

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма	«Економічна кібернетика»
Галузь знань	05 «Соціальні та поведінкові науки»
Спеціальність	051 «Економіка»

Форма навчання: очна (денна)
очна (денна) / дистанційна

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА
на тему «Моделювання витрат страхових компаній»
(назва теми)

здобувача Медведєва Данила Димитрійовича
(ПІБ)

(підпис)

Науковий керівник: кандидат економічних наук, доцент
Луняк Ірина Вікторівна

(підпис)

**Робота допущена до захисту перед екзаменаційною комісією
з атестації здобувачів вищої освіти (ЕК)**

В.о. завідувача кафедри: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

(підпис)

Київ 2024

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ МОДЕЛЮВАННЯ ВИТРАТ СТРАХОВИХ КОМПАНІЙ	6
1.1. Страхова компанія: види та етапи формування страхових витрат	6
1.2. Теоретичний базис кореляційно-регресійного аналізу	10
РОЗДІЛ 2. МОДЕЛЮВАННЯ РІВНЯ МЕДИЧНИХ ВИТРАТ ЗА ДОПОМОГОЮ ЛІНІЙНОЇ РЕГРЕСІЇ	25
2.1. Формування та аналіз статистичної бази	25
2.2. Розробка економетричної моделі рівня медичних витрат	34
2.3. Покращення якості моделі	40
2.4. Прогнозування рівня медичних витрат за допомогою множинної регресійної моделі	48
ВИСНОВКИ	52
ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ	53

ВСТУП

Актуальність теми. Формування доходів страхових компаній, специфіка їх витрат, склад і порядок нарахування страхових резервів – все це пов'язано з особливостями страхування як сфери бізнесу. Бінарний характер діяльності страховиків, які поєднують проведення самого страхування й виконання ролі активного інвестора, обумовлює специфіку його витрат.

Нині великий інтерес викликають різні математичні моделі економічних процесів. Це твердження справедливе для моделей пов'язаних із проблемами так званої актуарної математики, предметом вивчення якої є страхова справа та робота страхових компаній зокрема. З одного боку існування ризиків для життя, власності, довкілля спричинило бурхливий розвиток страхової індустрії, яка має забезпечувати фінансове покриття непередбачуваних втрат своїх клієнтів. З іншого боку, в центрі всіх вище перелічених подій лежить незаперечна присутність випадковості, що тягне за собою широке застосування методів теорії ймовірності та математичної статистики у сфері моделювання процесів страхування. Зокрема для вивчення деяких загальних питань функціонування страхових компаній і знаходження таких величин як імовірність банкрутства компанії за відомого стартового капіталу, розподілу кількості клієнтів компанії, функції кореляції кількості клієнтів, їхнього математичного очікування тощо, можуть бути застосовані методи регресійного аналізу.

Медичне страхування займає одне з провідних місць у соціально-економічних процесах підтримання та відтворення людського капіталу. Однак у вітчизняній економічній науці розвитку теорії медичного страхування практично не приділяється уваги, незважаючи на практичну значущість таких досліджень. Наявні розв'язання задач актуарної математики, зазвичай, присвячені окремим питанням роботи компанії. Одним із таких

актуальних питань є вивчення кількісної залежності факторів, які впливають на суму медичних витрат страхових компаній.

Аналіз останніх досліджень і публікацій. Питання, які пов'язані з функціонування ринку та компаній в страхуванні в цілому, та зокрема медичних страхових компаній неодноразово ставали предметом ґрунтовних досліджень провідних вітчизняних науковців, таких як: О. Подра, Н. Петришин, О. Сосновська, Г. Баліцька, А. Супрун, О. Бессінна, О. Журавка, О. Пахненко, В. Бухтіарова. Страхові компанії також проводять власні дослідження виявлення факторів впливу на розміри збільшення/зменшення сум витрат. Проте, незважаючи на вагомі наукові дослідження, недостатньо уваги приділяється питанням моделювання витрат компаній, які здійснюють свою діяльність на ринку, зокрема медичних страхових послуг.

Мета і завдання дослідження. Мета виконання кваліфікаційної бакалаврської роботи полягає в розробці та побудові економетричної регресійної моделі залежності розміру виплат страхової компанії бенефіціарам, з метою виявлення значущих факторів.

Досягнення поставленої мети зумовило вирішення таких завдань:

- розглянути особливості функціонування страхових компаній;
- дослідити види страхових витрат, виплат;
- проаналізувати теоретичні аспекти застосування кореляційно-регресійного аналізу та регресійних моделей;
- підготувати інформаційну базу для проведення дослідження;
- провести аналіз статистичних даних страхової компанії;
- побудувати множинну регресійну модель та перевірити її ефективність;
- застосувати критерії для перевірки моделі на мультиколінерність та гомо/гетероскедастичність;
- провести прогнозні розрахунки за моделлю.

Об'єктом дослідження є витрати страхової компанії.

Предметом дослідження є кореляційно-регресійний аналіз та регресійні моделі залежності впливу факторів на розмір витрат страхових компаній.

Процеси обробки та аналізу даних, а також моделювання та прогнозування були реалізовані за допомогою програмного середовища RStudio.

Теоретична та практична значущість отриманих результатів. Теоретична частина даної роботи присвячена огляду базового математичного інструментарію моделювання – простій та множинній регресіям, зокрема етапам її побудови та способам покращення. Практичні результати дослідження дозволяють страховим компаніям покращити процес ціноутворення на свої послуги для різних клієнтських груп та/або за різними демографічними показниками. Крім цього, експортувавши коефіцієнти регресійної моделі будь-яка медична страхова компанія зможе побудувати власну функцію прогнозування витрат, і використовувати її в режимі реального часу.

Інформаційною базою дослідження слугували статистичні дані як українських відомств, зокрема Національної комісії, що здійснює державне регулювання у сфері фінансових послуг та низки міжнародних організацій – Світового банку, зокрема бази даних «World Development Indicators», Міжнародної організації European Insurance, а також загальнодоступних статей у мережі Інтернет.

Структура роботи. Кваліфікаційна бакалаврська робота містить вступ, три розділи, висновки, перелік використаних джерел.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ОСНОВИ МОДЕЛЮВАННЯ ВИТРАТ СТРАХОВИХ КОМПАНІЙ

1.1. Страхова компанія: види та етапи формування страхових витрат

Страхування почало розвиватися при переході до ринкової економіки і являє собою відносно новий вид діяльності в Україні. В умовах колишньої соціалістичної економіки потреба в страхуванні була мінімальною, тому що населення страхувало своє життя і майно не в масовому порядку.

Сучасний страховий ринок України позиціонується як одна зі значущих ланок фінансової системи і є головним інструментом, який сприяє консолідації інвестиційних ресурсів для забезпечення соціального розвитку у країні. Страхування здійснюється за допомогою страхових інститутів, які беруть на себе всі ризики в разі настання страхового випадку. Інакше кажучи, страхування - це вид економічних відносин щодо захисту інтересів фізичних та юридичних осіб у разі настання непередбачуваних страхових випадків за рахунок грошових фондів, які формуються страховиками [13, 14].

Формування звіту про фінансові результати страхових організацій має низку особливостей, обумовлених сукупністю видів їх діяльності. Структура звіту про фінансові результати відрізняється від аналогічної форми звітності інших економічних суб'єктів. Доходи, витрати і прибуток організації розкриваються за видами страхування (страхування життя, страхування інше, ніж страхування життя). Страхові організації займаються не тільки страховою діяльністю, а й інвестиційною, яку також відображено в бухгалтерській звітності.

Метою страхової організації є отримання певного доходу в результаті здійснення страхової та іншої діяльності. Фінансовий результат визначається як різниця між доходами та витратами. Залежно від галузевої специфіки та стратегії страхових організацій доходи групуються таким чином [20, 21]:

- доходи від страхової діяльності;
- доходи від інвестицій;
- операційні доходи;
- інші доходи.

До доходів від страхової діяльності відносять страхові премії (внески) за договорами страхування, співстрахування та перестрахування [20]. За своєю економічною природою страхові премії є виручкою від реалізації послуг. До доходів від страхової діяльності відносять дохід у вигляді відшкодування частки збитків, а також суми повернення страхових резервів, відрахованих у попередні періоди. Сама страхова організація може виступати в ролі посередника й отримувати комісійні винагороди і тантьєми (комісія з прибутку у вигляді відсотка).

Доходи від інвестиційної діяльності безпосередньо не пов'язані зі страхуванням і виступають у вигляді джерела капіталовкладень в активи. До складу входять доходи від розміщення страхових резервів і власного капіталу.

Операційні та інші доходи можуть бути пов'язані як з процесом страхування, так і з іншою діяльністю, не пов'язаною з наданням страхових послуг. До першої групи належать суми відсотків, нарахованих на депо премій за ризиками перестрахування, винагорода за надання послуг, штрафи, неустойки за порушення умов договорів. До другої групи включають доходи у вигляді орендної плати за користування майном, від реалізації основних засобів і нематеріальних активів та інші надходження.

У своїй діяльності витрати страхових компаній формують собівартість страхових послуг і можуть бути згруповані таким чином [4]:

- витрати від страхової діяльності;

- витрати від інвестиційної діяльності;
- управлінські витрати;
- операційні витрати;
- інші витрати, не пов'язані зі страхуванням (надзвичайні витрати, покриття збитків).

До складу витрат, пов'язаних зі здійсненням страхової діяльності, включають суми відрахувань до страхових резервів, страхові виплати за договорами страхування, співстрахування та перестрахування, суми страхових премій, винагороди та тантьєми, витрати на оплату за надання страхових послуг фізичними та юридичними особами [20, 28].

До інвестиційних витрат відносять усі витрати, пов'язані з управлінням і продажем об'єктів інвестицій.

Управлінські витрати не пов'язані з операціями страхування, співстрахування та перестрахування. До них відносять витрати на адміністративно-управлінський апарат, витрати на відрядження та представницькі витрати, витрати на добровільне страхування, загальногосподарські витрати, амортизаційні відрахування, витрати на оплату послуг сторонніх організацій тощо.

До операційних витрат відносять витрати, пов'язані зі сплатою відсотків за кредитами, облігаціями та іншими цінними паперами, витрати з надання послуг страховими агентами тощо.

Для того щоб медична страхова компанія могла заробляти гроші, необхідно, щоб сума щорічних внесків перевищувала витрати на медичне обслуговування бенефіціарів. Отже, страховики вкладають багато часу та грошей у розробку моделей, які точно прогнозують медичні витрати застрахованого населення.

Медичні витрати важко оцінити, оскільки найдорожчі випадки трапляються рідко і здаються випадковими. Проте деякі ситуації є більш поширеними для певних верств населення. Наприклад, рак легенів частіше

трапляється у курців, ніж у некурців, а від хвороб серця частіше страждають огрядні люди. Тому, метою практичної частини моєї роботи є аналіз, моделювання та прогнозування витрат на медичне обслуговування заданими про клієнтів страхової компанії. Ці оцінки можуть бути використані для створення страхових таблиць, згідно з якими сума щорічних внесків встановлюється вищою або нижчою залежно від очікуваних витрат на лікування.

1.2. Теоретичний базис кореляційно-регресійного аналізу

Математичні співвідношення допомагають розібратися в багатьох аспектах повсякденного життя. Наприклад, вага тіла залежить від кількості калорій, що споживаються; дохід - від освіти та досвіду роботи; цифри опитувань допомагають оцінити шанси на переобрання в президенти.

Якщо уявити такі закономірності у вигляді чисел, то все стане зрозумілішим. Наприклад, зайві 250 кілокалорій, спожиті щодня, можуть призвести до збільшення ваги майже на кілограм на місяць; щороку досвід роботи зростає, що може бути приводом для підвищення зарплати на 1000 доларів на рік; а президент, найімовірніше, буде переобраний, якщо економіка стабільна. Очевидно, що ці рівняння не ідеально підходять для кожної ситуації, але можна очікувати, що в більшості випадків вони досить точні.

Поняття регресії. Регресія передбачає визначення взаємозв'язку між однією залежною кількісною змінною (прогнозоване значення) та однією або кількома незалежними кількісними змінними (предикторами). Як впливає з назви, залежна змінна залежить від значення незалежної змінної або змінних. Найпростіші форми регресії припускають, що взаємозв'язок між незалежною і залежною змінними є лінійним.

Термін «регресія», застосовуваний для опису процесу підгонки прямих ліній до даних, з'явився завдяки дослідженням у галузі генетики сера Френсіса Гальтона (Sir Francis Galton) наприкінці XIX століття. Гальтон виявив, що у низькорослих або, навпаки, дуже високих батьків зазвичай народжувалися сини, зріст яких був ближчим до середнього. Він назвав це явище «регресією до середнього».

Пряму лінію можна описати у формі рівняння з кутовим коефіцієнтом виду $y = a + bx$. У цьому рівнянні y позначає залежну змінну, а x - незалежну змінну. Кутовий коефіцієнт b вказує, наскільки сильний нахил прямої в міру збільшення x . Позитивні значення відповідають прямим із нахилом догори, а від'ємні - з нахилом донизу.

Коефіцієнт a називають зміщенням, оскільки він визначає точку, в якій пряма перетинає вертикальну вісь Y . Точка перетину вказує значення y для $x = 0$. На рис. 1.1 показано приклади прямих.

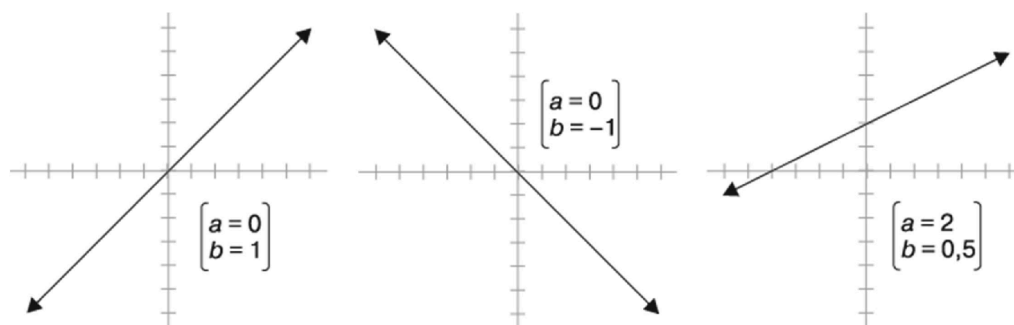


Рисунок 1.1 — Варіанти прямих з різними коефіцієнтами нахилу і точками перетину [7]

Дані моделей регресійних рівнянь використовують аналогічний формат рівнянь із кутовим коефіцієнтом. Завдання машини полягає в ідентифікуванні значень a і b таким чином, щоб пряма найкращим чином пов'язувала надані значення x зі значеннями y . Не завжди вдається побудувати одну функцію, яка б ідеально пов'язувала ці значення, тому в машини також має бути якийсь спосіб кількісної оцінки допустимої похибки.

Регресійний аналіз застосовується для розв'язання різних завдань - це найпоширеніший метод машинного навчання. Він може використовуватися як

для пояснення минулого, так і для екстраполяції майбутнього і може бути застосований для вирішення практично будь-якого завдання. Ось деякі приклади використання цього методу:

- вивчення різних характеристик для жителів у різних країнах і окремих людей у межах наукових досліджень у галузі економіки, соціології, психології, фізики та екології;
- кількісна оцінка причинно-наслідкових зв'язків між подією та її наслідками в таких випадках, як клінічні випробування ліків, технічні випробування на безпеку або маркетингові дослідження;
- побудова моделей, які можна використовувати для прогнозування поведінки з урахуванням відомих критеріїв: прогнозування страхових позовів, збитків від стихійних лих, результатів виборів або рівня злочинності тощо.

Регресійні методи також використовуються для статистичної перевірки гіпотез - така перевірка дає змогу визначити, з якою ймовірністю припущення є істинним чи хибним відповідно до спостережуваних даних. Оцінки інтенсивності та стабільності співвідношень, отримані завдяки регресійній моделі, дають інформацію, яка може бути використана для того, аби визначити, чи є спостереження випадковістю [17].

Регресійний аналіз має не один алгоритм реалізації. Швидше, це загальна назва для великої кількості методів, які можна адаптувати практично до будь-якого завдання машинного навчання.

Розглянемо прості моделі лінійної регресії - тих, у яких використовуються прямі лінії. Випадок, коли існує тільки одна незалежна змінна, називається простою лінійною регресією. За наявності двох і більше незалежних змінних має місце множинна лінійна регресія, або просто множинна регресія. Обидва ці методи потребують наявності єдиної залежної змінної, яка вимірюється безперервно.

Регресія може також використовуватися для інших типів залежних змінних і навіть для деяких завдань класифікації.

Наприклад, логістична регресія застосовується для моделювання результату події в бінарному вигляді (0 або 1), а регресія Пуассона, названа на честь французького математика Симона Пуассона (Sime'on Poisson), моделює результат у вигляді цілих чисел. Поліноміальна логістична регресія моделює не бінарний результат і тому може використовуватися для класифікації. Оскільки у всіх регресійних методах застосовуються одні й ті самі статистичні принципи, після того, як ви опануєте лінійний випадок, вивчити інші варіанти не становитиме труднощів.

Багато специфічних регресійних методів належать до класу узагальнених лінійних моделей (Generalized Linear Models, GLM). Використовуючи GLM, можна поширити лінійні моделі на інші шаблони за допомогою функції зв'язку, яка визначає складніші форми залежності між x і y . Це дає змогу застосовувати регресію практично до будь-якого типу даних [1, 23].

Почнемо з базового випадку простої лінійної регресії. Незважаючи на назву, цей метод не такий уже й простий під час розв'язання складних завдань. З наступного розділу ми дізнаємося, як застосування моделі простої лінійної регресії могло б запобігти жахливій катастрофі.

Проста лінійна регресія. На початку січня 1986 року семеро членів екіпажу американського космічного шатла «Челленджер» загинули через відмову ракетного прискорювача, що призвело до його катастрофічного руйнування. Згодом експерти швидко назвали головну причину трагедії - температура запуску. Гумові кільця ущільнювачів, які забезпечують герметичність з'єднань ракети, ніколи не випробовували при температурі нижче 40 °F (4 °C), а погода в день запуску була нетипово холодною - нижче нуля. Ця аварія стала прикладом того, наскільки важливими є аналіз і візуалізація даних. Невідомо, яка інформація була доступна інженерам-ракетобудівникам та особам, які приймали рішення про запуск. Але не можна заперечувати, що якби в їхньому розпорядженні були більш якісні дані, після ретельного аналізу катастрофи цілком можна було б уникнути [5].

Інженери майже напевно знали, що в умовах низьких температур компоненти стають більш крихкими і менш придатними для якісної герметизації, що, ймовірно, може стати причиною витоку палива. Однак, з огляду на тиск політиків, які вимагали продовжувати запуск, інженерам потрібні були дані, які б підтверджували їхню гіпотезу. Тут їм би дуже знадобилася регресійна модель, що демонструє залежність між температурою і відмовами кілець ущільнювачів, яка могла б спрогнозувати ймовірність відмови з урахуванням очікуваної температури в момент запуску.

Щоб побудувати регресійну модель, вчені могли б використовувати дані про температуру запуску і проблеми з компонентами, записані під час 23 попередніх успішних запусків шаттла. Руйнування компонентів може вказувати на одну з двох проблем. Перша з них - ерозія - виникає, коли надлишкове тепло спалює кільце ущільнювача. Друга - продування - виникає, коли крізь кільце ущільнювача, що недостатньо прилягає, протікають - «продувають» його - гарячі гази.

Оскільки шатл мав загалом шість первинних ущільнюючих кілець, за рейс могло відбуватися до шести аварій. Ракета могла пережити одне або кілька таких руйнувань або ж бути знищеною лише одним, але кожне додаткове руйнування підвищувало ймовірність катастрофічної відмови. На рис. 1.2 показано графік руйнувань первинних кілець ущільнювачів, зафіксованих протягом попередніх 23 запусків, залежно від температури під час старту.

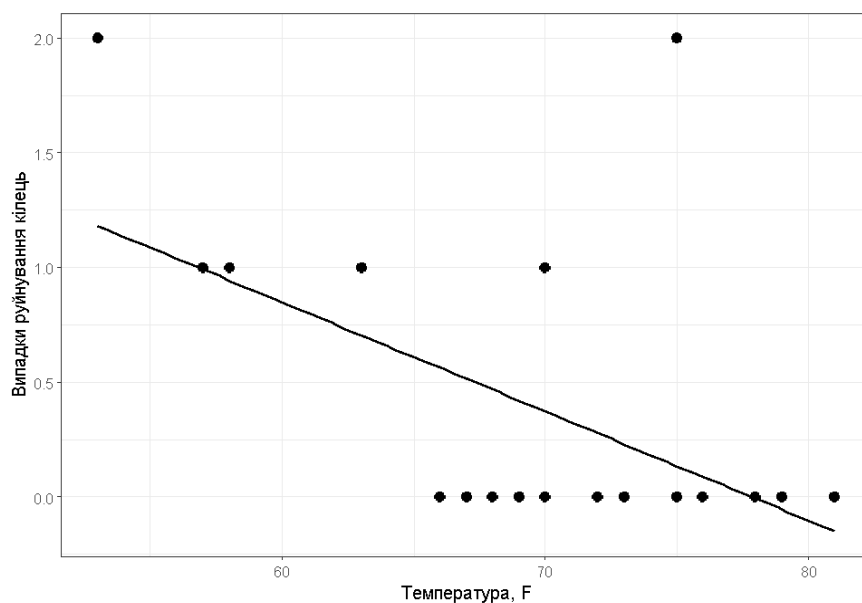


Рисунок 1.2 — Візуалізація пошкоджень ущільнюючих кілець космічного шатла в залежності від температури під час старту

Джерело: [5]

Досліджуючи цей графік, можна простежити тенденцію: під час запусків, що відбувалися за вищих температур, було зафіксовано менше зруйнованих кілець ущільнювачів. Крім того, під час найхолоднішого запуску (53 °F) відбулося два таких руйнування - за жодного іншого запуску подібної кількості руйнувань не спостерігалось.

З урахуванням цієї інформації той факт, що «Челленджер» мав стартувати за температури, нижчої більш ніж на 20 градусів по Фарингейту, видається тривожним. Але щоб відповісти на це запитання, можна проаналізувати до простої лінійної регресії.

Проста лінійна регресійна модель визначає взаємозв'язок між залежною змінною та однією незалежною змінною-предиктором за допомогою лінійної залежності, яка описується таким рівнянням:

$$Y = \alpha + \beta X, \quad (1.1)$$

де Y – результативна змінна (відгук);

X – незалежна змінна (предиктор);

α, β – коефіцієнти (параметри) регресії.

Якщо не зважати на грецькі літери, то це рівняння практично не відрізняється від представленого раніше рівняння з кутовим коефіцієнтом. Зсув (альфа) показує, в якій точці пряма перетинає вісь Y , а нахил (бета) - швидкість зміни y при збільшенні x . Для даних про запуск шатла нахил вказуватиме на очікувану зміну частоти відмов ущільнюючих кілець за умови підвищення температури в момент запуску на один градус.

Грецькі літери часто використовують у статистиці для позначення змінних, які є параметрами статистичної функції. Тому регресійний аналіз означає оцінку значень параметрів. Оцінки параметрів для альфа та бета зазвичай позначаються як a і b , хоча іноді деякі з цих термінів використовуються як взаємозамінні. Припустимо, що оцінка параметрів регресії в рівнянні для даних запуску шатла відома і становить $a = 3,70$ і $b = -0,048$. Тоді повне лінійне рівняння має вигляд $y = 3,70 - 0,048x$. Поки що, не звертаючи уваги на те, як було отримано ці числа, ми можемо зобразити пряму, як показано на рис. 1.2.

Як видно на рис. 1.2, при 60°F ми прогнозуємо менш як одну відмову кілець ущільнювачів; а при 50°F - очікуємо близько 1,3 відмови. Якщо використати цю модель для екстраполяції до 31°F - прогнозованої температури під час запуску «Челленджера», - то отримаємо приблизно $3,70 - 0,048 \times 31 = 2,21$ події відмови кілець ущільнювачів.

Припускаючи, що кожна відмова ущільнювального кільця однаковою мірою може призвести до катастрофічного витoku пального, це означає, що запуск «Челленджера» при 31°F був майже втричі ризикованішим, ніж за звичайного запуску за 60°F , і більше ніж у вісім разів ризикованішим, ніж запуск за 70°F .

З графіка на рис. 1.2 видно, що пряма не проходить точно через усі точки даних. Замість цього вона перетинає область даних більш-менш рівномірно, тож деякі прогнози виявляються нижче або вище лінії.

Оцінка методом найменших квадратів. Щоб обчислити оптимальні оцінки як звичайний метод найменших квадратів (Ordinary Least Squares, OLS).

При OLS-регресії нахил і зміщення обирають таким чином, щоб сума квадратів похибок (Sum of the Squared Errors, SSE) була мінімальною. Помилки, їх також називають нев'язками, є вертикальними відстанями між прогнозованим значенням і фактичним значенням y . Оскільки помилки можуть бути завищеними або заниженими, вони можуть бути представлені позитивними або негативними значеннями. На рис. 1.3 показано помилки для кількох точок.

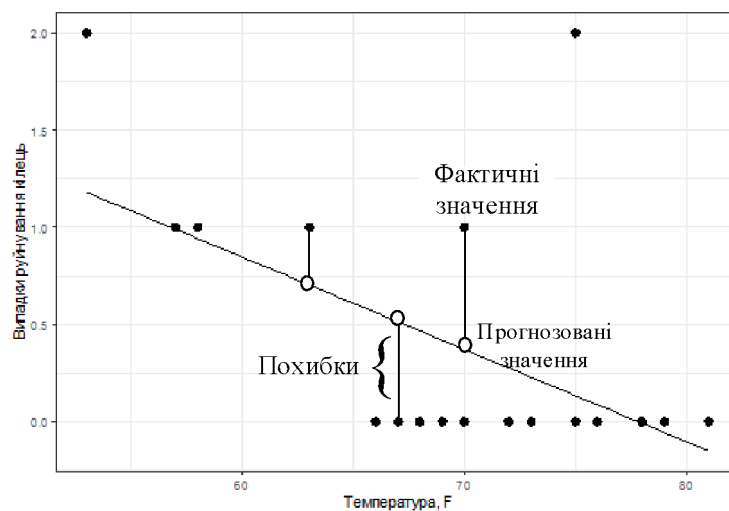


Рисунок 1.3 — Прогнозовані значення регресії відрізняються від фактичних; різниця називається нев'язкою

Джерело: складено автором на основі даних [5]

У математичних термінах мета OLS-регресії може бути описана як задача мінімізації такого рівняння [24]:

$$\sum (y_i - \hat{y}_i)^2 = \sum y_i^2 \rightarrow \min, \quad (1.2)$$

де y_i - фактичним значенням результативної змінної y ;

$\hat{y}_i$ - прогнозоване (розрахункове) значенням результативної змінної y .

Простіше кажучи, це рівняння визначає e (помилку) як різницю між фактичним значенням y і прогнозованим значенням $\hat{y}$. Значення помилок підносять до квадрату, щоб виключити від'ємні значення, і підсумовують за всіма даними.

Розв'язок цього рівняння для a залежить від значення b :

$$a = \bar{y} - b\bar{x}, \quad (1.3)$$

де $\bar{y}$ – середнє значення результативної змінної y ;

$\bar{x}$ – середнє значення незалежної змінної x .

Якщо пропустити доведення, то можна показати, що значення b , яке призводить до мінімальної квадратичної помилки, обчислюється наступним чином:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \quad (1.4)$$

Якщо поділити це рівняння на складові частини, то його можна дещо спростити. Знаменник дуже схожий на дисперсію x , яка позначається як $\text{Var}(x)$. Дисперсія обчислюється як середньоквадратичне відхилення від середнього значення x і може бути виражена наступним чином:

$$\text{Var}(x) = \frac{\sum (x_i - \bar{x})^2}{n}, \quad (1.5)$$

де $\text{Var}(x, y)$ – варіація (або дисперсія) між змінними x та y .

Чисельник є сумою відхилень усіх значень даних від середнього значення x , помноженого на відхилення цієї точки від середнього значення y . Це схоже на коваріаційну функцію для x і y , яка позначається як $\text{Cov}(x, y)$. Коваріаційна формула має такий вигляд:

$$Cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}, \quad (1.6)$$

де $Cov(x, y)$ – коваріація між змінними x та y .

Якщо розділити коваріаційну функцію на функцію дисперсії, то n у чисельнику і знаменнику взаємно скоротяться, і можна буде переписати формулу для b таким чином:

$$b = \frac{Cov(x, y)}{Var(x, y)} \quad (1.7)$$

Оскільки, усі розрахунки в наступному розділі дипломної роботи будуть проводитись в програмному середовищі RStudio, то є сенс паралельно розглядати як в ньому проводити регресійний аналіз. Враховуючи формулу (1.6) легко обчислити значення b , використовуючи вбудовані R-функції. Застосуємо їх до даних про запуски шатлів, щоб обчислити регресію.

```
>
> launch <- read.csv("challenger.csv")
> str(launch)
'data.frame':  23 obs. of  4 variables:
 $ distress_ct      : int  0 1 0 0 0 0 0 1 1 ...
 $ temperature     : int  66 70 69 68 67 72 73 70 57 63 ...
 $ field_check_pressure: int  50 50 50 50 50 50 100 100 200 200 ...
 $ flight_num      : int  1 2 3 4 5 6 7 8 9 10 ...
> b <- cov(launch$temperature, launch$distress_ct) / var(launch$temperature)
> b
[1] -0.04753968
> a <- mean(launch$distress_ct) - b * mean(launch$temperature)
> a
[1] 3.698413
>
```

Рисунок 1.4 — Розрахунок параметрів регресії в RStudio за даними про запуски шатлів

Джерело: [5]

З рис. 1.4 видно, що дані про запуски шатлів будуть збережені у фреймі даних `launch`, незалежна змінна x називається температура (`temperature`), а залежна змінна y - `distress_ct` (кількість зруйнованих кілець ущільнювачів).

Для оцінювання параметрів a і b можна скористатися R-функціями `cov()`, `var()` та `mean()`.

Обчислювати параметри рівняння регресії вручну - це не найкращий спосіб. Тому в R є вбудована функція автоматичного обчислення регресійних моделей. Саме її буде використано для розрахунків в наступному розділі моєї роботи.

Кореляцією між двома змінними називається число, яке показує, наскільки тісний їх взаємозв'язок і близький до лінійної залежності. Іноді кореляцією називають коефіцієнтом кореляції Пірсона, який був введений у XX столітті математиком Карлом Пірсоном (Karl Pearson) [23]. Значення коефіцієнта кореляції знаходиться в межах від -1 до +1. Максимальне і мінімальне значення вказують на ідеальну лінійну залежність та тісний взаємозв'язок, тоді як кореляція, близька до нуля, означає відсутність лінійної залежності.

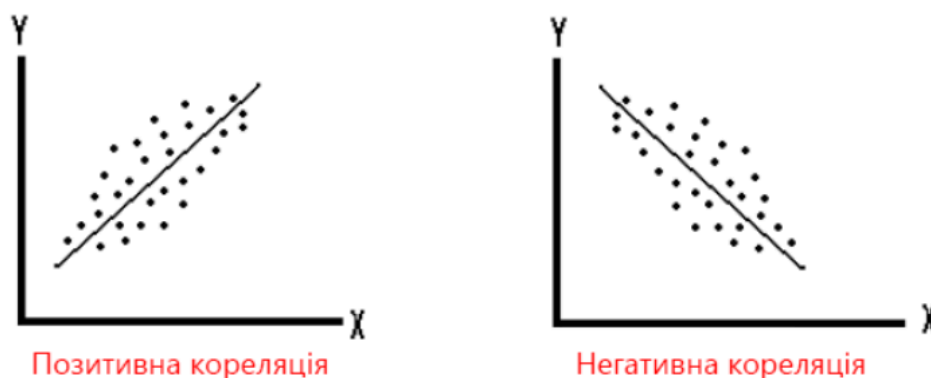


Рисунок 1.5 — Графічне представлення позитивної та негативної кореляції

Джерело: [8]

Кореляція Пірсона описується наступною формулою:

$$\rho = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}, \quad (1.8)$$

де ρ - коефіцієнт кореляції Пірсона;

σ_x , σ_y - середньоквадратичні відхилення відповідно x та y .

За допомогою формули (1.7) можна обчислити кореляцію між температурою та кількістю відмов ущільнюючих кілець при старті шатла. На мові R є вбудовані функції `cov()` - коваріаційна функція, `sd()` - функція стандартного відхилення та `cor()` - функція кореляції. Результат зберігаємо у змінній r - цією літерою зазвичай в науковій літературі [17, 23] позначають кореляцію.

```
> r <- cov(1launch$temperature, 1launch$distress_ct) /
+ (sd(1launch$temperature) * sd(1launch$distress_ct))
> r
[1] -0.5111264
> cor(1launch$temperature, 1launch$distress_ct)
[1] -0.5111264
> |
```

Рисунок 1.6 — Розрахунок в RStudio коефіцієнта кореляції між температурою при старті шатла та кількістю відмов ущільнюючих кілець

Джерело: [5]

Як видно, з рис. 1.6, кореляція між температурою і кількістю пошкоджень ущільнюючих кілець дорівнює -0,51. Негативна кореляція означає, що при підвищенні температури навколишнього середовища кількість зруйнованих ущільнюючих кілець зменшується, а при зменшенні - навпаки. Для інженерів NASA, які займаються вивченням цього питання, це було б явним показником того, що запуск за низької температури може призвести до проблем.

Кореляція також свідчить про відносну інтенсивність взаємозв'язку між температурою і руйнацією ущільнюючих кілець. Оскільки - 0,51 знаходиться посередині між відсутністю кореляції та максимальною негативною кореляцією -1, це означає, що існує помірно сильна негативна лінійна залежність.

Існує низка емпіричних правил, які використовуються для інтерпретації інтенсивності кореляції. За одним із методів статус «слабкий» присвоюється значенням від 0,1 до 0,3, «помірний» - значенням у діапазоні від 0,3 до 0,5 і

«сильний» - значенням, що перевищують 0,5 (це також стосується аналогічних діапазонів від'ємних кореляцій). Однак, ці порогові значення можуть бути занадто суворими або занадто слабкими, залежно від конкретних цілей. Часто кореляцію потрібно інтерпретувати залежно від контексту. Наприклад, для даних які стосуються людини, кореляцію 0,5 можна вважати дуже високою; а для згенерованих механічними процесами даних кореляція 0,5 може бути дуже слабкою [17].

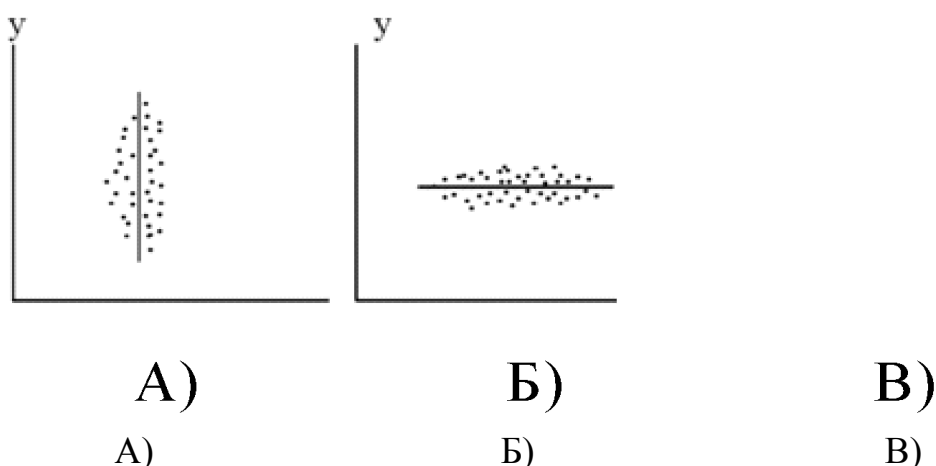


Рисунок 1.7 — Графічне представлення трьох прикладів (умовних) прояву кореляції (А, Б – слабка кореляція, В – відсутня кореляція)

Джерело: [8]

Існує вислів, що кореляція не має на увазі причинно-наслідковий зв'язок. Річ у тім, що кореляція описує тільки зв'язок між парою змінних, проте можуть бути й інші пояснення, які не були виміряні. Наприклад, може існувати сильний зв'язок між очікуваною тривалістю життя і часом, проведеним за переглядом фільмів, але перш ніж лікарі почнуть рекомендувати всім дивитися більше фільмів, необхідно виключити інше пояснення: у молодих людей, які дивляться багато фільмів, імовірність смерті менша.

Вимірювання кореляції між двома змінними дає змогу швидко перевірити наявність лінійної залежності між незалежними змінними та залежною змінною. Це стає ще важливішим, коли будується регресійні моделі з великою кількістю предикторів.

Множинна лінійна регресія.

У більшості випадків при реальному аналізі є кілька незалежних змінних. Тому цілком імовірно, що в багатьох завданнях числового прогнозування ви будете використовувати множинну лінійну регресію. У табл. 1.1 подано переваги та недоліки множинної лінійної регресії.

Таблиця 1.1 — Переваги та недоліки множинної лінійної регресії

Переваги	Недоліки
- Найпопулярніший на даний момент підхід для моделювання кількісних даних. - Є можливість адаптувати для моделювання практично будь-якої задачі. - Має оцінки як розміру так і сили взаємозв'язку між ознаками (незалежними змінними) та результатом.	- Робить сильні припущення даних. - Досліднику повинен одразу вказувати формулу моделі. - Факторні (категоріальні) дані можна використовувати після додаткової обробки. - Не опрацьовує пропущені значення.

Джерело: складено автором на основі [10]

Множинну регресію можна трактувати як розширення простої лінійної регресії. В обох випадках завдання схожі: знайти значення коефіцієнтів нахилу, при яких похибка прогнозу лінійного рівняння була б мінімальною. Ключова відмінність множинної регресії полягає в наявності більшої кількості незалежних змінних (від двох і більше).

Моделі множинної регресії мають наступний вид:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \varepsilon_i, \quad (1.9)$$

де y – результативна змінна (предиктор);

β_0 – вільний член (зсув);

$x_1, x_2, \dots, x_i$ – незалежні змінні (ознаки);

$\beta_1, \beta_2, \dots, \beta_i$ – параметри (коефіцієнти) при незалежних змінних;

ε_i – похибка (залишок) моделі.

У формулі (1.9) залежна змінна y визначається як сума вільного члена (β_0) та кожної з i -ої ознаки, представленою добутком змінної x на його оціночне значення. Значення залишку моделі вказує на те, що жоден прогноз не є ідеальним. У кожній ознаці цієї моделі є свій коефіцієнт. Це дає змогу будь-якій ознаці x робити свій вплив на значення результату y . Іншими словами, при разі збільшенні i -ої ознаки на 1 одиницю – результат y змінюється на величину β_i . Отже, величина зсуву – це очікуване значення результативної змінної y , за умови що всі незалежні змінні дорівнюють нулю.

Щоб оцінити параметри регресії, кожне спостережуване значення залежної змінної y має бути пов'язане зі спостережуваними значеннями незалежних змінних x за допомогою наведеного вище рівняння регресії. На рис. 1.8 у графічному вигляді подано постановку задачі множинної регресії.

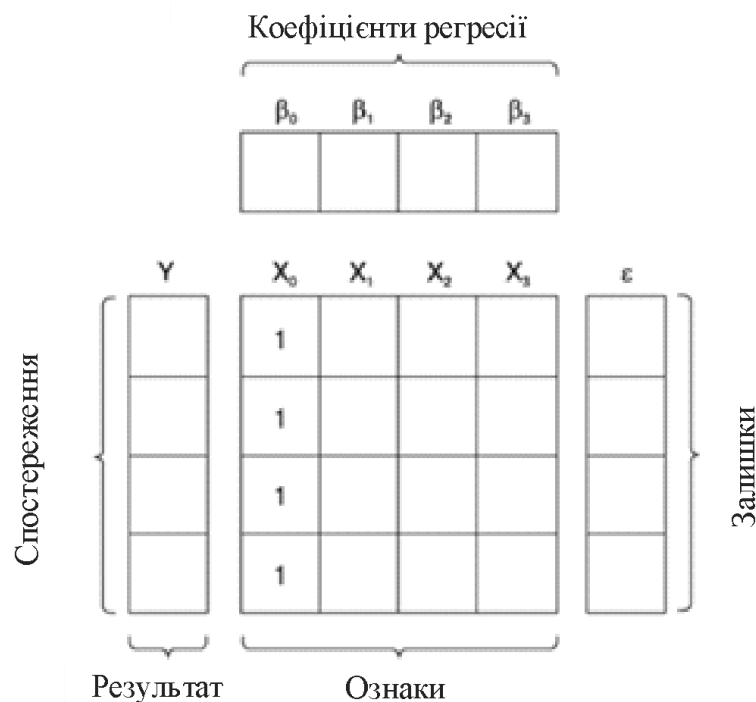


Рисунок 1.8 — Графічне представлення постановки задачі множинної регресії

Джерело: [17]

Рядки та стовпці даних на рис. 1.8 можна подати з використанням матричної форми - відповідні змінні виділено напівжирним, щоб підкреслити, що в кожній з них міститься кілька значень. Спрощена таким чином формула має такий вигляд:

$$\vec{Y} = \vec{\beta}X, \quad (1.9)$$

де $\vec{Y}$ – вектор значень результативної змінної;

$\vec{\beta}$ – вектор оцінених параметрів регресії;

X – матриця значень незалежних змінних.

У матричному записі залежною змінною є вектор Y , що містить по одному рядку для кожного прикладу. Незалежні змінні об'єднані в матрицю X , кількість стовпчиків у якій відповідає числу ознак плюс додатковий стовпець значень «1» для зсуву. У кожному стовпчику матриці є рядок для кожного прикладу. Коефіцієнт регресії та залишкова помилка тепер також є векторами.

В наступному розділі розглянемо практичний приклад побудови регресії.

РОЗДІЛ 2

МОДЕЛЮВАННЯ РІВНЯ МЕДИЧНИХ ВИТРАТ ЗА ДОПОМОГОЮ ЛІНІЙНОЇ РЕГРЕСІЇ

2.1. Формування та аналіз статистичної бази

Процедуру моделювання обсягу витрат на лікування розглянемо покроково.

Крок 1. Збір, дослідження та опис даних.

Для аналізу ми скористаємося набором даних в якому наведені майбутні обсяги медичних витрат для пацієнтів, які проживають у Сполучених Штатах. Цей датасет є синтетичним набором даних, які були створені на основі демографічної статистики, наданої Бюро перепису населення США, і, таким чином, приблизно відповідають реальним умовам.

Щоб завантажити дані для аналізу застосуємо функцією `read.csv()`. Ми можемо використовувати параметр `stringsAsFactors=TRUE`, тому що три номінальні змінні доцільно перетворити на фактори. За допомогою `str()` переглянемо завантажений датасет, з метою перевірки завантаження в правильному форматі.

```
> medical_payments <- read.csv("insurance.csv", stringsAsFactors = TRUE)
> str(medical_payments)
'data.frame': 1338 obs. of 7 variables:
 $ age      : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex      : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi      : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children: int   0 1 3 0 0 0 1 3 2 0 ...
 $ smoker   : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region   : Factor w/ 4 levels "northeast","northwest",...: 4 3 3 2 2 3 3 2 1 2 ...
 $ expenses: num  16885 1726 4449 21984 3867 ...
```

Рисунок 2.1 — Завантаження та перегляд структури вхідного датасету.

Джерело: розроблено автором на даних [21]

У файлі insurance.csv наведено інформацію по 1338 бенефіціарам, які зареєстровані в програмі страхування, та 7 ознак які характеризують його фізіологічний стан, а також загальні медичні витрати за програмою страхування за календарний рік. У файлі вказані наступні характеристики (показники) пацієнтів:

- age - вік основного бенефіціара (за винятком осіб, старших за 64 роки, оскільки їхні витрати на медицину зазвичай покриваються урядом) - ціле число;

- sex - стать страхувальника: чоловіча або жіноча;

- bmi - індекс маси тіла (ІМТ, або BMI), який дає змогу визначити, чи має людина недостатню або надмірну вагу. ІМТ обчислюється як квадрат відношення ваги людини (у кілограмах) до її росту (у метрах). Ідеальний ІМТ має значення в межах від 18,5 до 24,9;

- children – відображає кількість дітей або утриманців, на яких поширюється програма страхування - ціле число;

- smoker - категоріальна змінна, що набуває значень "так" або "ні" і вказує на те, палить/не палить застрахована особа;

- region - місце проживання одержувача страховки в США; може приймати значення від 1 до 4, залежить від чотирьох географічних регіонів: північний схід, південний схід, південний захід або північний захід.

Тепер важливо визначити, як розмір медичних виплат залежить від цих характеристик. Наприклад, можна очікувати, що люди похилого віку та особи які палять більше схильні до ризику великих медичних витрат. На відміну від багатьох інших методів машинного навчання під час регресійного аналізу залежність між ознаками зазвичай визначається користувачем, а не розпізнається автоматично. Деякі з потенційних залежностей будуть розглянуті далі.

Залежна змінна в нашій моделі - expenses, це обсяг витрат на медичне обслуговування, які покриваються медичною страховкою протягом року для кожної людини. Перед тим як будувати регресійну модель, потрібно

перевірити дані на нормальність. Для лінійної регресії залежна змінна не обов'язково повинна мати нормальний розподіл, проте часто модель виходить кращою, коли ця умова виконується. Подивимось на зведену статистику та побудуємо гістограму для її розподілу.

```
>
> summary(medical_payments$expenses)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1122   4740   9382   13270   16640   63770
> hist(medical_payments$expenses,col = "gray",
+       xlab = "Обсяг річних медичних витрат, дол.США", ylab="",
+       main = "")
> abline(v = mean(medical_payments$expenses), col = "blue")
>
```

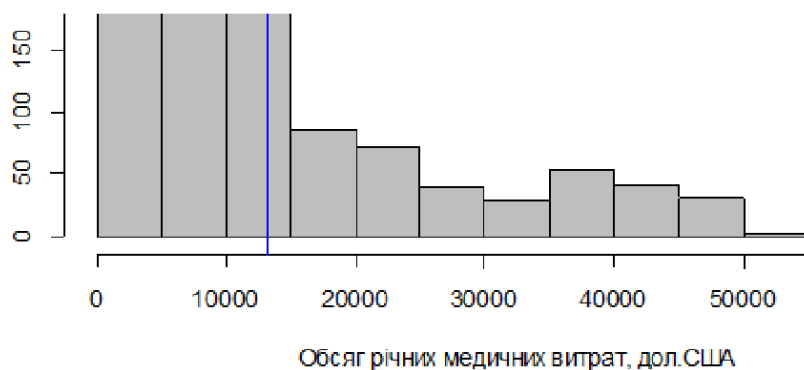


Рисунок 2.2 — Зведена статистика та гістограма розподілу обсягу витрат на медичне обслуговування (expenses)

Джерело: розраховано та побудовано автором на даних [21]

Як видно за результатами зведеної описової статистики (рис.2.2), середнє значення обсягу витрат значно більше медіани. Це означає, що розподіл витрат на страхування має зсув праворуч. Це свідчить про те, що для більшості людей у наших даних, щорічні медичні витрати складають від 0 до 15000 доларів. Проте хвіст цього розподілу тягнеться далеко за межі цих пікових значень. Такий розподіл є неідеальним для лінійної регресії, однак інформація про цей недолік, може допомогти розробити кращу модель.

```

> summary(medical_payments)
   age      sex      bmi      children      smoker      region      expenses
Min.  :18.00  female:662  Min.  :16.00  Min.  :0.000  no :1064  northeast:324  Min.  : 1122
1st Qu.:27.00  male  :676  1st Qu.:26.30  1st Qu.:0.000  yes: 274  northwest:325  1st Qu.: 4740
Median :39.00          Median :30.40  Median :1.000          southeast:364  Median : 9382
Mean   :39.21          Mean   :30.67  Mean   :1.095          southwest:325  Mean   :13270
3rd Qu.:51.00          3rd Qu.:34.70  3rd Qu.:2.000          Max.   :63770
Max.   :64.00          Max.   :53.10  Max.   :5.000

```

Рисунок 2.3 - Зведена статистика по усім змінним датасету `medical_paym`

Джерело: розраховано автором на даних [21]

Перш ніж вирішувати цю проблему, звернемо увагу на ще одну деталь. Для регресійних моделей потрібно, щоб усі ознаки були числовими, проте в нашому датафреймі (рис. 2.1) є три факторні. Зокрема, змінна стать (`sex`) набуває значень `male` і `female`, а змінна курці (`smoker`) має категорії `yes` та `no`. Наприклад, з результатів `summary(medical_payments)` (рис.2.3) відомо, що ознака `region` має чотири рівні, і варто ретельно проаналізувати їх розподіл.

```

> table(medical_payments$region)

northeast northwest southeast southwest
        324         325         364         325

```

Рисунок 2.4 — Кількість значень кожного рівня факторної змінної `region`

Джерело: розраховано автором

Як видно з рис. 2.4, дані розподіляються майже порівну між чотирма географічними регіонами. Далі в цьому розділі ці факторні змінні будуть враховані R-функціями лінійної регресії.

Вивчення залежностей між ознаками - матриця кореляції.

Перш ніж побудувати регресійну модель під дані, іноді корисно визначити, яким чином незалежні змінні пов'язані із залежною змінною та одна з одною. Швидко відповідь на це запитання дає матриця кореляції. Вона показує кореляцію для кожної пари змінних із заданого набору.

Для того щоб створити матрицю кореляції для числових змінних датасету, скористаємося функцією `cor()`. На перетині певного рядка та стовпчика відображається значення коефіцієнтів кореляції відповідної пари змінних. Діагональні значення завжди дорівнюють 1.00, оскільки кореляція

змінної до самої себе завжди ідеальна. Значення, які розташовані симетрично відносно діагоналі, ідентичні, оскільки ці кореляції однакові, тобто $\text{cor}(x,y)=\text{cor}(y,x)$.

```
> cor(medical_payments[c("age", "bmi", "children", "expenses")])
      age      bmi  children  expenses
age    1.0000000 0.10934101 0.04246900 0.29900819
bmi    0.1093410 1.00000000 0.01264471 0.19857626
children 0.0424690 0.01264471 1.00000000 0.06799823
expenses 0.2990082 0.19857626 0.06799823 1.00000000
>
```

Рисунок 2.5 — Матриця коефіцієнтів кореляції числових змінних датасету `medical_payments`

Джерело: розраховано та побудовано автором

Як видно, з рис. 2.5, жодна з кореляцій у цій матриці не має тісного зв'язку, проте є кілька помітних залежностей. Зокрема, між віком (`age`) та індексом маси тіла (`bmi`) є слабка позитивна кореляція, а це означає, що при збільшенні віку людини - її маса тіла збільшується. Є також позитивна кореляція між ознаками `age` та обсягом витрат на медичне обслуговування (`expenses`), `bmi` та `expenses`, а також `children` і `expenses`. Ці залежності означають, що із зростанням віку, збільшенням маси тіла та кількості народжених дітей – очікувані медичні витрати при страхуванні зростають. Потрібно виявити ці залежності чіткіше, для цього побудуємо регресійну модель.

Візуалізація взаємозв'язків між ознаками: матриця розсіювання. Спочатку проведемо візуалізація залежностей між числовими об'єктами за допомогою діаграм розсіювання. Можна побудувати матрицю розсіювання (скорочено `SPLOM`, від `ScatterPLOt Matrix`), це набір діаграм розсіювання, представлених у вигляді сітки. Зазвичай, така матриця розсіювання застосовується для виявлення закономірностей серед трьох і більше змінних. Вона не є справжньою багатовимірною візуалізацією, оскільки одночасно розглядаються лише дві ознаки, проте надає загальне представлення про взаємозв'язки між даними.

Для побудови матриці чотирьох числових ознак - age, bmi, children і expenses - скористаємося функцією *pair()*. Вона входить до складу стандартного пакета R і надає базові функціональні можливості побудови та візуалізації матриць розсіювання:

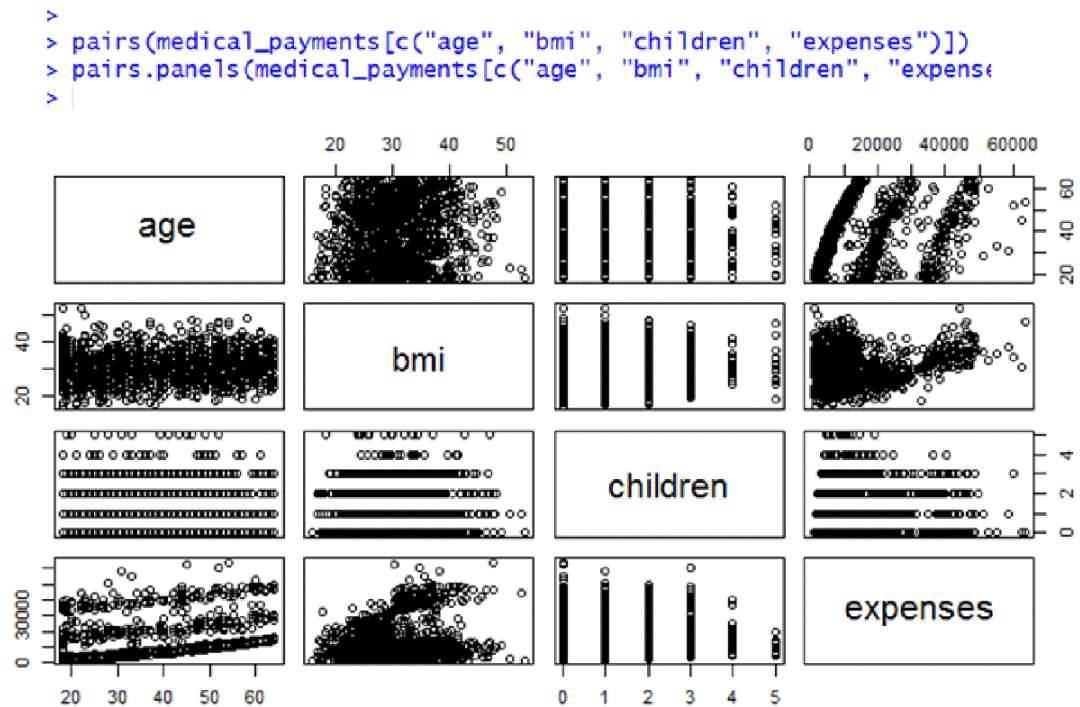


Рисунок 2.6 — Матриця розсіювання взаємозв'язків між числовими ознаками (вверху – скрипт побудови)

Джерело: розраховано та побудовано автором

На рис. 2.6 показано матрицю розсіювання, на перетині кожного рядка і стовпчика містяться діаграма розсіювання для пари відповідних змінних. Діаграми, які розташовані над та під головної діагоналі, є транспозиційними, оскільки для них осі X і Y міняються місцями.

Деякі з них виглядають як випадкові хмари точок, проте на інших, схоже, помітні деякі тенденції. Взаємозв'язок між змінними age та expenses представлено кількома відносно прямими лініями, разом з тим для залежності між expenses та bmi можна виділити дві групи точок. На інших діаграмах важко виявити будь-які тенденції.

Якщо до діаграми (рис. 2.6) додати більше інформації, то можна знайти більше корисної інформації щодо взаємозв'язків. За допомогою функції *pair.panels()* (пакет *psych*) можна побудувати розширену матрицю розсіювання (скрипт для його побудови наведено на рис. 2.6). Це дозволить отримати більш інформативну матрицю розсіювання, як показано на рис. 2.7.

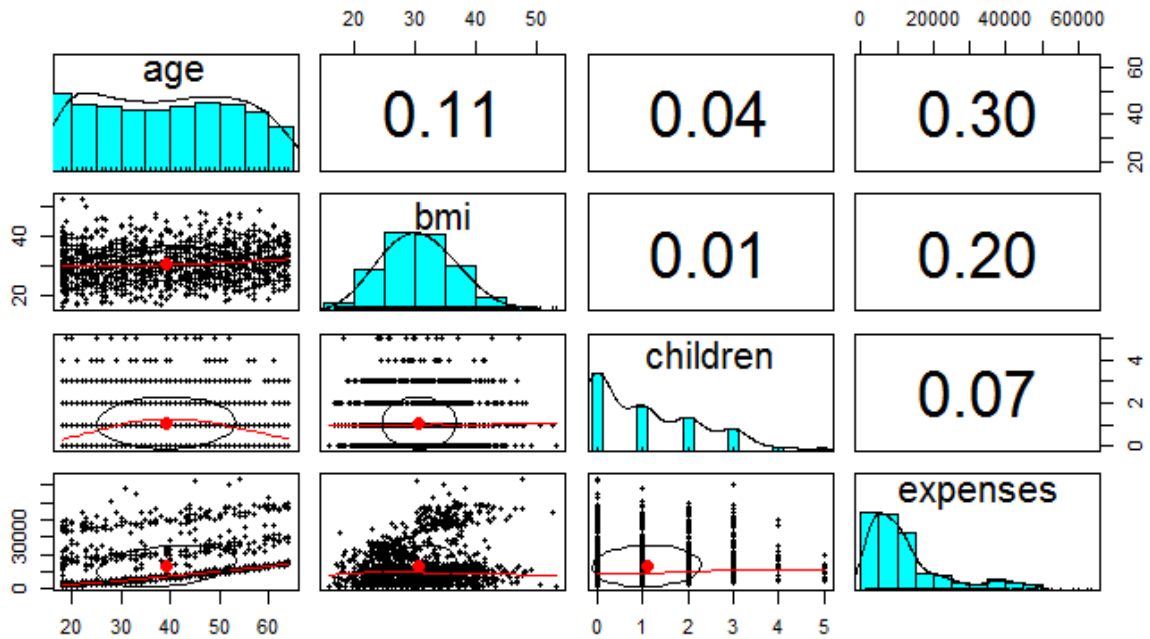


Рисунок 2.7 — Розширена та деталізована матриця розсіювання

Джерело: розраховано та побудовано автором

Діаграма на рис.2.7 відрізняється від діаграми розсіювання (з рис.2.6) тим, що діагоналі замінюються матрицями кореляції. Тепер на діагоналі містяться гістограми, що відображають розподіл значень для кожної ознаки. Нарешті, діаграми розсіювання, розташовані нижче діагоналі, представлені з додатковою візуальною інформацією.

Об'єкти овальної форми, присутні на кожній діаграмі розсіювання, називаються еліпсом кореляції. Він візуалізує силу та тісноту кореляції. Чим більше витягнутий еліпс, тим сильніша кореляція. Наявність майже ідеального кола, як у залежності між *bmi* і *children*, вказує на дуже слабку кореляцію (у цьому випадку 0,01). Еліпс для залежності між *age* і *expenses* витягнутий більше, що свідчить про сильну кореляцію (0,30); точка в центрі еліпса відображає середні значення змінних по осях X і Y.

Криві, зображені на діаграмах розсіювання, називаються LOESS-кривими. Вони вказують на загальну залежність між змінними осей X і Y. Це найкраще пояснити на прикладі. Крива для змінних age і childre має вигляд перевернутої літери U, пік якої відповідає середньому віку. Це означає, що найстаріші та наймолодші люди в даній вибірці мають менше дітей, які враховуються в програмі страхування, ніж люди середнього віку.

Оскільки ця тенденція є нелінійною, то цей висновок не міг бути зроблений на підставі лише коефіцієнтів кореляції. А ось LOESS-крива для age і bmi, навпаки, є прямою з незначним нахилом - це означає, що з віком маса тіла збільшується, - утім, цю залежність ми вже вивели на підставі матриці кореляції.

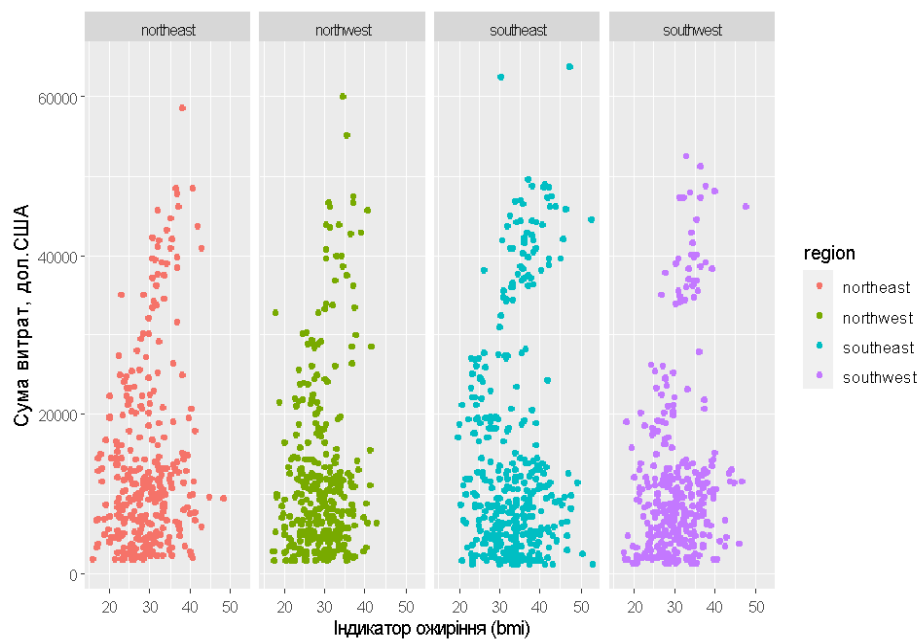
Крім того, можна візуалізувати попарні взаємозв'язки між ознаками в розрізі цілочислових або категоріальних змінних. В R це легко зробити за допомогою функцій пакету ggplot2.

```
>
> ggplot(data = medical_payments, aes(x = bmi, y = expenses)) +
+   geom_point(aes(color = region)) +
+   facet_grid(~region)+
+   labs(x = "Індикатор ожиріння (bmi)", y = "Сума витрат, дол.США")
>
> ggplot(data = medical_payments, aes(x = bmi, y = expenses)) +
+   geom_point(aes(color = children)) +
+   facet_grid(~children)+
+   labs(x = "Індикатор ожиріння (bmi)", y = "Сума витрат, дол.США")
> |
```

Рисунок 2.8 — Скрипт для побудови точкових діаграм зв'язків між індексом зайвої ваги та обсягом медичних витрат

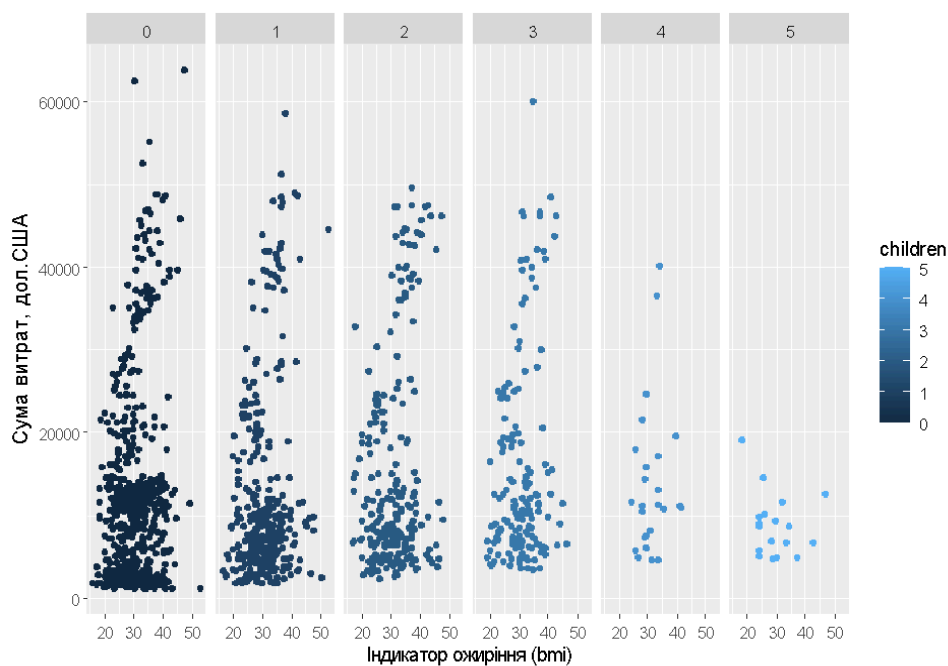
Джерело: складено автором

Наприклад, Нижче наведено скрипт для побудови точкових діаграм.



А)

Джерело: розраховано та побудовано автором



Б)

Рисунок 2.9 — Точкова діаграма зв'язку між індикатором зайвої ваги та обсягом медичних витрат (А – в розрізі регіонів, Б – в розрізі кількості дітей)

Джерело: розраховано та побудовано автором

З діаграм видно, що зв'язок між індикатором зайвої ваги та обсягом медичних витрат практично не залежить від регіону в якому мешкає бенефіціар

(рис. 2.9, А). А для застрахованих осіб які мають чотири або п'ять дітей цей зв'язок послаблюється (рис. 2.9, Б).

2.2. Розробка економетричної моделі рівня медичних витрат

Крок 2. Побудова та оцінка моделі

Щоб побудувати модель лінійної регресії для даних за допомогою функції мови R, потрібно використати функцію `lm()`. Вона входить до складу пакета `stats` і є стандартним пакетом RStudio, який завантажується за замовчуванням. Ця функція створює модель та зберігає її під вказаною назвою, яку можна використовувати для подальших досліджень. Основні складові та параметри цієї функції наведено в табл. 2.1.

Таблиця 2.1 - Синтаксис використання функції `lm()` для моделювання множинної регресії

<i>Побудова моделі</i>	
Конструкція скрипта	<code>m <- lm(dv ~ iv, data = train_data)</code>
Параметри	dv – залежна (результативна) змінна з датасету mydata iv – R-формула, яка визначає незалежні змінні з датасету mydata, які потрібно використовувати в моделі m train_data – дата фрейм навчальної вибірки (з датасету mydata), на якому будується модель.
Результат скрипта	Створюється об'єкт у вигляді списку, в якому зберігається аовна інформація побудованої моделі
<i>Прогнозування</i>	
Конструкція скрипта	<code>p <- predict(m, test_data)</code>
Параметри	m – модель, яка побудована функцією <code>lm</code> test_data - дата фрейм тестової вибірки (з датасету mydata)
Результат скрипта	Створюється вектор, який містить прогнозні значення результативної змінної

Джерело: складано автором самостійно на основі [22]

Наступний скрипт (Рис. 2.10) дозволяє побудувати модель лінійної регресії, яка встановлює залежність сумарних медичних витрат від усіх (шести) незалежних змінних. Модель зберігаємо під назвою `expenses_m`. У синтаксисі R-формули для опису моделі використовується символ "тільда" (~); - ліворуч від неї стоїть залежна змінна `expenses`, а праворуч – можна використати крапку, яка означає що враховуються всі незалежні змінні.

```
> expenses_m <- lm(expenses ~ age + children + bmi + sex + smoker + region,
+                   data = medical_payments)
> expenses_m

Call:
lm(formula = expenses ~ age + children + bmi + sex + smoker +
    region, data = medical_payments)

Coefficients:
(Intercept)          age      children          bmi      sexmale
   -11941.6         256.8         475.7         339.3        -131.4
smokeryes regionnorthwest regionsoutheast regionsouthwest
   23847.5        -352.8        -1035.6        -959.3
```

Рисунок 2.10 — Коефіцієнти побудованої множинної регресії `expenses_m`

Джерело: розраховано автором

Отже, отримана регресія має вид:

$$\begin{aligned} \text{Expenses} = & -11941.6 + 256.8 \text{ age} + 475.7 \text{ children} + 339.3 \text{ bmi} - 131.4 \text{ sexmale} + \\ & -131.4 \text{ sexmale} + 23847.5 \text{ smokeryes} - 352.8 \text{ regionnorthwest} - \\ & -1035.6 \text{ regionsoutheast} - 959.3 \text{ regionsouthwest} \end{aligned}$$

Змінна `Intercept` (-11941.6) - це вільний член (або зсув) регресії, який відображає прогнозоване значення сумарних витрат, якщо незалежні змінні будуть дорівнювати нулю. Однак у більшості випадках вільний член сам по собі має мало сенсу, оскільки рідко буває така ситуація, коли всі ознаки дорівнюють нулю. Зокрема, в нашому випадку не існує людей з нульовим віком та нульовим коефіцієнтом ІМТ. Отже, зсув не має прикладної інтерпретації, тому його часто ігнорують.

Коефіцієнти вказують на передбачуване збільшення/зменшення витрат якщо відповідну ознаку збільшити на одиницю за умови, що всі інші значення

залишаються постійними. Наприклад, очікується, що збільшення віку людини на один рік призводить до зростання обсягу медичних витрат в середньому на 256,8 долара за умови, що всі інші змінні залишаються сталими. Аналогічним чином, поява кожної наступної дитини призводить до додаткових медичних витрат, які в середньому становлять 475,7 долара на рік, а збільшення індексу зайвої ваги на одну одиницю спричиняє збільшення щорічних витрат на медичне обслуговування в середньому на 339,3 долара за інших рівних умов.

Крім того, з рис 2.10 видно, що у формулі моделі вказано тільки шість ознак, а у результатах, крім Intercept, вказано змінні номінальні змінні з фіктивним кодуванням. Наприклад, змінна *region* представлена в моделі трьома категоріями (*regionnorthwest*, *regionsoutheast* та *regionsouthwest*), а змінна *sex*, лише *sexmale*. Це сталося тому, що функція *lm()* автоматично застосовує фіктивне кодування до всіх включених у модель змінних, які мають тип фактор.

Таке фіктивне кодування дає змогу розглядати номінальну ознаку як числову, створюючи двійкову змінну для кожної категорії цієї ознаки. Фіктивна змінна набуває значення 1, якщо спостереження відповідає певній категорії, або 0 – якщо ні. Наприклад, ознака *sex* має дві категорії: *male* і *female*. Тоді вона розбивається на дві бінарні змінні, яким R присвоює імена *sexmale* і *sexfemale*. Для спостережень, де *sex=male*, змінна *sexmale=1*, а *sexfemale=0*; і навпаки, якщо *sex=female*, то *sexmale=0* і *sexfemale=1*. Аналогічне кодування застосовується до змінних із трьома і більше категоріями. Так, R розділяє ознаку *region* з чотирма категоріями на чотири фіктивні змінні: *regionnorthwest*, *regionsoutheast*, *regionsouthwest* і *regionnortheast*.

При додаванні такої фіктивної змінної в регресійну модель одна категорія завжди буде еталонною, а інші інтерпретуються відносно неї. Зокрема, в нашій R-моделі (*expenses_m*) автоматично створено змінні *sexfemale*, *smokerno* і *regionnortheast*, які трактуються наступним чином - жінки, які не палять і проживають у північно-східному регіоні, і стають еталонною групою. Отже, для чоловіків щорічні медичні витрати на 131,4

долара менше, ніж у жінок; а курці витрачають на медицину в середньому на 23847,5 долара на рік більше, ніж некурці. Для всіх трьох регіонів, які залишилися, у цій моделі коефіцієнти є негативним. Це означає, що еталонна група - північно-східний регіон - має в середньому найбільші витрати на медицину.

Зазвичай, в R автоматично за еталонний береться перший з можливих рівнів факторної змінної. Якщо потрібно віддати перевагу іншій категорії (тобто змінити автоматичний еталонний рівень), то можна скористатися функцією `relevel()`, щоб вказати еталонну групу вручну.

У результатах (рис. 2.10), отриманих за допомогою лінійної регресійної моделі, простежується певна логіка: похилий вік, паління та ожиріння, як правило, спричиняють проблеми зі здоров'ям, разом з тим збільшення дітей у сім'ї може призвести до частішого відвідування лікаря та звернення за профілактичною допомогою (щеплення та щорічні огляди).

Далі перевіримо якість побудованої моделі, тобто наскільки добре вона відповідає даним.

Крок 3. Визначення ефективності моделі

Розглянуті вище коефіцієнти (параметри) моделі (рис. 2.10) показують, як незалежні змінні пов'язані із залежною змінною, але не показують її якість. Щоб оцінити ефективність отриманої моделі, використаємо стандартну функцію `summary()`. На рис. 2.11 представлені отримані результати характеристики моделі `expenses_m` (цифри в колах проставлено в ілюстративною метою).

На перший погляд, результат, отриманий за допомогою функції `summary()`, може здатися неймовірним, проте легко зрозуміти, на чому він заснований. У таблиці в пронумерованих мітках видно, що є три ключові способи оцінити ефективність нашої моделі, тобто її відповідність даним.

У розділі `Residuals (1)` (рис. 2.11) подано зведену статистику залишків розрахункових значень результативної змінної, деякі з них досить суттєві. Оскільки залишкове значення дорівнює різниці фактного і прогнозованого

значень, максимальна помилка становить 29981,7. Це означає, що як мінімум для одного спостереження модель занижила прогнозовані витрати майже на 30000 доларів. З іншого боку, 50% помилок (Median) відносяться до значень 1-го і 3-го кварталу (перша і третя квартилі), тому більшість прогнозів перебувають у межах від -2850,9 до 1383,9 долара.

```
> summary(expenses_m)

Call:
lm(formula = expenses ~ age + children + bmi + sex + smoker +
    region, data = medical_payments)

Residuals:
    Min       1Q   Median       3Q      Max 
-11302.7 -2850.9  -979.6   1383.9 29981.7 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11941.6     987.8  -12.089  < 2e-16 ***
age             256.8       11.9   21.586  < 2e-16 ***
children       475.7       137.8    3.452 0.000574 ***
bmi            339.3       28.6   11.864  < 2e-16 ***
sexmale       -131.3       332.9   -0.395 0.693255
smokeryes     23847.5      413.1   57.723  < 2e-16 ***
regionnorthwest -352.8      476.3   -0.741 0.458976
regionsoutheast -1035.6     478.7   -2.163 0.030685 *
regionsouthwest -959.3      477.9   -2.007 0.044921 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494 
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Рисунок 2.11 — Статистичні характеристики множинної регресії expenses_m отримані функцією summary()

Джерело: розраховано автором самостійно

В розділі Coefficients (2) (рис. 2.11) для кожного обчисленого коефіцієнта регресії визначено p -значення ($\text{Pr}(>|t|)$), яке є оцінкою ймовірності того чи є він статистично значущим. Невеликі p -значення припускають, що істинний коефіцієнт навряд чи дорівнює нулю, отже, малоймовірно, що ця ознака не пов'язана із залежною змінною. Деякі з p -значень (рис.2.10) позначені зірочками (***), які вказують на рівень значущості даної оцінки. Рівень значущості є порогом, обраним до побудови моделі, який буде

використовуватися для позначення "реальних", а не випадкових результатів; р-значення, нижчі за рівень значущості, вважаються статистично незначущими. Якщо в моделі мало таких значень, це може бути причиною для занепокоєння, оскільки вказує на те, що обрані незалежні змінні не дуже добре описують результат. У нашій моделі є кілька статистично значущих змінних, які логічно пов'язані з результатом.

В третьому секторі результатів (3) на рис 2.11, коефіцієнт детермінації дає змогу оцінити, наскільки добре модель загалом пояснює варіацію значення залежної змінної. Ця величина схожа на коефіцієнт кореляції, оскільки чим ближче її значення до 1,0, тим краще модель пояснює дані. Оскільки в моделі коефіцієнт детермінації (Adjusted R-squared) дорівнює 0,7494, це означає, що модель пояснює майже 75 % змін залежної змінної. Оскільки чим більше в моделі ознак, тим краще вони пояснюють варіацію залежної змінної. Коефіцієнт детермінації є значенням, скоригованим за рахунок накладення штрафів на моделі з великою кількістю незалежних змінних. Це корисно для порівняння ефективності моделей з різною кількістю пояснювальних змінних.

На підставі описаних вище показників ефективності можна сказати, що модель `expenses_m` працює досить добре. Нерідко регресійні моделі реальних даних мають доволі низький коефіцієнт детермінації; значення 0,75, по суті, доволі гарний результат. Трохи турбує величина деяких помилок, але це пояснюється характером даних про медичні витрати. Однак, можна підвищити ефективність цієї моделі, побудувавши модель іншим способом.

2.3. Покращення якості моделі

Крок 4. Підвищення ефективності моделі

Як уже згадувалося, головна відмінність регресійного моделювання від інших методів машинного навчання полягає в тому, що регресія зазвичай залишає користувачеві вибір ознак і специфікацію моделі. Отже, якщо ми знаємо про те, як ознака пов'язана з результатом, то можемо використати цю інформацію в специфікації моделі і, можливо, підвищити її ефективність. Далі розглянемо три способи підвищити якість моделі на основі зміни її специфікації за рахунок:

- додавання нелінійних залежностей;
- трансформації змінних, зокрема перетворення числової змінної на двійковий індикатор, тобто бінарну змінну;
- введення в модель ефектів взаємодії.

Специфікація моделі: за рахунок додавання нелінійних залежностей.

У лінійній регресії передбачається, що зв'язок між незалежною та залежною змінними є лінійним, проте це не завжди так. Наприклад, вплив віку на медичні витрати не може бути постійним для всіх вікових значень; для літніх груп населення лікування може стати занадто дорогим. Ці висновки підтверджує точкова діаграма на рис. 2.12.

```
> ggplot(data = medical_payments, aes(x = age, y = expenses)) +
+   geom_point(aes(color = bmi30_1)) +
+   labs(x = "Вік застрахованої особи, роки", y = "Сума витрат, дол.США")
> |
```

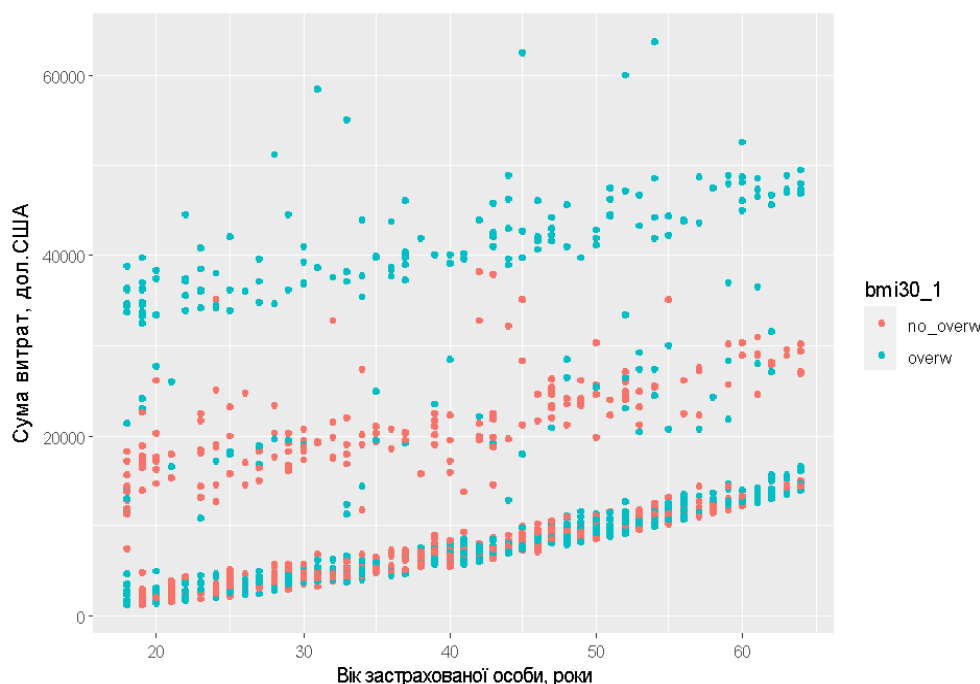


Рисунок 2.12 — Точкова діаграма залежності між віком застрахованої особи та обсягами медичних витрат (градація кольору – факторна змінна індикатору зайвої ваги)

Джерело: розраховано та побудовано автором

Як зазначалось вище в формулі (1.8), просте рівняння регресії має такий вигляд:

$$Y = \beta_0 + \beta_1 X$$

Щоб врахувати нелінійні залежності, потрібно до рівняння регресії додати член вищого порядку, представивши модель у вигляді полінома. Фактично, ми будемо моделювати наступні залежності:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2, \quad (2.1)$$

Де Y - результативна змінна;

$\beta_0, \beta_1, \beta_2$ – параметри регресії;

X, X^2 - незалежна змінна, відповідно в першій та другій степені.

Різниця між цим рівнянням (2.1) та простою лінійною регресією полягає в тому, що вводиться додатковий коефіцієнт, призначений для відображення впливу ознаки в квадраті x^2 . Це дає змогу представити вплив віку як функцію квадрата віку.

Для того щоб додати в модель нелінійну залежність, наприклад, для віку, потрібно просто створити ще одну змінну і ввести її в R-формулу:

```
> medical_payments$age2 <- medical_payments$age^2
> medical_payments$bmi30 <- ifelse(medical_payments$bmi >= 30, 1, 0)
> medical_payments$bmi30_1 <- ifelse(medical_payments$bmi >= 30, "overw", "no_overw")
>
```

Рисунок 2.13 — Скрипт для створення трьох нових змінних age2, bmi30

Джерело: складено автором

Далі, коли ми будемо будувати поліпшену модель, додамо у формулу `lm()` як `age`, так і `age2`, у формі `expenses~age+age2`. Це дасть змогу розділити в моделі лінійний і нелінійний вплив віку на медичні витрати.

Трансформація: перетворення числової змінної на двійковий індикатор.

Припустимо, що вплив ознаки не є накопичувальним, скоріше, він починає даватися взнаки лише після досягнення певного порогового значення. Наприклад, індикатор маси тіла може жодним чином не впливати на обсяг медичних витрат для людей із нормальною вагою, але сильно збільшуватись для людей з зайвою вагою (тобто для ІМТ зі значеннями 30 і більше).

Щоб врахувати в моделі цей вплив, створимо бінарну змінну індикатора ваги тіла, яка дорівнює 1, якщо ІМТ перевищує значення 30, і 0 - якщо не перевищує. Потім оціночний бета-коефіцієнт для цієї бінарної ознаки вказуватиме на середній чистий вплив на медичні витрати для тих осіб, чий ІМТ дорівнює або перевищує 30, відносно тих, у кого ІМТ менше 30.

Щоб створити ознаку, можна скористатися функцією `ifelse()`, яка перевіряє зазначену умову для кожного елемента вектора і повертає значення залежно від того, чи є ця умова істинною або хибною. Для ІМТ, більшого або такого, що дорівнює 30, ми повертатимемо 1, в іншому разі – 0. Скрипт створення змінної `bmi30` наведено на рис. 2.12.

Потім можна включити змінну `bmi30` у нашу покращену модель або замінити нею вихідну змінну `bmi`, або на додаток до неї, залежно від того, чи вважаємо ми, що вплив ожиріння доповнює окремий лінійний вплив ІМТ. Зокрема в нашому прикладі, ми включимо в остаточну модель обидві ознаки.

Щоб перевірити, чи варто включати змінну в модель, то звичайною практикою є перевірити її *p*-значення. Якщо змінна не є статистично значущою, тоді її можна відкинути.

Специфікація моделі: додавання ефектів взаємодії

Досі ми розглядали тільки індивідуальний внесок кожної ознаки в результат. Але що, якщо окремі ознаки мають спільний вплив на залежну змінну? Наприклад, куріння й ожиріння можуть чинити шкідливий вплив окремо, але розумно припустити, що їхній сукупний ефект може виявитися гіршим, ніж від кожного з них загалом.

Коли дві ознаки мають сукупний ефект, це називається взаємодією. Якщо ми підозрюємо, що дві змінні взаємодіють, можна перевірити цю гіпотезу, додавши їхню взаємодію в модель.

Візуалізуємо взаємозв'язок між індикатором маси тіла та сумами медичних витрат через призму того, чи є бенефіціар курцем чи ні (рис. 2.15).

```
>
> ggplot(data = medical_payments, aes(x = bmi, y = expenses)) +
+   geom_point(aes(color = smoker)) +
+   labs(x = "Індикатор ожиріння (bmi)", y = "Сума витрат, дол.США")
>
```

Рисунок 2.14 — Скрипт побудови для створення трьох нових змінних `age2`, `bmi30`

Джерело: складено автором

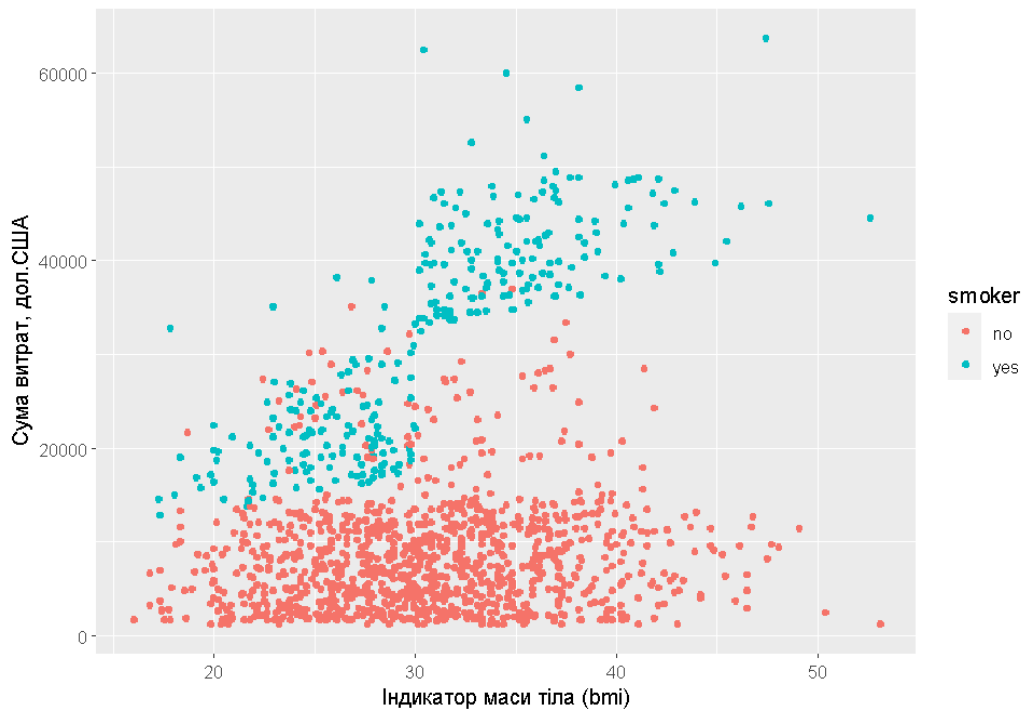


Рисунок 2.15 — Точкова діаграма залежності між індикатором маси тіла та обсягами медичних витрат (градація кольору за змінною палить/непалить)

Джерело: розраховано та побудовано автором

Ефекти взаємодії задаються з використанням синтаксису R-формули. Щоб задати взаємодію між індикатором зайвої ваги (bmi30) та індикатором куріння (smoker), можна подати формулу у вигляді: `expenses~bmi30*smoker`.

Оператор `*` - це скорочення, яке вказує на те, що модель треба представити як: `expenses~bmi30+smokeryes+bmi30:smokeryes`.

Оператор двокрапки (`:`) у розгорнутому вигляді вказує на те, що `bmi30:smokeryes` являє собою взаємодію між двома змінними. Розширена форма також автоматично містить і окремі змінні `bmi30` і `smokeryes`, і їхню взаємодію.

Взаємодії ніколи не повинні включатися в модель без додавання кожної з взаємодіючих змінних. Якщо ви завжди створюєте взаємодії за допомогою оператора `*`, це не проблема, оскільки R автоматично додає всі необхідні компоненти.

Збираємо описані способи зміни специфікації початкової моделі (рис. 2.11) все разом і будуємо покращену регресійну модель. Знаючи, яким чином

медичні витрати можуть залежати від характеристик пацієнта, ми розробили, більш точну формулу регресії. У підсумку ми внесли наступні поліпшення:

- додали нелінійну змінну для віку (age2);
- створили індикатор ожиріння (bmi30);
- ввели ефект взаємодії між зайвою вагою та курінням (bmi30:smokeryes).

Ми будемо навчати модель, як і раніше, використовуючи функцію `lm()`, однак цього разу додамо створені змінні та ефект взаємодії (рис. 2.16). Результати мають наступний вид:

```
> expenses_m2 <- lm(expenses ~ age + age2 + children + bmi + sex +
+                     bmi30*smoker + region,
+                     data = medical_payments)
> summary(expenses_m2)
```

Call:

```
lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 *
    smoker + region, data = medical_payments)
```

Residuals:

Min	1Q	Median	3Q	Max
-17297.1	-1656.0	-1262.7	-727.8	24161.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	139.0053	1363.1359	0.102	0.918792
age	-32.6181	59.8250	-0.545	0.585690
age2	3.7307	0.7463	4.999	6.54e-07 ***
children	678.6017	105.8855	6.409	2.03e-10 ***
bmi	119.7715	34.2796	3.494	0.000492 ***
sexmale	-496.7690	244.3713	-2.033	0.042267 *
bmi30	-997.9355	422.9607	-2.359	0.018449 *
smokeryes	13404.5952	439.9591	30.468	< 2e-16 ***
regionnorthwest	-279.1661	349.2826	-0.799	0.424285
regionsoutheast	-828.0345	351.6484	-2.355	0.018682 *
regionsouthwest	-1222.1619	350.5314	-3.487	0.000505 ***
bmi30:smokeryes	19810.1534	604.6769	32.762	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared: 0.8664, Adjusted R-squared: 0.8653
F-statistic: 781.7 on 11 and 1326 DF, p-value: < 2.2e-16

Рисунок 2.16 — Статистичні характеристики розширеної множинної регресії `expenses_m2` отримані функцією `summary()`

Джерело: розраховано автором самостійно

Порівняння двох моделей (рис. 2.11 та 2.16) допомагає визначити, чи покращилась якість регресійної моделі після внесених змін. В моделі `expenses_m2` порівняно з першою моделлю коефіцієнт детермінації підвищився з 0,75 до 0,87. Аналогічно скоригований коефіцієнт детермінації,

який враховує той факт, що модель стала складнішою, підвищився з 0,75 до 0,87. Тепер наша модель пояснює 87 % варіації обсягів витрат на лікування. Крім того, наше припущення про функціональну форму моделі (квадратичну регресію), підтверджується. Коефіцієнт при змінній найвищого порядку (age2) є статистично значущим, як і індикатор ожиріння bmi30, але змінна віку age – не статистично значуща.

Крім того, якщо перевірити ці дві моделі на мультиколінерність (наявність тісного зв'язку між незалежними змінними) та гетероскедастичність (тобто нерівномірному розмаху залишків для різних рівнів змінної) – то висновки щодо кращої моделі стають суперечливими.

```
> vif(expenses_m)
              GVIF Df GVIF^(1/(2*Df))
age          1.016843 1          1.008386
children    1.004008 1          1.002002
bmi         1.106682 1          1.051990
sex         1.008900 1          1.004440
smoker      1.012067 1          1.006015
region      1.098928 3          1.015847
> vif(expenses_m2)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
              GVIF Df GVIF^(1/(2*Df))
age          47.810746 1          6.914531
age2         47.769222 1          6.911528
children     1.102586 1          1.050041
bmi          2.957411 1          1.719712
sex          1.010946 1          1.005458
bmi30        3.019059 1          1.737544
smoker       2.134714 1          1.461066
region       1.107873 3          1.017220
bmi30:smoker 2.392643 1          1.546817
> |
```

Рисунок 2.17 — Перевірка моделей expenses_m та expenses_m2 на мультиколінеарність

Джерело: розраховано автором

В R функція vif() дозволяє виконати перевірку вже готової моделі на ознаки мультиколінерності. Якщо значення GVIF дорівнює/більше 5, то відповідна змінна викликає мультиколінерність. Як видно з рис. 2.17 в моделі expenses_m2 змінні age та age2 викликають мультиколінерність.

В R функція bptest () дозволяє перевірити залишки моделі на гомоскедастичність/гетероскедастичності за тестом Уайта. Згідно цього тесту нульова гіпотеза (H0) стверджує що присутня гомоскедастичність,

альтернативна (H1) присутня гетероскедастичність (або інакше гомоскедастичність відсутня) [10].

```
> bptest(expenses_m, data = medical_payments)

studentized Breusch-Pagan test

data: expenses_m
BP = 121.59, df = 8, p-value < 2.2e-16

> bptest(expenses_m2, data = medical_payments)

studentized Breusch-Pagan test

data: expenses_m2
BP = 4.4852, df = 11, p-value = 0.9535

> |
```

Рисунок 2.18 — Перевірка моделей expenses_m та expenses_m2 на гетероскедастичність за тестом Уайта

Джерело: розраховано автором

Як видно з рис. 2.18 в моделі expenses_m присутня гомоскедастичність, а в моделі expenses_m2, як ми вважали покращеній, присутня гетероскедастичність.

Взаємозв'язок між зайвою вагою і курінням (змінна bmi30:smoker) має велике значення для визначення суми витрат. Окрім збільшення витрат у разі куріння на суму понад 13 402 долари, зайва вага збільшує медичні витрати страхової компанії на 114,8 долари. Але, якщо беніфіціар є курцем і страждає від ожиріння, то витрати на таку особу будуть складати 19784 доларів на рік. Це може свідчити про те, що куріння посилює захворювання, пов'язані з ожирінням.

Отже, регресійне моделювання робить деякі сильні припущення про дані. Для числового прогнозування ці припущення не особливо важливі, оскільки цінність моделі не ґрунтується на тому, чи дійсно вона відображає базовий процес, - нас просто хвилює точність прогнозів. Щоб зробити точні висновки на підставі коефіцієнтів регресійної моделі, необхідно виконати діагностичні тести, щоб переконатися, що припущення регресійної моделі не були порушено [2].

2.4. Прогнозування рівня медичних витрат за допомогою множинної регресійної моделі

Вивчивши оціночні коефіцієнти регресії та статистику відповідності, можна використовувати модель для прогнозування витрат медичного страхування. Щоб проілюструвати цей процес, спочатку застосуємо модель до вхідних тренувальних даних, використовуючи функцію `predict()` (рис. 2.19). У результаті прогнози будуть збережені у фреймі даних `medical_payments` як нові вектори з ім'ям `pred` або `pred2` отриманими відповідно за різними моделями. Потім можна обчислити кореляцію між прогнозованими та фактичними витратами на медицину.

```
> medical_payments$pred <- predict(expenses_m, medical_payments)
> medical_payments$pred2 <- predict(expenses_m2, medical_payments)
> cor(medical_payments$pred2, medical_payments$expenses)
[1] 0.9307999
> cor(medical_payments$pred, medical_payments$expenses)
[1] 0.8665613
> |
```

Рисунок 2.19 — Розрахунок кореляції між прогнозованими та фактичними витратами на медицину

Джерело: розраховано автором

Кореляція в 0,93 передбачає дуже сильну лінійну залежність між прогнозованими і фактичними значеннями. Це хороший знак - отже, модель дуже точна. Корисно також подати результат у вигляді діаграми розсіювання, яку наведено нижче.

```
> plot(medical_payments$pred2, medical_payments$expenses)
>   abline(a = 0, b = 1, col = "red", lwd = 3, lty = 2)
>
> plot(medical_payments$pred, medical_payments$expenses)
>   abline(a = 0, b = 1, col = "green", lwd = 3, lty = 2)
> |
```

Рисунок 2.20 — Скрипт для побудови точкових діаграм розсіювання прогнозованих та фактичних витрат

Джерело: розраховано автором

На рис. 2.20 наведено скрипт для побудови цих діаграм розсіювання.

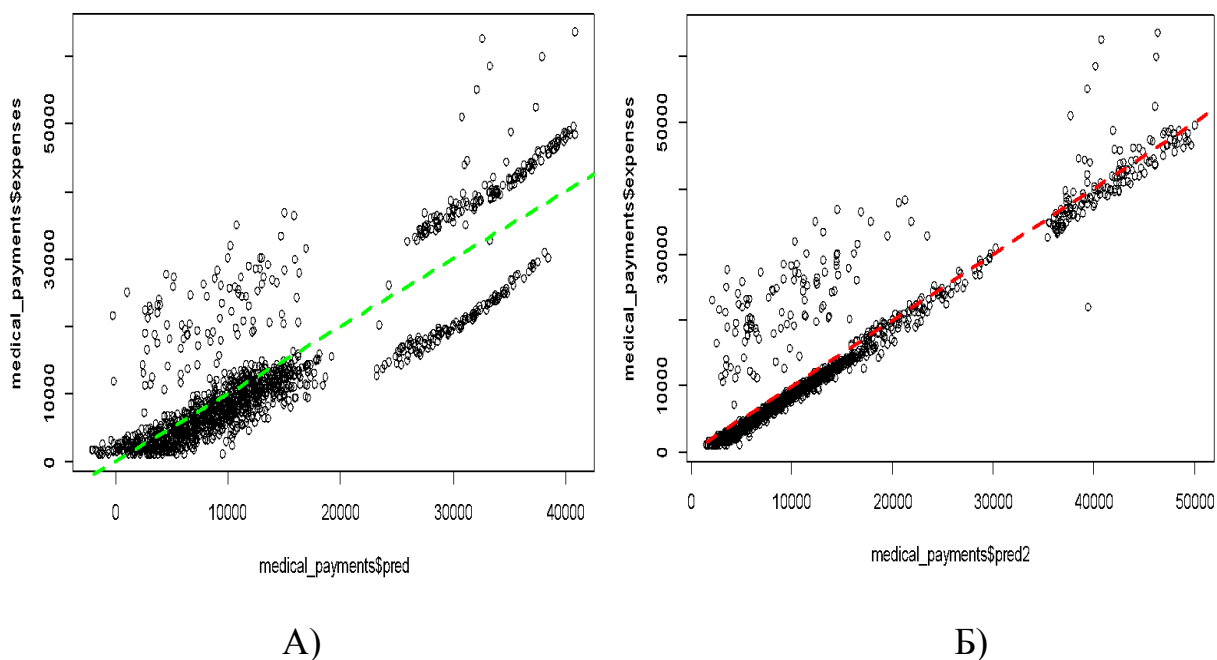


Рисунок 2.21 — Точкові діаграми розсіювання прогнозованих та фактичних витрат (А – за моделлю `expenses_m`, Б - за моделлю `expenses_m2`)

Джерело: розраховано та побудовано автором

На цій діаграмі розсіювання (рис. 2.21) точки, що потрапляють на діагональну пунктирну лінію $y = x$ або розташовані поруч із нею, вказують на прогнози, які дуже близькі до фактичних значень. Точки, розташовані вище діагональної прямої, відповідають випадкам, коли фактичні витрати виявилися більшими, ніж очіувалося, а точки, розташовані нижче цієї лінії, - випадкам, коли витрати були меншими, ніж очіувалося. Тут видно, що невелика кількість пацієнтів зі значно більшими, ніж очіувалося, медичними витратами врівноважується великою кількістю пацієнтів із витратами, трохи меншими за очікувані.

Тепер припустимо, що ми хочемо спрогнозувати витрати для нових учасників програми страхування. Для цього потрібно надати функції `predict()` фрейм даних з інформацією про нових бенефіціарів. Якщо пацієнтів багато, можна створити файл у форматі електронної таблиці CSV для завантаження в R; якщо ж пацієнтів небагато, то можна просто створити фрейм даних у самій функції `predict()`.

Наприклад, для того, щоб оцінити витрати на страхування 40-річного чоловіка, який не палить, але має надлишкову вагу (індикатор складає 35) і проживає з трьома дітьми на північному сході США, можна провести наступні розрахунки:

```
> predict(expenses_m2,
+         data.frame(age = 40, age2 = 40^2, children = 3,
+                   bmi = 35, sex = "male", bmi30 = 1,
+                   smoker = "no", region = "northeast"))
1
9536.573
> predict(expenses_m2,
+         data.frame(age = 40, age2 = 40^2, children = 3,
+                   bmi = 35, sex = "female", bmi30 = 1,
+                   smoker = "no", region = "northeast"))
1
10033.34
>
```

Рисунок 2.22 — Розрахунок прогнозованої суми витрат для чоловіків та жінок з однаковими вхідними даними

Джерело: розраховано автором

Як видно з рис. 2.22, використовуючи це значення (9536,57), страхова компанія, вирішить встановити ціну для такої демографічної групи на рівні близько 9600 доларів на рік, або 800 доларів щомісяця, щоб гарантувати собі беззбитковість. Нижня частина скрипта на рис 3.4 розраховує цей же показник для жінки з такими самими характеристиками. Різниця між цими двома значеннями становить $9536.57 - 10033.34 = -496,769$ і дорівнює коефіцієнту регресії при змінній *sexmale*. За оцінками, у середньому чоловіки за інших рівних умов витрачають на медицину приблизно на 496 доларів на рік менше, ніж жінки. Це свідчить про те, що прогнозовані витрати є сумою кожного з коефіцієнтів регресії, помноженого на відповідне значення у фреймі даних прогнозування. Наприклад, використовуючи коефіцієнт регресії моделі 678,6017 при змінній кількість дітей, можна припустити, що ті, хто не має дітей, витрачають на медичне обслуговування на $3 \times 678,6017 = 2035.805$ доларів менше.

Виконавши аналогічні операції ще для кількох клієнтських сегментів, страхова компанія може розробити вигідну структуру ціноутворення для різних демографічних показників.

Експорт коефіцієнтів регресійної моделі дає змогу побудувати власну функцію прогнозування. Одним із можливих варіантів використання такої функції було б впровадження регресійної моделі в базу даних клієнтів для прогнозування в режимі реального часу.

ВИСНОВКИ

У даній роботі було розглянуто теоретичні засади функціонування страхових компаній, зокрема процес формування медичних витрат. Запропонований алгоритм виявлення впливу факторів на суми витрат страхових компаній дозволяє розробляти вигідну структуру ціноутворення для різних демографічних груп та клієнтських сегментів. Експорт коефіцієнтів регресійної моделі дасть змогу будувати власну функцію прогнозування.

На основі отриманої моделі страхова компанія отримала наступні висновки. Взаємозв'язок між зайвою вагою і курінням (змінна `bmi30:smoker`) має велике значення для визначення суми витрат. Окрім збільшення витрат у разі куріння на суму понад 13 402 долари, зайва вага збільшує медичні витрати страхової компанії на 114,8 долари. Але, якщо беніфіціар є курцем і страждає від ожиріння, то витрати на таку особу будуть складати 19784 доларів на рік. Це може свідчити про те, що куріння посилює захворювання, пов'язані з ожирінням.

Отже, регресійне моделювання робить деякі сильні припущення про дані. Для числового прогнозування ці припущення не особливо важливі, оскільки цінність моделі не ґрунтується на тому, чи дійсно вона відображає базовий процес, - нас просто хвилює точність прогнозів. Щоб зробити точні висновки на підставі коефіцієнтів регресійної моделі, можна виконати додаткові діагностичні тести, щоб переконатися, що припущення регресійної моделі не були порушено. В даній роботі цього не було зроблено.

Крім того, одним із можливих варіантів використання такої функції було б впровадження регресійної моделі в базу даних клієнтів для прогнозування в режимі реального часу. Це створить зручний інтерфейс для клієнтів і дозволить страховій компанії налагодити зворотній зв'язок з ними.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Кобець В. Економетрика в RStudio. Київ. 2021. 132 с.
2. Allison P.D. Multiple Regression: A Primer. London, Pine Forge Press, 1998. 220 p. URL: <https://onlinelibrary.wiley.com/doi/epdf/10.1002/sim.895> (дата звернення: 05.05.2024)
3. Самошкіна І. Д. Становлення та розвиток медичного страхування в Україні. *Причорноморські економічні студії*. Одеса : ПУ Причорноморський НДІ економіки та інновацій, 2020. Вип. 54. С.187–192
4. Zhuravka F., Zhuravka O., Bondarenko E. Voluntary health insurance as a source of funding for the health care system: the world's experience and Ukraine. *Insurance Markets and Companies*. 2020. Vol. 11. P. 61– 80. URL: [https://doi.org/10.21511/ins.11\(1\).2020.07](https://doi.org/10.21511/ins.11(1).2020.07) (дата звернення: 05.05.2024)
5. Dalal S.R., Fowlkes E.B., Hoadley B. Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*. 1989. Vol. 84. P. 945–957. URL: <https://utstat.toronto.edu/reid/sta2101f/2020/shuttle.pdf> (дата звернення: 05.05.2024)
6. Tufte E.R., Cheshire C.T. Visual Explanations: Images And Quantities, Evidence And Narrative. Graphics Press, 1997. URL: <https://faculty.salisbury.edu/~jtanderson/teaching/cosc311/fa21/files/tufte.pdf> (дата звернення: 05.05.2024)
7. Robison W., Boisjoly R., Hoeker D., Young S. Tufte and the Morton Thiokol engineers on the Challenger. *Science and Engineering Ethics*, 2002. Vol. 8.P. 59–81. URL: https://ils.unc.edu/courses/2016_fall/inls161_004/materials/task04.spreadsheets/FINRobison.pdf (дата звернення: 05.05.2024)

8. Відкритий посібник з відкритих даних: для громадських організацій, журналістів, і не тільки.... Київ, 2016. URL: <https://socialdata.org.ua/manual/manual4/> (дата звернення: 05.05.2024)
9. Наконечний С.І., Терещенко Т.О., Романюк Т.П. Економетрія: Підручник. – К.: КНЕУ, 2006. -528 с.
10. Гур'янова Л. С., Клебанова Т. С., Прокопович С. В. Прикладна економетрика : навч. посіб. : у двох частинах. Частина 2 : [Електронне видання]. Харків : ХНЕУ ім. С. Кузнеця, 2016. 252 с. URL: http://www.repository.hneu.edu.ua/bitstream/123456789/19842/1/2016-%D0%93%D1%83%D1%80_%D1%8F%D0%BD%D0%BE%D0%B2%D0%B0_%D0%9B_%D0%A1__%D0%9A%D0%BB%D0%B5%D0%B1%D0%B0%D0%BD%D0%BE%D0%B2%D0%B0_%D0%A2_%D0%A1_%D1%82%D0%B0_%D1%96%D0%BD.PDF (дата звернення: 05.05.2024)
11. Третяк Д. Д. Сутність та соціально-економічне значення медичного страхування. *Інвестиції: практика та досвід*. 2019. №12. С. 43–47
12. Sokyrrko E. S., Kobzar I. A., Khanaliyeva I. S. The foreign experience in development of medical insurance and possibilities its application in Ukraine. *Economy and Society*. 2018. №19. URL: <https://doi.org/10.32782/2524-0072/2018-19-171> (дата звернення: 05.05.2024)
13. Сова О. Ю. Сучасний стан та перспективи розвитку медичного страхування в Україні. *Вчені записки Університету «КРОК»*. 2018. Випуск 49. С. 80–92
14. Гатаулліна Е. І. Тенденції розвитку добровільного медичного страхування в Україні. *Науковий вісник Херсонського державного університету*. 2018. Випуск 28. Частина 2. С. 45–48
15. Офіційний сайт Світового банку. URL: <https://www.worldbank.org/uk/country/ukraine> (дата звернення: 05.05.2024)
16. European Insurance. URL: <https://www.insuranceeurope.eu/insurancedata> (дата звернення: 05.05.2024)
17. Економетрія: Навч. посіб./ В.І. Жлуктенко, Н.К. Водзянова, С.С. Савіна, О.В. Колодінська. — К.: Вид-во Європ. Ун-ту, 2005. — 552 с.

18. Інформація про стан і розвиток страхового ринку. Офіційний сайт Національної комісії, що здійснює державне регулювання у сфері фінансових послуг. URL: <https://www.nfp.gov.ua/ua/Ohliad-strakhovoho-ryнку.html> (дата звернення: 05.05.2024)

19. Insurance Top. URL: https://forinsurer.com/insurancetop_64 (дата звернення: 05.05.2024)

20. Журавка О. С., Пахненко О. М., Бухтіарова А. Г. Страхування: навчальний посібник. Суми, 2020. – 350 с.

21. Гаманкова О. О. Фінанси страхових організацій : навч. посіб. Київ, КНЕУ, 2007. 328 с.

22. Журнал «Insurance TOP». Сайт інформаційного агентства 3(67). URL: https://forinsurer.com/insurancetop_67 (дата звернення: 05.05.2024).

23. Єрємін М.О. Економетрика: Конспект лекцій. Харків: УкрДУЗТ, 2015. – 40 с. URL: <http://lib.kart.edu.ua/bitstream/123456789/2798/1/%D0%9A%D0%BE%D0%BD%D1%81%D0%BF%D0%B5%D0%BA%D1%82%20%D0%BB%D0%B5%D0%BA%D1%86%D1%96%D0%B9.pdf> (дата звернення: 05.05.2024)

24. Friedman J., Tibshirani R., Hastie T., The elements of statistical learning. Data mining, inference and prediction, Springer series in Statistics, 2015. URL: <https://hastie.su.domains/Papers/ESLII.pdf> (дата звернення: 05.05.2024)

25. Ієрусалимов В., Фесенко А., Марчак Д. Дослідження: аналіз ефективності підтримки застрахованих осіб Фондом соціального страхування України. 2022. 29с. <https://www.rada.gov.ua/news/razom/225965.html> (дата звернення: 05.05.2024)

26. Лісовська О.О. Функціонування ринку медичних страхових послуг в період пандемії COVID-19 в Україні. *East European Scientific Journal*, vol.2(66), 2021. pp.7-13. URL: <http://www.economy.nayka.com.ua/?op=1&z=8268> (дата звернення: 05.05.2024).

27. Консолідовані звітні дані страхового ринку України / Офіційний сайт Національної комісії, що здійснює державне регулювання у сфері ринків

фінансових послуг. URL:
<https://www.nfp.gov.ua/ua/Konsolidovani-zvitni-dani.html> (дата звернення:
05.05.2024).

28. Кострач Л. М., Рудь Л. О. Тенденції розвитку страхових компаній в Україні. *Збірник наукових праць Національного університету державної податкової служби України*. 2015. №2. С. 135–153.

29. Офіційна безкоштовна платформа програмних інструментів мови R.
URL: <https://cran.r-project.org/> (дата звернення: 05.05.2024)

30. Репозиторій машинного навчання UCI Machine Learning Repository
URL: <http://mlr.cs.umass.edu/ml/> (дата звернення: 05.05.2024)