, Warszawa, Poland).

Публікації. Основні результати дисертації опубліковані у 8 наукових роботах, у тому числі, 1 статті у фаховому виданні України (*категорія А, індексовано у Scopus*), 2 публікації у зарубіжних наукових періодичних виданнях *індексовано у Scopus*), 5 тез доповідей у матеріалах наукових конференцій.

Структура дисертації. Дисертаційна робота складається зі вступу, 4 розділів, висновків, списку використаних джерел, що містить 117 найменувань, та додатків. Загальний обсяг дисертаційної роботи становить 138 сторінок, у тому числі 123 сторінок основного тексту, ілюстрованого 17 рисунками і 8 таблицями.

РОЗДІЛ 1
ОГЛЯД ПІДХОДІВ ДО ЗНАХОДЖЕННЯ ОЦІНОК ПАРАМЕТРІВ РЕГРЕСІЇ ПРИ НЕГАУСОВОМУ РОЗПОДІЛІ ПОМИЛОК

1.1. Задачі та моделі регресійного аналізу

Задачі регресійного аналіз є одними із найдавніших у галузі математичної статистики і їх історія сягає вже більше двохсот років [1]. Найбільш ранньою формою регресії була лінійна залежність для якої застосовувався метод найменших квадратів, розробником якого вважають Гауса, хоча сам термін «найменші квадрати» походить із більш ранньої роботи, що була опублікована Лежандром у 1805 році [2]. Як Лежандр так і Гаус застосували цей метод до проблеми визначення орбіт небесних тіл навколо Сонця на основі астрономічних спостережень. У 1821 році Гаус опублікував [3] подальший розвиток теорії найменших квадратів, включно із версією нині відомої теореми Гауса-Маркова, яка є фундаментальною теоремою в області загальних лінійних моделей. Безпосередньо сам термін «регресія» виник дещо пізніше. Вперше його запровадив британський біолог Френсіс Гальтон, який на початку XX-го ст. займався вивченням спадковості. Одним з його спостережень було те, що діти високих батьків були вищими за середній рівень, але не такими високими, як їх батьки. Ця «регресія до посередності» й дала назву сукупності відповідних статистичних моделей та методів.

Основою регресійного аналізу є моделювання взаємозв'язку між деякою залежною змінною Y (яку також називають змінною відповіді або цільовою змінною) та однією або декількома (вектор) незалежними змінними $\boldsymbol{X}$ (також відомими як «зміні передбачування» (англ. «predictor variable»), «регресори» (англ. «regressor»), «описові змінні» (англ. «explanatory variable»). Залежна змінна моделюється (1.1) як адитивна взаємодія певного функціоналу $R(\boldsymbol{\theta}, \boldsymbol{X})$ (зазвичай детермінованого) від незалежних змінних і відповідних інформативних параметрів (коефіцієнтів) регресії $\boldsymbol{\theta}$ та випадкової складової (регресійних помилок) ξ, що представляє варіації залежної змінної, які не пояснюються функцією від незалежних змінних.

$$Y = R(\boldsymbol{\theta}, \boldsymbol{X}) + \xi. \tag{1.1}$$

Припущення про характер розподілу випадкових регресійних помилок називаються гіпотезою породження даних. Ця гіпотеза грає центральну роль у виборі критерію оцінки якості моделі і, як наслідок, в способі налаштування її параметрів.

Якщо регресійна модель адекватно відображає справжній зв'язок між змінною відповіді та незалежними змінними, ця модель може бути використана для прогнозування залежної змінної, виявлення важливих незалежних змінних та встановлення причинно-наслідкових зв'язків між ними. Модель регресії часто в значній мірі спирається на деякі припущення, що повинні виконуватися. Регресійний аналіз іноді критикують при застосуванні їх у тих випадках, коли неможливо перевірити відповідні припущення. Одним із важливих факторів такої критики є той факт, що регресійну модель легше піддавати критиці, ніж знайти той метод, що відповідає моделі регресії [4].

Існує декілька способів класифікації існуючих типів регресійних моделей. Один із них візуально представлений на рис.1.1.

Розрізняють одновимірну (однофакторну) і багатовимірну (багатофакторну) регресію, тобто залежність цільової змінної від однієї або декількох незалежних змінних. За типом детермінованої функції, що пов'язує залежну змінну із незалежними змінними та інформативними параметрами також розрізняють лінійну і нелінійну регресію. При цьому можливі додаткові варіації.

- Регресійна модель може бути загалом нелінійною щодо незалежних змінних, але при цьому представляти собою суперпозицію у вигляді лінійної комбінації із ваговими коефіцієнтами (параметрами) від набору деяких функцій від незалежних змінних. Прикладом такого типу моделей є поліноміальна регресія, яка розглядається у даній роботі.

- Деякі нелінійні залежності, наприклад, експоненціального або логарифмічного типу, шляхом певних функціональних перетворень можуть бути перетворені в лінійні (лінеаризовані). Проте використання цього підходу вимагає обережності, оскільки змінюється ступінь впливу значень даних та структура помилок моделі, що необхідно враховувати при оцінюванні параметрів та інтерпретації отриманих результатів.

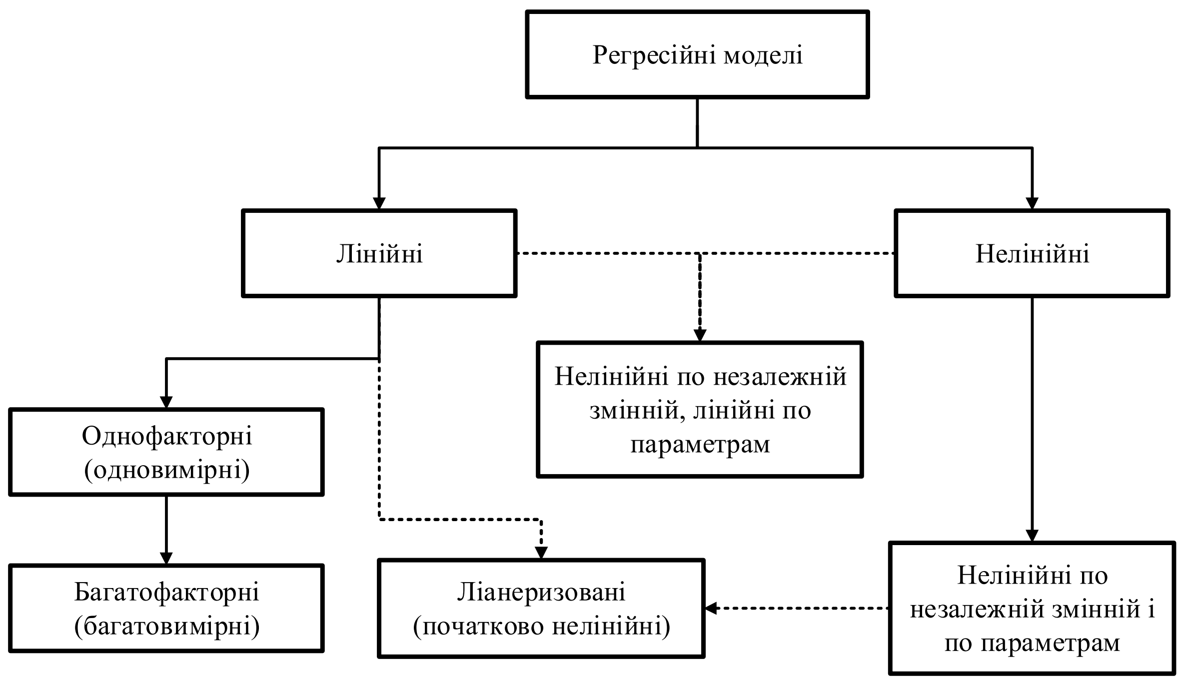


Рисунок 1.1 – Типи регресійних моделей

Окрім того, розрізняють параметричні і непараметричні регресії. Хоча точну межу між цими термінами провести складно і в даний час немає загальноприйнятого критерію відмінності одного типу моделей від іншого. Зокрема, вважається, що лінійні моделі є параметричними, а моделі, що включають усереднення залежної змінної по простору незалежних змінних – непараметричними [5].

На сьогодні регресійний аналіз є одним із найбільш часто використовуваних на практиці статистичних підходів. Його застосування можна знайти в багатьох наукових і прикладних галузях, включаючи медицину, біологію, сільське господарство, економіку, машинобудування, соціологію, геологію та ін. При вирішенні задач регресійного аналізу постають такі основні питання [4]:

- Як вибрати структуру і тип детермінованої складової регресійної моделі і якому саме сімейству вона повинна належати?
- Яка гіпотеза породження даних про розподіл випадкової складової регресійної?
- Якою цільовою функцією оцінити якість апроксимації?
- Як відшукати інформативні параметри моделі та яким повинен бути алгоритм оптимізації параметрів?

Інформативні параметри регресії необхідно оцінювати так, щоб модель максимально відповідала даним. Значення параметрів оцінюються із використанням певного критерію. Найбільш часто таким критерієм є мінімізація квадрату помилок. Але іноді специфіка реальних даних призводить до необхідності використання інших критеріїв та альтернативних методів оцінювання параметрів. Зокрема, в даній роботі і розглядається одна із таких ситуацій, пов’язана із специфікою розподілу випадкової складової регресійних моделей.

1.2. Програмні засоби для регресійного аналізу

Очевидно, що на нинішньому етапі розвитку комп’ютерних наук та інформаційних технологій опрацювання даних неможливо вже уявити без застосування певних спеціалізованих програмних засобів. Для вирішення задач регресійного аналізу розроблено велика кількість різноманітних програмних комплексів, серед яких виділимо найбільш відомі комерційні продукти [6]:

- Система статистичного аналізу (SAS), розроблена SAS Institute, Inc., є одним із популярних програмних засобів, який може бути використаний для проведення регресійного аналізу. Система SAS - це інтегрована система програмних продуктів, яка дозволяє користувачам здійснювати введення даних та управління даними, створювати статистичні графіки, проводити широкий спектр статистичного аналізу, отримувати дані із платформи сховища даних та забезпечувати динамічний інтерфейс з іншим програмним забезпеченням тощо.
- Програмне забезпечення S-PLUS, розроблене компанією Insightful Inc., є ще одним з популярних програмних засобів, яке використовується аналітиками в різних наукових галузях. Це програмне забезпечення є потужним обчислювальним інструментом, що охоплює широкий спектр методів статистики. S-PLUS пропонує широку колекцію спеціалізованих модулів, які надають додаткову функціональність S-PLUS у таких сферах, як прогнозування волатильності, оптимізація, аналіз клінічних випробувань, екологічна статистика та аналіз просторових даних. Крім того, у користувачів є можливість написання програми або функції для S-PLUS за допомогою спеціалізованої мови S.
- Статистичний пакет для соціальних наук SPSS також є одним із найбільш широко використовуваних програм для статистичного аналізу в галузі соціальних наук. Це одне з найкращих програмних засобів, що використовується дослідниками ринку, охорони здоров'я, освіти та ін.

Проте в останні кілька десятиліть саме MATLAB від Mathworks залишався домінуючою мовою для наукових обчислень та аналізу даних. Відомий своїм зручним інтерфейсом та гарною підтримкою він надав науковому та промисловому співтовариству складний і багатогранний набір програмних засобів, що дозволяє дослідникам легко керувати, обробляти та візуалізувати набори даних. Фактично

MATLAB прийнятий більшістю світових лабораторій та університетів як стандартний інструмент для комп'ютерної математики та моделювання і активно використовується для вирішення задач із сфери «науки про дані». Модуль MATLAB «Statistics and Machine Learning Toolbox» містить дуже велику кількість інструментів для опису, аналізу та моделювання даних. В цьому модулі імплементована велика кількість різноманітних моделей, методів та обчислювальних алгоритмів регресії та класифікації, які дозволяють робити висновки з даних та будувати прогнозні моделі або інтерактивно за допомогою спеціалізованого додатку «Classification and Regression Learner apps», або шляхом написання програмного коду із використанням засобів автоматизованого машинного навчання [7].

Проте необхідно відзначити достатньо високу ціну подібних комерційних продуктів, що часто спонукає до все більш широкого використання безкоштовних засобів, що базуються на високорівневих мовах програмування. Наприклад, найпопулярнішими мовами «Науки про дані» є R та Python. Хоча Python є більш зручним для написання нових алгоритмів, що досягається за рахунок його вищої гнучкості, простого синтаксису і чітко декларованого стилю, що полегшують як написання коду, так і пошук помилок [8]. З іншого боку, R як мова програмування, що створена статистиками для статистиків, має переваги, коли потрібно провести складний статистичний аналіз даних, якісно та гарно візуалізувати, а також здійснити маніпуляції з відносно великими базами даних за допомогою інструментів, що вже існують. А завдяки тому, що R - відкритий ресурс із потужною спільнотою, таких інструментів різного рівня складності надзвичайно багато. Більшість відомих методів регресійного аналізу вже розроблені із застосуванням мови R та розміщені у відкритих репозиторіях, основним із яких є Comprehensive R Archive Network (CRAN) https://cran.r-project.org/ [9].

Саме широке поширення програмних продуктів MATLAB та R в галузі опрацювання даних спричинило їх вибір як інструменту для імплементації результатів даного дисертаційного дослідження.

1.3. Класичні методи оцінювання параметрів регресії

В класичному регресійному аналізі для знаходження оцінок інформативних параметрів застосовуються два основні методи:

- метод найменших квадратів (МНК).
- метод максимальної правдоподібності (ММП).

Широке поширення МНК обумовлено двома основними факторами. По перше, алгоритмічною простотою його реалізації. По друге, тим відомим фактом, що при виконанні ряду умов оцінки звичайного МНК є оптимальними у сенсі забезпечення мінімуму їх дисперсії. Умови ж, при яких можна використовувати ММП є більш жорсткими, оскільки цей метод вимагає явного завдання певного виду розподілу ймовірностей статистичних даних. З іншого боку, ММП більш універсальний і зазвичай більш ефективний, оскільки використовує біль повний ймовірнісний опис регресійної моделі. Проте його застосування з алгоритмічної точки зору призводить до більш складних рішень.

Коротко розглянемо теоретичні основи цих методів оцінювання.

1.3.1. Метод найменших квадратів

Метод найменших квадратів полягає у знаходженні таких оптимальних значень параметрів $\boldsymbol{\theta}$ регресійної моделі (1.1), при яких мінімізується сума квадратів регресійних залишків, тобто

$$MSE = \frac{1}{N}\sum_{v=1}^{N}[y_v - R_v(\boldsymbol{\theta},\boldsymbol{X})]^2 \to min, \qquad (1.2)$$

де N – обсяг вибірки даних.

У випадку коли детермінована складова регресії є лінійною функцією відносно вектору параметрів $\boldsymbol{\theta}$ і матриці регресорів $\boldsymbol{X}$, тобто $R(\boldsymbol{\theta},\boldsymbol{X}) = \boldsymbol{X\theta}$, то величина MSE є відстанню від вектору $\boldsymbol{Y}$ до вектору $\boldsymbol{X\theta}$. Тоді відшукання МНК оцінок $\widehat{\boldsymbol{\theta}}$ еквівалентно задачі відшуканню такої точки$\boldsymbol{X}\widehat{\boldsymbol{\theta}}$, яка лежить найближче (в евклідової метриці) до $\boldsymbol{Y}$ і знаходиться при цьому в просторі стовпців матриці регресорів $\boldsymbol{X}$. Результуюче рішення відповідно до методу найменших квадратів складається з пошуку коренів системи нормальних виду [5]

$$(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\theta}^T = \frac{1}{N}\sum_{v=1}^{N}\boldsymbol{X}^T\boldsymbol{Y}^T, \qquad (1.3, a)$$

де результуючі МНК-оцінки можна записати як

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{N}\sum_{v=1}^{N}\boldsymbol{X}^T\boldsymbol{Y}^T. \qquad (1.3, б)$$

У випадку коли детермінована складова регресії $R(\boldsymbol{\theta},\boldsymbol{X})$є нелінійною функцією від для Q -розмірного параметру $\boldsymbol{\theta}$ для знаходження мінімуму функції MSE, формується система рівнянь із часткових похідних по кожному p -му елементу вектору параметрів

$$\sum_v R_v(\boldsymbol{\theta},\boldsymbol{X})\frac{\partial}{\partial\theta_p}R_v(\boldsymbol{\theta},\boldsymbol{X}) = 0, p = \overrightarrow{1,Q}. \qquad (1.4)$$

Оскільки в загальному випадку функції MSE не має єдиного мінімуму [10], то застосовується ітераційна процедура покрокового наближення $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{(k+1)} + \Delta\widehat{\boldsymbol{\theta}}$. На кожному k -у кроці ітерації на основі використання наближення рядами Тейлора нелінійна регресійна модель лінеаризується відносно параметрів $\boldsymbol{\theta}^{(k)}$

$$R_v(\boldsymbol{\theta},\boldsymbol{X}) \approx R_v\left(\boldsymbol{\theta}^{(k)},\boldsymbol{X}\right) + \sum_p \frac{\partial}{\partial\theta_p}R_v\left(\boldsymbol{\theta}^{(k)},\boldsymbol{X}\right)\left[\boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}\right] \approx R_v\left(\boldsymbol{\theta}^{(k)},\boldsymbol{X}\right)\sum_p H_{v,p}\,\Delta\boldsymbol{\theta}, \qquad (1.5)$$

де $H_{v,p}$ - елементи матриці Якобі $\boldsymbol{H}$, що залежать від компонентів θ_pвекторного параметрів $\boldsymbol{\theta}$ при фіксованих значеннях регресорів $\boldsymbol{X}$.

Результуюча нормалізована система рівнянь може бути представлена у вигляді

$$(\boldsymbol{H}^T \boldsymbol{H})\Delta\boldsymbol{\theta}^T \approx \frac{1}{N}\sum_{v=1}^{N} \boldsymbol{H}^T \Delta\boldsymbol{Y}^T. \tag{1.6}$$

Для знаходження оптимальних МНК-оцінок параметрів нелінійних регресійних моделей як розв'язки системи (1.6) використовуються градієнтні чисельні методи, метод Ньютона-Гауса або алгоритм Левенберга-Марквардта [11-13].

1.3.2. Метод максимальної правдоподібності

Альтернативою мінімізації квадратичного функціоналу (1.2) методу найменших квадратів є пошук максимуму функції правдоподібності. Загалом, термін «максимум правдоподібності» вперше був використаний в роботі Фішера майже 100 років тому [14].

В загальному ММП дозволяє відновлювати щільність розподілу ймовірностей по вибіркових значеннях $\boldsymbol{Z} = \{z_1, z_2 \ldots z_N\}$, якщо загальний вигляд щільності імовірнісного розподілу описується апріорно відомою функцією $p(\boldsymbol{Z}, \boldsymbol{\theta})$, що залежить від векторного параметру $\boldsymbol{\theta}$. ММП-оціночні компонентів параметра $\boldsymbol{\theta}$ знаходяться за умови максимізації функціоналу виду

$$\widehat{\boldsymbol{\theta}} = arg \ \ \underset{\boldsymbol{\theta}}{min}[L(\boldsymbol{Z}, \boldsymbol{\theta})], \tag{1.7}$$

де

$$L(\boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{v=1}^{N} p(z_v, \boldsymbol{\theta}), \tag{1.8}$$

Шукані оцінки компонентів векторного параметру можуть бути отримані шляхом чисельного розв'язку системи нелінійних рівнянь

$$\sum_{v=1}^{N} \frac{\partial}{\partial \theta_p} ln\, p\, (z_v, \boldsymbol{\theta}) = 0, p = \overline{1,Q} \tag{1.9}$$

Наприклад, у випадку задачі знаходження оцінок параметрів регресії загального виду (1.1) за умови, що випадкова складова ξ має центрований нормальний (гаусовий) розподіл

$$p(z,\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-z^2}{2\sigma^2}\right),$$

функціонал (1.8) можна представити у вигляді

$$L(\boldsymbol{X},\boldsymbol{\theta}) = \prod_{v=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-[y_v - R_v(\boldsymbol{\theta},\boldsymbol{X})]^2}{2\sigma^2}\right), \qquad (1.10)$$

Очевидно, що точка екстремуму функції правдоподібності буде співпадати із екстремумом функції

$$ln\, L\,(\boldsymbol{X},\boldsymbol{\theta}) = \sum_{v=1}^{N} \left[\frac{1}{2} ln(2\pi) + ln(\sigma) + \frac{1}{2\sigma^2} [y_v - R_v(\boldsymbol{\theta},\boldsymbol{X})]^2\right]. \quad (1.11)$$

При цьому важливим фактом є те, що точка максимуму (1.11) співпадає із точкою мінімуму (1.2), а це означає, що при гаусовому характері помилок регресійної моделі ММП-оцінки вироджуються в лінійні МНК-оцінки для в незалежності від типу детермінованої складової регресійної залежності (1.1) [15].

1.3.3. Властивості оцінок параметрів регресії

Як вже зазначалося вище в класичному регресійному аналізі зазвичай вводяться певні теоретичні припущення щодо властивостей регресійних моделей. Зокрема:

1) Математичне сподівання випадкової складової ξ є тотожним нулю $E[\xi] = 0$.
2) Дисперсія випадкової складової ξ є постійною (гомоскедакстичною) для всіх спостережень $D[\xi] = const$.

3) Окремі відліки випадкової складової ξ некорельовані між собою $E\left[\xi_i \xi_j\right] = 0, \quad i \neq j$.
4) Регресори $\boldsymbol{X}$ і відліки випадкової складової ξ незалежні для всіх спостережень $E\left[x_{v,i} \xi_v\right] = 0$.
5) Розподіл відліків випадкової складової ξ описується нормальним (гаусовим) законом.

Перші 4-и припущення складають умови відомої теореми Гауса-Маркова, при виконанні яких МНК-оцінкам параметрів лінійної регресії притаманна незміщеність, слушність та ефективність. Тобто такі оцінки матимуть найменшу можливу дисперсію в класі лінійних незміщених оцінок. Порушення однієї із умов теореми Гауса-Маркова призводить до порушення ефективності оцінок, тобто в класі незміщених оцінок можна знайти такі, які мають меншу дисперсію [16].

Останнє припущення про гаусовий закон розподілу випадкової складової регресійних моделей є, можливо, найбільш дискусійним. Це можна пов’язати із тим, що загальне теоретичне припущення про нормалізацію розподілу випадкових величин доводиться в рамках центральної граничної теореми. А відповідно до її умов таке припущення можна вважати правдоподібним лише якщо випадкові величини породжуються в результаті впливу великої кількості незалежних випадкових факторів, кожен з яких не обов'язково має нормальний розподіл [17]. Проте в реальних задачах розподіл статистичних даних (у тому числі помилок регресійних моделей) часто в тій чи іншій мірі відрізняється від нормального.

Багато дослідників зазначають, що результати спостережень, і взагалі статистичні дані, мають властивості, які призводять до того, що моделювати їх необхідно випадковими величинами з розподілами, більш-менш відмінними від нормального закону [18]. При цьому в одних випадках такі розподіли будуть істотно відрізняються від гаусової ідеалізації, а в інших нормальні розподіли можуть розглядатися як деяка базова апроксимація. Зокрема, як показують результати багатьох чисельних експериментів, здійснених при дослідженні законів розподілу різного роду похибок вимірювання, в більшості ситуацій їх розподіл суттєво відрізняється від гаусової моделі [19]. Тому в залежності від ступеня такої розбіжності може виникнути необхідність пошуку альтернативних моделей та методів оцінювання, які б враховували специфіку даних і забезпечували більш точні рішення. Подібна проблема, пов'язана із можливою специфікою негаусового розподілу регресійних помилок, і розглядається у даному дисертаційному дослідженні.

1.4. Способи оцінювання параметрів умовах негаусового розподілу регресійних помилок

Постійне порушення різноманітних теоретичних умов (подібних до припущень теореми Гауса-Маркова або припущення про нормалізацію розподілів помилок в рамках центральної граничної теореми) при опрацюванні реальних статистичних даних призвело до того, що у другій половині XX-го століття сформувалося ряд науково-прикладних напрямів, присвячених подоланню або мінімізації подібних проблем.

Відомо, що з прикладної точки зору коли існує певна параметрична модель, яка, як передбачається, є хорошим наближенням до реальної ситуації, то не можна бути точно впевненим, що вона ідеально правильна. Тому будь-яка статистична процедура повинна мати наступні бажані властивості [20]:

- Ефективність: вона повинна мати досить хороший (оптимальний або майже оптимальний) коефіцієнт корисної дії на передбачуваній моделі.
- Стійкість: вона повинна бути надійною в тому сенсі, що невеликі відхилення від припущень моделі повинні лише незначно погіршувати результативність, тобто кількісна характеристика (наприклад, величина асимптотичної дисперсії оцінок або рівня потужності статистичного тесту) повинен бути близькою до номінального значення, розрахованого на моделі.

Статистичні методи, як направлені на забезпечення властивості ефективності на основі урахувань специфіки моделей зазвичай називають адаптивними. А методи, які в основному направлені на забезпечення стійкості – робастними. І хочу не існує чіткої межі, яка розділяє ці методи, проте такий поділ обумовлений тим, що застосування механізмів, які направлені на досягнення цих двох властивостей може призводити до внутрішніх протиріч між ними.

Існуючі підходи до оцінювання параметрів регресії, призначені для опрацювання даних в умовах негаусового характеру розподілу регресійних помилок також можна поділити на декілька груп (див. рис. 1.2).

До першої групи відносяться робастні процедури та методи, які орієнтовані на забезпечення стійкості оцінок параметрів при наявності у вибірках екстремальних значень (так званих «викидів»), тобто спостережень, які сильно відрізняються від інших. З точки зору статистики, розподіл даних, що містять викиди, схожі до нормальним (іноді їх називають «забруднені», але має більш «важкі» хвости. Однією із перших стратегій для таких ситуацій – це «очищення» викидів (цензурування вибірок) [21]. І хоча початково, такий підхід дійсно мав право на існування, проте, з розвитком регресійного аналізу, стало зрозуміло, що він не працює для багатьох задач, зокрема при багатовимірній регресійній моделі. Це спричинило до розробки великої кількості підходів, реальне застосування яких стало можливим саме завдяки комп'ютерним технологіям. Серед них виділимо ряд найбільш поширених сімейств робастних оцінок [22–24]:

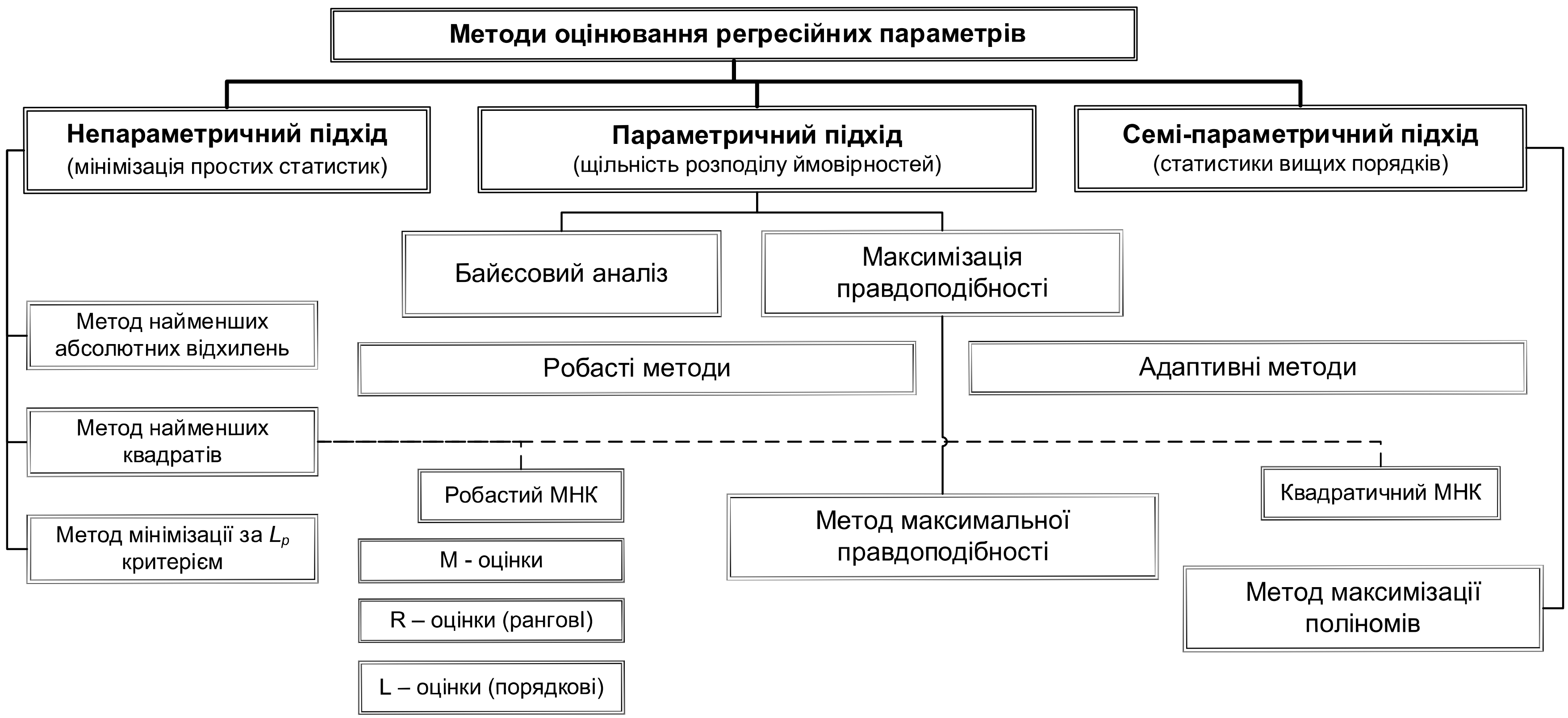


Рисунок 1.2 – Класифікація методів оцінювання параметрів регресії

- M-оцінки, засновані на ідеї методу максимальної правдоподібності.
- R-оцінки, засновані на аналізі рангів кожного спостереження.
- L-оцінки, які є лінійними функціями порядкових статистик, стійкі до великої кількості видів (до 50% спостережень).

Проте існує цілий ряд ситуацій, наприклад, в біологічних [25] або технічних [26] системах, коли відмінність від гаусових моделі породжена не «шкідливими» поодинокими викидами, а специфікою самих даних. У цьому випадку може бути використано підхід, заснований на нормалізуючих перетвореннях: Бокса-Кокса [27], Джонсона [28] та ін. Цей спосіб дозволяє шляхом нелінійної трансформації вхідних даних здійснити перехід до гасової моделі помилок, що навіть при невеликих обсягах вибірок забезпечує можливість перевірки гіпотез і розрахунку довірчих інтервалів інформативних параметрів з використанням традиційних t-статистик Стьюдента. Основним недоліком такого підходу є те, що разом із регресійними помилками відбувається нелінійна трансформація детермінованої складової регресійних моделей. А такі процеси можуть суттєво ускладнити як підбір необхідних моделей так і інтерпретацію кінцевих результатів регресійного аналізу.

Ще одним способом покращення точності оцінок регресійних параметрів для негаусових даних є заміна квадратичної функції втрат на іншу. Ця ідея не є новою, оскільки ще у 1793 році Лаплас запропонував метод оцінювання на основі мінімізації абсолютних відхилень (MAB), відомий також як оцінювач регресії у просторі L_1 [29]. Такий оцінювач порівняно із МНК є більш ефективним для лептокуртичних розподілів помилок із більш тяжкими ніж у гаусового закону хвостами. Узагальненням такого підходу є побудова регресійних оцінювачів у просторі L_p, розвинутий в роботах [30–32]. Відомо, що кожна із вищезазначених функцій втрат має відповідні моделі розподілів при яких оцінки максимальної правдоподібності є еквівалентними оцінкам методів, що мінімізують відповідні втрати. Зокрема, оцінювач МНК відповідає ММП, коли розподіл помилок є нормальним; оцінювач MAB еквівалентний ММП за розподілу Лапласа, а оцінювач

L_p відповідає ММП, коли регресійні помилки описуються експоненціальним степеневим розподілом (ЕСР) [32].

У тих випадках, коли основним критерієм є саме ефективність (мінімізація дисперсії) оцінок параметрів регресійних моделей застосовується саме адаптивний підхід, який часто базується на різних модифікаціях методу максимальної правдоподібності [32–42]. Як вже зазначалося раніше, ключовим моментом застосування ММП є апріорне визначення типу розподілу ймовірностей. Для вирішення задач регресійного аналізу використовується велика кількість різноманітних як асиметричних [33–36] так і симетричних типів розподілів [32, 37–41]. Крім того, спеціально розроблені моделі розподілів, що дозволяють змінювати величину коефіцієнтів асиметрії, ексцесу і тяжкість хвостів регресійних помилок [42–45].

Очевидно, що при використанні параметричного підходу із реалізаційної точки зору вирішення загальної задачі ускладнюється тим, що окрім підбору адекватної ймовірнісної моделі розподілу помилок необхідно здійснювати сумісне оцінювання їх параметрів, невизначеність яких суттєво впливає на точність оцінок інформативних параметрів регресії. При виконанні припущення, що помилки моделі не залежать від регресорів може бути застосований підхід, який передбачає ряд послідовних кроків. Спочатку здійснюється оцінювання параметрів регресії простим методом, який не потребує додаткової інформації про специфіку ймовірнісного розподілу статистичних даних. Потім на основі аналізу регресійних залишків здійснюється ідентифікація типу випадкової складової моделі та знаходять оцінки її параметрів. І на останньому етапі знаходять уточнені (адаптивні) оцінки параметрів із урахуванням апостеріорних властивостей похибок. Цей підхід використовується і в даному дослідженні.

Адаптивне підхід може базуватися не лише на максимізації правдоподібності, яка потребує визначення закону розподілу ймовірностей і знаходження оцінок його параметрів. Це є окремою і часто складною в обчислювальному плані задачею. Альтернатива може базуватися на використанні ймовірнісного опису регресійних моделей у вигляді статистик вищих порядків, наприклад, моментів або кумулянтів. Відомо [46], що такий спосіб опису негаусових випадкових величин є компромісним з точки зору складності і повноти відображення ймовірнісних властивостей. Прикладами його використання для вирішення завдань регресійного аналізу, є роботи [47–51].

Використання моментно-кумулянтного опису у поєднанні із адаптивним підходом до оцінювання лежить в основі квадратичної модифікації МНК [52]. Різні варіації цього метода оцінювання проявляють свою ефективність при суттєвій асиметрії розподілів, використовуючи як додаткову інформацію значення апостеріорних оцінок коефіцієнтів асиметрії та ексцесу регресійних залишків [53–55]. За нашою думкою, цей підхід до оцінювання параметрів регресії концептуально є найближчим до методів, заснованих на апаратові стохастичних поліномів Кунченка, який використовується у даному дослідженні.

1.5. Застосування стохастичних поліномів Кунченка при вирішенні статистичних задач

1.5.1. Стохастичні поліноми та їх властивості

Відомо, що відправною точкою фундаментальних досліджень проф. Ю.П.Кунченка щодо розробки апарату стохастичних поліномів були ряди Вольтера, які він застосував для розробки нового методу знаходження оцінок невідомих параметрів випадкових величин та процесів, відомого як метод максимізації поліномів (ММПл) [56]. В основі цього методу лежить одна із базових властивостей стохастичних поліномів, яка полягає у тому, що їх математичне сподівання, як функція оцінюваного параметра, при певним чином підібраних коефіцієнтах має глобальний максимум в околі дійсного значення цього параметра. Основною перевагою ММПл є те, що зі зростанням порядку полінома дисперсія оцінок зменшується й асимптотично прагне до ефективних значень [57].

Теоретична можливість використання модифікації ММПл, розробленої для знаходження оцінок векторного параметру при неоднаково-розподілених вибіркових значеннях для оцінювання параметрів регресійних моделей вперше була задекларована в роботі [58] ще майже 30 років тому. Проте достатньо довгий час цей метод знаходив використання лише для задач, дотичних до однієї специфічної галузі – статистичної радіотехніки, серед них можна виділити:

- оцінювання параметрів негаусових та полігаусових випадкових величин [59–61];
- оцінювання кута надходження гармонічного сигналу на антенну решітку за умови дії негаусових завад [62, 63];
- оцінювання параметрів постійних, імпульсних, гармонічних та радіосигналів на фоні адитивних негаусових завад [64–68];
- оцінювання параметрів постійного та радіосигналів на фоні мультиплікативних та адитивно-мультиплікативних негаусових завад [69, 70].

Серед сукупності отриманих в рамках наукової школи проф. Ю.П.Кунченка результатів застосування ММПл в можна роботи [71–74], присвячені знаходженню оцінок параметрів лінійних та поліноміальних трендів при негаусовій стохастичній компоненті, які є найближчими до даного дослідження.

Іншим прикладним напрямком застосування апарату стохастичних поліномів є задачі, пов’язані із виявленням та розпізнаванням сигналів. Отримані на поточний момент результати умовно можна поділити на дві окремі групи.

В основі результатів першої групи лежить властивість стохастичних поліномів, яка полягає в їх здатності зменшувати дисперсію випадкових величин та процесів при розкладі їх у просторі з порідним елементом (просторі Кунченка) [75]. Його застосування відкриває перспективи для побудови нових методів виявлення розладки (різких змін властивостей) випадкових послідовностей при вирішенні задач ймовірнісної діагностики [76], а також при розпізнаванні сигналів та образів [77–80].

Друга група базується на представленні логарифма відношення правдоподібності у вигляді стохастичних поліномів з коефіцієнтами, оптимальними за моментними критеріями якості синтезу вирішних правил перевірки статистичних гіпотез [81, 82]. На основі цього підходу розроблена велика кількість моделей та методів, призначених для вирішення задач виявлення сигналів на фоні негаусових завад з різними ймовірнісно-кореляційними властивостями [83–86].

Аналіз сукупності результатів застосування апарату стохастичних поліномів Кунченка свідчить про їх орієнтацію на вирішення тих задач, специфікою яких є негаусовий характер статистичних даних. Результуючі моделі, методи засновані на використанні нелінійних (поліноміальних) перетворень, що приводить до зменшення ймовірностей помилок або дисперсії оцінок, порівняно з більш простими лінійними алгоритмами, які є оптимальними для гаусової моделі (наприклад МНК). З іншого боку, поліноміальні алгоритми, в загальному випадку, можуть бути менш ефективними в порівнянні з алгоритмами, заснованими на параметричних процедурах максимізації правдоподібності. Проте на відміну від параметричного підходу, який потребує наявності апріорної інформації про розподіл моделі даних, алгоритми, що базуються на властивостях стохастичних поліномів, використовують неповний імовірнісний опис у вигляді статистик вищих порядків, наприклад, послідовності моментів або кумулянтів. Це значно зменшує обсяг необхідної апріорної інформації, спрощує процеси синтезу та аналізу алгоритмів, а також дозволяє реалізовувати можливість адаптивності.

Таким чином аналіз зазначених реалізаційних властивостей апарату стохастичних поліномів Кунченка дозволяє сформулювати наукову гіпотезу про потенційну ефективність застосування даного математичного апарату при вирішенні задач, пов’язаних із знаходженням оцінок регресійних параметрів у випадку негаусового розподілу їх помилок.

1.5.2 Особливості застосування стохастичних поліномів та моментно-кумулянтного опису

Практичне використання стохастичних поліномів Кунченка у поєднанні із частковим описом статистиками вищих порядків для вирішення задачі знаходження адаптивних оцінок векторних параметрів має певні особливості.

Зокрема, при використанні як базових функції стохастичних поліномів степеневих перетворень результуючі аналітичні вирази містять залежність від математичних сподівань таких перетворень, тобто моментів (моментних функцій) відповідного порядку. Зокрема, при застосуванні ММПл степені S для оцінювання параметрів виникає необхідність наявності часткового опису ймовірнісної складової моделей у вигляді моментів m_i до $2S$-го порядку. Проте з точки зору аналізу ефективності поліноміальних методів як показано в роботах [59, 60] є більш зручним і конструктивним використання саме кумулянтного опису. Це пояснюється тим, що кумулянти, на відміну від моментів, мають чітко виражений самостійний статистичний зміст і можуть бути задані до певної міри незалежно один від одного, виконуючи при цьому роль деяких «нормальних координат» статистичного опису. Друга перевага пов'язана з тим, що врахування вищих порядків кумулянтів дає можливість просто описати будь-яку ступінь негаусовості випадкових величин [46].

Як вже відзначалося вище, відсутність інформації про апріорні значення таких параметрів може бути достатньо просто компенсована на основі апостеріорних оцінок цих параметрів. Зокрема, оцінки початкових моментів достатньо просто отримати, використовуючи усереднені статистики виду

$$\widehat{m}_i = \frac{1}{N}\sum_{v=1}^{N} x_v^i,\ i = \overrightarrow{1,2S}. \tag{1.12}$$

І хоча подібний спосіб подолання апріорної невизначеності з алгоритмічної точки зору є дуже простим, проте існує проблема того, що з ростом порядку моментів i суттєво зростає дисперсія оцінок відповідних параметрів [87, 88]. Таке зростання в свою чергу збільшує ступінь розузгодженості між істинними і оціненими значеннями параметрів моделі. Мінімізація ступеня такої невизначеності може бути досягнута лише за рахунок збільшення обсягу навчальних даних, необхідних для забезпечення адаптивності. Проте в реальних задачах це часто не є можливим. Саме зазначений фактор обмежує використання в даному дослідження поліномів лише 2-го та 3-го (для симетрично-розподілених випадкових величин) порядку.

Оскільки потенційна ефективність стохастичних поліномів проявляється лише при суттєво негаусовому характері статистичних даних, то окремо виникає задача, пов'язана із перевіркою статистичної гіпотези про нормалізацію. Крім того, як буде показано далі, при виборі степені стохастичного полінома додатково виникає необхідність перевірки гіпотези про симетрію розподілів. Це відкриває ще один потенційний аспект використання моментно-кумулянтного опису при реалізації адаптивного оцінювання параметрів регресії із застосуванням методу максимізації поліномів (див. підрозділ 2.6). Зокрема, серед множини статистичних критеріїв [89], що можуть бути застосовані для вирішення подібних задач відзначимо тест Харке-Бера [90], який базується на статистиці виду

$$JB = \frac{N}{6}\left(\hat{\gamma}_3^2 + \frac{1}{4}\hat{\gamma}_4^2\right), \tag{1.13}$$

де оцінки кумулянтних коефіцієнтів можна легко отримати із (1.12) [46].

Реалізація тесту Харке-Бера, який широко застосовується в задачах аналізу регресійних залишків, імплементовано як базову функцію в пакеті MATLAB, а також він присутній в бібліотеці «moments» для мови R [91].

Висновки

Проведений огляд інформаційних джерел за тематикою дисертаційного дослідження показує, що не зважаючи на більш ніж двохсотлітню історію регресійного аналізу задача підвищення точності оцінювання параметрів на основі урахування специфіки регресійних моделей залишається актуальною. На сучасному етапі розвитку науково-технічного прогресу вирішення подібних задач неможливо уявити без використання комп’ютерних технологій та спеціалізованих програмних засобів основу яких, складають відповідні математичні моделі та обчислювальні методи.

Результати аналізу свідчать, що використання класичного методу найменших квадратів забезпечує отримання оптимальних оцінок лише при виконанні ряду теоретичних припущень, зокрема при гаусовому (нормальному) розподілі помилок. В ситуаціях коли ефект нормалізації відсутній можуть бути застосовані робастні або адаптивні підходи. Перший підхід орієнтований на забезпеченні стійкості результуючих оцінок щодо відхилень від теоретичних припущень про властивості моделей. Метою другого підходу, який використовується у даному дослідженні, є забезпечення ефективності (мінімізації дисперсії) оцінок регресійних параметрів.

Практичне використання параметричних методів адаптивного оцінювання, ускладнюється суттєвими вимогами до наявності апріорної інформації про тип та значення параметрів розподілу регресійних помилок. Одним із компромісних підходів до вирішення задач, пов’язаних із опрацюванням статистичних даних із відмінним від гаусової ідеалізації характером, є напрямок, який базується на застосуванні апарату стохастичних поліномів Кунченка. Сукупність наявних результатів по застосуванню цього апарату для вирішення різноманітних задач статистичної радіотехніки: виявлення, розпізнавання сигналів та оцінювання їх параметрів свідчить, що використання нелінійних перетворень, надає додаткові важелі для підвищення точності порівняно із лінійними методами, які оптимізовані для гаусових моделей. Крім того, застосування апарату стохастичних поліномів базується на використанні спрощеного опису у вигляді статистик вищих порядків, що суттєво спрощує (порівняно з параметричним підходом) процес синтезу адаптивних обчислювальних алгоритмів, здатних враховувати специфічні властивості реальних даних.

РОЗДІЛ 2
СИНТЕЗ МОДЕЛЕЙ ТА МЕТОДІВ ПОЛІНОМІАЛЬНОГО ОЦІНЮВАННЯ ПАРАМЕТРІВ РЕГРЕСІЇ

2.1 Загальна постановка задачі

Нехай є регресійна модель спостережень у вигляді суми детермінованої і випадкової складової

$$y_v = R_v(\boldsymbol{\theta}, \boldsymbol{X}) + \xi_v, v = \overrightarrow{1, N}, \tag{2.1}$$

що описує залежність значень цільової змінної $\boldsymbol{Y} = \{y_1, y_2, \dots, y_N\}$ від в загальному випадку множини $\boldsymbol{X}$ незалежних змінних (P-факторного регресора), сукупність яких можна представити у вигляді матриці розмірності PхN виду

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,N} \\ x_{2,1} & x_{2,2} & \dots & x_{2,N} \\ \dots & \dots & \dots & \\ x_{P,1} & x_{P,2} & \dots & x_{P,N} \end{pmatrix}.$$

Детермінована складова $R_v(\boldsymbol{\theta}, \boldsymbol{X})$ містить Q-мірний вектор інформативних параметрів $\boldsymbol{\theta} = \{a_0, a_1, \dots a_{Q-1}\}$.

Відліки випадкової складової (помилки) регресійної моделі ξ_v представляють собою послідовність незалежних і однаково-розподілених випадкових величин, що володіють властивістю центрованості ($E\{\xi\} = 0$) та гомоскедактичності $D\{\xi\} = const$. Імовірнісні властивості помилки регресійної моделі суттєво відрізняються від гаусового (нормального) закону.

Загальна задача полягає в знаходженні оцінок складових вектору інформативного параметру $\boldsymbol{\theta}$ на основі статистичного аналізу статистичних даних $\boldsymbol{Y}$та $\boldsymbol{X}$.

2.2 Застосування методу максимізації полінома при знаходженні оцінок векторного параметру при неоднаково розподілених даних

Для вирішення поставленої задачі регресійного аналізу можна скористатися модифікацією ММПл для статистичного оцінювання векторного параметру при неоднаково розподілених даних [58]. В його основі лежить властивість максимізації функціонала у вигляді стохастичного полінома загального виду

$$L_{SN} = \sum_{v=1}^{N} \sum_{i=1}^{S} \phi_i(y_v) \int^{a} k_{iv}(a)dz - \sum_{i=1}^{S} \sum_{v=1}^{N} \int^{a} \Psi_{iv} k_{iv}(a)dz, \qquad (2.2)$$

Припускається, що випадкові величини y_v описуються послідовністю математичних сподівань

$$\Psi_{iv} = E\{\phi_i(y_v)\},\ i = \overrightarrow{1,S},\ v = \overrightarrow{1,N}, \qquad (2.3)$$

і є двічі диференційовані по параметру a, що оцінюється.

Стохастичний поліном L_{SN} виду (2.2) володіє двома основними властивостями [58]:

1) для будь-якого порядку S при асимптотичному зростанні $v \to \infty$ обсягу вибірки $\boldsymbol{Y}$ поліном L_{SN} як функція параметра a приймає максимум в околиці істинного значення цього параметра;
2) при різних вибірках $\boldsymbol{Y}$ відхилення максимуму полінома L_{SN} від істинного значення параметра a буде мати мінімальну дисперсію (для відповідної порядку полінома S).

За аналогією до методу максимальної правдоподібності оцінку параметра a можна знаходити із вирішення рівняння виду

$$\frac{d}{da} L_{SN}\Big|_{a=\hat{a}} = \sum_{i=1}^{S} \sum_{v=0}^{N} k_{iv}[\phi_i(y_v) - \Psi_{iv}]\Big|_{a=\hat{a}} = 0. \qquad (2.4)$$

Оптимальні коефіцієнти k_{iv}, що максимізують функціонал (2.2) знаходяться з розв'язання системи лінійних алгебраїчних рівнянь

$$\sum_{j=1}^{S} k_{jv} F_{(i,j)v} = \frac{d}{da} \Psi_{iv},\ i = \overrightarrow{1,S},\ v = \overrightarrow{1,N}, \tag{2.5}$$

де $F_{(i,j)v} = \Psi_{(i,j)v} - \Psi_{iv}\Psi_{jv},\ \Psi_{(i,j)v} = E\{\phi_i(y_v)\phi_j(y_v)\},\ i,j = \overrightarrow{1,S}$.

Даний підхід до оцінювання може бути поширений на випадок знаходження оцінок векторного параметра $\boldsymbol{\theta} = \{a_0, a_1, \dots a_{Q-1}\}$. В такій ситуації необхідно використовувати Q поліномів $L_{SN}^{(p)},\ p = \overrightarrow{0, Q-1}$ загального виду (2.2) для кожної компоненти a_p векторного параметра.

Кожний p-ий стохастичний поліном $L_{SN}^{(p)}$ як функція параметра a_p при відомих значеннях інших складових вектору $\boldsymbol{\theta}$ при $n \to \infty$ також має максимум в околиці істинного значення параметра a_p. Таким чином, шукані оцінки параметра можуть бути знайдені як розв'язок системи рівнянь виду

$$f_{SN}^{(p)}(y_v, x_v) = \sum_{i=1}^{S} \sum_{v=1}^{N} k_{iv}^{(p)} [\phi_i(y_v) - \Psi_{iv}]\Big|_{a_p = \hat{a}_p} = 0,\ p = \overrightarrow{0, Q-1}. \tag{2.6}$$

При використанні поліномів степені $S \geq 2$ знаходження ММПЛ-оцінок векторного параметру у переважній більшості випадків (аналогічно до ситуації із використанням ММП) потребує застосування чисельних методів розв’язку систем нелінійних рівнянь. Зокрема. часто використовується підхід, який базується на ітераційній чисельній процедурі Ньютона-Рафсона [58]. В його основі лежить принцип лінеаризації шляхом розкладу лівої частини кожного нелінійного рівняння системи (2.6) в ряд Тейлора в околі істинного значення вектору $\boldsymbol{\theta}$. Обмежившись першими двома членами ряду можна записати у матричній формі лінійну систему

$$\widehat{\boldsymbol{\theta}}_{(S)}^{(k+1)} = \widehat{\boldsymbol{\theta}}_{(S)}^{(k)} - \left[\boldsymbol{Z}_S(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})\right]^{-1}\boldsymbol{F}_S(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)}), \tag{2.7}$$

яка може бути використана для ітераційного пошуку оціночних значень. Для отримання системи (8) необхідно обчислити матрицю-стовпець $\boldsymbol{F}_S(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$, що складена з елементів лівої частини кожного нелінійного рівняння системи (2.6) та квадратну матрицю $\boldsymbol{Z}_S(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)}) = \boldsymbol{H}_S(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)}) + \boldsymbol{J}_S(\widehat{\boldsymbol{\theta}}^{(k)})$ із складовими

$$H_{SN}^{(p,q)} = \sum_{i=1}^{S} \frac{\partial}{\partial a_q} k_{iv}^{(p)} \left[\sum_{v=1}^{N} \phi_i(y_v) - N\Psi_{iv}\right] \tag{2.8}$$

і елементами матриці кількості добутої інформації$\boldsymbol{J}_S(\widehat{\boldsymbol{\theta}}^{(k)})$ виду

$$J_{SN}^{(p,q)} = \sum_{v=1}^{N} \sum_{i=1}^{S} k_{i,v}^{(p)} \frac{\partial}{\partial a_q} \Psi_{iv}, p, q = \overrightarrow{0, Q-1}. \tag{2.9}$$

Для старту ітераційної процедури припускається наявність деякого початкового наближення $\boldsymbol{\theta}^{(1)} = \left\{a_0^{(1)}, a_1^{(1)}, \dots, a_{Q-1}^{(1)}\right\}^T$ в якості якого можуть бути обрані «грубі» оцінки, знайдені більш простим оцінювачем або лінійні ММПЛ-оцінки при степені $S = 1$ [58].

2.3 Знаходження ММПл-оцінок параметрів лінійної багатофакторної регресії

Лінійна багатофакторна (множинна) залежність, що описує взаємозв'язок між цільовою змінною і адитивною композицією множини регресорів з певними ваговими коефіцієнтами, є найбільш простим, але при цьому найпоширенішим типом регресійних моделей. Крім того, багато нелінійних залежностей (степеневих, експоненціальних, логарифмічних та ін.) шляхом відповідних перетворень також можуть бути зведені до лінійної регресійної моделі. Наслідком цього є поширене використання лінійної регресії в технічних, екологічних, медичних економічних, і інших додатках.

Як вже зазначалося вище, найбільш поширеним способом знаходження оцінок параметрів лінійної регресії є використання звичайного методу найменших квадратів, застосування якого зводиться до вирішення системи лінійних рівнянь відносно інформативних параметрів, що оцінюються.

2.3.1. Постановка задачі оцінювання параметрів лінійної регресії

Модифікуємо загальну постановку задачі із підрозділу 2.1 для випадку лінійної регресійної моделі. У цьому разі регресійна модель спостережень описує залежність значень цільової змінної $\boldsymbol{Y} = \{y_1, y_2, \dots, y_N\}$ у вигляді лінійної комбінації між собою незалежних змінних

$$y_v = a_0 + \sum_{p=1}^{Q-1} a_p x_{p,v} + \xi_v = \sum_{p=0}^{Q-1} a_p x_{p,v} + \xi_v,\ v = \overrightarrow{1,N}. \quad (2.10)$$

Для уніфікації запису матрицю багатофакторного регресору $\boldsymbol{X}$ представимо у вигляді

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,2} & \dots & x_{1,N} \\ 1 & x_{2,2} & \dots & x_{2,N} \\ \dots & \dots & \dots & \\ 1 & x_{Q-1,2} & \dots & x_{Q-1,N} \end{pmatrix},$$

ввівши колонку із елементів $x_{0,v} = 1$ як формальних регресорів для постійного коефіцієнта a_0, значення якого також підлягає оцінюванню. Таким чином, реальна кількість векторів регресорів є на одиницю меншою за кількість елементів вектору $\boldsymbol{\theta} = \{a_0, a_1, \dots a_{Q-1}\}$, що містить інформативні параметрів (коефіцієнти степеневого полінома) детермінованої складової лінійної регресійної моделі (2.10).

Крім того, введемо додаткову вимогу, що випадкові помилки ξ_v регресійної моделі можуть бути описані за допомогою послідовності центральних моментів μ_r або кумулянтів κ_r (кумулянтних коефіцієнтів γ_r). При такому описі кумулянт другого порядку κ_2 співпадає із дисперсію μ_2 випадкової складової, а кумулянтні коефіцієнти вищих порядків $\gamma_r = \frac{\kappa_r}{\kappa_2^{\frac{r}{2}}}$ чисельно описують ступінь відмінності від гаусового розподілу.

2.3.2. Застосування методу максимізації полінома при степені *S*=1 для оцінювання параметрів лінійної регресії

Для вирішення поставленої задачі із застосуванням методу максимізації полінома при формування стохастичного полінома (2.2) як базисні функції використаємо степеневі перетворення вибіркових значень $\boldsymbol{Y}$, тобто

$$\phi_i(y_v) = y_v^i. \quad (2.11)$$

У цьому випадку послідовність математичних сподівань (2.3) є сукупністю початкових моментів відповідного порядку

$$\Psi_{iv} = E\{y_v^i\} = \alpha_{iv},\ \ i = \overrightarrow{1,S},\ v = \overrightarrow{1,N}, \quad (2.12)$$

а функції $F_{(i,j)v} = \alpha_{(i+j)v} - \alpha_{iv}\alpha_{jv}$ називають центрованими корелянтами [58].

При використанні стохастичного полінома порядку $S = 1$ ММПл-оцінки елементів векторного параметра $\boldsymbol{\theta}$ детермінованої складової регресійної моделі (2.10) можуть бути знайдені з рішення системи Q рівнянь виду:

$$\sum_{v=1}^{N} \left\{ k_{1,v}^{(p)} \left[y_v - \sum_{p=0}^{Q-1} a_p x_{p,v} \right] \right\} = 0,\ \ p = \overrightarrow{0,Q-1}, \quad (2.13)$$

де оптимальні коефіцієнти $k_{1,v}^{(p)}$ знаходяться як розв'язання системи лінійних алгебраїчних рівнянь (2.5), яка з урахуванням (2.12) трансформується в систему

$$\sum_{j=1}^{S} k_{jv} F_{(i,j)v} = \frac{d}{da_p} \alpha_{iv},\ i = \overrightarrow{1,S},\ v = \overrightarrow{1,N},\ p = \overrightarrow{0,Q-1} \qquad (2.14)$$

і які можна представити у вигляді

$$k_{1,v}^{(p)} = \frac{\partial}{\partial a_p} \left[\sum_{p=0}^{Q-1} a_p x_{p,v} \right] = \frac{x_{p,v}}{\mu_2},\ p = \overrightarrow{0,Q-1}.$$

Після певних трансформацій (2.13) можна представити у матричному вигляді як систему лінійних рівнянь

$$(\boldsymbol{X}^T \boldsymbol{X}) \boldsymbol{\theta}^T = \frac{1}{N} \sum_{v=1}^{N} \boldsymbol{X}^T \boldsymbol{Y}^T, \qquad (2.15, a)$$

звідки ММПл-оцінками шуканого параметра є

$$\widehat{\boldsymbol{\theta}}_{(1)} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \frac{1}{N} \sum_{v=1}^{N} \boldsymbol{X}^T \boldsymbol{Y}^T. \qquad (2.15, б)$$

Відзначимо, що система (2.15) для знаходження ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(1)}$ параметрів лінійної регресії (2.10) при *S=1* є тотожною відповідній лінійній системі МНК-оцінок (1.3).

Такі оцінки є оптимальними (за критерієм мінімуму дисперсії) лише для ситуації, коли помилки регресійній моделі мають гаусовий розподіл. У випадку, коли ймовірнісний характер помилок відрізняється від гаусового закону існують альтернативні методи оцінювання. Тому нижче наведемо новий підхід до нелінійного оцінювання параметрів лінійної регресії (2.10), який базується на використанні степеневих стохастичних поліномів більш високого порядку.

2.3.3. Застосування методу максимізації полінома при степені *S*=2 для оцінювання параметрів лінійної регресії

При застосуванні методу максимізації полінома при степені S=2 для опису випадкової складової регресійних моделей загального виду (2.1) необхідні співвідношення для перших 4-х початкових моментів, які можна представити у вигляді:

$$\alpha_{1v} = R_v(\boldsymbol{\theta}, \boldsymbol{X}),\ \alpha_{2v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^2 + \mu_2,$$
$$\alpha_{3v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^3 + 3R_v(\boldsymbol{\theta}, \boldsymbol{X})\mu_2 + 3\mu_2^2 + \mu_3,$$
$$\alpha_{4v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^4 + 6[R_v(\boldsymbol{\theta}, \boldsymbol{X})]^2\mu_2 + 4R_v(\boldsymbol{\theta}, \boldsymbol{X})\mu_3 + \mu_4,$$

Використовуючи загальний вираз (2.6) із урахуванням виразу (2.11), сформуємо систему із Q рівнянь для пошуку ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(2)}$ детермінованої складової регресійної моделі (2.10) у вигляді:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}\left[y_v - \sum_{p=0}^{Q-1} a_p x_{p,v}\right] + k_{2,v}^{(p)}\left[(y_v)^2 - \left[\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right)^2 + \mu_2\right]\right]\right\} = 0,\ p = \overrightarrow{0, Q-1}, \quad (2.16)$$

де оптимальні коефіцієнти $k_{i,v}^{(p)}, i = \overrightarrow{1,2}$ забезпечують мінімізацію дисперсії оцінок компонентів шуканого параметра при використанні ступеня полінома $S = 2$. Ці коефіцієнти знаходяться як розв'язок відповідної системи виду (2.14) і можуть бути представлені як функції, що залежать від центральних моментів випадкової складової регресійної моделі:

$$k_{1,v}^{(p)} = \frac{\mu_4 - \mu_2^2 + 2\mu_3\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right)}{\mu_2(\mu_4 - \mu_2^2) - \mu_3^2} x_{p,v},\ \ k_{2,v}^{(p)} = -\frac{\mu_3^2}{\mu_2(\mu_4 - \mu_2^2) - \mu_3^2} x_{p,v}. \quad (2.17)$$

Підставивши коефіцієнти (2.17) у (2.16), після певних перетворень систему рівнянь для знаходження оцінок можна записати у вигляді:

$$\sum_{v=1}^{N}\left\{x_{p,v}\left[A_2\left(\sum_{p=0}^{Q-1}a_p x_{p,v}\right)^2+B_{2,v}\left(\sum_{p=0}^{Q-1}a_p x_{p,v}\right)+C_{2,v}\right]\right\}=0,\ p=\overrightarrow{0,Q-1},\tag{2.18}$$

де

$$A_2=\mu_3,\ B_{2,v}=\mu_4-\mu_2^2-2y_v\mu_3,\ C_{2,v}=y_v^2\mu_3-y_v(\mu_4-\mu_2^2)-\mu_2\mu_3.\tag{2.19}$$

Очевидно, що при степені полінома $S=2$ ММПЛ-оцінки можуть бути знайдені лише чисельно, наприклад, із застосуванням ітераційної процедури Ньютона-Рафсона. Використовуючи вирази для оптимальних коефіцієнтів (2.17), необхідні для формування відповідної системи рівнянь, елементи матриці $\boldsymbol{Z}_2(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$ можна представити у вигляді

$$Z_{2N}^{(p,q)}=\sum_{v=1}^{N}\sum_{v=1}^{N}\left\{x_{p,v}x_{q,v}\left[2A_2\left(\sum_{p=0}^{Q-1}a_p x_{p,v}\right)+B_{2,v}\right]\right\}.\tag{2.20}$$

Як вже зазначалося вище, в якості початкового наближення можуть бути використані оцінки, знайдені із вирішення лінійної системи виду (2.15), формування якої не потребує додаткової інформації про властивості помилок регресійної моделі.

2.3.4. Застосування методу максимізації полінома при степені S=3 (симетричний випадок) для оцінювання параметрів лінійної регресії

Аналіз виразів (2.17 – 2.19) показує, що в разі симетрії розподілу (або хоча б тотожності $\mu_3=0$) випадкової складової помилок регресійної моделі застосування ММПл при $S=2$ є недоцільним, оскільки система (2.18) вироджується в систему лінійних рівнянь виду (2.13), еквіваленту системі для звичайного МНК. Але навіть у випадку симетрії (за умови суттєвої відмінності від гаусового розподілу) можна застосовувати стохастичний поліном більш високого ступеня $S=3$.

При застосуванні методу максимізації полінома при степені S=3 для опису симетрично розподіленої випадкової складової регресійних моделей загального виду (2.1) необхідні співвідношення для перших 6-х початкових моментів, які можна представити у вигляді:

$$\alpha_{1v} = R_v(\boldsymbol{\theta}, \boldsymbol{X}),\ \alpha_{2v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^2 + \mu_2,\ \alpha_{3v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^3 + 3R_v(\boldsymbol{\theta}, \boldsymbol{X})\mu_2,$$
$$\alpha_{4v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^4 + 6[R_v(\boldsymbol{\theta}, \boldsymbol{X})]^2\mu_2 + 3\mu_2^2 + \mu_4,$$
$$\alpha_{5v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^5 + 10[R_v(\boldsymbol{\theta}, \boldsymbol{X})]^3\mu_2 + 5(3\mu_2^2 + \mu_4)R_v(\boldsymbol{\theta}, \boldsymbol{X}),$$
$$\alpha_{6v} = [R_v(\boldsymbol{\theta}, \boldsymbol{X})]^6 + 15[R_v(\boldsymbol{\theta}, \boldsymbol{X})]^4\mu_2 + 45[R_v(\boldsymbol{\theta}, \boldsymbol{X})]^2\mu_2^2 + 15\mu_2^3 +$$
$$+ 15[R_v(\boldsymbol{\theta}, \boldsymbol{X})]^2\mu_4 + 15\mu_2\mu_4 + \mu_6$$

Використовуючи загальний вираз (2.6) із урахуванням виразу (2.11), сформуємо систему із Q рівнянь для пошуку ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(3)}$ детермінованої складової регресійної моделі (2.10) у вигляді:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}\left[y_v - \sum_{p=0}^{Q-1} a_p x_{p,v}\right] + k_{2,v}^{(p)}\left[y_v^2 - \left(\left[\sum_{p=0}^{Q-1} a_p x_{p,v}\right]^2 + \mu_2\right)\right] +\right.$$
$$\left. + k_{3,v}^{(p)}\left[y_v^3 - \left(\left[\sum_{p=0}^{Q-1} a_p x_{p,v}\right]^3 + 3\mu_2 \sum_{p=0}^{Q-1} a_p x_{p,v}\right)\right]\right\} = 0, \quad p = \overline{0, Q-1}, \quad (2.21)$$

де оптимальні коефіцієнтів $k_{i,v}^{(p)}$, $i = \overline{1,3}$, що забезпечують мінімізацію дисперсії оцінок компонентів шуканого параметра можуть бути представлені у вигляді:

$$k_{1,v}^{(p)} = \frac{3\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right)^2(\mu_4 - 3\mu_2^2) + 3\mu_4\mu_2 - \mu_6}{\mu_2^{-2}(\mu_4^2 - \mu_2\mu_6)} x_{p,v},$$
$$k_{2,v}^{(p)} = \frac{-3(\mu_4 - 3\mu_2)}{\mu_2^{-2}(\mu_4^2 - \mu_2\mu_6)}\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right) x_{p,v},$$
$$k_{3,v}^{(p)} = \frac{\mu_4 - 3\mu_2}{\mu_2^{-2}(\mu_4^2 - \mu_2\mu_6)} x_{p,v},\ p = \overline{0, Q-1}. \quad (2.22)$$

Підставляючи коефіцієнти (2.22) в (2.21), після певних перетворень систему рівнянь для знаходження оцінок можна записати у вигляді:

$$\sum_{v=1}^{N}\left\{x_{p,v}\left[A_3\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right)^3 + B_{3,v}\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right)^2 + C_{3,v}\left(\sum_{p=0}^{Q-1} a_p x_{p,v}\right) + D_{3,v}\right]\right\} = 0,\ p = \overrightarrow{0, Q-1}, \quad (2.23)$$

де

$$A_3 = 1,\ B_{3,v} = -3y_v,\ C_{3,v} = 3y_v^2 - \frac{\mu_6 - 3\mu_4\mu_2}{\mu_4 - 3\mu_2^2},\ D_{3,v} = y_v \frac{\mu_6 - 3\mu_4\mu_2}{\mu_4 - 3\mu_2^2} - y_v^3. \quad (2.24)$$

Очевидно, що як і у попередньому випадку, при степені полінома $S = 3$ ММПл-оцінки можуть бути знайдені лише за допомогою чисельного рішення систем рівнянь (2.23). При застосуванням ітераційної процедури Ньютона-Рафсона необхідні елементи матриці $\boldsymbol{Z}_3(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$ по аналогії до (2.20) можна представити у вигляді

$$Z_{3N}^{(p,q)} = \sum_{v=1}^{N}\sum_{v=1}^{N}\left\{x_v^{p+g}\left[3A_3\left(\sum_{p=0}^{Q-1} a_p x_v^p\right)^2 + 2B_{3,v}\left(\sum_{p=0}^{Q-1} a_p x_v^p\right) + C_{3,v}\right]\right\}.$$

Детальний аналіз виразів (2.24) також показує, що в разі коли розподіл випадкової складової помилок регресійної моделі відповідає нормальному закону, то застосування ММПл при $S = 3$ стає недоцільним, оскільки система (2.23) вироджується в систему лінійних рівнянь виду (2.13), еквіваленту системі для звичайного МНК. Ц підтверджується також результатами аналізу ефективності ММПл-оцінок при $S = 3$, які наведені у підрозділі 3.2.3.

2.4 Знаходження ММПл-оцінок параметрів поліноміальної регресії

Оцінювання параметрів поліноміальної регресії є однією із самих ранніх задач статистичного аналізу [1]. За більш ніж два століття модель поліноміальної регресії набула широкого поширення при вирішенні різноманітних прикладних задач технічного, геофізичного, біомедичного, економічного та інших аспектів людської діяльності.

Як і у випадку із лінійною багатофакторною регресією у переважній більшості для знаходження оцінок параметрів поліноміальної регресії застосовують варіації методу найменших квадратів. Це пов’язано із тим фактом, що поліноміальну регресію, яка надає можливість моделювати нелінійні залежності, формально із математичної точки зору можна трактувати як окремий різновид множинної лінійної регресії. Проте побудова та інтерпретація моделі поліноміальної регресії має свої особливості. Це обумовлено тим, що на відміну від лінійної багатофакторної регресії, де окремі змінні є незалежними між собою, поліноміальні регресори є сильно корельованими між собою. Це ускладнює інтерпретацію оцінок коефіцієнтів при підборі поліноміальної регресії.

Інші особливості поліноміальної регресійної моделі полягають у наступному:

- застосування таких моделей дозволяє адекватно описувати нелінійно-розподілені дані із складними взаємозв'язками;
- за рахунок можливості вибору порядку полінома забезпечується гнучкість в плані контролю над процесами моделюванням змінних об'єкта.

Важливо фактором також є те, що при завищенні порядку апроксимаційного полінома, результуюча модель може бути «перенавчена» на тренувальних даних і показувати гіршу прогнозну властивість на тестових. Тому на практиці зазвичай застосовують поліноми невисокого (не більше 3-4-го) порядку [92].

2.4.1. Постановка задачі оцінювання параметрів поліноміальної регресії

Модифікуємо постановку задачі із підрозділу 2.3.1 для випадку поліноміальної однофакторної регресійної моделі. У цьому разі регресійна модель спостережень описує залежність значень цільової змінної $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_N\}$ у вигляді лінійної комбінації деяких функціональних перетворень від змінних вектору $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\}$ виду

$$y_v = \sum_{p=0}^{Q-1} a_p \varphi_p(x_v) + \xi_v, v = \overrightarrow{1, N}. \qquad (2.25)$$

Як базис детермінованої складової (2.25) теоретично можуть використовуватись різні функціональні залежності, але найбільш поширеним є саме степеневі функції

$$\varphi_p(x) = x^p, p = \overrightarrow{0, Q-1}. \qquad (2.26)$$

Необхідно також зазначити, що поняття «степеневі стохастичні поліноми» як і присутність терміну «максимізація поліномів» у назві методу оцінювання, який використовується у даній роботі, ніяким чином не пов'язані із поліноміальним типом детермінованої складової регресійної моделі. Це абсолютно різні типи поліномів з точки зору їх функціонального призначенням та практичного застосування.

2.4.2. Застосування методу максимізації полінома при степені *S*=1 для оцінювання параметрів поліноміальної регресії

Використовуючи аналогію між багатофакторною лінійною і поліноміальною однофакторною регресійними моделями можна записати систему для пошуку ММПл-оцінки (при степені полінома $S = 1$) елементів векторного параметра $\boldsymbol{\theta}$ детермінованої складової регресійної моделі (2.25) у наступному вигляді

$$\sum_{v=1}^{N} \left\{ k_{1,v}^{(p)} \left[y_v - \sum_{p=0}^{Q-1} a_p \varphi_p(x_v) \right] \right\} = 0, \qquad p = \overrightarrow{0,Q-1}, \qquad (2.27)$$

де оптимальний коефіцієнт

$$k_{1,v}^{(p)} = \frac{\varphi_p(x_v)}{\mu_2}, p = \overrightarrow{0,Q-1}.$$

Якщо використати як базисні функції поліноміальної регресійної моделі (2.25) степеневі перетворення (2.26), то після певних трансформацій (2.27) можна записати у матричному вигляді:

$$(\boldsymbol{B}^T\boldsymbol{B})\boldsymbol{\theta}^T = \frac{1}{N}\sum_{v=1}^{N} \boldsymbol{B}^T\boldsymbol{Y}^T, \tag{2.28}$$

де $\boldsymbol{B}$ – відома матриця Вандермонда [93] з усередненими елементами $b_{v,p} = x_v^p$ виду

$$\boldsymbol{B} = \begin{pmatrix} 1 & \overline{x} & \overline{x^2} & \dots & \overline{x^{Q-1}} \\ \overline{x} & \overline{x^2} & \overline{x^3} & \dots & \overline{x^Q} \\ \overline{x^2} & \overline{x^3} & \overline{x^4} & \dots & \overline{x^{Q+1}} \\ \dots & \dots & \dots & \dots & \dots \\ \overline{x^{Q-1}} & \overline{x^Q} & \overline{x^{Q+1}} & \dots & \overline{x^{2(Q-)1}} \end{pmatrix}, \tag{2.29}$$

Таким чином стає очевидним, що система (2.27) як і система (2.13) для знаходження ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(1)}$ параметрів лінійної регресії при *S=1* є еквівалентною лінійній системі МНК-оцінок. Тому нижче наведемо результати застосування нового підходу до оцінювання параметрів поліноміальної регресії (2.25), який базується на використанні степеневих стохастичних поліномів більш високого порядку.

2.4.3. Застосування методу максимізації полінома при степені *S*=2 для оцінювання параметрів поліноміальної регресії

Використовуючи вже зазначену вище аналогію між багатофакторною лінійною і поліноміальною однофакторною регресійними моделями модифікуємо систему рівнянь для пошуку ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(2)}$ детермінованої складової регресійної моделі (2.25) і представимо її у вигляді:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}\left[y_v-\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right]+k_{2,v}^{(p)}\left[(y_v)^2-\left[\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)^2+\mu_2\right]\right]\right\}=0,$$

$$p=\overrightarrow{0,Q-1}, \quad (2.29)$$

де оптимальні коефіцієнти $k_{i,v}^{(p)}$, $i=\overrightarrow{1,2}$ можуть бути записані наступним чином:

$$k_{1,v}^{(p)}=\frac{\mu_4-\mu_2^2+2\mu_3\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)}{\mu_2(\mu_4-\mu_2^2)-\mu_3^2}\varphi_p(x_v),$$

$$k_{2,v}^{(p)}=-\frac{\mu_3^2}{\mu_2(\mu_4-\mu_2^2)-\mu_3^2}\varphi_p(x_v). \quad (2.30)$$

Підставивши коефіцієнти (2.30) у (2.29), після певних перетворень систему рівнянь для знаходження оцінок можна записати у вигляді:

$$\sum_{v=1}^{N}\left\{\varphi_p(x_v)\left[A_2\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)^2+B_{2,v}\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)+C_{2,v}\right]\right\}=0,\ p=\overrightarrow{0,Q-1}, \quad (2.31)$$

де позначення A_2, $B_{2,v}$, $C_{2,v}$ такі ж як і в (2.19).

Як і раніше подібні оцінки можуть бути знайдені лише чисельно. Використовуючи вирази для оптимальних коефіцієнтів (2.30), необхідні для формування відповідної системи рівнянь, елементи матриці $\boldsymbol{Z}_2(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$ можна представити у вигляді

$$Z_{2N}^{(p,q)}=\sum_{v=1}^{N}\sum_{v=1}^{N}\left\{\varphi_p(x_v)\varphi_q(x_v)\left[2A_2\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)+B_{2,v}\right]\right\}.$$

Як початкові наближення можуть бути використані оцінки, знайдені із вирішення лінійної системи виду (2.28), формування якої не потребує додаткової інформації про властивості помилок регресійної моделі.

2.4.4. Застосування методу максимізації полінома при степені *S*=3 (симетричний випадок) для оцінювання параметрів поліноміальної регресії

В разі симетрії розподілу (або хоча б тотожності $\mu_3 = 0$) випадкової складової помилок поліноміальної регресійної моделі (як і випадку із лінійною регресією) застосування ММПл при $S = 2$ також є недоцільним, оскільки результуюча система рівнянь стає еквівалентною МНК.

При застосуванні методу максимізації полінома при степені *S*=3 у випадку симетрично розподіленої випадкової складової регресійних моделей систему із Q рівнянь для пошуку ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(3)}$ подібно до (2.2.1) можна представити у вигляді:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}\left[y_v-\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right]+k_{2,v}^{(p)}\left[y_v^2-\left(\left[\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right]^2+\mu_2\right)\right]+\right.$$
$$\left.+k_{3,v}^{(p)}\left[y_v^3-\left(\left[\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right]^3+3\mu_2\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)\right]\right\}=0,\quad p=\overrightarrow{0,Q-1}, \tag{2.32}$$

де оптимальні коефіцієнтів $k_{i,v}^{(p)}$, $i=\overrightarrow{1,3}$, що забезпечують мінімізацію дисперсії оцінок

$$k_{1,v}^{(p)}=\frac{3\left(\sum_{p=0}^{Q-1}a_px_{p,v}\right)^2(\mu_4-3\mu_2^2)+3\mu_4\mu_2-\mu_6}{\mu_2^{-2}(\mu_4^2-\mu_2\mu_6)}\varphi_p(x_v),$$
$$k_{2,v}^{(p)}=\frac{-3(\mu_4-3\mu_2)}{\mu_2^{-2}(\mu_4^2-\mu_2\mu_6)}\left(\sum_{p=0}^{Q-1}a_px_{p,v}\right)\varphi_p(x_v),$$
$$k_{3,v}^{(p)}=\frac{\mu_4-3\mu_2}{\mu_2^{-2}(\mu_4^2-\mu_2\mu_6)}\varphi_p(x_v),\ p=\overrightarrow{0,Q-1}. \tag{2.33}$$

Підставляючи коефіцієнти (2.33) в (2.32), після певних перетворень систему рівнянь для знаходження оцінок можна записати у вигляді:

$$\sum_{v=1}^{N}\left\{\varphi_p(x_v)\left[A_3\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)^3+B_{3,v}\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)^2+\right.\right.$$
$$\left.\left.C_{3,v}\left(\sum_{p=0}^{Q-1}a_p\varphi_p(x_v)\right)+D_{3,v}\right]\right\}=0,\ p=\overrightarrow{0,Q-1}, \quad (2.34)$$

де позначення A_3, $B_{3,v}$, $C_{3,v}$, $D_{3,v}$ такі ж як і в (2.24).

Як і раніше, при $S=3$ ММПл -оцінки можуть бути знайдені лише за допомогою чисельного рішення систем рівнянь (2.36). При застосуванням ітераційної процедури Ньютона-Рафсона необхідні елементи матриці $\boldsymbol{Z}_3(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$ можна представити у вигляді

$$Z_{3N}^{(p,q)}=\sum_{v=1}^{N}\sum_{v=1}^{N}\left\{x_v^{p+g}\left[3A_3\left(\sum_{p=0}^{Q-1}a_p x_v^p\right)^2+2B_{3,v}\left(\sum_{p=0}^{Q-1}a_p x_v^p\right)+C_{3,v}\right]\right\}.$$

2.5 Знаходження ММПл-оцінок параметрів нелінійної регресії

Відомо, що спосіб моделювання, заснований на лінійних регресійних моделями в багатьох реальних ситуаціях не забезпечує адекватний результат. Тому достатньо часто виникає необхідність використання більш складних, нелінійних залежностей, зокрема, базованих на експоненціальних, логарифмічних, тригонометричних та інших функціях або їх комбінаціях. Зазначимо, що частина із них шляхом процедури лінеаризації може бути зведена до лінійної багатофакторної регресі. Крім того, сукупність нелінійних регресійних моделей можна умовно поділити на два класи:

1) регресії, що є нелінійними щодо включених в аналіз пояснювальних змінних (регресорів), але є лінійними щодо інформативних параметрів, які підлягають оцінюванню;
2) регресії, що є нелінійними як щодо включених в аналіз пояснювальних змінних (регресорів), так і нелінійними щодо інформативних параметрів, які підлягають оцінюванню.

До першого класу відносяться поліноміальні регресії, знаходження ММПл-оцінок параметрів яких було розглянуто у попередньому підрозділі. Нижче розглянемо особливості застосування методу максимізації поліномів для оцінювання параметрів моделей нелінійних регресій, що відносяться до другого класу. Зазначимо, що з математичної точки зору такі регресійні моделі підпадають під опис зальний виразом (2.1), як і сама постановка загальної задачі, викладена у підрозділі 2.1

2.5.1. Застосування методу максимізації полінома при степені *S*=1 для оцінювання параметрів нелінійної регресії

При використанні стохастичного полінома порядку $S = 1$ ММПл-оцінки елементів векторного параметра $\boldsymbol{\theta}$ детермінованої складової нелінійної регресійної моделі (2.1) можуть бути знайдені з рішення системи Q рівнянь виду:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}[y_v - R_v(\boldsymbol{\theta}, \boldsymbol{X})]\right\} = 0, \qquad p = \overrightarrow{0, Q-1}. \tag{2.35}$$

де оптимальні коефіцієнти $k_{1,v}^{(p)}$ знаходяться як розв'язання системи лінійних алгебраїчних рівнянь (2.14), і які можна представити у вигляді

$$k_{1,v}^{(p)} = \frac{1}{\mu_2}\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta}, \boldsymbol{X}), p = \overrightarrow{0, Q-1}. \tag{2.36}$$

Використаємо припущення аналогічне до того, що використовується при застосуванні звичайного МНК при оцінюванні параметрів нелінійної регресії (див. підрозділ 1.3.1). Воно полягає у тому, що детермінована складова нелінійної моделі (2.1) може бути апроксимована лінійною функцією, побудованою на основі ряду Тейлора першого порядку (1.5). Тоді із урахуванням (2.36) після певних трансформацій можна представити у матричному вигляді систему лінійних рівнянь

$$(\boldsymbol{H}^T\boldsymbol{H})\Delta\boldsymbol{\theta}^T \approx \frac{1}{N}\sum_{v=1}^{N}\boldsymbol{H}^T\Delta\boldsymbol{Y}^T, \tag{2.37}$$

яка є еквівалентна відповідній МНК-системі (1.6). Таким чином, як і у для МНК при знаходженні лінійних ММПл-оцінок використовується матриця $\boldsymbol{H}$, що складається із відповідних часткових похідних (2.36).

Як відомо [10] оцінки, що отримуються із (2.37) на основі використанням відповідних чисельних ітераційних процедур (градієнтного спуску, методу Ньютона-Гауса або алгоритму Левенберга-Марквардта) можуть мати певний зсув, але їх можна використовувати як перше наближення для більш складних методів оцінювання, зокрема ММПл при більш високих степенях полінома.

2.5.2. Застосування методу максимізації полінома при степені *S*=2 для оцінювання параметрів нелінійної регресії

Використовуючи загальну формулу (2.6) представимо систему рівнянь для пошуку ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(2)}$ детермінованої складової нелінійної регресійної моделі (2.1) у вигляді:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}[y_v - R_v(\boldsymbol{\theta},\boldsymbol{X})] + k_{2,v}^{(p)}\left[(y_v)^2 - \left[\left(R_v(\boldsymbol{\theta},\boldsymbol{X})\right)^2 + \mu_2\right]\right]\right\} = 0,\, p = \overrightarrow{0, Q-1}, \quad (2.38)$$

де оптимальні коефіцієнти $k_{i,v}^{(p)}$, $i = \overrightarrow{1,2}$ можуть бути записані наступним чином:

$$k_{1,v}^{(p)} = \frac{\mu_4 - \mu_2^2 + 2\mu_3 R_v(\boldsymbol{\theta},\boldsymbol{X})}{\mu_2(\mu_4 - \mu_2^2) - \mu_3^2}\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta},\boldsymbol{X}),$$

$$k_{2,v}^{(p)} = -\frac{\mu_3^2}{\mu_2(\mu_4 - \mu_2^2) - \mu_3^2}\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta},\boldsymbol{X}). \quad (2.39)$$

Підставивши коефіцієнти (2.39) у (2.38), після певних перетворень систему рівнянь для знаходження оцінок можна записати у вигляді:

$$\sum_{v=1}^{N}\left\{\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta},\boldsymbol{X})\left[A_2\left(R_v(\boldsymbol{\theta},\boldsymbol{X})\right)^2+B_{2,v}R_v(\boldsymbol{\theta},\boldsymbol{X})+C_{2,v}\right]\right\}=0,\ p=\overrightarrow{0,Q-1}, \tag{2.40}$$

де позначення A_2, $B_{2,v}$, $C_{2,v}$ такі ж як і в (2.19).

Елементи матриці $\boldsymbol{Z}_2(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$, необхідні для чисельного розв'язку системи (2.40), можна представити у вигляді

$$Z_{2N}^{(p,q)}=\sum_{v=1}^{N}\sum_{v=1}^{N}\left\{\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta},\boldsymbol{X})\frac{\partial}{\partial a_q}R_v(\boldsymbol{\theta},\boldsymbol{X})\left[2A_2R_v(\boldsymbol{\theta},\boldsymbol{X})+B_{2,v}\right]\right\}.$$

Як початкові наближення можуть бути використані наближені оцінки, знайдені із вирішення лінійної системи виду (2.37).

2.5.3. Застосування методу максимізації полінома при степені *S*=3 (симетричний випадок) для оцінювання параметрів нелінійної регресії

При застосуванні методу максимізації полінома при степені S=3 у випадку симетрично розподіленої випадкової складової нелінійних регресійних моделей систему із Q рівнянь для пошуку ММПл-оцінок $\widehat{\boldsymbol{\theta}}_{(3)}$ можна представити у вигляді:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}[y_v-R_v(\boldsymbol{\theta},\boldsymbol{X})]+k_{2,v}^{(p)}[y_v^2-([R_v(\boldsymbol{\theta},\boldsymbol{X})]^2+\mu_2)]+\right.$$
$$\left.+k_{3,v}^{(p)}\left[y_v^3-\left([R_v(\boldsymbol{\theta},\boldsymbol{X})]^3+3\mu_2R_v(\boldsymbol{\theta},\boldsymbol{X})\right)\right]\right\}=0,\quad p=\overrightarrow{0,Q-1}, \tag{2.41}$$

де оптимальні коефіцієнти

$$k_{1,v}^{(p)}=\frac{3\left(R_v(\boldsymbol{\theta},\boldsymbol{X})\right)^2(\mu_4-3\mu_2^2)+3\mu_4\mu_2-\mu_6}{\mu_2^{-2}(\mu_4^2-\mu_2\mu_6)}\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta},\boldsymbol{X}),$$

$$k_{2,v}^{(p)}=\frac{-3(\mu_4-3\mu_2)}{\mu_2^{-2}(\mu_4^2-\mu_2\mu_6)}R_v(\boldsymbol{\theta},\boldsymbol{X})\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta},\boldsymbol{X}),$$

$$k_{3,v}^{(p)} = \frac{\mu_4 - 3\mu_2}{\mu_2^{-2}(\mu_4^2 - \mu_2\mu_6)} \frac{\partial}{\partial a_p} R_v(\boldsymbol{\theta}, \boldsymbol{X}), p = \overrightarrow{0, Q-1}. \quad (2.42)$$

Підставляючи коефіцієнти (2.42) в (2.41), після певних перетворень систему рівнянь для знаходження оцінок можна записати у вигляді:

$$\sum_{v=1}^{N} \left\{ \frac{\partial}{\partial a_p} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \left[A_3 \left(R_v(\boldsymbol{\theta}, \boldsymbol{X}) \right)^3 + B_{3,v} \left(R_v(\boldsymbol{\theta}, \boldsymbol{X}) \right)^2 + C_{3,v} R_v(\boldsymbol{\theta}, \boldsymbol{X}) + D_{3,v} \right] \right\} = 0,$$

$$p = \overrightarrow{0, Q-1}, \quad (2.43)$$

де позначення A_3, $B_{3,v}$, $C_{3,v}$, $D_{3,v}$ такі ж як і в (2.24).

Як і раніше, при $S = 3$ ММПл -оцінки можуть бути знайдені лише за допомогою чисельного рішення систем рівнянь (2.43). При застосуванням ітераційної процедури Ньютона-Рафсона необхідні елементи матриці $\boldsymbol{Z}_3(\boldsymbol{Y}/\widehat{\boldsymbol{\theta}}^{(k)})$ можна представити у вигляді:

$$Z_{3N}^{(p,q)} = \sum_{v=1}^{N} \sum_{v=1}^{N} \left\{ \frac{\partial}{\partial a_p} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \frac{\partial}{\partial a_q} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \left[3A_3 \left(R_v(\boldsymbol{\theta}, \boldsymbol{X}) \right)^2 + 2B_{3,v} R_v(\boldsymbol{\theta}, \boldsymbol{X}) + C_{3,v} \right] \right\}.$$

2.6 Адаптивне оцінювання параметрів регресії із застосуванням методу максимізації полінома

Отримані результати попередніх підрозділів свідчать про доцільність застосування методу максимізації поліномів лише при відмінності розподілу випадкової складової регресійних моделей від гаусового закону. Нагадаємо той важливий з практичної точки зору факт, що для отримання як ММПл так і ММП-оцінок необхідна додаткова апріорна інформація про властивості помилок регресійної моделі, яка в реальних умовах зазвичай відсутня.

Відомо, що при виконанні припущення, що помилки моделі не залежать від регресорів може бути застосовано адаптивний підхід [94]. Адаптивність розуміється у тому сенсі, що застосовується ряд послідовних уточнюючих кроків. На першому етапі здійснюється оцінювання параметрів детермінованої складової регресійної моделі звичайним МНК, який не враховує специфіку ймовірнісного розподілу помилок. Після видалення детермінованої складової здійснюється ідентифікація типу випадкової складової регресійної моделі та знаходять оцінки її параметрів. На третьому етапі знаходять уточнені (адаптивні) оцінки інформативних параметрів регресії моделі із урахуванням ймовірнісних властивостей похибок. Даний підхід базується на тій властивості, що неадекватність гаусової моделі не є критичним фактором із тієї точки зору, що МНК-оцінки регресійних параметрів все одно залишаються незміщеними і слушними, хоча і перестають бути оптимальними. А оскільки метод найменших квадратів за своєю суттю є лінійним то ймовірні властивості регресійних залишків практично не відрізняються від властивостей вихідної випадкової складової регресійній моделі [4].

Нагадаємо вже вище зазначений факт, що задача знаходження оціночних значень статистик вищих порядків методологічно є суттєво простішою ніж задача ідентифікації і оцінювання параметрів щільності розподілу. В цьому сенсі суттєво спрощується вирішення задачі подолання апріорної невизначеності щодо імовірнісних характеристик моделі порівняно з підходом, який базується на максимальній правдоподібності. Крім того ступінь невизначеності оцінок моментів (в залежності від їх порядку) теж може бути меншою невизначеності оцінок параметрів щільності розподілу. І як буде показано в розділі 4 цей фактор може призводити до додаткового покращення (відносно ММП) результуючих адаптивних ММПл-оцінок інформативних параметрів регресії.

З урахуванням вище викладеного, узагальнимо алгоритм (графічне представлення блок-схеми якого наведено на рис.2.1), для знаходження адаптивних ММПЛ-оцінок параметрів регресії у вигляді

1 крок – знаходження лінійних МНК-оцінок параметрів регресії;

2 крок – формування регресійних залишків і знаходження апостеріорних оцінок моментів до 4-го порядку;

3 крок – перевірка гіпотези про гаусовий розподілі регресійних МНК-залишків (в разі її підтвердження алгоритм завершується);

4 крок – перевірка гіпотези про симетрію розподілу регресійних МНК-залишків (в разі її підтвердження здійснюється перехід до кроку 6);

5 крок – знаходження ММПл-оцінок з використанням полінома ступеня $S = 2$ і завершення алгоритму;

6 крок – знаходження оцінок моментів 6-го порядку регресійних МНК-залишків;

7 крок – знаходження ММПл -оцінок з використанням полінома ступеня $S = 3$ і завершення алгоритму.

При цьому зазначимо, що отримані апостеріорні оцінки статистик вищих порядків (зокрема кумулянтних коефіцієнтів асиметрії і ексцесу) можуть бути використані для перевірки гіпотези про гаусовий розподіл і симетрію регресійних помилок на основі статистичних тестів Харке-Бера [90].

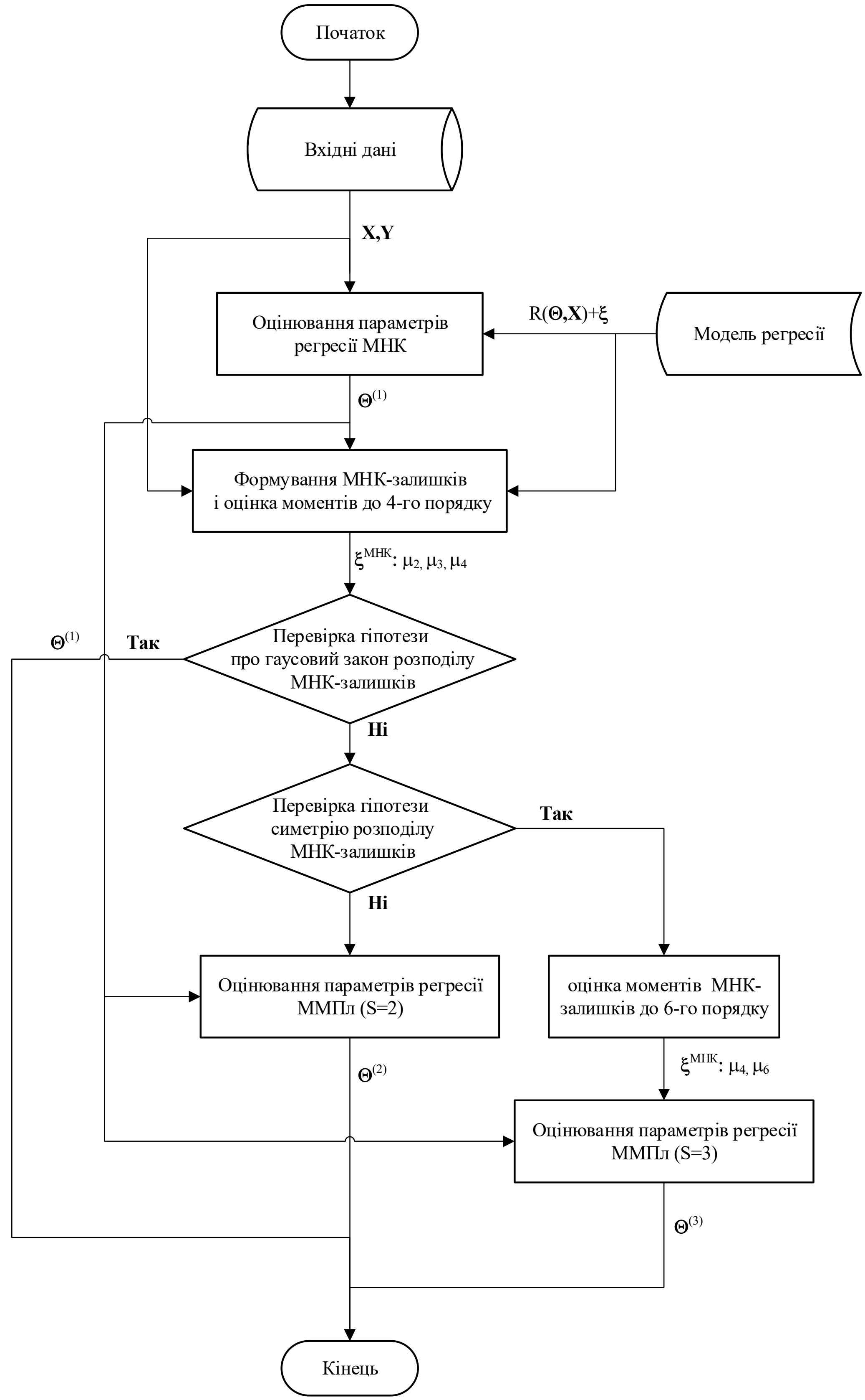


Рисунок 2.1 – Блок-схема алгоритму адаптивного оцінювання параметрів регресії методом максимізації поліномів

Висновки

В даному розділі на основі використання апарату стохастичних поліномів Кунченка та часткового опису статистиками вищих порядків здійснено синтез обчислювальних методів та алгоритмів адаптивного оцінювання параметрів регресійних моделей лінійного, поліноміального і нелінійного типу. На основі сукупності отриманих результатів можна зробити наступні висновки:

1. Загальна задача знаходження оцінок інформативних параметрів регресійних моделей із застосуванням методу максимізації полінома, може бути зведена до розв’язання системи нелінійних (степеневих) стохастичних рівнянь із застосуванням чисельних ітераційних процедур Ньютона-Рафсона.
2. При використанні стохастичних поліномів степені S=1 системи рівнянь методу максимізації полінома є еквівалентними відповідним системам, що отримуються із застосуванням методу найменших квадратів, для кожного із типів розглянутих регресійних моделей.
3. Використання стохастичних поліномів при степені S=2 для оцінювання параметрів регресії є доцільним лише для випадку асиметричних негаусових розподілів регресійних помилок.
4. При симетричному характері розподілу регресійних помилок для оцінювання параметрів доцільним є застосування стохастичних поліномів лише при степені S=3.
5. За відсутності апріорної інформації про властивості випадкової складової регресійної моделей, що є характерним для реальних ситуацій, може бути застосований адаптивний підхід, заснований на використанні апостеріорних оцінок статистик вищих порядків регресійних МНК-залишків.

РОЗДІЛ 3
АНАЛІЗ ТЕОРЕТИЧНОЇ ЕФЕКТИВНОСТІ ПОЛІНОМІАЛЬНИХ ОЦІНОК ПАРАМЕТРІВ РЕГРЕСІЇ

3.1 Асимптотичні дисперсії ММПл-оцінок векторного параметра при неоднаково розподілених даних

Відомо, що основним критерієм ефективності оцінок інформативних параметрів регресії виступають їх дисперсії. Як вже зазначалося раніше, оцінки векторних параметрів, які знаходяться із застосуванням методу максимізації поліномів, в загальному випадку володіють властивістю слушності і є асимптотично незміщеними. Для отримання аналітичних виразів, що описують дисперсії ММПл-оцінок векторного параметру $\boldsymbol{\theta}$, використовується матриця кількості добутої інформації про компоненти параметру при застосуванні стохастичних поліномів загального виду (2.6) порядку S [58]. При моментному описі така матриця $\boldsymbol{J}_S(\boldsymbol{\theta})$ складається із елементів

$$J_{SN}^{(p,q)} = \sum_{v=1}^{N} \sum_{i=1}^{S} \sum_{j=1}^{S} k_{iv}^{(p)} k_{jv}^{(q)} F_{(i,j)v} = \sum_{v=1}^{N} \sum_{i=1}^{S} k_{i,v}^{(p)} \frac{\partial}{\partial a_q} \alpha_{iv}, p, q = \overrightarrow{0, Q-1}. \quad (3.1)$$

У статистичному сенсі кількість добутої інформації концептуально є поняттям близьким до кількості інформації по Фішеру. Показано, що дисперсії $\sigma^2_{(a_p)S}$ ММПл-оцінок складових векторного параметру $\boldsymbol{\theta}$ в асимптотичному випадку (при $N \to \infty$) можуть бути отримані як елементи головної діагоналі варіаційної матриці $\boldsymbol{V}_S(\boldsymbol{\theta})$, яка є оберненою до матриці, складеної із елементів матриці (3.1)

$$\boldsymbol{V}_{ММПл}(\boldsymbol{\theta}) = \begin{bmatrix} J_{SN}^{(1,q)} & J_{SN}^{(2,q)} & \dots & J_{SN}^{(1,q)} \\ J_{SN}^{(2,1)} & J_{SN}^{(2,2)} & \dots & J_{SN}^{(2,q)} \\ \dots & \dots & \dots & \dots \\ J_{SN}^{(p,1)} & J_{SN}^{(p,2)} & \dots & J_{SN}^{(p,q)} \end{bmatrix}^{-1}. \quad (3.2)$$

Ще одна важлива властивість ММПл полягає у тому, що при збільшенні числа членів S стохастичного полінома (1) дисперсія оцінок зменшується, оскільки кількість добутої інформації асимптотично (при $S \to \infty$) прямує до інформації по Фішеру [58].

3.2 Порівняльний аналіз теоретичної ефективності ММПл-оцінок із МНК-оцінками

3.2.1. Точність лінійних ММПл-оцінок параметрів регресії при моментному описі

Синтезовані у розділі 2 обчислювальні алгоритми оцінювання параметрів регресії базуються на застосуванні як базисних функції МММл степеневих перетворень (2.11). Це визначає необхідність використання для ймовірнісного опису випадкових складових регресійних моделей сукупність відповідних початкових моментів випадкових послідовностей (2.12).

Використовуючи результати підрозділів 2.3.2, 2.4.2, 2.5.1 щодо застосування лінійного варіанта методу максимізації полінома можемо записати вираз для кількості добутої інформації (при степені S=1) про значення оцінок параметрів для лінійної багатофакторної регресії

$$J_{1N}^{(p,q)} = \frac{1}{\mu_2} \sum_{v=1}^{N} \left[x_{v,p} x_{v,q} \right], p, q = \overrightarrow{0, Q-1}, \qquad (3.3, a)$$

для поліноміальної регресії

$$J_{1N}^{(p,q)} = \frac{1}{\mu_2} \sum_{v=1}^{N} \left[x_v^{p+q} \right], p, q = \overrightarrow{0, Q-1} \qquad (3.3, б)$$

та нелінійної регресії

$$J_{1N}^{(p,q)} = \frac{1}{\mu_2} \sum_{v=1}^{N} \left[\frac{\partial}{\partial a_p} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \frac{\partial}{\partial a_q} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \right], p, q = \overrightarrow{0, Q-1}, \qquad (3.3,\ в)$$

де μ_2 – центральний момент 2-го порядку (дисперсія) випадкових помилок ξ_v відповідної регресійної моделі.

Зазначимо, отриманий у розділі 2 факт, що результуючі вирази для знаходженні ММПл-оцінки параметрів регресійних моделей при степені *S*=1 є еквівалентними МНК-оцінкам. А отже у цьому випадку співпадає і точність оцінок обох методів. Тому отримані вирази (3.3) можуть бути використані у подальшому для порівняльного аналізу ефективності ММПл-оцінок, які отримуються із застосуванням поліномів біль високого порядку.

3.2.2. Відносна ефективність квадратичних оцінок ММПл-оцінок при асиметрично-розподілених моделях регресійних помилок

Наведені результати підрозділів 2.3.3, 2.4.3, 2.5.2 свідчать, що застосування квадратичного варіанта методу максимізації полінома для знаходження оцінок параметрів регресії є доцільним лише у випадку асиметрії розподілу випадкової складової регресійних помилок. Одним із математичних критеріїв цієї властивості може виступати відмінність величини центрального моменту 3-го порядку від нуля $\mu_3 = 0$.

Використовуючи отримані вирази для оптимальних коефіцієнтів $k_{i,v}^{(p)}$, $i = \overrightarrow{1,2}$, що забезпечують мінімізацію дисперсії оцінок компонентів шуканого параметра при степені *S*=2, можемо записати відповідні вирази для кількості добутої інформації про значення оцінок параметрів для лінійної багатофакторної регресії

$$J_{2N}^{(p,q)} = \frac{\mu_2^3 - \mu_2\mu_4}{\mu_2(\mu_2^3 - \mu_2\mu_4 + \mu_3^2)} \sum_{v=1}^{N} \left[x_{v,p} x_{v,q} \right], p, q = \overrightarrow{0, Q-1}, \qquad (3.4,\ a)$$

для поліноміальної регресії

$$J_{2N}^{(p,q)} = \frac{\mu_2^3 - \mu_2\mu_4}{\mu_2(\mu_2^3 - \mu_2\mu_4 + \mu_3^2)} \sum_{v=1}^{N} \left[x_v^{p+q}\right], p, q = \overrightarrow{0, Q-1} \qquad (3.4, б)$$

та нелінійної регресії

$$J_{2N}^{(p,q)} = \frac{\mu_2^3 - \mu_2\mu_4}{\mu_2(\mu_2^3 - \mu_2\mu_4 + \mu_3^2)} \sum_{v=1}^{N} \left[\frac{\partial}{\partial a_p} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \frac{\partial}{\partial a_q} R_v(\boldsymbol{\theta}, \boldsymbol{X})\right], p, q = \overrightarrow{0, Q-1}, \qquad (3.4, в)$$

де μ_2, μ_3, μ_4 – центральні моменти випадкових помилок ξ_v відповідної регресійної моделі.

Зіставлення відповідних виразів (3.3) і (3.4) дозволяє зробити два основні висновки:

1) вирази, які описують кількість добутої інформації при степені S=2 для різних типів регресійних моделей відрізняться між собою однаковим ваговим коефіцієнтом, який залежить від центральних моментів випадкової складової регресійної моделі;

2) ефективність квадратичних ММПл-оцінок відносно лінійних ММПл-оцінок (а відповідно і МНК-оцінок) параметрів регресії залежить від значень статистик вищих порядків (до 4-го порядку).

Таким чином, можна записати співвідношення для відповідних варіаційних матриць загального виду (3.2), що містять асимптотичні дисперсії оцінок параметрів регресії, у вигляді

$$\boldsymbol{V}_{(ММПл2)} = g_{\frac{ММПл2}{ММПл1}} \boldsymbol{V}_{(ММПл1)} = g_{\frac{ММПл2}{МНК}} \boldsymbol{V}_{(ММПл1)}, \qquad (3.5)$$

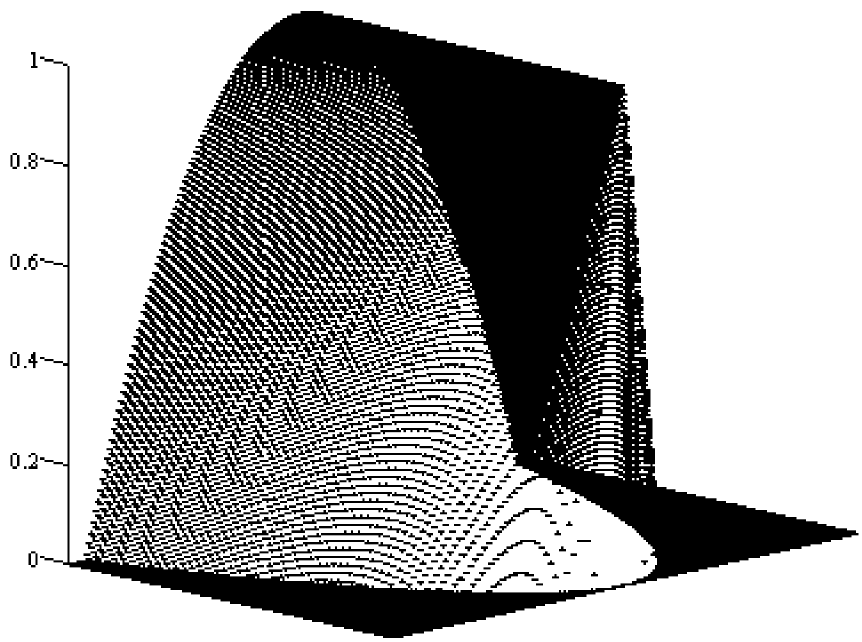
де коефіцієнт зменшення дисперсії оцінок, який описує залежність відносної точності, може бути представлений у вигляді

$$g_{\frac{ММПл2}{МНК}} = 1 - \frac{\gamma_3^2}{2 + \gamma_4}. \qquad (3.6)$$

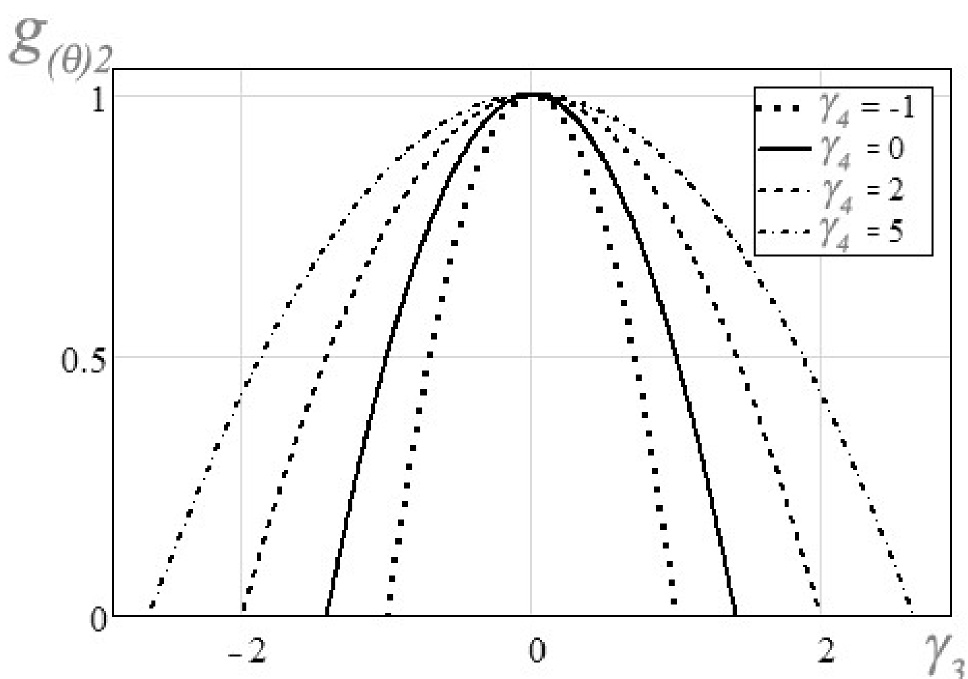
Перехід у виразі (3.6) від моментного опису до кумулянтного обумовлений тим фактом, що відхилення значень кумулянтних коефіцієнтів вищих порядків γ_r від нуля показує міру відмінності від гаусового розподілу. Крім того, кумулянтні коефіцієнти вищих порядків не можуть приймати довільних значень. Зокрема, для кумулянтних коефіцієнтів асиметрії та ексцесу відомою є нерівністю $\gamma_4 + 2 \geq \gamma_3^2$ [46]. З урахуванням цієї нерівності, на основі аналізу (3.6) можна зробити висновок, що коефіцієнт зменшення дисперсії $g_{\frac{\text{ММПл2}}{\text{МНК}}}$ є безрозмірною величиною, що належить діапазону $(0; 1]$.

На рис.3.1 зображена залежність коефіцієнту зменшення дисперсії $g_{\frac{\text{ММПл2}}{\text{МНК}}}$ від кумулянтних коефіцієнтів асиметрії γ_3 та ексцесу γ_4. Оскільки коефіцієнт зменшення дисперсії є функціями двох змінних, то множина їх значень представляє собою поверхню (рис.3.1, *а*). Для більшої наочності крім 3D-графіка на рис.3.1,*б* також представлені розрізи, побудовані як залежності від коефіцієнта асиметрії γ_3 при фіксованих значеннях γ_4.

Таким чином, відносне зменшення асимптотичної дисперсії ММПЛ-оцінок є однаковим для всіх компонентів векторного параметра. Загалом відносна ефективність залежить від ступені відмінності розподілу регресійних помилок від гаусового закону, що чисельно виражається перш за все відмінністю величини абсолютного значення (модуля) коефіцієнта асиметрії γ_3. При її збільшенні величина відносного зменшення дисперсії може бути достатньо суттєвою і навіть асимптотично прагнути до нуля при наближенні абсолютного значення коефіцієнту асиметрії до межі області допустимих значень, тобто $|\gamma_3| \rightarrow \sqrt{2 + \gamma_4}$.



а)



б)

Рисунок 3.1 – Залежність коефіцієнту зменшення дисперсії $g_{\frac{\text{ММПл2}}{\text{МНК}}}$ від кумулянтних коефіцієнтів асиметрії γ_3 та ексцесу γ_4

3.2.3. Відносна ефективність кубічних ММПл-оцінок при симетрично-розподілених моделях регресійних помилок

Використовуючи результати підрозділів 2.3.4, 2.4.4, 2.5.3 щодо застосування кубічного варіанта методу максимізації полінома для випадку симетрично розподіленої випадкових регресійних помилок, можемо записати відповідні вирази для кількості добутої інформації при *S*=3 про значення оцінок параметрів для лінійної багатофакторної регресії у вигляді

$$J_{2N}^{(p,q)} = \frac{9\mu_2^3 - 6\mu_2\mu_4 + \mu_6}{\mu_2\mu_6 - \mu_4^2} \sum_{v=1}^{N} \left[x_{v,p} x_{v,q} \right], \; p, q = \overrightarrow{0, Q-1}, \tag{3.7, а}$$

для поліноміальної регресії

$$J_{2N}^{(p,q)} = \frac{9\mu_2^3 - 6\mu_2\mu_4 + \mu_6}{\mu_2\mu_6 - \mu_4^2} \sum_{v=1}^{N} \left[x_v^{p+q} \right], \; p, q = \overrightarrow{0, Q-1} \tag{3.7, б}$$

та нелінійної регресії

$$J_{2N}^{(p,q)} = \frac{9\mu_2^3 - 6\mu_2\mu_4 + \mu_6}{\mu_2\mu_6 - \mu_4^2} \sum_{v=1}^{N} \left[\frac{\partial}{\partial a_p} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \frac{\partial}{\partial a_q} R_v(\boldsymbol{\theta}, \boldsymbol{X}) \right], \; p, q = \overrightarrow{0, Q-1}, \tag{3.7, в}$$

де μ_2, μ_4, μ_6 – парні центральні моменти випадкових помилок ξ_v відповідної регресійної моделі.

Таким чином, як і в попередньому випадку ефективність кубічних ММПл-оцінок відносно лінійних ММПл-оцінок (а відповідно і МНК-оцінок) не відрізняться для різних типів регресійних моделей і визначається однаковим ваговим коефіцієнтом

$$\boldsymbol{V}_{(ММПл3)} = g_{\frac{ММПл3}{ММПл1}} \boldsymbol{V}_{(ММПл1)} = g_{\frac{ММПл3}{МНК}} \boldsymbol{V}_{(ММПл1)}, \tag{3.8}$$

який залежить від парних центральних моментів випадкової складової регресійної моделі (до 6-го порядку).

Для компактності запису та більш зручного трактування результатів аналогічно до (3.6) здійснимо перехід від моментного до кумулянтного опису і представимо коефіцієнт зменшення дисперсії ММПл оцінок при *S*=3 у вигляді

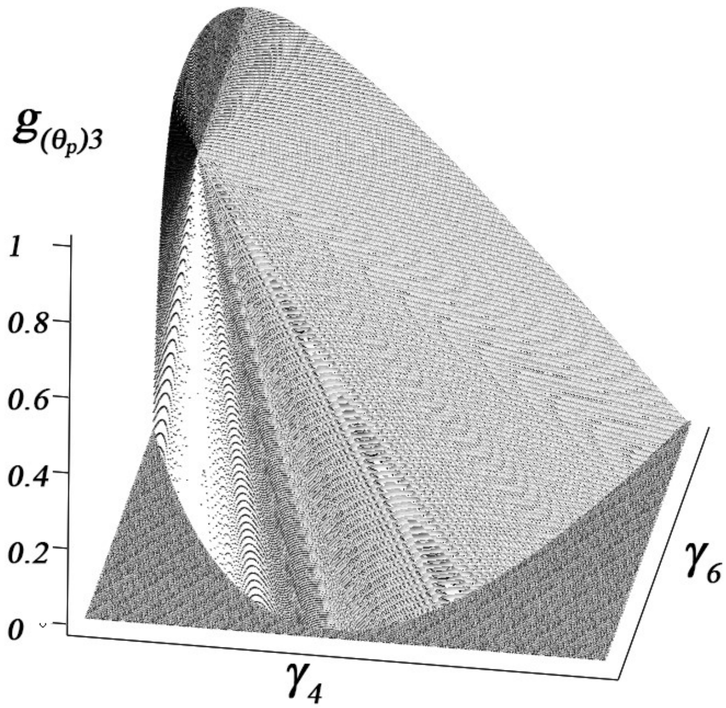
$$g_{\frac{ММПл3}{МНК}} = 1 - \frac{\gamma_4^2}{6+9\gamma_4+\gamma_6}. \qquad (3.9)$$

Як вже було зазначено вище, для значень кумулянтних коефіцієнтів вищого порядку існують певні області допустимих значень. Зокрема, для симетрично розподілених випадкових величин обмеження для кумулянтних коефіцієнтів 4-го і 6-го порядку визначається нерівностями: $\gamma_4 > -2$ та $\gamma_6 + 9\gamma_4 + 6 > \gamma_4^2$ [46]. Враховуючи ці нерівності, можна зробити висновок, що коефіцієнт зменшення дисперсії $g_{\frac{\text{ММПл3}}{\text{МНК}}}$ також є безрозмірною величиною, що належить діапазону $(0; 1]$.

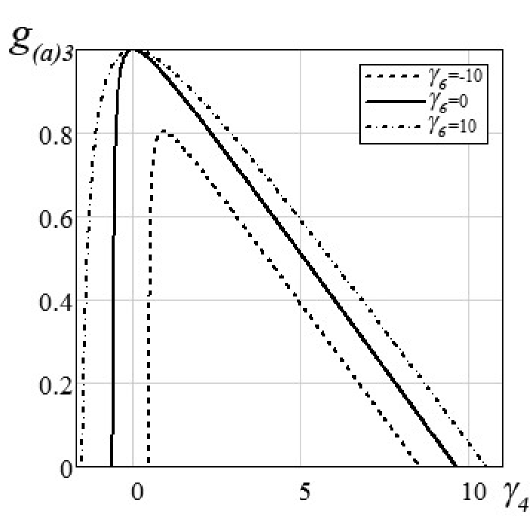
На рис.3.2,*б* зображена залежність (у вигляді 3D-графіка) коефіцієнту зменшення дисперсії $g_{\frac{\text{ММПл3}}{\text{МНК}}}$ від кумулянтних коефіцієнтів 4-го γ_4 та 6-го γ_6 порядків. Для більшої наочності на рис.3.2,*б* представлені розрізи, побудовані як залежності від коефіцієнта γ_4 при фіксованих значеннях γ_6, а на рис.3.2,*в* залежності від коефіцієнта γ_6 при фіксованих значеннях γ_4.

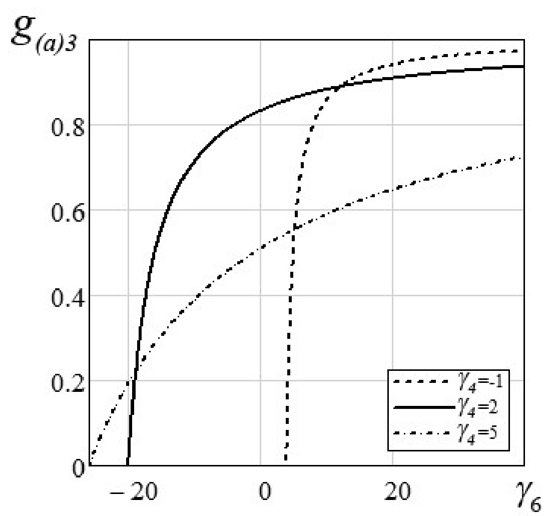
Аналіз представлених залежностей показує, що точність ММПЛ-оцінок при $S = 3$ може може бути суттєво більшою за точність МНК-оцінок. Зокрема, при наближенні величини кумулянтних коефіцієнтів до кордону області допустимих значень $|\gamma_4| \to \sqrt{\gamma_6 + 9\gamma_4 + 6}$ дисперсія ММПл-оцінок асимптотично прагне до нуля.

Відносне зменшення дисперсій оцінок не спостерігається лише у тому випадку, коли коефіцієнт ексцесу тотожний нулю $\gamma_4 = 0$, що зокрема відповідає гаусовому розподілу регресійних помилок.



а)





б) *в)*

Рисунок 3.2 – Залежність коефіцієнту зменшення дисперсії $g_{\frac{\text{ММПлз}}{\text{МНК}}}$ від кумулянтних коефіцієнтів γ_4 та γ_6

3.3 Порівняльний аналіз теоретичної ефективності ММПл-оцінок із ММП-оцінками

Як вже зазначалося раніше, однією із основних альтернатив методу найменших квадратів в регресійному аналізі є підхід, що базується на методі максимальної правдоподібності. Його застосування потенційно є більш точним, оскільки дозволяє враховувати ймовірнісні властивості випадкової складової регресійної моделі. Але платою за це є необхідність вирішення додаткових задач ідентифікації щільності розподілу регресійних помилок та оцінювання її параметрів.

3.3.1. Застосування експоненціального степеневого розподілу для опису регресійних помилок

Здійснимо порівняльний аналіз ефективності цього підходу використавши ЕСР як ймовірнісну модель регресійних помилок. Вибір цієї моделі обумовлений тим фактором, що вона адекватно підходить для опису великого числа процесів та явищ, які спостерігаються в реальності. Про її універсальність говорить той факт, що розподіл Лапласа, гаусовий та рівномірний розподіли є окремими випадками ЕСР при певних значеннях його параметру форми. Тому в різних літературних джерелах можна зустріти й альтернативні назви: «узагальнений нормальний розподіл», «узагальнений гаусовий розподіл», «узагальнений розподіл помилок», «розподіл Суботіна» та ін. [95–97].

ЕСР застосовується як ймовірнісна модель при вирішення великої кількості різноманітних практичних задач: для опису похибок вимірювальних приладів і результатів вимірювання [19, 98], випадкових шумів та завад при опрацюванні радіо [99], відео [100] та аудіо [101] сигналів, статистичних даних медичного [102], біологічного [103], економічного [104] походження, в управлінні ризиками [105] тощо. Подібне широке поширення ЕСР спонукало до розробки ряду спеціалізованих програмних продуктів, орієнтованих на комп'ютерне статистичне моделювання. Серед них виділимо бібліотеку «normalp» [31] для мови програмування R, що використовується у даній роботі.

З урахуванням центрованості (нульового математичного сподівання) регресійних помилок модель ЕСР описується функцією виду

$$w(\xi) = \frac{1}{2\sigma_\beta p^{\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right)} exp\left(-\frac{|\xi|^\beta}{\beta\sigma_\beta^\beta}\right), \quad (3.10)$$

де значення параметрів масштабу σ_β і форми β.

На рис.3.3 наведено графіки щільності ЕСР, побудовані при різних значеннях параметру форми β. Очевидно, що при $\beta = 2$ розподіл (3.10) відповідає гаусовому закону. Для значень $\beta < 2$ функція (3.10) має лептокуртичний (гостровершинний) характер з пологими спадами, при $\beta = 1$ відповідає розподілу Лапласа. Із зростанням величини $\beta > 2$ розподіл стає платокуртичним, тобто більш плосковершинним, а в граничному випадку (при $\beta \to \infty$) трансформується в рівномірний закон розподілу.

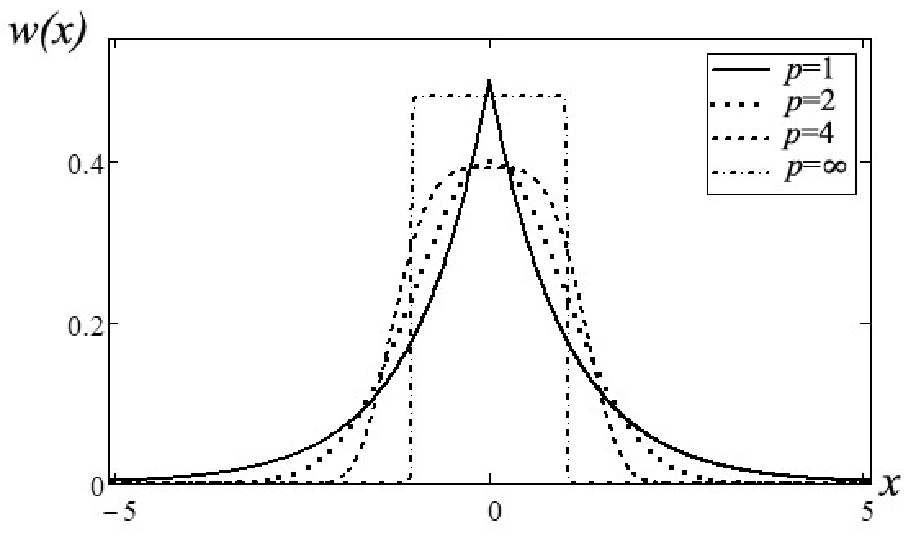


Рисунок 3.3 – Експоненціальний степеневий розподіл при різних значеннях параметра форми β.

3.3.2. Відносна точність оцінок параметрів регресії при експоненціальному степеневому розподілі помилок

Використовуючи (3.10) логарифмічну функцію правдоподібності, яка приймає максимум в околі істинних значень інформативних параметрів $\boldsymbol{\theta}$ детермінованої складової $R_v(\boldsymbol{\theta}, \boldsymbol{X})$ регресійної моделі, можна записати у вигляді

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}) = -N\,log(2) + N\,log(\sigma) - \sum_{v=1}^{N}\left[\frac{\Gamma\left(1+\frac{1}{\beta}\right)|y_v - R_v(\boldsymbol{\theta}, \boldsymbol{X})|}{\sigma}\right]^{\beta}, \quad (3.11)$$

де σ нормалізоване значення середньоквадратичного відхилення

$$\sigma = \sqrt{\mu_2} = \beta^{\frac{1}{\beta}}\sigma_\beta \Gamma\left(\frac{1+1}{\beta}\right). \quad (3.12)$$

Для аналітичного розрахунку величини дисперсії ММП-оцінок, використаємо величину кількості інформації по Фішеру, яка розрахована для регресійних моделей з ЕСР в роботі [106]

$$J_N^{(p,q)} = \frac{\Gamma\left(\frac{1}{\beta}\right)\Gamma\left(2-\frac{1}{\beta}\right)}{\sigma^2}\sum_{v=1}^{N}\left[\frac{\partial}{\partial a_p}R_v(\boldsymbol{\theta}, \boldsymbol{X})\frac{\partial}{\partial a_q}R_v(\boldsymbol{\theta}, \boldsymbol{X})\right], p, q = \overrightarrow{0, Q-1}. \quad (3.11)$$

Із урахуванням (3.3) відповідна варіаційна матриця може бути представлена у вигляді $\boldsymbol{V}_{(MLE)} = g_{\frac{ММП}{МНК}}\boldsymbol{V}_{(МНК)}$, де

$$g_{\frac{ММП}{МНК}} = \frac{\Gamma[1+\beta^{-1}]^2}{\Gamma[2-\beta^{-1}]\Gamma[3\beta^{-1}]}. \quad (3.12)$$

Оскільки ЕСР має симетричний характер, то для отримання оцінок регресійних параметрів необхідно застосовувати кубічну модифікацію методу максимізації поліномів. Для отримання аналітичних виразів, що описують дисперсії ММПл-оцінок, використаємо взаємозв'язок між центральними моментами μ_i і параметрами ЕСР [31]:

$$\mu_i = \Gamma[(i+1)\beta^{-1}]\Gamma[\beta^{-1}]^{\frac{i-2}{2}}\Gamma[3\beta^{-1}]^{-\frac{i}{2}}\sigma_\beta^2. \qquad (3.13)$$

Використовуючи вирази (3.13) та (3.7) для кількості добутої інформації можна показати, що елементи варіаційної матриці ММПл-оцінок при $S = 3$ відрізняються від елементів матриці лінійних МНК-оцінок (3.3) на коефіцієнт, який може бути представлено як функція від параметру форми ЕСР

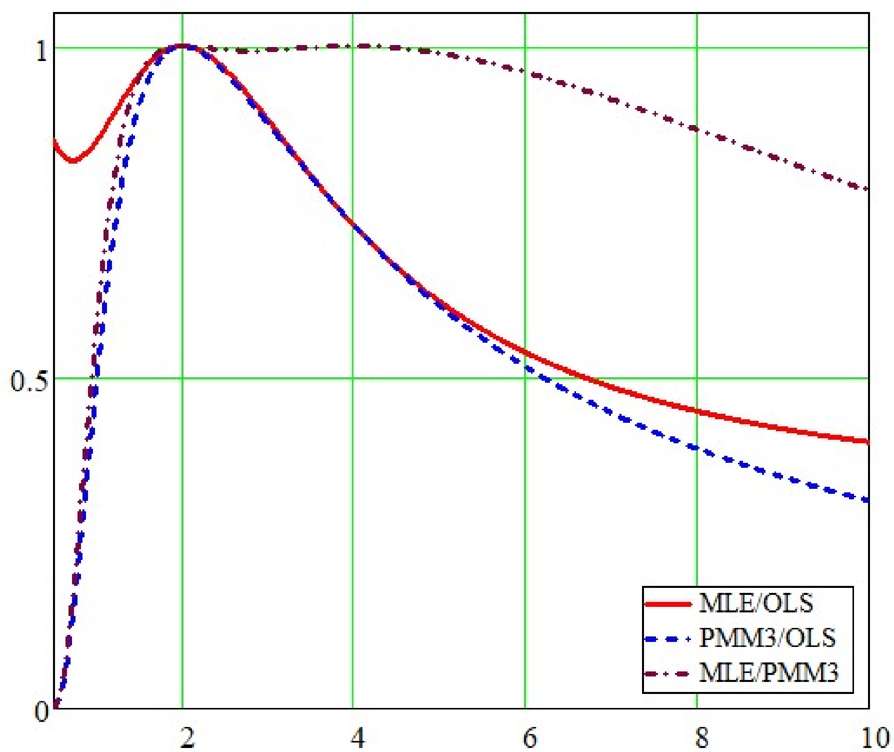


Рисунок 3.4 – Залежність коефіцієнту зміни дисперсії $g_{\frac{MLE}{OLS}}$, $g_{\frac{PMM3}{OLS}}$ і $g_{\frac{MLE}{PMM3}}$ від параметра форми β ЕСР.

$$g_{\frac{PMM3}{OLS}} = \frac{\Gamma[p^{-1}]^2(\Gamma[5\beta^{-1}]^2 - \Gamma[3\beta^{-1}]\Gamma[7\beta^{-1}])}{\Gamma[3\beta^{-1}](9\Gamma[3\beta^{-1}]^3 - 6\Gamma[\beta^{-1}]\Gamma[3\beta^{-1}]\Gamma[5\beta^{-1}] + \Gamma[\beta^{-1}]^2\Gamma[7\beta^{-1}])}. \qquad (3.14)$$

Аналіз залежностей, представлених на рис.3.4, свідчить, що для лептокуртичних (гостровершинних) розподілів при $\beta < 2$ відносна ефективність ММП може бути суттєво вищою за інші методи.

У випадку гаусового розподілу при $\beta = 2$ точність усіх трьох методів (ММП, ММПл і МНК) є однаковою. При зростанні величини параметру форми $\beta > 2$ для достатньо широкого діапазону значень теоретична ефективність ММП і ММПл між собою фактично не відрізняється, але може суттєво перевищувати МНК. Для суттєво платокуртичних (плосковершинних) розподілів при $\beta > 5$ відносна ефективність ММП щодо ММПл поступово зростає.

Висновки

Проведено теоретичний аналіз ефективності поліноміальних оцінок параметрів регресії порівняно із оцінками методу найменших квадратів та максимальної правдоподібності. Отримані наступні основні результати:

1. Дисперсії оцінок параметрів регресійних залежностей, отриманих із застосуванням методу максимізації поліномів, в загальному випадку не перевищують, а в цілому можуть бути значно менші за дисперсії оцінок, знайдених методом найменших квадратів.
2. Ступінь ефективності ММП-оцінок (відносно МНК) не залежить від типу регресійних моделей та є однаковою для всіх складових векторного параметру, що оцінюється.
3. Теоретична величина коефіцієнтів зменшення дисперсії ММПл-оцінок залежить від ступеня негаусовості випадкової складової регресійної моделі.
4. При використанні стохастичних поліномів степені S=2 і асиметричному характері регресійних помилок коефіцієнт зменшення дисперсії (відносно МНК) залежить лише від величини їх коефіцієнтів асиметрії та ексцесу.
5. При симетричному характері регресійних помилок і використанні стохастичних поліномів степені S=3 коефіцієнт зменшення дисперсії

(відносно МНК) залежить від величини їх коефіцієнтів 4-го та 6-го порядків.

6. Зростання відносної точності ММПл-оцінок може бути достатньо суттєвим при наближенні величини кумулянтних коефіцієнтів до меж своїх областей допустимих значень.
7. При ЕСР помилок теоретична точність ММПл-оцінок параметрів регресії у цілому є меншою за точність ММП-оцінок. Проте при платокуртичному (плосковершинному) характері розподілів для достатньо широкого діапазону зміни величини параметру форми ЕСР теоретична точність ММПл-оцінок і ММП-оцінок фактично не відрізняється.

РОЗДІЛ 4
ПРОГРАМНІ ЗАСОБИ СТАТИСТИЧНОГО МОДЕЛЮВАННЯ ПОЛІНОМІАЛЬНОГО ОЦІНЮВАННЯ ПАРАМЕТРІВ РЕГРЕСІЇ ТА РЕЗУЛЬТАТИ ЇХ ЗАСТОСУВАННЯ

4.1 Структура та функціонал програмних засобів поліноміального регресійного аналізу

Сукупність отриманих у попередніх розділах результатів у вигляді відповідних моделей та методів, заснованих на використанні апарату стохастичних поліномів Кунченка та статистик вищих порядків, формують математичний фундамент обчислювальних алгоритмів, що лежать в основі розроблених в рамках даного дисертаційного дослідження програмних засобів, які можуть бути застосовані для ефективного вирішення задач оцінювання інформативних параметрів регресійних залежностей в умовах негаусового розподілу випадкових помилок. Необхідно також зазначити універсальність отриманих результатів, які не прив’язані до специфіки вирішення окремих прикладних задач регресійного аналізу, а орієнтовані на широке коло можливих типів детермінованих складових регресійних моделей: лінійних, поліноміальних та нелінійних.

Загалом, алгоритмічна реалізація компонентів розроблених програмних засобів, що включає опис відповідних моделей та побудову обчислювальних процедур, характеризується відносною простотою. Вона може бути здійснена як за допомогою однієї з мов програмування високого рівня (в даному дисертаційному дослідженні це мова R), так і на основі використання спеціалізованих пакетів комп’ютерної математики та моделювання (в даному в даному дисертаційному дослідженні використано засоби MATLAB).

На рис. 4.1. наведена структура розробленого програмного комплексу, в якому можна виділити три окремі функціональні модулі:

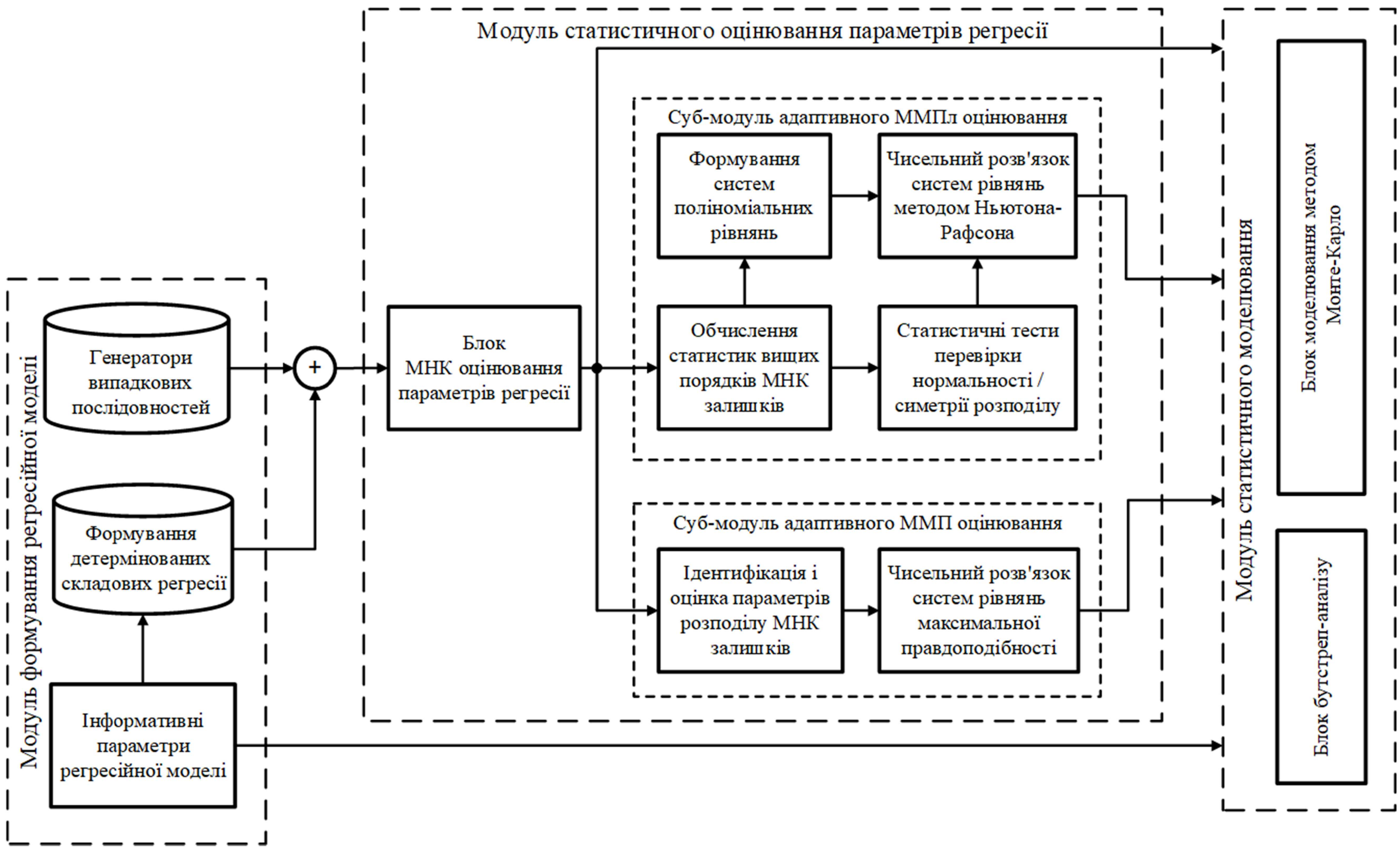


Рисунок 4.1 – Структура програмного комплексу

1) Модуль формування регресійної моделі, що містить блоки генерації детермінованої і випадкової складової регресійних моделей із заданими значеннями інформативних параметрів;
2) Модуль статистичного оцінювання параметрів регресії із застосуванням трьох методів: найменших квадратів, максимальної правдоподібності і максимізації поліномів. Останні два складають суб-модулі адаптивного варіанту оцінювання, що враховують статистичні властивості регресійних МНК-залишків. Суб-модуль ММПл містить блоки статистичного оцінювання моментів вищих порядків та перевірки статистичних гіпотез нормальності (гаусовості) та симетрії розподілу регресійних МНК-залишків, а також оригінальні блоки формування систем рівнянь максимізації поліномів та блок чисельного розв’язку таких нелінійних систем із використанням методу Ньютона-Рафсона.
3) Модуль статистичного моделювання, що містить блок, який реалізує процедуру багаторазового тестування на основі методу Монте-Карло [107], а також блок на основі технології розмноження вибірок, відомий як бутстреп-аналіз [108].

Фактично комплекс розроблених програмних засобів містить в собі оригінальні файли MATLAB (m-функцій) та їх аналоги, написані на мові R, що вирішують задачі, які безпосередньо пов’язані із знаходженням адаптивних поліноміальних оцінок параметрів регресії. А при вирішенні задач формування регресійних моделей та знаходження оцінок їх параметрів класичними методам (МНК і ММП) використовуються наявні у відповідних бібліотеках програмних пакетів вбудовані функції.

Для дослідження ефективності застосування синтезованих поліноміальних методів оцінювання також створені файли-сценарії, що реалізують процедуру багаторазових випробувань методом Монте-Карло, результати застосування яких представлені у наступному підрозділі. А у Додатку А також наведені результати бутстреп-аналізу при дослідженні працездатності синтезованих методів оцінювання параметрів регресійних моделей при опрацюванні реальних статистичних дані, взятих із відкритих джерел (репозиторіїв).

4.2 Результати статистичного моделювання оцінювання параметрів регресії методом Монте-Карло

Нагадаємо, що результати розділу 3, які описують ступінь ефективності синтезованих у розділі 2 поліноміальних моделей та методів знаходження оцінок параметрів регресії отримані за двох теоретичних обмежень. По перше, це те, що розраховані аналітичні вирази, які описують залежність величин дисперсії ММПл-оцінок від параметрів випадкової складової регресійних моделей отримані для асимптотичного випадку, коли обсяг вибірки прямує до нескінченності. По друге, передбачається наявність апріорної інформації про властивості регресійних помилок (у вигляді значень певної кількості їх відповідних моментів або кумулянтів). Тому подібні результати потребують додаткової верифікації з метою перевірки їх достовірності та ефективності застосування в реальних умовах. Одним із найпоширеніших шляхів для такої експериментальної перевірки є використання процедури статистичного моделювання, заснованої на методі Монте-Карло. Сутність цього методу полягає у проведенні багаторазових випробувань, зокрема, оцінювання інформативних параметрів деякої моделі із визначеними (фіксованими) властивостями, але при різних штучно сформованих вибірках випадкових (стохастичних) складових [107].

Оскільки основним критерієм ефективності, що характеризує точність статистичних оцінок деякого інформативного параметра a є їх дисперсія σ_a^2, то експериментальне значення цих характеристик можна обчислити як середнє квадратів відхилення сукупності M оціночних значень $\hat{a}^{(j)}$ від істинного $a^{(0)}$ значення цього параметру:

$$\hat{\sigma}_a^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\hat{a}^{(j)} - a^{(0)}\right)^2. \tag{4.1}$$

Таким чином, експериментальними характеристиками відносної ефективності ММПл при використанні стохастичних поліномів степені S щодо класичних МНК і ММП-оцінок параметрів будуть відповідні коефіцієнти

$$\hat{\boldsymbol{g}}_{\frac{\text{ММПл}\boldsymbol{S}}{\text{МНК}}} = \frac{\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{a}(\text{ММПл}\boldsymbol{S})}}{\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{a}(\text{МНК})}}, \quad \hat{\boldsymbol{g}}_{\frac{\boldsymbol{ММПлS}}{\boldsymbol{МНК}}} = \frac{\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{a}(\boldsymbol{ММПлS})}}{\hat{\boldsymbol{\sigma}}^2_{\boldsymbol{a}(\boldsymbol{ММП})}}.$$

Очевидно, що на точність оцінювання параметрів впливає розмір вибірки статистичних даних N, а на точність визначення дисперсії оцінок параметрів кількість експериментів M. Експерименти здійснюються за умови однакових типів і параметрів регресійних моделей, а також властивостей їх випадкових складових.

4.2.1. Статистичне моделювання оцінювання параметрів лінійної регресії

4.2.1.1. Постановка задачі статистичного моделювання

Результати теоретичного аналізу точності поліноміального оцінювання векторного параметру $\boldsymbol{\theta} = \{a_0, a_1, \dots a_{Q-1}\}$ лінійної регресії загального виду (2.10) свідчать, що відносна ефективність не залежить від розмірності моделі Q і є однаковою для всіх елементів вектору параметрів a_p, $p = \overrightarrow{0,Q-1}$. Тому для реалізації статистичного моделювання можна використати найпростішу лінійну регресійну залежність виду

$$y_v = a_0 + a_1 x_v + \xi_v, \; v = \overrightarrow{1,N}, \tag{4.2}$$

де значення інформативних параметрів a_0 і a_1 детермінованої складової регресійної моделі вважаються апріорно невідомими і підлягають оцінюванню.

Для отримання сукупності експериментальних значень коефіцієнтів відношення дисперсії оцінок $\widehat{\boldsymbol{g}}_{\frac{\text{ММПл}s}{\text{МНК}}}$ необхідно здійснити M експериментів при різних вибірках ξ_v, $v = \overrightarrow{1, N}$, але за умови їх однакових ймовірнісних властивостей.

4.2.1.2. Результати застосування методу максимізації полінома при степені *S*=2

При використанні ММПл степені *S=2* в якості асиметричної випадкової складової (похибки) регресійної моделі використано послідовності незалежних і однаково-розподілених випадкових величин, що мають експоненціальний, логнормальний, гамма та розподіл Вейбула. При цьому значення параметрів випадкової складової регресійної моделі, які необхідні для знаходження адаптивних ММПл-оцінок вважалися апріорно невідомими, а у відповідності до алгоритму адаптивного оцінювання, представленого у підрозділі 2.6, використовувалися апостеріорні оцінки моментів до 4-го порядку регресійних МНК-залишків.

Для знаходження уточнених значень адаптивних ММПл-оцінок реалізується шляхом вирішення систем відповідних поліноміальних рівнянь на основі застосовування чисельної процедури Ньютона-Рафсона. Її використання забезпечує при середній кількості 3-4 ітерацій високу відносну точність на рівні 10^{-5}.

В таблиці 4.1 представлено сукупність результатів статистичного моделювання оцінювання параметрів регресійної моделі виду (4.2), які отримані для $M = 10^4$ експериментів.

Таблиця 4.1 – Результати статистичного моделювання Монте-Карло оцінювання параметрів лінійної регресії при асиметрії регресійних помилок

Розподіл		Теоретичне значення параметрів			Результати статистичного моделювання					
		γ_3	γ_4	$g_{\frac{PMM2}{OLS}}$	$\hat{g}_{\frac{ММПл2}{МНК}}$					
					$N = 20$		$N = 50$		$N = 200$	
					a_0	a_1	a_0	a_1	a_0	a_1
Гама	$\alpha = 0.5$	2.83	12	0.43	0.53	0.38	0.51	0.35	0.48	0.33
Експоненціальний (Гама, $\alpha = 1$)		2	6	0.5	0.66	0.56	0.59	0.46	0.56	0.42
Гама	$\alpha = 2$	1.41	3	0.6	0.75	0.68	0.68	0.59	0.67	0.57
	$\alpha = 4$	1	1.5	0.71	0.88	0.85	0.81	0.74	0.78	0.71
Логнормальний $\sigma^2 = 0.1, \mu = 1$		1	1.86	0.74	0.89	0.85	0.82	0.76	0.79	0.73
Вейбула $a = 1, b = 2$		0.63	0.25	0.82	0.97	0.96	0.91	0.88	0.87	0.83

Аналіз даних, наведених у таблиці 4.1, показує значну кореляцію між результатами аналітичних розрахунків та статистичним моделюванням. Очевидно, що із збільшенням початкового обсягу вибірки розбіжність між теоретичним та експериментальним значеннями коефіцієнта зменшення дисперсії (навіть для малих значеннях обсягу вибірок $N = 20$ різниця не перевищує 20%) зменшується.

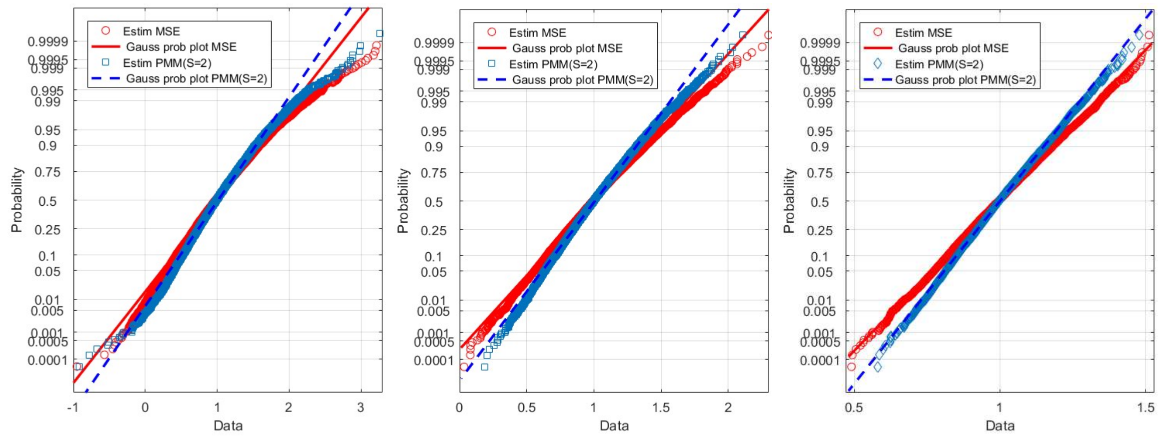
Іншим важливим результатом статистичного моделювання є перевірка припущення тієї важливої з прикладної точки зору властивості, що із збільшенням обсягу вибірок N розподіл ММПл-оцінок асимптотично наближається до гаусового закону. Обґрунтованість цієї гіпотези досліджується за допомогою критерію Лілієфорса на основі статистики Колмогорова-Смірнова [109], яка вбудована в MATLAB. У таблиці 4.2 представлені результати статистичних випробувань у вигляді числових значень величини критерію $LSTAT$ тесту Лілієфорса при різних величинах обсягу вибірок N. Якщо, величина статистики менше порогу $LSTAT < CV$, то нульова гіпотеза (адекватність гаусової моделі) для даного критичного рівня

$p - value$ є дійсною. Результати в таблиці 4.2 отримані для різних типів розподілу помилок регресійної моделі та розміру дат вибірки $N = 20,\ 50,\ 200$ при фіксованому рівні значущості нульової гіпотези $p - value = 0.05$ та $M = 10^4$ експериментів.

Таблиця 4.2 – Результати тестування адекватності гаусового розподілу ММПл-оцінок (для $S = 2$) на основі тесту Лілієфорса.

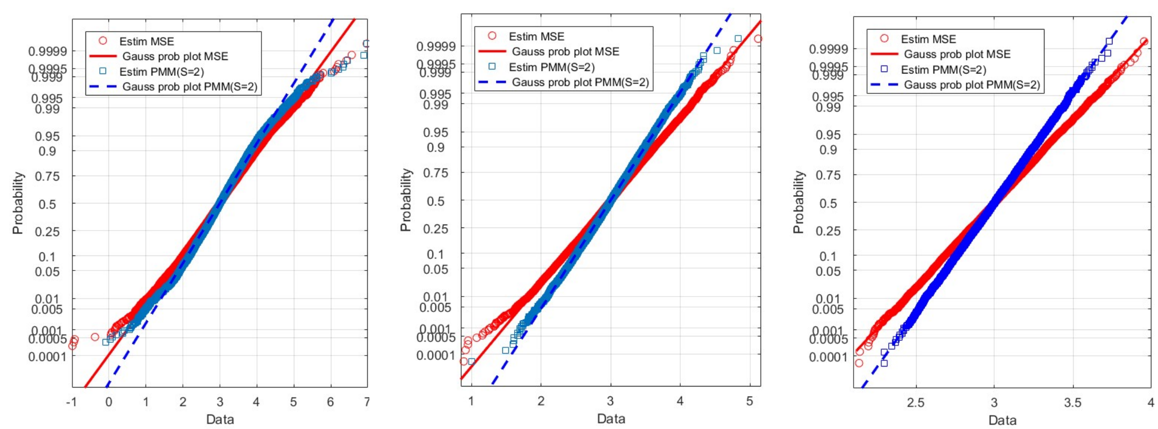
<table>
<tr><th colspan="2" rowspan="4">Розподіл</th><th colspan="7">Результати статистичного моделювання</th></tr>
<tr><th colspan="6">LSTAT</th><th rowspan="3">CV</th></tr>
<tr><th colspan="2">$N = 20$</th><th colspan="2">$N = 50$</th><th colspan="2">$N = 200$</th></tr>
<tr><th>a_0</th><th>a_1</th><th>a_0</th><th>a_1</th><th>a_0</th><th>a_1</th></tr>
<tr><td>Гама</td><td>$\alpha = 0.5$</td><td>0.062</td><td>0.057</td><td>0.041</td><td>0.019</td><td>0.023</td><td>0.009</td><td rowspan="6">0.009</td></tr>
<tr><td colspan="2">Exponential
(Гама, $\alpha = 1$)</td><td>0.051</td><td>0.039</td><td>0.026</td><td>0.014</td><td>0.021</td><td>0.008</td></tr>
<tr><td rowspan="2">Гама</td><td>$\alpha = 2$</td><td>0.031</td><td>0.022</td><td>0.021</td><td>0.009</td><td>0.009</td><td>0.007</td></tr>
<tr><td>$\alpha = 4$</td><td>0.028</td><td>0.016</td><td>0.016</td><td>0.007</td><td>0.008</td><td>0.006</td></tr>
<tr><td colspan="2">Логнормальний
$\sigma^2 = 0.1, \mu = 1$</td><td>0.021</td><td>0.016</td><td>0.014</td><td>0.006</td><td>0.008</td><td>0.005</td></tr>
<tr><td colspan="2">Вейбула
$a = 1, b = 2$</td><td>0.019</td><td>0.013</td><td>0.013</td><td>0.006</td><td>0.007</td><td>0.004</td></tr>
</table>

Крім того, за результатами статистичного моделювання на рис.4.2 та рис.4.3 наведено приклади, що демонструють розподіл експериментальних значень оцінок МНК та ММПл (при $S = 2$) компонентів векторного параметру $\boldsymbol{\theta} = \{a_0, a_1\}$. У цих прикладах результуючі графіки побудовані при $M = 10^4$ експериментах для розміру вибірок $N = \ 20{,}50{,}200$, що містять оцінки параметрів $a_0 = 1$ та $a_1 = 3$ для моделі регресійних помилок з гамма-розподілом (при $\alpha = 2$).



а) *б)* *в)*

Рисунок 4.2 – Гаусові ймовірнісні графіки, що апроксимують експериментальні значення оцінок параметра a_0 лінійної регресії при Гамма-розподілі помилок: *а) N=20; б) N=50; в) N=200*



а) *б)* *в)*

Рисунок 4.3 – Гаусові ймовірнісні графіки, що апроксимують експериментальні значення оцінок параметра a_1 лінійної регресії при Гамма-розподілі помилок: *а) N=20; б) N=50; в) N=200*

Аналіз цих даних та інших результатів експериментальних досліджень показує, що у випадку коли розподіл регресійних помилок суттєво відрізняється від гаусової моделі (великі абсолютні значення асиметрії та ексцесу), нормалізація розподілу оцінок спостерігається лише для досить великого розміру початкових елементів вибірки, тобто $N > 100$. При цьому важливим фактом є то, що, не зважаючи на наявність нелінійних перетворювань (при обчисленні квадратичної статистики для знаходження оцінок ММПл при $S = 2$), швидкість нормалізації емпіричного розподілу ММПл-оцінок не зменшується, а в окремих випадках, навіть перевищує динаміку нормалізації МНК-оцінок.

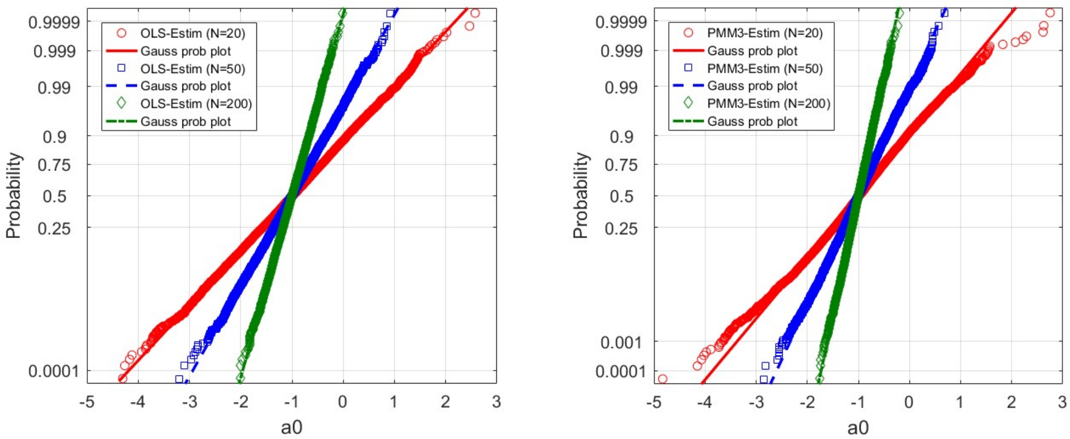
4.2.1.3. Результати застосування методу максимізації полінома при степені *S*=3 (симетричний випадок)

При використанні ММПл степені $S=3$ в якості симетрично-розподіленої випадкової похибки використано випадкові величини, що мають рівномірний, трапецієподібний, трикутний та розподіл Лапласа. Як і у попередньому випадку значення параметрів випадкової складової регресійної моделі вважалися апріорно невідомими, а для адаптивного оцінювання використовувалися апостеріорні оцінки парних моментів до 6-го порядку регресійних МНК-залишків.

В таблиці 4.3 представлено сукупність результатів статистичного моделювання оцінювання параметрів регресійної моделі виду (4.2) із симетрично-розподіленими помилками, які отримані для $M = 10^4$ експериментів.

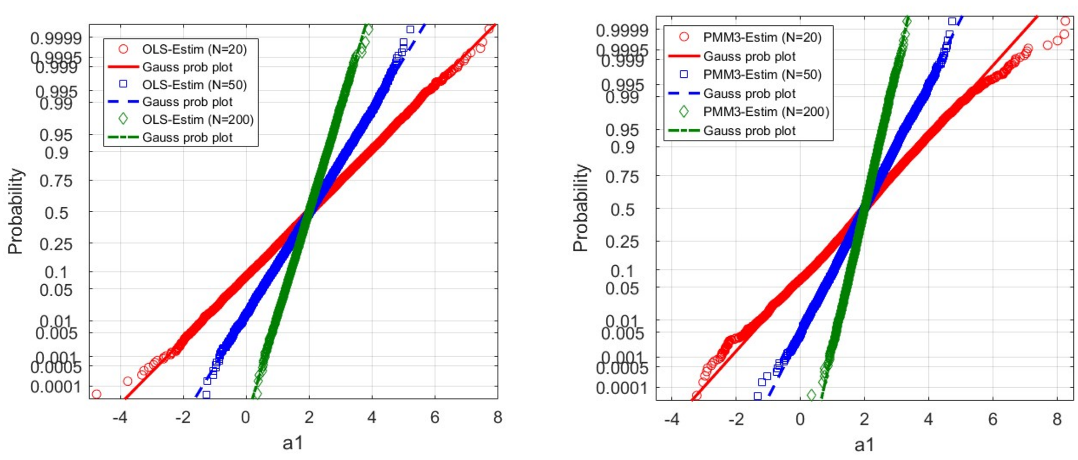
Аналіз результатів табл.4.3 показує, що відносна ефективність ММПл-оцінок є приблизно однаковою для обох оцінюваних параметрів a_0 і a_1, що відповідає теоретичним висновкам розділу 3.1. Ступінь ефективності ММПл відносно МНК залежить від властивостей регресійних помилок і в більшій мірі проявляється для розподілів, що мають плосковершиний характер та від’ємні значення коефіцієнту ексцесу γ_4. Зокрема, для моделей регресійних помилок у яких $\gamma_4 <$ -1 зменшення дисперсії становить від 15% (для малих вибірок $N = 20$) до кількох разів (з ростом N або при зменшенні γ_4).

Крім того, за результатами статистичного моделювання на рис.4.4 та рис.4. наведено приклади, що демонструють розподіл експериментальних значень оцінок МНК та ММПл (при $S = 3$) компонентів векторного параметру $\boldsymbol{\theta} = \{a_0, a_1\}$. У цих прикладах результуючі графіки побудовані при $M = 10^4$ експериментах для розміру вибірок $N = 20,50,200$, що містять оцінки параметрів $a_0 = -1$ та $a_1 = 2$ для моделі регресійних помилок з Трапецієвидним розподілом (при $\beta = 0.5$).



a) *b)*

Рисунок 4.4 – Гаусові ймовірні графіки, що апроксимують експериментальні значення оцінок параметрів *a_0* лінійної регресії при трапецієвидному розподілу помилок: *а*) МНК-оцінки; *б*) ММПл-оцінки



a) *b)*

Рисунок 4.5 – Гаусові ймовірні графіки, що апроксимують експериментальні значення оцінок параметрів *a_1* лінійної регресії при трапецієвидному розподілу помилок: *а*) МНК-оцінки; *б*) ММПл-оцінки

Таблиця 4.3 – Результати статистичного моделювання Монте-Карло оцінювання параметрів лінійної регресії при симетрії регресійних помилок.

Розподіл		Теоретичне значення параметрів			Результати статистичного моделювання					
		γ_4	γ_6	$g_{\frac{PMM2}{OLS}}$	$\hat{g}_{\frac{ММПл3}{МНК}}$					
					$N = 20$		$N = 50$		$N = 200$	
					a_0	a_1	a_0	a_1	a_0	a_1
Арксинусний		-1.3	8.2	0.2	0.51	0.53	0.31	0.32	0.22	0.23
Рівномірний		-1.2	6.9	0.3	0.67	0.67	0.44	0.45	0.33	0.33
Трапеціє-видний	$\beta = 0.75$	-1.1	6.4	0.36	0.70	0.70	0.50	0.50	0.40	0.40
	$\beta = 0.5$	-1	5	0.55	0.85	0.84	0.68	0.69	0.57	0.58
	$\beta = 0.25$	-0.7	2.9	0.76	1.01	1.00	0.89	0.90	0.80	0.81
Трикутний		-0.6	1.7	0.84	1.08	1.09	0.97	0.97	0.88	0.89
Лапласа		3	30	0.86	1.47	1.50	0.86	0.87	0.85	0.85

Аналіз цих даних та інших результатів експериментальних досліджень показує, що для похибки регресії, розподіл якої суттєво відрізняється від гаусової моделі, нормалізація розподілу оцінок спостерігається лише для досить великого розміру вихідних елементів вибірки. Це пояснюється наявністю додаткових нелінійних перетворень (квадратичної та кубічної статистики при знаходженні ММПл-оцінок для $S = 3$), що дещо уповільнює динаміку нормалізації емпіричного розподілу ММПл-оцінок порівняно в МНК-оцінками.

4.2.2. Статистичне моделювання оцінювання параметрів поліноміальної регресії

Для реалізації статистичного моделювання оцінювання параметрів поліноміальної регресії використано квадратичну степеневу залежність виду

$$y_v = a_0 + a_1 x_v + a_2 x_v^2 + \xi_v,\ v = \overrightarrow{1, N}, \tag{4.3}$$

де задані істинні значення інформативних параметрів $a_0 = 1$, $a_1 = 2$ і $a_2 = -4$ детермінованої складової регресійної моделі вважаються апріорно невідомими і підлягають оцінюванню.

В якості моделей випадкових складових поліноміальної регресії використані ті ж самі розподіли, що при моделюванні оцінювання лінійної регресії моделі. Сукупність експериментальних значень коефіцієнтів відношення дисперсії, які отримані при $M = 10^4$ експериментів, представлено в табл. 4.3 та табл. 4.4.

Наведені в табл.4.3 і табл.4.4 результати свідчать, що експериментальні значення величини відношення дисперсії ММПл до дисперсії МНК оцінок в певній мірі відрізняються від теоретичних. Але як вже зазначалося раніше, це пояснюється тим, що теоретичні значення коефіцієнтів зменшення дисперсії отримані для асимптотичного випадку, тобто при $N \to \infty$. Крім того, на точність отримання адаптивних ММПл-оцінок впливає невизначеність апостеріорних оцінок параметрів (моментів відповідного порядку) регресійних залишків, яка також зменшується при зростанні обсягу вибірок. Проте в цілому аналіз представлених даних свідчить про адекватність аналітичних розрахунків, оскільки максимальна відносна розбіжність між теоретичними і експериментальними значеннями коефіцієнтів відношення дисперсії при $N = 200$ не перевищує до 2-3%.

Іншим важливим результатом статистичного моделювання є підтвердження теоретичної властивості, яка полягає у тому що відносне зростання точності ММПл-оцінок є однаковим для усіх складових векторного параметру (коефіцієнтів детермінованої складової поліноміальної регресії). Більше того, воно практично не залежить від порядку Q детермінованої складової регресійної моделі. Ця властивість також підтверджується суттєвою кореляцією із результатами моделювання оцінювання параметрів лінійної однофакторної регресії, представленими у підрозділі 4.2.1.

Таблиця 4.4 – Результати статистичного моделювання Монте-Карло оцінювання параметрів квадратичної регресії.

Розподіл		Теоретичні значення			Результати статистичного моделювання Монте-Карло								
		γ_3	γ_4	$g_{\frac{PMM2}{OLS}}$	$\hat{g}_{\frac{PMM2}{OLS}}$								
					$N = 50$			$N = 100$			$N = 200$		
					a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2
Гамма	$\alpha = 0.5$	2.83	12	0.42	0.44	0.38	0.37	0.43	0.35	0.35	0.43	0.36	0.36
Експоненціальний (Гамма, $\alpha = 1$)		2	6	0.5	0.57	0.53	0.54	0.54	0.46	0.46	0.53	0.47	0.47
Гамма	$\alpha = 2$	1.41	3	0.6	0.64	0.64	0.63	0.63	0.59	0.6	0.63	0.59	0.59
	$\alpha = 4$	1	1.5	0.71	0.75	0.74	0.75	0.73	0.72	0.71	0.74	0.72	0.71
Логнормальний $\sigma^2 = 0.1, \mu = 1$		1	1.86	0.74	0.77	0.76	0.76	0.75	0.74	0.74	0.74	0.75	0.75
Вейбула $a = 1$, $b = 2$		0.63	0.25	0.82	0.91	0.85	0.85	0.84	0.84	0.84	0.83	0.82	0.82

Таблиця 4.5 – Результати статистичного моделювання Монте-Карло оцінювання параметрів квадратичної регресії (симетричний випадок).

Розподіл		Теоретичні значення			Результати статистичного моделювання Монте-Карло								
		γ_4	γ_6	$g_{\frac{PMM3}{OLS}}$	$\hat{g}_{\frac{PMM3}{OLS}}$								
					$N = 50$			$N = 100$			$N = 200$		
					a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2
Рівномірний		-1.2	6.9	0.3	0.55	0.53	0.54	0.41	0.4	0.4	0.35	0.35	0.35
Трапеція	$\beta = 0.75$	-1.1	6.4	0.36	0.61	0.58	0.58	0.47	0.47	0.46	0.41	0.41	0.41
	$\beta = 0.5$	-1	5	0.55	0.75	0.73	0.74	0.66	0.66	0.66	0.59	0.58	0.58
	$\beta = 0.25$	-0.7	2.9	0.76	0.95	0.94	0.94	0.86	0.86	0.87	0.81	0.82	0.82
Трикутний		-0.6	1.7	0.84	1	1	1	0.95	0.94	0.94	0.9	0.9	0.89
Лапласа		3	30	0.86	1.4	1.66	1.72	0.86	0.86	0.85	0.83	0.82	0.82

4.2.3. Статистичне моделювання оцінювання параметрів регресії із помилками, що мають експоненціальний степеневий розподіл

Порівняльний аналіз теоретичної ефективності поліноміальних оцінок із ефективністю оцінок максимальної правдоподібності на прикладі моделі ЕСР помилок, здійснений у підрозділі 3.3, також потребує експериментальної верифікації. Проте специфіка цієї задачі потребувала модифікації програмних засобів статистичного моделювання. Його реалізація була побудована на основі використання спеціалізованого модуля «normalp package» для мови R, що містить набір функцій для генерації випадкових величин з ЕСР, а також статистичного оцінювання параметрів лінійної регресії з ЕСР помилками методом максимальної правдоподібності [31].

Оскільки як вже зазначалося раніше теоретичні значення коефіцієнтів відношення дисперсії оцінок є однаковими для всіх компонентів векторного параметру регресії (що також експериментально підтверджено результатами підрозділів 4.2.1 та 4.2.2), то для спрощення представлення і аналізу результатів використані геометричні усереднення виду:

$$\hat{g}_{\frac{\text{ММП}}{\text{МНК}}} = \left(\prod_{p=0}^{Q-1} \frac{\hat{\sigma}^2_{(a_p)\text{ММП}}}{\hat{\sigma}^2_{(a_p)\text{МНК}}} \right)^{\frac{1}{Q}}, \qquad \hat{g}_{\frac{\text{ММПлз}}{\text{МНК}}} = \left(\prod_{p=0}^{Q-1} \frac{\hat{\sigma}^2_{(a_p)\text{ММПлз}}}{\hat{\sigma}^2_{(a_p)\text{МНК}}} \right)^{\frac{1}{Q}}, \tag{4.4}$$

де $\hat{\sigma}^2_{(a_p)\text{МНК}}$, $\hat{\sigma}^2_{(a_p)\text{ММП}}$, $\hat{\sigma}^2_{(a_p)\text{ММПлз}}$ – усереднені на основі M експериментів значення дисперсії оцінок параметрів, що отримуються із застосуванням відповідного методу.

Зазначимо, що статистичне моделювання здійснювалося для двох принципово відмінних ситуацій з точки зору наявності апріорної інформації щодо характеристик випадкової складової регресійної моделі. У першому (ідеалізованому з практичної точки зору) випадку величина параметрів ЕСР розподілу вважалися апріорно відомими і їх значення використовувалися при знаходженні ММП-оцінок. Ця інформація також дозволяла обчислювати на основі

(3.13) значення парних моментів до 6-го порядку, необхідні для знаходження ММПл-оцінок при степені $S = 3$. Очевидно, що в реальних умовах подібна апріорна інформація практично відсутня. Тому у другому випадку застосовується адаптивний підхід. Як вже було зазначено вище, в його основі лежить гіпотеза про схожість розподілу випадкової складової регресійної моделі і регресійних залишків, які отримуються при застосуванні лінійних методів оцінювання, зокрема, МНК. Таким чином, замість апріорних значень параметрів, необхідних як для застосування ММП так і ММПл, можуть бути використані їх апостеріорні оцінки, знайдені по регресійним МНК-залишкам.

Для здійснення статистичного моделювання оцінювання параметрів в якості детермінованої складової регресійної моделі використано двофакторну лінійну залежність

$$y_v = a_0 + a_1 x_{1,v} + a_2 x_{2,v} + \xi_v,\ v = \overline{1, N}, \tag{4.5}$$

із вектором параметрів $\boldsymbol{\theta} = \{1{,}3, -1\}$, які вважалися невідомими і підлягали оцінюванню. В табл. 4.5 та табл. 4.6 представлена сукупність результатів статистичного моделювання для різних значень величини параметра форми ЕСР $\beta = 1{,}2{,}3{,}5{,}10$ та обсягів вибірки $N = 20{,}50{,}100{,}200$, що отримані при $M = 10^4$ багаторазових експериментів.

Таблиця 4.6 – Результати статистичного моделювання Монте-Карло оцінювання параметрів лінійної регресії із ЕСР помилками за наявності апріорної інформації

β	Результати статистичного моделювання									
	$\hat{g}_{\frac{ММПл3}{МНК}}$				$g_{\frac{ММПл3}{МНК}}$	$\hat{g}_{\frac{ММП}{МНК}}$				$g_{\frac{ММП}{МНК}}$
	N					N				
	20	50	100	200		20	50	100	200	
10	0.98	0.56	0.46	0.44	0.4	0.74	0.49	0.39	0.36	0.31
5	0.88	0.72	0.66	0.64	0.61	0.84	0.71	0.65	0.63	0.61
3	0.95	0.91	0.9	0.89	0.89	0.95	0.91	0.9	0.88	0.88
2	1	1	1	1	1	1	1	1	1	1
1	1.57	1.21	1.19	0.89	0.85	0.91	0.73	0.66	0.61	0.5

Таблиця 4.7 – Результати статистичного моделювання Монте-Карло оцінювання параметрів лінійної регресії із ЕСР помилками за відсутності апріорної інформації

<table>
<tr><td rowspan="4">β</td><td colspan="10">Результати статистичного моделювання</td></tr>
<tr><td colspan="4">$\hat{g}_{\frac{PMM3}{OLS}}$</td><td rowspan="3">$g_{\frac{PMM3}{OLS}}$</td><td colspan="4">$\hat{g}_{\frac{MLE}{OLS}}$</td><td rowspan="3">$g_{\frac{MLE}{OLS}}$</td></tr>
<tr><td colspan="4">N</td><td colspan="4">N</td></tr>
<tr><td>20</td><td>50</td><td>100</td><td>200</td><td>20</td><td>50</td><td>100</td><td>200</td></tr>
<tr><td>10</td><td>0.87</td><td>0.61</td><td>0.49</td><td>0.43</td><td>0.4</td><td>1.06</td><td>0.64</td><td>0.47</td><td>0.39</td><td>0.31</td></tr>
<tr><td>5</td><td>0.98</td><td>0.8</td><td>0.69</td><td>0.65</td><td>0.61</td><td>1.15</td><td>0.86</td><td>0.75</td><td>0.7</td><td>0.61</td></tr>
<tr><td>3</td><td>1.11</td><td>0.97</td><td>0.94</td><td>0.91</td><td>0.89</td><td>1.26</td><td>1.08</td><td>0.99</td><td>0.93</td><td>0.88</td></tr>
<tr><td>2</td><td>1.35</td><td>1.15</td><td>1.07</td><td>1.03</td><td>1</td><td>1.22</td><td>1.06</td><td>1.04</td><td>1.02</td><td>1</td></tr>
<tr><td>1</td><td>3.2</td><td>1.17</td><td>0.82</td><td>0.82</td><td>0.85</td><td>0.91</td><td>0.72</td><td>0.66</td><td>0.61</td><td>0.5</td></tr>
</table>

Аналіз теоретичних та експериментальних значень коефіцієнтів відношення дисперсії показує, те що існує певна кореляція аналітичних розрахунків та результатів, отриманих шляхом статистичного моделювання. Виникнення різниці обумовлено тим, що вирази, які описують дисперсії ММП і ММПл-оцінок, отримані для асимптотичного випадку (при $N \to \infty$). Дані табл. 4.5 підтверджують, що зі збільшенням обсягу вибірки розбіжність між теоретичними та експериментальними даними зменшується.

Експериментальні результати також підтверджують зазначену раніше тезу, що на точність отримання ММП і ММПл-оцінок суттєво впливає фактор наявності/відсутності апріорної інформації щодо властивостей моделі помилок (параметрів ЕСР для ММП і парних центральних моментів до 6-го порядку для ММПл). Дані табл. 4.6 відображають важливий факт, що при відсутності апріорної інформації про властивості ЕСР для плосковершинних розподілів ($p > 2$) ефективність ММПл-оцінок може бути вищою за ММП-оцінки. Це можна пояснити тим, що в цих умовах вплив невизначеності величини параметру форми ЕСР на ММП є більш суттєвим ніж вплив невизначеності центральних моментів на ММПл. При цьому відносне зменшення дисперсії оцінок ММПл (порівняно з

ММП) залежить як від параметра форми так і від обсягу вибіркових значень і особливо є суттєвим саме при малих вибірках.

Подібні результати є особливо важливими з практичної точки зору, оскільки для переважної більшості реальних ситуацій апріорна інформація про істинні значення параметрів випадкової складової регресійної моделі відсутня. Тому саме для такої ситуації на рис.4.1 зображено границі, що розмежовують області найбільшої ефективності (на основі критерію мінімуму дисперсії) при застосуванні різних методів оцінювання: МНК, ММП та ММПл. Ці області також отримані шляхом статистичного моделювання методом Монте-Карло (для $M = 10^4$ експериментів) при різних значеннях параметра форми ЕСР β та обсягів вибірки N.

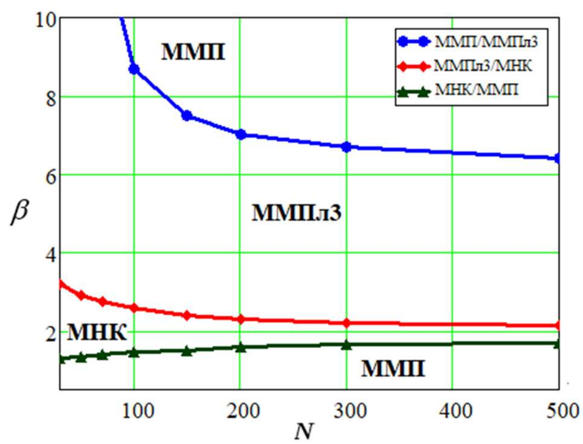


Рисунок 4.6 – Області ефективності методів знаходження оцінок параметрів лінійної регресії із ЕСР помилками

На основі сукупності отриманих результатів статистичного моделювання можна зробити наступні висновки:

- При розподілі регресійних помилок близькому до гаусового ($\beta = 2$) найбільш ефективним (як з точки зору точності так і простоти реалізації) є оцінки методу найменших квадратів.
- Для лептокуртичних (гостровершинних) розподілів із важкими хвостами загалом більш ефективним є застосування методу максимальної

правдоподібності. Для малих вибірок межа ефективності лежить біля $\beta \approx 1.3$ і плавно в асимптотиці (при зростанні N) збільшується до $\beta \approx 1.7$.

- Для платокуртичних (плосковершинних) розподілів для достатньо широкого діапазону значень параметру форми β найкращу ефективність демонструє метод максимізації поліномів. Для малих вибірок нижня межа ефективності (між МНК і ММПл) приблизно починається від $\beta \approx 3$ і параболічно в асимптотиці (при зростанні N) прямує до $\beta \approx 2$.
- Для великих значень параметру форми $\beta > 6$ (що відповідає ймовірнісним розподілам близьким для рівномірного закону) при великих обсягах вибірок більш ефективним знову є метод максимальної правдоподібності. Межа, яка розділяє ММП і ММПл ще в більшій мірі залежить від обсягу вибіркових значень, проте різниця в їх точності (відношенні дисперсії оцінок) складає всього одиниці відсотків.

4.2.4. Статистичне моделювання оцінювання параметрів нелінійної регресії

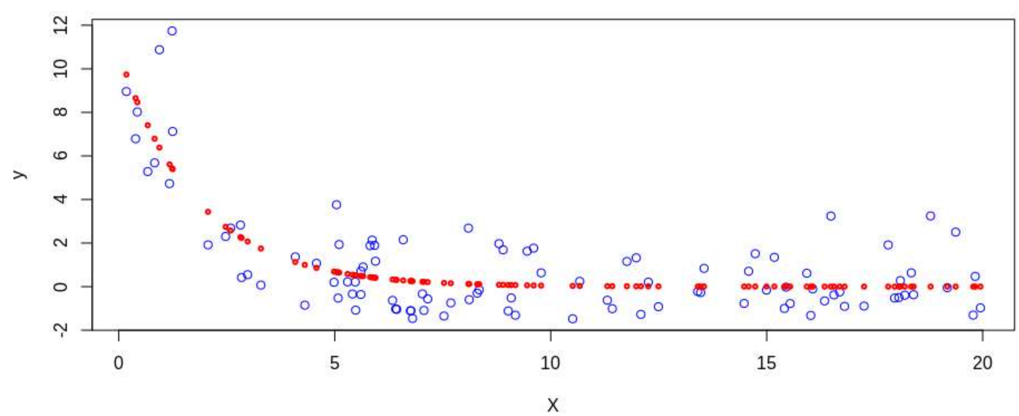
В даному підрозділі представлені результати порівняльного аналізу ефективності методу максимізації поліномів при степені *S=2* та квадратичної модифікації методу найменших квадратів (SLS) [52–54]. Дослідження проведені на прикладі оцінювання параметрів нелінійних регресійних моделей загального виду (2.1) для двох типів нелінійних детермінованих функцій

$$f_1(X,\vec{\theta}) = \theta_1 + exp(\theta_2 X), \qquad (4.6, a)$$

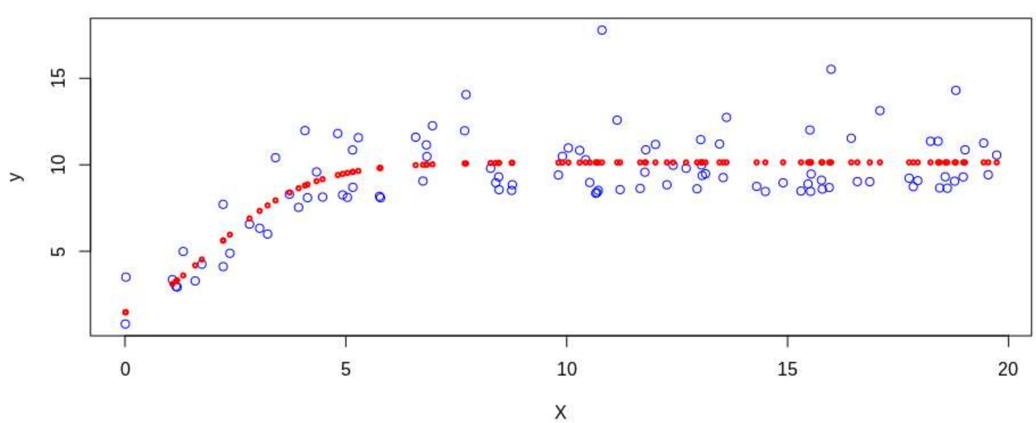
$$f_2(X,\vec{\theta}) = \theta_1/[1 + exp(\theta_2 + \theta_3 X)]. \qquad (4.6, б)$$

Математична постановка задачі оцінювання передбачає, що стохастична складова (випадкові помилки) має суттєво відмінний від гаусового асиметричний розподіл, проте апріорна інформація про параметри цього розподілу відсутня.

Для порівняльного аналізу ефективності (за критерієм відношення дисперсії оцінок) різних методів використано опосередкований підхід. Він полягав у початковому обчисленні шляхом статистичного моделювання методом Монте-Карло дисперсії МНК-оцінок і ММПл2-оцінок, що отримувались при однакових вихідних умовах (типах детермінованих функцій, значеннях їх параметрів, властивостях регресійних помилок, обсягах вибіркових даних) із відповідними експериментальними даними роботи [52]. Для нелінійної функції 1-го типу (4.6, *а*) використані значення параметрів $\theta_1 = 10$ і $\theta_2 = -0.6$ (див. рис.4.7, *а*), а для нелінійної функції 2-го типу (4.6, *б*) використані значення параметрів $\theta_1 = 10, \theta_2 = 1.5$ і $\theta_3 = -0.8$ (див. рис.4.7, *б*). В якості моделі помилок використано центрований розподіл хі-квадрат $\xi = (\chi(3) - 3)/\sqrt{3}$з параметром масштабу $\sigma^2 = 2$.



а)



б)

Рисунок 4.7 – Приклади нелінійних регресійних залежностей із асиметричним розподілом помилок

Серед інших дані роботи [52] містять оціночні значення дисперсії МНК-оцінок та SLS-оцінок, відношення яких і було використано для порівняння із відповідними відношенням дисперсії МНК-оцінок та ММПл2-оцінок, отриманими в рамках даного дослідження. На рис.4.8 (для регресійної моделі (4.6, *а*)) та рис.4.9 (для регресійної моделі (4.6, *б*)) наведена графічна візуалізація сукупності результатів проведеного статистичного моделювання

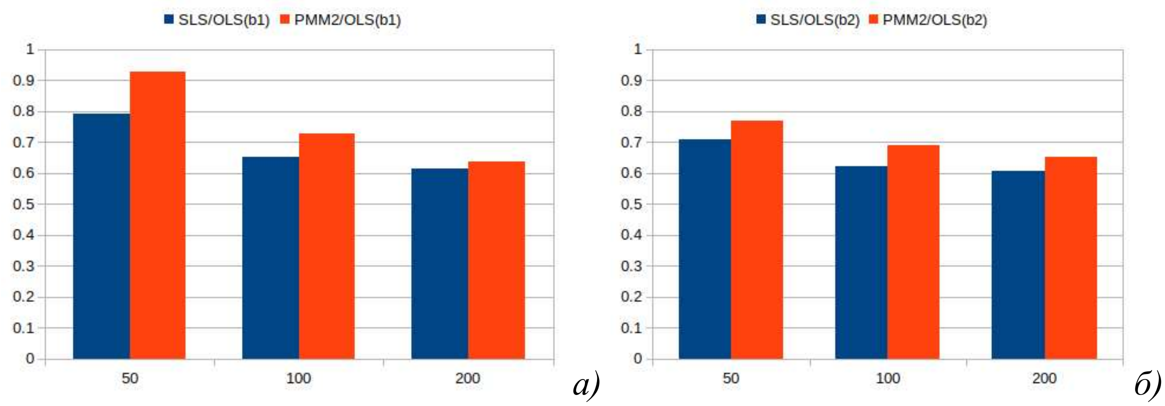
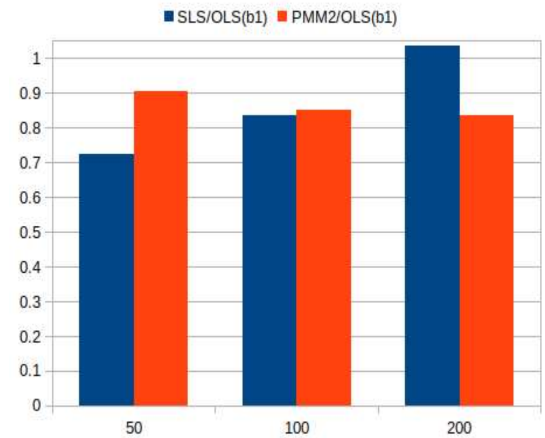
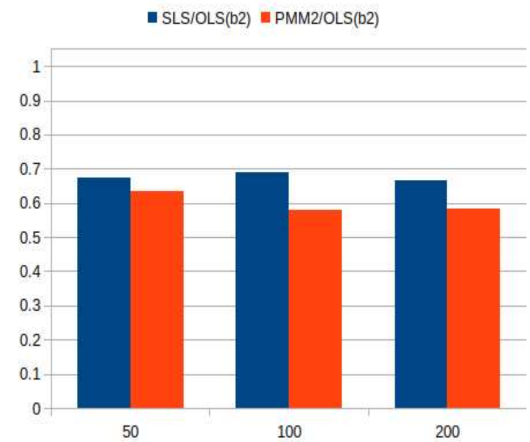


Рисунок 4.8 – Коефіцієнти відношення дисперсії оцінок нелінійної моделі 2-го типу

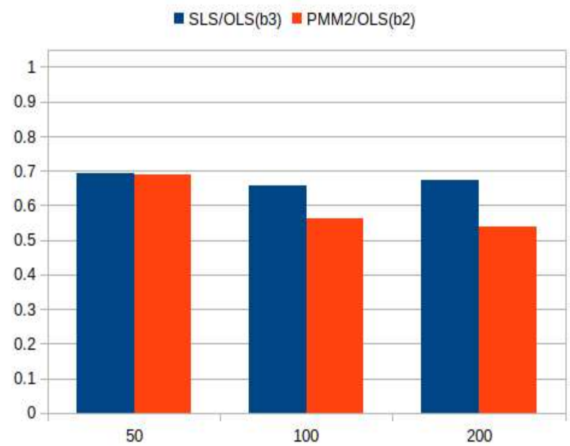
Аналіз наведених результатів дає можливість зробити висновки про те, що ефективність (величина дисперсії оцінок) методів SLS та ММПл2 у цілому перевищує точність класичного МНК при асиметричному характері розподілу випадкової складової регресійних моделей. Хоча ступінь виграшу має достатньо різний характер в залежності від вигляду регресійної моделі, типу їх параметрів та обсягу вибірковий значень. Зокрема, для моделі (4.6, *а*) відносна (щодо МНК) точність SLS-оцінок дещо вища (на 5-10%) за ММПл2-оцінки. Проте для моделі (4.6, *б*) навпаки, дещо точнішими (на 5-20%) є оцінки ММПл2. Важливим фактом також є те, що із зростанням обсягу вибірки відносна точність ММПл2-оцінок прямує до теоретично розрахованої величини.



а)



б)



в)

Рисунок 4.9 – Коефіцієнти відношення дисперсії оцінок нелінійної моделі 2-го типу

У цілому отримані результати ще раз підтверджують потенційну ефективність застосування апарату стохастичних поліномів Кунченка для вирішення практичних статистичних задач регресійного аналізу навіть в умовах апріорної невизначеності.

4.3. Прикладне застосування розроблених програмних засобів при вирішенні задачі регресійного аналізу

Як приклад для тестування запропонованої адаптивної процедури знаходження ММПл-оцінок параметрів нелінійної регресійної моделі загального виду (2.1) використаний набір даних отриманих при дослідженні технологічних характеристик розробленого FDM 3D принтера. Дані експерименти були здійснені в рамках держбюджетної НДР «Розробка технології та пристроїв адитивного виробництва індивідуальних хірургічних імплантатів та протезів з біосумісних полімерних матеріалів» (№ держ. реєстрації 0117U000937) [110].

Зокрема були використані дані дослідження механічних характеристик системи подачі філаменту та реологічних характеристик екструдера 3D принтера. Цей набір даних представляє собою залежність швидкості виходу пластикового філаменту через сопло екструдера у вигляді розплаву від керуючих параметрів, зокрема, заданої швидкості подачі філаменту в твердому стані та температури гарячої частини екструдера для різних термопластичних матеріалів. Аналіз отриманих експериментальних даних показав, що математична модель таких процесів достатньо адекватно може бути описана нелінійної регресійною залежністю у вигляді суми гармонічної і лінійної складової виду:

$$y_v = A\,sin(Fx_v + \phi) + a_0 + a_1 x_v + \xi_v, \tag{4.7}$$

де інформативні параметри $\boldsymbol{\theta} = \{A, F, \varphi, a_0, a_1\}$, що підлягають оцінюванню.

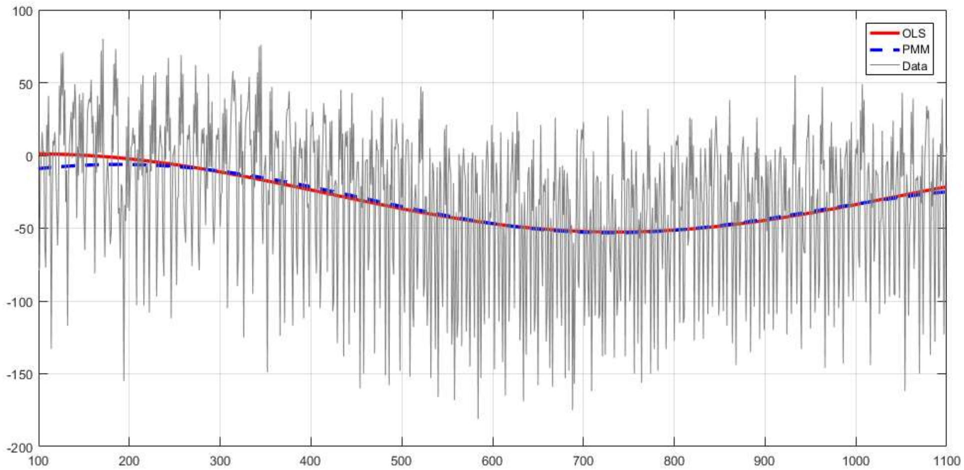
В табл. 4.7 узагальнені результати 3-х експериментів, які отримані для різних типів термопластичних матеріалів із використанням програмних засобів, реалізованих в системі MATLAB.

На рис.4.10 візуалізовано статистичні дані швидкості виходу пластикового філаменту через сопло екструдера (цільова змінна y) від швидкості подачі (регресор x) для $N = 1000$ (з урахуванням пропущених) пар даних та відновлені детерміновані функції регресійних залежностей (4.7) із оціночними значеннями інформативних параметрів, знайденими із застосуванням класичного МНК і квадратичного ММПл (при S=2).

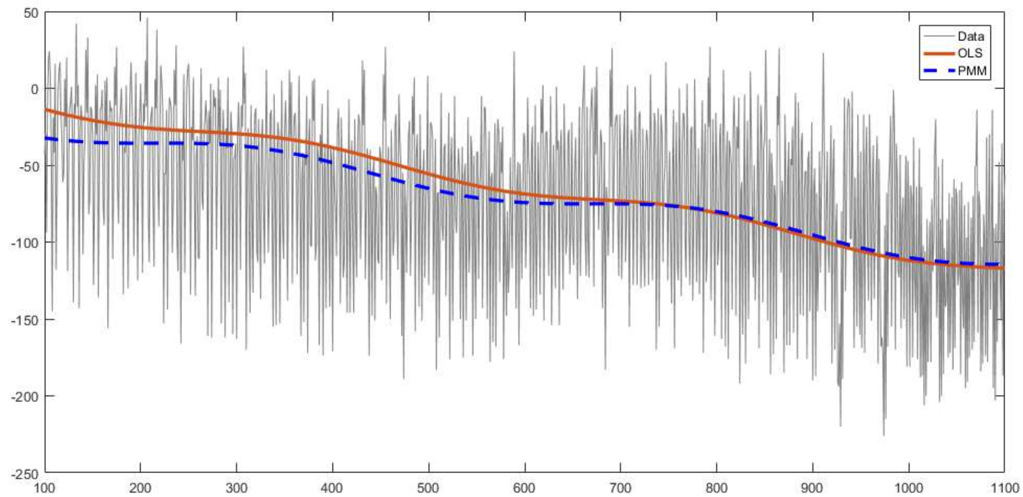
Асиметричність розподілу регресійних помилок моделі (4.7) візуально видно на рис.4.11, де представлений емпіричний розподіл частот регресійних МНК-залишків для різних експериментів.

Таблиця 4.8 – Результати експериментальних досліджень оцінювання параметрів нелінійної регресії

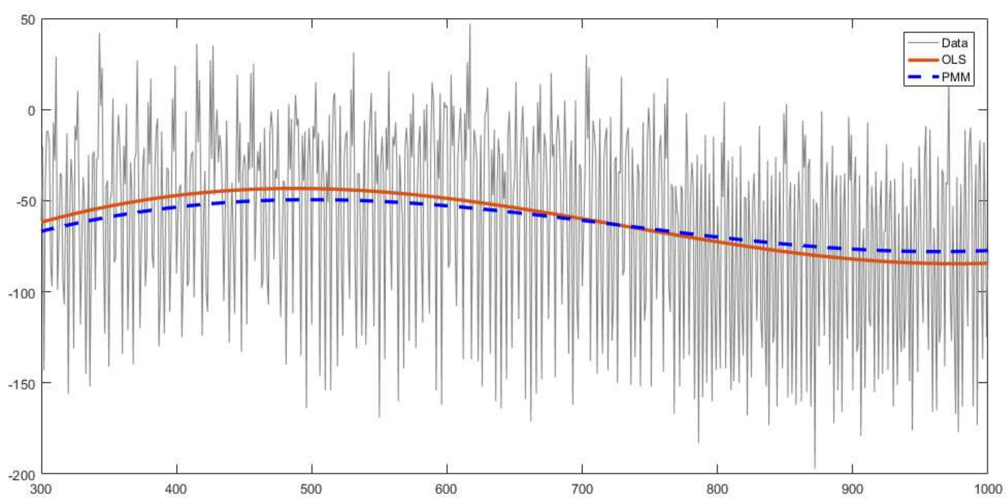
Оціночні значення		Експеримент № 1	Експеримент № 2	Експеримент № 3
МНК-оцінки інформативних параметрів	$\hat{A}^{МНК}$	25.52	4.87	60.07
	$\hat{F}^{МНК}$	0.005	0.0152	0.0043
	$\hat{\phi}^{МНК}$	7.35	8.56	0.0001
	$\hat{a}_0{}^{МНК}$	-24.38	0.003	-158.11
	$\hat{a}_1{}^{МНК}$	-0.0004	-0.1071	-0.129
Характеристики тесту Ярки-Бера	$JBSTAT$	81.5	35.17	47.81
	CV	5.9		
Параметри МНК-залишків	$\hat{\mu}_2$	2.32	3.02	2.94
	$\hat{\gamma}_3$	-0.69	-0.34	-0.52
	$\hat{\gamma}_4$	-0.18	-0.61	-0.71
ММПл-оцінки інформативних параметрів	$\hat{A}^{ММПл2}$	19.32	6.4	66.78
	$\hat{F}^{ММПл2}$	0.006	0.138	0.004
	$\hat{\phi}^{ММПл2}$	6.53	9.62	0.348
	$\hat{a}_0{}^{ММПл2}$	-22.36	-17.25	-181.95
	$\hat{a}_1{}^{ММПл2}$	-0.016	-0.087	0.162
Коефіцієнти зменшення дисперсії	$\hat{g}_{ММПл2/МНК}$	0.74	0.91	0.79



а)

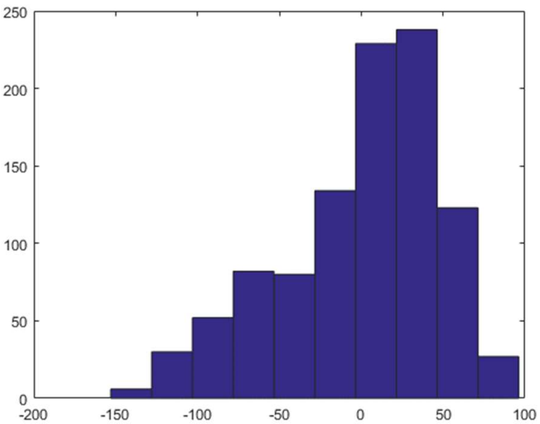


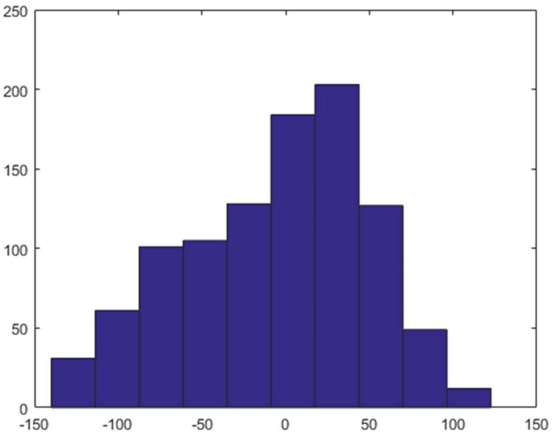
б)



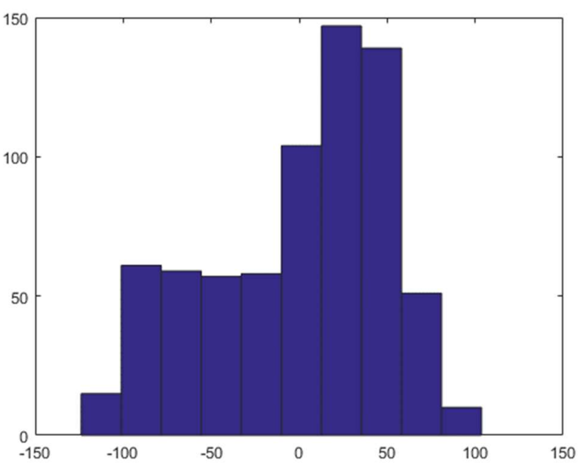
в)

Рисунок 4.10 – Експериментальні дані та оцінки регресії, на основі МНК і ММПл: *а*) експеримент №1; *б*) експеримент №2; *в*) експеримент №3





а) *б)*



в)

Рисунок 4.11 – Емпіричний розподіл частот регресійних МНК-залишків: *а*) експеримент №1; *б*) експеримент №2; *в*) експеримент №3

Гіпотеза про гаусовість МНК-залишків також спростовується вбудованим в MATLAB тестом Харке-Бера, який базується на аналізі величини кумулятивних коефіцієнтів асиметрії та ексцесу [90]. Оціночні значення статистики $JBSTAT$ та величина порогу CV при рівні значущості $\alpha_0 = 0.05$ відповідної гіпотези представлені у табл.4.7.

Всі ці фактори підтверджують доцільність застосування адаптивного оцінювання із використанням ММПл при степені *S=2.* Оціночні значення кумулянтних коефіцієнтів асиметрії $\hat{\gamma}_3$ та ексцесу $\hat{\gamma}_4$ регресійних МНК-залишків, отримані для достатньо великого об'єму вихідних даних ($N = 1000$) у відповідності до (3.6), дозволяють оцінити величини зменшення дисперсії $\hat{g}_{ММПл2/МНК}$ для кожного із експериментів.

Аналіз сукупності отриманих результатів експериментів № 1-3 підтверджує можливість і доцільність застосування методу максимізації поліномів для знаходження оцінок параметрів нелінійних регресійних моделей за умови негаусового розподілу помилок. Проведені розрахунки із використання реальних статистичних даних залежності, отриманих при в процесі дослідження режимів роботи 3D принтера показали, що застосування ММПл дозволяє зменшити на 10-25 % (порівняно із МНК) дисперсію оцінок інформативних параметрів нелінійної регресійної моделі у вигляді суми гармонічної і лінійної функцій.

Висновки

В рамках дисертаційного дослідження на основі використання засобів системи комп’ютерної математики та моделювання MATLAB та мови R створено програмний комплекс який надає можливість вирішувати широке коло задач, пов’язаних із процесами оцінювання інформативних параметрів регресійних моделей в умовах негаусового розподілу їх помилок. Функціонал розроблених програмних засобів включає вирішення двох типів задач. По перше, це безпосереднє вирішення задачі знаходження адаптивних оцінок параметрів різних типів регресійних залежностей (лінійних, поліноміальних, нелінійних) на основі використання методу максимізації поліномів. По друге, це задачі, пов’язані із реалізацією комп’ютерного статистичного моделювання з метою порівняльного аналізу ефективності поліноміальних оцінок з оцінками класичних методів найменших квадратів та максимальної правдоподібності.

Сукупність отриманих результатів статистичного моделювання на основі багаторазових випробувань методом Монте-Карло у цілому підтверджують теоретично доведену у розділі 3 ефективність поліноміальних оцінок відносно лінійних оцінок методу найменших квадратів для всіх проаналізованих типів регресійних моделей. На прикладі моделі помилок з експоненціальним степеневим розподілом також показано, що при відсутності апріорної інформації про значення параметрів регресійних помилок (характерної для реальних ситуацій) адаптивні оцінки методу максимізації поліномів можуть бути більш точними, порівняно із оцінками максимальної правдоподібності для достатньо широкого діапазону плосковершинних симетричних розподілів.

Розроблені програмні засоби були успішно застосовані при вирішенні прикладної задачі в рамках дослідження механічних характеристик системи подачі філаменту та реологічних характеристик екструдера 3D принтера. Результати застосування запропонованих моделей та методів оцінювання дозволили забезпечити зменшення на 10-25 % величини дисперсії оцінок інформативних параметрів нелінійних регресійних моделей.

ВИСНОВКИ

У дисертаційній роботі із застосуванням апарату стохастичних поліномів Кунченка та моментно-кумулянтного опису вирішена важлива науково-технічна задача – розробка методів математичного і комп'ютерного моделювання процесів адаптивного оцінювання параметрів регресійних залежностей, що дозволило підвищити точність процесів оцінювання на основі урахуванні характеристик негаусових регресійних помилок та створити алгоритмічні основи і програмні засоби їх реалізації. Зокрема, отримано такі теоретичні і практичні результати.

1. Аналіз задач підвищення ефективності процесів оцінювання інформативних параметрів регресії за умови негаусового розподілу регресійних помилок показав, що перспективним напрямком є підхід, який ґрунтується на застосуванні статистик вищих порядків та методу максимізації полінома. Перевагою цього підходу є можливість забезпечення адаптивності на основі урахування негаусовості випадкової складової регресійних моделей, що поєднується із простотою їх застосування. Це дозволяє створити нові математичні та алгоритмічні засади програмних засобів моделювання та статистичного опрацювання.

2. Досліджено властивості та обґрунтовано застосування апарату стохастичних поліномів Кунченка для модифікації регресійних моделей, що надало можливість коректно та компактно описувати ступінь негаусовості регресійних помилок та створило передумови для синтезу реалізаційно-простих обчислювальних алгоритмів адаптивного оцінювання, які враховують негаусовий характер статистичних даних.

3. Із використанням апарату стохастичних поліномів Кунченка та моментного опису здійснено синтез моделей, методів та обчислювальних алгоритмів адаптивного оцінювання параметрів регресійних моделей лінійного, поліноміального і нелінійного типу. Показано, що загальна задача, алгоритмічно може бути зведена до розв'язання системи нелінійних стохастичних рівнянь із застосуванням чисельних ітераційних процедур Ньютона-Рафсона. Отримані аналітичні вирази для стохастичних поліномів степені S=1–3 дозволяє приймати

рішення щодо доцільності їх застосування в залежності від ймовірнісного характеру регресійних помилок. Зокрема, лінійні (при S=1) ММП-оцінки є еквівалентними класичним МНК оцінкам, що є оптимальними для гаусового розподілу; використання квадратичних (при S=2) ММП-оцінок є доцільним лише для асиметричних розподілів; а при симетричному характері негаусових розподілів доцільним є використання кубічних (при S=3) ММП-оцінок в незалежності від типу детермінованої складової регресійних моделей.

4. Теоретично показано, що застосування запропонованого способу поліноміального оцінювання загалом забезпечує зменшення дисперсії отримуваних оцінок, порівняно із відомими оцінками МНК. При цьому ступінь ефективності не залежить від типу регресійних моделей та є однаковою для всіх складових параметрів регресії. Зростання точності досягається завдяки врахуванню негаусовості регресійних помилок. Кількісно величина ефективності ММПл-оцінок (відносно МНК) залежить від ступеня негаусовості, яка описується значеннями кумулянтних коефіцієнтів вищих порядків. На прикладі моделі з експоненціальним степеневим розподілом здійснено порівняльний аналіз точності поліноміальних оцінок відносно оцінок максимальної правдоподібності, результати якого свідчать близькість величини дисперсії оцінок для обох методів при плосковершинному характері розподілів регресійних помилок.

5. Розроблений програмний комплекс, його структура та набір програмних модулів забезпечують як безпосереднє вирішення задачі знаходження адаптивних оцінок параметрів різних типів регресійних залежностей (лінійних, поліноміальних, нелінійних) так і реалізацією комп’ютерного статистичного моделювання на основі методу Монте-Карло і бутстреп-аналізу. Сукупність отриманих результатів статистичного моделювання у цілому підтверджують теоретично доведену ефективність ММПл-оцінок відносно МНК-оцінок. На прикладі моделі ЕСР-помилок з також показано, що при відсутності апріорної інформації про значення параметрів регресійних помилок адаптивні ММПл-оцінки можуть бути більш точними порівняно із класичними оцінкам. При цьому максимально можливе зростання точності щодо МНК складає до 60%, а відносно ММП до 10%.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. J. D. Gergonne (1974) The application of the method of least squares to the interpolation of sequences. Historia Math., vol. 1, no. 4, pp. 439-447, doi: 10.1016/0315-0860(74)90034-2.
2. Legendre A. M. (1805) Nouvelles méthodes pour la détermination des orbites des comètes, Chez Firmin Didot, Paris, France
3. Gauss C. F. (1823) Theoria combinationis observationum erroribus minimis obnoxiae, Henricus Dieterich, Gottingen, Germany.
4. Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.
5. Стрижов В.В. Методы индуктивного порождения регрессионных моделей. М.: ВЦ РАН, 2008. 54 с
6. Xin Yan and Xiao Gang Su. (2009) Linear Regression Analysis: Theory and Computing. World Scientific Publishing Co., Inc., USA
7. Nylen, E. L., & Wallisch, P. (2017). Neural Data Science: A Primer with MATLAB® and PythonTM. Academic Press.
8. Bloice, M. D., & Holzinger, A. (2016). A tutorial on machine learning and data science tools with python. Machine Learning for Health Informatics, 435-480.
9. Sheather, S. (2009). A modern approach to regression with R. Springer Science & Business Media.
10. Seber G.A.F Wild C.J. (1989) Nonlinear Regression. New York: John Wiley and Sons
11. Nocedal J., Wright S.J. (1999) Numerical Optimization ,Springer, 1999
12. Björck, Å. (1996). Numerical methods for least squares problems. Society for Industrial and Applied Mathematics.
13. Levenberg, K. A Method for the Solution of Certain Problems in Last Squares. Quart. Appl. Math. 1944. Vol. 2. P. 164—168
14. Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594-604), 309-368

15. Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. Journal of Mathematical Psychology. 47 (1): 90–100. doi:10.1016/S0022-2496(02)00028-7
16. David, F. N.; Neyman, J. (1938). Extension of the Markoff theorem on least squares. Statistical Research Memoirs. 2: 105–116.
17. Gnedenko, B.V.; Kolmogorov, A.N. (1954). Limit distributions for sums of independent random variables. Cambridge: Addison-Wesley
18. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? / А.И. Орлов // Заводская лабораторія. – 1991. – Т. 57. – №. 7. – С. 64 – 66.
19. Новицкий П.В. Оценка погрешностей результатов измерений / П.В. Новицкий, И.А. Зограф. – Л.: Энергоатомиздат, 1985. – 248 с.
20. Huber, P. J., & Ronchetti, E. M. (2009). Robust Statistics. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470434697
21. Hodge, Victoria J.; Austin, Jim (2004), "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, 22 (2): 85–126, doi:10.1023/B:AIRE.0000045502.10941.a9
22. Huber, P. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics, 35*(1), 73-101. Retrieved March 7, 2021, from http://www.jstor.org/stable/2238020
23. Jureckova, J. (1971). Nonparametric Estimate of Regression Coefficients. The Annals of Mathematical Statistics
24. Jurečková, J. (1984). 21 M-, L- and R-estimators. In Handbook of Statistics (Vol. 4, pp. 463–485). Elsevier. https://doi.org/10.1016/S0169-7161(84)04023-2
25. Williams, M. S. (1997). A Regression Technique Accounting for Heteroscedastic and Asymmetric Errors. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(1), 108–129. https://doi.org/10.2307/1400643
26. Приходько, С., & Макарова, Л. (2014). Confidence interval of nonlinear regression of restoration time of network terminal devices. *Eastern-European Journal of Enterprise Technologies*, *3*(4(69)), 26-31. doi: http://dx.doi.org/10.15587/1729-4061.2014.24663

27. Box, G. E. P., and Cox, D. R. (1964), “An Analysis of Transformations Journal of the Royal Statistical Society, Ser. B, 26, 211-246.
28. Johnson, N. L. (1949), “Systems of Frequency Curves Generated by Methods of Translation,” Biometrika, 36, 149-176.
29. Barrodale, I., & Roberts, F. D. K. (1973). An Improved Algorithm for Discrete L1 Linear Approximation. SIAM Journal on Numerical Analysis, 10(5), 839–848. doi:10.1137/0710069
30. Zeckhauser R & Thompson M (1970). Linear regression with non-normal error terms. The Review of Economics and Statistics 52:280–286
31. Mineo A (1989). The norm-p estimation of location, scale and simple linear regression parameters. Lecture notes in statistics. Statistical Modelling Proceedings, 222–233. Trento.
32. Agrò G (1992). Maximum likelihood and Lp-norm estimators. Statistica Applicata 4:171–182.
33. Zeckhauser, R., & Thompson, M. (1970). Linear Regression with Non-Normal Error Terms. The Review of Economics and Statistics, 52(3), 280–286.
34. Islam, M. Q., Tiku, M. L., & Yildirim, F. (2001). Nonnormal Regression. I. Skew Distributions. Communications in Statistics - Theory and Methods, 30(6), 993–1020. https://doi.org/10.1081/STA-100104347
35. Marazzi, A., & Yohai, V. J. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. Journal of Statistical Planning and Inference, 122(1–2), 271–291. https://doi.org/10.1016/j.jspi.2003.06.011
36. Tiku, M. L., & Akkaya, A. D. (2010). Estimation in multifactor polynomial regression under non-normality. Pakistan Journal of Statistics, 26(1), 49–68.
37. Galea, M., Paula, G. A., & Bolfarine, H. (1997). Local influence in elliptical linear regression models. Journal of the Royal Statistical Society Series D: The Statistician, 46(1), 71–79.
38. Liu, S. (2002). Local influence in multivariate elliptical linear regression models. Linear Algebra and Its Applications, 354(1–3), 159–174. https://doi.org/10.1016/S0024-3795(01)00585-7

39. Ganguly, S. S. (2014). Robust Regression Analysis for Non-Normal Situations under Symmetric Distributions Arising In Medical Research. Journal of Modern Applied Statistical Methods, 13(1), 446–462. https://doi.org/10.22237/jmasm/1398918480

40. Bartolucci, F., & Scaccia, L. (2005). The use of mixtures for dealing with non-normal regression errors. Computational Statistics and Data Analysis, 48(4), 821–834. https://doi.org/10.1016/j.csda.2004.04.005

41. Seo, B., Noh, J., Lee, T., & Yoon, Y. J. (2017). Adaptive robust regression with continuous Gaussian scale mixture errors. *Journal of the Korean Statistical Society*, *46*(1), 113–125. https://doi.org/10.1016/j.jkss.2016.08.002

42. Tiku, M. L., Islam, M. Q., & Selçuk, A. S. (2001). Nonnormal regression. II. Symmetric distributions. *Communications in Statistics - Theory and Methods*, 30(6), 1021–1045. https://doi.org/10.1081/STA-100104348

43. Andargie, A. A., & Rao, K. S. (2013). Estimation of a linear model with two-parameter symmetric platykurtic distributed errors. Journal of Uncertainty Analysis and Applications, 1(1), 1–19.

44. Atsedeweyn, A. A., & Srinivasa Rao, K. (2014). Linear regression model with generalized new symmetric error distribution. Mathematical Theory and Modeling, 4(2), 48–73. https://doi.org/10.1080/02664763.2013.839638

45. Zhu, D., & Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. Journal of Econometrics, 148(1), 86–99. https://doi.org/10.1016/j.jeconom.2008.09.038

46. Малахов А.Н. Кумулянтный анализ негауссовских случайных процессов и их преобразований / А.Н Малахов. – М.: Сов. радио, 1978. – 376 с.

47. Pal, M. (1980). Consistent moment estimators of regression coefficients in the presence of errors in variables, J. Econometrics, 14, 349–364.

48. Van Montfort, K.; Mooijaart, A. and de Leeuw, J. (1987). Regression with errors in variables: estimators based on third order moments, Statist. Neerlandica, 41(4), 223–237.

49. Dagenais, M. G., & Dagenais, D. L. (1997). Higher Moment Estimators for Linear Regression Models with Errors in the Variables. Journal of Econometrics, 76(1–2), 193–221. https://doi.org/10.1016/0304-4076(95)01789-5
50. Cragg, J. G. (1997). Using Higher Moments to Estimate the Simple Errors-in-Variables Model. The RAND Journal of Economics, 28, S71. https://doi.org/10.2307/3087456
51. Gillard, J. (2014). Method of Moments Estimation in Linear Regression with Errors in both Variables. Communications in Statistics - Theory and Methods, 43(15), 3208–3222. https://doi.org/10.1080/03610926.2012.698785
52. Wang, L., & Leblanc, A. (2007). Second-order nonlinear least squares estimation. Annals of the Institute of Statistical Mathematics, 60(4), 883–900. doi:10.1007/s10463-007-0139-z
53. Chen, X., Tsao, M., & Zhou, J. (2012). Robust second-order least-squares estimator for regression models. *Statistical Papers*, *53*(2), 371–386. https://doi.org/10.1007/s00362-010-0343-4
54. Kim, Mm., & Ma, Y. (2013). The efficiency of the second-order nonlinear least squares estimator and its extension, 64(4), 751–764. https://doi.org/10.1007/s10463-011-0332-y.
55. Huda, S., & Mukerjee, R. (2018). Optimal designs with string property under asymmetric errors and SLS estimation. Statistical Papers, 59(3), 1255–1268. https://doi.org/10.1007/s00362-016-0819-y
56. Кунченко Ю.П. Нелинейная оценка параметров негауссовских радиофизических сигналов. / Ю.П. Кунченка. – К.: Вища школа. – 1987. – 191 с.
57. Кунченко Ю.П. Стохастические полиномы. / Ю.П. Кунченка. – К.: Наук. думка, 2006. – 275 с.
58. Кунченко Ю.П. Оценка параметров случайных величин методом максимизации полинома / Ю.П Кунченко, Ю.Г Лега. – К.: Наукова думка. – 1992. – 192 с.
59. Кунченко Ю.П. Полиномиальные оценки параметров близких к гауссовским случайных величин. ЧІТІ. Стохастические полиномы, их свойства и

применение для нахождения оценок параметров / Ю.П. Кунченко // – Черкассы: ЧИТИ, 2001. – 133 с.

60. Кунченко Ю.П., Полиномиальные оценки параметров близких к гауссовским случайных величин. Ч.II. Оценка параметров близких к гауссовским случайных величин / Ю.П. Кунченко, С.В. Заболотный. – Черкассы: ЧИТИ, 2001. – 251 с.
61. Чепинога, А. Оценивание параметров полигауссовых моделей методом максимизации полинома / А. Чепинога, С. В. Заболотний, Е. Бурдукова // Восточно-Европейский журнал передовых технологий. – 2014. – Т. 2, – № 4(68). – С. 43-46.
62. Kunchenko Y.P.The accuracy of the joint estimation of parameters of signal by the antenna arrays at non-Gaussian interference / Y.P Kunchenko, V.O. Danik, T.V. Vorobkalo // Proceeding of the 3rd International Conference on Antenna Theory and Techniques. – Sevastopol, Ukraine, 1999. – P. 217 – 218.
63. Кунченко Ю.П. Определение угла прихода волны многоэлементной антенной решеткой на фоне негауссовских помех / Ю.П Кунченко, В.А. Даник, Т.В. Воробкало // Радиотехника. – Харьков, – 2001. – №117. – С. 26 – 30.
64. Kunchenko Y.P. The parameters estimation of signals in the presence of the non-Gaussian interferences / Y.P. Kunchenko // Proceeding of Second International Symposium on Communication Systems Networks and Digital Signal Processing. – Bournemouth University (UK), 2000. – P. 241 – 245.
65. Кунченко Ю.П. Метод максимизации полинома и его приложения в радиотехнике / Ю.П Кунченко. А.С Гавриш // Радиоэлектроника и информатика. – Харьков: –№ 2. –2000. – С. 7 – 13.
66. Гавриш А.С. Построение высокоточных измерителей параметров гармонического сигнала при воздействии негауссовских помех / А.С. Гавриш // Радиоэлектроника и информатика. – Харьков: 2000. – № 1. – С. 8 – 13.
67. Gavrish O.S. Polynomial parameter estimations of a constant signal received at the first-type skewness interferences / O.S. Gavrish A.V Honcharov, Y.P. Kunchenko // Proceedings of the IVth International Hutsulian Workshop on Mathematical Theories

and their Applications in Physics & Technology. – Kyiv: TIMPANI, 2004. – P. 287 – 298.

68. Гавриш О.С. Метод аналізу асимптотичних властивостей адаптивних алгоритмів вимірювання параметрів радіосигналу при ексцесній заваді з використанням навчальної вибірки / О.С. Гавриш, С.В. Заболотній, О.В. Бурдукова // Радиоэлектроника и информатика. – Харьков. – № 3. 2013. – С. 7 – 10.
69. Гавриш О.С. Поліноміальне оцінювання фази радіосигналу при асиметрично-ексцесній мультиплікативній заваді / О.С. Гавриш, С.В. Заболотній, В.В. Коваль // Радиоэлектроника и информатика. – № 1. – 2011. – С. 12 – 15.
70. Гаврыш А.С. Итеративные процедуры решения уравнений максимизации полинома при неодинаково распределенной выборке / А.С. Гаврыш, С.В. Заболотный, В.В. Коваль // Радиоэлектроника и информатика. – №2. – 2011. – С. 6 – 7.
71. Нелинейное оценивание параметров полиномиальных трендов при негауссовской стохастической компоненте / В.О. Селин, С.В. Заболотный // Наукоемкие технологии в инфокоммуникациях: обработка и защита информации : коллективная монография / под ред. В.М. Безрука, В.В. Баранника. – Х.: Компания СМИТ, 2013. – 398 с.
72. Заболотній С.В. Дослідження ефективності нелінійних обчислювальних алгоритмів оцінювання параметрів лінійного тренду в умовах дії асиметрично-ексцесних завад / С.В. Заболотній, О.С. Гавриш, В.О. Селін // Вісник ЧДТУ. – 2008, – № 1. – С. 43 – 46.
73. Заболотній С.В. Моделювання оцінювання параметрів лінійного тренду методом максимізації полінома в умовах дії негаусових завад / С.В. Заболотній, В.О. Селін // Вісник ЧДТУ. – 2009, – № 4. – С. 45 – 48.
74. Заболотній С.В. Рекурентне нелінійне оцінювання параметрів поліноміального тренду при негаусовій стохастичній компоненті / С.В. Заболотній, О.С. Гавриш, В.О. Селін // Вісник ЧДТУ. – 2011, – № 4. – С. 101 – 103.

75. Кунченко Ю.П. Полиномы приближения в пространстве с порождающим элементом. / Ю.П. Кунченко. – К.: Наук. думка, – 2003. – 243 с.
76. Заболотній, С. В. (2014). Застосування розкладу в просторі з порідним елементом для вирішення задач ймовірнісної діагностики. *Eastern-European Journal of Enterprise Technologies*, *4*(4(70), 28–35. https://doi.org/10.15587/1729-4061.2014.26195
77. Заболотній С. Статистичне розпізнавання образів на основі розкладу в просторі з порідним елементом / С. Заболотній // Вісник Національного університету "Львівська політехніка". – 2009. – № 638 : Комп'ютерні науки та інформаційні технології. – С. 118-123.
78. Чертов О.Р. Поліноми Кунченка для розпізнавання образів // О.Р. Чертов // Вісник НТУУ "КПІ": Інформатика, управління та обчислювальна техніка. – 2009. – №. 51. – С. 103 – 108.
79. Чертов О.Р. Застосування поліномів Кунченка для аналізу статистичних даних / О.Р. Чертов, Д.Ю. Тавров // Східноєвропейський журнал передових технологій. – 2010. – № 4/4 (46). – С. 70 – 75.
80. Чертов О. Р. Зіставлення з шаблоном на базі поліномів Кунченка в демографічних даних // О.Р. Чертов // Наукові праці Чорноморського державного університету імені Петра Могили. Сер.: Комп'ютерні технології. – 2011. – №. 160, Вип. 148. – С. 36 – 43.
81. Кунченко Ю.П. Використання стохастичних поліномів для перевірки простих стохастичних гіпотез / Ю.П. Кунченко // Вісник ЧІТІ – Черкаси: ЧІТІ, – 1996, – № 2. – С. 37 – 47.
82. Kunchenko Y.P. A Moment Performance Criterion of a Decision Making for Testing Simple Statistical Hypothesis / Y.P. Kunchenko // IEEE, International Symposium on Information Theory, June-July, 1997. – Ulm, 1997 – P. 231.
83. Палагін В.В. Поліноміальне вирішення задач розпізнавання випадкових сигналів / В.В. Палагін, О.М. Жила // Вісник ЧДТУ. – 2008. – №2. – С. 31 – 35.
84. Лега Ю.Г. Обнаружители радиосигналов на фоне асимметричных негауссовских помех, оптимальних по моментному критерию типа Неймана-

Пирсона / Ю.Г. Лега, В.В. Палагин, С.А. Лелеко // Вісник ЧДТУ. – 2009. – №3. – С. 76 – 81.

85. Палагин В.В. Полиномиальные алгоритмы обнаружения сигналов на фоне коррелированных негауссовских помех / В.В. Палагин // Электроника и связь. – 2010. – №6, часть 1. – С. 20 – 26.
86. Палагін В.В. Нелінійні алгоритми виявлення радіосигналів на тлі адитивно–мультиплікативних негаусівських завад / В.В. Палагін // Східно-Європейський журнал передових технологій. – 2012. №6/11(60). – С. 23-28.
87. Cramér, H. (2016). *Mathematical Methods of Statistics (PMS-9)* (Vol. 9). Princeton university press.
88. Берегун В.С., Гармаш О.В., Красильников А.И., (2014). Среднеквадратические ошибки оценок кумулянтных коэффициентов пятого и шестого порядков. *Электронное Моделирование*, *36*(1), 17–28.
89. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. – М.: Физматлит. – 2006. – 816 с.
90. Jarque, C. M., & Bera, A. K. (2012). A Test for Normality of Observations and Regression Residuals. International Statistical Review, 55(2), 163–172.
91. Komsta L, Novomestky F (2007). moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests. R package version 0.11, URL http://CRAN.R-project.org/package=moments.
92. Норманн Р. Дрейпер, Гарри Смит. Прикладной регрессионный анализ, 3-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2007. – 912 С.
93. Macon, N.; A. Spitzbart (February 1958). "Inverses of Vandermonde Matrices". *The American Mathematical Monthly*. **65** (2): 95–100. doi:10.2307/2308881
94. Stone, C. J. (1975). Adaptive Maximum Likelihood Estimators of a Location Parameter. The Annals of Statistics, 3(2), 267–284. https://doi.org/10.1214/aos/1176343056
95. Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, *32*(7), 685–694. https://doi.org/10.1080/02664760500079464

96. Giller, G. L. (2013). A Generalized Error Distribution. *SSRN Electronic Journal*, *November*. https://doi.org/10.2139/ssrn.2265027

97. Krasilnikov, A. I. (2019). Family of Subbotin Distributions and its Classification. Èlektronnoe Modelirovanie, 41(3), 15–32. doi:10.15407/emodel.41.03.015

98. Lindsey, J.K. (1999). Multivariate Elliptical Contoured Distributions for Repeated Measurements. Biometrics 55, 1277- 1280.

99. Шелухин, О.И. (1999), Негауссовские процессы в радиотехнике, Радио и связь, Москва

100. Sharifi, K., & Leon-Garcia, A. (1995). Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. IEEE Transactions on Circuits and Systems for Video Technology, 5(1), 52-56

101. Dominguez-Molina, J.A., Gonzalez-Farias, G. and Rodriguez-Dagnino, R.M. (2001), A practical procedure to estimate the shape parameter in the generalized Gaussian distribution, available at: http://www.cimat.mx/reportes/enlinea/I-01-18_eng.pdf

102. Saatci, E., & Akan, A. (2010). Respiratory parameter estimation in non-invasive ventilation based on generalized Gaussian noise models. Signal processing, 90(2), 480-489.

103. Olosunde, A. A. (2013). On exponential power distribution and poultry feeds data: a case study. Journal of The Iranian Statistical Society, 12(2), 253-270.

104. Giacalone, M. (2020). A combined method based on kurtosis indexes for estimating p in non-linear L p-norm regression. *Sustainable Futures*, *2*, 100008. https://doi.org/10.1016/j.sftr.2020.100008

105. Chan, J., Choy, B., & Walker, S. (2018). On The Estimation Of The Shape Parameter Of A Symmetric Distribution. Stephen, On the Estimation of the Shape Parameter of a Symmetric Distribution (January 1, 2018).

106. Ferreira, M. A., & Salazar, E. (2014). Bayesian reference analysis for exponential power regression models. Journal of Statistical Distributions and Applications, 1(1), 12. doi:10.1186/2195-5832-1-12

107. Kroese, D. P.; Brereton, T.; Taimre, T.; Botev, Z. I. (2014). Why the Monte Carlo method is so important today. WIREs Comput Stat. 6 (6): 386–392. doi:10.1002/wics.1314

108. Efron, B. (1987). Better Bootstrap Confidence Intervals. Journal of the American Statistical Association, 82(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

109. Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399-402

110. НДР «Розробка технології та пристроїв адитивного виробництва індивідуальних хірургічних імплантатів та протезів з біосумісних полімерних матеріалів» (№ держ. реєстрації 0117U000937)

111. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

112. Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning* (pp. 236-243).

113. Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287-1294.

114. C.D. Keeling, T. P. Whorf, Scripps Institution of Oceanography (SIO), University of California, La Jolla, California USA 92093-0220. ftp://cdiac.esd.ornl.gov/pub/maunaloa-co2/maunaloa.co2.

115. Ferreira, M. A., & Salazar, E. (2014). Bayesian reference analysis for exponential power regression models. Journal of Statistical Distributions and Applications, 1(1), 12. doi:10.1186/2195-5832-1-12

116. Levine DM, Krehbiel TC, Berenson ML (2000). Business Statistics: A First Course. Prentice-Hall

117. Mineo, a M., & Ruggieri, M. (2005). A software tool for the exponential power distribution: The normalp package. Journal of Statistical Software, 12(4), 1–24. https://doi.org/http://dx.doi.org/10.18637/jss.v012.i04

ДОДАТОК А

Результати застосування поліноміальних моделей, методів та програмних засобів при регресійному аналізі реальних даних

А.1. Оцінювання параметрів лінійної регресії при асиметричному розподілі помилок

В даному експерименті для тестування розроблених моделей та адаптивних методів поліноміального оцінювання параметрів регресії використано набір даних Auto MPG із репозиторію UCI [111]. Цей набір даних представляє собою залежність витрати палива (галон на милю) в міському циклі, яка виступає у ролі цільової змінної, від різних характеристик автомобілів: потужності, кількості-циліндрів та робочого об’єму двигуна, прискорення і ваги, та ін. [112].

На рис.А.1,*а* представлена залежність величини витрати палива MPG (змінна $\boldsymbol{Y}$) від одного з параметрів (прискорення – Acceleration, регресор $\boldsymbol{X}$) для $N = 392$ (з урахуванням пропущених даних) типів автомобілів. Візуальний аналіз цього рисунка дозволяє припустити наявність лінійної залежності між параметрами $\boldsymbol{Y}$ і $\boldsymbol{X}$, яку можна описати регресійної моделлю виду (2.10) при $Q = 2$. МНК-оцінок параметрів такої моделі є значення: $\hat{a}_0{}^{(1)} = 4.83$; $\hat{a}_1{}^{(1)} = 1.2$. Отримане на основі тесту Бройша-Пагана [113] значення p-value на рівні 0.01, фактично не спростовує гіпотезу про гомоскедастичність регресійних МНК-залишків і коректність лінійної залежності.

Асиметричність розподілу помилок моделі візуально видно на рис.А.1,*б*, де зображений емпіричний розподіл частот регресійних МНК-залишків. Гіпотеза про гаусовість МНК-залишків також спростовується вбудованим в MATLAB тестом Харке-Бера ($JBSTAT = 17.4$ при пороговому значенні $CV = 5.8$ на фіксованому рівні значущості $alfa = 0.05$), який базується на аналізі величини кумулянтних коефіцієнтів асиметрії та ексцесу [90].

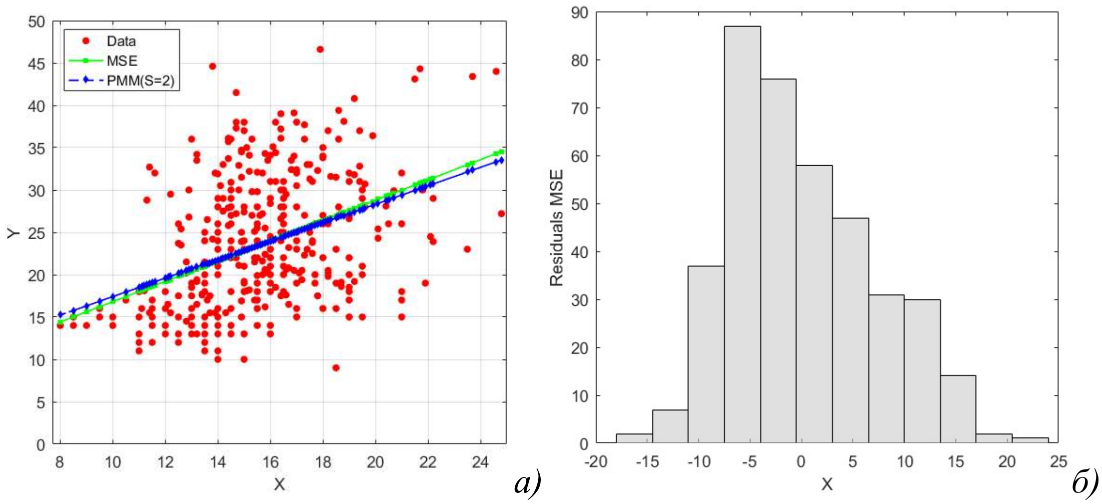


Рисунок А.1 – Побудова лінійної регресійної моделі із асиметричними помилками: *а)* експериментальні дані та оцінки регресії, на основі МНК і адаптивного ММПл; *б*) емпіричний розподіл частот регресійних МНК-залишків

Отримані оціночні значення параметрів $\hat{\kappa}_2 = 50$, $\hat{\gamma}_3 = 0.49$ і $\hat{\gamma}_4 = -0.32$ регресійних МНК-залишків (з урахуванням досить великого обсягу вихідних даних $N = 392$) дозволяють оцінити величину зменшення дисперсії отриманих ММПл-оцінок параметрів регресії ($\hat{a}_0{}^{(2)} = 6.5$; $\hat{a}_1{}^{(2)} = 1.09$) на рівні $\hat{g}_{(\theta_p)2} = 0.86$.

Відомо, що для задач регресійного аналізу іноді застосовуються не точкові оцінки параметрів, інтервальні, які легше інтерпретуються. З урахуванням дотримання умови нормалізації розподілу МНК-оцінок (для великого N), отримуємо, що при рівні довіри 0.95 інтервал $(0.82;\ 8.84)$ перекриває діапазон значень параметра регресії a_0, а інтервал $(0.94;\ 1.45)$ - параметра a_1. При цьому, значення величини оцінки коефіцієнта зменшення дисперсії, а також отриманих точкових значень ММПл-оцінок, дозволяє для заданого рівня довірчої ймовірності обґрунтовано скоригувати і зменшити на$1 - \sqrt{0.86} \approx 7$ % ширину довірчого інтервалу, отримавши його межах $(2.8;\ 10.22)$ для параметра a_0, і $(0.85;\ 1.33)$ для параметра a_1.

Для верифікації коректності отриманих результатів був проведений статистичний експеримент, заснований на методі бутстрепінга (Bootstrapping) [108]. За допомогою розроблених із застосуванням MATLAB програмних засобів початкова вибірка була розмножена на 10^4 бутстреп-вибірок з поверненням. Для кожної з них були знайдені МНК і ММПл оцінки параметрів лінійної регресії. Емпіричне розподіл цих оцінок представлено на рис.А.2 у вигляді boxplot-графіків, верхня і нижня границями яких є 2.5% і 97.5% перцентиль (percentile) відповідно.

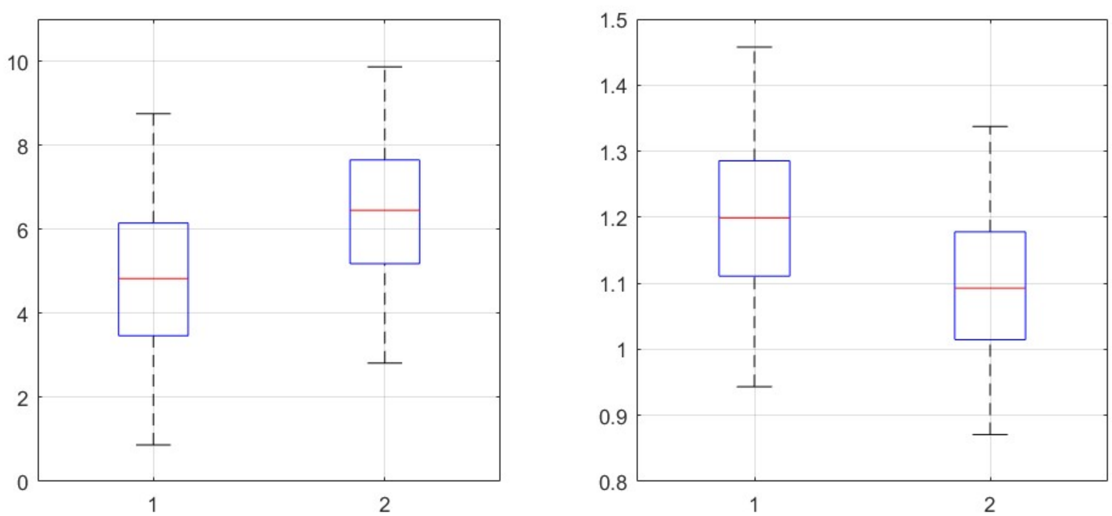


Рисунок А.2 – Емпіричний розподіл бутстреп-оцінок параметрів лінійної регресії

Візуальний аналіз, наведених на рис.А.2, границь 95% довірчих інтервалів, отриманих в результаті статистичного бутстреп-моделювання, і порівняння з отриманими вище розрахунковими значеннями інтервальних оцінок свідчить про їх істотною близькості. Це в цілому свідчить про достовірність аналітичних розрахунків.

А.2. Оцінювання параметрів лінійної регресії при симетричному розподілі помилок

Протестуємо розроблений в підрозділі 2.3.4 алгоритм знаходження адаптивних ММПл-оцінок параметрів лінійної регресії при симетрично-розподілених помилках з використанням реальних даних. Цей набір даних представляє собою залежність атмосферної концентрації CO^2, отриманих із зразків повітря, зібраних в обсерваторії Мауна-Лоа, Гаваї [114]. Для експерименту використана частина цих даних обсягом $N = 336$ (щомісячні вимірювання з січня 1970 по вересень 1997 року), які адекватно (див. рис.А.3, *а*) описуються лінійної регресійної моделлю виду (2.10) при $Q = 2$.

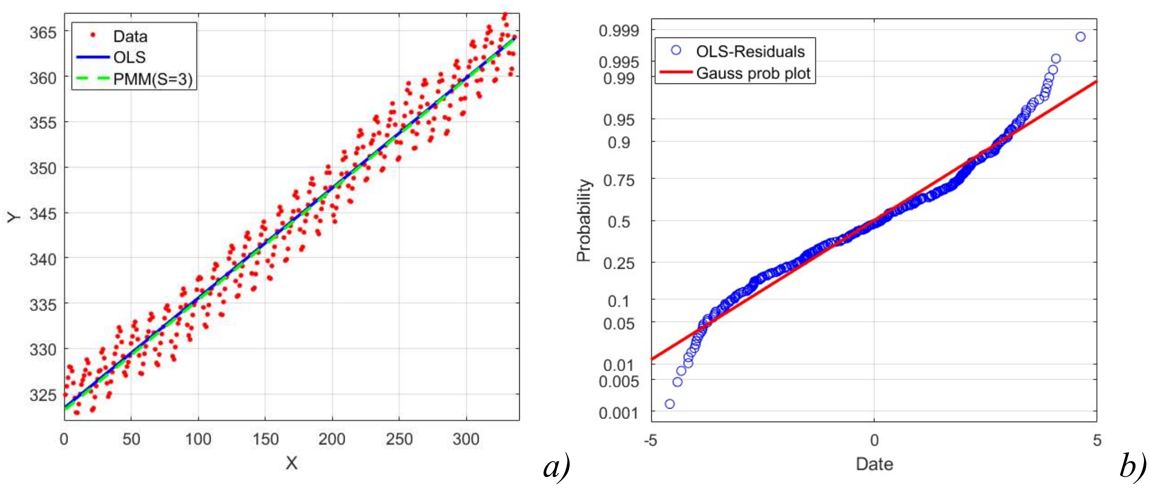


Рисунок А.3 – Побудова лінійної регресійної моделі із симетричними помилками:
а) експериментальні дані та оцінки регресії;
б) розподіл (Q-Q plot) МНК-залишків

МНК-оцінками параметрів такої моделі є значення: $\hat{a}_0{}^{(1)} = 323.411$; $\hat{a}_1{}^{(1)} = 0.121$. Гіпотеза про гаусовість МНК-залишків спростовується на основі використання тесту Харке-Бера ($JBSTAT = 15.6$ при пороговому значенні $CV = 5.8$ та фіксованому рівні достовірності $alfa = 0.05$) [90]. Симетрію розподілу помилок моделі візуально видно на рис.4, б, де представлена імовірнісний графік (Q-Q plot) регресійних МНК-залишків і підтверджується малою абсолютною величиною оцінки їх коефіцієнта асиметрії $\hat{\gamma}_3 = -0.15$.

Оціночні значення парних кумулянтних коефіцієнтів: $\hat{\gamma}_4 = -1$ і $\hat{\gamma}_6 = 5.2$ регресійних МНК-залишків дозволяють у відповідності до (3.9) досить точно (з урахуванням досить великого обсягу вихідних даних N) визначити величину коефіцієнта зменшення дисперсії $\hat{g}_{(\theta_p)3} = 0.5$. Відзначимо, що самі значення уточнених ММПл-оцінок ($\hat{a}_0{}^{(3)} = 323.156$; $\hat{a}_1{}^{(3)} = 0.122$) відрізняються від значень МНК-оцінок не значно. Однак суттєве зменшення дисперсії дозволяє будувати більш вузькі довірчі інтеграли, які зазвичай використовуються при регресійному аналізі. З огляду на дотримання умови нормалізації розподілу МНК-оцінок (при великому обсязі вибірки N), отримуємо, що при рівні довіри 0.95 інтервал $(322.936;\ 323.886)$ накриває значення параметра регресії a_0, а інтервал $(0.119;\ 0.124)$ - параметра a_1. При цьому, значення величини оцінки коефіцієнта зменшення дисперсії, а також отриманих точкових значень ММПл-оцінок, дозволяє для заданого рівня довірчої ймовірності обґрунтовано скоригувати і зменшити на $\left(1 - \sqrt{0.5}\right) 100 \approx 30\ \%$ ширину довірчого інтервалу, отримавши його в межах $(322.82;\ 323.491)$ для параметра a_0, і $(0.12;\ 0.123)$ для параметра a_1.

Для верифікації коректності отриманих результатів був проведений статистичний експеримент, заснований на методі бутстрепінга [108]. За допомогою розроблених із застосуванням MATLAB програмних засобів початкова вибірка була розмножена на 10^4 бутстреп-вибірок з поверненням. Для кожної з них були знайдені МНК і ММПл оцінки параметрів лінійної регресії. Емпіричне розподіл цих оцінок представлено на рис.А.2 у вигляді boxplot-графіків, верхня і нижня границями яких є 2.5% і 97.5% перцентиль (percentile) відповідно.

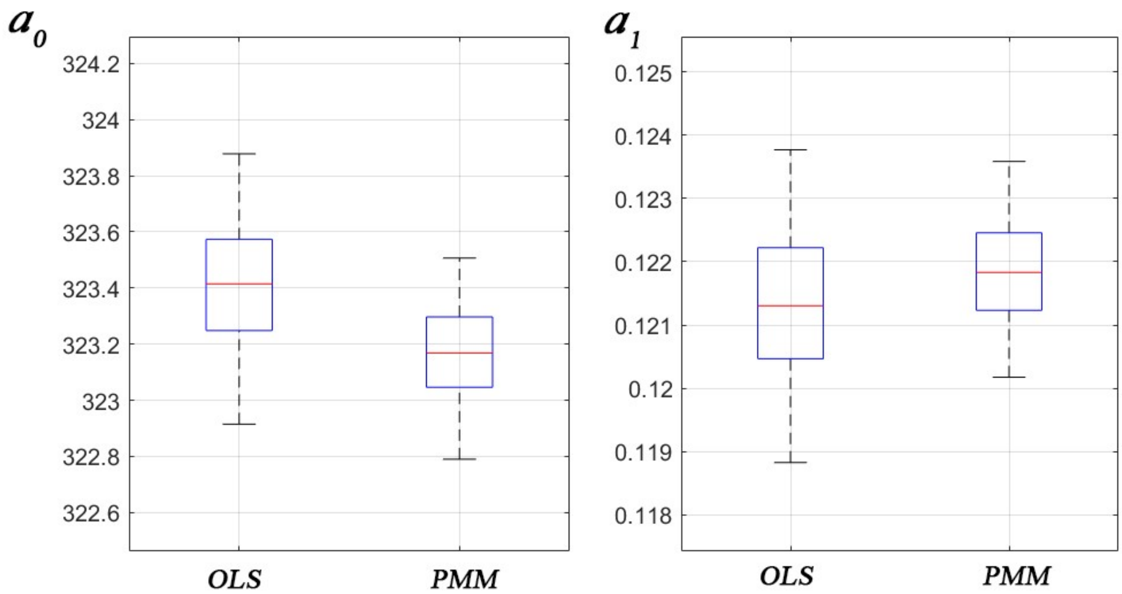


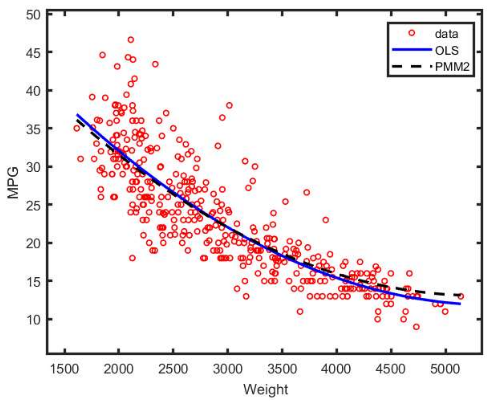
Рисунок А.4 – Емпіричний розподіл бутстреп-оцінок параметрів лінійної регресії

Візуальний аналіз, наведених на рис.А.4, границь 95% довірчих інтервалів, отриманих в результаті статистичного бутстреп-моделювання, і порівняння з отриманими вище розрахунковими значеннями інтервальних оцінок свідчить про їх істотною близькості. Це в цілому свідчить про достовірність аналітичних розрахунків.

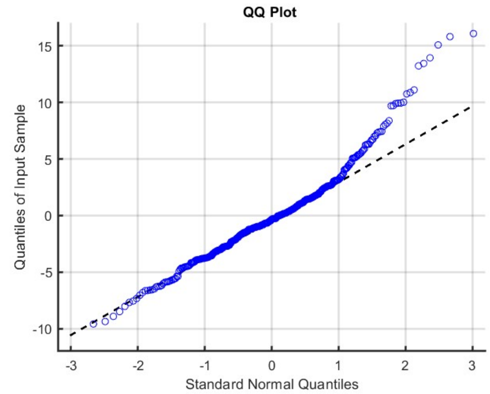
А.3. Оцінювання параметрів поліноміальної регресії при асиметричному та симетричному розподілі помилок

Як приклад для тестування розроблених в підрозділах 2.4.2 та 2.4.3 адаптивних алгоритмах для пошуку ММПл-оцінок параметрів поліноміальної регресії використано той же набір даних Auto MPG із сховища UCI [112].

А.3.1. У першому експерименті вирішується задача знаходження оцінок параметрів регресійної залежності цільової змінної (витрат палива MPG) від ваги автомобіля (параметр Weight) (див. рис.А.5,a) для $N = 392$ автомобілів. Візуальний аналіз цього рисунку свідчить про наявність квадратної залежності $MPG = a_0 + a_1 Weight + a_2 Weight^2$.



а)



б)

Рисунок А.5 – Побудова квадратичної регресійної моделі із асиметричними помилками: *а)* експериментальні дані та оцінки регресії, на основі МНК і адаптивного ММПл; *б*) розподіл (Q-Q plot) МНК-залишків

Оцінки МНК (з межами довірчих інтервалів на рівні 95%) параметрів цієї моделі є значеннями:

$$\hat{a}_0{}^{(1)} = 62.3\ (56.4, 68.1);\ \hat{a}_1{}^{(1)} = -18.5 * 10^{-3}(-22.4 * 10^{-3},\ -14.6 * 10^{-3});\ \hat{a}_2{}^{(1)} = 1.7 * 10^{-6}(1.1 * 10^{-6}, 2.3 * 10^{-6}).$$

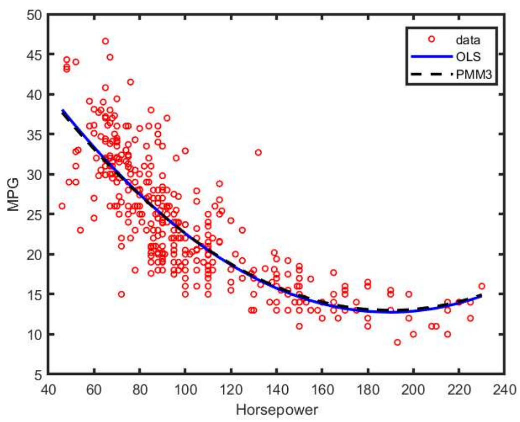
Асиметрія розподілу помилок моделі візуально видно на рис.А.5, *б*, де представлено графік Q-Q регресії МНК-залишків. Гіпотеза гаусового розподілу щодо МНК-залишків також спростовується тестом Харкі-Бери ($JBSTAT = 93.2$ при пороговому значенні $CV = 5.8$ на фіксованому рівні значущості $alfa = 0.05$).

Розрахункові значення $\hat{\gamma}_3 = 0.8$ і $\hat{\gamma}_4 = 1.8$ регресійних МНК-залишків (з урахуванням досить великого обсягу вихідних даних N) відповідно до (3.6) дозволяють оцінити величину зменшення дисперсії ММПл-оцінок (порівняно з МНК) на рівні $\hat{g}_2 = 0.83$. В результаті ММПл-оцінки (з межами довірчих інтервалів на рівні 95%) параметрів цієї моделі приймають значення:

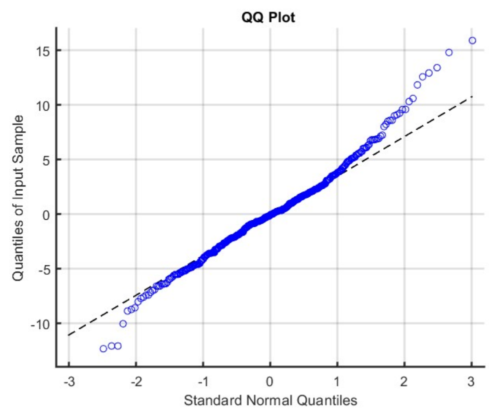
$$\hat{a}_0{}^{(2)} = 60.7\ (55.3, 66);\ \hat{a}_1{}^{(2)} = -18 * 10^{-3}(-21.5 * 10^{-3},\ -14.4 * 10^{-3});$$

$$\hat{a}_2{}^{(2)} = 1.7 * 10^{-6}(1.1 * 10^{-6}, 2.2 * 10^{-6}).$$

А.3.1. У другому експерименті вирішується задача оцінки параметрів залежності витрати палива (параметр MPG) від потужності автомобіля (параметр *Horspower*). Візуальний аналіз на рис.А,6, *а* свідчить про існування квадратичної залежності $MPG = a_0 + a_1 Horsepower + a_2 Horsepower^2$.



а)



б)

Рисунок А.6 – Побудова квадратичної регресійної моделі із симетричними помилками: *а)* експериментальні дані та оцінки регресії, на основі МНК і адаптивного ММПл; *б*) розподіл (Q-Q plot) МНК-залишків

Оцінки МНК (з межами довірчих інтервалів на рівні 95%) параметрів цієї моделі є значеннями:

$$\hat{a}_0{}^{(1)} = 56.9\ (53.36, 60.44);\ \hat{a}_1{}^{(1)} = -46.6 * 10^{-2}(-52.7 * 10^{-2},\ -40.5 * 10^{-2});$$

$$\hat{a}_2{}^{(1)} = 12.3 * 10^{-4}(9.9 * 10^{-4}, 14.7 * 10^{-4}).$$

Гіпотеза про гаусовість МНК-залишків спростовується тестом Харке-Бера ($JBSTAT = 30.6$ при пороговому значенні $CV = 5.8$ на фіксованому рівні значущості $alfa = 0.05$). При цьому візуальний аналіз QQ-графіку регресійних

МНК-залишків, представлений на рис.А.6,*б* вказує на симетрію розподілу. Це чисельно підтверджується близькістю до нуля значення коефіцієнта асиметрії$\hat{\gamma}_3 \approx 0.2$. У цій ситуації для адаптивного уточнення оцінок доцільно використовувати кубічний варіант ММПл при ступені $S = 3$.

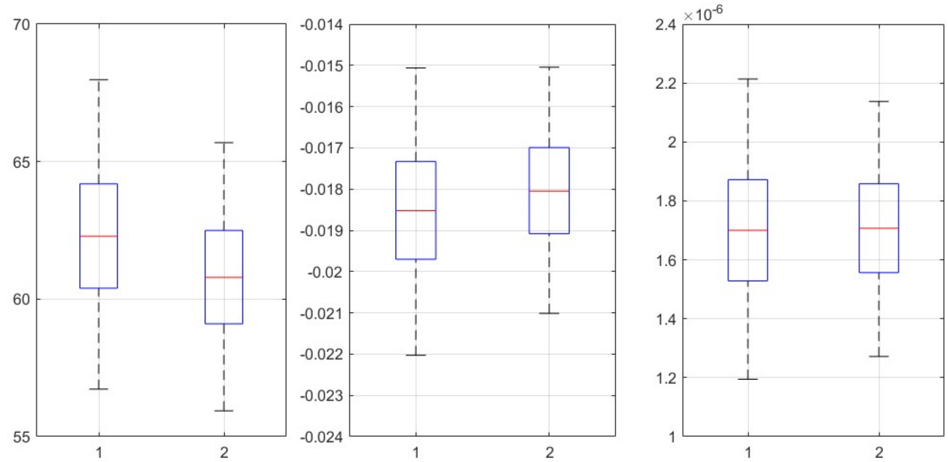


Рисунок А.7 – Емпіричний розподіл бутстреп-оцінок параметрів квадратичної регресії $MPG = a_0 + a_1 Weight + a_2 Weight^2$

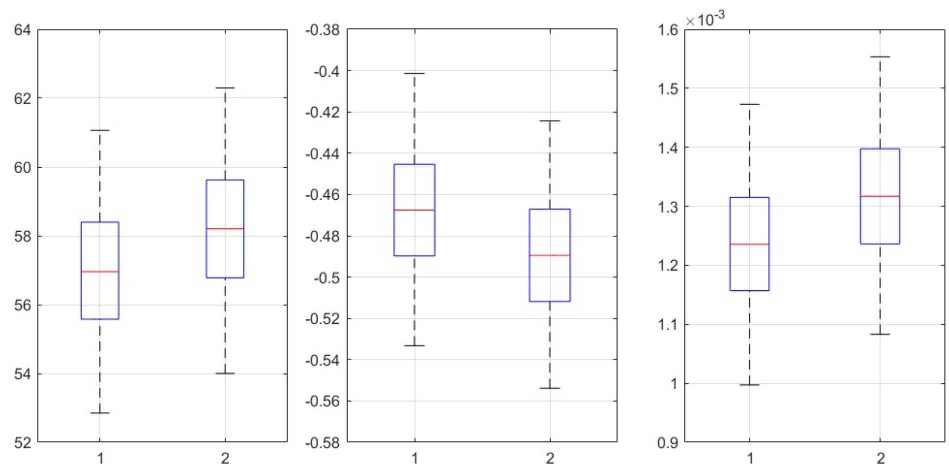


Рисунок А.8 – Емпіричний розподіл бутстреп-оцінок параметрів квадратичної регресії $MPG = a_0 + a_1 Horsepower + a_2 Horsepower^2$

Розрахункові значення $\hat{\gamma}_4 = 1.3$ і $\hat{\gamma}_6 = -1.9$ регресійних МНК-залишків (з урахуванням досить великого обсягу вихідних даних$N = 392$) відповідно до (3.9) дозволяють оцінити величину зменшення дисперсії ММПл-оцінок для$S = 3$

порівняно з МНК на рівні $\hat{g}_3 = 0.89$. В результаті ММПл-оцінками (з межами довірчих інтервалів на рівні 95%) параметрів цієї моделі є значення:

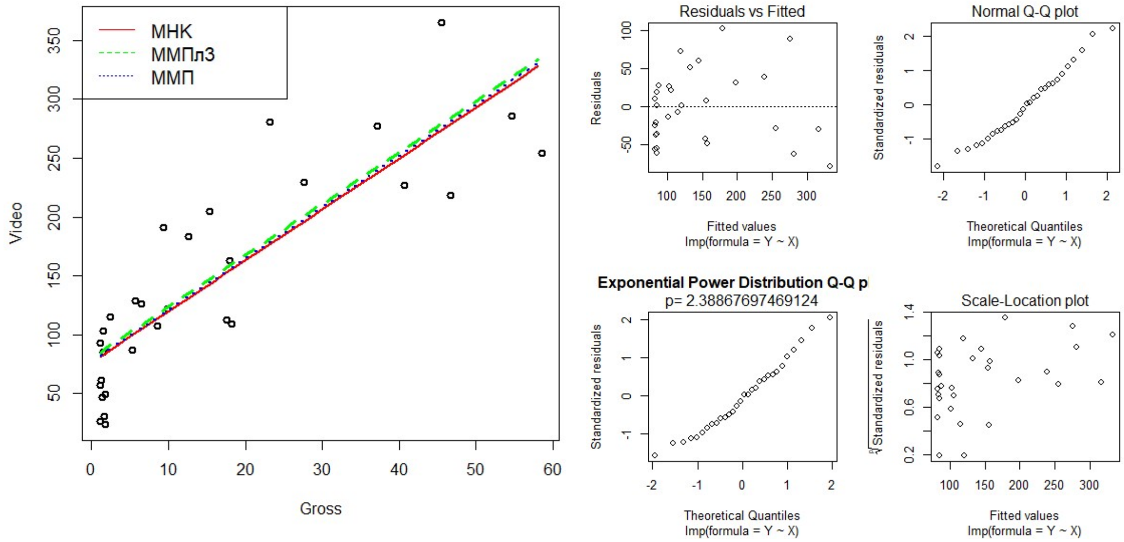
$$\hat{a}_0{}^{(2)} = 58.2\ (54.9, 61.5);\ \hat{a}_1{}^{(2)} = -48.9 * 10^{-2}(-54.7 * 10^{-2},\ -43.1 * 10^{-2});$$
$$\hat{a}_2{}^{(1)} = 13.2 * 10^{-4}(10.9 * 10^{-4}, 15.3 * 10^{-4}).$$

Для верифікації коректності оцінки параметрів поліноміальних регресійних моделей також був здійснений статистичний експеримент, заснований на бутстреп-аналізі. За допомогою розроблених із застосуванням MATLAB програмних засобів початкова вибірка була розмножена на 10^4 бутстреп-вибірок з поверненням. Для кожної з них були знайдені МНК і ММПл оцінки параметрів лінійної регресії. Емпіричне розподіл цих оцінок представлено на рис.А.7 та рис.А.8 у вигляді бокс-плот графіків, верхня і нижня границями яких є 2.5% і 97.5% перцентиль (percentile) відповідно.

Візуальний аналіз сукупності представлених результатів підтверджує різницю між відповідними оцінками параметрів МНК та ММПл, яка полягає у зсуві центрів їх довірчих інтервалів та зміні ширини.

А.4. Оцінювання параметрів лінійної регресії при ЕСР помилок

Для демонстрації працездатності запропонованого підходу адаптивного оцінювання параметрів регресійних моделей в умовах ЕСР регресійних помилок використаємо приклад з реальними даними, що розглядається в роботі [115], де також досліджується застосування ЕСР як моделі помилок лінійної регресії. Цей набір, взятий з [116], містить дані про прибуток у касах у мільйонах доларів (Gross) та кількість проданих домашніх відео у тисячах (Video) для вибірки з 30 фільмів. Задача полягає у визначені залежності кількість проданих відеозаписів від прибутку у касах, яка адекватно описується лінійною регресією (див. рис. А.9,*а*)



а) *б)*

Рисунок А.9 – Лінійна регресійна модель із ЕСР помилками:

а) експериментальні дані і залежності, побудовані різними методами;

б) характеристики МНК-залишків при використанні ЕСР

МНК-оцінками параметрів такої регресійної моделі є значення: $\hat{a}_0^{(\text{МНК})} = 76.54$; $\hat{a}_1^{(\text{МНК})} = 4.33$. Використовуючи для оцінювання параметрів регресії методом максимальної правдоподібності та візуалізації результатів (див. рис.А.9,*б*) додатковий функціонал пакету “normalp package” мови R [117] отримуємо оціночні значення: $\hat{a}_0^{(\text{МНК})} = 77.41$; $\hat{a}_1^{(\text{МНК})} = 4.37$. Підставивши оцінку параметру форми ЕСР регресійних МНК-залишків $\hat{\beta} = 2.39$ в (3.12) отримуємо оцінку коефіцієнта $\hat{g}_{\frac{\text{ММП}}{\text{МНК}}} = 0.97$. Оцінюючи величину кумулянтних коефіцієнтів 4-го $\hat{\gamma}_4 = -0.59$ та 6-го $\hat{\gamma}_4 = 2.14$ порядків і використовуючи (3.9) знаходимо оціночне значення коефіцієнту $\hat{g}_{\frac{\text{ММПл3}}{\text{МНК}}} = 0.88$. При цьому, уточненими адаптивними ММПл-оцінок параметрів регресійної моделі є: $\hat{a}_0^{(\text{ММПл3})} = 79.99$; $\hat{a}_1^{(ММПл)} = 4.37$.

У цілому можна констатувати перевищення точності ММПл-оціннок порівняно з ММП-оцінками приблизно на 10% за умови відсутності апріорної інформації про параметри ЕСР регресійних помилок.

ДОДАТОК Б

Список публікацій в яких опубліковані основні наукові результати дисертації

1. Zabolotnii, S., Warsza, Z. L., & Tkachenko, O. (2018). Polynomial Estimation of Linear Regression Parameters for the Asymmetric PDF of Errors. Automation 2018. Advances in Intelligent Systems and Computing, 743, 758–772. doi:10.1007/978-3-319-77179-3_75 (**Scopus**)
2. Zabolotnii, S. W., Warsza, Z. L., & Tkachenko, O. (2020). Estimation of Linear Regression Parameters of Symmetric Non-Gaussian Errors by Polynomial Maximization Method. Automation 2019. Advances in Intelligent Systems and Computing, 920, 636–649. doi:10.1007/978-3-030-13273-6_59 (**Scopus**)
3. Zabolotnii, S., Khotunov, V., Chepynoha, A., & Tkachenko, O. (2021). Estimating parameters of linear regression with an exponential power distribution of errors by using a polynomial maximization method . *Eastern-European Journal of Enterprise Technologies*, *1*(4 (109), 64–73. https://doi.org/10.15587/1729-4061.2021.225525 (**Scopus**)

Список публікацій які засвідчують апробацію матеріалів дисертації:

4. Заболотній С. В. Поліноміальні адаптивні процедури регресійного аналізу із використанням моделей негаусових помилок на основі статистик вищих порядків/ С.В.Заболотній, О.М.Ткаченко // Тези доповідей IV Міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи) – 2017» (ComInt – 2017): Київ, 16-18 травня 2017 р. – К: КНУ ім.Т.Шевченка, – 2017. С. 113-114
5. Заболотній С. В. Застосування методу максимізації полінома для оцінювання параметрів однофакторної лінійної регресії при негаусовому розподілі помилок / С.В.Заболотній, Ткаченко О.М. // Тези доповідей VI Міжнародної науково-практичної конференції “Обробка сигналів і негаусівських процесів – 2017” (ОСНП-2017): Черкаси, 24-26 травня 2017 р. – Черкаси: ЧДТУ, – 2017. С. 74-76.
6. Заболотній С.В. Аналіз ефективності поліноміальних оцінок параметрів лінійної регресії при симетричному розподілі негаусових помилок / С.В

Заболотній, М.П. Рудь, О.М. Ткаченко // Сучасні прилади, матеріали і технології для неруйнівного контролю і технічної діагностики машинобудівного і нафтогазопромислового обладнання, 14 - 16 листопада 2017: Тези доповідей 8-ма міжнародна н/т конф. – Івано-Франківськь, 2017. – С.130-131

7. Заболотній С.В. Особливості поліноміального оцінювання параметрів регресії при негаусовому симетричному розподілі помилок / С.В Заболотній, М.П. Рудь, О.М. Ткаченко // Автоматика та комп'ютерно-інтегровані технології у промисловості, телекомунікаціях, енергетиці та транспорті, 16 - 17 листопада 2017: Матеріали Всеукраїнської науково-практичної інтернет-конференції. – Кропивницький, 2017. – С.177-179
8. Заболотній С.В., Рудь М.П., Ткаченко О.М. Застосування методу максимізації полінома для оцінювання параметрів нелінійних регресійних моделей // Праці VII Міжнародної науково-практичної конференції «Обробка сигналів і негаусівських процесів», присвяченої пам'яті професора Ю.П. Кунченка: Тези доповідей. [Електронний ресурс]. – Черкаси: ЧДТУ, 2019, С. 76-79.

ДОДАТОК В

Документи про впровадження дисертаційної роботи

«ЗАТВЕРДЖУЮ»
Ректор
Черкаського державного
технологічного університету,
____________________ Григор О.О.

«___»________________ 2021р.

ДОВІДКА
про впровадження в освітній процес та наукову діяльність
результатів дисертаційної роботи Ткаченка Олександра Миколайовича

Дисертаційне дослідження Ткаченка О.М. проводилось у рамках держбюджетної НДР «Розробка технології та пристроїв адитивного виробництва індивідуальних хірургічних імплантатів та протезів з біосумісних полімерних матеріалів» (№ держ. реєстрації 0117U000937) .

Основні результати дисертаційної роботи впроваджені в освітній процес Черкаського державного технологічного університету при вивченні дисципліни «Теорія нелінійної статистичної радіотехніки», яка викладається студентам магістерського освітнього рівня спеціальності 172 Телекомунікації та радіотехніка на кафедрі радіотехніки, телекомунікаційних і робототехнічних систем.

До лекційних курсів та лабораторних робіт включено результати, що отримані автором:

- алгоритми побудови математичних моделей регресійних залежностей при негаусовому характері помилок з використанням моментно-кумулянтного опису;
- методи знаходження адаптивних поліноміальних оцінок параметрів регресійних моделей з використанням моделей негаусових помилок;
- створений в середовищі MATLAB програмний комплекс комп'ютерного статистичного моделювання, застосування якого надає можливості для дослідження ефективності поліноміальних методів регресійного аналізу.

Проректор з НДР та міжнародних зв'язків,
д.т.н., професор Е.В.Фауре

Декан ФЕТР,
к.т.н., доцент А.М.Чорній

Зав. кафедри РТРС
д.т.н., професор В.В. Палагін