Елементи статистичного аналізу взаємозв'язків

П Л А Н

Вступ

1. *Загальне поняття про лінійну регресію*
2. *Оцінка параметрів лінійної регресії*
3. *Коефіцієнти кореляції та детермінації*
4. *Перевірка моделі на адекватність*

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

Руденко В.М. Математична статистика. - К.: Центр учбової літератури, 2017. - 304 с.

Лозовий Б.Н., Пушак Я.С. Теорія ймовірностей і елементи математичної статистики. Навч. посібник. – К.: Ліра-К, 2018. – 276с.

Копич І.М., Сороківський В.М., Теорія ймовірностей та математична статистика. Навч. посібник. – К.: Ліра-К, 2018 – 382с.

Вступ

Одним із найважливіших завдань статистики є вивчення взаємозв'язків та залежностей в соціально-економічних явищах та процесах. Для такого вивчення застосовують методи кореляційного та регресійного аналізу.

Кореляційний і регресійний аналіз

– *це суміжні розділи математичної статистики, які призначені для вивчення статистичної залежності між випадковими величинами*.

Що таке статистична залежність між випадковими величинами?

Між випадковими величинами не завжди існує функціональна залежність.Часто зі зміною однієї величини змінюється тільки закон розподілу іншої величини (тобто, із зміною однієї з них змінюється середнє значення іншої). Такий зв'язок називається ***статистичним****.*

Завдання кореляційного аналізу:

вимір щільності зв'язку між ознаками, оцінка факторів, що мають найбільший вплив на результативну ознаку

Завдання регесійного аналізу:

встановлення форми залежності між ознаками, визначення функції регресії та використання її для оцінки невідомих значень результативної ознаки

Типові економіко-математичні моделі, які розробляє та вивчає кореляційно-регресійний аналіз

- *виробничі функції*
- *функції попиту різних груп споживачів*
- *функції переваги споживачів*
- *міжгалузеві моделі виробництва*
- *моделі розподілу і споживання продукції*
- *моделі загальної економічної рівноваги*

Етапи кореляційно-регресійного аналізу

- *Формулювання теорії чи гіпотези*
- *Розробка регресійної моделі для перевірки цієї гіпотези*
- *Оцінка параметрів обраної моделі*
- *Перевірка моделі на адекватність та висновки*
- *Прогнозування на основі запропонованої моделі*
- *Застосування запропонованої моделі*

1. Загальне поняття про лінійну регресію

- *аналіз зв’язку між змінними - використовується термін **«регресія»***

- *вимір тісноти зв’язку між змінними - термін **«кореляція»***

- *Функція, що відображає статистичний зв'язок між змінними, називається рівнянням регресії*

- *Якщо таке рівняння зв'язує лише дві змінні, це рівняння парної (простої) регресії*

$$Y = f(x)$$

- *Якщо воно відображає залежність результативної змінної від двох або більше незалежних змінних - це рівняння множинної регресії.*

$$Y = f(x_1, x_2, \ldots, x_n)$$

- *Прості лінійні регресійні моделі встановлюють лінійну залежність між двома змінними (парна регресія)*

Наприклад

- *між ціною на товар та попитом на нього*
- *витратами на рекламу та обсягом продукції, що випускається*
- *витратами на споживання та валовим національним продуктом (ВНП).*

- *Лінійна регресія є найпоширенішим і простим видом залежності між економічними змінними і часто служить початковою (відправною) точкою економічного аналізу*

- *При цьому одна із змінних вважається залежною змінною (результативною ознакою y) та розглядається як функція від незалежної змінної (факторної ознаки x).*

проста регресійна модель має вигляд

$$y = \alpha_0 + \alpha_1 x + e \qquad (1)$$

де y ***— вектор спостережень за залежною змінною*** $y = \{ y_1, y_2 \dots y_n \}$

x — вектор спостережень за незалежною змінною $x = \{ x_1, x_2 \dots x_n \}$

α_0, α_1 ***— теоретичні параметри регресійної моделі***

e — ***вектор помилок*** $e = \{ e_1, e_2, \ldots e_n \}$

є випадковою величиною, що характеризує відхилення *y* ***від теоретичної регресії і вказує на стохастичну суть регресійної залежності***

Помилки ще називають відхиленнями або залишками

В рівнянні (1) складова $\alpha_0 + \alpha_1 x$ ***називається системною складовою регресії (вона обумовлена лінійною залежністю між змінними).***

$$e = \{ e_1, e_2, \ldots e_n \}$$

називається випадковою складовою регресії (обумовлена впливом випадкових факторів)

Таким чином

рівняння (1) є лінійною регресійною моделлю, яку ще можна трактувати і як пряму на площині, де α_0 ***— перетин з віссю ординат,*** α_1 ***— нахил до осі абсцис, (якщо не враховувати випадкову величиу*** e ***).***

Як знайти явний вид регресійної залежності?

Треба знайти (оцінити) невідомі параметри цієї моделі α_0 ***та*** α_1

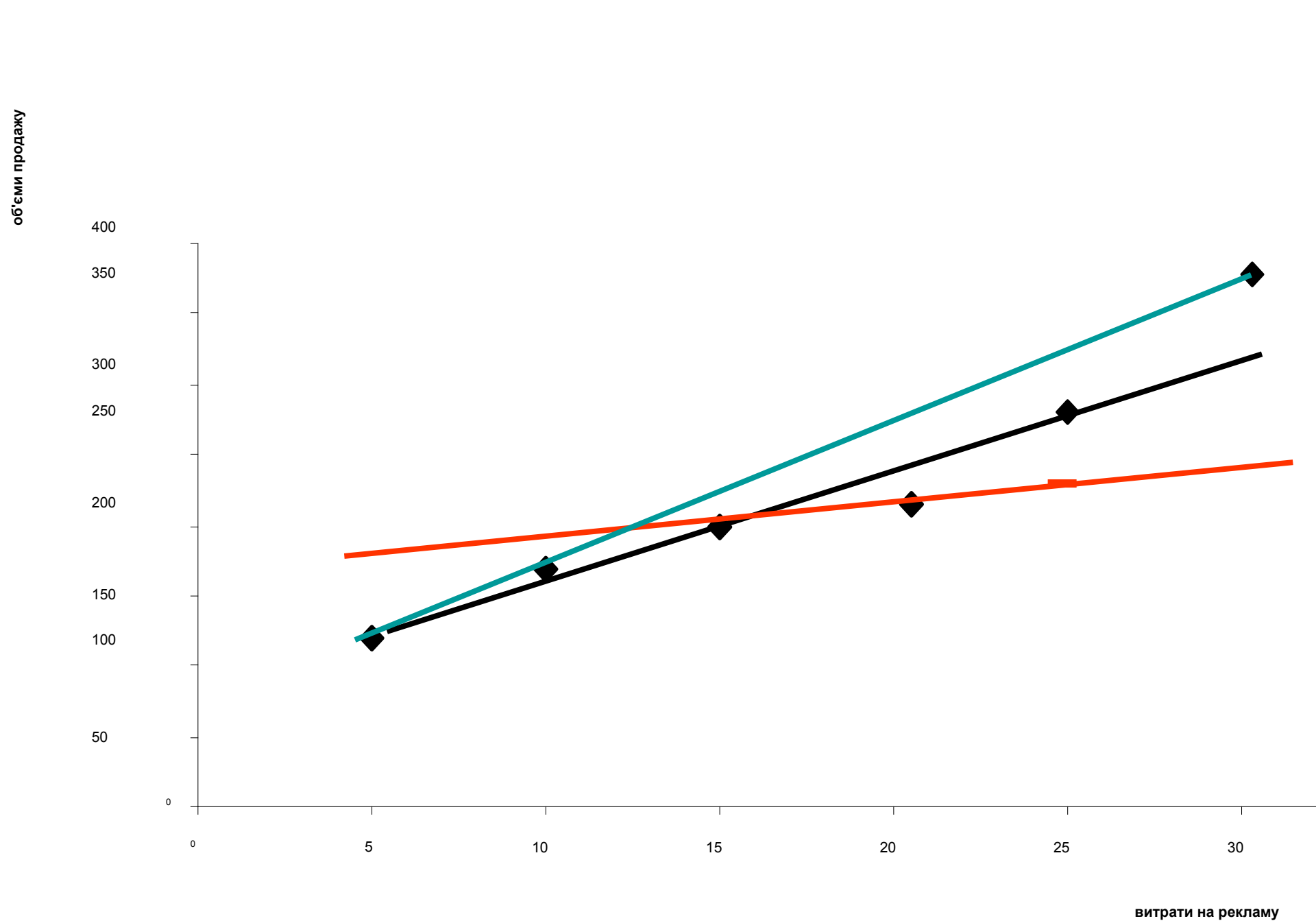
Як це зробити?

Розглянемо приклад

Приклад

Підприємство вирішило проаналізувати зв'язок між витратами на рекламу продукції та відповідними щомісячними об'ємами продаж цієї продукції. Для такої оцінки зафіксовані дані за шість місяців діяльності.

Витрати на рекламу (Х), тис.грн.	***5***	***10***	***15***	***20***	***25***	***30***
Щомісячні об'єми продажу продукції млн.грн. (У)	***120***	***170***	***200***	***220***	***280***	***350***


об'єми продажу
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
витрати на рекламу

- ***Існує безліч прямих, які можна провести через дані точки***

- ***Яку пряму обрати?***
- ***Обирають за критерієм мінімізації суми квадратів відхилень (застосовують метод найменших квадратів - МНК)***

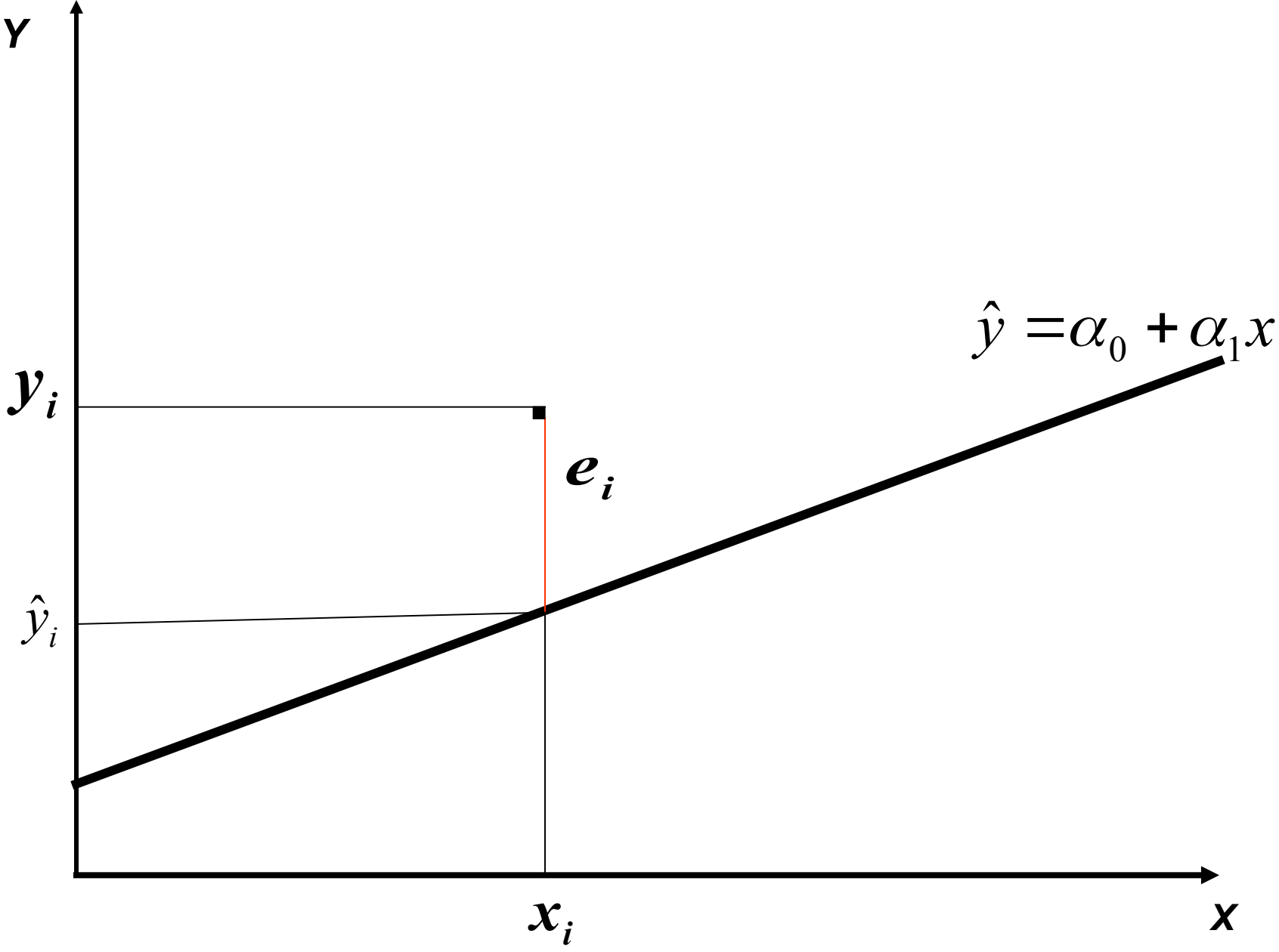
2. Оцінка параметрів лінійної регресії

Оцінка параметрів лінійної регресії проводиться за даними вибірки методом найменших квадратів (МНК)

Покажемо алгоритм цієї оцінки.

На малюнку оберемо конкретну пряму

$$\hat{y} = \alpha_0 + \alpha_1 x$$



$$e_i = y_i - \hat{y}_i = y_i - \alpha_0 - \alpha_1 x_i$$

Пряму треба проводити так, щоб сума квадратів помилок була мінімальною, тобто:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = f(\alpha_0, \alpha_1) \to \min$$

Необхідною умовою існування мінімума неперервно диференційованої функції двох змінних є рівність нулю її частинних похідних:

$$\begin{cases} \dfrac{\partial f}{\partial \alpha_0} = -2\sum (y_i - \alpha_0 - \alpha_1 x_i) = 0 \\ \dfrac{\partial f}{\partial \alpha_1} = -2\sum (y_i - \alpha_0 - \alpha_1 x_i) x_i = 0 \end{cases}$$

$$\begin{cases} n\alpha_0 + \alpha_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \alpha_0 \sum_{i=1}^{n} x_i + \alpha_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Остання система називається нормальною

Розв”язавши її, отримаємо

$$\alpha_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2} = \frac{\mathrm{cov}(x,y)}{D(x)}$$

$$\alpha_0 = \bar{y} - \alpha_1 \bar{x},$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

вибірковий кореляційний момент випадкових величин X і Y.

$$D(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$$

вибіркова дисперсія X

3. Коефіцієнти кореляції та детермінації

Коефіцієнт кореляції

Для того, щоб оцінити щільність лінійного зв'язку між змінними x та y, встановити значимість впливу змінної x на y, слід обчислити коефіцієнт кореляції r_{xy} ***між x та y***

$$-1 \le r_{xy} \le 1$$

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2}} = \frac{\operatorname{cov}(x, y)}{\sqrt{D(x)} \cdot \sqrt{D(y)}}$$

$r_{xy} > 0$ - *зв'язок між змінними прямий*

$r_{xy} < 0$ - *зв'язок між змінними зворотній*

$r_{xy} \to 0$ *-лінійного зв'язку між змінними*

x та y немає взагалі або він дуже слабкий – це означає, що проста лінійна регресія не найкращим чином описує досліджуваний економічний об'єкт

$r_{xy} \rightarrow \pm 1$ *- існує сильний (майже функціональний) лінійний зв'язок між змінними x та y, тобто*

лінійна регресія досить точно описує поведінку досліджуваного економічного об'єкта

Коефіцієнт детермінації

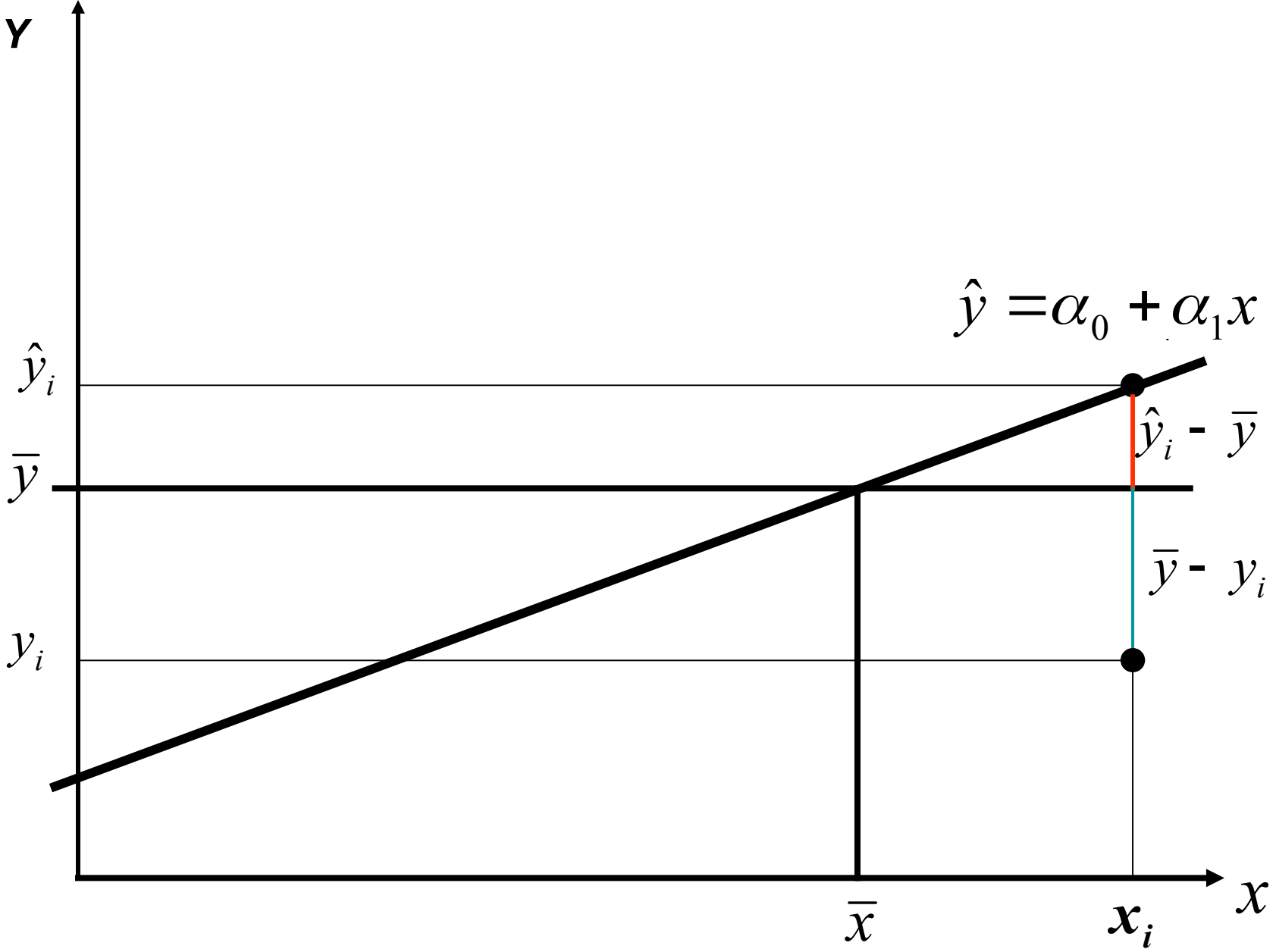
- ***Поряд з коефіцієнтом кореляції використовується для вимірювання щільності зв'язку між двома та більше показниками***
- ***Є критерієм для перевірки достовірності (адекватності) моделі – тобто дає відповідь на питання: чи справді y залежить від x, а не реалізується під впливом різноманітних випадкових факторів***

Прослідкуємо, як аналітично виводиться цей коефіцієнт.

Для цього розглянемо поняття

«Декомпозиція дисперсій»

(одне з центральних питань статистики)



$$(\hat{y}_i - y_i) = (\hat{y}_i - \bar{y}) + (\bar{y} - y_i)$$

$$(\hat{y}_i - y_i) = (\hat{y}_i - \bar{y}) + (\bar{y} - y_i)$$

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \qquad (*)$$

$y_i - \bar{y}$ - *загальне відхилення*

$\hat{y}_i - \bar{y}$ - *відхилення, яке пояснює регресія*

$y_i - \hat{y}_i = e_i$ - *відхилення, яке не пояснює регресія (помилка)*

Якщо піднести до квадрата обидві частини та підсумувати, то одержимо наступне $(\star)$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ ***-загальна сума квадратів***

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ ***- сума квадратів, що пояснює регресію;***

$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ***- сума квадратів помилок***

SST = SSR + SSE

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \sigma^2_{заг}$$ - *загальна дисперсія*

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \sigma^2_{пом}$$ - *дисперсія помилок*

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n} = \sigma^2_{регр}$$ - *дисперсія, що пояснює регресію*

Таким чином, ми розклали загальну дисперсію на дві частини: дисперсію, що пояснює регресію, та дисперсію помилок (або дисперсію випадкової величини):

$$\sigma^2_{заг} = \sigma^2_{пом} + \sigma^2_{регр}$$

Поділимо обидві частини попереднього виразу на $\sigma^2_{заг}$ ***і одержимо***

$$1 = \frac{\sigma^2_{пом}}{\sigma^2_{заг}} + \frac{\sigma^2_{регр}}{\sigma^2_{заг}}$$

Частина дисперсії, що пояснює регресію, називається коефіцієнтом детермінації

$$R^2 = \frac{\sigma^2_{регр}}{\sigma^2_{заг}} = \frac{SSR}{SST}$$

Коефіцієнт детермінації використовується як критерій адекватності моделі, бо є мірою того, на скільки в регресійній моделі y залежить від x (на скільки незалежна змінна x впливає на y)

Коефіцієнт детермінації обчислюється за формулою

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SSR}{SST}$$

$$0 \le R^2 \le 1$$

4. Перевірка моделі на адекватність

Якщо R^2 ***близький до 1, то модель адекватна***

Якщо R^2 ***близький до нуля, то модель неадекватна***

Якщо значення коефіцієнта детермінації має не явно виражене граничне значення (наприклад, 0,5), то зробити однозначний висновок про адекватність моделі неможливо

В цьому випадку треба залучити статистичний критерій.

Найпоширенішим з таких критеріїв є критерій Фішера (або F - критерій)

Перевірка моделі на адекватність за допомогою F - критерія Фішера

1. Розраховуємо F-відношення: $F_{(1,n-2)} = \frac{MSR}{MSE}$

2. Задаємо рівень значущості α

   (рівнем значущості наз. імовірність допустити помилку першого роду (імовірність відкинути правильну гіпотезу))

   α задають наперед (0,1; 0,05; 0,01)

   Якщо $\alpha = 0{,}05$ - це означає, що правильну гіпотезу ми ризикуємо відкинути у 5-ти випадках зі 100)

3. ***За статистичними таблицями F-розподілу Фішера з (1, n-2) ступенями вільності та рівнем значущості*** α ***знаходимо критичне значення критерію Фішера*** $F_{кр}$

4. ***Якщо*** $F > F_{кр}$ ***,то побудована нами регресійна модель адекватна реальній дійсності***

   Якщо $F < F_{кр}$ ***,то - неадекватна***