- związek stochastyczny (losowy), probabilistyczny $Y = \alpha_0 + \alpha_1 X + \xi$

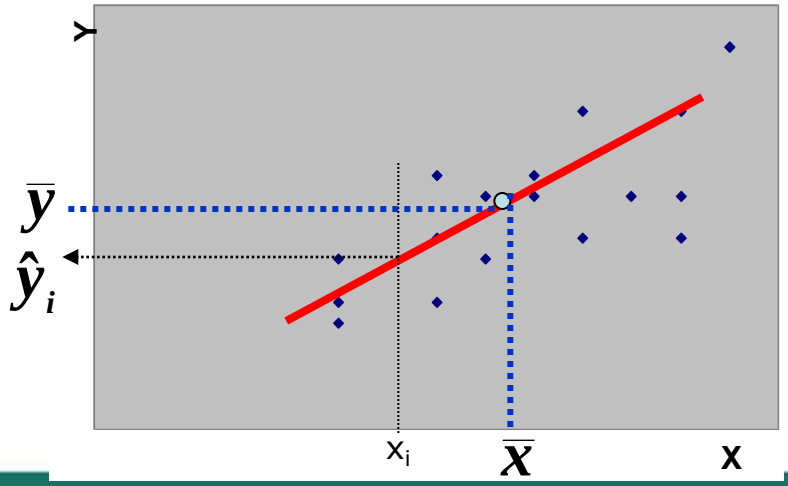
KAŻDEJ WARTOŚCI x_i ODPOWIADA CAŁY ZBIÓR WARTOŚCI y_i TWORZĄCYCH OKREŚLONY ROZKŁAD

- związek statystyczny $\hat{y}_i = a_0 + a_1 x_i + \xi_i$

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

ξ — obrazuje rozrzut

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



$$y_i = \alpha_0 + \alpha_1 \cdot x_{1i} + \xi_i$$

ZAŁOŻENIA STANDARDOWEGO MODELU REGRESJI LINIOWEJ

- Zmienna objaśniana – y - jest zmienną losową; rozkład tej zmiennej opisuje zbiór wartości, które może ona przyjmować (w danym momencie obserwujemy tylko jedną wartość).
- Wartość oczekiwana rozkładu zmiennej y dla obserwacji „*i*”:

$$E(y_i / x_{1i}) = \alpha_0 + \alpha_1 \cdot x_{1i} \qquad i = 1, \ldots, n$$

$$\alpha_0, \alpha_1 \text{ - nieznane parametry}$$

- Wariancja y_i przy danych x_{1i}, x_{2i} jest stała:

$$var(y_i / x_{1i}) = \sigma^2$$

$$\sigma^2 \text{ - nieznany parametr}$$

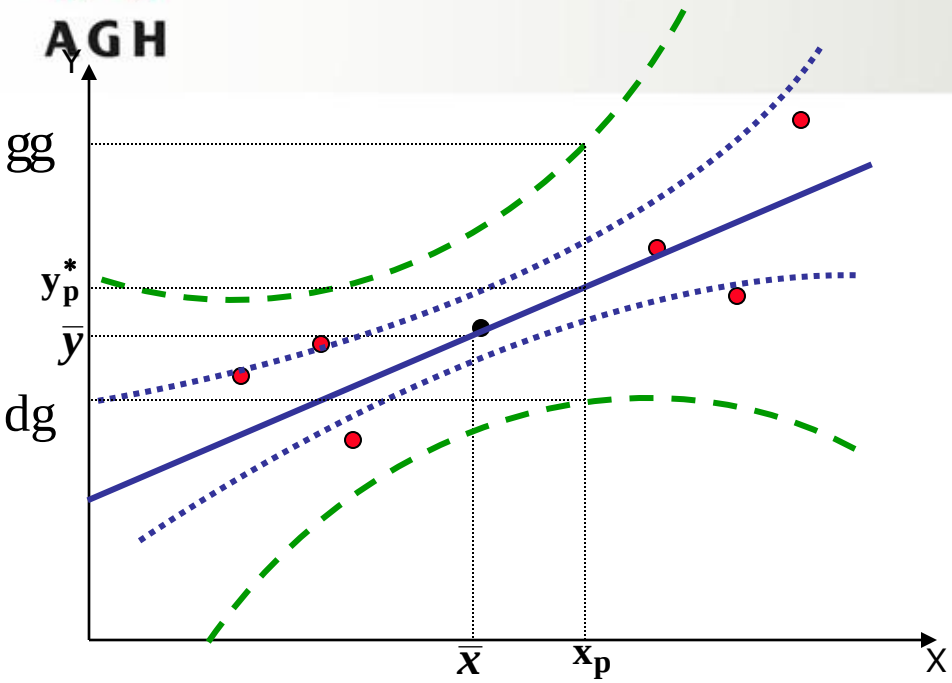
Wariancja mierzy stopień wpływu na zmienną y czynników innych niż x_1 (zmienne pominięte); stałość wariancji implikuje, że dyspersja łącznego wpływu zmiennych pominiętych nie zmienia się w czasie.

- Składnik losowy równania

Każdy składnik losowy ξ_i ma (przy ustalonych x_{1i}, x_{2i}) wartość oczekiwaną równą zero i wariancję σ^2.

$$\xi_i = y_i - E(y_i / x_{1i})$$

Krzywe von Neymana



obserwacje (dane empiryczne)

środek ciężkości próbki

prosta regresji (dla próbki)

$$\hat{y}_i = a_0 + a_1 x_i$$

krzywe wyznaczające pas ufności, w którym z prawdopobieństwem 1-α znajduje się nieznana prosta regresji I rodzaju (dla populacji)

$$E(Y/X) = \alpha_0 + \alpha_1 X$$

krzywe wyznaczające przedziałowe prognozy wartości zmiennej Y dla danego x_p

y_p^* prognoza punktowa uzyskana przez wstawienie x_p do równania

g_g, d_g przedział, w którym z szansą 1-α mieści się nieznana wartość y_i dla i-tej nowej jednostki spoza próbki

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

Estymatorem wariancji σ^2 jest s^2 $\quad s^2 = \frac{SSE}{n-2}$

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

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

Estymator a_i współczynnika regresji α_1 : $\quad \sigma^2_{\alpha_1} = \frac{\sigma^2}{S_{xx}} \qquad s^2_{a1} = \frac{s^2}{S_{xx}}$

Analiza współczynnika regresji $\quad P(a_1 - t_{\alpha/2;n-2} \cdot s_{a_1} < \alpha_1 < a_1 + t_{\alpha/2;n-2} \cdot s_{a_1}) = 1 - \alpha$

Estymacja wartości oczekiwanej y dla danej wartości X:

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

Przedział ufności dla prognozy y_p

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

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

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

Pełny zapis równania regresji liniowej

$$\hat{y}_i = a_0 + a_1 x_i + \xi_i \qquad r = r_{xy}$$
$$s(a_0) \quad s(a_1) \quad s(y)$$

parametry strukturalne i stochastyczne

$\hat{y}_i$ — zmienna zależna, zmienna-skutek, zmienna objaśniana

y_i — zaobserwowane wartości zmiennej zależnej

x_i — zaobserwowane wartości zmiennej niezależnej

a_0 — oszacowana wartość wyrazu wolnego

a_1 — oszacowana wartości 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 to reszty $e_i = y_i - \hat{y}_i$

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

$s(a_0)$ — błąd oszacowania wyrazu wolnego; służy do budowy przedziału ufności dla nieznanej wartości wyrazu wolnego dla populacji oraz do weryfikacji jego istotności

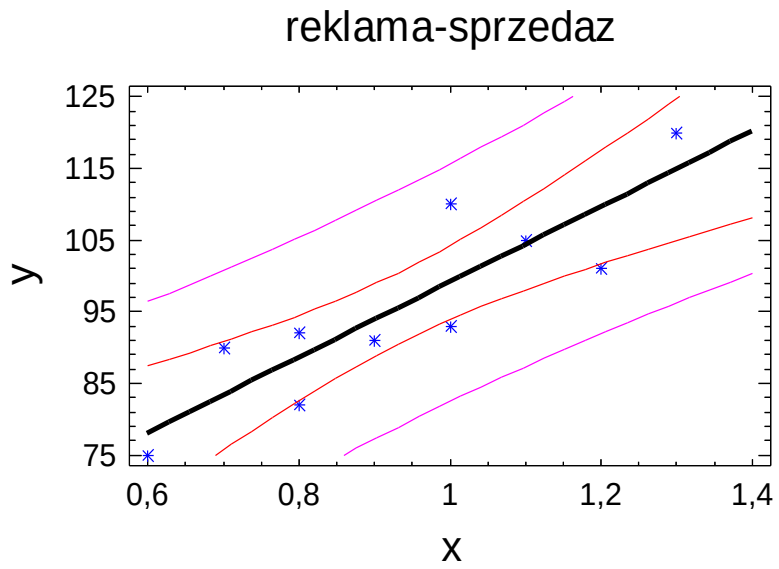
$s(a_1)$ — błąd oszacowania współczynnika regresji; służy do budowy przedziału ufności dla nieznanej wartości α_1 współczynnika regresji dla populacji oraz do weryfikacji jego istotności

s(y) lub s — błąd resztowy; jest odchyleniem standardowym składnika losowego ξ;

Przykład

Czy istnieje związek pomiędzy wydatkami na reklamę (x_i) a wielkością sprzedaży (y_i)?
Wydatki na reklamę i sprzedaż w mln zł.

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



lp.	y_i	x_i	x_i^2	$x_i \cdot y_i$
1	101	1,2	1,44	121,2
2	92	0,8	0,64	73,6
3	110	1	1,00	110
4	120	1,3	1,69	156
5	90	0,7	0,49	63
6	82	0,8	0,64	65,6
7	93	1	1,00	93
8	75	0,6	0,36	45
9	91	0,9	0,81	81,9
10	105	1,1	1,21	115,5
Suma	**959**	**9,4**	**9,28**	**924,8**

$$a_1 = \frac{\sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n}}{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}} = \frac{924,8 - \frac{9,4 \cdot 959}{10}}{9,28 - \frac{(9,4)^2}{10}} = 52,57$$

$$a_0 = \bar{y} - a_1 \bar{x} = 95,9 - 52,57 \cdot 0,94 = 46,49$$

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

Współczynnik determinacji

$$r^2 = \frac{SSTR}{SSTO} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{1226,93}{1600,9} = 0,875$$

Współczynnik zbieżności

$$\varphi^2 = \frac{SSE}{SSTO} = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{373,973}{1600,9} = 0,125$$

Błąd standardowy reszt

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{373,973}{10-2}} = 6,84$$

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

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

$$99{,}06 \pm 2.306\sqrt{46{,}75\left[\frac{1}{10} + \frac{(1{,}0 - 0{,}94)^2}{0{,}444}\right]}$$

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

Prognozowanie wartości y dla x=1,0

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

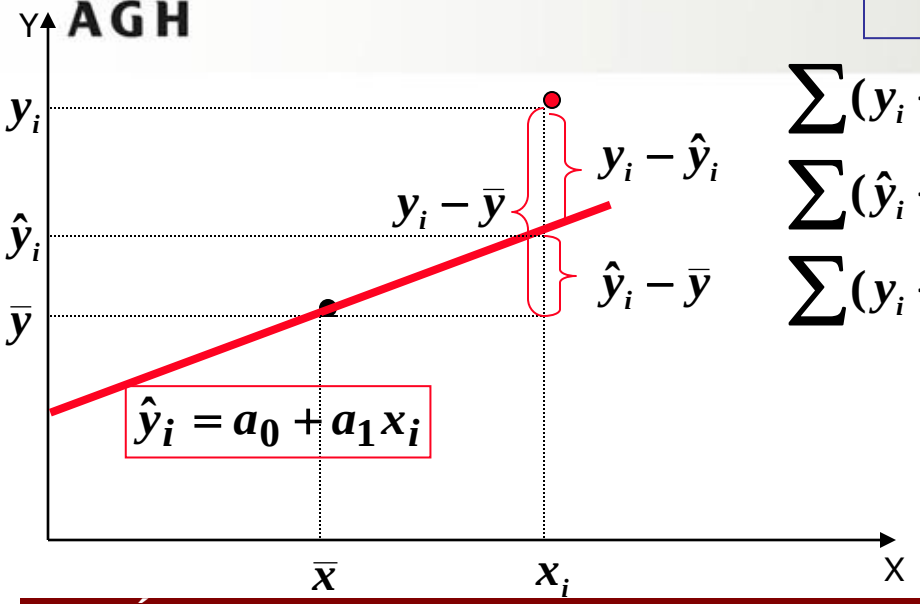
Prognoza przedziałowa: $P(82{,}46 < \hat{y}/x = 1{,}0 < 115{,}66) = 0{,}95$

Przedział ufności dla współczynnika regresji

$$\mathbf{P(28{,}90 < \alpha_1 < 76{,}24) = 0{,}95}$$

ANALIZA WARIANCJI

SSTO = SSTR + SSE



$\sum (y_i - \bar{y})^2$ = SSTO (zmienność całkowita)

$\sum (\hat{y}_i - \bar{y})^2$ = SSTR (zmienność wyjaśniona)

$\sum (y_i - \hat{y}_i)^2$ = SSE (zmienność niewyjaśniona)

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: $a_1 = 0$
H_1: $a_1 \neq 0$

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

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

Regresja krzywoliniowa

W wielu przypadkach dane układają się w zależności nieliniowe:

- gdy mają postać szeregu czasowego
- gdy dane przekrojowe układają się w smugę nieliniową

- gdy krzywoliniowa funkcja wielu zmiennych lepiej opisuje rzeczywistość niż funkcja liniowa; (tego nie widać, która lepsza można poznać tylko po R^2)

Do opisu takich zjawisk stosujemy rozmaite funkcje krzywoliniowe:

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

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

2. wielomiany różnego stopnia (ich fragmenty) $\mathbf{y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \xi \qquad (\alpha_2 > 0)}$

- funkcje potęgowe wielu zmiennych $\mathbf{y = \alpha_0 \cdot x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdot x_3^{\alpha_3} \ldots \xi}$
- funkcje wykładnicze wielu zmiennych $\mathbf{y = e^{\alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2} \cdot \xi}$

ABY MOŻNA BYŁO STOSOWAĆ MNK, FUNKCJE TE MUSZĄ BYĆ SPROWADZONE DO POSTACI LINIOWEJ

$$y = \alpha_0 x_1^{\alpha_1} \xi \quad \Rightarrow \quad \ln y = \ln\alpha_0 + \alpha_1 \cdot \ln x + \ln\xi$$

$$\ln y = y'; \quad \ln x = x' \quad \ln\alpha_0 = \alpha_0' \quad \ln\xi = \xi'$$

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

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 (współczynniki regresji i wyraz wolny), natomiast wszystkie parametry stochastyczne dotyczą tylko transformaty (R^2, φ^2)

ETAPY BUDOWY MODELU EKONOMETRYCZNEGO

1. Sformułowanie modelu

a. wybór zmiennych: $y, x_1, x_2,...$

b. wybór postaci matematycznej modelu: liniowa, potęgowa,...

2. Zebranie danych statystycznych (różne źródła)

3. Estymacja parametrów modelu:

a. parametrów strukturalnych: $a_0, a_1, a_2,...$

b. parametrów stochastycznych: $s(a_i), s(y), R^2, R$

4. Weryfikacja modelu (przy użyciu hipotez i testów statystycznych)

MODEL BEZ WERYFIKACJI NIE MA ŻADNEJ WARTOŚCI

***5*. Interpretacja modelu**

- wyciągnięcie wniosków dla celów zarządzania
- sprzedanie go klientowi

ETAP 1a WYBÓR ZMIENNYCH

• zmienna objaśniana Y: według zainteresowań (na ćwiczeniach), według polecenia szefa (w przedsiębiorstwie), według życzenia klienta (w firmie konsultingowej)

• zmienne objaśniające X_i; wybrane zmienne muszą mieć dużą zmienność (V>30%)

• najczęstszy błąd — *„masło maślane”* prowadzące do związku funkcyjnego i nie dające żadnej informacji o zmiennej objaśnianej

model bez sensu: wynagrodzenie = f(płacy, premii i dodatku stażowego)

ETAP 1b. WYBÓR POSTACI MATEMATYCZNEJ

• modele przyczynowo-skutkowe — najbardziej zalecane jest równoczesne prowadzenie obliczeń dla dwu postaci:

— liniowej $y = \sum a_i x_i + \xi$

— potęgowej $y = \prod x_i^{a_i} \varepsilon$ $\quad \ln y = \sum a_i \ln x_i + \xi$

— stosuje się też modele nieliniowe o narzuconej postaci nieliniowej, których parametry ustala się przez programowanie liniowe lub innymi metodami

• modele tendencji rozwojowej:

— funkcja liniowa

— proste funkcje nieliniowe

— wielomiany

— modele kombinowane: trend + wahania okresowe

ETAP 3. ESTYMACJA PARAMETRÓW MODELU

Cel etapu: wyznaczenie parametrów strukturalnych i stochastycznych

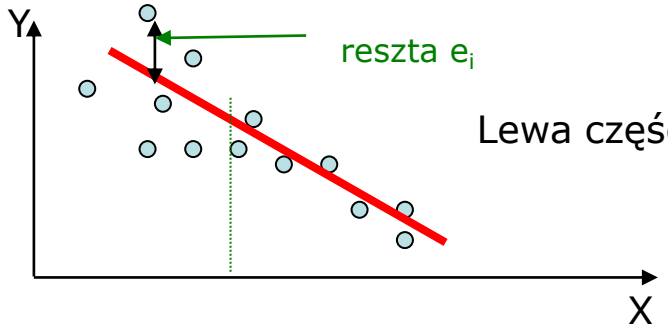
Estymacja: szacowanie parametrów populacji na podstawie próbki

Metody estymacji: MNK i inne

Skutki niedotrzymania założeń MNK i środki zaradcze

1. Model nieprzydatny; niekiedy absurdalny (źle uwarunkowane dane)

2.



Lewa część zbioru ma dużą wariancję, a prawa — wariancję małą

3. Jeśli reszty e_i są ze sobą powiązane (skorelowane) tzn. że występuje *autokorelacja składnika losowego* (najczęściej zjawisko występuje przy szeregach czasowych)

Oznacza to, że istnieje istotna zależność: $e_t = f(e_{t-j}) \quad t = 1,2,...$

Występowanie autokorelacji powoduje nieprzydatność modelu

4. Składnik losowy jest skorelowany ze zmienną objaśniającą, wtedy gdy została pominięta jakaś ważna zmienna - przyczyna

WYKAZ ETAPÓW WERYFIKACJI MODELU

4.1. Badanie istotności korelacji

4.2. Badanie wyrazistości modelu

4.3. Badanie istotności parametrów

4.4. Badanie składnika losowego

- Badanie symetrii skł. losowego
- Badanie losowości skł. losowego
- Badanie stacjonarności skł. los.
- Badanie wartości oczekiwanej skł. losowego
- Badanie autokorelacji skł. losowego
- Badanie heteroskedastyczności skł. losowego
- Badanie normalności skł. losowego

ETAP 4.1. Badanie istotności korelacji

Celem etapu jest sprawdzenie, czy istnieje w populacji generalnej powiązanie pomiędzy zmienną Y i wszystkimi zmiennymi objaśniającymi

Istotność korelacji weryfikuje się przez postawienie następujących hipotez dla współczynnika korelacji dla populacji generalnej:

$H_0 : \rho = 0$ Brak korelacji, nie ma powiązania...

$H_1 : \rho \neq 0$ Korelacja istotna, jest powiązanie...

- testem *t* Studenta (dla regresji dwóch zmiennych)
- testem *F* Fishera
- testem *R* Wallace'a-Snedecora

TEST STUDENTA

$$t_{obl} = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2} \qquad t_{tabl} = t_{\alpha/2;n-2}$$

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

TEST FISHERA

$$F_{obl} = \frac{MSTR}{MSE} = \frac{R^2}{1-R^2}\frac{n-k}{k-1} \qquad F_{tabl} = F_{\alpha;k-1;n-k}$$

Źródło zmienności	*Liczba stopni swobody*	*Suma kwadratów*	*Średni kwadrat*	*Statystyka F*
Model (czynniki)	k-1	SSTR	MSTR	$F_{obl} = \frac{MSTR}{MSE}$
Błąd (reszta)	n–k	SSE	MSE	
Razem	n–1	SSTO		

TEST WALLACE'A-SNEDECORA

Odczyt R_{tabl} z tablicy testu *R* Wallace'a-Snedecora

Stopnie swobody	Liczba zmiennych					
	2		3		4	
	0,05	0,01	0,05	0,01	0,05	0,01
8	0,632	0,765	0,726	0,827	0,777	0,860
18	0,444	0,561	0,532	0,633	0,587	0,678
28	0,361	0,463	0,439	0,530	0,490	0,573

Reguła decyzyjna:

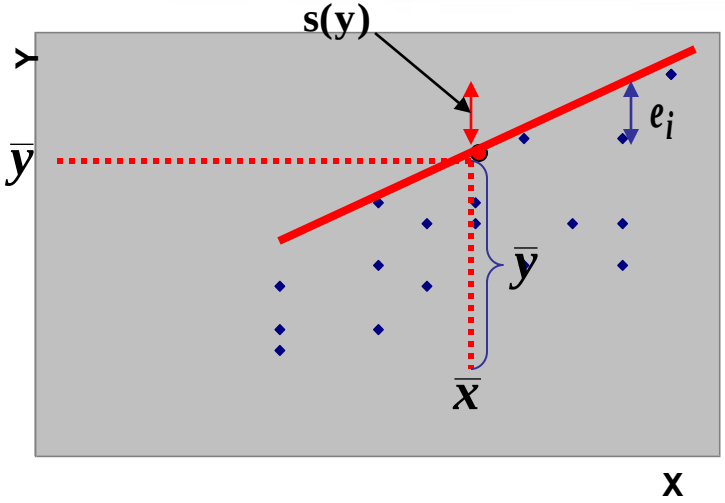
- jeżeli $R_{obl} > R_{tabl}$, model jest poprawny, korelacja istotna
- jeżeli $R_{obl} < R_{tabl}$, model jest niepoprawny, trzeba zmienić albo zestaw zmiennych objaśniających albo jego postać matematyczną

Rola współczynnika determinacji R^2

- korelacja może być istotna przy małym R i bardzo małym R^2 (r=0,4; R^2=0,16 co oznacza, że tylko 16% zmienności zmiennej Y jest wyjaśnione przez zmienną objaśniającą)
- małe R^2 oznacza niski stopień wyjaśnienia rzeczywistości i stanowi zagrożenie dla modelu
- należy dążyć (poprzez odpowiedni dobór zmiennych-przyczyn i postaci matematycznej modelu) do jak największego R^2 (dla postaci pierwotnej)
- wysoka wartość R^2 świadczy o dobrym poznaniu badanego zjawiska
- wysoka wartość R^2 bardzo często wynika jednak ze złego dobrania zmiennych objaśniających (silnie powiązane ze sobą — „masło maślane”)

KORELACJA POZORNA — Przyczyny...Trzeba unikać wartości bezwzględnych (ludność, liczba kin, wielkość produkcji)

ETAP 4.2. Badanie wyrazistości modelu



Wyrazistość modelu dana jest wzorem

$$V_{obl} = \frac{s(y)}{\bar{y}} 100\ \%$$

Współczynnik zmienności losowej V_{obl}<30% (w przeciwnym przypadku rozrzut danych jest zbyt duży)

Uwaga: gdy $\bar{y}$ jest bliskie 0 trudności w ustaleniu czy model poprawny czy niepoprawny

ETAP 4.3. Badanie istotności parametrów (współczynników) modelu

Zmienna (czynnik)	Wartość oszacowana	Błąd oszacowania	Statystyka t_{obl}
Wyraz wolny	a_0	$s(a_0)$	$t(a_0)$
Czynnik X_1	a_1	$s(a_1)$	$t(a_1)$
Czynnik X_2	a_2	$s(a_2)$	$t(a_2)$
Czynnik X_3	a_3	$s(a_3)$	$t(a_3)$
Współczynniki: determinacji R^2, zbieżności φ^2, błąd resztowy s(y)			

weryfikacja hipotezy: H_0: $\alpha_i=0$ wobec H_1: $\alpha_i \neq 0$

$$t_{obl}(a_i) = \frac{a_i - 0}{s(a_i)} \qquad t_{tabl} = t_{\alpha/2;n-k}$$

- jeżeli $|t_{obl}(a_i)|>t_{tabl}(a_i)$, odrzucamy hipotezę zerową; parametr jest istotny z błędem równym co najwyżej α
- jeżeli $|t_{obl}(a_i)|<t_{tabl}(a_i)$, nie ma podstaw do odrzucenia hipotezy zerowej; parametr jest nieistotny

Odrzucając H_0 → ZMIENNA X_i MA WPŁYW NA ZMIENNĄ Y

Badanie symetrii składnika losowego

Badanie symetrii: dla n>30 test *z* (r-d normalny); dla n<30 test *t*-Studenta

$$H_0 : \frac{n_1}{n} = \frac{1}{2}$$

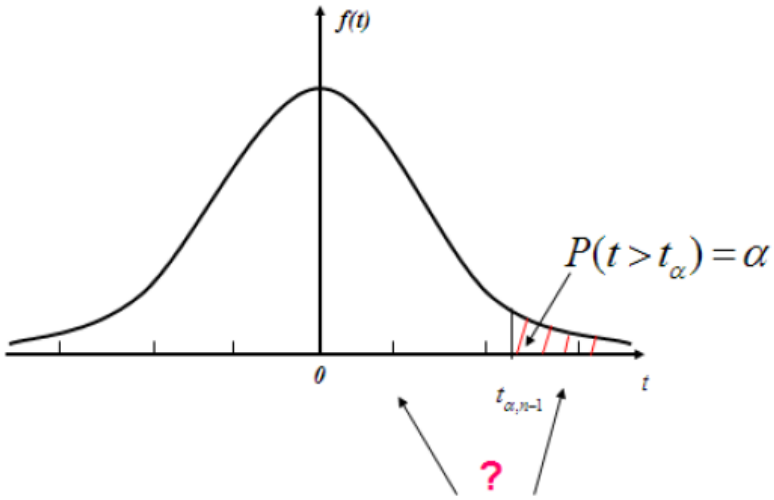
$$H_1 : \frac{n_1}{n} \neq \frac{1}{2}$$

$$t_{obl} = \frac{\left|\frac{n_1}{n} - \frac{1}{2}\right|}{\sqrt{\frac{\frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)}{n-1}}}$$

n_1 – liczba reszt dodatnich (lub ujemnych)
n - liczność próby

Brak symetrii wymaga zmiany matematycznej postaci modelu

Test prawostronny! $t_{\alpha,\nu=n-1}$



$$t = \frac{\left|\frac{m}{n} - \frac{1}{2}\right|}{\sqrt{\frac{\frac{m}{n}\left(1 - \frac{m}{n}\right)}{n-1}}}$$

Badanie losowości składnika losowego

Badanie losowości przeprowadza się testem *t*-Studenta lub testem serii

Test serii: $H_0 : \xi_t$ jest składnikiem losowym $\Leftrightarrow H_0 : Y = f(x_1, x_2, \ldots x_{k-1})$

$H_1 : \xi_t$ nie jest losowy $\Leftrightarrow H_1 : Y \neq f(x_1, x_2, \ldots x_{k-1})$

a) wartościom $e_t>0$ nadajemy symbol *A;* liczba symboli *A* – n_1

b) wartościom $e_t<0$ nadajemy symbol *B*; liczba symboli *B* - n_2

Otrzymujemy podciągi czyli serie z kolejnych symboli *A* lub *B*

c) Liczba wszystkich serii (podciągów) - *k*.

Badanie wartości oczekiwanej składnika losowego

weryfikacja hipotezy:

$$H_0 : EV(\xi) = 0$$
$$H_1 : EV(\xi) \neq 0$$

Celem etapu jest sprawdzenie, czy odchylenie od *„0”* nie jest zbyt duże (służy do tego test t-Studenta)

Badanie heteroskedastyczności składnika losowego

Heteroskedastyczność – niejednorodność wariancji składnika losowego w obrębie próby

weryfikacja hipotezy: $H_0 : \sigma^2(\xi) = const$

$H_1 : \sigma^2(\xi) \neq const$

Skutki – niespełnienie założeń MNK

Testowanie homoskedastyczności (heteroskedastyczności)

1. Test White'a (najbardziej ogólny)
2. Test Harrisona-McCabe'a
3. Test Goldfelda-Quandta

Badanie autokorelacji składnika losowego

Składnik losowy ξ nie jest czysto losowy, lecz zależy od wskaźnika i, czyli zmienne losowe ξ_i są zależne od poprzednich wartości $\xi_{t-\tau}$.

Autokorelacja to korelacja wartości zmiennej ξ z jej wartościami z okresów wcześniejszych o jeden lub więcej okresów.
Na ogół autokorelację można wyrazić w postaci relacji:

$$\xi_i = f(\xi_{i-1}, \xi_{i-2}, \ldots, \xi_{i-\tau}) \qquad e_i = f(e_{i-k}) \quad i = 1,2,\ldots$$

W praktyce przyjmuje się, że funkcja f jest funkcją liniową, a maksymalne opóźnienie τ wynosi jeden lub dwa (rząd autokorelacji).

Estymator współczynnika autokorelacji ρ_1 (rzędu pierwszego,

$$r_1 = \frac{\sum_{i=2}^{n}(e_i - \bar{e}_i)(e_{i-1} - \bar{e}_{i-1})}{\sqrt{\sum_{i=2}^{n}(e_i - \bar{e}_i)^2 \sum_{i=2}^{n}(e_{i-1} - \bar{e}_{i-1})^2}}$$

$e_i = y_i - \hat{y}_i$	e_{i-1}
e_1	-
e_2	e_1
e_3	e_2
e_4	e_3
e_5	e_4

Skutki: estymatory są nieefektywne, estymator wariancji ξ jest obciążony co prowadzi do niedoszacowania błędów

Badanie autokorelacji można przeprowadzić:

- testem R istotności korelacji

$$r_1 = \frac{\sum_{i=2}^{n}(e_i - \bar{e}_i)(e_{i-1} - \bar{e}_{i-1})}{\sqrt{\sum_{i=2}^{n}(e_i - \bar{e}_i)^2 \sum_{i=2}^{n}(e_{i-1} - \bar{e}_{i-1})^2}}$$

- testem Durbina-Watsona

Test Durbina-Watsona służy do sprawdzenia hipotezy: $H_0 : \rho_1 = 0$ $H_1 : \rho_1 < 0 \; lub \; H_1 : \rho_1 > 0$

Statystyka d:

$$d_{obl} = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}$$

Na podstawie tablic Durbina-Watsona wyznaczamy dwie wartości krytyczne: d_L i d_U, dla określonej liczności próby *(n)* i określonej ilości zmiennych objaśniających *(k).*

Reguła decyzyjna:

- jeżeli $d_{obl} < d_L$ – wnioskujemy, że zachodzi dodatnia autokorelacja,
- jeżeli $d_L < d_{obl} < d_U$ – wynik niczego nie przesądza,
- jeżeli $d_U < d_{obl} < 4-d_U$ – nie ma podstaw do odrzucenia H_0 – brak autokorelacji,
- jeżeli $4-d_U < d_{obl} < 4-d_L$ – wynik niczego nie przesądza,
- jeżeli $d_{obl} > 4-d_L$ – wnioskujemy, że zachodzi ujemna autokorelacja.

Badanie normalności składnika losowego

Celem etapu jest stwierdzenie, czy reszty mają rozkład normalny

Stosuje się testy nieparametryczne:

- λ - Kołmogorowa-Smirnowa lub χ^2 test

Powyższe testy wymagają bardzo dużej próby (podział zbioru reszt na klasy wartości, gdzie $n_i >= 5$)

TEST Jargue'a-Bery (JB)

Krok 1. szacowanie wartości obciążonego estymatora odchylenia standardowego składnika losowego $s = \sqrt{\frac{1}{n}\sum_i^n e_i^2}$

Krok 2. szacowanie wartości miary asymetrii rozkładu reszt (skewness) $A = \frac{1}{n}\sum_i^n \frac{e_i^3}{s^3}$

Krok 3. szacowanie wartości miary kurtozy rozkładu reszt $K = \frac{1}{n}\sum_i^n \frac{e_i^4}{s^4}$

Krok 4. wyliczanie wartości statystyki JB

Statystyka JB ma rozkład χ^2 dla $\nu = 2$ $JB = \frac{n-k}{6}\left(A^2 + \frac{1}{4}(K-3)^2\right)$ k – ilość zmiennych objaśniających

Reguła decyzyjna:

- jeżeli JB>$\chi^2_{\alpha,2}$ to H_0 o normalności składnika losowego odrzucamy (prawostronny obszar odrzucenia!!)
- jeżeli JB<$\chi^2_{\alpha,2}$ nie ma podstaw do odrzucenia H_0

INTERPRETACJA MODELU

- INTERPRETUJĄC MODEL (RÓWNANIE REGRESJI) NALEŻY UŻYWAĆ WYŁĄCZNIE PROSTEJ TERMINOLOGII EKONOMICZNEJ ZROZUMIAŁEJ DLA KLIENTA
- NIE NALEŻY UŻYWAĆ TERMINOLOGII MATEMATYCZNEJ ZROZUMIAŁEJ TYLKO DLA TWÓRCÓW MODELU
- INTERPRETOWAĆ WOLNO TYLKO MODEL ZWERYFIKOWANY
- CAŁY TRUD MODELOWANIA NIE MOŻE BYĆ „SZTUKĄ DLA SZTUKI”, LECZ MA SŁUŻYĆ UZYSKANIU KONKRETNYCH WNIOSKÓW PRAKTYCZNYCH

Ocena jakościowa i ilościowa

Na podstawie znaków stojących przy współczynnikach r oraz a_i możemy stwierdzić, że wpływ jest:

• dodatni (*im lepsze zaliczenie - tym lepszy wynik egzaminu; im więcej wydatków na reklamę - tym większa sprzedaż; itd.)*

• ujemny *(im więcej zatrudnionych - tym gorszy wynik finansowy; im mniej braków - tym wyższy zysk; im mniejsza absencja - tym wyższe wynagrodzenie; itd.)*

Przykład

Model liniowy

Dane zawarte w poniższej tablicy uzyskano z pewnego złoża gazu ziemnego, na którym znajduje się 8 odwiertów produkcyjnych. Dla każdego odwiertu podano początkowe dopuszczalne wydobycie gazu i efektywną miąższość pokładu produktywnego w tych odwiertach. Podejrzewamy, że istnieje zależność pomiędzy początkowym dopuszczalnym wydobyciem gazu a efektywną miąższością.

Numer odwiertu	Początkowe wydobycie dopuszczalne [m3/min]	Miąższość efektywna pokładu [m]
1	132	18,3
2	302	76,0
3	42	33,0
4	300	58,0
5	77	10,0
6	75	15,6
7	150	30,0
8	17,5	15,7

$$a_1 = \frac{\sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n}}{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}} = \frac{50868{,}35 - \frac{256{,}6 \cdot 1095{,}5}{8}}{12053 - \frac{(256{,}6)^2}{8}} = 4{,}114$$

$$a_0 = \bar{y} - a_1 \bar{x} = 136{,}9 - 4{,}115 * 32{,}075 = 4{,}97$$

l.p.	y_i	x_i	$x_i * y_i$
1	**132**	**18,3**	2415,6
2	**302**	**76,0**	22952
3	**42**	**33,0**	1386
4	**300**	**58,0**	17400
5	**77**	**10,0**	770
6	**75**	**15,6**	1170
7	**150**	**30,0**	4500
8	**17,5**	**15,7**	274,75
Suma	1095,5	256,6	50868,35

$$\hat{y}_i = 4{,}97 + 4{,}114 x_i + \xi_i$$

r = 0,874

36,3 0,93 57,8