E K O N O M E T R I A

LITERATURA:

Ekonometria, red. M. Krzysztofiak, PWE, Warszawa 1984

Ekonometria, S. Bartosiewicz, PWE, Warszawa 1978

Matematyczne Techniki Zarz□dzania, skrypt AGH, rozdzia□V

Statystyka w zarz□dzaniu, Aczel A. D., PWN, Warszawa 2000

Zespó□realizuj□cy przedmiot:

dr in□ Alicja Byrska R□pa□a - wyk□adowca

dr Izabela Stach

mgr in□. Mateusz Wiernek

Zaj□cia: Wyk□ady - 30 godz.

Laboratorium -15 godz.

□wiczenia - 15 godz.

Ekonometria - 1

Ekonometria — nauka o mierzeniu związków występujących między zjawiskami lub procesami ekonomicznymi a innymi zjawiskami (innymi zjawiskami ekonomicznymi, przyrodniczymi, technicznymi, demograficznymi i socjologicznymi) w celach poznawczych i dla prognozowania (Bartosiewicz S.)

— nauka zajmująca się ustalaniem, za pomocą metod matematyczno-statystycznych, ilościowych prawidłowości zachodzących w życiu gospodarczym

Specyficzne warunki prowadzenia badań ekonometrycznych
- brak możliwości powtórzenia eksperymentu (nie działają prawa statystyki matematycznej)
- zaostrzone kryteria matematyczne (n>100)
- trudności z danymi: dostępność, ilość, wiarygodność, porównywalność

NARZĘDZIEM BADAWCZYM EKONOMETRII JEST MODEL EKONOMETRYCZNY

– opis fragmentu ekonomicznej rzeczywistości, uwzględniający tylko istotne jej elementy;

– konstrukcja myślowa, która w uproszczony sposób przedstawia funkcjonowanie lub wzrost gospodarki lub jej części – def. dla ekonomii

– konstrukcja formalna, która za pomocą pewnego równania lub układu równań przedstawia zasadnicze powiązania występujące pomiędzy rozpatrywanymi zjawiskami ekonomicznymi (społeczno-ekonomicznymi).

Model ekonometryczny za pomocą równań przedstawia zależności występujące pomiędzy zmiennymi.

ELEMENTY MODELU:
- *Zmienne,*
- *Parametry*
- *Elementy losowe*

WŁAŚCIWOŚCI MODELU EKONOMETRYCZNEGO:

- spójność ekonomicznych i społecznych zjawisk i procesów;
- powiązanie (skorelowanie) za pomocą formalnych konstrukcji zjawisk społeczno-ekonomicznych wchodzących do wyodrębnionego systemu;
- mierzalność zjawisk;
- jednoznaczność formalna w zapisie, odczytywaniu i interpretacji uzyskanych wyników;
- jest zasadą, że każde równanie modelu przedstawia mechanizm kształtowania się jednej i tylko jednej zmiennej.

Jeżeli więc model ma przedstawić mechanizm kształtowania się jednej tylko zmiennej, będzie składał się z jednego równania.

Jeżeli natomiast celem modelu będzie opis mechanizmu kształtowania się *G* zmiennych, to musi składać się z *G* równań.

Terminologia

formalny podział zmiennych
- zmienna objaśniana (Y)
- zmienne objaśniające (X1, X2...)

ze względu na właściwości teoretyczne i praktyczne modelu
- zmienne endogeniczne
- zmienne egzogeniczne

ze względu na opóźnienie w czasie
- zmienne bieżąco współzależne
- zmienne z góry ustalone

KLASYFIKACJA MODELI EKONOMETRYCZNYCH

I. Klasyfikacja według wnoszonej informacji:

- modele przyczynowo-skutkowe

y — skutek
X_i — przyczyny

$$y = f(x_1, x_2, ..., x_{k-1})$$

- modele tendencji rozwojowej

y — analizowane zjawisko
t — czas

$$y = f(t)$$

II. Klasyfikacja według stopnia uwzględniania czasu:

- modele statyczne
- modele dynamiczne

III. Klasyfikacja według liniowości:

- modele liniowe
- modele nieliniowe (konieczna transformacja liniowa)

IV. Klasyfikacja według powiązania równań:

Jest to podział ze względu na zależności powiązania między różnymi zmiennymi endogenicznymi modelu.

Postać strukturalna modelu:

$$Y_{1,t} = \alpha_{11} Y_{1,t-1} + \alpha_{12} X_{1,t} + \alpha_1 + \xi_{1,t}$$

$$Y_{2,t} = \alpha_{21} Y_{1,t} + \alpha_2 + \xi_{2,t}$$

$$Y_{3,t} = \alpha_{31} Y_{1,t} + \alpha_{32} Y_{2,t} + \alpha_3 + \xi_{3,t}$$

Y_1, Y_2, Y_3 – zmienne endogeniczne
X_1 – zmienna egzogeniczna
a_{ij} – współczynnik przy j-tej zmiennej endogenicznej w i-tym równaniu

$$A = \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} Y1 & Y2 & Y3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ -\alpha_{21} & 1 & 0 \\ -\alpha_{31} & -\alpha_{32} & 1 \end{bmatrix} \end{array}$$

• modele proste: macierz A jest macierzą diagonalną i każdy element przekątnej równy 1 (macierz diagonalna: elementy poza przekątną=0);
nie występują związki między nie opóźnionymi zmiennymi endogenicznymi

$$Y_{1,t} = f(x_{1,t}, x_{2,t}, x_{3,t})$$
$$Y_{2,t} = f(x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t})$$
$$Y_{3,t} = f(x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t})$$

JEDNO RÓWNANIE LUB KILKA ODDZIELNYCH

• modele rekurencyjne: macierz A jest macierzą trójkątną; zmienna Y_{jt} zależeć może od nie opóźnionych zmiennych endogenicznych, których wskaźnik bieżący jest mniejszy od *j*, od zmiennych endogenicznych opóźnionych oraz od zmiennych egzogenicznych.

$$Y_{1,t} = f(x_{1,t}, x_{2,t}, x_{3,t})$$
$$Y_{2,t} = f(Y_{1,t}, Y_{2,t-1}, x_{1,t}, x_{2,t}, x_{3,t})$$
$$Y_{3,t} = f(Y_{2,t}, Y_{3,t-1}, x_{1,t}, x_{2,t}, x_{3,t})$$

• modele o równaniach współzależnych: macierz A dowolna; istnieje sprzężenie zwrotne między zmiennymi endogenicznymi w okresie *t*.

ETAPY BUDOWY MODELU EKONOMETRYCZNEGO

1. Sformułowanie problemu
    a. wybór zmiennych: $y, x_1, x_2,...$
    b. wybór postaci matematycznej modelu: liniowa, potęgowa,...
2. Zebranie danych statystycznych (różne źródła)
3. Selekcja zmiennych objaśniających (celem podziału na dwie grupy — nadające się do modelu i niepotrzebne w nim)
4. Estymacja parametrów modelu:
    a. parametrów strukturalnych: $a_0, a_1, a_2,...$
    b. parametrów stochastycznych: $s(a_i), s(y), R^2, R$
5. Weryfikacja modelu (przy użyciu hipotez i testów statystycznych)
6. Interpretacja modelu
    - wyciągnięcie wniosków dla celów zarządzania
    - sprzedanie go klientowi

ANALIZA REGRESJI I KORELACJI

- umożliwia badanie wpływu czynników mierzalnych, takich jak: zużycie materiałów, wielkość produkcji, ilość placówek wychowania pozaszkolnego, ilość spożywanego alkoholu itd.

- umożliwia ustalanie przyczyn zachowania się danego zjawiska:
- jest to bardzo popularna metoda, zgodna z naszą intuicją
- obliczenie parametrów modelu dokonuje się metodą najmniejszych kwadratów (MNK)
- stosuje się: estymację, testowanie hipotez, analizę wariancji itd.

Podstawowe pojęcia i terminy

KORELACJA — fakt powiązania, współzależności, związku zmiennych ze sobą

WSPÓŁCZYNNIK KORELACJI — liczba określająca siłę i kierunek tego związku

- współczynnik korelacji liniowej dwu zmiennych: r lub r_{xy}

Współczynnik *r* niesie dwie informacje poprzez swój znak i moduł

$$-1 \le \mathbf{r} \le 1$$

$$0 \le |\mathbf{r}| \le 1$$

Znak informuje o kierunku zależności

Moduł informuje o sile zależności

- współczynnik korelacji liniowej wielu zmiennych (korelacji wielokrotnej lub wielorakiej): R

$$0 \le \mathbf{R} \le 1$$

Interpretacja:

- im wyższa wartość *R*, tym silniejsza współzależność (*R*=0: brak korelacji, *R*=1: zależność funkcyjna, nie ma składnika losowego)
- *R* określa siłę powiązania zmiennej *Y* z wszystkimi zmiennymi X_i, bez względu na to jak poszczególne z nich są skorelowane z *Y*
- współczynnik korelacji cząstkowej dwu zmiennych $\mathbf{r_{x_i x_j}}$ $\mathbf{r_{y,x_i}}$

REGRESJA — statystyczna metoda modelowania związków między zmiennymi; opisuje ją funkcja odzwierciedlająca powiązanie zmiennych (czynników)

- w mowie potocznej regresja to cofanie się, spadek, zanik
- skąd się wzięło to słowo w statystyce?

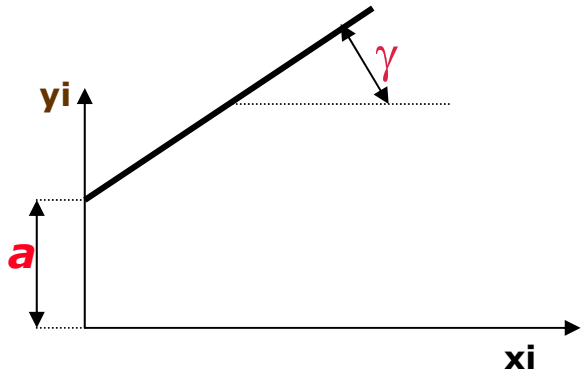
WSPÓŁCZYNNIK REGRESJI — liczba stojąca przy każdej zmiennej *X*, określająca jej wpływ na zmienną *Y*

$$\mathbf{y_i = a + bx_i + \xi}$$

a — wyraz wolny (stała), współrzędna punktu przecięcia z osią *Y*

b — współczynnik regresji, tangens kąta γ nachylenia prostej

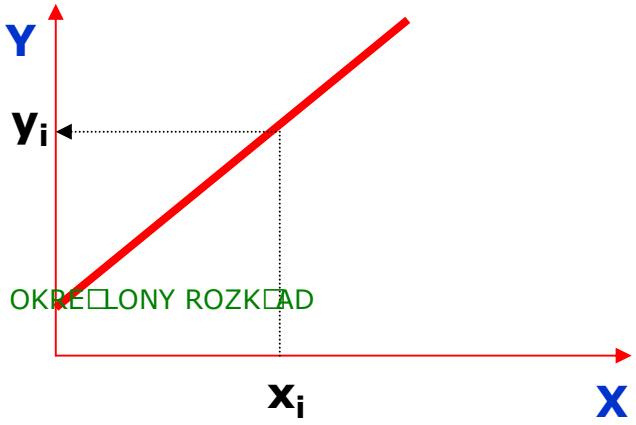
ξ — składnik losowy $N(0, \sigma^2)$



Trzy rodzaje związków pomiędzy Y i X

- związek funkcyjny (deterministyczny) $y_i = a + bx_i$

KAŻDEJ WARTOŚCI x_i ODPOWIADA JEDNA I TYLKO JEDNA WARTOŚĆ y_i



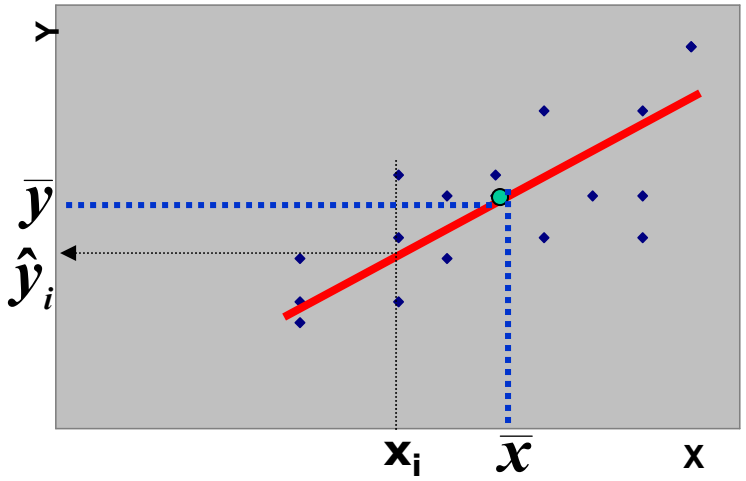
- związek stochastyczny (losowy), probabilistyczny

KAŻDEJ WARTOŚCI x_i ODPOWIADA CAŁY ZBIÓR WARTOŚCI y_i TWORZĄCYCH OKREŚLONY ROZKŁAD

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

- związek statystyczny

$$\hat{y}_i = a + bx_i + \xi$$



$\hat{y}_i$ — średnia rozkładu dla ustalonej wartości x_i

ξ — obrazuje rozrzut

$\bar{x}, \bar{y}$ — środek ciężkości zbioru

Funkcja regresji I i II rodzaju

- regresja I rodzaju dotyczy populacji (jest nieznana) $\mathbf{Y} = \alpha_0 + \alpha_1\mathbf{X}_1 + \alpha_2\mathbf{X}_2 + \ldots + \varepsilon$

- regresja II rodzaju dotyczy próbki (jest znana) $\mathbf{y} = \mathbf{a}_0 + \mathbf{a}_1\mathbf{x}_1 + \mathbf{a}_2\mathbf{x}_2 + \ldots + \xi$

Współczynniki regresji to α_i oraz a_i; tak jak przy estymacji innych parametrów mamy to do czynienia z estymatorami, ich odchyleniami standardowymi (czyli błędami oszacowania) oraz z wartościami oszacowanymi.

Regresja liniowa I rodzaju

Załóżmy, że mamy dany rozkład zmiennej losowej dwuwymiarowej.

Przyjmuje ona wartości $(x_i; y_j)$ z prawdopodobieństwem P_{ij},

a odpowiednie rozkłady brzegowe mają postać $h(x_i)$ i $g(y_j)$.

Zmienna losowa dwuwymiarowa

Tablica dwudzielna

	y_1	y_2	y_i	y_m	Suma Rozkład brzegowy $h(x_i)$
x_1	P_{11}	P_{12}	P_{1j}	P_{1m}	P_1
x_2	P_{21}	P_{22}	P_{2j}	P_{2m}	P_2
x_i	P_{i1}	P_{i2}	P_{ij}	P_{im}	P_i
x_n	P_{n1}	P_{n2}	P_{nj}	P_{nm}	P_n
Suma Rozkład brzegowy $g(y_j)$	$P_{\cdot 1}$	$P_{\cdot 2}$	P_j	P_m	1

Dla rozkładu warunkowego zmiennej losowej Y względem X wartość oczekiwana dla rozkładu dyskretnego i ciągłego dana jest wzorem:

$$E(Y/X=x_k)=\sum_j y_j \cdot P(Y/X=x_k)=\sum_j y_j \cdot \frac{P_{ij}}{h(x_k)}$$

$$E(Y/X=x)=\int_{-\infty}^{+\infty} y \cdot f(y/x)dy=\int_{-\infty}^{+\infty} y \cdot \frac{f(x,y)}{h(x)}dy$$

Definicja

Zbiór punktów (x,y) spełniających równanie: y=E(Y/X=x) nazywamy linią regresji I rodzaju zmiennej losowej Y względem X.

RÓWNANIE REGRESJI
(model deterministyczny)

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

RÓWNANIE REGRESJI
(model probabilistyczny)

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

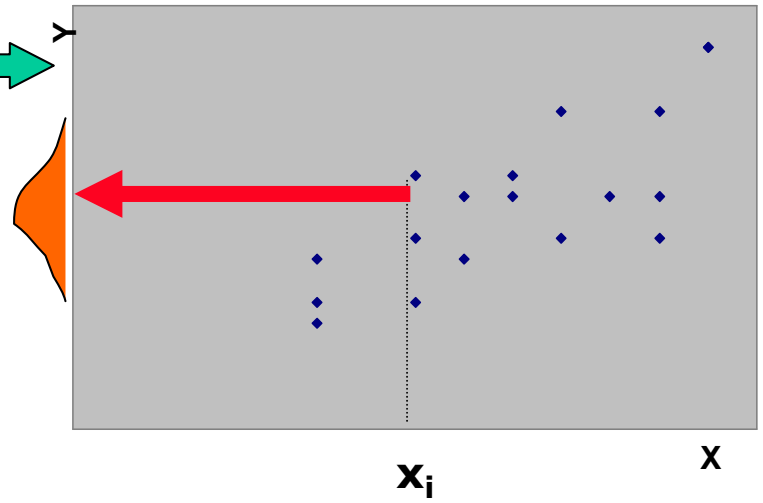
Regresja liniowa II rodzaju

KA□DEJ WARTO□CI x_i ODPOWIADA CA□Y ZBIÓR WARTO□CI y_i TWORZ□CYCH OKRE□LONY ROZK□AD

Obserwacja rzeczywisto□ci

DANE

Lp.	x_i	y_i
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
.....		



Je□eli ten rozk□ad jest normalny to zale□no□□Y(X) jest liniowa.

$$y_i = a + bx_i + \xi$$

Wydruk komputerowy równania regresji

Zmienna (czynnik)	Warto□□ oszacowana	B□□d oszacowania	Statystyka t_{obl}	Rzeczywisty po-ziom istotno□ci P
Wyraz wolny	a_0	$s(a_0)$	$t(a_0)$	$P(a_0)$
Czynnik X_1	a_1	$s(a_1)$	$t(a_1)$	$P(a_1)$
Czynnik X_2	a_2	$s(a_2)$	$t(a_2)$	$P(a_2)$
Czynnik X_3	a_3	$s(a_3)$	$t(a_3)$	$P(a_3)$
Wspó□czynniki: determinacji R^2, zbie□no□ci φ^2, b□□d resztowy s(y) i inne				

Pe□ny zapis równania regresji

$$\hat{y}_i = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \xi_i$$

$$\hat{y}_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + a_3 x_{3i} + \xi \qquad R^2(R)$$

$$s(a_0)\ \ s(a_1)\ \ s(a_2)\ \ s(a_3)\ \ s(y) \qquad \varphi^2$$

Y — zmienna zale□na, skutek, zmienna objaśniana, endogeniczna

y_i — zaobserwowane warto□ci zmiennej zale□nej dla obiektów (jednostka próby)

X_k — zmienne niezale□ne, przyczyny, zmienne objaśniaj□ce - egzogeniczne

x_{ki} — zaobserwowane warto□ci zmiennych niezale□nych

parametry strukturalne i stochastyczne

a_0 — oszacowana warto□□wyrazu wolnego, estymator

a_i... — oszacowane warto□ci wspó□czynników regresji; okre□laj¹ wp□yw poszczególnych zmiennych X_i na zmienn¹ Y

ξ — sk□adnik losowy, reprezentuj□cy rozrzut punktów wokó□p□aszczyz-ny regresji; sk□adnik ten jest zmienn¹ losow¹, jego warto□ci nazywaj¹ si□reszty $e_i = y_i - \hat{y}_i$

a jego rozk□ad jest rozk□adem normalnym o $E(\xi)=0$ i $V(\xi)=s^2(y)$

$s(a_0)$ — b□□d oszacowania wyrazu wolnego; s□u□y do budowy przedzia□u ufno□ci dla nieznanej warto□ci wyrazu wolnego α_0 dla populacji oraz do weryfikacji istotno□ci α_0 (H_0: $\alpha_0=0$)

$s(a_i)$ — b□□dy oszacowania wspó□czynników regresji; s□u¿¹ do budowy przedzia□u ufno□ci dla nieznanych warto□ci α_i wspó□czynników regresji dla populacji oraz do weryfikacji ich istotno□ci (H_0: $\alpha_i=0$)

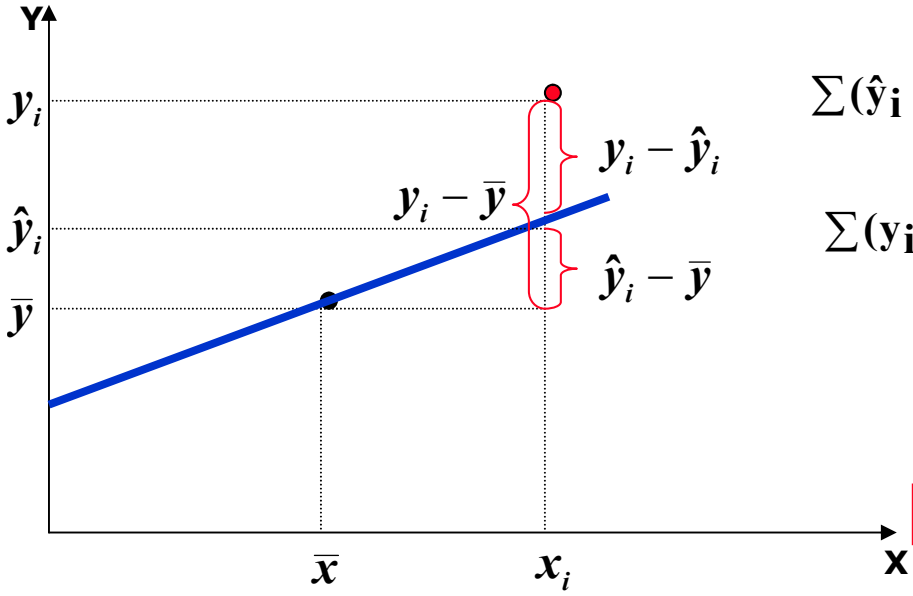
$s(y)$ — b□□d resztowy; odchylenie standardowe sk□adnika losowego ξ; okre□la □redni¹ wielko□□reszty e_i

$R^2(r^2)$— wspó□czynnik determinacji; okre□la jaka cz□□ zmienno□ci ca□ko-witej SSTO zosta□a wyja□niona przez równanie regresji

φ^2 — wspó□czynnik zbie□no□ci (zgodno□ci); okre□la jaka cz□□ zmien-no□ci ca□kowitej SSTO nie zosta□a wyja□niona przez równanie regresji

$\sum(y_i - \bar{y})^2$ = SSTO (zmienno□□ca□kowita)

$\sum(y_i - \bar{y})^2$ = SSTO (zmienno□□ca□kowita)



$\sum(\hat{y}_i - \bar{y})^2$ = SSTR (zmienno□□wyja□niona)

$\sum(y_i - \hat{y}_i)^2$ = SSE (zmienno□□niewyja□niona)

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

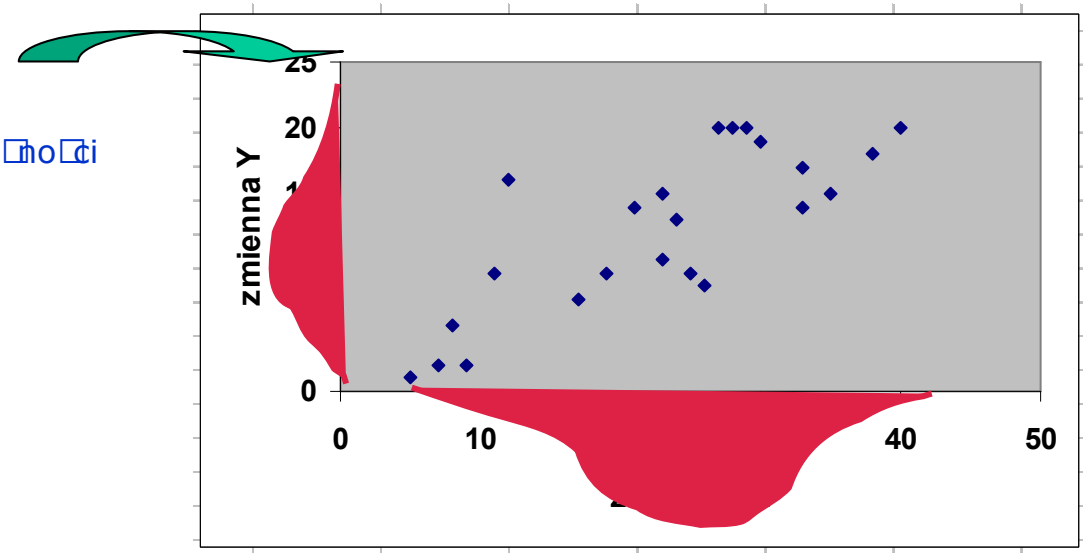
SSTO = SSTR + SSE

$$R^2 = \frac{SSTR}{SSTO} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \qquad \varphi^2 = \frac{SSE}{SSTO} = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad s(y) = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-k}}$$

□ród□o Zmienno□ci	*Liczba stopni swobody*	*Suma kwadratów*	*□redni kwadrat*	*Statystyka F*
Model (czynniki)	k–1	SSTR	MSTR	$F_{obl} = \frac{MSTR}{MSE}$
B□□d (reszta)	n–k	SSE	MSE	
Razem	n–1	SSTO		

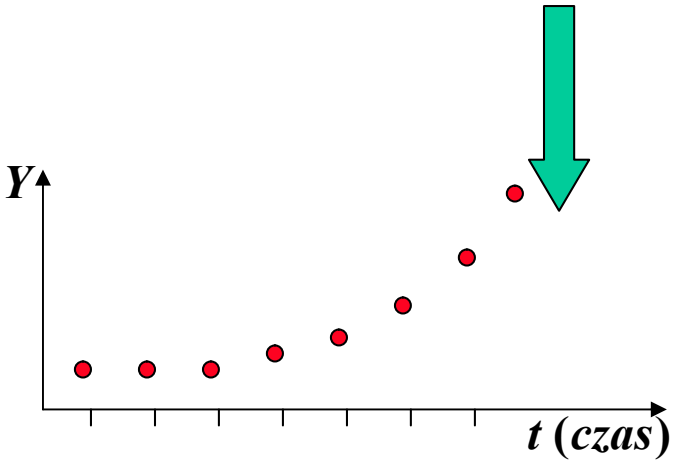
Regresja krzywoliniowa

Kiedy wystpuje regresja liniowa? — gdy obie zmienne maj rozkad normalny!

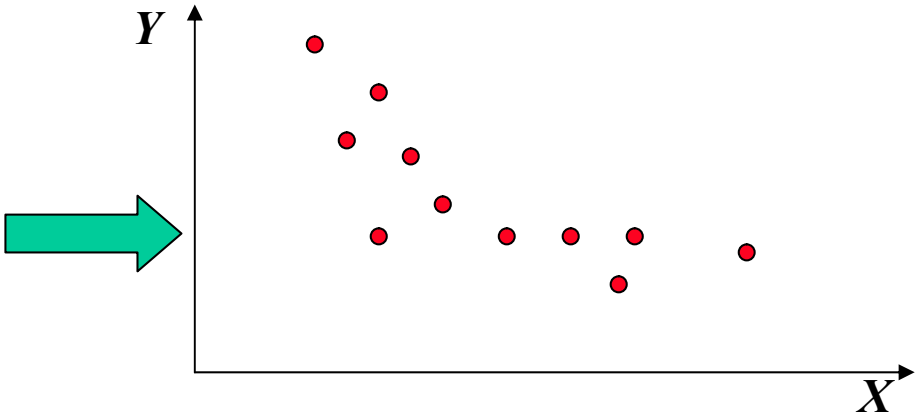


W wielu przypadkach dane ukadaj si w zalenoci nieliniowe:

- gdy maj posta szeregu czasowego



- gdy dane przekrojowe ukadaj si w smug nieliniow



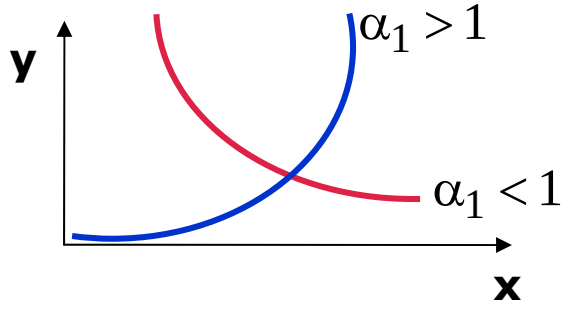
- gdy krzywoliniowa funkcja wielu zmiennych lepiej opisuje rzeczywisto ni funkcja liniowa; (tego nie wida, która lepsza mona pozna tylko po R^2)

Do opisu takich zjawisk stosujemy rozmaite funkcje krzywoliniowe:

1. proste funkcje (rosnące lub malejące) dwu zmiennych:

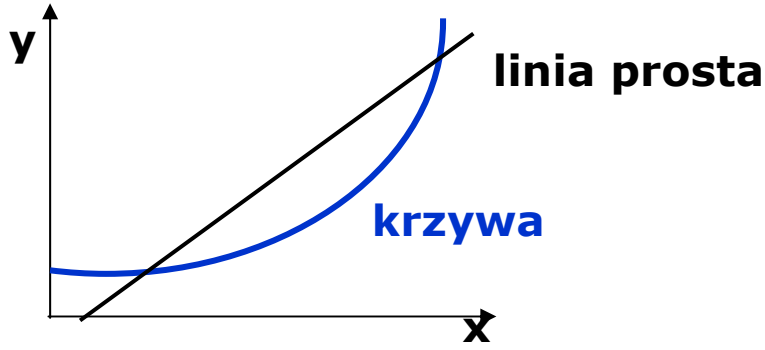
- wykładnicze $y = \alpha_0 \cdot e^{\alpha_1 \cdot x} \cdot \xi$

- potęgowe itp. $y = \alpha_0 \cdot x^{\alpha_1} \cdot \xi$



2. wielomiany różnego stopnia (ich fragmenty)

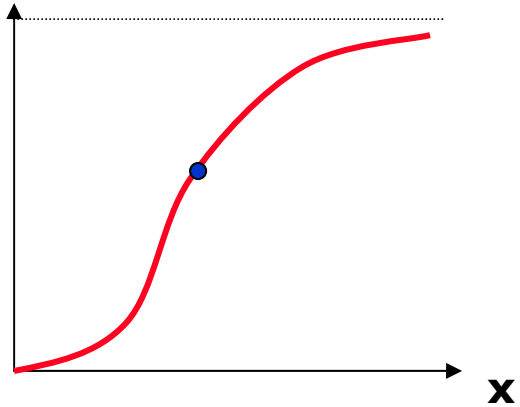
$$\hat{y} = a_0 + a_1 x + a_2 x^2 \quad (a_2 > 0)$$



3. funkcje bardziej złożone:

- krzywe logistyczne

$$y = \frac{e^{(\alpha_0 + \alpha_1 \cdot x)}}{1 + e^{(\alpha_0 + \alpha_1 \cdot x)}}$$



- funkcje potęgowe wielu zmiennych

$$y = a_0 x_1^{a_1} x_2^{a_2} x_3^{a_3} \ldots \varepsilon$$

- funkcje wykładnicze wielu zmiennych

$$y = e^{\alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2} \cdot \xi$$

ABY MOŻNA BYŁO STOSOWAĆ METODĘ NAJMNIEJSZYCH KWADRATÓW, FUNKCJE TE MUSZĄ BYĆ SPROWADZONE DO POSTACI LINIOWEJ

1. Uliniowienie przez podstawianie np.

$$\ln y = \alpha_0 + \alpha_1 \cdot \ln x + \xi$$

$$\ln y = y'; \qquad \ln x = x'$$

$$y' = \alpha_0 + \alpha_1 \cdot x' + \xi$$

2. Transformacja logarytmiczna

$$y_i = a_0 x_1^{a_1} x_2^{a_2} x_3^{a_3} \ldots \varepsilon \quad \Rightarrow \quad \ln y_i = \ln a_0 + a_1 \ln x_1 + a_2 \ln x_2 + a_3 \ln x_3 + \ldots + \xi$$

Kolejność czynności przy estymacji funkcji regresji krzywoliniowej:
1. zebranie danych empirycznych
2. dobranie modelu (funkcji nieliniowej)
3. transformacja modelu do liniowego (logarytmowanie — transformata)
4. przeliczenie danych na układ liniowy (robi to komputer)
5. oszacowanie równania regresji liniowej
6. retransformacja do postaci pierwotnej (odlogarytmowanie)

Retransformacji podlegają tylko parametry strukturalne, natomiast wszystkie parametry stochastyczne dotyczą tylko transformaty

Metody estymacji równania regresji

- klasyczna metoda najmniejszych kwadratów (KMNK) w wielu wariantach obliczeniowych
- podwójna MNK
- regresje specjalne: grzbietowa *(ridge regression),* odporna (*robust)* itd.
- metoda najwi□kszej wiarygodno□ci

Klasyczna metoda najmniejszych kwadratów (KMNK)

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

$$\hat{y}_i = a + bx_i \qquad SSE = \min \sum (y_i - a - bx_i)^2$$

$$bn + a\sum x_i = \sum y_i$$

$$b\sum x_i + a\sum x_i^2 = \sum x_i y_i$$

Wersja 2. Metoda „sigma prim” (uproszczona reguła obliczeniowa różnicy kwadratów)

$$\Sigma' y^2 = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

Wersja 3. Metoda mnożników Gaussa, posługuje się formularzami obliczeniowymi opartymi o wartości „sigma prim”.

Wersja 4. Metoda przekształceń Jordana

Wersja 5. Metoda macierzowa

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \ldots + a_{k-1} x_{k-1} + \xi$$

$$a = (X^T X)^{-1} X^T y$$

$$a = \begin{bmatrix} a_0 \\ a_1 \\ \ldots \\ a_{k-1} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdot & x_{k-1,1} \\ 1 & x_{12} & \cdot & x_{k-1,2} \\ \cdot & \ldots & \cdot & \ldots \\ 1 & x_{1n} & \cdot & x_{k-1,n} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_n \end{bmatrix}$$

X^TX — współczynniki układu r. n.

X^Ty — prawe strony układu r. n.

$$s^2 = \frac{1}{n-k}[y^T y - (X^T y)^T a]$$

$$D^2(a) = s^2 (X^T X)^{-1}$$

na głównej przekątnej tej macierzy znajdują się wariancje $s^2(a_0)$, $s^2(a_1)$...

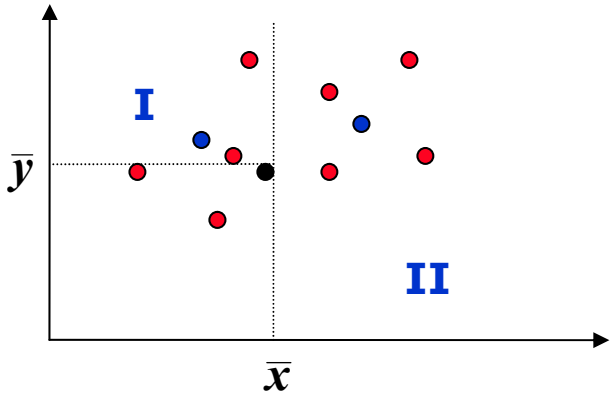
Wersja 6. Metoda uproszczona Hellwiga

Dzielimy zbiór na 2 podzbiory i wyznaczamy ich środki ciężkości

$\bar{x}_I, \bar{y}_I$

$\bar{x}_{II}, \bar{y}_{II}$

po czym budujemy prostą przechodzącą przez te punkty



ANALIZA REGRESJI LINIOWEJ

Dla poznania rzeczywistości często konieczne jest badanie kilku zmiennych losowych równocześnie, wraz ze zwróceniem uwagi na ich wzajemne powiązania.

$$y_i = a + bx_i + \xi$$

Linia regresji II rodzaju zmiennej Y względem X

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Metoda najmniejszych kwadratów MNK pozwala wyznaczyć współczynniki a i b prostej, która najlepiej pasuje do zmierzonych punktów.

Wzory na obliczanie wyrazu wolnego *a* i współczynnika regresji *b* KMNK:

$$b = \frac{S_{xy}}{S_{xx}} \quad \Rightarrow \quad S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \qquad S_{xx} = \sum_i (x_i - \bar{x})^2$$

$$\Rightarrow \quad b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Uproszczona reguła obliczania sumy kwadratów odchyleń SS

$$S_{xy} = \sum_i x_i y_i - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n}$$

$$S_{xx} = \sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}$$

KAŻDEJ WARTOŚCI x_i ODPOWIADA CAŁY ZBIÓR WARTOŚCI y_i TWORZĄCYCH OKREŚLONY ROZKŁAD a parametrami tego rozkładu są E(Y/X) i wariancja σ^2

Estymatorem wariancji σ^2 jest s^2

$$s^2 = \frac{SSE}{n-2} \qquad SSE = S_{yy} - b \cdot S_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2 \quad \Rightarrow \quad S_{yy} = \sum_i y_i^2 - \frac{\left(\sum_i y_i\right)^2}{n}$$

Estymator współczynnika regresji

$$E(b) = \beta_1 \qquad \sigma_b^2 = \frac{\sigma^2}{S_{xx}} \quad \Rightarrow \quad s_b^2 = \frac{s^2}{S_{xx}}$$

Analiza współczynnika regresji

$$P(b - t_{\alpha/2;n-2} \cdot s_b < \beta_1 < b + t_{\alpha/2;n-2} \cdot s_b) = 1 - \alpha$$

- weryfikujemy hipotezę H_0: $\beta_1 = 0$ wobec H_1: $\beta_1 \neq 0$

Test Studenta (t)

$$t = \frac{b}{s / \sqrt{S_{xx}}}$$

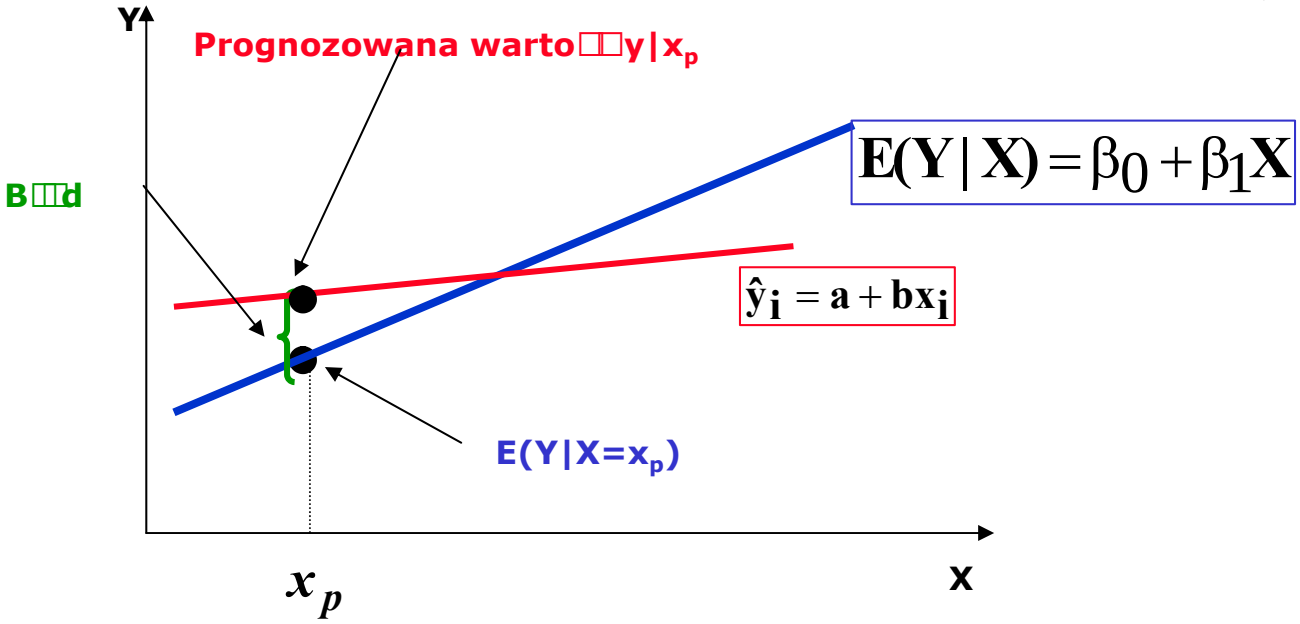
Estymacja E(Y/X) warto□ci oczekiwanej y dla danej warto□ci X $\mathbf{E(Y \mid X = x_p) = \beta_0 + \beta_1 X}$

$$\sigma^2_{\hat{y}_p} = \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]$$

estymator

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]$$

$$\mathbf{s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}}$$



- przedzia□ufno□ci dla $\mathbf{E(Y|X=x_p)}$

$$\mathbf{P(y_p - t_{\alpha/2;n-2} \cdot s_{y_p} < E(Y/x_p) < y_p + t_{\alpha/2;n-2} \cdot s_{y_p}) = 1 - \alpha}$$

- **przedział ufności dla prognozy y_p**

$$\sigma^2_{y-\hat{y}_p} = \sigma^2 + \sigma^2_{\hat{y}_p} = \sigma^2\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}\right]$$

Z zależności:

$$\sigma^2_{\hat{y}_p} = \sigma^2\left[\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}\right] \quad \Rightarrow \quad s_{\hat{y}_p} = s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$$

$$\sigma^2 \quad \Rightarrow \quad s^2 = \frac{SSE}{n-2}$$

obliczamy:

$$s_{y-\hat{y}_p} = \sqrt{s^2\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}\right]}$$

$$P(\hat{y}_p - t_{\alpha/2;n-2}\cdot s_{y-\hat{y}_p} < y/x_p < \hat{y}_p + t_{\alpha/2;n-2}\cdot s_{y-\hat{y}_p}) = 1-\alpha$$

Wydruk komputerowy równania regresji

Zmienna (czynnik)	Warto□□ oszacowana	B□□d oszacowania	Statystyka t_{obl}	Rzeczywisty poziom istotno□ci P
Wyraz wolny Czynnik X	a b	s(a) s(b)	t(a) t(b)	P(a) P(b)
Wspó□czynniki: korelacji liniowej Persona *r*, determinacji r^2, zbie□no□ci φ^2, b□□d resztowy *s(y)* i inne				

Pe□ny zapis równania regresji liniowej

$$\hat{y}_i = a + bx + \xi$$

$$s(a) \quad s(b)$$

r

parametry strukturalne i stochastyczne

Y — zmienna zale□na, zmienna-skutek, zmienna objaśniana

y_i — zaobserwowane warto□ci zmiennej zale□nej dla jednostek próbki

X — zmienna niezale□ne, zmienna-przyczyny, zmienna objaśniaj□ca

x_i — zaobserwowane warto□ci zmiennej niezale□nej

a — oszacowana warto□□wyrazu wolnego

b — oszacowana wartość współczynnika regresji; określa wpływ zmiennej X na zmienną Y

ξ — składnik losowy, reprezentujący rozrzut punktów wokół prostej regresji; składnik ten jest zmienną losową, jego wartości nazywają się reszty

$$e_i = y_i - \hat{y}_i$$

a jego rozkład jest rozkładem normalnym o $E(\xi)=0$ i $V(\xi)=s^2(y)$

s(a) — błąd oszacowania wyrazu wolnego; służy do budowy przedziału ufności dla nieznanej wartości wyrazu wolnego dla populacji oraz do weryfikacji jego istotności (H_0: $\beta_0 = 0$)

s(b) — błąd oszacowania współczynnika regresji; służy do budowy przedziału ufności dla nieznanej wartości β_1 współczynnika regresji dla populacji oraz do weryfikacji jego istotności (H_0: $\beta_1 = 0$)

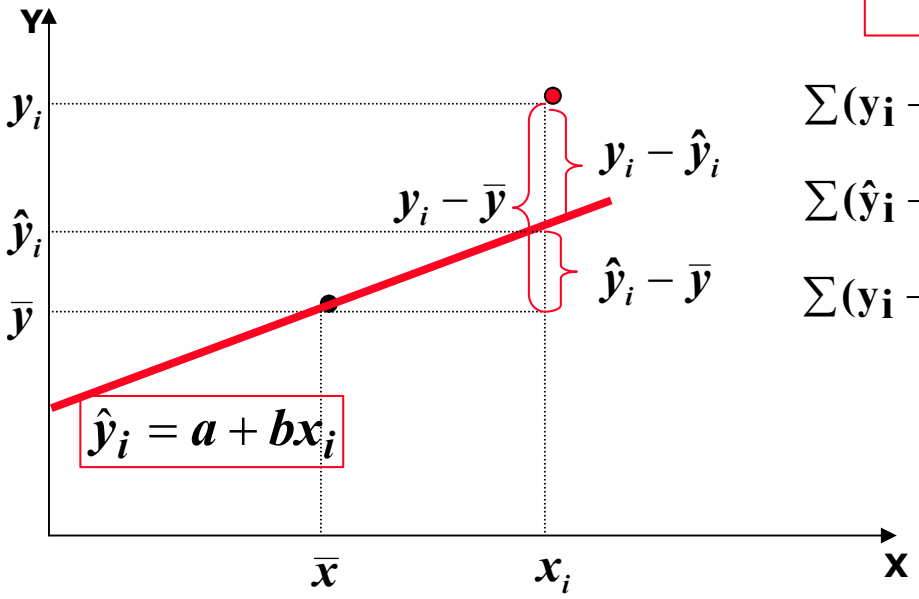
s(y) — błąd resztowy; jest odchyleniem standardowym składnika losowego ξ; określa średnią wielkość reszty e_i

r^2 — współczynnik determinacji; określa jaka część zmienności całkowitej SSTO została wyjaśniona przez równanie regresji

$$y_i = a + bx_i + \xi$$

φ^2 — współczynnik zbieżności (zgodności); określa jaka część zmienności całkowitej SSTO nie została wyjaśniona przez równanie regresji

ANALIZA WARIANCJI



SSTO = SSTR + SSE

$\sum(y_i - \bar{y})^2$ = SSTO (zmienno□□ca□kowita)

$\sum(\hat{y}_i - \bar{y})^2$ = SSTR (zmienno□□wyja□niona)

$\sum(y_i - \hat{y}_i)^2$ = SSE (zmienno□□niewyja□niona)

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

□ród□o Zmienno□ci	*Liczba stopni swobody*	*Suma kwadratów*	*□redni kwadrat*	*Statystyka F*
Model (czynniki)	2–1	SSTR	MSTR	$F_{obl} = \frac{MSTR}{MSE}$
B□□d (reszta)	n–2	SSE	MSE	
Razem	n–1	SSTO		

Przykład

Wpływ wydatków na reklamę na wielkość sprzedaży

Miesiąc	Wydatki na reklamę (X) (mln zł)	Wartość sprzedaży (Y) (mln zł)
1.	1,2	101
2.	0,8	92
3.	1,0	110
4.	1,3	120
5.	0,7	90
6.	0,8	82
7.	1,0	93
8.	0,6	75
9.	0,9	91
10.	1,1	105

Regression Analysis - Linear model: Y = a+bX

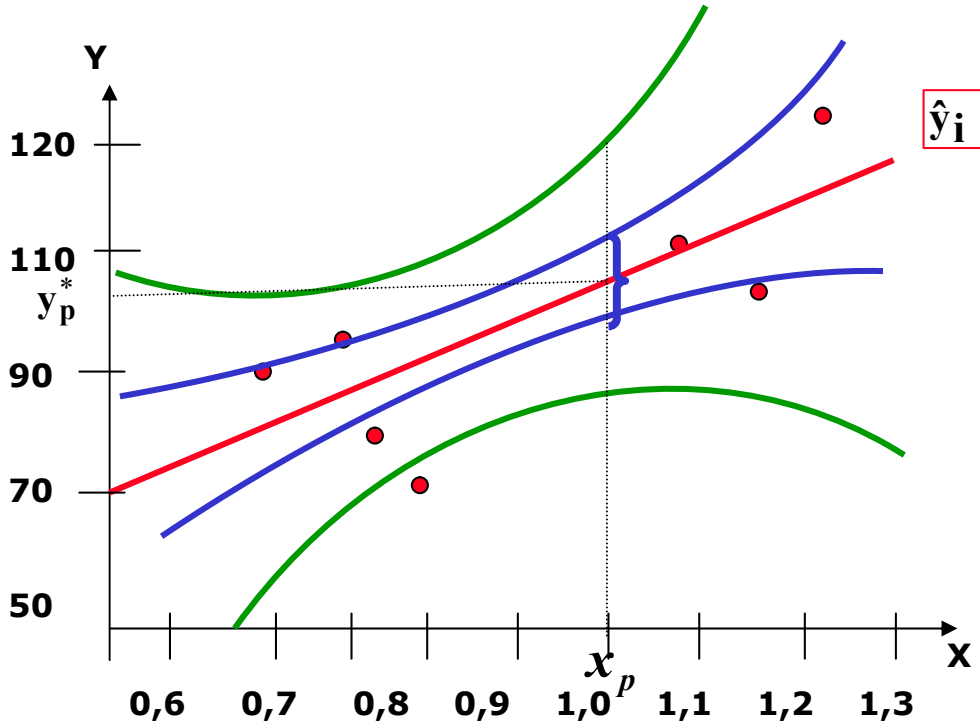
Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	46.4865	9.8846	4.7029	0.00154
Slope	52.5676	10.2609	5.1231	0.00090

Correlation Coefficient = 0.8754 R-squared = 76.64 (%)
Stnd. Error of Est. = 6.83715

$$\hat{y}_i = 46{,}49 + 52{,}57x_i + \xi \quad r=0{,}88$$

9,88 10,26 6,84

Estymacja E(y/ x=1,0) wartości oczekiwanej y dla x_p=1,0



$$\hat{y}_i = 46{,}49 + 52{,}57x$$

$$P(93{,}88 < E(Y / x = 1{,}0) < 104{,}24) = 0{,}95$$

Prognozowanie wartości y dla x=1,0

Prognoza punktowa: $\hat{y} = 46{,}49 + (52{,}57)(1{,}0) = 99{,}06$

Prognoza przedziałowa: $P(82{,}46 < \hat{y} / x = 1{,}0 < 115{,}66) = 0{,}95$

Badanie parametrów strukturalnych modelu

Przedział ufności dla współczynnika regresji

$$P(28,90 < \beta_1 < 76,24) = 0,95$$

Interpretacja:

Zmiana miesięcznych wydatków na reklamę o jedną jednostkę (1 mln zł) spowoduje zmianę wielkości sprzedaży w przedziale od 28,9 do 76,24 mln zł.

ANALIZA WARIANCJI

Źródło Zmienności	Liczba stopni swobody	Suma kwadratów	Średni kwadrat	Statystyka F
Model (czynniki)	1	1226,9	1226,9	$F_{obl} = \frac{MSTR}{MSE} = 26,25$
Błąd (reszta)	8	374,0	46,7	
Razem	9	1600,9		

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

$$t^2 = \left(\frac{b}{s/\sqrt{S_{xx}}}\right)^2 = \frac{MSTr}{MSE} = F$$

$F_{1;8;0,025} = 7,57$

Wniosek: 77% zmienności *y* wyjaśnia wyestymowany model