Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Кафедра обчислювальної математики факультету кібернетики

Аналіз динаміки та факторів впливу на кількість аспірантів в Україні

Текстова частина до курсової роботи
за спеціальністю „Прикладна математика”

Керівник курсової роботи
к.ф.-м.н., асистент Затула
Дмитро Васильович

Виконав студент групи ОМ-3
Кроча Кирило Геннадійович
01.06.2023 р.

Київ 2023

ЗМІСТ

Вступ

Актуальність теми: Аналіз якості вищої освіти завжди актуальна тема, їй присвячено багато досліджень по усьому світу, але оскільки оцінка якості освіти є суб'єктивною, різні дослідження у цій області орієнтуються на різні критерії оцінювання, деякі більш ефективні, деякі менш ефективні, тому аналіз якості освіти ніколи не втрачає актуальність, достатньо лише обрати доцільні критерії оцінювання. Також через те, що дослідження буде проводитись за допомогою регресійного аналізу, тема є актуальною, бо дозволить прогнозувати майбутні оцінки, а також робити висновки про залежності оцінки від певних параметрів.

Мета і завдання дослідження. Метою цього є дослідження якості вищої освіти за допомогою регресійного аналізу залежності кількості аспірантів від різноманітних даних, які будуть обратись в залежності від логіки, та доступності у відкритих джерелах.

Об'єкт дослідження. Об'єктом цього дослідження є кількість аспірантів в Україні.

Методи дослідження. У дослідженні використовуються методи регресійного аналізу, що представлені у пакеті «Аналіз даних» у програмі Excel.

Практичне значення одержаних результатів. Отримані результати можуть використовуватись у секторі освіти для збільшення кількості аспірантів за допомогою зміни параметрів, залежність від яких буде знайдена, а також для прогнозування майбутньої кількості аспірантів за цими параметрами.

Розділ 1. Теоретичні відомості

1.1 Теоретична частина. Парна регресія

Парна регресія – це модель лінійної регресії, яка є спробою пояснити зміни у залежній змінній Y за допомогою змін у незалежній змінній X. Це означає побудову залежності вигляду(якщо вона існує):

$$Y_k = \beta_0 + \beta_1 * X_k + \varepsilon_k$$

Основною задачею регресійного аналізу є саме оцінювання коефіцієнтів у цьому рівнянні, а також оцінка похибки ε_k, для того, щоб для вибірки даних створити рівняння, за яким можливо оцінювати Y:

$$\widehat{Y_k} = \widehat{\beta_0} + \widehat{\beta_1} * X_k$$

Ця залежність будується за методом найменших квадратів – ми мінімізуємо суму квадратів відхилень значень із вибірки від нашої лінії.

Для обчислення коефіцієнтів $\widehat{\beta_0}$, $\widehat{\beta_1}$ можна користуватись формулами(тут $\bar{y}$ – це середнє значення по вибірці):

$$\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} * \bar{x}$$

Але на практиці це завжди робиться автоматично за допомогою засобів для аналізу даних, що обирає дослідник.

Слід також зауважити, що інколи рівняння регресії записують наступним чином:

$$\bar{y} = \bar{y} + r\frac{\sigma_x}{\sigma_y}(x - \bar{x})$$

Де ${\sigma^2}_x = \frac{1}{n}\sum(x_i - \bar{x})^2$, ${\sigma^2}_y = \frac{1}{n}\sum(y_i - \bar{y})^2$ – вибіркові дисперсії, а $r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x * \sigma_y}$ – вибірковий коефіцієнт кореляції[4].

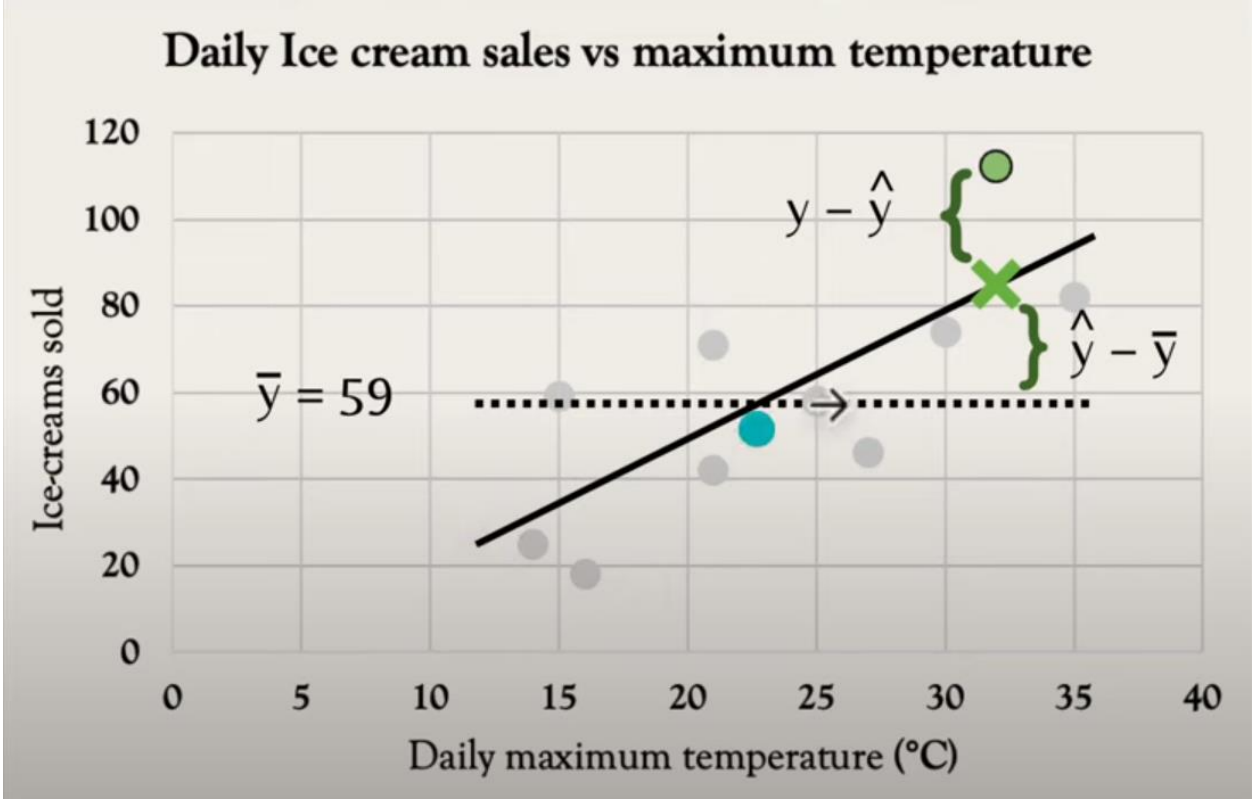
Дуже важливою частиною регресійного аналізу є значення R^2-коефіцієнту детермінації, яке характеризує частину змін у залежній змінній, яку ми можемо пояснити за допомогою незалежного(-их) параметру(-ів). Для пояснення цієї величини введемо декілька позначень:

SST= $\sum(y_i - \bar{y_i})^2$ - Sum of Squares Total (Загальна сума квадратів)

SSR= $\sum(\hat{y_i} - \bar{y_i})^2$ – Sum of Squares due to Regression (Регресійна сума квадратів)

SSE = $\sum(y_i - \hat{y_i})^2$– Sum of Squares due to Error (Залишкова сума квадратів).

Практичне значення цих величин можна побачити на малюнку:



Мал.1.1

Очевидно, що SST= SSE+ SSR, а у найкращому випадку(ідеальна модель регресії) SST = SSR За визначенням, $R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE}$. Тому ця величина буде коливатись від 0 до 1, даючи нам гарне уявлення про відсоток коливань, що пояснюються нашою регресією, даючи нам кількісну оцінку якості нашої моделі.

Для того, щоб регресія відображала реальну залежність Y від X, коефіцієнт β_1 має бути відмінним від нуля. Саме тому нам треба перевірити гіпотезу $H_0: \beta_1 = 0$ при альтернативі $H_0: \beta_1 \neq 0$. Якщо з'ясується, що $\beta_1 \neq 0$, то регресію визнаємо значущою[4]. Для прийняття або відхилення гіпотези використовують так звану F-статистику, яка може бути обчислена як $\frac{SSR}{SSE/2}$, та також автоматично обчислюється будь-яким засобом роботи з даними. Для отриманої F-статистики також автоматично рахується p-критерій, який показує ймовірність отримати значення більше за F. Зазвичай береться рівень значущості 0,05(але може бути змінений в залежності від задачі), тоді при p-критерії меншому за рівень значущості ми можемо із відповідним рівнем значущості відкинути нуль-гіпотезу.

Зауваження 1. При побудові парної лінійної регресії не обов'язково шукати регресію у вигляді прямої лінії, можна розглядати довільну функцію від X, зробивши відповідну заміну змінних.

Зауваження 2. При побудові прямої регресії припускається, що $\varepsilon_k \sim N(0; \sigma^2)$, тому варто перевіряти нормальність розподілу залишків. При

виконанні припущення модель називається нормальною парною регресією, а незсуненою та ефективною оцінкою параметру σ^2 є[5]

$$\widehat{\sigma^2} = \sum \frac{(y_i - \hat{y})^2}{n-2}$$

1.2 Теоретична частина. Множинна регресія

Ціль побудови моделі множинної лінійної регресії – побудувати лінійну залежність даних Y від $X_1, X_2, \dots, X_n$ – незалежних параметрів, таким чином, щоб графік отриманої прямої проходив максимально близько до усіх точок. Побудована модель у найпростішому випадку виглядає наступним чином:

$$\widehat{Y\_k} = \widehat{\beta_0} + \widehat{\beta_1} * X_{1k} + \cdots + \widehat{\beta_n} * X_{nk}, \text{ k=1..p}$$

Але як і випадку парної регресії можлива заміна змінних.

Загалом усі назви означень переносяться із розділу парної регресії, та продовжують означати ті самі, або схожі речі, тому наведемо зміни у означеннях:

По-перше, тепер ми маємо не список точок, а матрицю з x_{ij} з розмірністю n на p, також тепер при умові rank(X)=p, та при припущенні нормальності розподілу похибок маємо наступну незміщену та ефективну оцінку σ^2[5]:

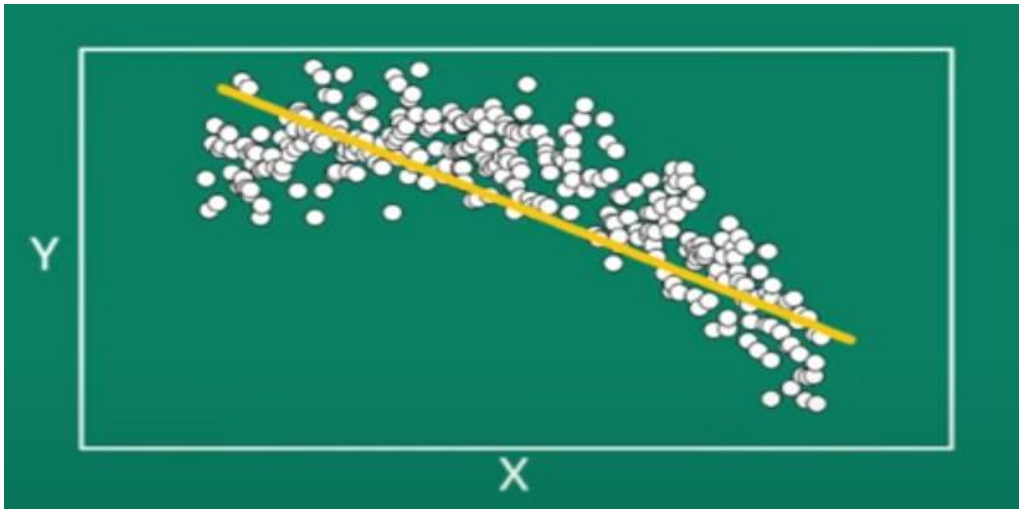
$$\widehat{\sigma^2} = \sum \frac{(y_i - \hat{y})^2}{n-p}$$

Незначна зміна у визначені F-статистики – тепер вона визначається як $\frac{SSR/(p-1)}{SSE/(n-p)}$.

Також тепер при додаванні змінних, що не впливають на залежну величину, коефіцієнт детермінації все одно може збільшуватись тому вводиться модифікований коефіцієнт $AdjR^2 = 1 - \left[(1-R^2)\left(\frac{n-1}{n-k-1}\right)\right]$, який не має цієї властивості, таким чином використовуючи різницю модифікованих коефіцієнтів детермінації до та після введення параметру можна розуміти, чи впливає введений параметр на регресію.

При множинній регресії, на відміну від парної робиться певна кількість припущень, які дозволяють гарантувати надійність числових оцінок коефіцієнтів нашої моделі. Список таких припущень може відрізнятись в залежності від літератури, але загалом кожен з таких списків означає схожі до інших речі, тому наведемо одну з варіацій такого списку:

Припущення при побудові моделі. Лінійність. Це припущення означає що коефіцієнти в рівнянні регресії беруть участь як лінійні доданки, тобто що ми використовуємо підходящу функціональну форму. Це не означає, що ми не можемо використовувати заміну змінних, навпаки, функціональне відношення, що визначено дослідником, має бути коректним. Приклад некоректно визначеної залежності(тут проблему вирішить введення квадрату незалежної змінної у регресію):



Які проблеми випливають при відсутності цього припущення? Це припущення є дуже важливим, оскільки якщо воно не виконується, то усі дані отриманої регресії некоректні – від оцінки коефіцієнтів до стандартної похибки.

Як виявляти проблему? Головний та найпростіший спосіб – за допомогою графіку залишків візуально визначати, чи існує краще наближення іншою функцією.

Як боротися з проблемою? Єдиним способом боротьби з відсутністю лінійності є вибір іншого типу залежності

Припущення при побудові моделі. Відсутність гетероскедастичності.

Відсутність гетероскедастичності(heteroscedasticity) означає, що ми припускаємо, що похибки нашої моделі мають постійну дисперсію.

Які проблеми випливають при відсутності цього припущення? Це припущення не є настільки важливим, як попереднє, тому в цілому модель залишиться коректною, тільки на стандартні похибки вже не можна покладатися.

Як виявляти проблему? У цьому може допомогти Goldfeld-Quant test, який полягає у тому, що ми розбиваємо дані на дві частини, та будуємо модель регресії для кожної з частин, та порівняти їх дисперсії, тобто маємо $F = \frac{SSE_A/(n_A-2)}{SSE_B/(n_B-2)}$, що розподілена як F-статистика F_{n_A-2,n_B-2}, тому при p-значенні, пов'язаному з цим розподілом, меншому за 0,05 ми маємо гетероскедастичність у моделі.

Як боротися з проблемою? Найпопулярнішим інструментом є стандартні похибки Вайта, які обчислюються наступним чином:

$V(\hat{\beta}) = (X^T X)^{-1}\left(X^T \begin{bmatrix} e_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e_n^2 \end{bmatrix} X\right)(X^T X)^{-1}$, тут e_k- похибки у k-ій точці.

Припущення при побудові моделі. Незалежність похибок.

Незалежність похибок, або відсутність автокореляції означає, що ми не можемо передбачити наступне значення похибки за попереднім.
Які проблеми випливають при відсутності цього припущення? При невиконанні цього припущення на стандартні похибки не можна покладатися.
Як виявляти проблему?Durbin-Watson test, що полягає у обчисленні наступної статистики: $dw = \frac{(e_2-e_1)^2+\cdots+(e_n+e_{n-1})^2}{e_1^2+\cdots e_n^2}$. Це число буде лежати в межах від 0 до 4, а межі у яких присутня автокореляція визначаються за таблицями[6].
Як боротися з проблемою? Гарантованого способу не існує, але треба перевірити модель на наявність «опущених» змінних, тобто таких, від яких наша залежна змінна залежить, але яких нема у моделі.

Припущення при побудові моделі. Нормальність похибок.
Це припущення вже було описано в розділі про парну регресію, воно означає, що ми припускаємо, що наші похибки мають нормальний розподіл.
Які проблеми випливають при відсутності цього припущення? Якщо нормальність порушена, та мир маємо мальку кількість спостережень(n<10), то на стандартні похибки не можна покладатися, але при більшій кількості спостережень ця проблема перестає впливати на похибки.
Як виявляти проблему? Найпростішим способом є порівняння гістограми похибок із гістограмою нормального розподілу.
Як боротися з проблемою? Змінити тип функціональної залежності у регресії.

Припущення при побудові моделі. Відсутність мультиколінеарності.
Мультиколінеарність виникає, коли незалежні змінні у нашій моделі корелюють між собою, що не є гарним для нашої моделі.
Які проблеми випливають при відсутності цього припущення? При наявності мультиколінеарності як коефіцієнти, так і стандартні похибки є ненадійними. Також при дуже близьких до 1 значеннях кореляції між змінними стає неможливим взагалі побудувати модель.

Як виявляти проблему? Найпростішим є простий підрахунок та побудова кореляційної матриці для усіх параметрів, та подальше виключення даних, що мають зависоке значення кореляції з іншими параметрами(зазвичай про виключення треба задуматись при значеннях, що за модулем перевищують 0,9), при цьому вибір, який з двох параметрів виключити, робить дослідник в залежності від задачі, характеристик парної регресії по кожному з параметрів, та своєї інтуїції.

Інший, більш надійний спосіб – підрахунок коефіцієнту інфляції дисперсії(VIF), який полягає у тому, що почергово кожен параметр, щодо якого ми маємо підозру про мультиколінеарність, ми розглядаємо як залежну

величину від інших параметрів. Так, наприклад, якщо ми перевіряємо параметр X_p, то будуємо модель множинної регресії:

$$X_p = \gamma_0 + \gamma_1 * X_1 + \cdots + \gamma_n * X_n$$

Далі для отриманої моделі ми рахуємо ${R_p}^2$, і за допомогою наведеної нижче формули рахуємо VIF:

$$VIF = \frac{1}{1 - {R_p}^2}$$

Загалом проблемними є значення VIF більші за 10, але це не чітке правило, тому завжди рішення щодо проблемності змінної приймається дослідником.
Як боротися з проблемою? Видаляти одну із корелюючих змінних.

Припущення при побудові моделі. Відсутність залежності від «опущених» змінних.

Це припущення означає, що за межами моделі нема змінних, що впливають і на залежну змінну, і на незалежні параметри.
Які проблеми випливають при відсутності цього припущення? При наявності «опущених» змінних модель може без проблем застосовуватись для передбачення майбутніх значень залежної змінною, проблеми можуть виникати при поясненні причинно-наслідкових зв'язків.
Як виявляти проблему? Єдиним інструментом виявлення опущених змінних є інтуїція та логіка дослідника.
Як боротися з проблемою? Вводом у модель опущених змінних.

ВИСНОВОК

Отже, ми розглянули основні теоретичні підстави та методи побудови регресійних моделей, тепер можемо перейти до практичної частини – регресійного аналізу досліджуваної величини.

Розділ 2. Кількість аспірантів

2.1 Загальний огляд

Кількість аспірантів в Україні будемо розглядати як залежну величину від змінних, список яких наведено нижче:

1. Населення
2. Кількість університетів та академій
3. Відсоток населення з підключенням до інтернету
4. Курс долара до гривні
5. Курс євро до гривні
6. Частка кількості промислових підприємств, що впроваджували інновації
7. Очікувана тривалість життя
8. Кількість працюючих людей віком 15-70 років
9. Викиди забруднюючих речовин та діоксиду вуглецю в атмосферне повітря

Усі дані відібрані для України у період 2000-2020 років з відкритих джерел, список яких наведено у кінці роботи.

Графік кількості аспірантів по рокам:

Мал. 2.1 Графік кількості аспірантів

2.2 Описова статистика

Почнемо з обчислення основних статистичних характеристик для наших даних. Використовуючи вбудовані функції Excel, маємо:

Описова статистика	
Медіана	28412
Максимум	34653
Мінімум	22829
Середнє	28763,7619
Середньоквадратичне відхилення	3961,646046
Асиметрія	0,132146691

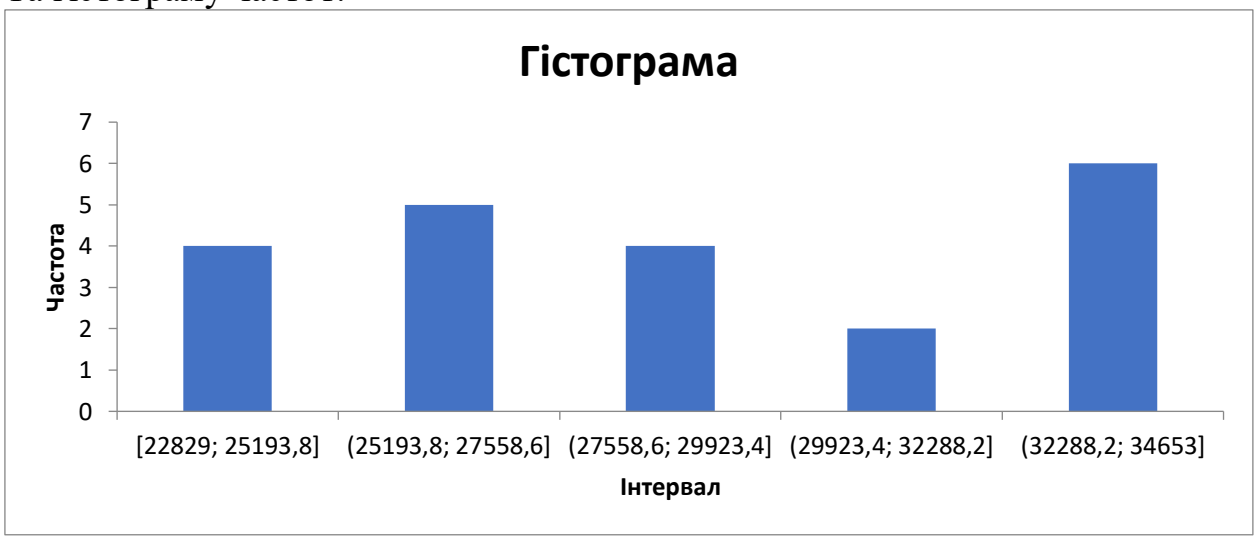
Табл. 2.1

В цілому дані мають доволі велике відхилення – майже 14 відсотків від середнього. Також з асиметрії бачимо, що значення кількості аспірантів у незначній мірі зміщені в сторону мінімуму.
Тепер побудуємо гістограму і таблицю частот, для обчислення кількості інтервалів скористаємось м із варіантів формули Стерджеса: $[ln(n) + 1]$, де n – кількість вимірів. Таким чином, маємо 5 інтервалів розбиття, кожен довжиною $\frac{(x_{max} - x_{min})}{n}$. Обчислюючи частоту потрапляння в кожний з проміжків, отримуємо:

Інтервал	*Частота*
[22829; 25193,8]	4
(25193,8; 27558,6]	5
(27558,6; 29923,4]	4
(29923,4; 32288,2]	2
(32288,2; 34653]	6

табл.2.2

Та гістограму частот:



Мал. 2.2 Гістограма частот для кількості аспірантів

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

Перейдемо до більш детального розгляду параметрів, залежність від яких ми будемо досліджувати.

Розділ 3. Населення

3.1 Загальний огляд

Населення є однією із головних характеристик будь-якої держави, тому цілком природньо припустити, що від нього буде залежати кількість аспірантів. Тим паче, логіка підказує, що оскільки ми беремо саме кількість, а не відсоток, то чим більше людей в країні, тим більші усі кількісні характеристики. Тому цікаво буде перевірити це твердження більш формалізовано. Маємо таблицю населення в тисячах осіб за період 2000-2020рр. Джерелом статистичної інформації є Державна служба статистики України.[1]. Графік населення по рокам:

Мал.3.1, Населення

3.2 Описова статистика

Обчислимо основні статистичні характеристики нашої таблиці, для цього скористаємось функціями Excel, отримуємо наступні дані:

Описова статистика	
Медіана	45 962,90
Максимум	49 429,80
Мінімум	41 902,40
Середнє значення	45660,91429
Середньоквадратичне відхилення	2351,681654
Асиметрія	-0,284724619

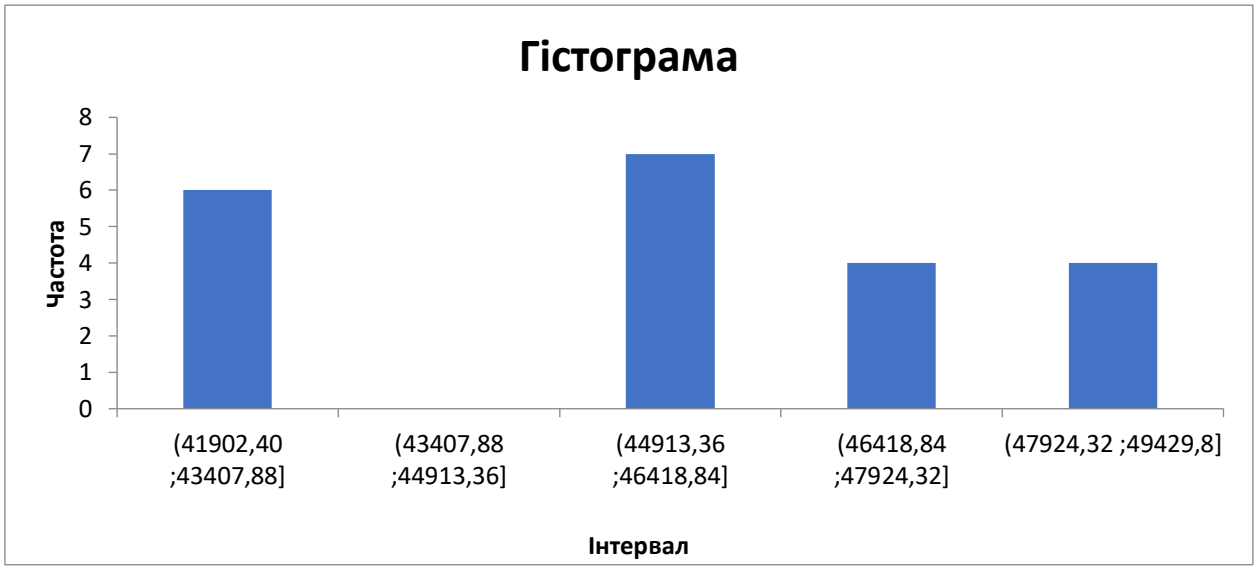
Табл.3.1, Описова статистика населення.

З від'ємності асиметрії бачимо, що в середньому значення знаходяться трохи ближче до максимуму, щоб подивитись на це більш детально, побудуємо гістограму розподілу. Для цього обчислимо кількість інтервалів для розбиття за формулою Стерджеса: $[ln(n) + 1]$, де n – кількість вимірів. Отже, маємо 5

інтервалів розбиття, кожен довжиною $\frac{(x_{max}-x_{min})}{n}$. Обчислюючи частоту потрапляння в кожний з проміжків, отримуємо:

Інтервал	*Частота*
(41902,40 ;43407,88]	6
(43407,88 ;44913,36]	0
(44913,36 ;46418,84]	7
(46418,84 ;47924,32]	4
(47924,32 ;49429,8]	4

Табл.2.2 Таблиця частот для населення



Мал.3.2 Гістограма частот населення

Є дуже слабка схожість на гістограму нормального розподілу, яка скоріш за все не виявиться корисною в дослідженні.

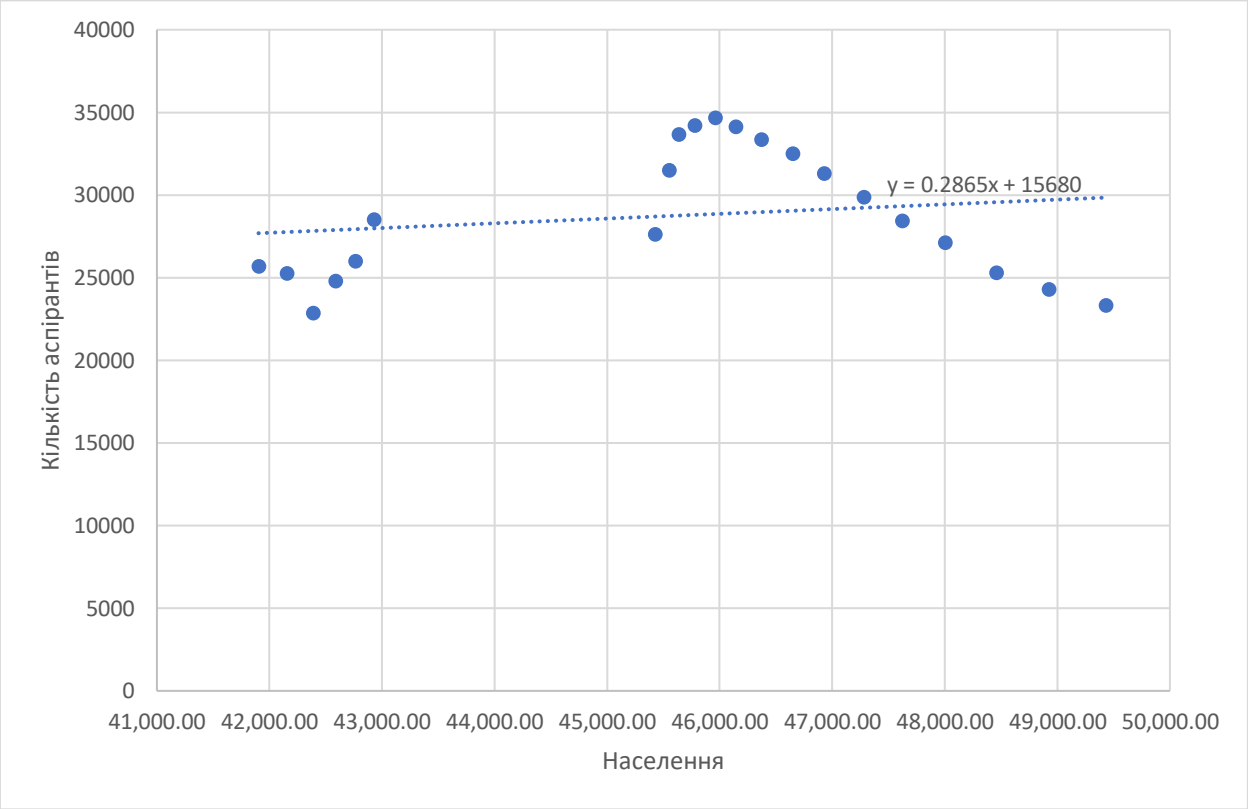
3.3 Побудова моделі парної регресії

Перш ніж будувати модель множинної регресії, побудуємо парну модель для населення та кількості аспірантів, для того, щоб оцінити значущість, та подивитись, чи треба взагалі включати в множинну регресію кількість населення.

Скористаємось надбудовою «Аналіз даних» та обчислимо кореляцію між населенням та кількістю аспірантів, отримуємо 0,170088995-дуже слабку кореляцію, що не підтверджує наше припущення про те, що ці величини мають бути тісно пов’язані.

Перейдемо до побудови моделі. За допомогою згаданої вище надбудови «Аналіз даних» будуємо модель парної лінійної регресії,

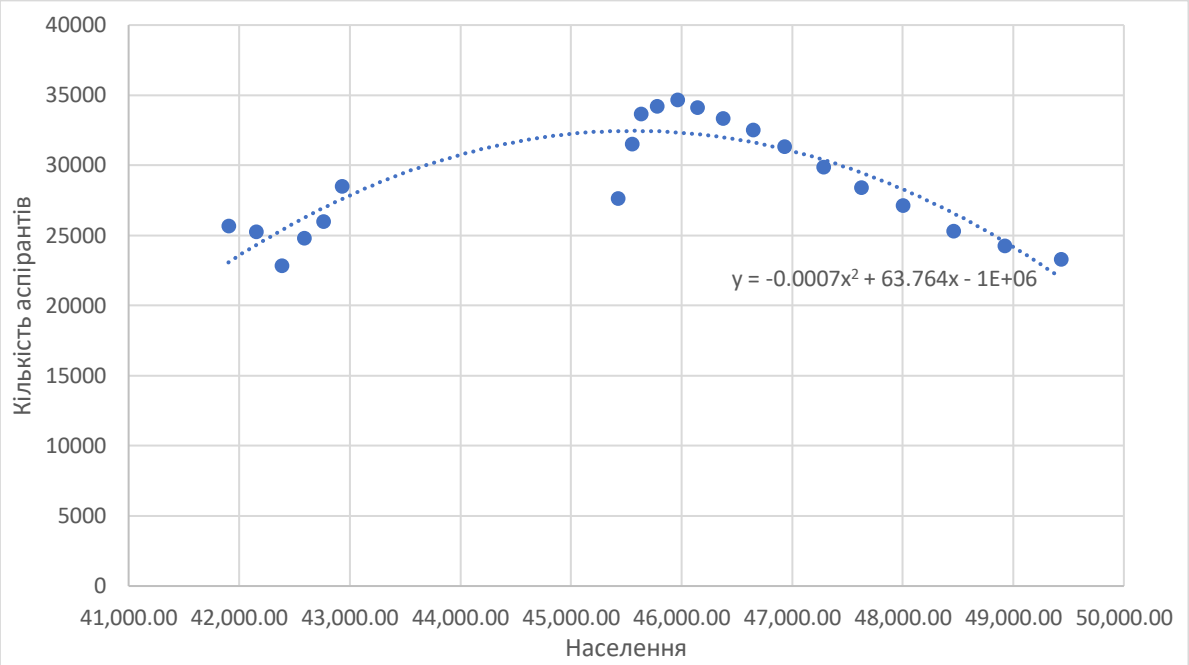
отримуємо наступний графік залежності та пряму регресії:



Бачимо, що пряма не відображає ніякої залежності між даними, це змушує нас задуматись про те, що тип функціональної залежності ми обрали невірно, а це буде значити, що усі числові характеристики регресії є ненадійними, тому нема жодного сенсу говорити про отримані оцінки.

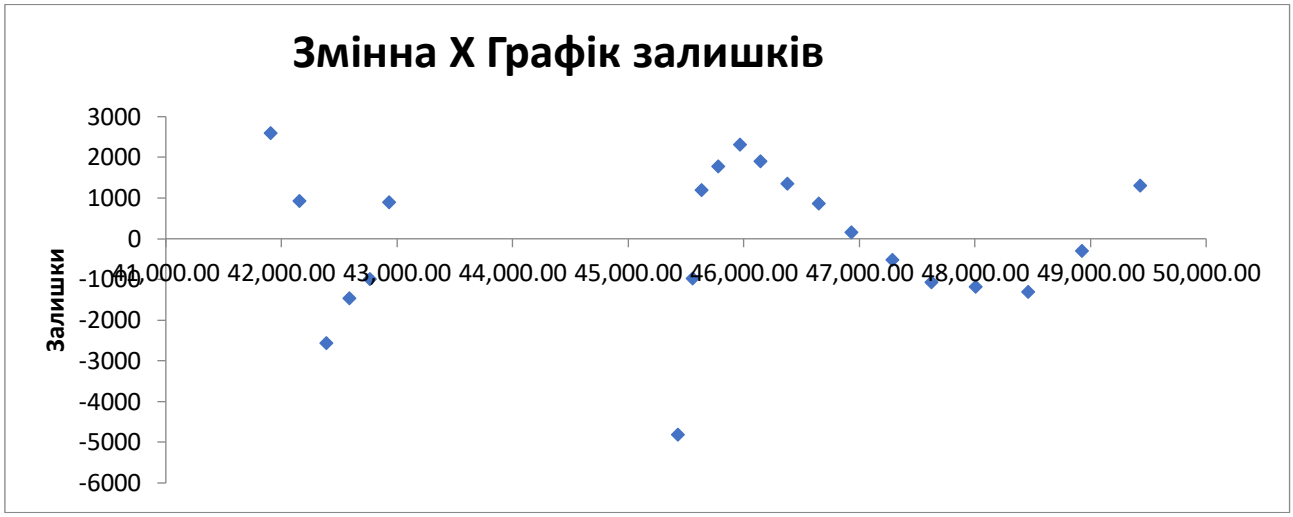
Візуальний аналіз не дає чіткої відповіді, чи є якась інша функціональна залежність, тому спробуємо перевірити найпоширеніші варіанти, почнемо із залежності виду $\hat{Y} = \hat{\beta}_0 + \widehat{\beta_1} * X + \widehat{\beta_2} * X^2$, обчислюючи значення X^2 за допомогою вбудованих функцій Excel, та включаючи його у

регресію, маємо:



Очевидно, що це вже набагато краще наближення потенційної функціональної залежності. Перевіримо, чи виконуються наші припущення. Перше, що приходить на розум при погляді на графік – можлива автокореляція. Кодуємо функцію для обчислення статистики тесту Дюрбіна-Вотсона, отримуємо значення 1,209687653, в таблиці маємо значення(для рівня значущості 0,05) критичних точок 1.13 та 1.54, наше значення потрапляє усередину інтервалу, це означає, що нам не пощастило, і тест не може сказати що автокореляції зовсім нема, як і того, що вона суттєво вплине на результат регресії. Таким чином рішення чи суттєва автокореляція, залишається робити мені самостійно. Оскільки я не бачу кращої функціональної залежності, а також того, що на обидві величини впливає інший фактор(оскільки я вважаю, що населення – це доволі глобальна характеристика, на яка не має тенденції залежати від чогось іншого), то я буду вважати, що автокореляція несуттєва.

З іншими припущеннями легше – нормальність має виконуватись через достатньо велику кількість спостережень, опущених параметрів я не помічаю, а модель найкраща з усіх простих відомих мені залежностей. За графіком залишків, бачимо що дисперсію можна вважати константою:



Отже, робимо висновок, що модель парної регресії коректна, тому оцінки усіх параметрів будуть надійними.

Перейдемо до аналізу отриманих числових значень. Значення коефіцієнту детермінації складає 0,794177506, це означає, що побудована нами модель пояснює коливання кількості аспірантів на 79%, що є дуже гарним результатом. Причому можна із впевненістю сказати, що ми відхиляємо нуль-гіпотезу. оскільки p-критерій має надзвичайно мале значення - 6,62881E-07, отже побудована нами модель є значущою. Перевіряючи це на практиці, можна побачити, що ми можемо передбачати значення кількості аспірантів за побудованою моделлю із середньою точністю 5,16%.

ВИСНОВОК

У результаті попереднього регресійного аналіз ми вже отримали доволі якісну регресійну модель: *$Y= -0,0007*X^2 + 63,764*X - 1E+06$(тут Y – це кількість аспірантів(осіб), а X – населення(тис. осіб)),* яка дозволяє не тільки сказати, що існує залежність між кількістю аспірантів, але і передбачати майбутні значення кількості аспірантів із непоганою точністю.

Розділ 4. Кількість університетів та академій

4.1 Загальний огляд

Дуже природним буде дослідити залежність(яка за логікою має існувати) між кількістю університетів та аспірантів, тому в побудову множинної регресії також включимо цей параметр. Графік зміни кількості університетів та академій за роками:

Мал. 4.1 Графік зміни кількості університетів

4.2 Описова статистика

Як і для попередніх параметрів, обчислюємо основні характеристики:

Описова статистика	
Медіана	330
Максимум	353
Мінімум	277
Середнє значення	322,2857
Середньоквадратичне відхилення	27,57561
Асиметрія	-0,50451

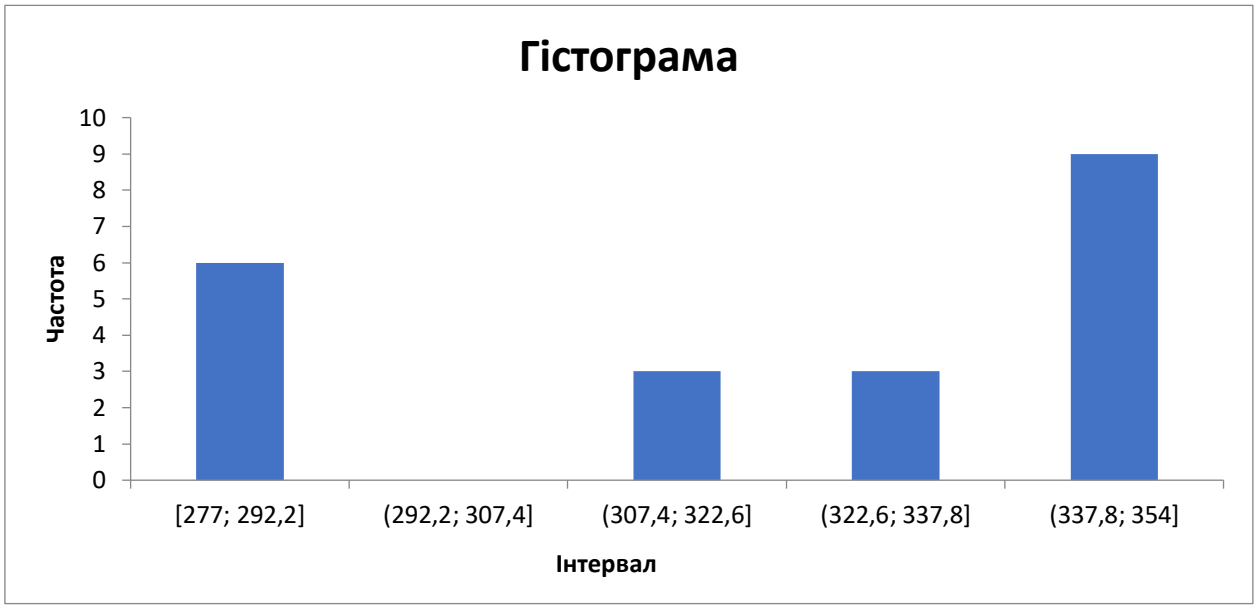
Табл. 4.1

Бачимо, що при помірному середньоквадратичному відхиленні значення загалом знаходяться ближче до мінімуму.

За вже наведеними формулами побудуємо таблицю та гістограму частот:

Табл. 4.2

Інтервал	*Частота*
[277; 292,2]	6
(292,2; 307,4]	0
(307,4; 322,6]	3
(322,6; 337,8]	3
(337,8; 354]	9



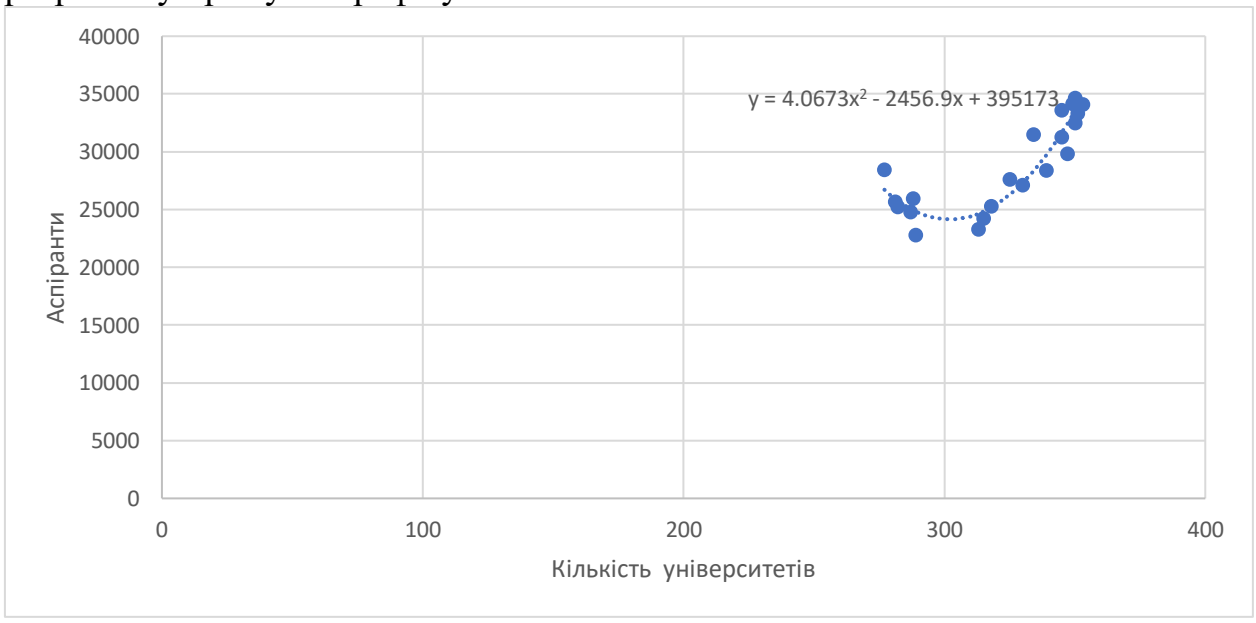
Мал. 4.2. Гістограма частот

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

4.3 Побудова моделі парної регресії

Як і раніше рахуємо кореляцію, та отримуємо значення 0,79, що означає сильну кореляцію між досліджуваними величинами.

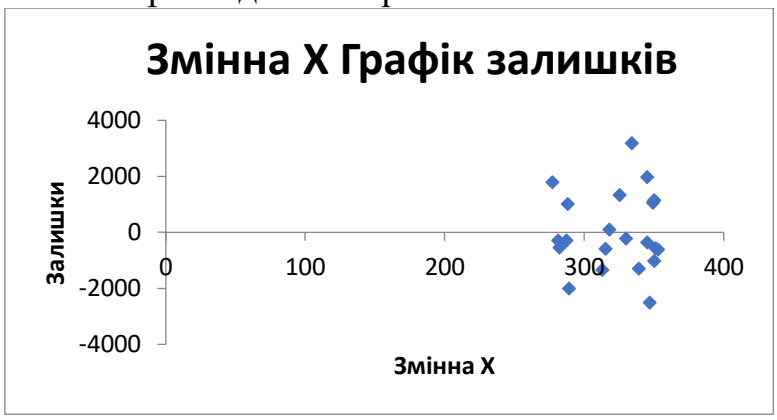
Перейдемо до побудови моделі, на цей раз по графіку залежності очевидно, що кількість аспірантів має залежати від квадрату кількості університетів, тому одразу додаємо цей параметр в модель, та отримуємо наступну регресійну пряму на графіку залежності:



Мал. 4.3 Графік залежності

Це наближення очевидно є дуже гарним, що підтверджують числові характеристики регресії – модель є значущою бо значення p-критерію

дорівнює 6,9653E-09. Також гарним є значення коефіцієнту детермінованості - 0,875931949, отже коливання кількості аспірантів можна пояснити коливаннями кількості університетів на 87%. Всі отримані характеристики є гарними, але необхідно також перевірити виконання припущень. Дивлячись на графік залишків, не виконає ніяких сумнівів, що автокореляція відсутня, а похибки розподілені нормально та мають постійну дисперсію:



Мал. 4.4 Графік залишків

Вважати, що квадратична модель не підходить, як і те що існують опущені змінні, в даному випадку немає сенсу. Отже, нашу модель можна вважати коректною. Дійсно, наша модель на практиці прогнозує кількість аспірантів із середньою точністю 3,83%, що підтверджує правильність побудованої моделі.

ВИСНОВОК

У результаті попереднього регресійного аналіз ми вже отримали доволі якісну регресійну модель: $Y = 4{,}0673 * X^2 - 2456{,}9 * X + 395173$ *(тут Y – це кількість аспірантів(осіб), а X – кількість університетів(штук)),* яка дозволяє не тільки сказати, що існує залежність між кількістю аспірантів та числом університетів, але і передбачати майбутні значення кількості аспірантів із гарною точністю.

Розділ 5. Відсоток населення з підключенням до інтернету

5.1 Загальний огляд

Відсоток підключених до інтернету хоча і не перше, від чого хочеться дослідити залежність кількості аспірантів, але очевидно, що чим більше людей мають доступ до інтернету, то тим більше мають доступ до технологій дистанційного навчання, тому цілком можливо, що цей параметр дійсно буде впливати на досліджувані нами дані. Для початку подивимось на графік(дані отримано з світового банку відкритих даних[3]):

Мал. 5.1 Графік доступу до інтернету по рокам

5.2 Описова статистика

Основні статистичні характеристики даних:

Описова статистика	
Медіана	23,3
Максимум	75
Мінімум	0,7
Середнє значення	28,41428571
Середньоквадратичне відхилення	25,58992547
Асиметрія	0,443042501

Табл. 5.1

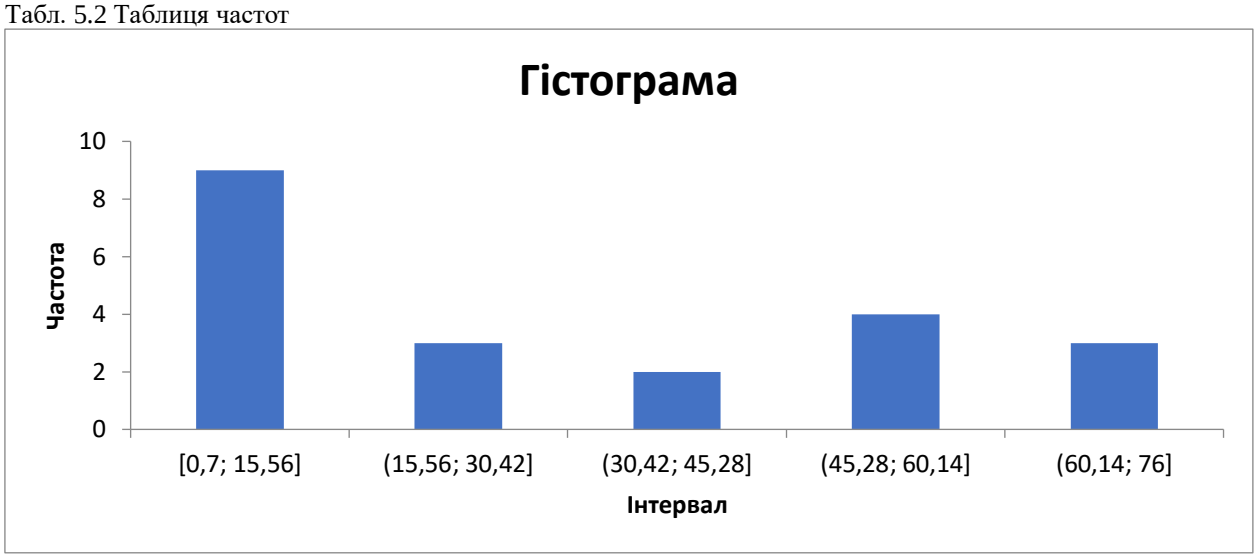
Бачимо, що на цей раз дані мають дуже велике середньоквадратичне відхилення – майже таке велике, як і середнє значення.

Тепер подивимось на таблицю та гістограму частот:

Інтервал	*Частота*
[0,7; 15,56]	9
(15,56; 30,42]	3
(30,42; 45,28]	2

(45,28; 60,14]	4
(60,14; 76]	3

Табл. 5.2 Таблиця частот



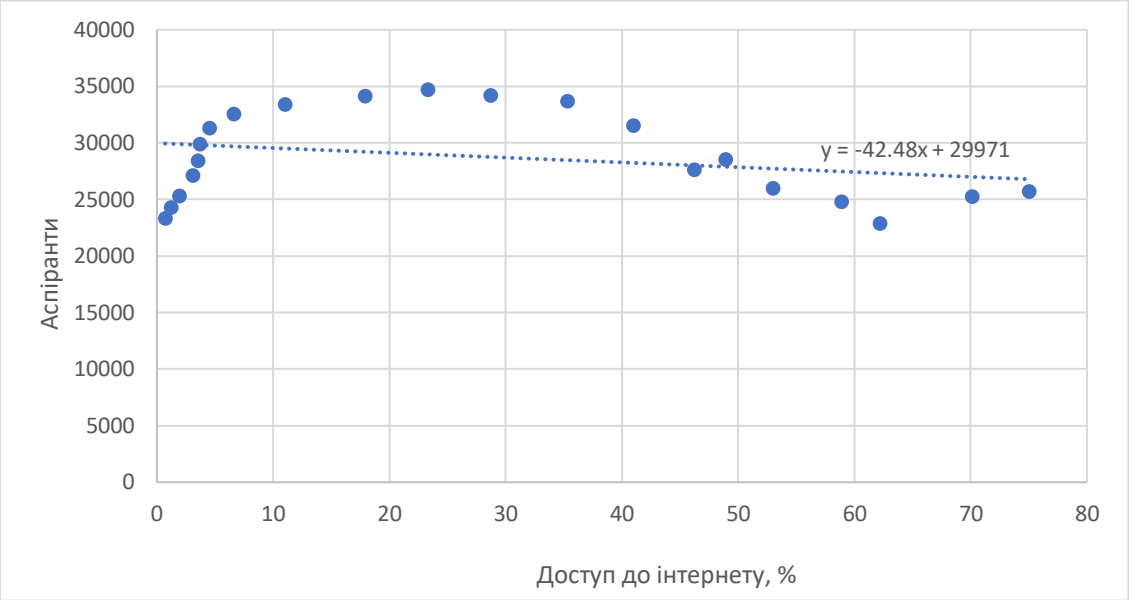
Мал.5.2. Гістограма частот

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

5.3 Побудова моделі парної регресії

Кореляція між величинами низька зворотна: -0,274396754.

Будуємо парну регресію, оскільки на графіку нема чіткої залежності у вигляді відомої мені простої функції, тому спробуємо побудувати лінійну залежність. В результаті отримуємо досить погане наближення із значенням p-критерію 0,228697474, що означає(оскільки ми вважаємо рівень значущості $\alpha = 0{,}05$), що ми маємо прийняти нуль-гіпотезу про те, що усі коефіцієнти в нашому розкладі мають бути нульові, тому ми робимо висновок, що доля людей з доступом до мережі інтернет не впливає на кількість аспірантів. Спроба побудувати залежність від квадрату також не дає достатнього значення p-критерію, для того, щоб відхилити нуль гіпотезу. Залишається вірогідність існування більш складної поліноміальної залежності 3-го порядку, або вище, але оскільки досліджувати подібну залежність доволі важко, і нема основи припускати, що цей фактор сильно впливає на кількість аспірантів, мною було прийнято рішення про видалення з розгляду цього параметру.



Мал.5.3. Графік залежності

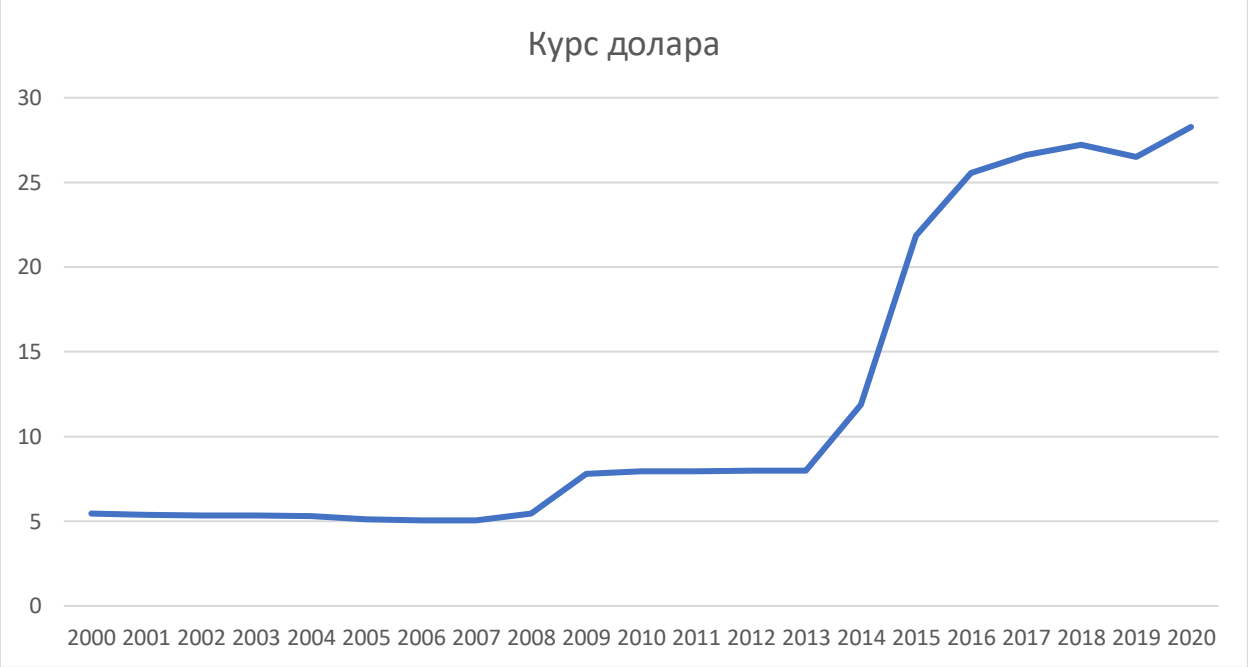
ВИСНОВОК

У результаті попереднього регресійного аналіз ми встановили відсутність лінійної або квадратичної залежності між доступом до інтернету та кількістю аспірантів, що дозоляє нам виключити з розгляду у моделі множинної регресії цей параметр, Таким чином, попередній аналіз дозволив нам спростити задачу.

Розділ 6. Курс долара

6.1 Загальний огляд

Курси валют є універсальною величиною, залежність від якої досліджувати майже завжди доречно, тому також у розгляд введемо курси євро та долара. Для обох величин будемо брати середньорічне значення курсу до гривні. Джерело інформації – Національний Банк України[2]. Графік курсу по роках:



Мал.6.1 Графік курсу долара

6.2 Описова статистика

Основні статистичні характеристики курсу долара:

Описова статистика	
Медіана	7,9347
Максимум	28,2746
Мінімум	5,05
Середнє	12,14428571
Середньоквадратичне відхилення	9,194783474
Асиметрія	0,969600934

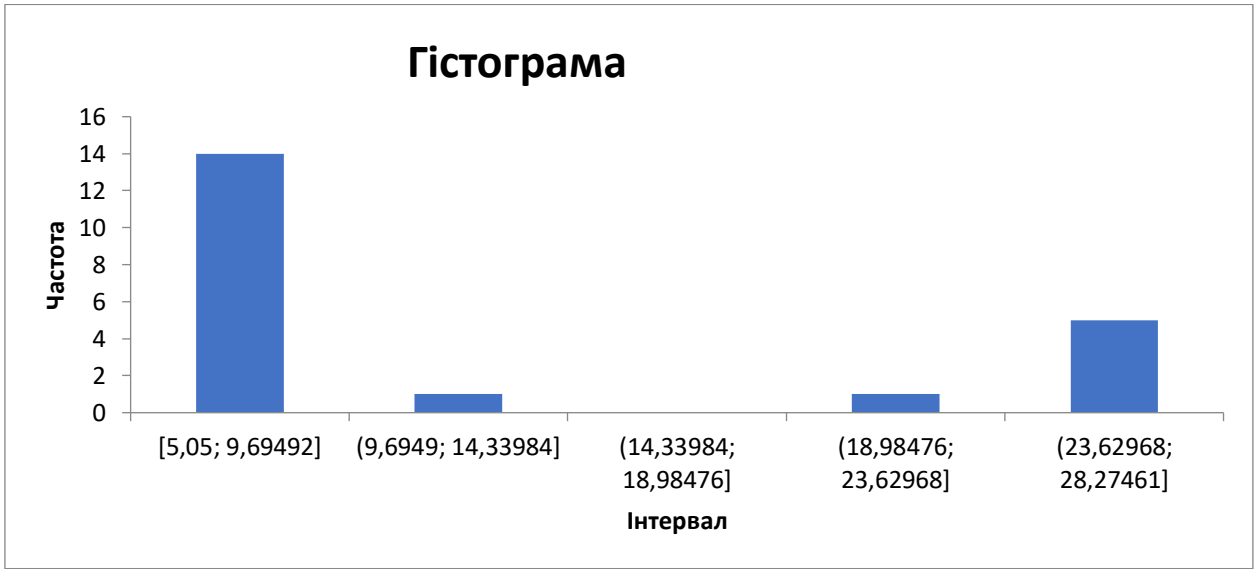
Табл. 6.1 Описова статистика курсу долара

З усіх даних виділяється дуже високий коефіцієнт асиметрії та велике середньоквадратичне відхилення, що каже про те що дані сильно розкидані, та сильно зміщенні в сторону мінімуму.

Таблиця та гістограма частот:

Інтервал	*Частота*
[5,05; 9,69492]	14
(9,6949; 14,33984]	1
(14,33984; 18,98476]	0
(18,98476; 23,62968]	1
(23,62968; 28,27461]	5

Табл. 6.2 Таблиця частот курсу долара

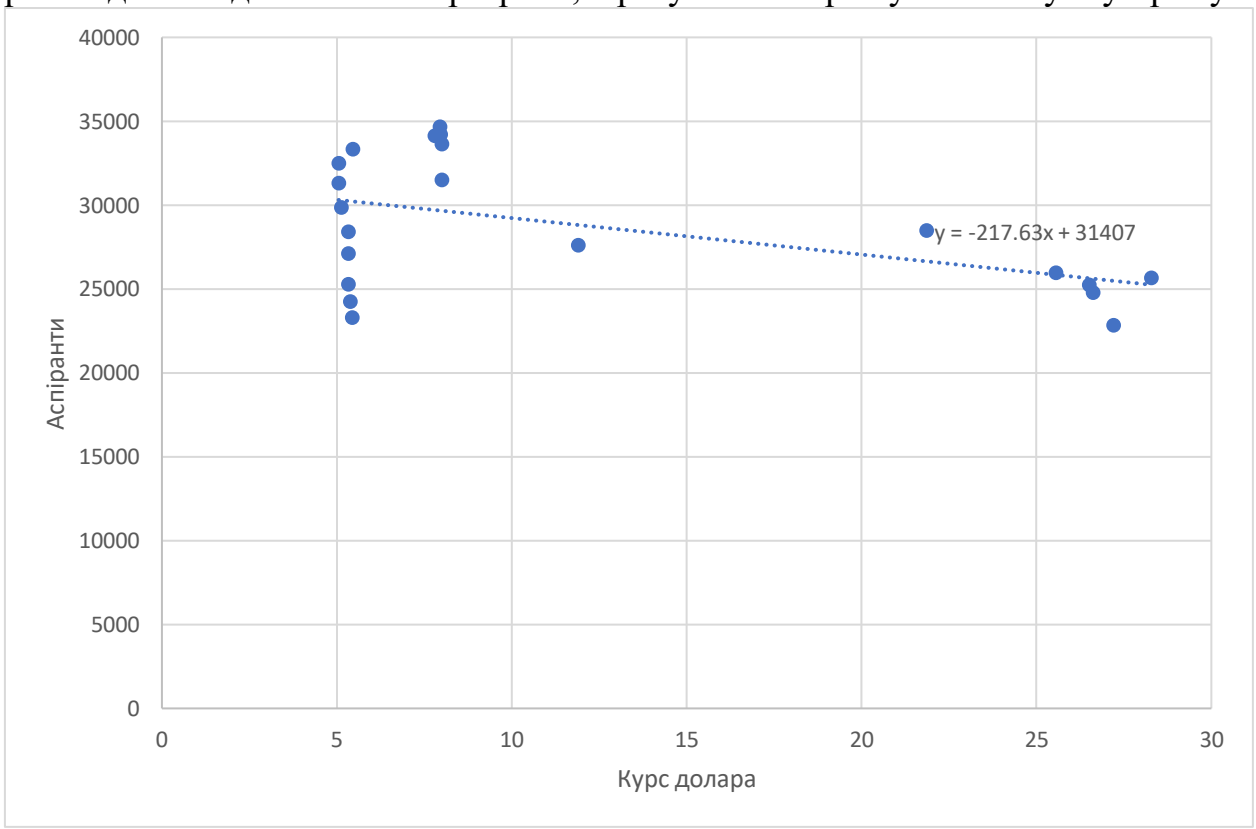


Мал. 6.2 Гістограма частот курсу долара

Схожості із нормальним розподілом не спостерігається, що загалом природньо для курсів валют, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

6.3 Побудова моделі парної регресії

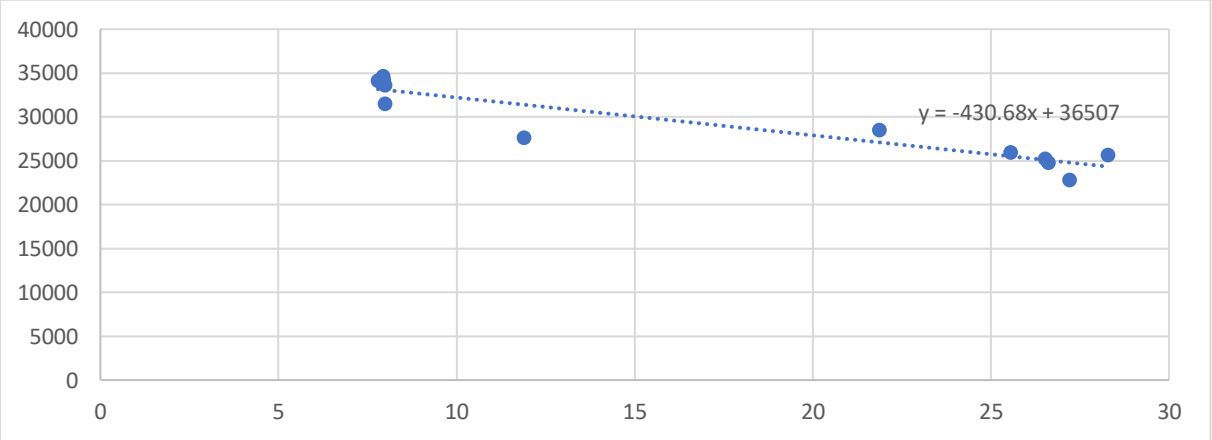
При обчисленні кореляції отримуємо значення -0,505115102, це свідчить про помірну обернену кореляцію між досліджуваними величинами. На графіку залежності не видно будь-якої явно вираженої залежності, тому ми будемо розглядати модель лінійної регресії, в результаті отримуємо наступну пряму:



Мал.6.2 Графік залежності

Як можна побачити, вона хоч і відображає якусь залежність між нашими величинами, але в значно меншій мірі ніж попередні випадки, і це підтверджують отримані нами числові характеристики регресії – p-критерій має значення 0,019512799, значить ми можемо відхилити нуль-гіпотезу, а R^2 має занчення 0,255141266, тобто лише 25% коливань кількості аспірантів пояснюються курсом долара. Жодні з наших припущень не порушені, тому модель є коректною. На практиці бачимо, що ми можемо намагатись за курсом долара передбачити кількість аспірантів, але це не дає надзвичайної точності – 9,53%.

На графіку залежності також видно, що наш графік містить багато точок, що мають дуже близькі значення курсу долара. Це може погіршувати нашу модель. Спробуємо виключити з розгляду такі точки, тобто розглядатимемо тільки точки із значенням курсу, більшим за 5. Це дійсно покращує модель, оскільки тепер значення коефіцієнту детермінації близьке до 0,86, а p-критерій також покращився, отже модель залишилася значущою. Таким чином, ми отримали кращу модель для меншої вибірки даних, що може бути корисним у прогнозуванні, оскільки дані, на основі яких побудована ця модель, є новішими(беруться починаючи з 2009 року), а значить більш актуальними. Отримана лінія регресії:



Мал.6.3 Графік залежності

ВИСНОВОК

У результаті попереднього регресійного аналіз ми вже отримали наступну регресійну модель: *Y = -217,6328841* X + 31406,75783(тут Y – це кількість аспірантів(осіб), а X – курс долара до гривні),* яка дозволяє не тільки сказати, що існує залежність між кількістю аспірантів і курсом долара, але і передбачати майбутні значення кількості аспірантів із достатньою точністю. Отже, ми можемо сказати, що чим вищий курс долара, тим менше буде аспірантів. Також, обмеживши вибірку, ми змогли побудувати більш значущу модель, що може виявитись корисним у інших практичних задачах.

Розділ 7. Курс євро

7.1 Загальний огляд

В цілому курс євро обирався з тих самих міркувань, що і курс долару, та з того самого джерела. Графік курсу по роках:

Мал.7.1 Графік курсу євро

6.2 Описова статистика

Основні статистичні характеристики курсу євро:

Описова статистика	
Медіана	10,5229
Максимум	34,7396
Мінімум	4,8152
Середнє значення	14,4221
Середньоквадратичне відхилення	10,49414
Асиметрія	0,922654

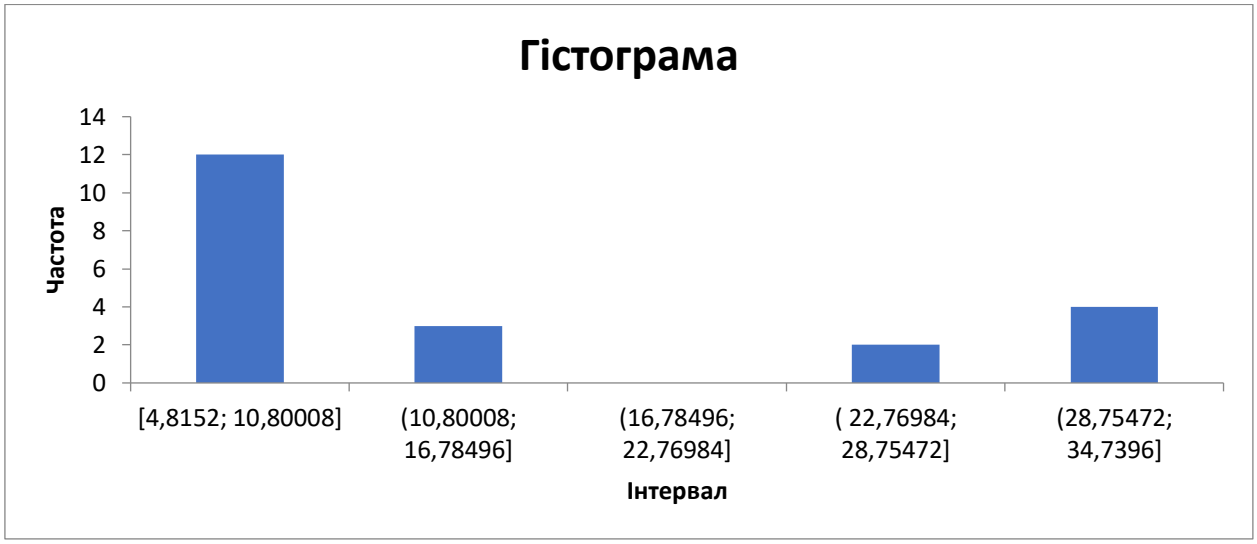
Табл. 7.1 Описова статистика курсу євро

Так само, як і у випадку долара, на основі цих даних, робимо висновок про сильне зміщення до мінімуму.

Таблиця та гістограма частот:

Інтервал	*Частота*
[4,8152; 10,80008]	12
(10,80008; 16,78496]	3
(16,78496; 22,76984]	0
(22,76984; 28,75472]	2
(28,75472; 34,7396]	4

Табл. 7.2 Таблиця частот курсу євро

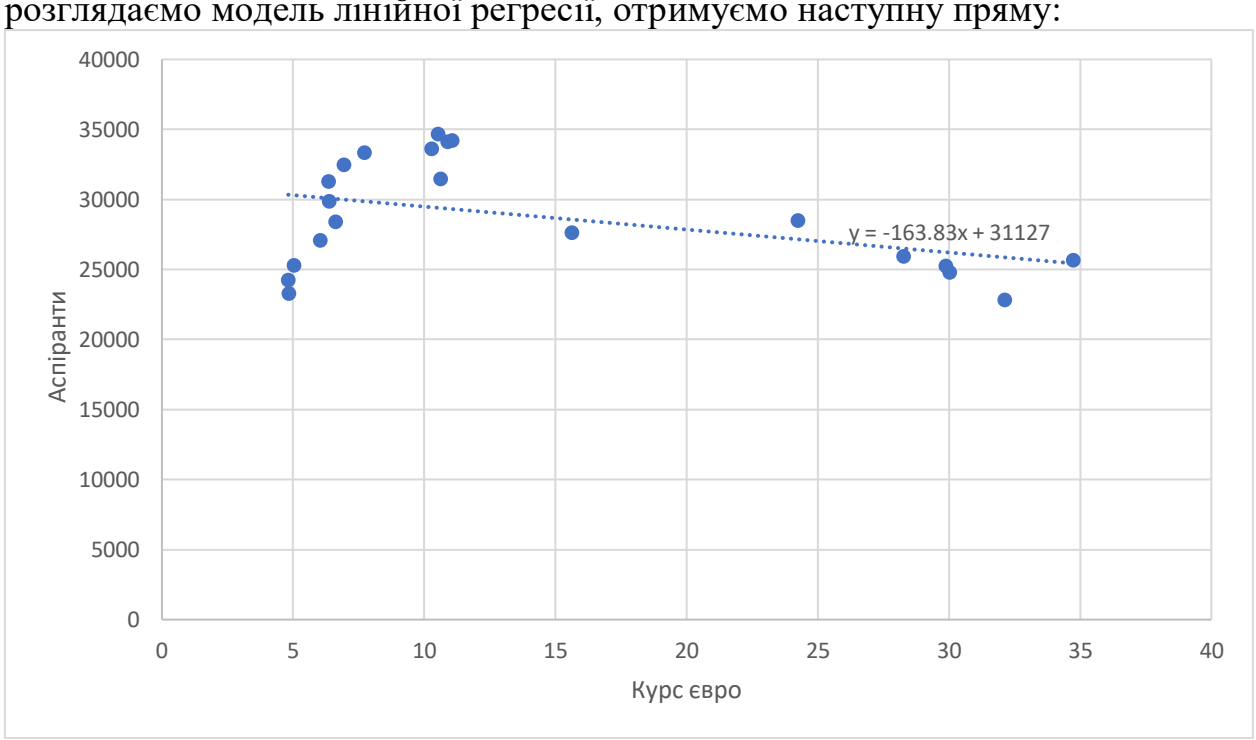


Мал. 7.2 Гістограма частот курсу євро

Схожості із нормальним розподілом не спостерігається, що загалом природньо для курсів валют, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

7.3 Побудова моделі парної регресії

При обчисленні кореляції отримуємо значення -0,4339789, це значення показує більш слабку обернену кореляцію, ніж у випадку долара, так само розглядаємо модель лінійної регресії, отримуємо наступну пряму:



Мал. 7.3

Для євро значення p-критерію ще більше, ніж у випадку долара, але все ж дозволяє відхилити нуль гіпотезу, оскільки дорівнює 0,049343482. Критерій детермінованості тут також менший, ніж у долара - 0,188337684. Припущення регресії не порушені, тому модель є коректною. Очікувано,

передбачення кількості аспірантів за курсом євро дає гірші результати, ніж за доларом, в середньому похибка складає 10,06%.

ВИСНОВОК

У результаті попереднього регресійного аналіз ми вже отримали наступну регресійну модель: *Y = -163,8314535* X + 31126,55473(тут Y – це кількість аспірантів(осіб), а X – курс євро до гривні)* , яка дозволяє не тільки сказати, що існує залежність між кількістю аспірантів і курсом євро, але і передбачати майбутні значення кількості аспірантів із достатньою точністю. Отже, ми можемо сказати, що чим вищий курс євро, тим менше буде аспірантів, але не настільки менше, як у випадку з доларом.

Розділ 8. Частка кількості промислових підприємств, що впроваджували інновації

8.1 Загальний огляд

З логічної точки зору, частка інноваційно активних підприємств має впливати на загальне бажання людей брати участь у наукових працях, оскільки зацікавленість компанії у дослідженні означає більші гроші у науковому секторі. Саме тому цікаво побачити, чи є якийсь зв'язок між кількістю людей, що ідуть до аспірантури, та кількістю інноваційно активних підприємств. Дані представлені у відсотках від загальної кількості підприємств, та були взяті з сайту ДССУ[1]. Графік даних по роках:

Мал. 8.1. Графік частки інноваційно активних підприємств

8.2 Описова статистика

Основні статистичні характеристики отриманих даних представлені у таблиці:

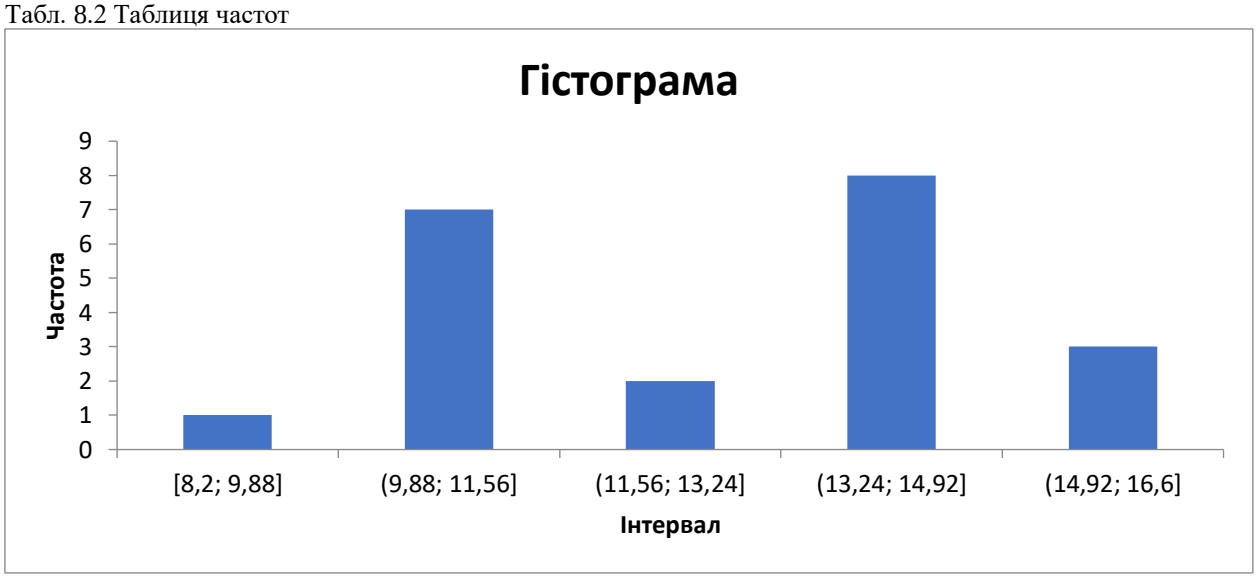
Описова статистика	
Медіана	13,6
Максимум	16,6
Мінімум	8,2
Середнє значення	12,87619
Середньоквадратичне відхилення	2,204066
Асиметрія	-0,33619

Табл. 8.1. Опис даних

Бачимо, що на відміну від більшості даних у цій роботі, частка інноваційно активних підприємств виділяється порівняно низьким середньоквадратичним відхиленням. Також можна відмітити слабкий зсув даних в сторону максимуму. Побудуємо гістограму частот для наших даних:

Інтервал	*Частота*
[8,2; 9,88]	1
(9,88; 11,56]	7
(11,56; 13,24]	2
(13,24; 14,92]	8
(14,92; 16,6]	3

Табл. 8.2 Таблиця частот



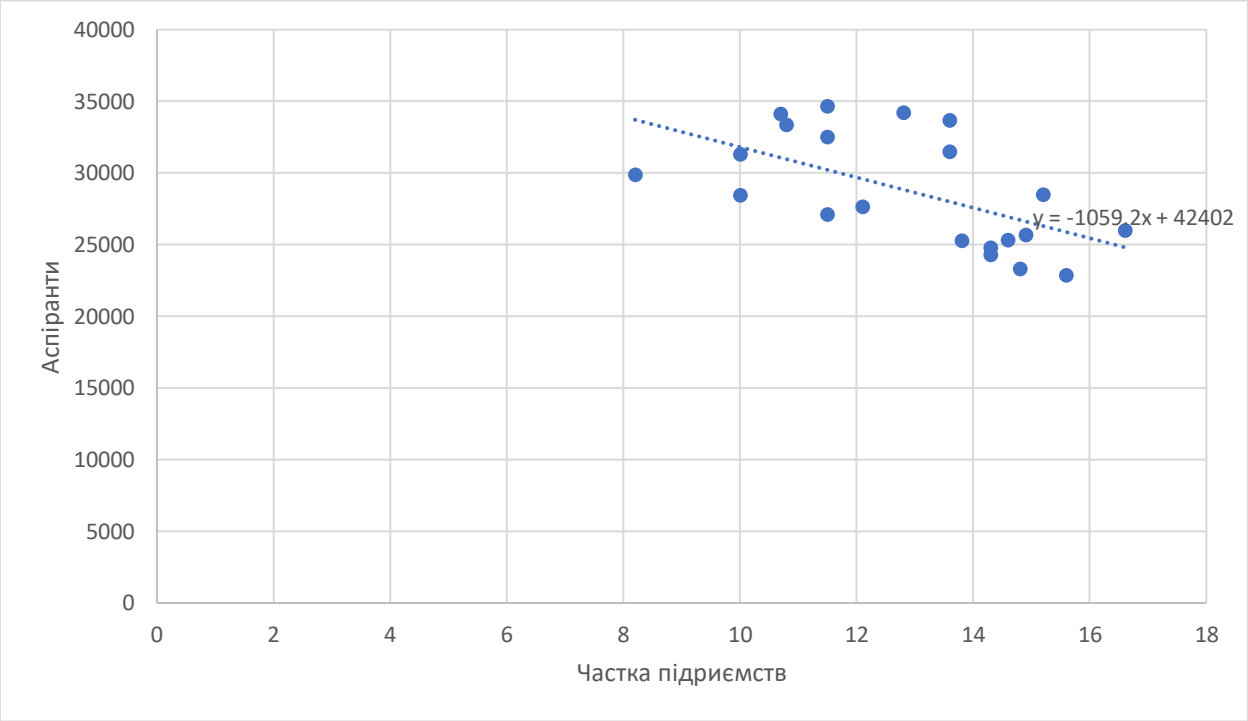
Мал. 8.2 Гістограма частот

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

8.3 Побудова моделі парної регресії

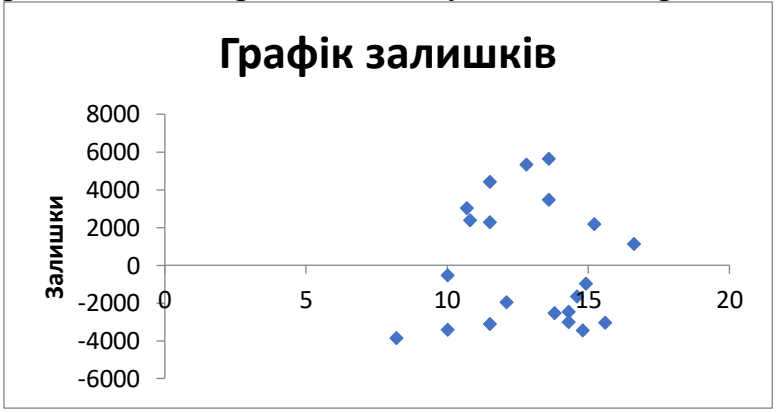
Значення кореляції між часткою інноваційно активних підприємств та кількістю аспірантів складає – 0,58 , це означає помірну зворотну кореляцію, яку можна побачити на графіку залежності, та в цілому проста лінійна регресія здається найефективнішою моделлю в даному випадку, тому в

результаті регресійного аналізу отримуємо наступну пряму:



Мал. 8.3

Р-критерій має значення 0,004937177, що дозволяє нам із впевненістю відхилити нуль гіпотезу. Але критерій детермінованості доволі низький - 0,347267605, тобто лише 34% коливань кількості аспірантів ми можемо пояснити за допомогою нашої моделі. Опущені змінні відсутні, а модель є найбільш підходящою для наших даних. За графіком залишків можемо побачити, що гетероскадичніть та автокореляція відсутні, а самі залишки розподілені нормально, тому модель є коректною:



Мал. 8.3

На практиці, побудрвана модель дозволяє передбачати значення кількості аспірантів із середньою точністю 9,87%, що не є найкращим результатом, але в цілому логічним і припустимим.

ВИСНОВОК

У результаті попереднього регресійного аналіз ми вже отримали наступну регресійну модель: *Y = -1059,213016* X + 42402,390453(тут Y – це кількість аспірантів(осіб), а X – частка інноваційних підприємств(%)),*

яка дозволяє не тільки сказати, що існує залежність між кількістю аспірантів і часткою інноваційно активних підприємств, але і передбачати майбутні значення кількості аспірантів із достатньою точністю. Отже, ми можемо сказати, що чим більша частка підприємств, що вводять інновації, тим менше буде аспірантів, що цілком пояснюється логікою, оскільки люди віддають більшу перевагу заробітку, який зростає із збільшенням інноваційно активних підприємств, аніж аспірантурі.

Розділ 9. Очікувана тривалість життя

9.1 Загальний огляд

Якщо люди більше живуть, то, як підказує логіка, тим більше часу вони можуть витратити на навчання. Але такі припущення слід перевірити, тому в нашу регресійну модель також включимо очікувану тривалість життя в роках, інформація взята із сайту ДССУ[1]. Графік отриманих даних по рокам:

Мал. 9.1 Графік тривалості життя

9.2 Описова статистика

Основні статистичні характеристики отриманих даних представлені у таблиці:

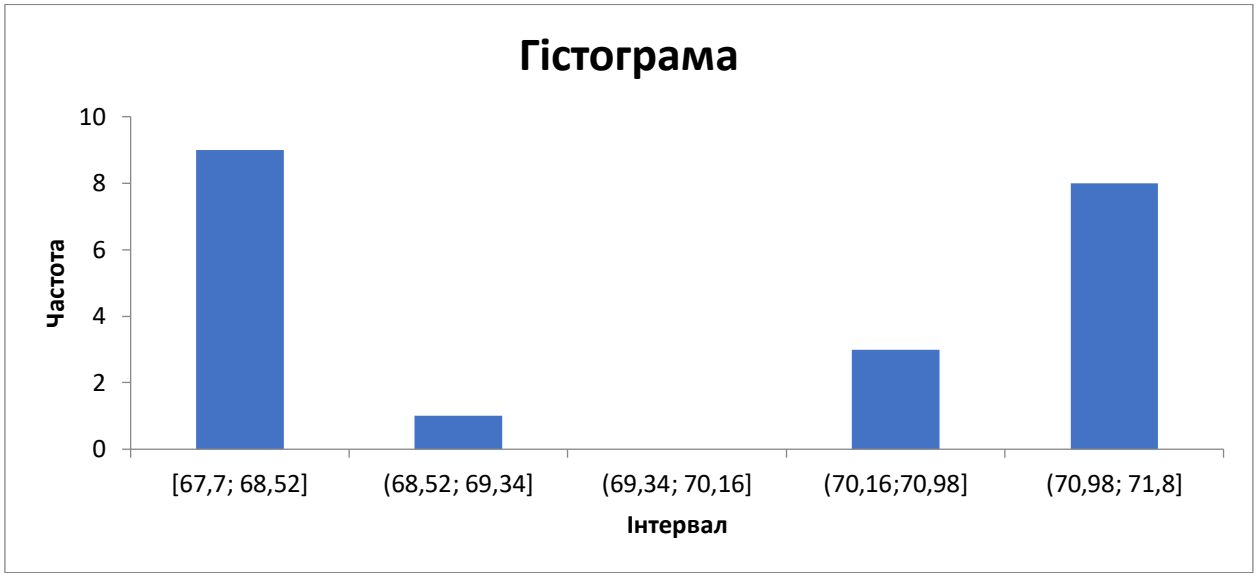
Описова статистика	
Медіана	70,3
Максимум	71,8
Мінімум	67,7
Середнє значення	69,78571429
Середньоквадратичне відхилення	1,604769676
Асиметрія	-0,065196049

Табл 9.1 Опис тривалості життя

Дані виділяються низьким середньоквадратичним відхиленням, та дуже низькою асиметрією. Побудуємо гістограму частот:

Інтервал	*Частота*
[67,7; 68,52]	9
(68,52; 69,34]	1
(69,34; 70,16]	0
(70,16;70,98]	3
(70,98; 71,8]	8

Табл. 9.2 Таблиця частот тривалості життя



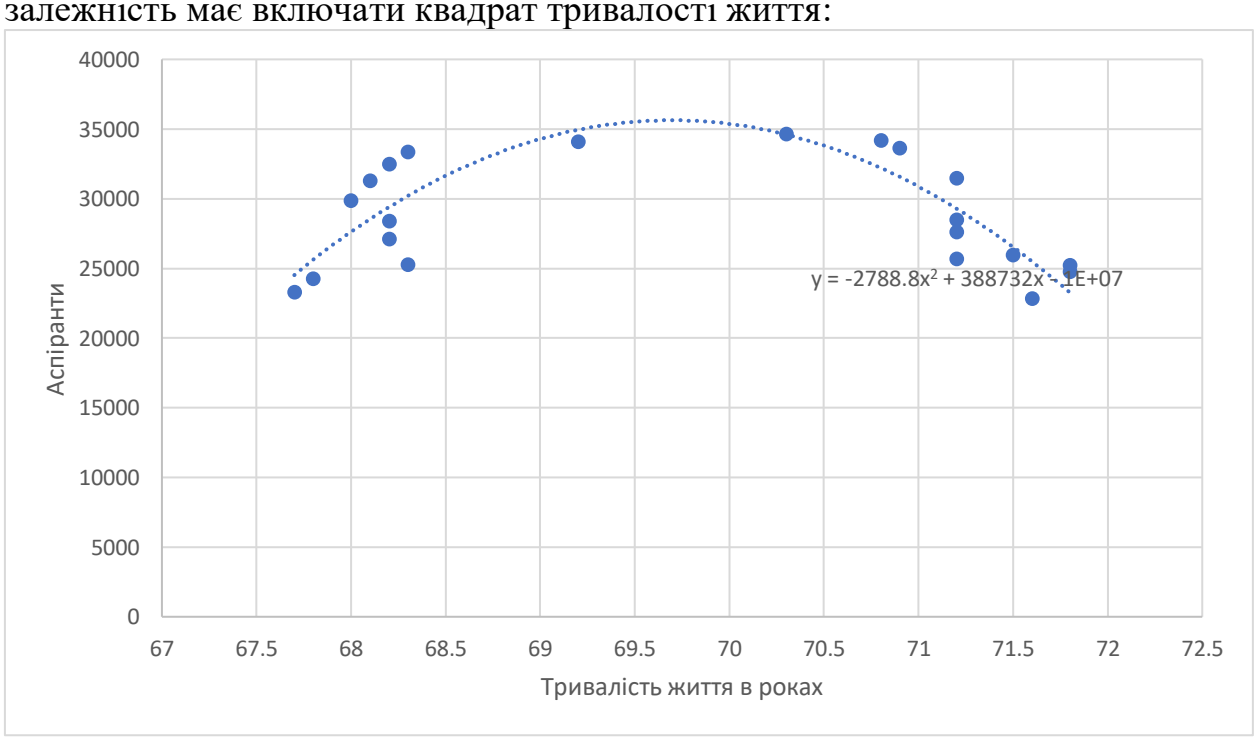
Мал. 9.2 Гістограма частот тривалості життя

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

9.3 Побудова моделі парної регресії

Отримуємо надзвичайно мале значення кореляції, можна сказати, що вона відсутня.

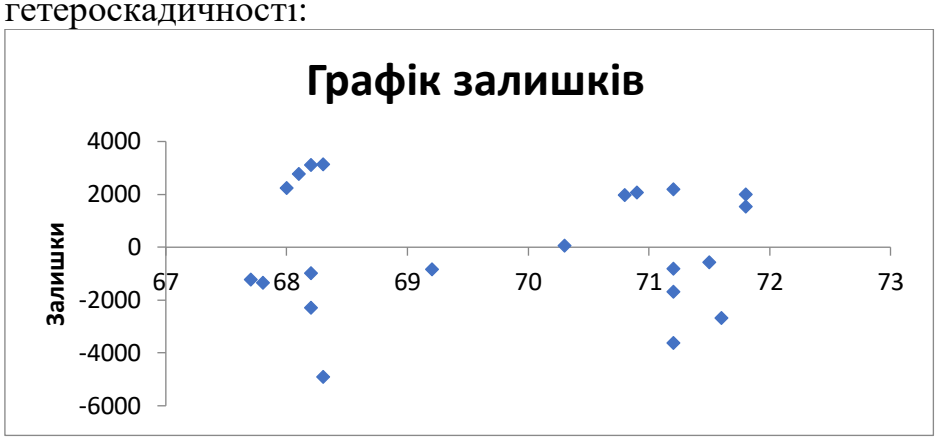
Перейдемо до побудови моделі регресії, за графіком залежності очевидно, що залежність має включати квадрат тривалості життя:



Мал. 9.3

В результаті проведеного регресійного аналізу отримуємо модель із дуже малим(порядку 10^{-5}) значенням p-критерію та коефіцієнтом детермінованості 0,649796245, що є непоганим результатом. Модель є

найбільш підходящою із усіх можливих, опущені змінні відсутні, за графіком залежності очевидна нормальність, відсутність автокореляції та гетероскадичності:



Мал. 9.4

Отже, побудована модель є коректною, а усі її характеристики надійними. Перевіряючи на практиці точність передбачення кількості аспірантів за побудованою моделлю отримаємо середню похибку 7%, що є задовільним значенням.

ВИСНОВОК

У результаті попереднього регресійного аналіз ми вже отримали наступну регресійну модель: $Y = -2788{,}8 * X^2 + 388732 * X - 13510614{,}8008685$ *(тут Y – це кількість аспірантів(осіб), а X – тривалість життя(у роках)),* яка дозволяє не тільки сказати, що існує залежність між кількістю аспірантів і тривалістю життя але і передбачати майбутні значення кількості аспірантів із достатньою точністю.

Розділ 10. Кількість працюючих людей віком 15-70 років

10.1 Загальний огляд

Логічно припустити, що чим більше людей працює, тим у меншої кількості буде час на аспірантуру, але це припущення треба перевірити методами математичної статистики, тому ми візьмемо дані про кількість працюючих людей(у тисячах) з сайту ДССУ[1]. Графік наших даних по рокам:

Мал. 10.1 Графік кількості працюючих людей

10.2 Описова статистика

Основні статистичні характеристики отриманих даних представлені у таблиці:

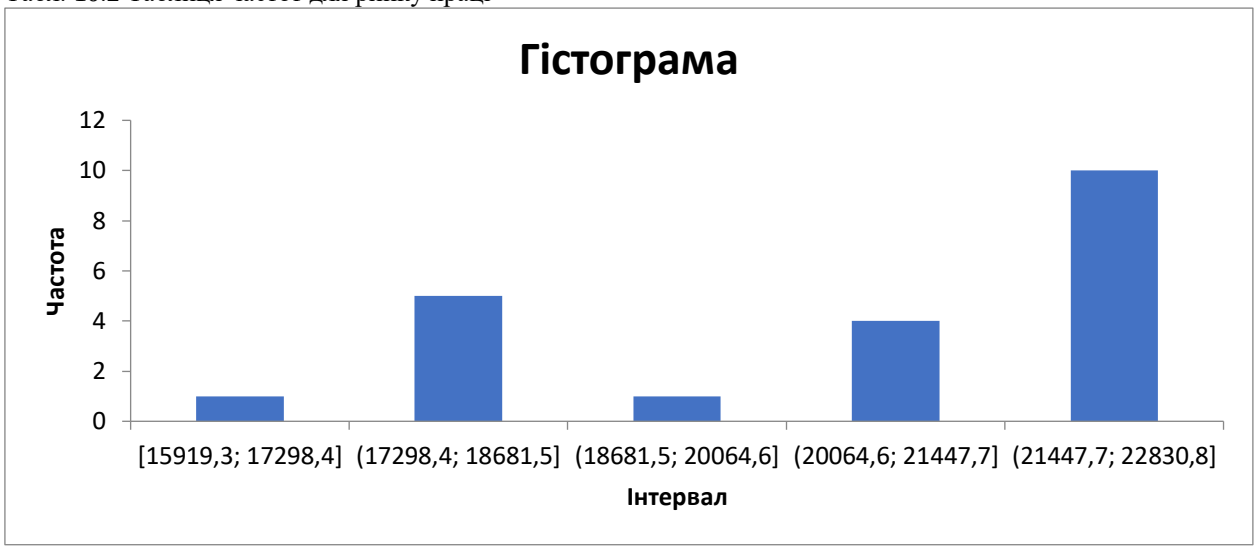
Описова статистика	
Медіана	20894,1
Максимум	22830,8
Мінімум	15915,3
Середнє значення	20591,72
Середньоквадратичне відхилення	2085,587
Асиметрія	-0,83939

Табл. 10.1 Описова статистика ринку праці

Дані мають помірне середньоквадратичне відхилення, та сильно зміщенні в сторону максимуму, що можна побачити на гістограмі частот:

Інтервал	*Частота*
[15919,3; 17298,4]	1
(17298,4; 18681,5]	5
(18681,5; 20064,6]	1
(20064,6; 21447,7]	4
(21447,7; 22830,8]	10

Табл. 10.2 Таблиця частот для ринку праці

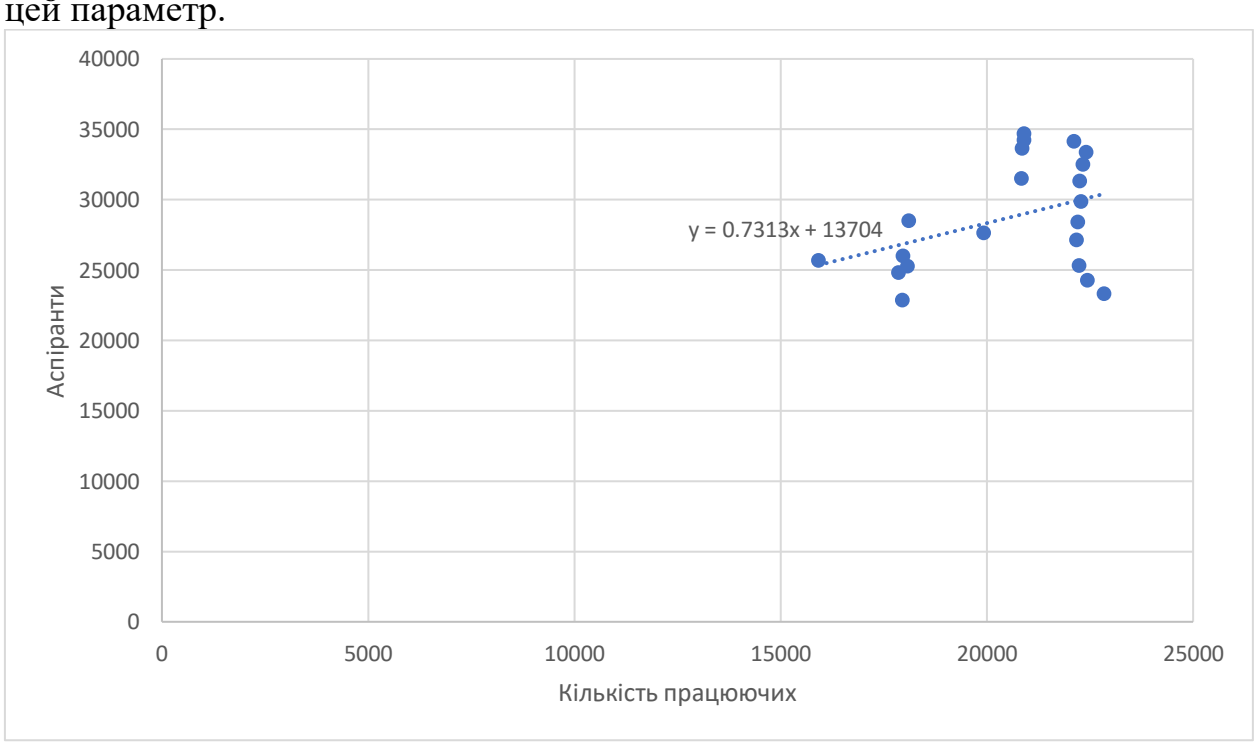


Мал. 10.2 Гістограма частот

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

10.3 Побудова моделі парної регресії

Отримуємо низьке значення кореляції - 0,385008551. За графіком залежності не видно інших варіантів побудови залежності, тому будуємо просту лінійну. Жодні з наших припущень не порушуються, отже модель є коректною. Отримуємо значення p-критерію 0,08, та, оскільки іншої функціональної залежності не видно, ми приймаємо нуль гіпотезу та виключаємо із розгляду цей параметр.



Мал. 10.3 Графік залежності та лінія тренду

ВИСНОВОК

У результаті попереднього регресійного аналіз ми встановили відсутність лінійної залежності між ринком праці та кількістю аспірантів, що дозоляє нам виключити з розгляду у моделі множинної регресії цей параметр, Таким чином, попередній аналіз дозволив нам спростити задачу.

Розділ 11. Викиди забруднюючих речовин та діоксиду вуглецю в атмосферне повітря

11.1 Загальний огляд

Загалом нема очевидних причин для того, щоб кількість аспірантів якось корелювала із забрудненням повітря, але це також означає, що ніхто не досліджував її наявність. Тому ми перевіримо наявність кореляції, на випадок, якщо вона існує. Викиди вимірюються у тисячах тон, дані взяті з сайту ДССУ[1]. Графік по рокам:

Мал. 11.1 Графік забрудненості повітря

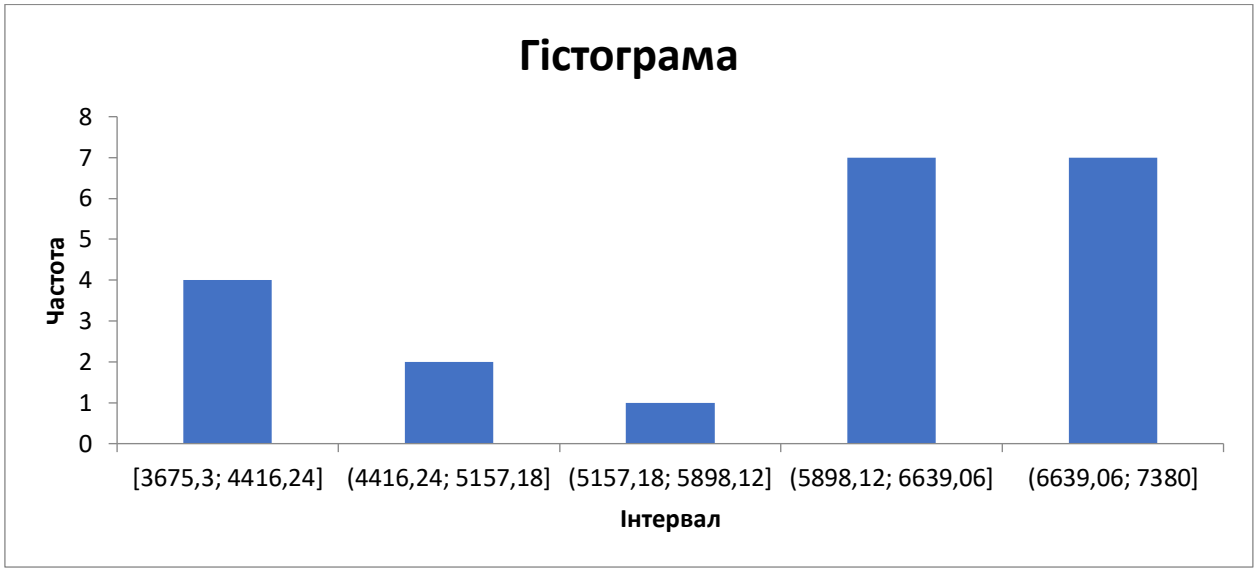
11.2 Описова статистика

Основні статистичні характеристики отриманих даних представлені у таблиці:

Описова статистика	
Медіана	6191,3
Максимум	7380
Мінімум	3675,3
Середнє значення	5859,014286
Середньоквадратичне відхилення	1165,139445
Асиметрія	-0,628135355

Табл. 11.1 Описова статистика забруднення повітря

У даних бачимо велике середньоквадратичне відхилення, та помірний зсув даних до максимального значення. Побудуємо гістограму частот:

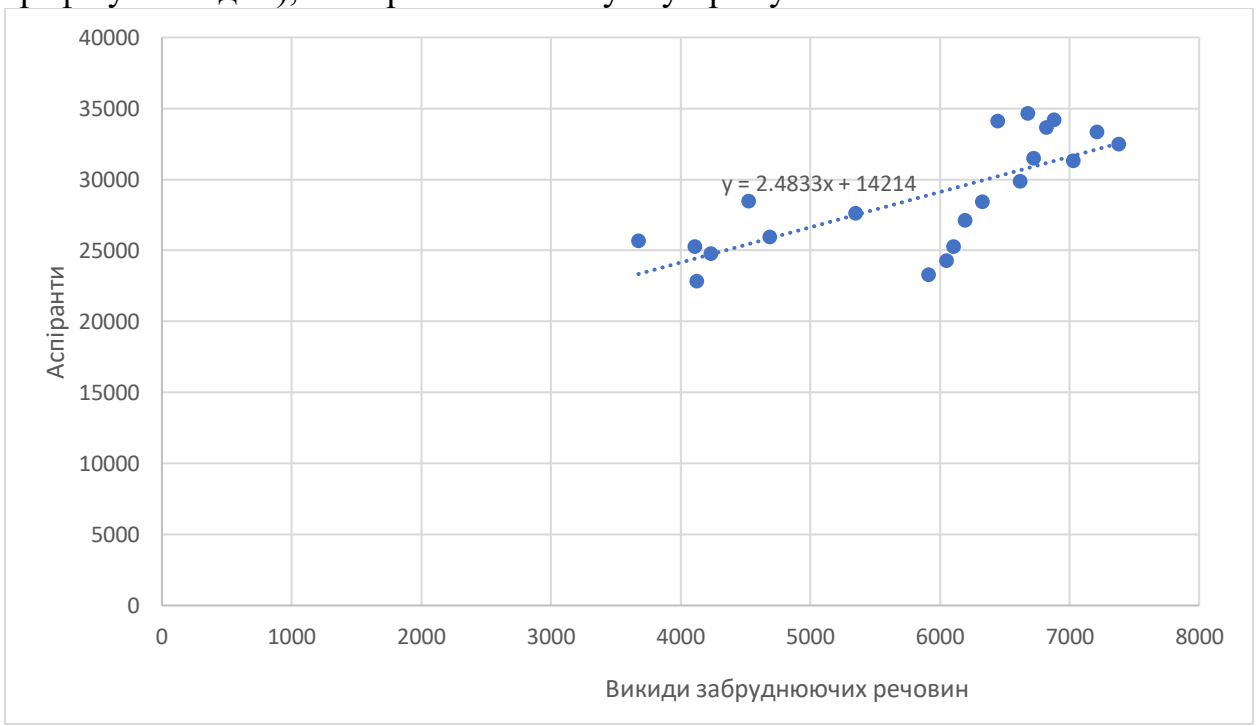


Мал. 11.2 Гістограма частот

Схожості із нормальним розподілом не спостерігається, тому треба бути обережним при інших дослідженнях цих даних, але для класичної регресії це не важливо.

11.3 Побудова моделі парної регресії

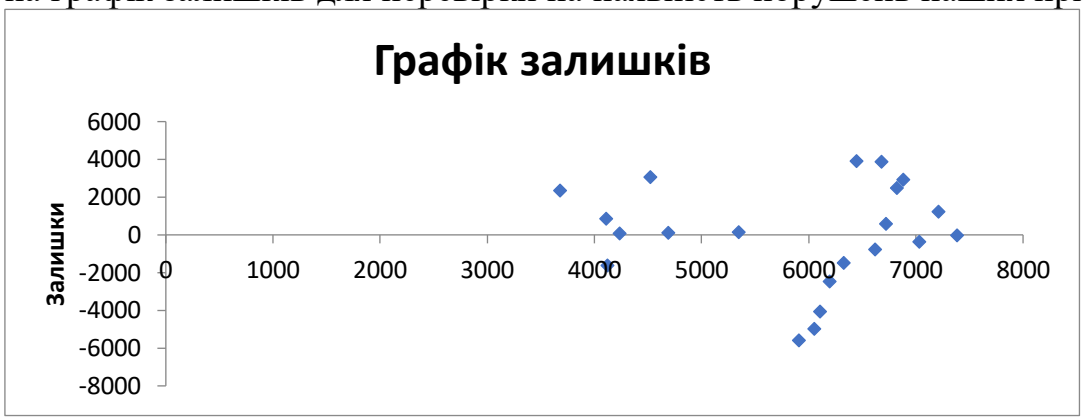
Рахуємо кореляцію, маємо значення 0,730362808, що означає сильну кореляцію. Якщо побудуємо модель лінійної регресії(оскільки іншої на графіку не видно), то отримаємо наступну пряму:



Мал. 11.3 Графік залежності та лінія тренду

Значення p-критерію - 0,000170368, що дозволяє нам назвати модель значущою, а коефіцієнт детермінованості рівний 0,533429831, каже що більше половини змін кількості студентів можна пояснити побудованою моделлю. Але за логікою ці дані не мають бути пов’язані, тому подивимось

на графік залишків для перевірки на наявність порушень наших припущень:



Мал. 11.4 Графік залишків

Бачимо, що в цілому ці залишки можна вважати нормально розподіленими з постійною дисперсією. Припущення лінійності також виконується, бо іншу модель підібрати неможливо. Передбачення модель робить із непоганою точністю – в середньому 7%. Отже, такий результат може свідчити про наявність опущеної змінної, яка впливає як на кількість аспірантів, так і на викиди в повітря забруднюючих речовин. Якщо це так, можливо ми зможемо побачити якісь підтвердження при побудові множинної регресії.

ВИСНОВОК

У результаті попереднього регресійного аналіз ми встановили наявність лінійної залежності між забрудненням повітря та кількістю аспірантів, що має вигляд *Y = X *2,4833+14214 (тут Y – це кількість аспірантів(осіб), а X – забруднення повітря(тис. тон)),*. Але це не є логічним результатом, тому можлива наявність опущеної змінної.

Розділ 12. Множинна лінійна регресія

12.1 Загальний огляд

Тепер, коли ми маємо певну інформацію про дані, а також зробили регресійний аналіз залежності кількості аспірантів від кожного з параметрів, ми можемо перейти до побудови множинної лінійної регресії, почнемо з побудови моделі, у якій кожен параметр буде входити як лінійний доданок

12.2 Побудова параметрів

Оскільки при побудові парної регресії для деяких параметрів ми виявили квадратичну залежність, то для того щоб побудувати в якій кожен доданок буде лінійним, нам треба видозмінити дані:

1.При досліджені залежності від населення отримали *Y= -0,0007*T1² + 63,764*T1 - 1E+06(тут Y – це кількість аспірантів(осіб), а T1 – населення(тис. осіб)),* виділимо повний квадрат, тоді отримуємо наше співвідношення у наступному вигляді:

-0,00069738230540663 (T1- 45562.75068032)² + 32456.06362805*

Таким чином, беремо перший параметр X1 =*(T1- 45562.75068032)²*

2. Виділяючи повний квадрат для залежності *Y = 4,0673 * T2² -2456,9 * T2 + 395173(тут Y – це кількість аспірантів(осіб), а T2 – кількість університетів(штук)),* отримуємо:

*4.0673 * (T2 - 302.03083126) ^ 2 + 24143.22533375*

Таким чином, беремо другий параметр X2 =*(T2- 302.03083126)²*

3. X3 = Курс долара

4. X4 = Курс євро

5. X5 = Частка підприємств, що впровадила інновації

6. Виділяючи повний квадрат із *Y = - 2788,8* T3² + 388732* T3 - 13510614,8008685(тут Y – це кількість аспірантів(осіб), а T3 – тривалість життя(у роках)),* отримуємо:

*- 2788.8 * (T3 - 69.6952094)²+ 35764.27113379*

Таким чином, беремо параметр X6 =*(T3 - 69.6952094)²*

7. X7 = Викиди у повітря забруднюючих речовин.

12.3 Аналіз мультиколінеарності

Ми використаємо коефіцієнт інфляції дисперсії як критерій перевірки на мультиколінеарність.. За допомогою пакету «Аналіз даних» для програми Excel будуємо кореляційну матрицю:

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*Y*
X1	1							
X2	-0,80634	1						
X3	0,521241	-0,55398	1					
X4	0,450984	-0,48548	0,993375	1				
X5	0,625375	-0,74897	0,644997	0,600079	1			
X6	0,717797	-0,7238	0,470071	0,399272	0,412272	1		
X7	-0,71211	0,756108	-0,92734	-0,90156	-0,66458	-0,55772	1	

Y	-0,89117	0,935912	-0,50512	-0,43398	-0,58929	-0,8061	0,730363	1

Табл. 12.1 Кореляційна матриця

Тепер підрахуємо VIF, будуючи для кожного параметру регресійну модель, маємо:

VIF X1 = 3.9016390335009379
VIF X2 = 12.1628483795405423
VIF X3 = 136.628348792496188165
VIF X4 = 98.887247120668252917
VIF X5 = 5.111094525035029596
VIF X6 = 4.050752114968136798
VIF X7 = 36.2579539779840268

Значення VIF більше за 10 мають параметри X2, X3, X4, X7.

У випадку X3 та X4 це легко пояснюється тим, що євро і долар пов'язані між собою економічно, і тому значення курсу однієї валюти можна пояснити значенням іншої. В даному випадку парна регресія з доларом мала кращі характеристики, тому мною прийнято рішення видалити змінну X4 з розгляду. Для X7 значення VIF = 36,3 може значити, що дійсно існує змінна, від якої залежить і викиди в атмосферу, і кількість аспірантів. Дійсно, за кореляційною матрицею можна побачити сильну кореляцію X7 та X3. На цей раз керуючись логічними міркуваннями, що курс долара це більш глобальний параметр, я видаляю з розгляду X7. Хоча X2 і є проблемною, це величина, утворена із кількості університетів, тому вона має давати гарний вклад у модель, тим паче значення VIF в неї на порядок менші, ніж у інших проблемних величин. З таких міркувань я трохи підніму поріг проблемності, і таким чином залишу X2 в моделі. Видаливши змінні, та знову порахувавши VIF для колишніх проблемних параметрів, отримуємо:

VIF X2 = 2.820377492072075296
VIF X3 = 1.286111670519916048

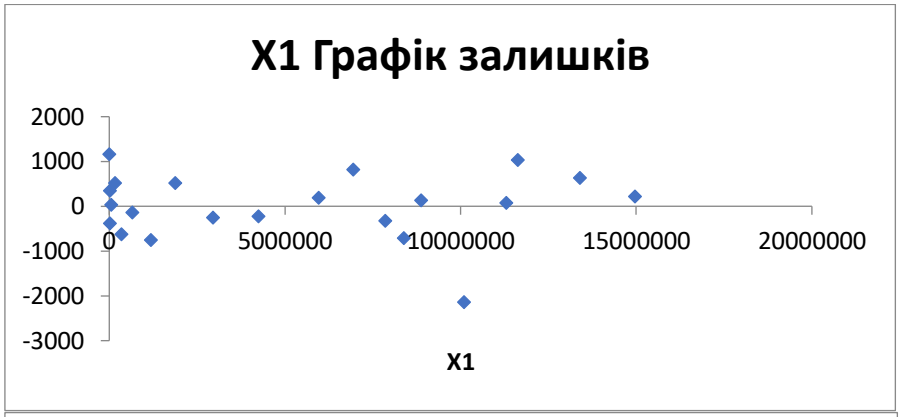
Це означає, що можна вважати, що мультиколінеарність серед обраних параметрів відсутня, тому можна перейти до побудови моделі множинної лінійної регресії.

12.4 Побудова моделі

Користуючись надбудовою «Аналіз даних» для MS Excel, будуємо модель множинної регресії для обраних параметрів:

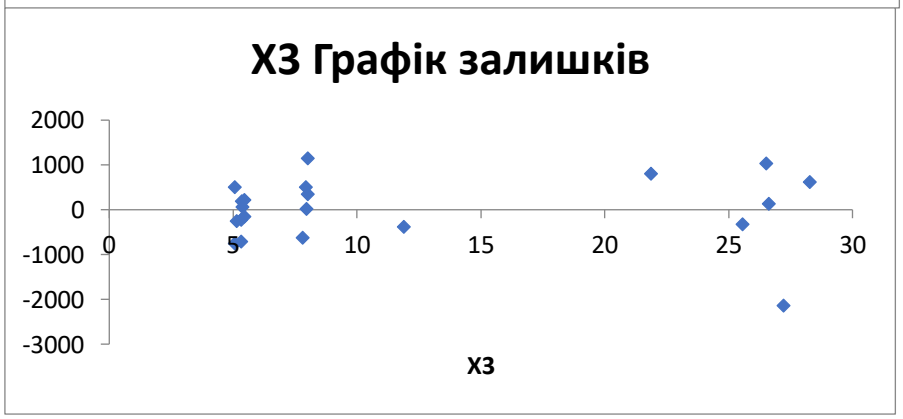
$$Y = -0{,}000286893 * X1 + 3{,}266600306 * X2 - 6{,}628394983 * X3 + 452{,}5187257 * X5 - 333{,}7296165 * X6$$

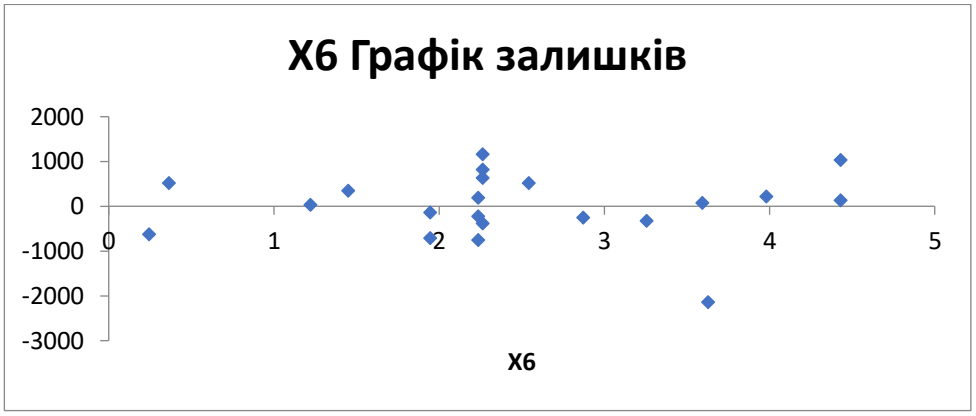
Ця модель має непоганими характеристиками – p-критерій порядку 10^{-10}, а коефіцієнт детермінованості рівний 0,966126647. За графіками залишків, наведеними нижче, можна побачити, що усі припущення виконуються:


X1 Графік залишків
2000
1000
0
-1000
-2000
-3000
0
5000000
10000000
15000000
20000000
X1


X2 Графік залишків
2000
1000
0
-1000
-2000
-3000
0
500
1000
1500
2000
2500
3000
X2


X5 Графік залишків
2000
1000
0
-1000
-2000
-3000
0
5
10
15
20
X5


X3 Графік залишків
2000
1000
0
-1000
-2000
-3000
0
5
10
15
20
25
30
X3



Але отриману модель можна покращити.

12.5 Покращення моделі

Зробити більш змістовною нашу модель можливо шляхом видалення незначущих змінних, тобто таких, для яких p-критерій став більшим за наш рівень надійності(в нашому випадку це найпоширеніший варіант - 0,05). Це змінні X3 та X6, для них в даній моделі ми не можемо відхилити нуль гіпотезу. Видалимо ці змінні, та побудуємо нову, покращену модель. Отримана модель має p-критерій порядку 10^{-12}, але $R^2 = 0{,}962118149$, що є гіршим результатом, ніж в попередній моделі. Як було сказано в теоретичній частині, R^2 має тенденцію збільшуватись при додаванні змінних, навіть якщо вони не є значущими, тому в даному випадку ми можемо скористатись модифікованим коефіцієнтом $AdjR^2$, який позбавлений цієї властивості. Дійсно, модифікований коефіцієнт покращився, але не набагато – лише на 0.000597587. Це свідчить про незначне, але все ж покращення нашої моделі. Як і в попередньому випадку, усі припущення регресії виконуються та модель є коректною.

Висновок

В результаті проведеного регресійного аналізу ми отримали значущу і надійну модель, за допомогою якої можна пояснити 96% коливань кількості студентів, використовуючи лише 3 параметри. Модель має наступний вигляд:

Аспіранти= 20093,38602 -
*0,000319593*X1*+3,534298912*X2+492,9434463**X5, де*

X1 =*(T1- 45562.75068032)²* *, де T1 – населення України(тис. осіб)*

X2 =*(T2- 302.03083126)²* *, деT2 – кількість університетів(штук)*

X5 = Частка підприємств, що впровадила інновації

Висновки

Було проведено регресійний аналіз залежності кількості аспірантів від обраних параметрів, та в результаті побудована модель множинної лінійної регресії, за допомогою якої можна як робити передбачення майбутньої кількості аспірантів(причому в отриманій нами моделі усі параметри є числами, які значно легше передбачити, ніж кількість аспірантів), так і зробити певні висновки про залежність цієї величини від параметрів. Найочевиднішим висновком є те, що із збільшенням частки інноваційно активних підприємств зростає і кількість аспірантів, причому на один відсоток зростання частки підприємств приходиться майже 493 аспіранта. Цей висновок має спонукати підприємців вводити інновації на своїх підприємствах, оскільки таким чином вони будуть покращувати стан вищої освіти в Україні. Залежність між аспірантам та населенням, як і з університетами, більш складна, але тим не менш значуща і здатна давати гарні прогнози із середньою похибкою в 2,05%.

Список використаних джерел

1. https://www.ukrstat.gov.ua/
2. https://bank.gov.ua/
3. https://data.worldbank.org/
4. http://applstat.univ.kiev.ua/ukr/docs/materials/matstat.pdf
5. Ивченко Г. И. Математическая статистика / Г. И. Ивченко, Ю. И. Медведев. – М.: Высшая школа, 1984.
6. https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf