НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

фізико-математичний факультет

кафедра математичного аналізу та теорії ймовірностей

«На правах рукопису»
УДК __519.2, 368.01____

До захисту допущено:
Завідувач кафедри
_____ Олег КЛЕСОВ
«___»_____20__ р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-професійною програмою «Страхова та фінансова математика»

зі спеціальності 111 «Математика»

на тему: «Узагальнені лінійні моделі в актуарній математиці»

Виконала:
студентка ІІ курсу, групи ОМ-41мп
Могилян Аліна Андріївна _____

Науковий керівник:
Доктор фізико-математичних наук, доцент
Василик Ольга Іванівна _____

Рецензент:
канд. фіз.-мат. наук, доцент
Київського національного університету
імені Тараса Шевченка
Яневич Тетяна Олександрівна _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____

Київ – 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
фізико-математичний факультет
кафедра математичного аналізу та теорії ймовірностей

Рівень вищої освіти – другий (магістерський)

Спеціальність – 111 «Математика»

Освітньо-професійна програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олег КЛЕСОВ
«__» _____ 20__ р.

ЗАВДАННЯ
на магістерську дисертацію
Могилян Аліні Андріївні

1. Тема дисертації «Узагальнені лінійні моделі в актуарній математиці», науковий керівник дисертації доктор фізико-математичних наук, доцент Василик Ольга Іванівна, затверджені наказом по університету від «06» листопада 2025 р. № 4843-с

2. Термін подання студенткою дисертації : «18» грудня 2025

3. Об’єкт дослідження: моделювання страхових ризиків за допомогою узагальнених лінійних моделей.

4. Предмет дослідження: побудова, аналіз та застосування узагальнених лінійних моделей та їх варіацій для визначення страхових тарифів та оцінювання збитків.

5. Перелік завдань, які потрібно розробити:

1) ознайомлення з літературою про узагальнені лінійні моделі;
2) проаналізувати технічні основи узагальненої лінійної моделі: функція зв’язку, випадкова та систематична частини;
3) побудова узагальненої лінійної моделі для страхових даних;

4) аналіз отриманих результатів.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу

7. Орієнтовний перелік публікацій:

Могилян А., Василик О. Узагальнені лінійні моделі в страховій математиці. Матеріали Двадцятої міжнародної наукової конференції імені академіка Михайла Кравчука, 17–20 листопада 2025 року, Київ, КПІ ім. Ігоря Сікорського, с.167-168.

8. Дата видачі завдання: «03» вересня 2025 року.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Опрацювання літератури по темі дисертації.	05.09.2025-17.09.2025	виконано
2.	Вивчення теоретичного матеріалу про узагальнену лінійну модель та її компоненти	18.09.2025-09.10.2025	виконано
3.	Побудова узагальнених лінійних моделей на основі страхових даних у застосунках RStudio та EXCEL.	10.10.2025-30.10.2025	виконано
4.	Аналіз отриманих результатів.	01.11.2025-10.11.2025	виконано
5.	Формування висновків	11.11.2025-20.11.2025	виконано
6.	Оформлення магістерської дисертації	21.11.2025-15.12.2025	виконано

Студентка Могилян А. А.

Науковий керівник дисертації Василик О.І.

Реферат

Магістерська дисертація: 54 сторінок, 21 першоджерел, 28 слайдів презентації.

Одним із найважливіших інструментів сучасної актуарної математики є узагальнена лінійна модель, яка дозволяє моделювати частоту страхових випадків, величину збитків, чисту премію та інші показники. У магістерській дисертації досліджується узагальнена лінійна модель, її компоненти та розширення класичної моделі. Особлива увага приділяється аналізу даних із страхових компаній, обробці пропусків, вибору розподілів та параметрів моделі.

Метою та завданням роботи є застосування методів теорії ймовірностей, математичної статистики, регресійного аналізу та актуарної математики для дослідження узагальнених лінійних моделей (УЛМ) та їх практичного використання в актуарних розрахунках. Вивчення теоретичних основ УЛМ, аналіз причин виникнення пропусків, методів підготовки страхових даних. Самостійною частиною роботи є побудова узагальнених лінійних моделей та їх застосування для заповнення пропусків у трикутнику розвитку з використанням програмного забезпечення EXCEL та середовища RStudio.

Об’єкт дослідження: моделювання страхових ризиків за допомогою узагальнених лінійних моделей.

Предмет дослідження: побудова, аналіз та застосування узагальнених лінійних моделей та їх варіацій для визначення страхових тарифів та оцінювання збитків.

Ключові слова: актуарна математика, страхування, узагальнена лінійна модель, функція зв’язку, експоненціальна сім’я, трикутник розвитку, дані з пропусками.

Abstract

Master's Thesis: 54 pages, 21 primary sources, 28 presentation slides.

One of the most important tools in modern actuarial mathematics is the generalized linear model (GLM), which makes it possible to model claim frequency, claim severity, pure premium, and other key indicators. In this master's thesis, we study generalized linear models, their components, and extensions of the classical model. Particular attention is paid to the analysis of insurance company data, handling missing values, and selecting appropriate distributions and model parameters.

The purpose and tasks of the thesis are to apply methods of probability theory, mathematical statistics, regression analysis, and actuarial mathematics to the study of generalized linear models (GLM) and their practical use in actuarial calculations. This includes studying the theoretical foundations of GLMs, analyzing the causes of missing data, and developing methods for preparing insurance datasets. An independent part of the work is the construction of generalized linear models and their application to fill in the missing data in the development triangle with the help of EXCEL software and RStudio environment.

Research object: modelling insurance risks using generalized linear models.

The subject of research: construction, analysis, and application of generalized linear models and their variations for determining insurance tariffs and estimating losses.

Keywords: actuarial mathematics, insurance, generalized linear model, link function, exponential family, development triangle, missing data.

Зміст

Вступ

Страхування є важливою частиною фінансової системи, оскільки допомагає людям убезпечитись від фінансових витрат у разі непередбачуваних подій. Для належної роботи страховим компаніям потрібно точно прогнозувати майбутні виплати та встановлювати обґрунтовані тарифи. Актуарні моделі допомагають робити такі прогнози, тому вони мають велике значення.

Одними з найбільш ефективних та поширених моделей є *узагальнені лінійні моделі (УЛМ)*. Вони дають можливість моделювати різні страхові показники: частоту страхових випадків, середній збиток, чисту премію тощо. Також за допомогою узагальнених лінійних моделей можна врахувати випадки з асиметрією розподілу збитків, нульовими значеннями виплат, використовуючи широкий клас розподілів.

Актуальність теми пов’язана з реальними ситуаціями, коли страхові дані можуть бути неповними та містити пропуски. Наприклад, під час побудови трикутника розвитку часто виникають пропущені значення. Пропуски можуть суттєво вплинути на результат. Використання узагальненої лінійної моделі дає можливість відновлювати пропуски надійнішим методом, ніж прості підстановки.

УЛМ складається з випадкової та систематичної частин. Випадкова частина описує тип розподілу змінної, а систематична частина пов’язує середнє значення з набором предикторів. Завдяки використанню функції зв’язку, модель можна застосувати до широкого спектру даних. Саме тому узагальнена лінійна модель широко застосовується у моделюванні збитків, оцінюванні частоти страхових подій та інших актуарних задачах.

У даній роботі досліджуються теоретичні основи побудови узагальнених лінійних моделей та розглядається їх практичне застосування до задач страхової математики.

У першому розділі даної роботи розглядаються причини виникнення пропусків в страхових даних, проблеми, які виникають при об’єднанні даних про страхові поліси і збитки, підготовка таких даних до подальшого моделювання.

У другому розділі аналізується структура експоненціальної сім’ї розподілів, описується роль випадкової та систематичної складової моделі, пояснюється значення функції зв’язку та інтерпретація параметрів. Також розглядаються підходи до оцінювання моделі, а саме критерії значущості та методи оцінки якості. Цей розділ формує фундамент для подальшого практичного застосування УЛМ.

У третьому розділі роботи розглядаються альтернативи класичної УЛМ. Сюди входять моделі зі змішаними ефектами (GLMM), моделі з прогнозуванням дисперсії (DGLM), узагальнені адитивні моделі (GAM), MARS, а також інші підходи, що дозволяють працювати з нелінійними залежностями. Це демонструє адаптацію узагальненої лінійної моделі до різних страхових завдань.

Останній розділ роботи пов’язаний із практичним застосуванням узагальненої лінійної моделі. У першому завданні наведений процес заповнення трикутника розвитку методом ланцюгових сходів. Застосування УЛМ дозволяє заповнити пропуски за допомогою програмного забезпечення. Також аналізуються результати відновлення та остаточні оцінки резервів. У другому завданні, на основі реальних даних, наданих страховою компанією, створюється трикутник розвитку та заповнюються пропущені дані, використовуючи УЛМ в програмному середовищі R. Розділ демонструє, як узагальнена лінійна модель приносить користь в обробці та аналізі реальних даних у страхуванні.

1. Причини виникнення пропусків

Очищення та підготовка даних - це не лише базовий крок у побудові моделі, але й один із найбільш трудомістких. Надійне моделювання починається з обробки інформації, тому актуарії повинні гарно орієнтуватись в цьому напряму, так як кожна компанія має свої особливості у збиранні, зберіганні та використанні даних.

Важливо розуміти, що процес підготовки даних рідко буває легким. Часто трапляється, що виявлена одна проблема веде до виявлення наступних. Результати побудови моделі можуть змусити повернутись назад, щоб перевірити підхід до обробки даних.

1. 1 Об’єднання даних про поліси та збитки

Одним із першочергових завдань під час створення моделі є об'єднання двох основних типів даних - страховий внесок на рівні ризику (демографічні характеристики поліса) та відомостей про збитки (тобто претензії). У більшості випадків саме така комбінація дозволяє побудувати точну модель класифікації ризиків. Для деяких ліній бізнесу може бути достатньо прикріпити збитки до записів поліса і моделювати на рівні поліса. Для інших - доцільніше моделювати на рівні окремих ризиків у межах поліса.[2]

Ідеальним варіантом є створення сукупності даних, де кожен рядок відповідає одному ризику в певному часовому інтервалі. Для деяких напрямів страхування цілком прийнятно працювати на рівні полісів, поєднуючи їх із пов’язаними ризиками. Водночас для інших видів страхування, таких як автострахування, доцільно деталізувати дані до рівня конкретного транспортного засобу або водія - тобто до рівня окремих ризиків.[1]

Однак на практиці дані про поліси та збитки часто зберігаються в окремих системах і підрозділах. Наприклад, андеррайтинг керує базами

полісів, тоді як інформацію про претензії обробляє інша команда. Через це зв’язок між цими джерелами може бути слабким або взагалі відсутнім. Одним із перших викликів для актуарія є виявлення, локалізація та об’єднання цих баз у цілісну структуру.

Якщо структура баз сучасна і синхронізована, об’єднання даних може виявитись доволі простим завданням. Проте в умовах застарілих або ізольовано побудованих систем можуть виникати численні труднощі - від різних форматів записів до серйозних логічних суперечностей.

Під час поєднання баз варто звернути увагу на кілька ключових питань:

- Чи синхронізовані у часі ці дані?

Якщо, наприклад, база полісів оновлюється щомісяця, а претензії — щоденно, то певна частина даних може бути некоректно прив'язана через часові розбіжності.

- Чи існує унікальний і стабільний ключ для об’єднання двох баз так, щоб кожен запис про збиток відповідав лише одному запису про поліс?

Ідеальна ситуація - коли кожний збиток однозначно відповідає одному полісу. Якщо ж зв’язок неоднозначний, можливі помилки: дублювання інформації або втрата даних.

- На якому рівні деталізації слід проводити об’єднання?

Це залежить від призначення моделі. Часто дані агрегують до рівня календарного року, що зручно як для інтерпретації, так і для моделювання сезонних ефектів. Наприклад, поліс, який діє з жовтня, покриває лише 25% відповідного року - цю пропорцію важливо враховувати.[1]

- До якого рівня деталізації слід узагальнювати набори даних перед злиттям?

Зазвичай це робиться на рівні поліса за рік. Наприклад, якщо поліс має дві претензії по 500 доларів, то підсумковий запис матиме значення "2" у полі кількості претензій і "1000" у полі загальних збитків.[1]

У деяких випадках об’єкти страхування мають складнішу структуру. Наприклад, у страхуванні комерційної нерухомості поліс може охоплювати кілька локацій. Тут можливе агрегування як за полісом, так і за кожною окремою локацією. Вибір рівня залежить від кількості таких випадків і мети моделі. Іноді доцільно зберігати найдрібніший рівень, аби забезпечити детальніший аналіз, але іноді для спрощення - об’єднувати дані вище, додаючи лише лічильник кількості об’єктів.[1]

Інші аспекти, що заслуговують уваги:

- Які поля можна безпечно видалити?

Часто зустрічаються змінні, що не мають практичного значення для моделі. Їх можна видалити, що дозволить пришвидшити подальшу обробку. Але видаляти дані слід дуже обережно: іноді пізніше може з’ясуватися, що певна змінна була цінною.

- Чи є дублікати або майже ідентичні поля?

Якщо дві змінні несуть однакову інформацію (або майже однакову), їх одночасна присутність може спричинити проблеми в моделі - це явище називають аліасингом (від англ. *aliasing* — накладення). У такому випадку краще залишити лише одне з полів.

- Чи не бракує важливої інформації?

Можливо, існують дані про страхувальника, які можуть бути корисними для прогнозування майбутніх збитків, і які збираються на етапі андеррайтингу, але не зберігаються для подальшого використання. Також трапляється, що цінні для моделі змінні взагалі не збираються. У таких

ситуаціях актуарію варто ініціювати обговорення щодо вдосконалення практик збору даних.

Загалом, підготовка даних - це не просто технічне завдання, а стратегічна робота. Успіх подальшого моделювання значною мірою залежить від якості, повноти й узгодженості початкових даних.

1. 2 Розподіл даних для моделювання

На початку будь-якого аналітичного проєкту дуже важливо розділити наявні дані щонайменше на дві групи. Основні з них - тренувальні дані (training set), який використовується для побудови моделі, та випробувальні дані (test set або holdout set), використовується для оцінки якості моделі [1]. Тренувальні дані потрібні для підбору змінних, перетворень, функцій зв’язку, параметрів і структури моделі. Випробувальна (тестова) вибірка дозволяє оцінити здатність моделі до узагальнення на нові дані.

Для чого це потрібно? Одна з причин полягає в тому, що спроба перевірити продуктивність будь-якої моделі на тому ж наборі даних, на якому модель була побудована, призведе до надмірно оптимістичних результатів. Такий підхід переоцінює її ефективність - модель добре «запам’ятовує» дані, а не вчиться узагальнювати закономірності. Якщо порівнювати її з іншими моделями, це створює хибну уяву про перевагу. Крім того, моделі можна зробити надмірно складними, додавши багато змінних, поліномів, функцій порогу або взаємодій. У результаті зростає ризик того, що модель точно описує випадкові коливання, а не справжні залежності.

У моделюванні за допомогою узагальнених лінійних моделей складність вимірюється кількістю ступенів свободи - це кількість параметрів, які підганяються під дані. Кожен параметр дає моделі більше гнучкості. Зв’язок між складністю і точністю ілюструє наступна тенденція:

спочатку зростання ступенів свободи зменшує помилку як на тренувальних, так і на тестових даних, але з певного моменту точність на нових даних починає падати. Це свідчить про перенавчання.[1]

Реальне уявлення про здатність моделі робити прогнози дає тестування на окремих даних. Тому важливо зберігати частину даних осторонь до кінця побудови моделі, а також продумати стратегію поділу ще до початку моделювання.

1. 2.1 Тренувальна та випробувальна вибірка

Найпростішим варіантом є розділення даних на дві частини: одна - тренувальна вибірка, інша - випробувальна вибірка. Тренувальні дані використовуються на всіх етапах побудови моделі: від первинного аналізу до остаточного налаштування. Випробувальні - лише після завершення побудови, щоб порівняти результати або оцінити ефективність.[1]

Типові пропозиції для поділу - 60/40 або 70/30. Більша навчальна частина дає змогу краще виявити структуру даних, але надто мала випробувальна вибірка може зробити оцінку ненадійною. Поділ можна зробити випадково або за часом. Якщо використовується часовий поділ, модель перевіряється не тільки на нових даних, а й у новому часовому періоді, що підвищує достовірність перевірки.[2]

Такий підхід особливо важливий у випадках, коли на результати впливають масові події (наприклад, природні катастрофи). Якщо подія потрапляє і в тренувальні, і в випробувальні дані - це спотворює результат. Тому поділ за часом допомагає уникнути таких перетинів.

1. 2.2 Тренувальна, валідаційна та випробувальна вибірки

Якщо обсяг даних дозволяє, варто розділити їх на три частини. Валідаційна вибірка використовується для уточнення моделі під час її побудови. Наприклад, модель створюється на тренувальній вибірці, перевіряється на валідаційній, після чого коригується, і лише потім -

остаточно оцінюється на випробувальній. У цьому випадку випробувальна вибірка не впливає на процес налаштування.[1]

Типовий розподіл: 40% тренувальні дані, 30% - валідаційні, 30% - випробувальні. Кожна частина повинна бути достатньо великою, щоб давати надійні оцінки.

1.2.3 Раціональне використання випробувальних даних

Використання випробувальної вибірки має бути обмеженим. Якщо її використовують часто - вона стає частиною побудови, а не незалежною оцінкою. Це знижує її цінність для прогнозування нових даних.

Тому слід чітко продумати, як саме використовувати кожну частину даних. Якщо є валідаційна вибірка, вона дає змогу вносити зміни до моделі. Але й її надмірне використання зменшує її незалежність. Основна робота з підбору структури моделі має ґрунтуватися на внутрішніх статистиках, як-от p-значення або F-критерій.

Щоб уникнути хаотичних змін у моделі, краще заздалегідь визначити рівні складності моделей, які будуть розглядатись. Наприклад: 1) залишити чинну модель; 2) змінити лише коефіцієнти; 3) додати нові змінні; 4) ввести взаємодії; 5) деталізувати категорії. Кожен рівень має більшу складність і вартість.[1]

Моделі всіх рівнів перевіряються (за можливості - спочатку на валідаційній), а остаточна версія - на випробувальній вибірці. Після вибору остаточної моделі, її слід перебудувати, використовуючи всі наявні дані, щоб параметри були оцінені максимально точно.

2. Узагальнені лінійні моделі

Узагальнені лінійні моделі (УЛМ) - це метод моделювання, який допомагає знайти залежність між тим, що ми хочемо передбачити (цільова змінна, позначається y), і факторами, що можуть на це впливати (пояснювальні змінні або предиктори). [6]

У практиці страхування майна чи нещасних випадків як цільову змінну зазвичай розглядають один з таких показників:

- **частота страхових подій** - скільки страхових випадків припадає на одиницю ризику;
- **середній збиток** - середня виплата на одну подію;
- **чиста страхова премія** - очікувані збитки на одиницю ризику;
- **коефіцієнт збитковості** - частка збитків від загальної суми нарахованої премії.

УЛМ дає можливість оцінити математичне сподівання числових показників.

В деяких застосуваннях цільовою змінною може бути подія або її відсутність. Наприклад: чи поновить клієнт свій страховий поліс або чи є підозра на шахрайство в заяві. В даних випадках УЛМ застосовується для оцінювання ймовірності настання події.

Пояснювальні змінні (предиктори) – це будь-які характеристики, які можуть вплинути на результат. Позначаються $x_1, x_2, \dots, x_p$, де p – кількість предикторів у моделі. У страхуванні це можуть бути, наприклад:

- для автомобілів: вік водія, марка авто, сімейний стан;
- для житла: тип будинку, його вік, страхова сума (AOI).[1]

2.1. Компоненти узагальненої лінійної моделі

Узагальнена лінійна модель базується на тому, що результат цільової змінної формується під впливом ***систематичної*** та ***випадкової*** складових.

Систематична складова - це частина варіації, яка пов’язана зі значеннями предикторів. Наприклад, якщо ми включили вік водія і він справді впливає на кількість аварій, то його ефект входить у цю частину.

Випадкова частина - це все інше, що впливає на результат, але чого модель не охоплює. Вона охоплює як чисту випадковість (тобто неконтрольовані або теоретично непередбачувані обставини), так і потенційно передбачувані ефекти від змінних, яких немає у моделі.[2] Наприклад, якщо вік водія не включили, хоча він важливий, його вплив потрапляє до випадкової складової.

Мета моделі - максимально точно пояснити поведінку цільової змінної через предиктори. Тобто, ми намагаємось зменшити частку випадковості і якомога більше змін пояснити систематично.

УЛМ ґрунтується на певних математичних припущеннях щодо обох частин - і систематичної, і випадкової. Далі розглядається кожна з них окремо, починаючи з випадкової.

2.1.1 Випадкова складова: експоненціальна сім’я розподілів

У межах узагальненої лінійної моделі (УЛМ) цільова змінна *y* трактується як випадкова величина, яка має певний ймовірнісний розподіл. При цьому вважається, що цей розподіл належить до експоненціальної сім’ї розподілів.

Експоненціальна сім’я - це клас розподілів, що мають корисні властивості для побудови УЛМ. До експоненціальної сім’ї належать,

зокрема, такі поширені розподіли, як нормальний, пуассонівський, біноміальний і гамма-розподіл.[1] Сюди ж належить і менш відомий, але дуже корисний у страхових задачах розподіл Твіді. Одним із ключових етапів побудови УЛМ є вибір відповідного розподілу та його характеристик.

В УЛМ припущення про **розподіл випадкової величини** y_i, , яка описує значення цільової змінної для i–го ризику, має вигляд:

$$y_i \sim Exponential\ (\mu_i, \varphi) \quad (1)$$

Термін "*Exponential*" не стосується конкретного розподілу, а використовується як позначення для будь-якого розподілу з експоненціальної сім'ї. Вираз в дужках відображає спільну рису, яка характерна для всіх розподілів цієї сім'ї, а саме кожен із них задається двома параметрами μ і φ, де:

- μ - математичне сподівання (середнє значення) розподілу випадкової величини;
- φ – параметр розсіювання, пов'язаний із дисперсією, але не є самою дисперсією.[1]

Параметр μ відіграє ключову роль - саме його значення модель намагається оцінити, і ця оцінка є основним прогнозом, який дає УЛМ.

Якщо не враховувати жодної додаткової інформації про ризик, то для всіх об'єктів прогноз μ буде однаковим і дорівнюватиме середньому значенню історичних даних. Але УЛМ дозволяє використати пояснювальні змінні, щоб отримати індивідуальні прогнози для кожного ризику, спираючись на статистичні зв'язки між предикторами та історичними значеннями цільової змінної.[5]

Індекс i у виразі μ_i вказує, що математичне сподівання може відрізнятись для кожного ризику. Натомість параметр розсіювання φ вважається сталим для всіх ризиків.

2.1.1 Систематична складова

Систематична складова УЛМ моделює залежність між математичним сподіванням цільової змінної μ_i (прогнозом моделі) та предикторами:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (2)$$

Це означає, що деяке *перетворення* параметра μ_i (позначене як $g(\mu_i)$) дорівнює *вільному члену* (позначеному β_0) плюс лінійна комбінація предикторів та *коефіцієнтів*, які позначені як $\beta_1, \ldots, \beta_p$. Ці коефіцієнти можна оцінити за допомогою програмного забезпечення для УЛМ.[5]

Функція $g(\cdot)$, яка застосовується до μ, називається *функцією зв'язку*, та її визначає дослідник.

Права частина рівняння (2) називається *лінійним предиктором*; після обчислення, вона дає значення $g(\mu_i)$ – тобто прогноз моделі, перетворений за допомогою обраної функції зв'язку. Значення $g(\mu_i)$ нас мало цікавить, наш головний інтерес полягає в значенні самого μ_i. Тому, після обчислення лінійного предиктора, прогноз моделі отримується шляхом застосування *оберненої* функції до результату, який представлений як $g(\cdot)$.

Функція зв'язку $g(\cdot)$ забезпечує гнучкість у встановленні залежності між прогнозом моделі та предикторами, оскільки дозволяє встановити не прямий зв'язок між середнім значенням і лінійним предиктором, а через певне перетворення.

Можливість використовувати функцію зв'язку відкриває більше варіантів для побудови моделі й, відповідно, дає більше шансів створити модель, що найкраще відображає реальність.

При використанні узагальнених лінійних моделей для побудови страхових тарифних планів додаткову перевагу отримують тоді, коли функцію зв'язку задають як натуральний логарифм (тобто $g(x) = \ln(x)$):

УЛМ з таким уточненням (її називають УЛМ із **лог-зв'язком**) має властивість формувати мультиплікативну структуру страхового тарифу [5]. Це випливає з того, що у випадку, коли функція зв'язку - логарифм, то рівняння (2) набуває вигляду:

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Щоб отримати μ_i, до обох частин рівняння застосуємо обернену функцію до натурального логарифма, тобто експоненту:

$$\begin{aligned}\mu_i &= \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \\ &= \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot \cdots \cdot \exp(\beta_p x_{ip})\end{aligned}$$

Отже, адитивні компоненти у лінійному предикторі перетворюються на мультиплікативні у прогнозі моделі.

Мультиплікативні моделі є найпоширенішим типом тарифної структури, що використовується для розрахунку страхових премій, завдяки ряду переваг порівняно з іншими підходами. Зокрема:

- вони прості й зручні у впровадженні;
- використання адитивних членів у моделі може призвести до від'ємних премій, що є нелогічним. Натомість мультиплікативна структура гарантує додатне значення премії без потреби вводити громіздкі виправлення, як-от правила про мінімальну премію;
- мультиплікативна модель має більш зрозумілу логіку. Наприклад, дивно стверджувати, що наявність порушення повинна збільшувати страхову премію на $500 незалежно від того, чи базова премія становить $1,000 чи $10,000. Набагато логічніше сказати, що надбавка за порушення становить 10%. [1]

З цих (та інших) причин моделі з логарифмічним зв'язком, які дають мультиплікативну структуру, є найбільш природним вибором у страхових розрахунках.

2.2 Дисперсія експоненціальної сім'ї

Деякі подробиці про експоненціальну сім'ю не є критично важливими з точки зору практичного застосування, тому їх розглядати не будемо. Водночас дуже важливим є розуміти перші два центральні моменти розподілів цієї сім'ї та те, як вони пов'язані з параметрами.

- **Середнє**: Для будь-якого розподілу з експоненціальної сім'ї, математичне сподівання дорівнює μ.
- **Дисперсія**: Вона має вигляд добутку параметра φ та функції від μ, яку позначають $V(\mu)$.

$$Var(y) = \varphi \cdot V(\mu)$$

Функція $V(\mu)$ варіюється залежно від обраного розподілу. У Таблиці 1 подано приклади цих функцій для різних випадків.[5]

Таблиця 1. Варіаційні функції експоненціальної сім'ї розподілів

Розподіл	Варіаційна функція $V(\mu)$	Дисперсія $\phi V(\mu)$
нормальний	1	ϕ
Пуассонівський	μ	$\varphi\mu$
гамма	μ^2	$\varphi\mu^2$
обернений гаусівський	μ^3	$\varphi\mu^3$
від'ємний біноміальний	$\mu(1+\kappa\iota)$	$\varphi\mu(1+\kappa\mu)$
біноміальний	$\mu(1-\mu)$	$\varphi\mu(1-\mu)$
Твіді	μ^p	$\varphi\mu^p$

Для нормального розподілу $V(\mu)$ - це просто константа, тому що дисперсія не залежить від μ. Але для більшості інших розподілів ця функція зростає, що є корисною властивістю в страхуванні: очікується, що вищий ризик (тобто більше значення μ) супроводжується більшою дисперсією.

Важливо також знати, що у межах узагальнених лінійних моделей (УЛМ) припускається, що φ однакове для всіх ризиків. Але це не означає, що сам рівень дисперсії не може змінюватися, ми все одно можемо моделювати зростання дисперсії зі зростанням ризику.

Таким чином, хоча φ є незмінним, водій із вищим ризиком має й більшу дисперсію, що логічно з погляду страхових ризиків.

Важливо пам’ятати, що сама варіаційна функція ще не є остаточною дисперсією. Щоб отримати справжнє значення дисперсії, її потрібно скоригувати, помноживши на параметр φ, який діє як коефіцієнт масштабування.

2.3 Значущість змінних

Для кожного предиктора, включеного в узагальнену лінійну модель, програмне забезпечення надає оцінку відповідного коефіцієнта. Водночас варто розуміти, що ці значення є приблизними, які є результатом випадкового процесу, оскільки вони отримані з даних, у яких наявна випадковість. Тобто, сам процес оцінювання має елемент випадковості.[8]

Якщо б ми аналізували інший набір даних з подібними властивостями, але з іншими результатами, отримані значення коефіцієнтів могли б відрізнятися.

У такій ситуації виникає потреба оцінити, наскільки отримане значення коефіцієнта наближається до свого справжнього (істинного) значення. Також важливо з’ясувати, чи дійсно предиктор має реальний вплив на результуючу змінну, чи ж спостережуване ненульове значення могло з’явитися випадково, навіть якщо насправді жодного ефекту немає, тобто справжній коефіцієнт дорівнює нулю.

Щоб допомогти відповісти на ці питання, інструменти для побудови УЛМ зазвичай надають додаткову статистичну інформацію для кожного параметра, зокрема:

- стандартну похибку;
- p-значення;
- довірчий інтервал.

2.3.1 Стандартна похибка

Як уже згадувалося, оцінка коефіцієнта, яку ми отримуємо, є випадковою величиною, бо залежить від вибіркових даних. Стандартна похибка - це оцінка середньоквадратичного відхилення цієї випадкової величини.

Наприклад, якщо стандартна похибка оцінки деякого коефіцієнта дорівнює 0.15, це означає, що якби ми неодноразово повторювали побудову моделі за вибіркою одного й того самого обсягу (з тими самими характеристиками, але різними вибірковими значеннями), то стандартне відхилення результуючих оцінок цього коефіцієнта було б приблизно рівним 0.15.

- Мала похибка свідчить про те, що оцінка, швидше за все, наближена до справжнього значення, отже, їй можна довіряти.
- Якщо ж похибка велика, це вказує на більшу невизначеність - отримана оцінка може бути ненадійною.

Зазвичай при збільшенні розміру вибірки точність зростає, а стандартна похибка зменшується - бо більше даних дають чіткішу картину.

Окрім цього, величина стандартної похибки залежить і від оцінки параметра φ: чим більше φ, тим більші стандартні похибки, оскільки зростає розсіювання результатів (вибіркових значень), і «сигнал» стає менш помітним на фоні «шуму».

2.3.2 P-значення

Статистика, що тісно пов’язана зі стандартною похибкою (та виводиться з неї) – це **p-значення** (*p-value*).

P-значення для певної оцінки коефіцієнта показує, наскільки ймовірно отримати таке саме або ще більше значення випадковим чином, якщо припустити, що істинне значення коефіцієнта дорівнює нулю.[1]

Уявімо, що для певної змінної в моделі коефіцієнт дорівнює 1.5, а відповідне p-значення - 0.0012. Це означає, що якщо справжній коефіцієнт насправді дорівнює нулю, то ймовірність отримати коефіцієнт 1.5 або ще більше значення випадково становить 0.0012.

За такої низької ймовірності можна зробити висновок, що спостережений ефект не є наслідком випадковості. Отже, коефіцієнт, найімовірніше, не дорівнює нулю, а саму змінну називають статистично значущою.

З іншого боку, якщо p-значення становить 0.52, це свідчить про те, що навіть у разі повної відсутності впливу змінної на результат, значення коефіцієнта на рівні 1.5 або вище могло з’явитися випадково. У цьому випадку немає підстав стверджувати, що змінна впливає на результат. Водночас це не гарантує, що змінна точно не має впливу - цілком можливо, що він є, але для його виявлення потрібно більше даних.

Статистичні тести значущості зазвичай формулюються у контексті нульової гіпотези, тобто припущення, що справжнє значення коефіцієнта дорівнює нулю. Якщо p-значення досить мале, ми відхиляємо цю гіпотезу й робимо висновок, що змінна має ненульовий вплив на очікуване значення результату.

У статистичній практиці часто використовується поріг 0.05: якщо p-значення менше або дорівнює 0.05, нульову гіпотезу зазвичай відкидають. Однак навіть при такому рівні 1 із 20 змінних може виявитися хибно

значущою - тобто буде прийнятою до моделі випадково. А в страхових моделях, де перевіряється велика кількість змінних, такий поріг може бути надто високим, щоб запобігти помилковим включенням у модель.[8]

2.3.3 Довірчий інтервал

Р-значення допомагає визначити, чи є підстави відхилити нульову гіпотезу, яка припускає, що істинне значення коефіцієнта дорівнює нулю. Але це лише один з можливих варіантів. Ми можемо так само перевіряти будь-яке інше значення, і тоді р-значення показуватиме, наскільки далеко отримана оцінка від гіпотетичного значення.

Тому логічно постає питання: які значення коефіцієнта не буде відхилено при певному рівні значущості (р-рівні)? Такий набір значень називається **довірчим інтервалом**. Він показує, у якому діапазоні, з певним рівнем надійності (довіри), може лежати справжнє значення коефіцієнта.[1]

Зазвичай довірчі інтервали пов’язують із рівнем значущості у відсотках. Наприклад, якщо p = 0.05, то ми говоримо про 95% довірчий інтервал.

Більшість програм для узагальнених лінійних моделей (наприклад, SAS) автоматично виводять саме 95% довірчі інтервали, але при потребі цей рівень можна змінити.

Приклад. Припустимо, що для певного предиктора УЛМ дає такі результати:

- оцінка коефіцієнта: 0.48;
- р-значення: 0.00056;
- 95% довірчий інтервал: [0.17; 0.79]

Оскільки р-значення низьке, ми маємо підстави відхилити гіпотезу, що коефіцієнт дорівнює нулю. Крім того, усі значення в інтервалі від 0.17

до 0.79 достатньо близькі до оцінки 0.48, що б при перевірці кожного з них як гіпотетичного значення, p-значення не перевищило б 0.05. Отже, якщо ми користуємось порогом 0.05, то гіпотези про значення в цьому інтервалі не відхиляються, і весь інтервал [0.17; 0.79] це допустимий діапазон для оцінки коефіцієнта.[1]

2.4 Типи предикторних змінних

У моделюванні ми працюємо з двома головними типами змінних - неперервними та категоріальними. Неперервні - це числові величини, які змінюються плавно, наприклад, вік або страхова сума. Категоріальні - це змінні, що розподіляють об’єкти по групах, як-от тип авто або регіон. Різні значення, яких може набувати категоріальна змінна, називаються рівнями.[8]

2.4.1 Типи предикторних змінних

Розглянемо неперервні змінні. З ними все досить просто: кожна неперервна змінна включається в УЛМ як є, і модель виводить для неї оцінку відповідного коефіцієнта. Цей коефіцієнт показує, як змінюється результат моделі, коли значення даної змінної збільшується на одиницю. Наприклад, якщо коефіцієнт додатній, то зростання змінної призводить до зростання прогнозованого значення, і навпаки.

Якщо ми використовуємо логарифмічний зв’язок (log link), то будуть незначні зміни в інтерпретації: замість зміни на фіксовану кількість, кожна одиниця приросту у змінній веде до сталого відсоткового зростання чи зменшення результату. Часто можна зустріти, що логарифмують саму змінну перед включенням у модель. Це роблять, що б модель стала ще зручнішою. Це дозволяє краще узгодити масштаб змінної зі шкалою лог-зв’язку.

Часто неперервні змінні логарифмують, щоб краще узгодити масштаб із моделлю. Наприклад, якщо логарифмувати страхову суму, то

коефіцієнт моделі показує, як буде змінюватись результат при зміні в кілька разів. Це дозволяє моделі гнучкіше реагувати на зростання змінної. Проте іноді логарифмування буває недоречним — наприклад, коли змінна має складну форму залежності або містить нулі.

2.5 Ваги

У багатьох випадках набір даних, що використовується в узагальненій лінійній моделі, містять не результати окремих ризиків, а середні значення для груп подібних ризиків. Наприклад, у таблиці збитковості один рядок може відповідати середній сумі збитку за кількома заявами, а у випадку з чистою премією рядок може представляти середню чисту премію для кількох однакових страхових випадків (наприклад, пов'язаних з одним клієнтом).[1]

У подібних ситуаціях зрозуміло, що рядки, які охоплюють більшу кількість ризиків, мають давати більший внесок у побудову моделі, адже вони ґрунтуються на більшій кількості інформації. Саме для цього в УЛМ є можливість використання змінної ваги, яка визначає, наскільки важливим є кожен запис у процесі оцінювання параметрів моделі.

Зазвичай вага позначається як ω і формально враховується в математичну частину УЛМ шляхом зміни припущень щодо дисперсії. Дисперсія для експоненціальної сім'ї має вигляд: $Var[y] = \varphi V(\mu)$. Якщо ж врахувати вагову змінну, то для спостереження i вона набуває вигляду:

$$Var[y_i] = \frac{\varphi V(\mu_i)}{\omega_i}$$

Тобто, чим більша вага, тим менша дисперсія. Таким чином, дисперсія обернено пропорційна до ваги.[5]

Якщо вага дорівнює кількості елементів, яку представляє кожен агрегований рядок, це узгоджується з очікуваннями щодо дисперсії.[5]

Адже для середнього значення n незалежних, однаково розподілених величин, дисперсія зменшується в n разів:

$$Var\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n}Var[X]$$

Отже, якщо рядок показує середній збиток за двома заявами,то його дисперсія повинна бути вдвічі меншою, ніж у випадку однієї заяви. Тому, задаючи вагу як кількість заяв, ми даємо моделі змогу правильно врахувати цю особливість.

2.6 Ймовірнісні розподіли, які застосовуються в УЛМ

У наступних підрозділах наведено опис кількох розподілів експоненціальної сім'ї, які застосовують в УЛМ, з особливим акцентом на ті типи цільових змінних, що зазвичай моделюють під час створення тарифних планів: середня величина збитку, частота, чиста премія та коефіцієнт збитковості.

2.6.1 Розподіли для страхових збитків

Для моделювання величини страхових збитків найчастіше застосовують гамма-розподіл та розподіл Вальда (обернений гаусівський розподіл).

Гамма-розподіл

Гамма-розподіл характеризується правосторонньою асиметрією, має різкий пік та довгий правий «хвіст», а значення гамма-розподіленої випадкової величини не можуть бути меншими за нуль. Такі характеристики відповідають реальним емпіричним розподілам страхових збитків, тому гамма-розподіл є найпопулярнішим розподілом, який використовується в УЛМ для цього типу даних.[1]

Варіаційна функція гамма-розподілу має вигляд:

$$V(\mu) = \mu^2.$$

Тобто дисперсія прямо пропорційна квадрату середнього значення. Це означає, що зі збільшенням очікуваного розміру збитку зростає і дисперсія, що є логічним і корисним у моделюванні страхових виплат.

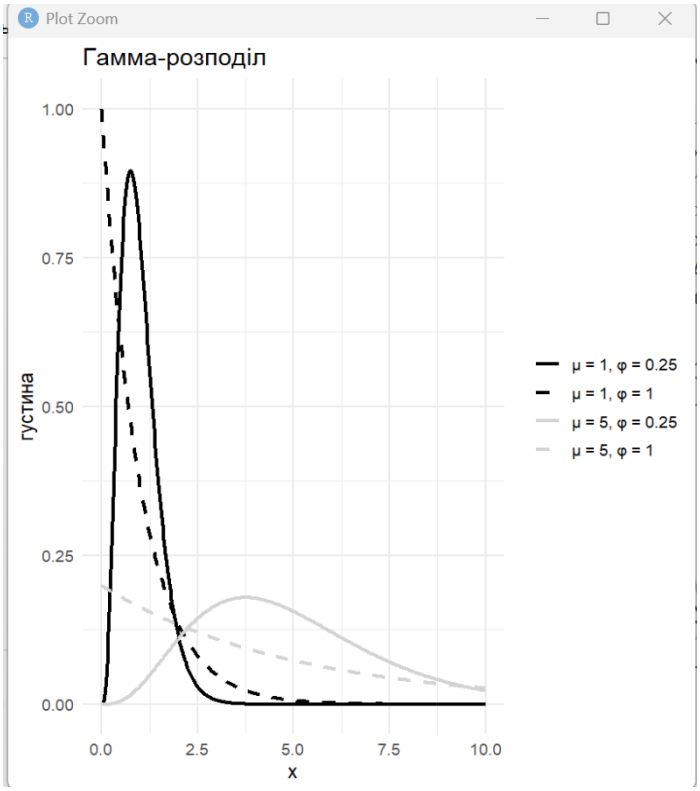

Рис. 1

На рисунку 1 наведено графіки кривих щільності ймовірності гамма-розподілу при різних значеннях параметрів μ (середнє) та φ (параметр розсіювання):

- Чорні лінії відповідають $\varphi = 1$ при середніх $\mu = 1$ та 5.
- Сірі лінії показують ті ж середні, але з меншими значеннями φ, що дає меншу дисперсію.

При цьому великий вплив на дисперсію відіграє саме μ. Навіть за однакового φ, розподіл з $\mu = 5$ (пунктирна лінія) має значно ширший розподіл, ніж з $\mu = 1$ (суцільна лінія). Це зумовлено квадратичною залежністю дисперсії від середнього, що є критичною властивістю при

моделюванні страхових виплат: більші очікувані виплати супроводжуються більшим розкидом.

Розподіл Вальда (обернений гаусівський розподіл)

Розподіл Вальда також має правосторонню асиметрію з мінімальною межею у нулі, тому його використовують для моделювання страхових виплат нарівні з гамма-розподілом. Проте він має свої особливості:

- **більш виражений пік**,
- **ширший хвіст**.

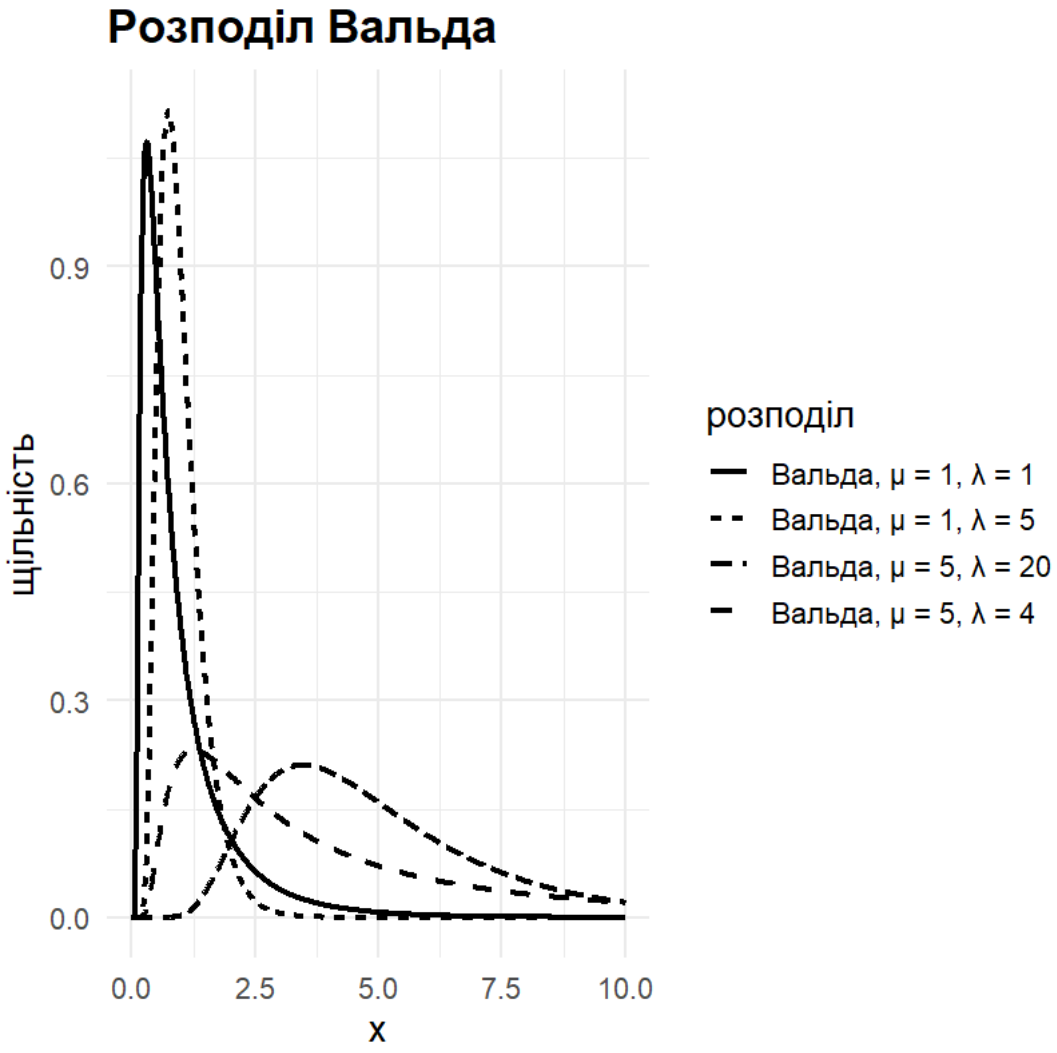

Рис. 2

Варіаційна функція для цього розподілу виглядає так:

$$V(\mu) = \mu^3,$$

тобто дисперсія тут зростає ще швидше зі збільшенням середнього значення порівняно з гамма-розподілом.[5]

На рисунку 2 показано порівняння щільностей розподілу Вальда (суцільні лінії) та гамма-розподілу (пунктирні лінії) з однаковими значеннями середнього та дисперсії. Видно, що розподіл Вальда:

- має вужчий та вищий пік,
- демонструє сильнішу асиметрію, ніж гамма-розподіл з тими ж параметрами.[1]

2.6.2 Розподіли для моделювання частоти збитків

При моделюванні частоти страхових випадків (наприклад, очікуваної кількості заяв на грошову одиницю премії) найчастіше використовується розподіл Пуассона. Іншим варіантом є від'ємний біноміальний розподіл.

Розподіл Пуассона

Розподіл Пуассона описує кількість подій, що відбуваються за певний проміжок часу, і широко застосовується у страхуванні для моделювання кількості позовів. Хоча цей розподіл дискретний (лише для цілих значень), в УЛМ він може приймати й дробові значення, що корисно при моделюванні частоти (наприклад, коли кількість страхових випадків ділиться на експозицію чи премію). У таких випадках вагу УЛМ зазвичай задають рівною знаменнику частоти.[1]

Для пуассонівського розподілу дисперсія має вигляд $Var[y] = \mu$, тобто зростає лінійно зі збільшенням математичного сподівання. Проте на практиці часто спостерігається надлишкова дисперсія, коли дисперсія перевищує середнє значення. Це відбувається через те, що крім випадковості самого процесу, є ще й різниця у рівнях ризику між страховими полісами.[5]

Щоб врахувати це, використовують пуассонівський розподіл з надлишковою дисперсією, який подібний до стандартного пуассонівського, але дозволяє параметру дисперсії φ набувати будь-яких додатних значень, а не лише 1. Це не змінює оцінки коефіцієнтів і прогнозів, проте коригує стандартні похибки та p-значення, роблячи діагностику моделі більш достовірною.

Від'ємний біноміальний розподіл

Інший спосіб врахувати надмірну дисперсію - це використати від'ємний біноміальний розподіл. Тут середнє значення пуассонівського розподілу розглядається як випадкова величина, яка має гамма-розподіл. Така комбінація призводить до від'ємного біноміального розподілу.

У ньому є додатковий третій параметр k (параметр надлишкової дисперсії), що збільшує дисперсію порівняно з пуасонівською. Варіаційна функція тут має вигляд:

$$V(\mu) = \mu(1 + k\mu),$$

тобто при $k \to 0$ від'ємний біноміальний розподіл наближається до пуасонівського. Таким чином, цей розподіл є гнучкішим інструментом для моделювання частоти страхових випадків, коли дисперсія сильно перевищує середнє.[1]

2.6.3 Розподіл для чистої премії: розподіл Твіді

Моделювання чистої премії (нетто-премії) або коефіцієнта збитковості на рівні окремого страхового полісу завжди було складним завданням. Це пояснюється тим, що більшість полісів взагалі не мають збитків, тобто значення таких виплат дорівнює нулю. А якщо збитки все ж виникають, їхні розміри розподілені нерівномірно: мають сильну правосторонню асиметрію. Тому функція щільності ймовірності, повинна мати в собі

велику ймовірність нульового результату та можливість великих виплат. Таким універсальним розподілом є **розподіл Твіді**.[1]

Розподіл Твіді має три параметри: середнє значення μ, параметр розсіювання φ та параметр степеня p. Значення p може бути будь-яким дійсним числом, за винятком проміжку від 0 до 1 (хоча значення 0 і 1 допустимі). Варіаційна функція для цього розподілу має вигляд:

$$V(\mu) = \mu^p.$$

Також при різних значеннях p розподіл Твіді набуває форми відомих розподілів: якщо $p = 0$, то це нормальний розподіл; при $p = 1$ - розподіл Пуассона; при $p = 2$ – гамма-розподіл; а при $p = 3$ - обернений гауссівський розподіл. Отже, розподіл Твіді об'єднує кілька розподілів експоненціальної сім'ї.[1]

Однак у страхуванні нас найбільше цікавлять значення, коли p знаходиться між 1 та 2. У цьому діапазоні розподіл Твіді утворює комбінацію розподіла Пуассона та гамма-розподілу. Це означає, що він одночасно враховує і кількість подій (частоту), і їхню величину, що чудово підходить для моделювання чистої премії або коефіцієнта збитковості. Розподіл Твіді можна уявити як «складний гамма-пуассонівський» розподіл. Це означає, що кількість страхових подій моделюється за допомогою розподілу Пуассона, а розмір кожної виплати – за допомогою гамма-розподілу. Середнє значення такого розподілу визначається так: $E[y] = \lambda \cdot \alpha \cdot \theta$, що відповідає базовій актуарній логіці: очікувана чиста премія дорівнює очікуваній кількості збитків, помноженій на їхній середній розмір.[5]

Від коефіцієнта варіації гамма-розподілу безпосередньо залежить параметр p. Коли коефіцієнт варіації прямує до нуля, p наближається до 1, а коли він стає дуже великим, p прямує до 2.

Параметр p можна визначити кількома способами. Деякі статистичні програми оцінюють його автоматично під час побудови моделі, але це потребує значних обчислень. Інший варіант – протестувати кілька значень p та обрати те, що оптимізує певну статистику, наприклад, логарифмічну правдоподібність або результати крос-валідації. Також багато актуаріїв обирають p на основі досвіду, беручи значення близько до 1.6-1.7, оскільки невеликі зміни цього параметра зазвичай не мають сильного впливу на прогноз.[1]

В УЛМ з розподілом Твіді параметри φ і p фіксуються для всіх спостережень, тоді як μ змінюється залежно від запису. Це означає, що модель припускає, що частота та середній збиток змінюються в одному напрямку, хоча на практиці це не завжди так.

2.6.3 Логістична регресія

У багатьох моделях ми хочемо передбачити цільову змінну, але вона не є числовим показником, а просто відображає, сталася певна подія чи ні. Такі змінні називають бінарними або дихотомічними. Наприклад:

- Чи продовжить клієнт страхову угоду.
- Чи перевищить нова заява про збиток визначений поріг.
- Чи буде подано регресну вимогу за збитком.

Для побудови такої моделі використовують дані про попередні подібні випадки, результат яких уже відомий. Тоді цільова змінна y_i набуває значення 1, якщо подія відбулася, і 0, якщо ні.[11]

В узагальненій лінійній моделі такі змінні моделюють за допомогою часткового випадку біноміального розподілу – розподілу Бернуллі. Прогноз моделі у цьому випадку – це математичне сподівання бінарної (індикаторної) випадкової величини, тобто ймовірність того, що подія станеться.[17]

Функція зв’язку

Для моделювання бінарної змінної потрібна спеціальна функція зв’язку. Чому не можна застосувати звичайну логарифмічну функцію зв’язку? Тому що лінійний предиктор в УЛМ (права частина рівняння) може мати будь-яке значення в інтервалі $[-\infty; +\infty]$, тоді як ймовірність – лише від 0 до 1. Тому потрібна функція зв’язку, яка відображає ймовірності з цього обмеженого проміжку на необмежену область.[1]

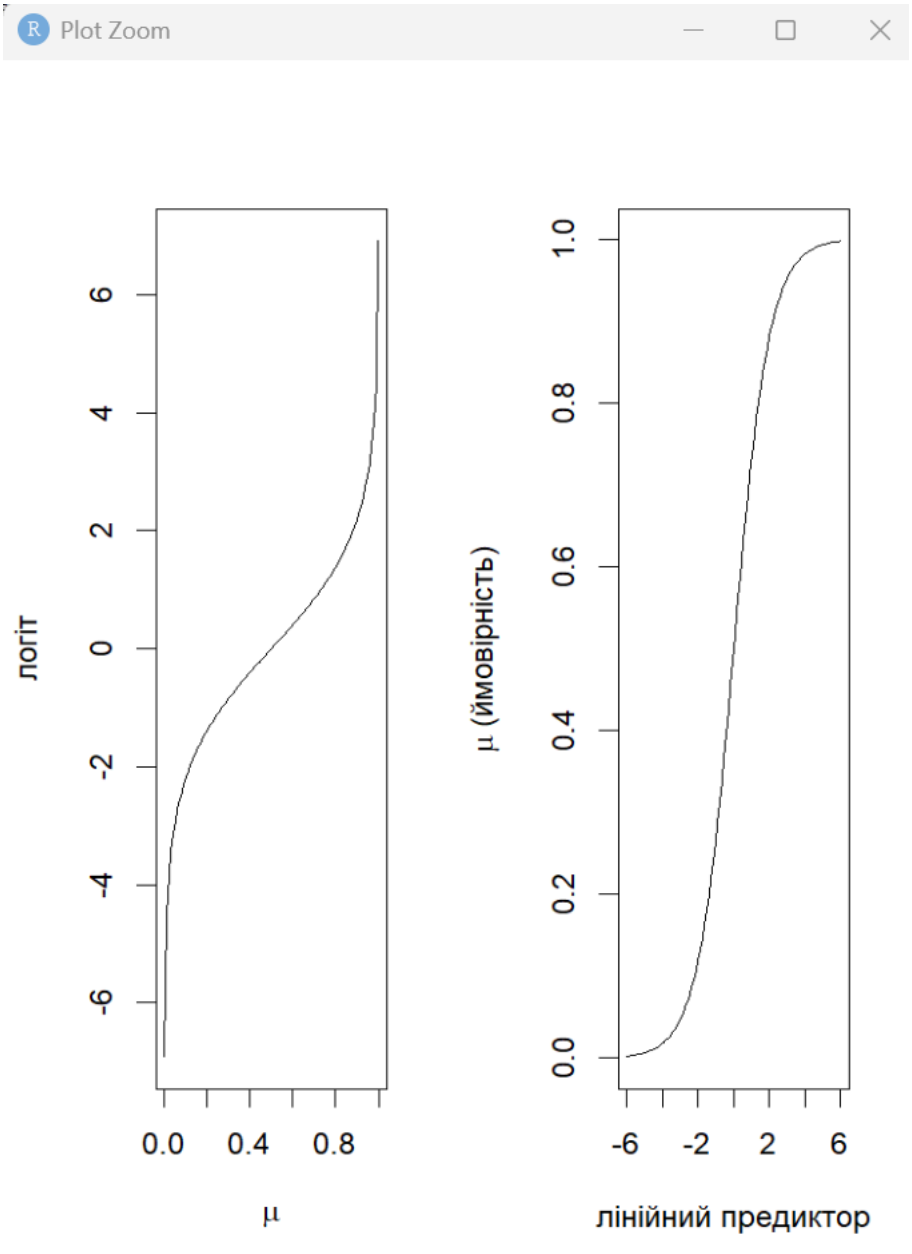

Рисунок 3. Логіт-функція (ліворуч) та її обернена функція - логістична функція (праворуч)

Для виконання цього завдання існує декілька функцій зв'язку, але найчастіше застосовується логіт-функція, яка визначається так:

$$g(\mu) = \ln \frac{\mu}{1-\mu}.$$

На лівій частині рисунка 3 зображено обернену функцію до логіт-функції, яка називається логістичною функцією та визначається так:

$$\frac{1}{1+e^{-x}}.$$

В узагальненій лінійній моделі ця функція перетворює значення лінійного предиктора у прогноз ймовірності. Ознакою низької ймовірності події є велике від'ємне значення предиктора, при високій ймовірності – предиктор додатній, при ймовірності 50% - рівний нулю.[1]

Модель логістичної регресії може бути підсумована наступним чином:

$$y_i \sim binomual(\mu_i)$$

$$\ln \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Інтерпретація результатів логістичної моделі

Логіт-функцію можна уявляти як натуральний логарифм відношення шансів. При цьому шанси визначаються як відношення ймовірності події до ймовірності її ненастання, тобто $\frac{\mu}{1-\mu}$. [1] Такий підхід дозволяє описати ймовірності у вигляді показників, що можуть набувати будь-яких додатних значень, проте сама ймовірність завжди лежить між 0 та 1. Наприклад, якщо подія майже гарантована, можна сказати, що «шанси мільйон до одного».

Якщо експонувати обидві частини рівняння УЛМ для логістичної регресії, отримаємо мультиплікативну форму, що безпосередньо описує шанси настання події. Завдяки цьому коефіцієнти УЛМ можна тлумачити

як вплив предикторів на шанси після експонування. Наприклад, якщо коефіцієнт для неперервної змінної дорівнює 0.24, то збільшення цієї змінної на одиницю збільшує шанси на $e^{0,24} - 1 = 27\%$. Так само, якщо 0.24 – це коефіцієнт для певного рівня категоріальної змінної, це означає, що шанси для цього рівня на 27% вищі, ніж для базового.[1]

3. Варіації узагальненої лінійної моделі

Узагальнена лінійна модель є хорошим інструментом, який може враховувати різні типи цільових змінних та зв'язки між ними. Незважаючи на це, будь-яка УЛМ має недоліки: прогнози базуються лише на лінійних зв'язках, погано працює з малими вибірками та при сильній кореляції між змінними, а також припускає, що випадкові змінні не залежать один від одного.

Існують сучасні методи, які можуть давати точніші прогнози, але для актуарних завдань зрозумілість моделі не менш важлива, ніж її точність. Тому було створено розширення УЛМ, які усувають частину її обмежень. Вони залишаються добре зрозумілими, але забезпечують більшу гнучкість, надійність та точність.

3.1 Узагальнені лінійні змішані моделі (GLMMs)

У звичайній узагальненій лінійній моделі коефіцієнти вважаються сталими, випадковим є лише результат. Це означає, що модель підлаштовується під дані, навіть коли їх мало.

Узагальнена лінійна змішана модель розширює УЛМ, дозволяючи деяким коефіцієнтам бути випадковими змінними. У цьому контексті предиктори з випадковими коефіцієнтами називають **випадковими ефектами**, а ті, що мають сталі коефіцієнти - **фіксованими ефектами**. На практиці, випадкові ефекти застосовуються для категоріальних змінних із багатьма рівнями, де бракує даних.[1]

Наприклад, у моделі автострахових збитків із трьома предикторами - вік, сімейний стан і територія — перші два можна задати як фіксовані ефекти, а територію як випадковий ефект.

Коефіцієнти фіксованих ефектів - β_1, β_2, а випадкових - $\gamma_1, \gamma_2, \dots, \gamma_{15}$.

Модель має вигляд:

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 z_1 + \cdots + \gamma_{15} z_{15}$$

$$y \sim gamma(\mu_i, \varphi)$$

$$\gamma \sim normal(\nu, \sigma)$$

GLMM одночасно враховує розподіл результатів і розподіл випадкових коефіцієнтів. Це створює ефект **стискання** - оцінки для рівнів із малою кількістю даних наближаються до середнього значення.[1]

Оцінювання GLMM проходить у два етапи:

1. спершу потрібно знайти параметри для фіксованих ефектів і розподілу випадкових коефіцієнтів;
2. потім, з використанням байєсівського підходу, оцінюють самі випадкові ефекти для всіх рівнів категоріальних змінних, враховуючи як оцінену випадковість параметра, так і обсяг даних.

Крім того, GLMM може враховувати кореляцію між спостереженнями. Наприклад, якщо дані містять кілька оновлень того самого страхового поліса за роки, і ці спостереження корельовані, можна ввести ID поліса як випадковий ефект. У цьому випадку модель оцінить ефект для кожного ID, хоча самі ці оцінки нас зазвичай не цікавлять — головне, щоб модель правильно врахувала кореляцію.[1]

3.2 УЛМ з моделюванням дисперсії (DGLMs)

У звичайній узагальненій лінійній моделі параметр розсіювання φ є сталим для всіх спостережень; в УЛМ з моделюванням дисперсії це

обмеження знімається, тому що кожне спостереження може мати власні значення φ_i та μ_i , які залежать від предикторів. Такі моделі називають **подвійними узагальненими лінійними моделями**.[1]

Математичний запис DGLM має вигляд:

$$y_i \sim Exponential(\mu_i, \varphi_i)$$

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

$$g_d(\mu_i) = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p \gamma_{ip}$$

В даних рівняннях $g_d(\cdot)$ - функція зв'язку для дисперсії (часто логарифмічна).

DGLM дозволяє змінювати дисперсію між спостереженнями, що робить модель гнучкішою та точнішою, особливо коли частина даних більш варіативна.

Алгоритм побудови:

1. Задати $\varphi_i = 1$.
2. Побудувати УЛМ для оцінювання коефіцієнтів β, використовуючи ваги $\frac{1}{\varphi_i}$. Якщо ми вже маємо вагу ω_i, слід використовувати $\frac{\omega_i}{\varphi_i}$.
3. Обчислити одиничне відхилення d_i.
4. Побудувати УЛМ для d_i (розподіл - гамма, предиктори - z).
5. Задати нові значення φ_i рівними прогнозам, отриманим на кроці 4.
6. Повторювати кроки 2–5, поки параметри моделі не перестануть істотно змінюватися між ітераціями (тобто, до досягнення збіжності моделі).[1]

Переваги DGLM:

- Моделює два параметри розподілу для кожного ризику, чим забезпечує гнучкішу форму кривої.

- Для розподілу Твіді враховує різний вплив предикторів на частоту та важкість подій.

3.3 Узагальнені адитивні моделі (GAMs)

У звичайній узагальненій лінійній моделі предиктори мають властивість лінійності. Для того, щоб врахувати нелінійні ефекти, можна застосувати перетворення змінних, але це потрібно робити вручну.

Узагальнена адитивна модель (**GAM**) - це модель, подібна до УЛМ, але вона враховує не лінійність безпосередньо. Її математичний запис має вигляд:

$$y_i \sim Exponential(\mu_i, \varphi)$$

$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip})$$

Як і в УЛМ, результат має розподіл з експоненціальної сім'ї. Головна відмінність: доданки у лінійному предикторі більше не є лінійними функціями предикторів. Функції $f_1, f_2, \ldots, f_p$ можуть бути будь-якими плавними кривими, що описують вплив предикторів. Їх форму визначає програма під час навчання моделі.

Слово «адитивна» в назві моделі означає, що предиктор складається з адитивних доданків, хоча вони не обов'язково лінійні. Як і в УЛМ, можна задати лог-зв'язок, який зробить модель мультиплікативною.

В УЛМ вплив змінної видно з коефіцієнта, а в узагальненій адитивній моделі ефекти аналізують графічно. Різні методи оцінювання функцій дозволяють регулювати ступінь згладжування — надмірна гнучкість може призвести до перенавчання.

3.4 Багатовимірні адаптивні регресійні слайни або моделі MARS

Ще один різновид УЛМ, який добре працює з нелінійними залежностями – це багатовимірні адаптивні регресійні слайни (MARS). На відміну від узагальненої адитивної моделі, яка підбирає плавні функції для предикторів, MARS моделі додають до звичайної УЛМ кусково-лінійні функції, автоматично визначаючи їх вигляд і точки розривів.[1]

Як і у GAM, у MARS є параметри гнучкості: чим вона більша, тим більше точок розриву з’являється, але й зростає ризик “перенавчання”.

Окрім природної здатності враховувати нелінійність, MARS має ще кілька корисних властивостей:

- Автоматичний вибір змінних - залишає лише статистично значущі предиктори.
- Пошук взаємодій - може виявляти не лише парні, а й багатоступеневі взаємодії між змінними та кусково-лінійними функціями.[1]

MARS також корисний навіть у випадках, коли кінцева модель має бути стандартною УЛМ. Він допомагає знайти можливі нелінійні перетворення чи взаємодії, які потім можна вручну додати до УЛМ. Однак, надмірне використання може призвести до випадкових закономірностей.

4. Практичне застосування УЛМ

4.1 Пошук пропущених даних у трикутнику розвитку

У цьому розділі проводиться аналіз даних, що містять інформацію про розміри страхових виплат за різні періоди. Дані були отримані з відкритих інтернет-джерел. Набір даних містить дату настання події (Accident date), дату подання заявки (ReportDate), загальні збитки (TotalLoss), сумарну виплату (TotalPayment), дату виплати (PaymentDate), закриття договору (CloseDate) за 2001-2010 роки настання страхового випадку. Дані містять 12011 позовів.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Simula	Occu	Claim	ClaimMa	Conve	AccidentDa	ReportDate	Covari	Line	Type	TotalLoss	ClaimLiability	Limit	Deduc	TotalPayment	PaymentDate	CloseDate
2	1	1	1	FALSE	NA	1/9/2001	7/8/2001	NA	Line 1	Type 1	6327.80999501795	TRUE	1,00E+07	0	6327.809995017	11/1/2001	11/1/2001
3	1	2	2	FALSE	NA	1/13/2001	6/24/2001	NA	Line 1	Type 1	5303.40102570024	TRUE	1,00E+07	0	5303.401025700	10/19/2001	10/19/2001
4	1	3	3	FALSE	NA	1/24/2001	6/30/2001	NA	Line 1	Type 1	5596.79088940776	TRUE	1,00E+07	0	5596.790889407	8/7/2001	8/7/2001
5	1	4	4	FALSE	NA	1/9/2001	3/13/2001	NA	Line 1	Type 1	4480.81361145459	TRUE	1,00E+07	0	4480.813611454	12/5/2003	12/5/2003
6	1	5	5	FALSE	NA	1/26/2001	8/28/2001	NA	Line 1	Type 1	5165.95522117711	TRUE	1,00E+07	0	5165.955221177	7/25/2009	7/25/2009
7	1	6	6	FALSE	NA	1/16/2001	3/12/2001	NA	Line 1	Type 1	5598.71053298209	TRUE	1,00E+07	0	5598.710532982	8/19/2004	8/19/2004
8	1	7	7	FALSE	NA	1/26/2001	1/27/2001	NA	Line 1	Type 1	4989.5762600284	TRUE	1,00E+07	0	4989.576260028	6/28/2025	6/28/2025
9	1	8	8	FALSE	NA	1/7/2001	2/28/2001	NA	Line 1	Type 1	4780.24270865942	TRUE	1,00E+07	0	4780.242708659	12/3/2001	12/3/2001
10	1	9	9	FALSE	NA	1/12/2001	1/22/2001	NA	Line 1	Type 1	3983.19647214058	TRUE	1,00E+07	0	3983.196472140	8/27/2001	8/27/2001
11	1	10	10	FALSE	NA	1/17/2001	8/17/2001	NA	Line 1	Type 1	4854.51318502013	TRUE	1,00E+07	0	4854.513185020	7/14/2002	7/14/2002
12	1	11	11	FALSE	NA	1/14/2001	5/16/2001	NA	Line 1	Type 1	3913.18412119817	TRUE	1,00E+07	0	3913.184121198	8/20/2001	8/20/2001
13	1	12	12	FALSE	NA	1/14/2001	12/4/2001	NA	Line 1	Type 1	5814.79897597526	TRUE	1,00E+07	0	5814.798975975	12/27/2003	12/27/2003
14	1	13	13	FALSE	NA	1/15/2001	5/4/2002	NA	Line 1	Type 1	3827.82874712577	TRUE	1,00E+07	0	3827.828747125	6/19/2004	6/19/2004
15	1	14	14	FALSE	NA	1/27/2001	4/11/2001	NA	Line 1	Type 1	3827.54267612659	TRUE	1,00E+07	0	3827.542676126	5/25/2001	5/25/2001
16	1	15	15	FALSE	NA	1/24/2001	3/20/2001	NA	Line 1	Type 1	5411.10536367597	TRUE	1,00E+07	0	5411.105363675	11/30/2002	11/30/2002
17	1	16	16	FALSE	NA	1/21/2001	8/16/2001	NA	Line 1	Type 1	6106.53861105327	TRUE	1,00E+07	0	6106.538611053	1/27/2004	1/27/2004
18	1	17	17	FALSE	NA	1/21/2001	5/3/2002	NA	Line 1	Type 1	4393.89996062128	TRUE	1,00E+07	0	4393.899960621	3/25/2004	3/25/2004
19	1	18	18	FALSE	NA	1/13/2001	2/3/2001	NA	Line 1	Type 1	5846.44879818178	TRUE	1,00E+07	0	5846.448798181	1/30/2005	1/30/2005
20	1	19	19	FALSE	NA	1/4/2001	1/21/2001	NA	Line 1	Type 1	5439.69449653924	TRUE	1,00E+07	0	5439.694496539	8/29/2004	8/29/2004
21	1	20	20	FALSE	NA	1/17/2001	1/28/2001	NA	Line 1	Type 1	4457.15337292134	TRUE	1,00E+07	0	4457.153372921	2/16/2001	2/16/2001
22	1	21	21	FALSE	NA	1/11/2001	2/12/2001	NA	Line 1	Type 1	6003.45923553127	TRUE	1,00E+07	0	6003.459235531	3/18/2001	3/18/2001

Рис. 4 Набір даних (перші 21 позовів)

Під час формування трикутника розвитку часто виникають пропущені дані. Наявність попусків ускладнює аналіз, оскільки метод ланцюгових сходів передбачає повний трикутник. Тому основним завданням буде відновити пропуски із використанням узагальненої лінійної моделі.

	12	24	36	48	60	72	84	96	108	120
2001	357848	766940	610542	482940	527326	574398	146342		227229	67948
2002	352118	884021	933894		445745	320996	527804	266172	425046	
2003	290507	1001799	926219	1016654	750816		495992	280405		
2004	310608		776189	1562400	272482	352053	206286			
2005	443160	693190	991983	769488	504851	470639				
2006	396132	937085		805037	705960					
2007	440832	847631	1131398	1063269						
2008	359480	1061648	1443370							
2009	376686	986608								
2010	344014									

Рис. 5 Трикутник розвитку з пропущеними даними

Якщо проігнорувати пропуски і оцінити необхідні резерви на основі тільки наявних даних, то отримаємо зміщені оцінки резервів. Оцінювати дані резерви буду методом ланцюгових сходів, але спочатку потрібно утворити трикутник з накопиченими сумами, а потім застосувати метод ланцюгових сходів для оцінювання остаточних збитків та резервів.

	Накопичені суми										
	12	24	36	48	60	72	84	96	108	120	Резерв
2001	357848	1124788	1735330	2218270	2745596	3319994	3466336	3466336	3693565	3761513	0
2002	352118	1236139	2170033	2170033	2615778	2936774	3464578	3730750	4155796	4232247,35	76451,35
2003	290507	1292306	2218525	3235179	3985995	3985995	4481987	4762392	5194009,64	5289560,29	527168,29
2004	310608	310608	1086797	2649197	2921679	3273732	3480018	3646680,08	3977180,26	4050345,74	570327,74
2005	443160	1136350	2128333	2897821	3402672	3873311	4267741,53	4472128,59	4877439,53	4967166,46	1093855,46
2006	396132	1333217	1333217	2138254	2844214	3156024,33	3477411,47	3643948,72	3974201,36	4047312,04	1203098,04
2007	440832	1288463	2419861	3483130	4212844,83	4674697,76	5150735,83	5397410,51	5886580,15	5994871,55	2511741,55
2008	359480	1421128	2864498	4111588,71	4972965,48	5518150,21	6080079,50	6371261,52	6948691,69	7076522,04	4212024,04
2009	376686	1363294	2379255,30	3415090,27	4130550,80	4583381,85	5050120,96	5291977,08	5771591,24	5877767,27	4514473,27
2010	344014	1086236,5	1895727,55	2721053,39	3291113,37	3651917,14	4023802,49	4216507,03	4598650,88	4683249,13	4339235,13
Суми		10506293	15956594	18791884	18515934	17389806	14892919	11959478	7849361	3761513	19048374,9
Суми без останнього даного	3327371	9142999	13092096	15308754	15671720	13516495	11412901	7197086	3693565		
Коефіцієнт розвитку	3,15754	1,7452254	1,435361	1,20949974	1,1096297	1,10183291	1,04789115	1,09063043	1,018396319		

Рис. 6 Трикутник розвитку з накопиченими сумами та оцінки резервів, отримані методом ланцюгових сходів

Сумарне значення резерву складає 19048374,9, але така оцінка є неповною, оскільки базується лише на спостереженнях без заповнення пропусків.

Наступним кроком знайдемо пропущені значення методом узагальненої лінійної моделі та порівняємо сумарні значення резервів з попереднім.

Модель для страхових виплат

- Кожну страхову виплату (збитки) з трикутнику розвитку, тобто $C_{i,j}$ можна записати в загальних термінах так:

$$C_{i,j} = r_j \cdot s_i \cdot x_{i+j} + \varepsilon_{i,j}$$

- $C_{i,j}$ - збитки, сплачені на кінець -го року розвитку сплати збитків за страховими подіями, що настали в i-му році настання збитків;
- r_j - фактор розвитку для j-го року, який репрезентує частку виплат в j-му році; кожне r_j не залежить від i;

- s_i - параметр, який міняється в залежності від початкового року і року виникнення збитків та репрезентує експозицію, наприклад, кількість позовів, що виникли в i-му році;
- x_{i+j} - параметр, який залежить від календарного року $k = i + j$, і, наприклад, репрезентує інфляцію;
- $\varepsilon_{i,j}$ - похибка.

Для побудови УЛМ використаємо припущення про гамма-розподіл виплат та логарифмічну функцію зв'язку (яка є стандартною для моделювання страхових виплат, коли вони мають таку мультиплікативну структуру, як наведено вище):

$$y_i \sim Gamma\ (\mu_i,\ \varphi)$$

$$\ln(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

де $\mu_i = Ey_i$ - очікуване значення виплати в трикутнику розвитку; x_1- рік настання збитку; x_2 - період розвитку.

Переведемо трикутник розвитку у таблицю та знайдемо функцію ln від значень.

Роки	Місяці	Значення	ln(Значення)
2001	12	357848	12,78786359
2001	24	766940	13,55016385
2001	36	610542	13,32210237
2001	48	482940	13,0876477
2001	60	527326	13,17557423
2001	72	574398	13,26107781
2001	84	146342	11,89370163
2001	108	227229	12,3337136
2001	120	67948	11,12649799
2002	12	352118	12,77172163
2002	24	884021	13,6922361
2002	36	933894	13,74711822
2002	60	445745	13,00750232
2002	72	320996	12,67918394

Рис.7 (перші 14 значень з 50)

За допомогою надбудови «Аналіз даних», знайдемо коефіцієнти β_0, β_1, β_2, що дозволить обчислити прогнозовані значення та заповнити пропуски.

Результат										
								розвитку	3,345078	1,7

Регресійна статистика	
Множинний R	0,459934
R-квадрат	0,211539
Нормований R-квадрат	0,177988
Стандартна похибка	0,559468
Спостереження	50

Дисперсійний аналіз

	df	*SS*	*MS*	*F*	*начимість F*
Регресія	2	3,946921	1,97346	6,304904	0,003753
Залишок	47	14,71119	0,313004		
Сумарно	49	18,65811			

	Коефіцієнт	*дартна по статисти*	*-Значення*		*ижнє 95%*	*ерхнє 95%*	*Нижнє 95,0%*	*ерхнє 95,0%*
β_0	-66,565	72,58295	-0,91709	0,363778	-212,583	79,45304	-212,58309	79,45304
β_1	0,039963	0,036183	1,104481	0,275007	-0,03283	0,112754	-0,0328272	0,112754
β_2	-0,00739	0,003049	-2,42278	0,019309	-0,01352	-0,00125	-0,0135207	-0,00125

Рис. 8 Результат використання функції «Регресія» в EXCEL

Отримані значення:

	12	24	36	48	60	72	84	96	108	120
2001	357848	766940	610542	482940	527326	574398	146342	325197,9	227229	67948
2002	352118	884021	933894	482495,8	445745	320996	527804	266172	425046	
2003	290507	1001799	926219	1016654	750816	420585,5832	495992	280405		
2004	310608	624022,363	776189	1562400	272482	352053	206286			
2005	443160	693190	991983	769488	504851	470639				
2006	396132	937085	618606,1829	805037	705960					
2007	440832	847631	1131398	1063269						
2008	359480	1061648	1443370							
2009	376686	986608								
2010	344014									

Рис. 9 Заповнені пропущені дані

Знайдемо сумарний резерв для даного трикутнику розвитку методом ланцюгових сходів, спочатку звівши до трикутника з накопиченими сумами. Отримаємо значення сумарного резерву 19900600.

	Накопичені суми										
	12	24	36	48	60	72	84	96	108	120	Резерв
2001	357848	1124788	1735330	2218270	2745596	3319994	3466336	3791533,9	4018762,9	4086710,9	0
2002	352118	1236139	2170033	2652528,8	3098273,8	3419269,8	3947073,8	4213245,8	4638291,8	4716714,64	78422,80
2003	290507	1292306	2218525	3235179	3985995	4406580,6	4902572,6	5182977,6	5605316,09	5700089,03	517111,45
2004	310608	934630,4	1710819,4	3273219,4	3545701,4	3897754,4	4104040,4	4394540,91	4752633,10	4832989,15	728948,79
2005	443160	1136350	2128333	2897821	3402672	3873311	4227702,15	4526955,97	4895838,10	4978615,41	1105304,41
2006	396132	1333217	1951823,2	2756860,2	3462820,2	3904215,46	4261434,23	4563075,75	4934901,12	5018338,90	1555518,72
2007	440832	1288463	2419861	3483130	4138942,18	4666520,69	5093487,08	5454024,66	5898449,62	5998178,78	2515048,78
2008	359480	1421128	2864498	4099899,46	4871838,49	5492837,10	5995407,86	6419787,03	6942907,80	7060296,33	4195798,33
2009	376686	1363294	2400690,65	3436061,16	4083011,10	4603460,25	5024656,89	5380322,39	5818741,66	5917123,15	4553829,15
2010	344014	1150754	2026418,02	2900372,12	3446461,23	3885771,28	4241302,49	4541519,00	4911587,80	4994631,41	4650617,41
Суми		11130315	17199223	20517008	20241058	18916910	16420023	13187757	8657054,8	4086710,9	19900600
Суми без останнього даного	3327371	9767021	14334725	17033878	16778238	15043599	12315982	8004779,8	4018762,9		
Коефіцієнт розвитку	3,345078	1,760949	1,4312803	1,1882824	1,127467	1,0914957	1,070784	1,0814857	1,0169077		

Рис. 10 Знаходження сумарного резерву

У випадку, коли пропущені дані були відновлені за допомогою узагальненої лінійної моделі, сумарний резерв виявився більшим. Це пов’язано з тим, що УЛМ враховує динаміку зміни збитків між періодами розвитку. В результаті прогнозовані значення мають тенденцію до зростання.

4.2 Застосування УЛМ для аналізу реальних даних

Для виконання даного завдання використаємо реальні дані зі страхової компанії. Отримані дані містять інформацію щодо:

- дати настання страхового випадку (Accident Date);
- дати виплати страхового відшкодування (Payment Date);
- суми виплати (Amount Paid).

Accident Date Дата настання Страхового Випадку	**Payment Date** Дата врегулювання страхової вимоги (виплати)	**Amount paid (for Paid) Сума страхового відшкодування (для виплаченої вимоги)**
29.07.2016	31.01.2019	87 319,25
27.03.2017	31.01.2019	37 738,29
02.11.2016	31.01.2019	650 000,00
12.05.2017	31.01.2019	246 408,70
20.10.2017	31.01.2019	15 286,77
06.04.2018	31.01.2019	74 436,38
18.02.2018	31.01.2019	559 761,74
25.11.2017	31.01.2019	650 000,00
08.03.2018	31.01.2019	51 086,82
13.10.2017	31.01.2019	104 554,32

Рис. 11 Перші 10 записів з 29942

Для побудови трикутника дані були згруповані за:

- роком настання збитку (accident year);
- періодом розвитку (development period) — різницею між роком виплати та роком збитку.

Побудову трикутника розвитку будемо виконувати за допомогою програмного забезпечення R. Код програми наведений у Додатку 1.

```
      DEV
ORIGIN         0        1       2         3         4        5        6        7
  2007        NA       NA      NA        NA        NA       NA       NA 42009.83
  2010        NA       NA      NA        NA 244153.91  71385.4 68266.32 16776.76
  2011        NA       NA      NA  413824.2  29095.23       NA       NA       NA
  2012        NA       NA 1505760  841240.4 796740.55 235466.1       NA       NA
  2013        NA  4197941 1670173  684816.1        NA       NA       NA       NA
  2014  40231751 11001298 2625832 2940708.4 317619.05 110076.6       NA       NA
  2015 298380876 13688468 7200553 3223202.8 430912.17       NA       NA       NA
  2016 494434172 27374233 6895613 2808449.2        NA       NA       NA       NA
  2017 781514453 44445441 6441970        NA        NA       NA       NA       NA
  2018 776609039 31510423      NA        NA        NA       NA       NA       NA
  2019 509529851       NA      NA        NA        NA       NA       NA       NA
>
```

Рис. 12 Трикутник розвитку

Отримано таблицю, у якій:

- **рядки** — роки настання збитків (Accident Year),
- **стовпці** — кумулятивні виплати за відповідні періоди розвитку (0, 1, 2, … років).

Верхня частина трикутника містить пропуски — це нормальне явище, оскільки останні роки ще не мають повного горизонту спостереження. Саме ці пропуски далі заповнюємо, використовуючи узагальнену лінійну модель.

Отриманий трикутник розвитку експортуємо в EXCEL, та будемо використовувати узагальнену лінійну модель для заповнення пропусків.

		Період розвитку							
		0	1	2	3	4	5	6	7
Рік, у якому стався збиток	2007								**42009,83**
	2010					**244153,9**	**71385,4**	**68266,32**	**16776,76**
	2011				**413824,2**	**29095,23**			
	2012			**1505760**	**841240,4**	**796740,6**	**235466,1**		
	2013		**4197941**	**1670173**	**684816,1**				
	2014	**40231751**	**11001298**	**2625832**	**2940708**	**317619,1**	**110076,6**		
	2015	**298380876**	**13688468**	**7200553**	**3223203**	**430912,2**			
	2016	**494434172**	**27374233**	**6895613**	**2808449**				
	2017	**781514453**	**44445441**	**6441970**					
	2018	**776609039**	**31510423**						
	2019	**509529851**							

Рис. 13 Трикутник розвитку у EXCEL

Запишемо значення із трикутника розвитку у таблицю та знайдемо функцію ln від значень.

Рік	Період	Значення	ln
2007	7	42009,83	10,64566
2010	4	244153,91	12,40555
2010	5	71385,4	11,17585
2010	6	68266,32	11,13117
2010	7	16776,76	9,72775
2011	3	413824,2	12,9332
2011	4	29095,23	10,27833
2012	2	1505760	14,22481
2012	3	841240,4	13,64263
2012	4	796740,55	13,58828
2012	5	235466,1	12,36932
2013	1	4197941	15,2501
2013	2	1670173	14,32844
2013	3	684816,1	13,43691
2014	0	40231751	17,51017
2014	1	11001298	16,21352
2014	2	2625832	14,78091
2014	3	2940708,4	14,89416
2014	4	317619,05	12,66861

Рис. 14. Перші 19 значень з 35

За допомогою надбудови «Аналіз даних», знайдемо коефіцієнти β_0, β_1, β_2, що дозволить обчислити прогнозовані значення та заповнити пропуски.

Результат

Регресійна статистика	
Множинний R	0,942797
R-квадрат	0,888867
Нормований R-квадрат	0,881921
Стандартна похибка	1,026987
Спостереження	35

Дисперсійний аналіз

	df	*SS*	*MS*	*F*	*ачимость F*
Регресія	2	269,9432	134,9716	127,9713	5,41E-16
Залишок	32	33,75046	1,054702		
Сумарно	34	303,6937			

	Коефіцієнт	*артна потатисти*	*-Значення*	*ижні 95%*	*ерхні 95%*	*ижні 95,0%*	*Верхні 95,0%*
β_0	-582,531	204,799	-2,8444	0,007694	-999,693	-165,369	-999,693	-165,3690081
β_1	0,298043	0,101556	2,934759	0,006131	0,09118	0,504907	0,09118	0,504906981
β_2	-1,06827	0,138663	-7,70404	8,79E-09	-1,35071	-0,78582	-1,35071	-0,785818994

Рис. 15 Результат використання функції «Регресія» в EXCEL

Отримані прогнозовані значення страхових виплат:

Рік, у якому стався збиток	Період розвитку							
	0	1	2	3	4	5	6	7
2007	6213817,4	2135090,2	733624,8	252076,1	86614,2733	29760,98	10225,98	**42009,83**
2010	15194077	5220740,2	1793865	616378,6	**244153,91**	**71385,4**	**68266,32**	**16776,76**
2011	20469769	7033487,1	2416732	**413824,2**	**29095,23**	98039,63	33686,77	11574,9
2012	27577289	9475656,6	**1505760**	**841240,4**	**796740,55**	**235466,1**	45383,51	15593,94
2013	37152683	**4197941**	**1670173**	**684816,1**	517870,489	177942,2	61141,58	21008,47
2014	**40231751**	**11001298**	**2625832**	**2940708**	**317619,05**	**110076,6**	82371,18	28303,04
2015	**298380876**	**13688468**	**7200553**	**3223203**	**430912,17**	322965,5	110972,1	38130,42
2016	**494434172**	**27374233**	**6895613**	**2808449**	1266301,18	435105,7	149503,9	51370,08
2017	**781514453**	**44445441**	**6441970**	4964984	1705986,65	586183,2	201414,7	69206,82
2018	**776609039**	**31510423**	19466988	6688928	2298339,84	789717,9	271349,9	93236,83
2019	**509529851**	76327252	26226321	9011459	3096370,07	1063924	365568,2	125610,5

Рис. 16 Заповнені пропущені дані

Аналіз отриманих результатів показує узгодженість побудованої моделі з даними (коефіцієнт детермінації R=0,89, p-значення становить $5.41*10^{-6}$).

Отже, завдяки узагальненій лінійній моделі можливо відновити пропущені дані, які часто виникають в актуарій математиці.

Висновок

У магістерській дисертації було здійснене дослідження узагальнених лінійних моделей та їх застосування в актуарній математиці. Розглянуто основи УЛМ, структуру експоненціальної сім'ї, функцію зв'язку та її роль, методи оцінювання параметрів. Це все дало уявлення про побудову та інтерпретацію моделей. На основі проведеної роботи, можемо сформувати висновок, що узагальнені лінійні моделі є потужним інструментом для моделювання частоти страхових подій, розміру страхових виплат (збитків), чистої премії та інших ключових страхових показників.

Особливу увагу в даній роботі було приділено даним з пропусками, що є типовим в актуарній математиці і впливає на точність розрахунків. Тому в практичній частині роботи побудовано узагальнену лінійну модель на основі реальних даних із використанням програмного забезпечення EXCEL та RStudio. Проведений аналіз підтвердив, що узагальнені лінійні моделі дозволяють відновити пропущені дані в трикутниках розвитку.

Отже, результати дослідження демонструють, що узагальнені лінійні моделі є ефективним та універсальним підходом для аналізу й прогнозування страхових ризиків. Вони забезпечують гнучкість, інтерпретованість, можливість роботи з неповними даними та здатність моделювати широкий спектр страхових показників, що підтверджує актуальність та практичну значущість використання УЛМ в актуарній практиці.

Список використаної літератури

1. Goldburd M., Khare A., Tevet D., Guller D. Generalized Linear Models for Insurance Rating. - 2nd ed. (2025 revision). - Arlington : Casualty Actuarial Society, 2025.
2. Anderson D., Feldblum S., Modlin C., Schirmacher D., Schirmacher E., Thandi N. A Practitioner's Guide to Generalized Linear Models [Електронний ресурс]. - 2007. - URL: https://www.casact.org/pubs/dpp/dpp04/04dpp1.pdf
3. Antonio K., Zhang Y. Nonlinear mixed models // Predictive Modeling Applications in Actuarial Science. - Vol. 1. - New York : Cambridge University Press, 2014. - Chap. 16.
4. Brockett P. L., Chuang S.-L., Pitaktong U. Generalized additive models and nonparametric regression // Predictive Modeling Applications in Actuarial Science. - Vol. 1. - New York : Cambridge University Press, 2014. - Chap. 15.
5. Clark D. R., Thayer C. A. A Primer on the Exponential Family of Distributions [Електронний ресурс]. - 2004. - URL: https://www.casact.org/pubs/dpp/dpp04/04dpp117.pdf
6. Dean C. G. Generalized Linear Models // Predictive Modeling Applications in Actuarial Science. - Vol. 1. - New York : Cambridge University Press, 2014. - Chap. 5.
7. McCullagh P., Nelder J. A. Generalized Linear Models. - 2nd ed. - Boca Raton : Chapman & Hall/CRC, 1989. - 532 p.
8. Dobson A. J., Barnett A. G. An Introduction to Generalized Linear Models. - 4th ed. - CRC Press, 2018.
9. Ohlsson E., Johansson B. Non-life Insurance Pricing with Generalized Linear Models. - Berlin : Springer, 2010.
10. De Jong P., Heller G. Z. Generalized Linear Models for Insurance Data. – Cambridge : Cambridge University Press, 2008. - 174 p.

11. Hilbe J. M. Practical Guide to Logistic Regression. - Boca Raton : Chapman & Hall/CRC, 2016.
12. Zeileis A., Kleiber C., Jackman S. Regression Models for Count Data in R // Journal of Statistical Software. - 2008. - Vol. 27(8). – P. 1–25.
13. Tweedie M. C. K. An Index which Distinguishes Between Some Important Exponential Families // Statistics Applications and Research. - 1984. - Vol. 24. - P. 87–99.
14. Nelder J. A., Wedderburn R. W. M. Generalized Linear Models // Journal of the Royal Statistical Society: Series A. - 1972. - Vol. 135(3). - P. 370-384.
15. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. - 2nd ed. - New York : Springer, 2009.
16. Wood S. N. Generalized Additive Models: An Introduction with R. - 2nd ed. - Boca Raton : CRC Press, 2017.
17. Fahrmeir L., Kneib T., Lang S. Regression: Models, Methods and Applications.- Berlin : Springer, 2013.
18. Cameron A. C., Trivedi P. K. Regression Analysis of Count Data. – 2nd ed. - Cambridge : Cambridge University Press, 2013.
19. Yee T. W. Vector Generalized Linear and Additive Models: With an Implementation in R. - New York : Springer, 2015.
20. R Core Team. R: A Language and Environment for Statistical Computing. - Vienna : R Foundation for Statistical Computing, 2025. - URL: https://www.R-project.org
21. Могилян А., Василик О. Узагальнені лінійні моделі в страховій математиці. Матеріали Двадцятої міжнародної наукової конференції імені академіка Михайла Кравчука, 17–20 листопада 2025 року, Київ, КПІ ім. Ігоря Сікорського, с.167-168.

Додаток 1

```r
library(readxl)
library(dplyr)
library(lubridate)
library(ChainLadder)
library(data.table)
# 1. Зчитування файлу
df <- read_excel("D:/статистика збитків.xlsx")
df
# 2. Усунення символів \r\n у назвах колонок
names(df) <- gsub("\r\n", "", names(df))
# 3. Підготовка полів ORIGIN, DEV, VALUE
df2 <- df %>%
  mutate(
    ORIGIN = year(AccidentDate),
    DEV = interval(AccidentDate, PaymentDate) %/% months(12),
    VALUE = AmountPaid
  ) %>%
  filter(DEV >= 0) %>%
  select(ORIGIN, DEV, VALUE)
# 4. Перетворення у data.table
DF <- as.data.table(df2)
# 5. Побудова трикутника (ChainLadder)
triData <- as.triangle(
  DF,
  origin = "ORIGIN",
  dev    = "DEV",
  value  = "VALUE"
)
triData

```