Кількісні методи досліджень в соціально-гуманітарних науках

Ставицький А.В.
Національний експерт з реформування вищої освіти України в рамках Болонського процесу, к.е.н, доц.
Київський національний університет імені Тараса Шевченка

Чому потрібні кількісні методи?

Незалежно від типу науково-пізнавальної діяльності, в основі будь-якого наукового методу лежать три основні принципи:

1. **Об'єктивність** передбачає відчуження суб'єкта пізнання від його об'єкта, тобто дослідник не дозволяє суб'єктивним уявленням впливати на процес наукового пізнання.
2. **Систематичність** передбачає впорядкованість науково-пізнавальної діяльності, тобто процес наукового пізнання виконується системним, впорядкованим чином.
3. **Відтворюваність**, тобто всі етапи і фази процесу наукового пізнання можна повторити (відтворити) під керівництвом інших дослідників, отримавши подібні, несуперечливі результати, і тим самим перевіривши їх достовірність. Якщо результати не відображаються, то вони ненадійні і, отже, не можуть вважатися достовірними.

Проти кількісного?

- Неможливо чисельно описати певні фактори
- Неможливо зібрати інформацію або неможливо представити її у чисельному форматі
- Мої дослідження не потребують чисел
- Традиції досліджень
- Числа ведуть до математики!!!

Перехід до кількісних методів

- **Реєстрація** - виявлення певної якості у явищ даного класу і підрахунок кількості за наявністю або відсутністю даної якості (наприклад, кількість успішних і неуспішних спостережень).
- **Ранжування** - розташування зібраних даних в певній послідовності (зменшення чи наростання зафіксованих показників), визначення місця в цьому ряду об'єктів, що вивчаються (наприклад, складання списку респондентів від кількості відповідей).
- Відомі чотири основні градації вимірювальних шкал:
  - 1) шкали найменувань (або номінальні);
  - 2) шкали порядку (або рангові);
  - 3) інтервальні шкали;
  - 4) шкали відносин.

Основні характеристики ряду

- Вибіркове середнє $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Дисперсія $\hat{S}^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$
- Середньоквадратичне відхилення

$$S = \sqrt{\hat{S}^2}$$

Кореляція

- В якій мірі дві змінні СПІЛЬНО змінюються? (Тобто, чи спричинить за собою збільшення однієї змінної збільшення або зменшення іншої, або не призведе до змін).
- Коефіцієнт корреляції характеризує силу зв'язку між змінними.
- Це просто параметр описової статистики

Коефіцієнт кореляції Пірсона (Pearson product-moment correlation coefficient *r*)

Karl Pearson (1857 –1936)

$$r = \frac{\sum z_{X_i} z_{Y_i}}{n-1}$$

$$z_{X_i} = \frac{X_i - \overline{X}}{s_X}$$

$$z_{Y_i} = \frac{Y_i - \overline{Y}}{s_Y}$$

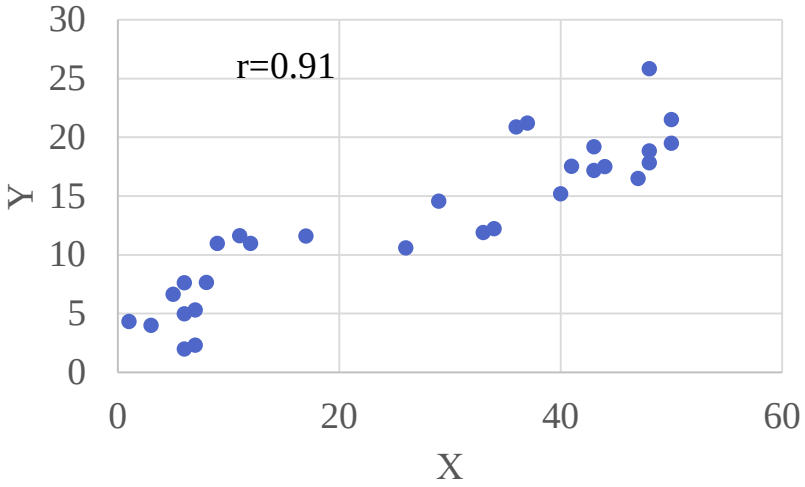
Припустимо, що досліджуємо ховрахів та хочемо дізнатися, чи є взаємозв’ящок між їх масою та довжиною хвоста

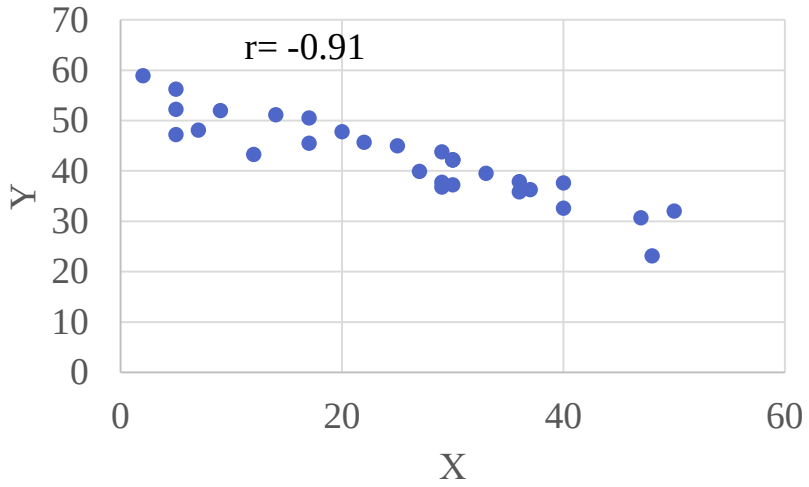
Змінні: 1. вага; 2. довжина хвоста.

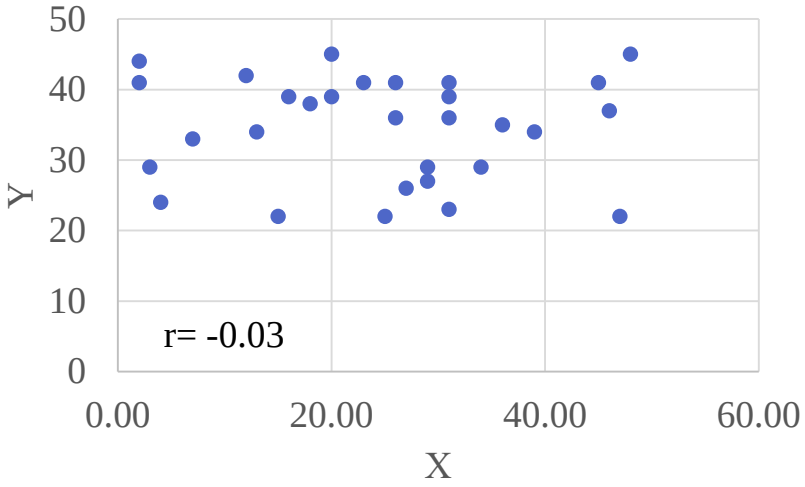
Великий коефіцієнт кореляції між масою тіла і довжиною хвоста дозволяє нам передбачати, що у великого ховраха, швидше за все, і хвіст буде довгим

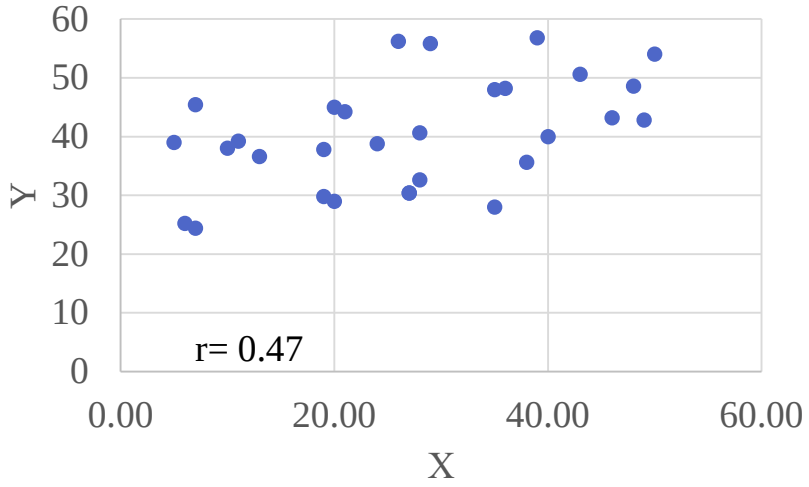
Коефіцієнт кореляції

1. Може приймати значення від -1 до +1
2. Знак коефіцієнта показує напрямок зв'язку (прямий або обернений)
3. Абсолютна величина показує силу зв'язку
4. Завжди заснований на парах чисел


r=0.91
Y
X
0
5
10
15
20
25
30
0
20
40
60


r= -0.91
Y
X
0
10
20
30
40
50
60
70
0
20
40
60


r= -0.03
Y
X
0
10
20
30
40
50
0.00
20.00
40.00
60.00


r= 0.47
Y
X
0
10
20
30
40
50
60
0.00
20.00
40.00
60.00

Ховрах	Вага	Хвіст
№1	72	160
№2	66	144
№3	68	154
№4	74	210
№5	68	182
№6	64	159
	68,7	168,2

$$r = \frac{\sum z_{X_i} z_{Y_i}}{n-1}$$

число рядків (ховрахів)

$$z_{X_i} = \frac{X_i - \overline{X}}{s_X}$$

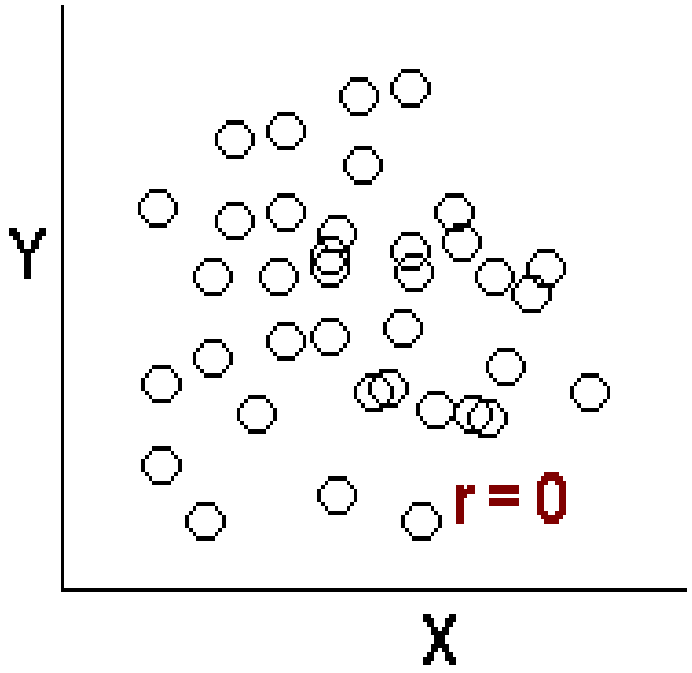
стандартне відхилення для ваги

$$z_{Y_i} = \frac{Y_i - \overline{Y}}{s_Y}$$

стандартне відхилення для хвоста

$$r = 0.68$$

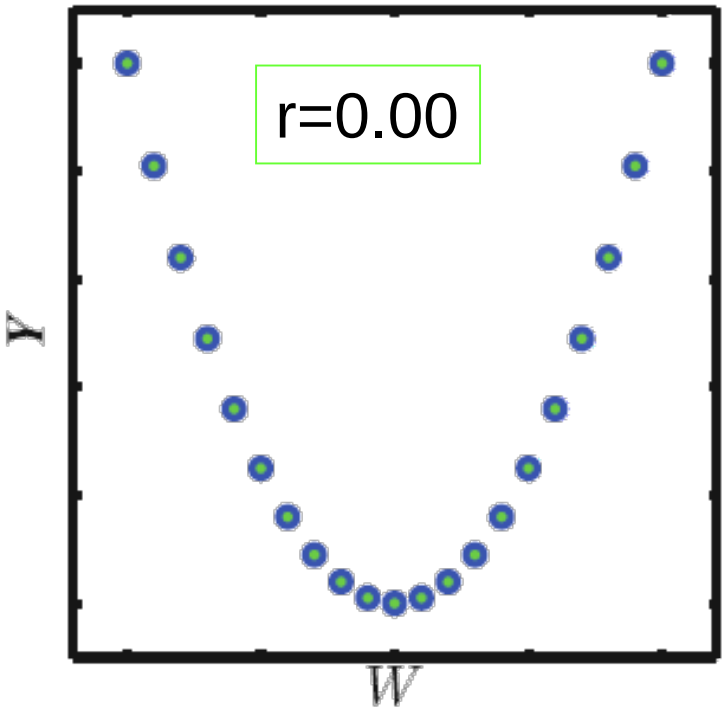
Створюється враження, що близький до нуля коефіцієнт кореляції говорить про те, що зв'язку між змінними немає або майже немає.



Тут її немає

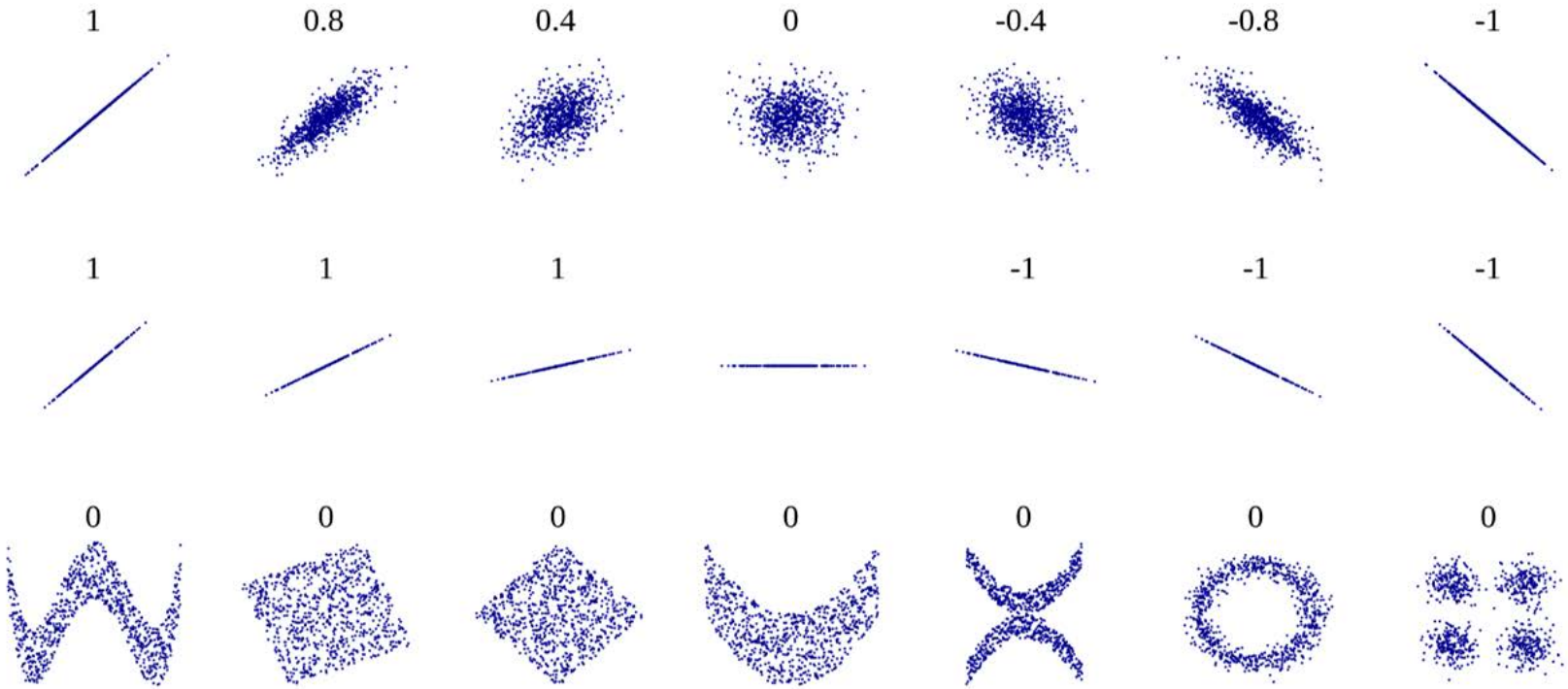
АЛЕ це не завжди так, є винятки.

Коефіцієнт кореляції Пірсона оцінює тільки лінійний зв'язок змінних!
Він не показує наявність нелінійного зв'язку!

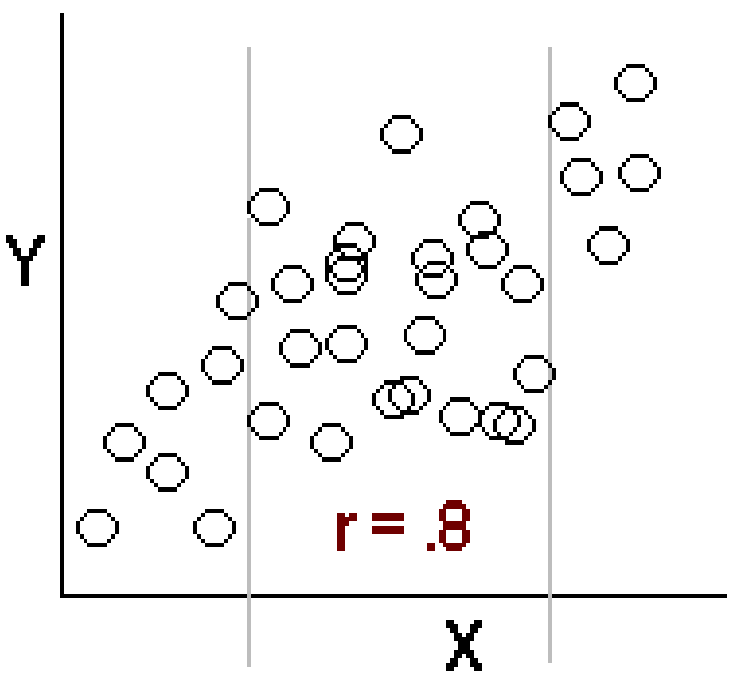


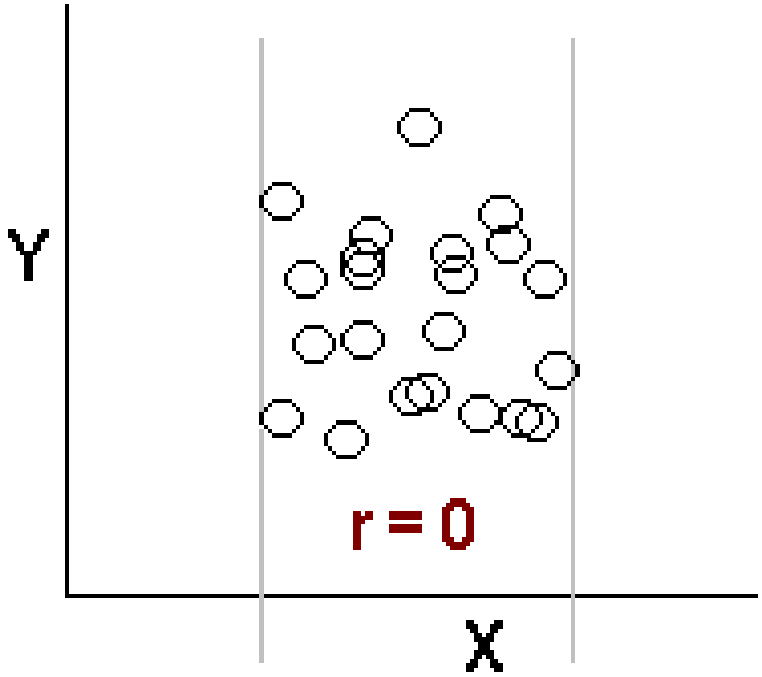
Тут зв'язок змінних дуже сильний, але r = 0.00

Приклади

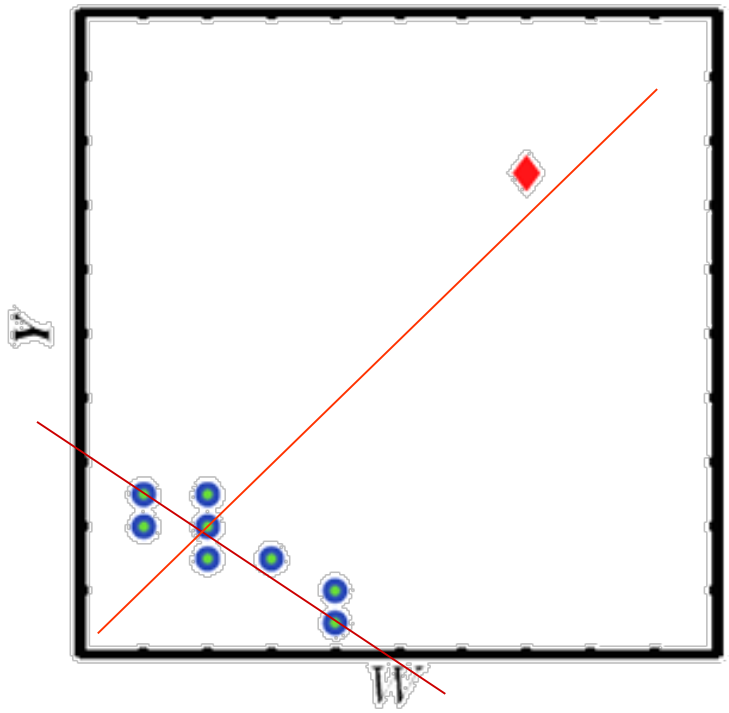


Необхідно, щоб у змінних була значна мінливість! Якщо сформувати вибірку з однотипних осіб, годі сподіватися виявити там кореляцію.





Коефіцієнт кореляції Пірсона дуже чутливий до викидів з вибірки.



Важливо!

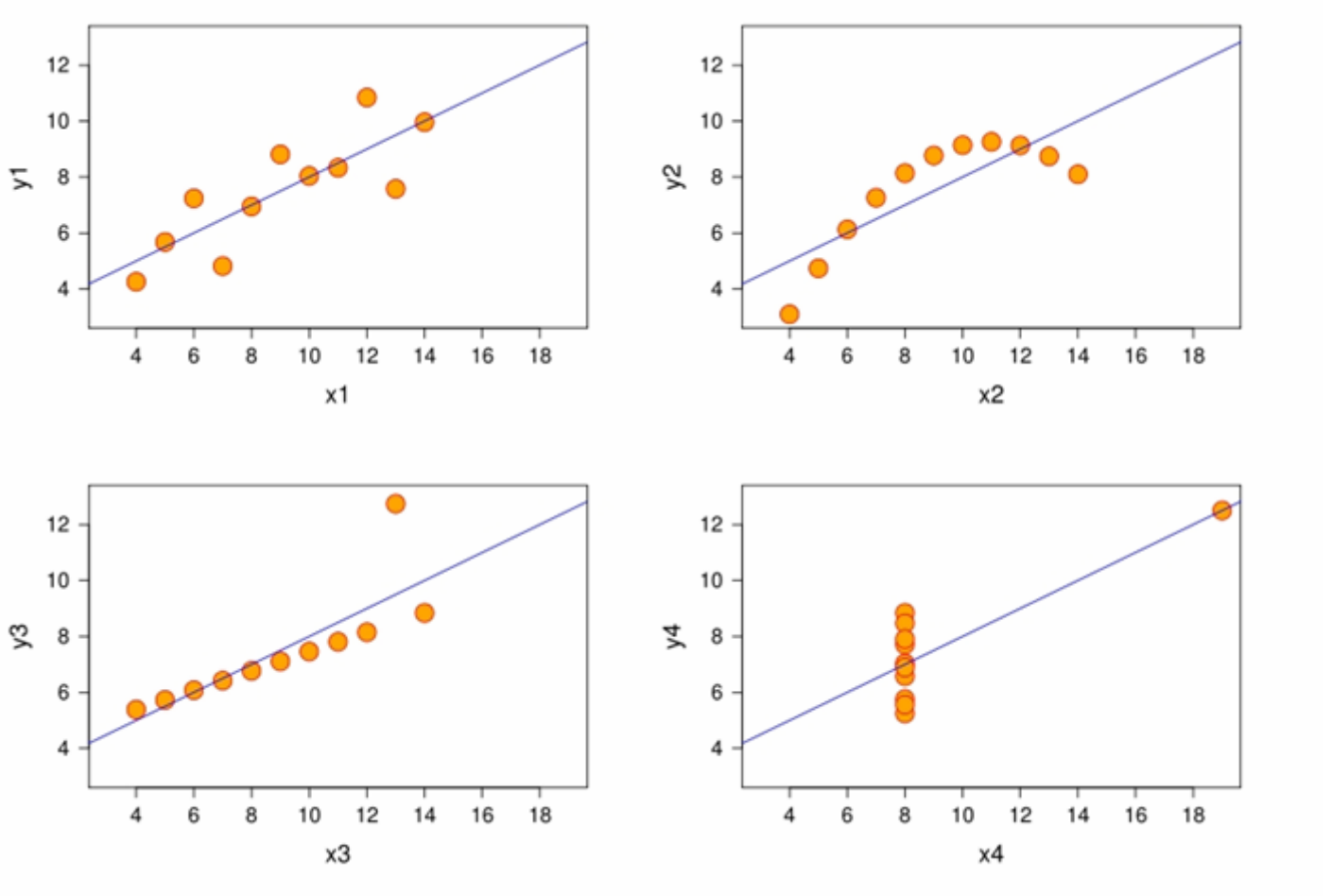
- Кореляція абсолютно не має на увазі наявність причинно-наслідкового зв'язку!

- Вона ВЗАГАЛІ НІЧОГО про нього НЕ ГОВОРИТЬ (навіть дуже великий r)

- Коефіцієнт кореляції Пірсона - параметр вибірки.Чи можемо ми на основі нього судити про популяцію?
- Просто дивлячись на коефіцієнт - НІ.

Кореляція між змінними всюди

r = **0.816**



Кореляційна матриця

- При використанні декількох рядів будується кореляційна матриця.
- Основний ризик – пряма інтерпретація коефіцієнтів

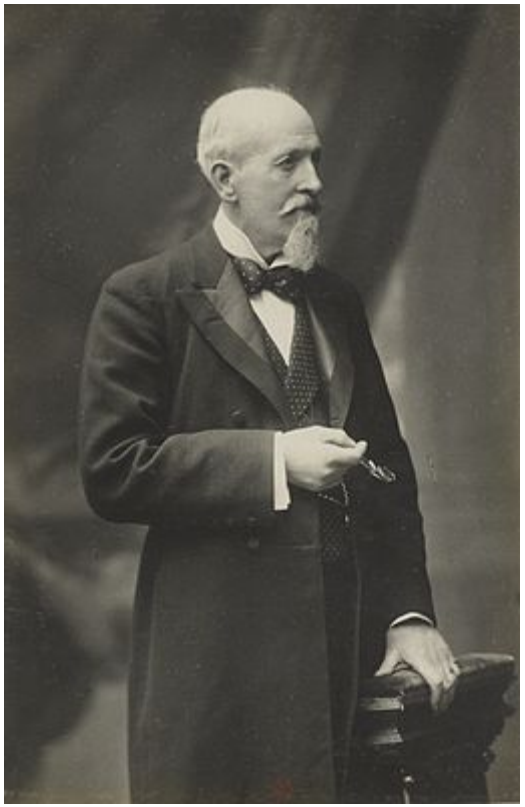
Що робити, якщо немає чисел?

Непараметричні тести для асоціацій

- Коефіцієнт кореляції Спірмена (Spearman rank order correlation)
- Коефіцієнт кореляції Кендалла (Kendall's coefficient of rank correlation, Kendall-τ)

Коефіцієнт кореляції Спірмена

- Чи пов’язані відстань до університету та рівень успішності студента?
- Чи пов’язані розмір заробітної плати та підтримка політичної партії?
- Чи пов’язані довжина тексту у реченні та сила голосу?



Charles Edward Spearman
1863-1945

Для нашої задачі не годиться коефіцієнт кореляції Пірсона, оскільки принаймні одна зі змінних рангова!

Коефіцієнт кореляції Спірмана:

1. Ранжируємо дані для кожної змінної від меншого до більшого;
2. Якщо зустрілися однакові значення (tied ranks), присвоюємо їм середні ранги;
3. Рахуємо різниці рангів в кожному рядку;
4. Рахуємо коефіцієнт r_s

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

різниці рангів

число рядків (розмір вибірки)

Коефіцієнт кореляції Кендалла

- Він оцінює різницю між ймовірністю того, що порядок даних в обох змінних однаковий, і ймовірністю того, що порядки різні.
- Рахується зовсім не так, як коефіцієнт Спірмана.
- Тільки для рангових змінних!

Sir Maurice George Kendall
1907–1983

- Чи чесно судді оцінюють спортсменів?
- Чи є зв’язок між рангом інвестицій та рангом обсягу випуску у філіалах корпорації?

Коефіцієнт кореляції Кендалла

- Нехай є набір пар (x_i, y_i), причому всі значення різні.
- Нехай S_1 – кількість узгоджених за рангом пар, S_2 – кількість неузгоджених за рангом пар.

$$\tau = \frac{S_1 - S_2}{\frac{1}{2} n(n-1)}$$

Властивості

- -1 ≤ Коефіцієнт Кендалла ≤ 1
- Чим ближче коефіцієнт до 1, тим вище кореляція.
- Чим ближче до нуля, тим менше зв'язок змінних (наприклад, згода експертів).
- Чим ближче коефіцієнт до -1, тим вище обернена кореляція.

Приклад – 1 (Коефіцієнт кореляції Пірсона)

№	X	Y
1	18.4	5.57
2	20.6	2.88
3	21.5	4.12
4	35.7	7.24
5	37.1	9.67
6	39.8	10.48
7	51.1	8.58
8	54.4	14.79
9	64.6	10.22
10	90.6	10.45

$$r = 0.683$$

Приклад – 2 (Коефіцієнт кореляції Спірмена)

№	X	Y	Ранг X	Ранг Y	D^2
1	18.4	5.57	1	3	4
2	20.6	2.88	2	1	1
3	21.5	4.12	3	2	1
4	35.7	7.24	4	4	0
5	37.1	9.67	5	6	1
6	39.8	10.48	6	9	9
7	51.1	8.58	7	5	4
8	54.4	14.79	8	10	4
9	64.6	10.22	9	7	4
10	90.6	10.45	10	8	4
					28

$$r = 0.830$$

Приклад – 3 (Коефіцієнт кореляції Кендалла)

№	X	Y	Ранг X	Ранг Y	S_1	S_2
1	18.4	5.57	1	3	7	2
2	20.6	2.88	2	1	8	0
3	21.5	4.12	3	2	7	0
4	35.7	7.24	4	4	6	0
5	37.1	9.67	5	6	4	1
6	39.8	10.48	6	9	1	3
7	51.1	8.58	7	5	3	0
8	54.4	14.79	8	10	0	2
9	64.6	10.22	9	7	1	0
10	90.6	10.45	10	8	0	0
					37	8

$$r = 0.644$$

Перевірка гіпотези про значимість

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$$t_{pr} = \rho \frac{\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

$$t_{table} = t\left(n-2;1-\alpha\right)$$

- Якщо $t_{pr} < t_{table}$, то коефіцієнт є значимим.

Питання

- Що робити, якщо цікавить питання не сили зв’язку, а того, як одна змінна впливає на іншу?

- Регресійний аналіз!

Розбіжність

- РЕГРЕССИЯ (regression) – прогнозування однієї змінної на підставі іншого. Одна змінна - незалежна (independent), а інша - залежна (dependent).
- Приклад: швидкість набору ваги у бегемота зростає із збільшенням тривалості годування; якщо довго годувати бегемота, то він швидше набирає вагу

- КОРЕЛЯЦІЯ (correlation) - показує, в якій степені дві змінні СПІЛЬНО змінюється. Не існує ні залежної, ні незалежної змінних, вони еквівалентні.
- Приклад: довжина хвоста у ховраха корелює позитивно з його масою тіла

Типи даних

- ✓ Часові ряди (Time series data)
- ✓ Перехресні дані (Cross-sectional data)
- ✓ Панельні дані (Panel data)

Часові ряди

- **Види**
  - кількісні (наприклад, обмінний курс, ціни акцій, випуск продукції),
  - якісні (наприклад, день тижня, стать людини).

- **Приклади часових рядів**

Ряди	**Частота**
GNP, безробіття	місячні чи квартальні
дефіцит бюджету	квартальні чи річні
Пропозиція грошей	тижневі чи місячні
Біржовий індекс	щохвилини

Приклади задач для часових рядів

1. Як кількість психічно хворих залежить від середньої зарплати в країні?
2. Як впливає новий медикамент на кількість днів лікування хворого?
3. Як зміна ставок НБУ впливає на курс валюти?

Перехресні дані

- Перехресні дані – дані, у яких розглядаються значення декількох змінних у один період часу:
  - Набір обсягу продажів всіх брокерів біржі за один день
  - Кількість пацієнтів по всіх лікарнях країни
  - Набір даних про кредитні рейтинги банків

Приклади задач для перехресних даних

- Залежність між розміром області та кількістю призерів Олімпіади
- Залежність між ВВП та ймовірністю дефолту за державними боргами

Панельні дані

- Панельні дані узагальнюють часові ряди та перехресні дані, використовуючи спостереження за декількома змінними декілька періодів.

Приклади задач для панельних даних

- Залежність між ВВП та обсягом державного боргу у всіх країнах ЄС
- Залежність між обсягом продажів та кількістю персоналу у всіх філіях компанії

Етапи моделювання

- Визначення проблеми чи гіпотези
- Розробка моделі
- Збір статистики
- Проведення описового аналізу даних
- Оцінка невідомих коефіцієнтів
- Оцінка якості та придатності моделі
- Використання моделі для прогнозу та аналізу

Мета

- Розробка статистичної моделі, яка дозволить прогнозувати значення залежної змінної на основі незалежних змінних.

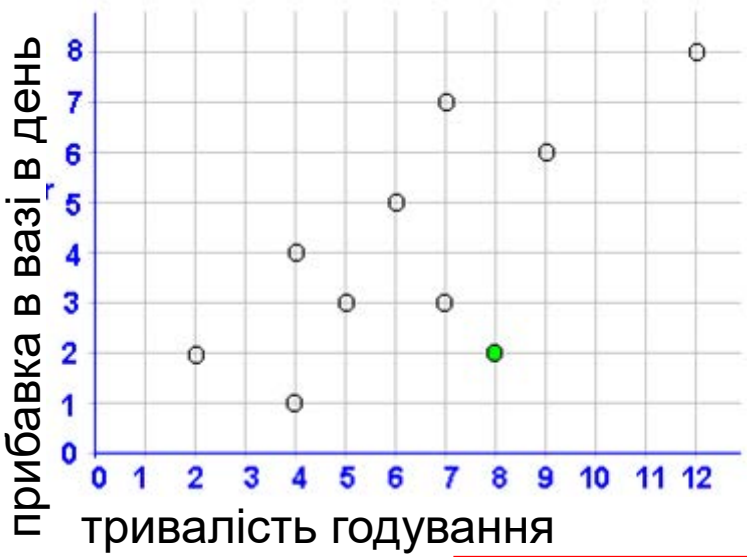
Приклад

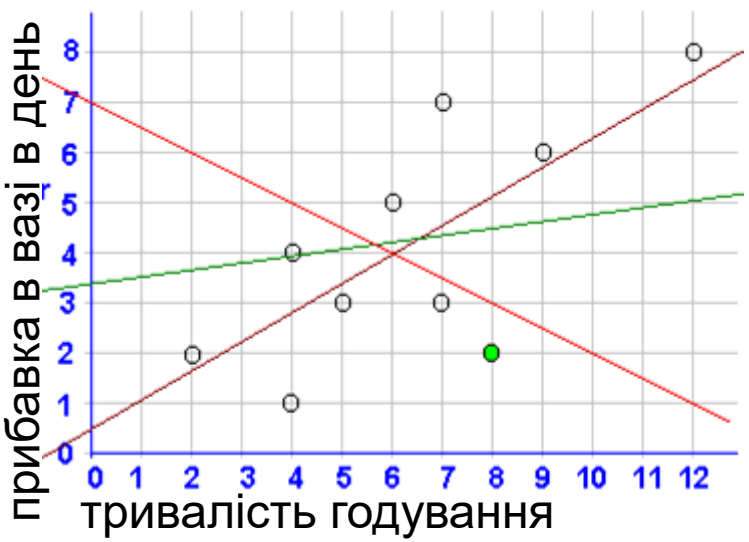
Нехай ми хочемо дізнатися, як пов'язана тривалість годування бегемотів в Африці зі швидкістю набору ваги у цих звірів?

У нас дві змінні:

- 1. тривалість годування в день (independent);
- 2. швидкість набору ваги в день (dependent)

Ми шукаємо пряму, яка найкращим чином буде передбачати значення Y на підставі значень X.





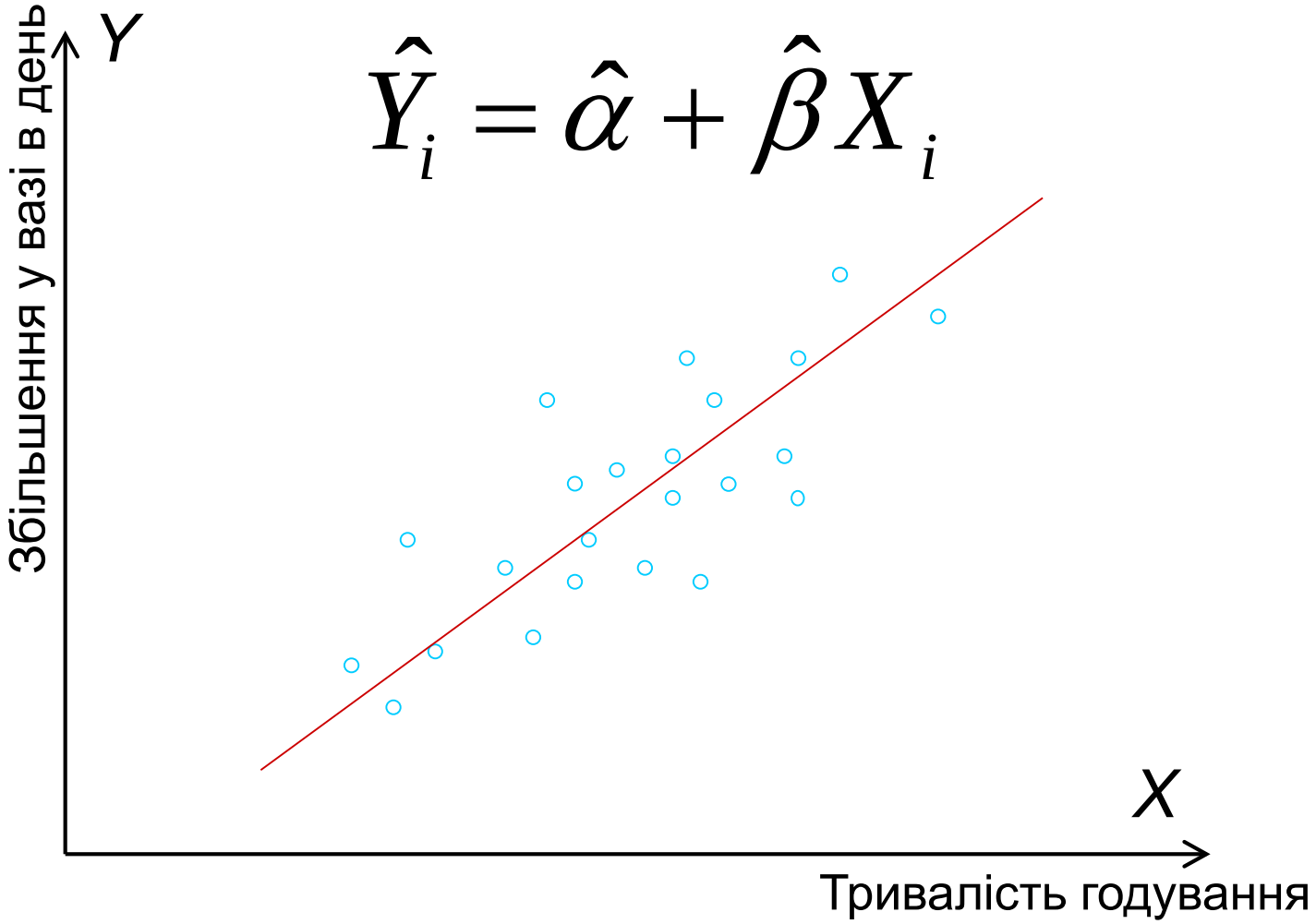
Проста лінійна регресія (*linear regression*)

- Y – залежна змінна
- X – незалежна змінна
- $\hat{\alpha}$ та $\hat{\beta}$ - коефіцієнти регресії

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

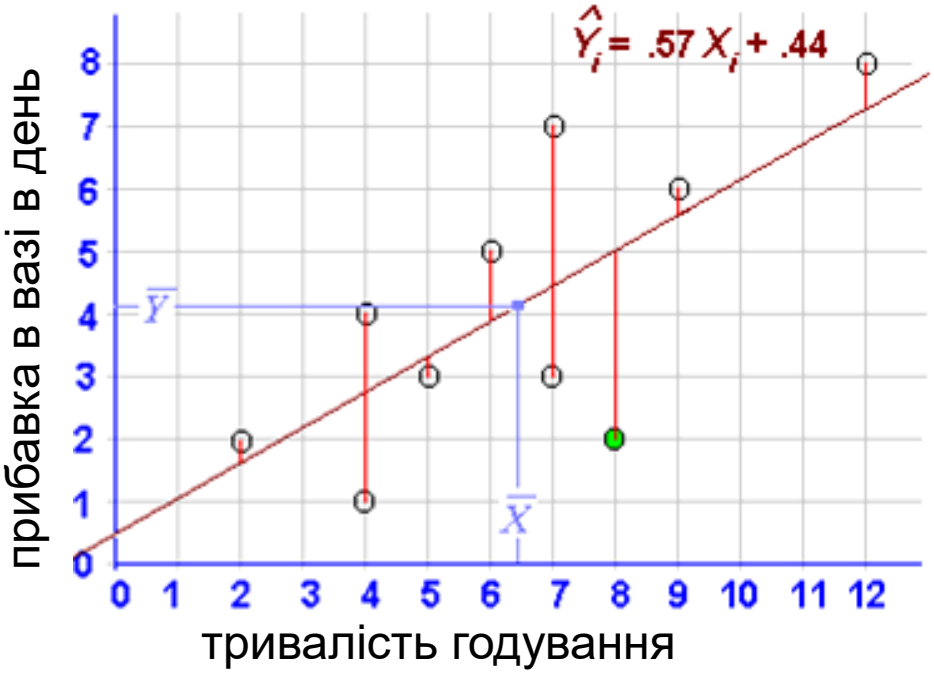
- $\hat{\beta}$ – характеризує НАХИЛ прямий; це найважливіший коефіцієнт;
- $\hat{\alpha}$ – визначає точку перетину прямої з віссю OY; не настільки істотний коефіцієнт (intercept).
- Яка розмірність коефіцієнтів?

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$



Помилка за моделлю (residual) = «залишки»

- e_i *позитивне для точок на прямій і негативне для точок під прямою.*



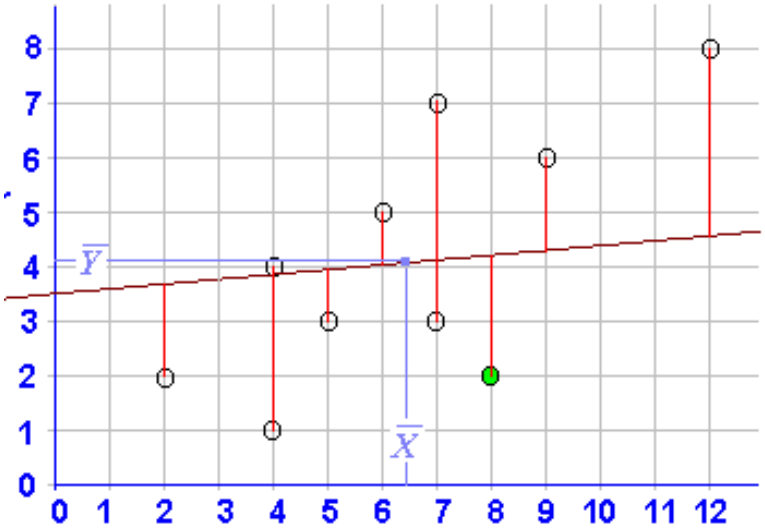
$$e_i = Y_i - \hat{Y}_i$$

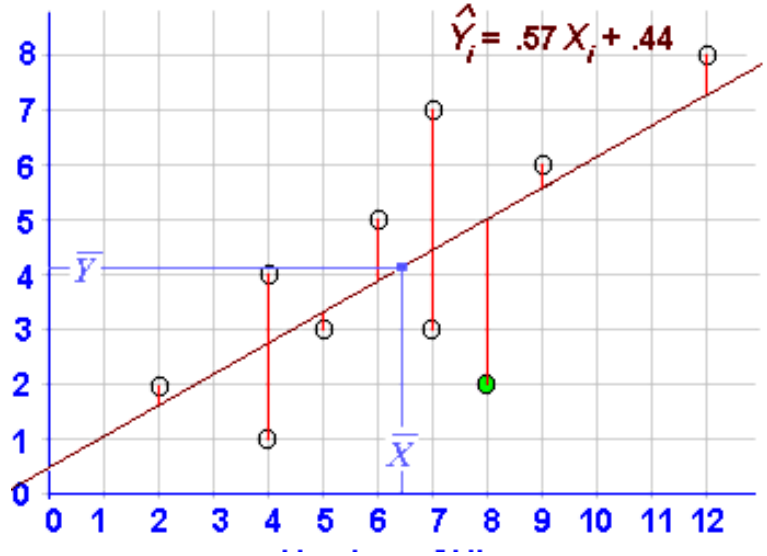
Метод найменших квадратів

Лінію регресії вибирають таку, щоб загальна сума квадратів помилок (residuals) була **найменшою.**

$$\sum e_i = 0$$

$$\sum e_i^2 \quad \text{- мінімальна}$$





Лінійна регресія

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t} + \ldots + \beta_{k-1} x_{k-1t} + \varepsilon_t, t = \overline{1,n}$$

y_t - залежна змінна;

$x_{1t}, x_{2t}, \ldots, x_{k-1t}$ - незалежні змінні;

ε_t - збурення.

Припущення

- **Лінійність -** Y лінійно залежить від набору X.
- **Незалежність похибок** – збурення незалежні з X.
- **Гомоскедастичність** – дисперсія збурень є константою для всіх X.
- **Нормальність** – збурення мають нормальний розподіл.

Метод найменших квадратів

- Визначає коефіцієнти регресії, при яких різниця між реальними даними (Y) та прогнозними (Y-hat) буде найменшою:

$$L = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \rightarrow \min$$

Програми

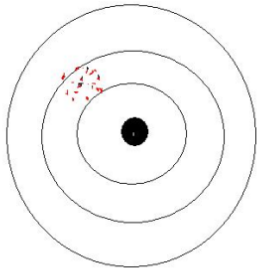
- MS Excel
- EViews
- Mathematica
- SPSS
- Statistica
- R/R-Studio
- Gretl
- MathLab
- …

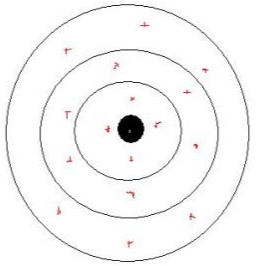
Бажані характеристики

- **Незміщеність** $E(\hat{\beta}) = \beta$
- **Ефективність**
  - Стандартна похибка/дисперсія має бути мінімальною:

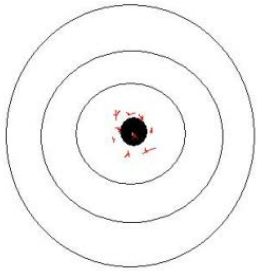
$$\text{var}(\hat{\beta}) = \frac{1}{\sum x_i^2}\sigma^2 = \frac{\sigma^2}{\sum x_i^2}$$

МНК мінізмізує σ^2 (дисперсію збурень)





- **Консистентність**
  - При збільшенні N стандартна похибка зменшується



Перетворення до лінійного вигляду

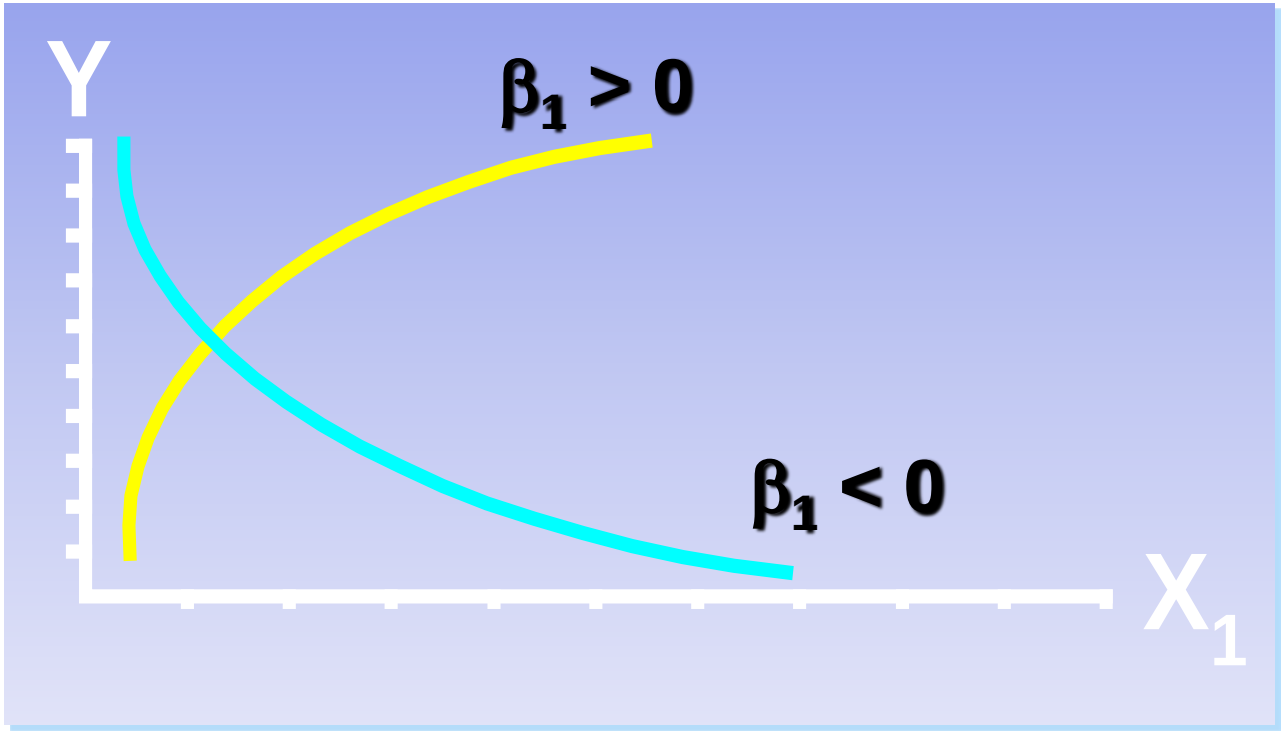
- Нелінійні моделі можуть бути виражені та оцінені у лінійній формі
- Необхідна трансформація даних

Популярні нелінійні регресійні моделі

- Експоненціальна: $(y = ae^{bx})$
- Степенева: $(y = ax^{b})$
- Модель постійного росту: $\left(y = \frac{ax}{b + x} \right)$
- Поліноміальна: $(y = a_0 + a_1 x + \ldots + a_m x^{m})$

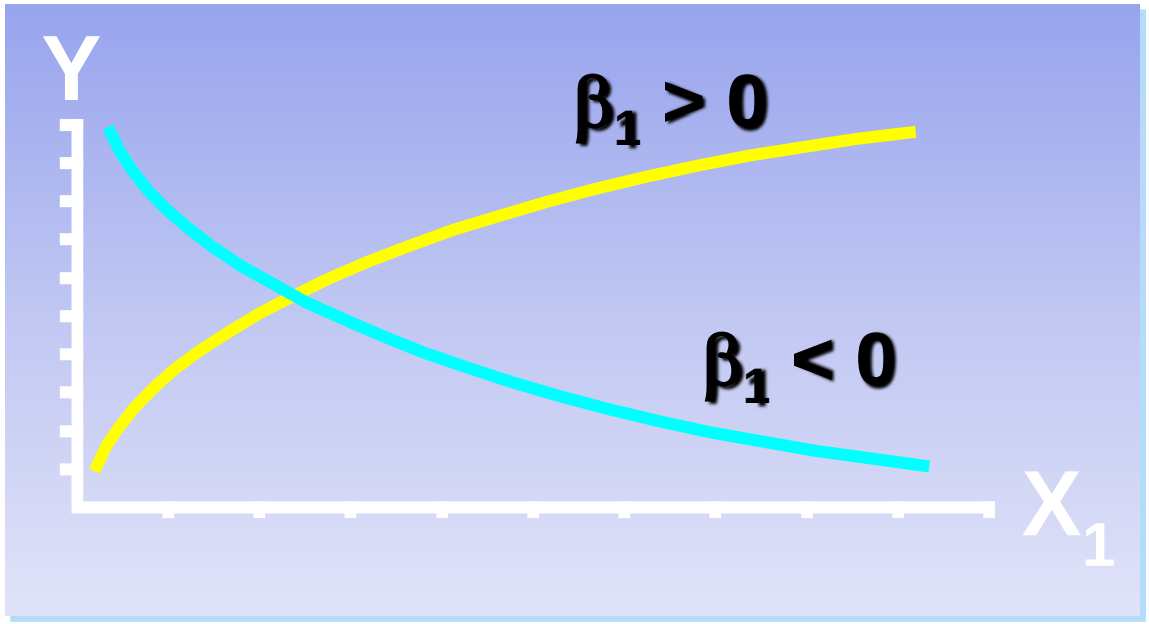
Логарифмічна трансформація

$$Y = \beta + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$$



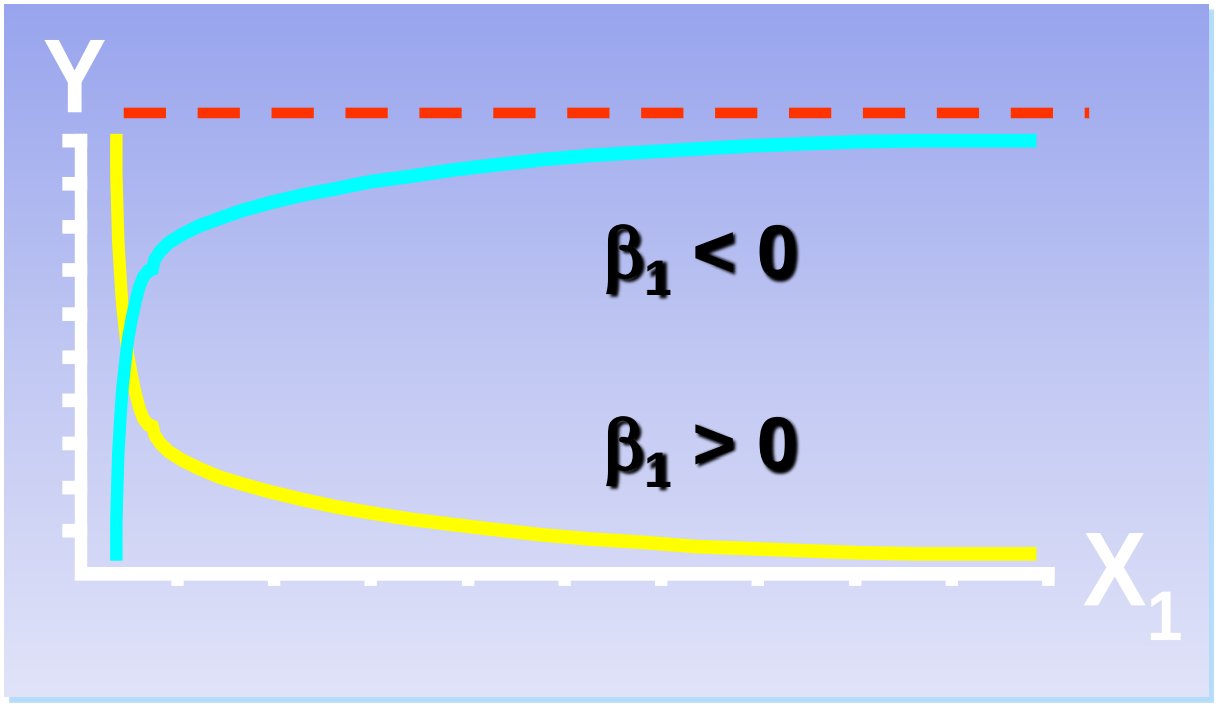
Степенева трансформація

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \beta_2 \sqrt{X_{2i}} + \varepsilon_i$$



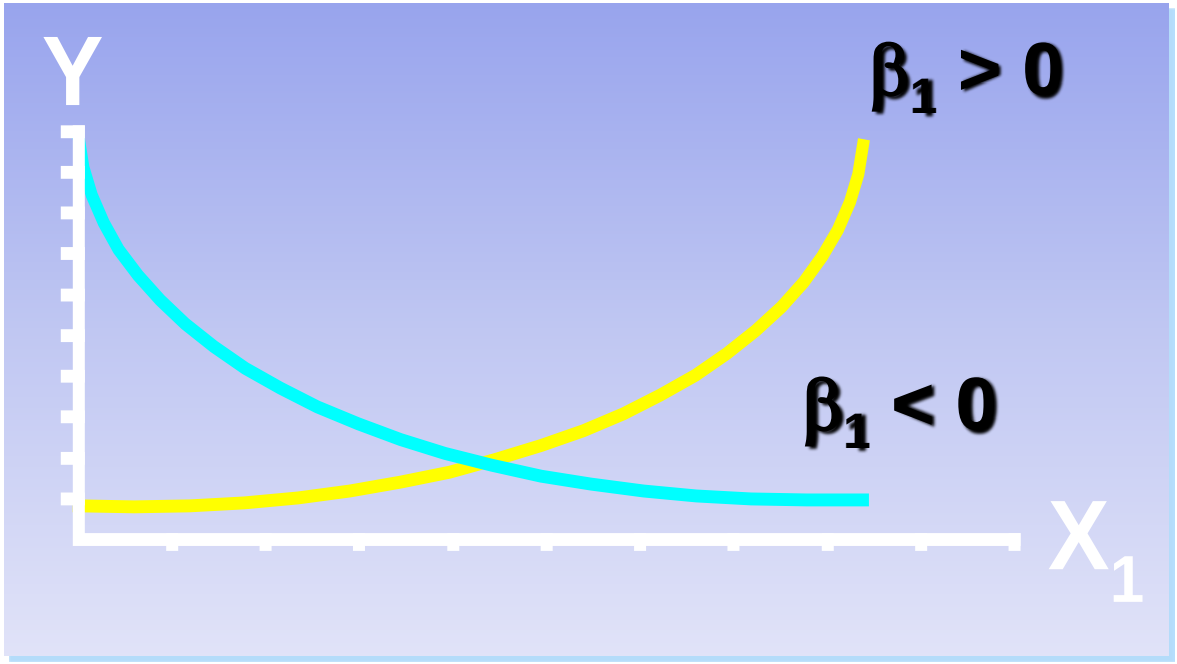
Обернена трансформація

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_{1i}} + \beta_2 \frac{1}{X_{2i}} + \varepsilon_i$$



Експоненційна трансформація

$$Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i$$



Моделі з фіктивними змінними

- Використовує незалежні змінні з 2 значеннями:
  - *наприклад, жінки-чоловіки, працює-безробітний тощо.*
- Змінні приймають значення 0 та 1
- Передбачається, що константа моделі різна, а інші коефіцієнти регресії - однакові.

Вибір тренду

Щоб правильно підібрати вид функції тренду або “кривої зростання”, треба знати різноманітні класи цих кривих і їхнє характерне поводження в залежності від зміни часу. Тоді, побудувавши графік часового ряду, іноді можна візуально підібрати потрібний клас кривих. При цьому слід враховувати:

•на якій стадії розвитку знаходиться процес (початкова стадія розвитку, стадія стабільного росту, стадія насичення);

•багато кривих на обмежених (можливо різних) ділянках мають схожі графіки, що може призвести до неправильної ідентифікації класу функцій.

Основні види функції тренду - 1

Лінійний тренд $f(t) = a_0 + a_1 t$

Квадратичний тренд $f(t) = a_0 + a_1 t + a_2 t^2$

Поліноміальний тренд $f(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n$

Експоненціальний тренд $f(t) = a_0 e^{a_1 t}$

Показниковий тренд $f(t) = a_0 t^{a_1}$

Основні види функції тренду - 2

Гіперболічний тренд

$$f(t) = \frac{a_0}{1 + a_1 t}$$

Логарифмічний тренд

$$f(t) = a_0 + a_1 \ln(t)$$

Логістичний тренд

$$f(t) = \frac{a_0}{1 + a_1 e^{a_2 t}}$$

Основні види функції тренду - 3

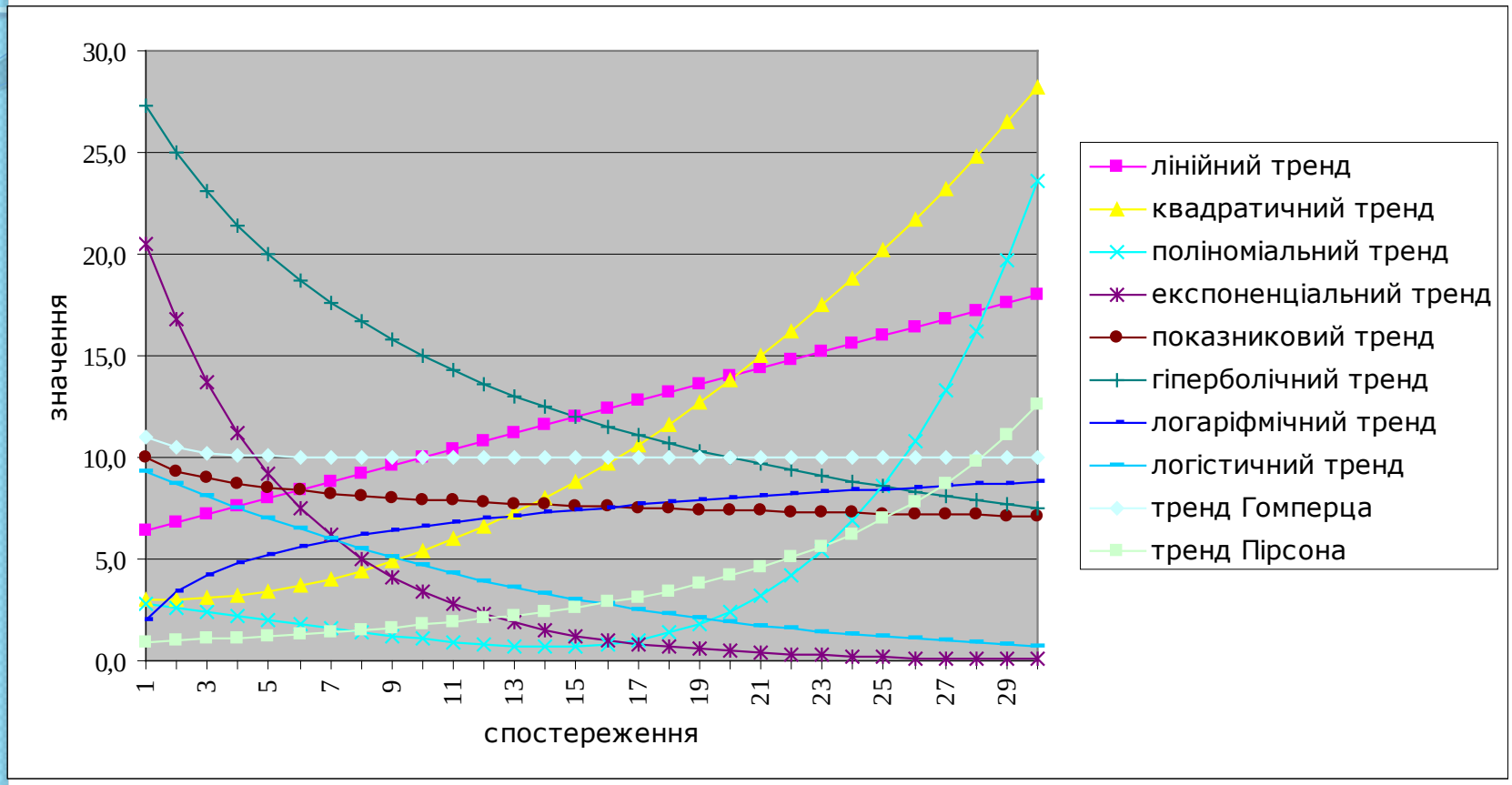
Тренд Гомперца

$$f(t) = a_0 a_1^{a_2^t}$$

Тренд Пірсона

$$f(t) = a_0 \left(1 - \frac{t - a_1}{a_2}\right)^{-a_3} \left(1 - \frac{t - a_4}{a_5}\right)^{-a_6}$$

Приклад


30,0
25,0
20,0
15,0
10,0
5,0
0,0
значення
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
спостереження
лінійний тренд
квадратичний тренд
поліноміальний тренд
експоненціальний тренд
показниковий тренд
гіперболічний тренд
логаріфмічний тренд
логістичний тренд
тренд Гомперца
тренд Пірсона

Виділення сезонності - 1

Фіктивні або бінарні змінні приймають тільки значення 0 або 1. Наприклад, якщо ми розглядаємо часовий ряд з квартальною структурою даних, то доцільним є розгляд такої моделі:

$$y_t = \beta_0 + \beta_1 q_1 + \beta_2 q_2 + \beta_3 q_3 + \varepsilon_t$$

де q_1 приймає значення 1, якщо відповідає першому кварталу року, 0 - в інших випадках;
q_2 - приймає значення 1, якщо відповідає другому кварталу року, 0 - в інших випадках тощо.

Виділення сезонності - 2

Таким чином,

$$q_1 = (1,0,0,0,1,0,0,0,1,0,\ldots)'$$

$$q_2 = (0,1,0,0,0,1,0,0,0,1,0,\ldots)'$$

$$q_3 = (0,0,1,0,0,0,1,0,0,0,1,0,\ldots)'$$

Виділення сезонності - 3

Тоді у перший квартал

$$y_t = \beta_0 + \beta_1 q_1 + \varepsilon_t$$

другий

$$y_t = \beta_0 + \beta_2 q_2 + \varepsilon_t$$

третій

$$y_t = \beta_0 + \beta_3 q_3 + \varepsilon_t$$

четвертий

$$y_t = \beta_0 + \varepsilon_t$$

Виділення сезонності - 4

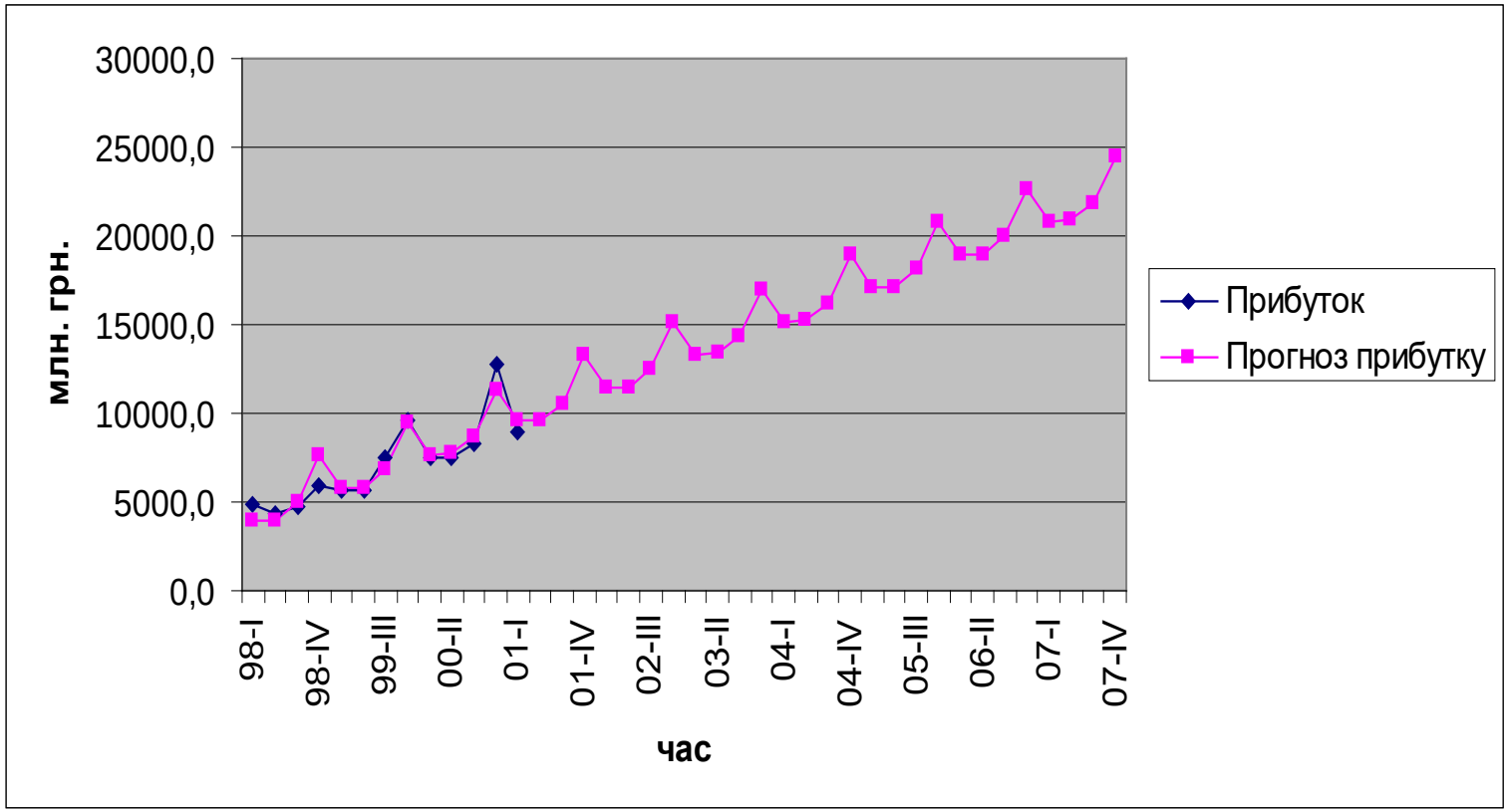
Вищенаведена модель оцінюється за звичайним МНК, знаходяться коефіцієнти

$$\beta_i \ , \ i = \overline{0,3}$$

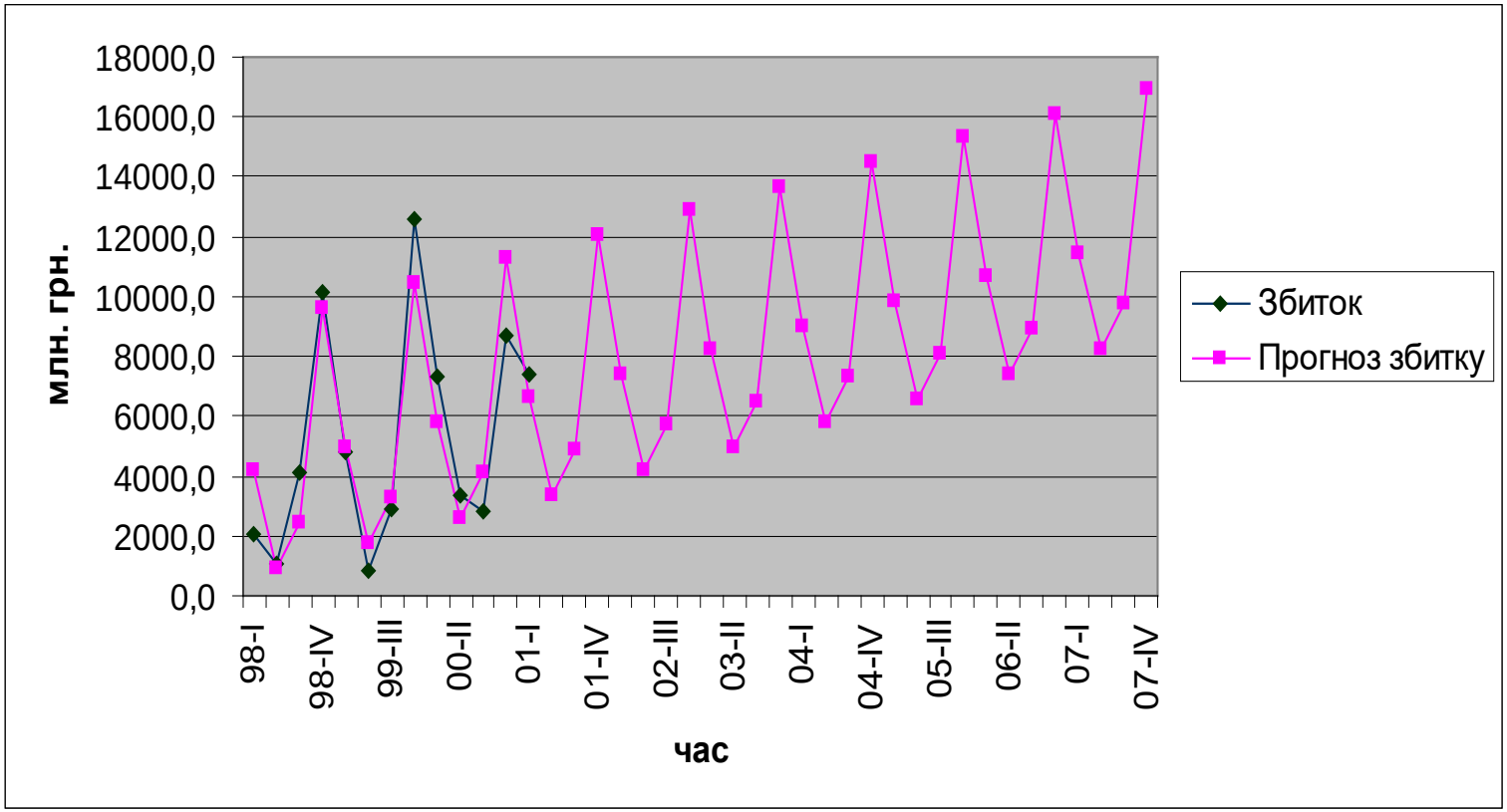
Можливо, часовий ряд крім сезонного компонента має і трендовий. В такому разі, модель можна розширити:

$$y_t = \beta_0 + \alpha t + \beta_1 q_1 + \beta_2 q_2 + \beta_3 q_3 + \varepsilon_t$$

Приклад - 1


30000,0
25000,0
20000,0
15000,0
10000,0
5000,0
0,0
млн. грн.
98-I
98-IV
99-III
00-II
01-I
01-IV
02-III
03-II
04-I
04-IV
05-III
06-II
07-I
07-IV
час
Прибуток
Прогноз прибутку

Приклад - 2


18000,0
16000,0
14000,0
12000,0
10000,0
8000,0
6000,0
4000,0
2000,0
0,0
млн. грн.
98-I
98-IV
99-III
00-II
01-I
01-IV
02-III
03-II
04-I
04-IV
05-III
06-II
07-I
07-IV
час
Збиток
Прогноз збитку

Лагові змінні

- Дані часових рядів за попередні періоди
- Дозволяють пояснити інертність ряду

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{2t-1} + \beta_4 y_{t-1} + \beta_5 y_{t-2} + \varepsilon_t$$

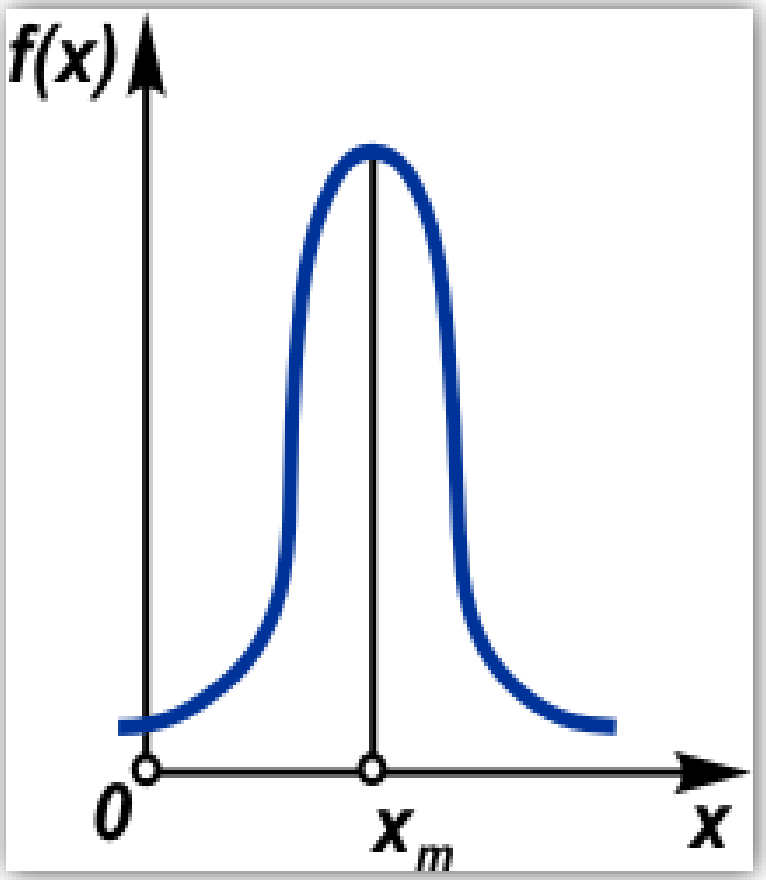
- Проблеми з автокореляцією!

Кроки для оцінки якості регресійної моделі

- Тест залишків на нормальність
- Тест значимості коефіцієнтів
- Тест адекватності моделі
- Тест на мультиколінеарність
- Тест на стійкість моделі
- Тест на автокореляцію залишків
- Тест на гетероскедастичність залишків
- Тест на специфікацію моделі
- Тест на стаціонарність даних

Тест залишків на нормальність

Критерій Жарке-Бера (Jarque-Bera statistics)



Критерій Жарке-Бера

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}\left((K-3)^2 \right) \right)$$

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{2}}$$

- ***S*** *- коефіцієнт асиметрії,*
- ***K*** *- ексцес.*

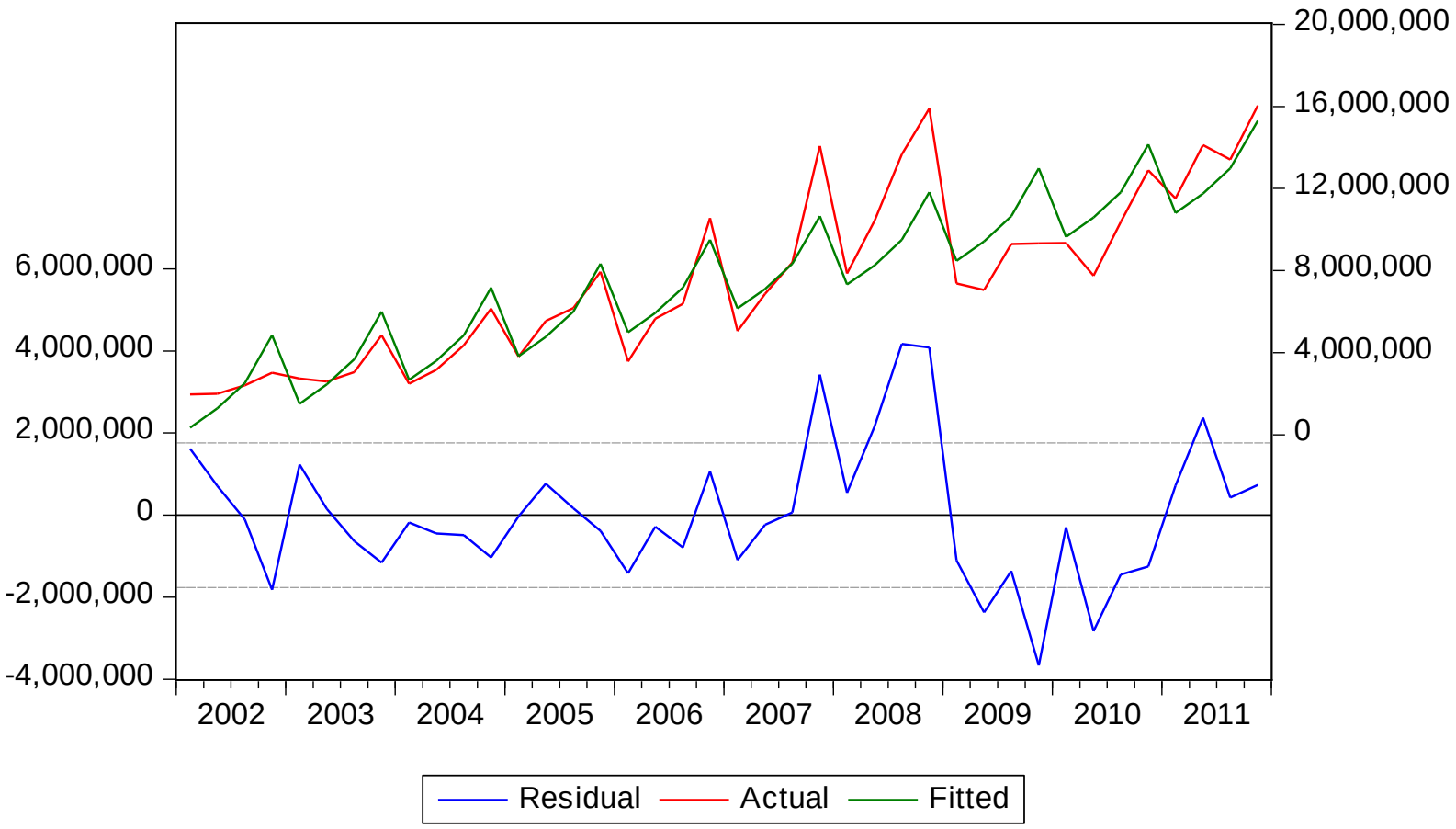
Приклад

Dependent Variable: TAX_ENT
Method: Least Squares
Date: 12/09/12 Time: 20:49
Sample: 2002Q1 2011Q4
Included observations: 40

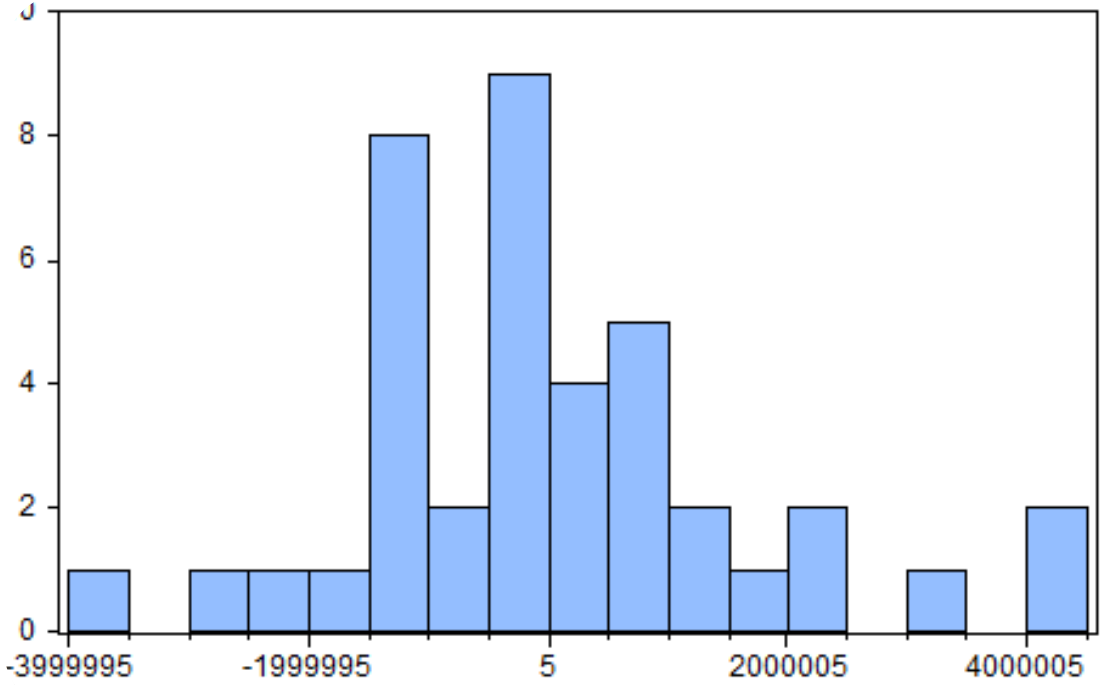
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3973770.	754540.7	5.266475	0.0000
@TREND	290525.1	24239.34	11.98568	0.0000
@SEAS(1)	-3627516.	791034.8	-4.585786	0.0001
@SEAS(2)	-2975920.	789175.7	-3.770922	0.0006
@SEAS(3)	-2032456.	788058.1	-2.579068	0.0143

R-squared	0.837415	Mean dependent var	7480035.
Adjusted R-squared	0.818834	S.D. dependent var	4138083.
S.E. of regression	1761318.	Akaike info criterion	31.71749
Sum squared resid	1.09E+14	Schwarz criterion	31.92860
Log likelihood	-629.3498	Hannan-Quinn criter.	31.79382
F-statistic	45.06800	Durbin-Watson stat	1.123746
Prob(F-statistic)	0.000000		

Залишки


20,000,000
16,000,000
12,000,000
8,000,000
4,000,000
0
6,000,000
4,000,000
2,000,000
0
-2,000,000
-4,000,000
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Residual
Actual
Fitted

Перевірка на нормальність



Series: Residuals	
Sample 2002Q1 2011Q4	
Observations 40	
Mean	-3.49e-11
Median	-204126.5
Maximum	4171075.
Minimum	-3659494.
Std. Dev.	1668551.
Skewness	0.590368
Kurtosis	3.696860
Jarque-Bera	3.132917
Probability	0.208783

Перевірка моделі на адекватність

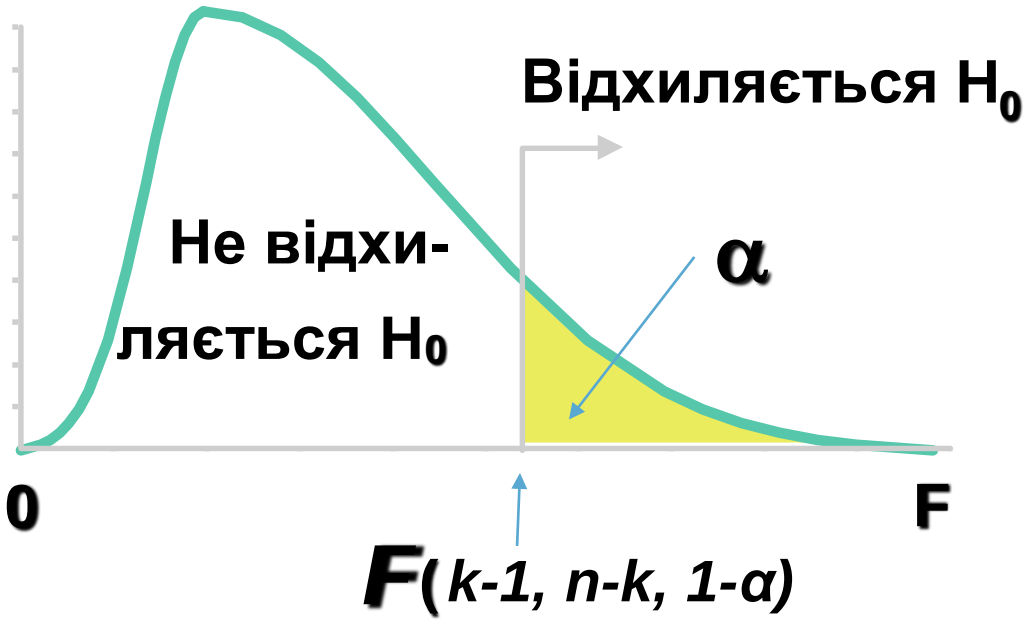
- Гіпотеза
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
    - *Немає зв'язку*
  - H_a: Хоча б один коефіцієнт не рівний 0
    - *Хоча б одна змінна впливає на Y*

$$F = \frac{RSS/(k-1)}{ESS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \overset{H_0}{\sim} F_{k,n-k}$$

Чому не використовують R^2?

Правило відхилення гіпотези

- Відхиляється H_0 на користь H_a, якщо F_{calc} попадає у зафарбовану область



- Відхилити H_0, якщо P-value = $P(F>F_{calc})<\alpha$

Приклад

Dependent Variable: TAX_ENT
Method: Least Squares
Date: 12/09/12 Time: 20:49
Sample: 2002Q1 2011Q4
Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3973770.	754540.7	5.266475	0.0000
@TREND	290525.1	24239.34	11.98568	0.0000
@SEAS(1)	-3627516.	791034.8	-4.585786	0.0001
@SEAS(2)	-2975920.	789175.7	-3.770922	0.0006
@SEAS(3)	-2032456.	788058.1	-2.579068	0.0143

R-squared	0.837415	Mean dependent var	7480035.
Adjusted R-squared	0.818834	S.D. dependent var	4138083.
S.E. of regression	1761318.	Akaike info criterion	31.71749
Sum squared resid	1.09E+14	Schwarz criterion	31.92860
Log likelihood	-629.3498	Hannan-Quinn criter.	31.79382
F-statistic	45.06800	Durbin-Watson stat	1.123746
Prob(F-statistic)	0.000000		

Тест на значення коефіцієнта

- Гіпотеза
  - H_0: $\beta_i = m$
  - H_a: $\beta_i \neq m$

Статистика

$$t = \frac{\hat{\beta}_i - m}{S_{\hat{\beta}_i}}$$

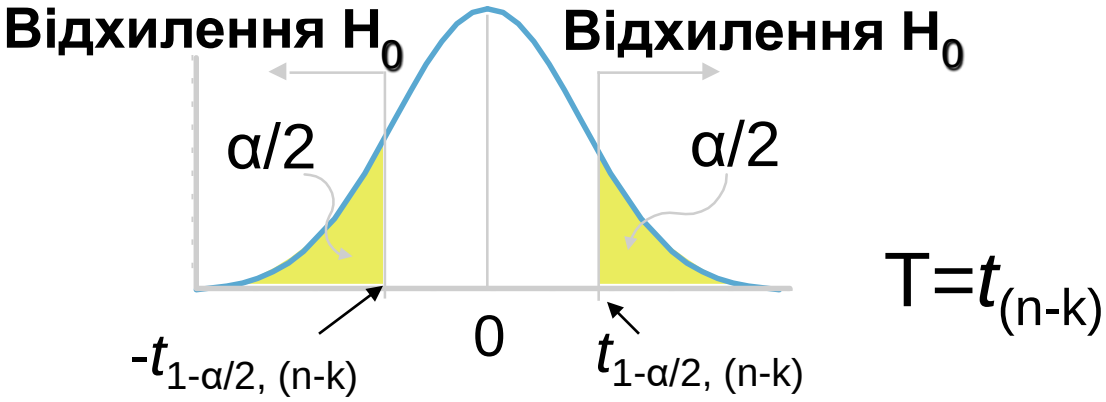
$$\text{де} \quad S_{\hat{\beta}_i} = \frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}}$$

$$S = \hat{\sigma} = \sqrt{\frac{ESS}{n-k}}$$

$$ESS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \sum_{i=1}^{k-1}\hat{\beta}_i X_i\right)\right]^2$$

Правило

- Відхилити H_0 на користь H_a, якщо t попадає в зафарбовану область



- Відхилити H_0, якщо P-value = $P(T>|t|) < \alpha$

Частковий випадок: значимість коефіцієнтів

- Гіпотеза
  - H_0: $\beta_i=0$
  - H_a: $\beta_i\neq0$

$$t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Приклад

Dependent Variable: TAX_ENT
Method: Least Squares
Date: 12/09/12 Time: 20:49
Sample: 2002Q1 2011Q4
Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3973770.	754540.7	5.266475	0.0000
@TREND	290525.1	24239.34	11.98568	0.0000
@SEAS(1)	-3627516.	791034.8	-4.585786	0.0001
@SEAS(2)	-2975920.	789175.7	-3.770922	0.0006
@SEAS(3)	-2032456.	788058.1	-2.579068	0.0143

R-squared	0.837415	Mean dependent var	7480035.
Adjusted R-squared	0.818834	S.D. dependent var	4138083.
S.E. of regression	1761318.	Akaike info criterion	31.71749
Sum squared resid	1.09E+14	Schwarz criterion	31.92860
Log likelihood	-629.3498	Hannan-Quinn criter.	31.79382
F-statistic	45.06800	Durbin-Watson stat	1.123746
Prob(F-statistic)	0.000000		

Тест Вальда (Wald test)

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_1 : \beta_1$ або β_2 або β_3 - не дорівнюють 0.

В матричній формі:

Гіпотеза: $Rb = r \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

Статистика:

$$F = \frac{(Rb - r)'\left[R \operatorname{cov}(b) R'\right]^{-1}(Rb - r)}{J}$$

Розрахувати F та порівняти з критичним F(J, n-k) з таблиці Фішера.

Мультиколінеарність

- Висока **кореляція** між змінними X
- Призводить до **нестабільності коефіцієнтів** моделі
- **Завжди** існує, питання лише у степені
- *Приклад*: Використання кількості кімнат та спалень як незалежних змінних в одній моделі

Визначення мультиколінеарності

- Критерій Фара-Глаубера
- VIF-тест

- Способи позбавлення
  - Збільшення кількості спостережень
  - Видалення однієї з корельованих змінних
  - Нормалізація змінних.

Приклад

$$\hat{s}_t = 0.4 + 0.8 y_t + 0.2 li_t - 0.1 si_t$$

$$(0.9)\ \ (1.2)\ \ \ (0.4)\ \ \ \ (0.1)$$

$$\overline{R}^2 = 0.98, (\text{стандартні похибки у дужках})$$

$(n = 60)$, де:

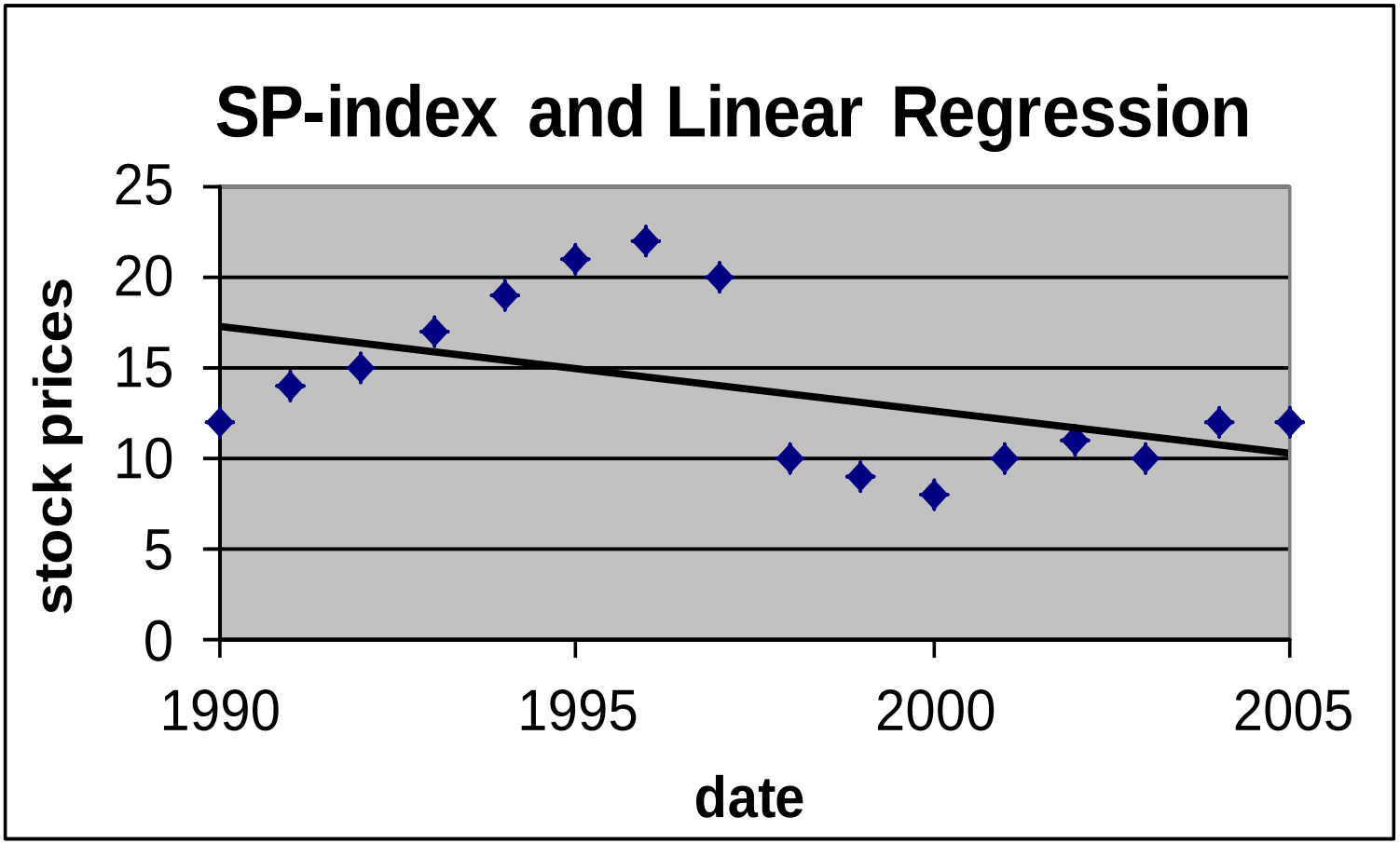
s_t – ціна акцій

y_t – ВВП

li_t – довгострокові ставки

si_t - короткострокові ставки

Тест на стабільність



Тест Чоу

- Перевіряє рівність коефіцієнтів у двох регресіях різних вибірок.

$$F = \frac{RSS_c - (RSS_1 + RSS_2)/k}{RSS_1 + RSS_2 / n - 2k} \sim F_{k,n-2k}$$

$$RSS_c - combined\_RSS$$

$$RSS_1 - pre-break\_RSS$$

$$RSS_2 - post-break\_RSS$$

Тести на автокореляцію

- **Durbin-Watson test** (перевіряє автокореляцію першого порядку)
- **Breusch-Godfrey Test** (перевіряє автокореляцію порядку q)

Статистика Дурбіна-Уотсона

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}, \text{ для n і k-1 d.f.}$$

- Явна автокореляція
- Автокореляція під питанням
- Відсутня автокореляція

Критерій Бройша-Годфрі

Процес авторегресії: AR(p)

$$\mu_t = \rho_1 \mu_{t-1} + \rho_2 \mu_{t-2} + \ldots + \rho_p \mu_{t-p} + \varepsilon_t$$

Гіпотеза

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_p = 0$$

Тест моделі:

$$\hat{\mu}_t = \delta_1 + \delta_2 X_{2t} + \ldots + \delta_k X_{kt} + \lambda_1 \hat{\mu}_{t-1} + \ldots + \lambda_p \hat{\mu}_{t-p} + \omega_t$$

Статистика

$$LM = (n\text{-}p) * R^2_{aux} \sim \chi^2_p$$

Тести на гетероскедастичність

- Типи тестів:
  - Регресійні: ***White test, Breusch–Pagan tests тощо***.
  - Загальні: ***the Goldfeld–Quandt test***

Тест Уайта

- Оцінка регресії Y відносно всіх змінних, розрахунок залишків $\varepsilon_1, \ldots, \varepsilon_n$
- Регресія ε_i^2 відносно константи, всіх змінних, їх квадратів та попарних добутків. Розрахунок R^2.
- Порівняння nR^2 з теоретичним Хі-квадрат з p степенями свободи.

Тест Голдфельда-Квондта – 1

Розділити n состережень на 2 групи розмірами n_1 та n_2

$$H_0 : \sigma_1^{\,2} = \sigma_2^{\,2} \quad \text{проти}$$

$$H_a : \sigma_1^{\,2} \neq \sigma_2^{\,2}$$

Оцінити регресію Y відносно
всіх змінних за вибіркою групи 1.

Оцінити регресію Y відносно
всіх змінних за вибіркою групи 2.

Тест Голдфельда-Квондта – 2

$$F_{calc} = \frac{\frac{RSS_1}{n_1 - k}}{\frac{RSS_2}{n_2 - k}} > 1$$

Порівняти F_{calc} з теоретичною F-статистикою з $(n_1 - k)$ та $(n_2 - k)$ степенями свободи.

Тест на специфікацію

$$F^{k}_{n-m-k+1} \sim \frac{\frac{R_1^2 - R_0^2}{k}}{\frac{1 - R_1^2}{n - m - k}}$$

Тест Рамсея (RESET)

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}_2 + \delta_2 \hat{y}_3 + \varepsilon$$

$H_0: \delta_1 = 0, \delta_2 = 0$

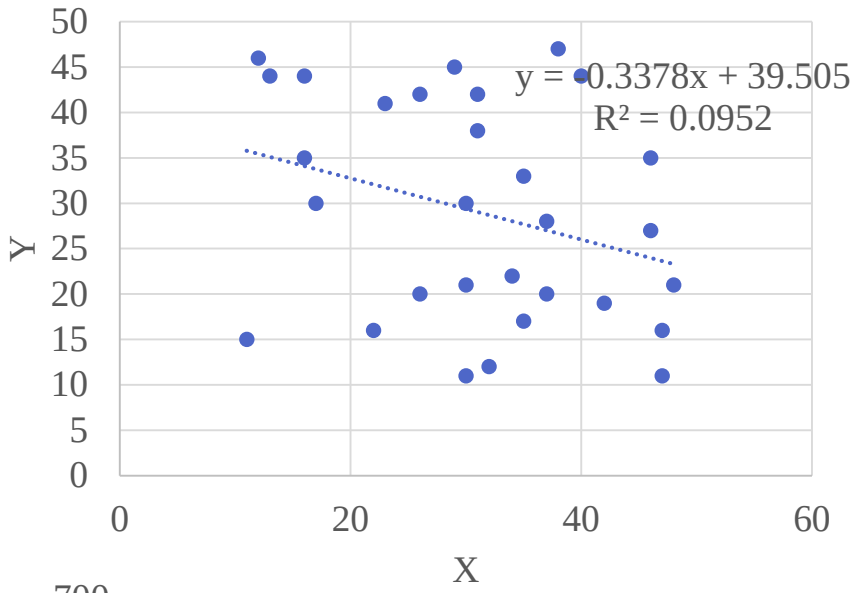
де $F \sim F_{2,n-k-3}$ або $LM \sim \chi^2(2)$.

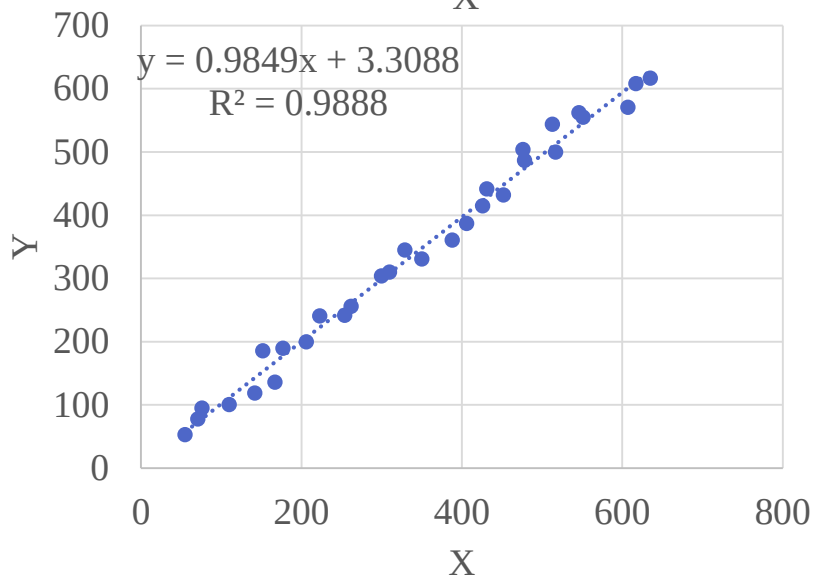
Стаціонарний процес

- Випадковий процес $X(t)$ називається стаціонарним, якщо всі його імовірнісні характеристики не міняються з часом t.

Чому стаціонарність важлива?

№	X	Y	T	X*	Y*
1	35	33	20	55	53
2	31	38	40	71	78
3	16	35	60	76	95
4	30	21	80	110	101
5	42	19	100	142	119
6	47	16	120	167	136
7	12	46	140	152	186
8	17	30	160	177	190
9	26	20	180	206	200
10	23	41	200	223	241
11	34	22	220	254	242
12	22	16	240	262	256
13	40	44	260	300	304
14	30	30	280	310	310
15	29	45	300	329	345
16	30	11	320	350	331
…					
30	35	17	600	635	617





Тест на стаціонарність

В загальному випадку модель перетворюємо до вигляду

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t$$

і перевіряємо гіпотезу

$$H^0: \quad \gamma = 0$$

$$H^1: \quad \gamma < 0$$

За допомогою стандартного економетричного апарату з односторонньою надійною областю (тест Дікі-Фулера).

Висновок

- Для аналізу бажано вибирати стаціонарні ряди даних
- Модель має відповідати основним характеристикам
- Модель має бути стійкою

Оформлення роботи з економетричними моделями

- Збільшений обсяг додатків
- Вимоги до опису даних
- Вимоги до подання результатів

Підготовка наукової роботи

- Вступ (Intoduction)
- Огляд літератури (Review of the Literature)
- Проблеми знаходження, вимірювання даних, їх опис (An Overview of the Data)
- Методологія дослідження (Methodology)
- Практичне застосування (Empirical Analysis)
- Дискусія (Discussion)
- Висновки (Concluding Remarks)

Проблеми знаходження, вимірювання даних, їх опис

- Конкретні засоби вимірювання, що дають можливість повторити експеримент
- Безпосередньо дані
- Зведена статистика (Summary Statistics):
  - Кількість спостережень
  - Середнє значення
  - Стандартне відхилення
  - Мінімальне значення
  - Максимальне значення
- Діаграми даних
- Висунення гіпотез про вплив показників

Методологія дослідження

- Запис конкретних гіпотез для перевірки
- Можлива трансформація даних
- Обґрунтування вибору моделі та її повний запис
- Методи оцінки та аналізу моделі
- Перевірка характеристик моделі

Практичне застосування

- Здійснення оцінки моделі та вказування її чисельних характеристик
- Доведення можливості використовувати її в дослідженні
- Порівняння різних моделей чи методів оцінки

Класифікація

- На основі значень коефіцієнтів регресій
- На основі рівня впливу (нормалізовані змінні, коефіцієнти еластичності)

Дискусія та висновки

- Аналіз можливості подальшого застосування моделі
- Обмеження щодо використання моделі на практиці
- Напрями удосконалення моделі
- Конкретні результати, отримані за допомогою аналізу моделі

Типові помилки

- Вступ не пов’язаний з моделюванням
- Огляд літератури не містить аналізу моделей за тематикою
- Слабкий огляд літератури чи аналіз застарілих джерел
- Відсутність конкретної моделі
- Математичні помилки у записі моделі чи її передумов
- Порушення вимог до оцінки моделі
- Відсутність конкретних гіпотез для перевірки
- Невідповідність моделі висунутій гіпотезі
- Відсутність даних про доведення адекватності моделі
- Висновки не пов’язані з результатами моделювання
- Порушення логічного змісту на користь моделі

Математичні помилки

- Використання нестаціонарних процесів для моделювання та формування висновків на їх основі
- Відсутність перевірки економетричних та статистичних якостей моделей
- Відсутність аналізу стійкості моделі у динаміці

Рекомендація

- У більшості випадків роботу читають фахівці, тому рекомендується отримати пораду спеціаліста з моделювання.

ОГЛЯД

Коефіцієнт кореляції

- Як правило, показує ступінь залежності між змінними.
- Вимірюється між -1 та 1

Лінійна регресія

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t} + \ldots + \beta_{k-1} x_{k-1t} + \varepsilon_t, t = \overline{1,n}$$

y_t - залежна змінна;

$x_{1t}, x_{2t}, \ldots, x_{k-1t}$ - незалежні змінні;

ε_t - збурення.

Припущення

- **Лінійність -** Y лінійно залежить від набору X.
- **Незалежність похибок** – збурення незалежні з X.
- **Гомоскедастичність** – дисперсія збурень є константою для всіх X.
- **Нормальність** – збурення мають нормальний розподіл.

Метод найменших квадратів

- Визначає коефіцієнти регресії, при яких різниця між реальними даними (Y) та прогнозними (Y-hat) буде найменшою:

$$L = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \rightarrow \min$$

Кроки для оцінки якості регресійної моделі

- Тест залишків на нормальність
- Тест значимості коефіцієнтів
- Тест адекватності моделі
- Тест на мультиколінеарність
- Тест на стійкість моделі
- Тест на автокореляцію залишків
- Тест на гетероскедастичність залишків
- Тест на специфікацію моделі
- Тест на стаціонарність даних

Стаціонарність даних

Зведення до стаціонарного процесу шляхом:

- Виділення тренду
- Взяття різниць

Спеціальні види даних

- Фіктивні змінні
  - Моделювання сезонності
  - Виділення тренду
- Лагові змінні

Типові помилки

- Вступ не пов’язаний з моделюванням
- Огляд літератури не містить аналізу моделей за тематикою
- Слабкий огляд літератури чи аналіз застарілих джерел
- Відсутність конкретної моделі
- Математичні помилки у записі моделі чи її передумов
- Порушення вимог до оцінки моделі
- Відсутність конкретних гіпотез для перевірки
- Невідповідність моделі висунутій гіпотезі
- Відсутність даних про доведення адекватності моделі
- Висновки не пов’язані з результатами моделювання

Математичні помилки

- Використання нестаціонарних процесів для моделювання та формування висновків на їх основі
- Відсутність перевірки економетричних та статистичних якостей моделей
- Відсутність аналізу стійкості моделі у динаміці

ПИТАННЯ?

Дякую за увагу!

Координатор ECTS Київського національного університету імені Тараса Шевченка

Член Національної команди експертів із реформування вищої освіти України

к.е.н., доц. А.В. Ставицький

a.stavytskyy@gmail.com

www.andriystav.cc.ua