

Jednorównaniowy model liniowy z jedną zmienną objaśniającą

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

gdzie:

Y - zmienna objaśniana, y_i - wartości (obserwacje) zmiennej Y ; $i = 1, \dots, n$ - numer obserwacji,

X - zmienna objaśniająca, x_i - wartości zmiennej X ,

β_0, β_1 - parametry strukturalne (ich przybliżoną wartość wyznacza się na podstawie obserwacji (x_i, y_i))

ε - składnik losowy.

Zakładamy, że

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

tzn. każda zaobserwowana wartość y_i jest funkcją liniową x_i z dokładnością do składnika losowego ε_i .

Zakładamy również, że x_i są ustalonymi wartościami (nielosowymi), takimi samymi w powtarzalnych próbach. Składniki losowe ε_i są losowymi zmiennymi niezależnymi o zerowej wartości przeciętnej i wariancji, która nie zależy od i (homoskedastyczność).

Aby wyznaczyć przybliżoną wartość parametrów strukturalnych β_0, β_1 na podstawie próby stosujemy metodę najmniejszych kwadratów (MNK).

MNK polega na wyznaczeniu takich przybliżeń

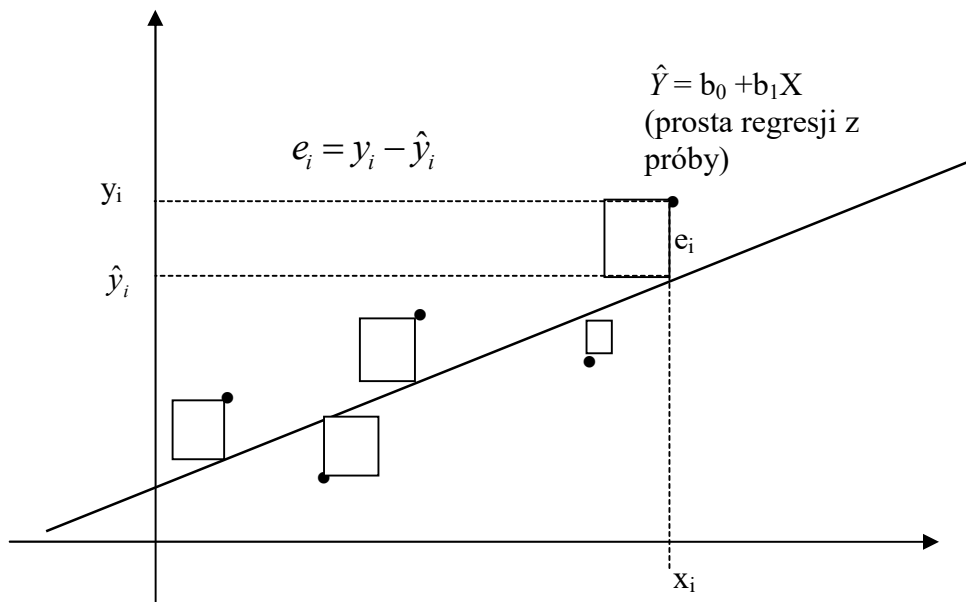
$$b_0 \approx \beta_0 \quad b_1 \approx \beta_1$$

aby dla danych obserwacji (x_i, y_i) suma kwadratów odchyłeń zaobserwowanych wartości y_i od wartości teoretycznych $\hat{y}_i = \beta_0 + \beta_1 x_i$ była minimalna, tzn. chcemy wyznaczyć minimum funkcji:

$$(*) \quad S(b_0, b_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

$e_i = y_i - \hat{y}_i$ nazywamy **resztami** modelu regresji

MNK:



Należy wyznaczyć prostą regresji tak aby suma pól kwadratów była minimalna.

Obliczając pochodne cząstkowe funkcji (*) i przyrównując do zera otrzymujemy (układ równań normalnych)

$$\frac{\partial S}{\partial \beta_0} = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)(-1) = -2 \left(\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i - n\beta_0 \right) = 0$$

$$\frac{\partial S}{\partial \beta_1} = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)(-x_i) = -2 \left(\sum_{i=1}^n y_i x_i - \beta_1 \sum_{i=1}^n x_i^2 - \beta_0 \sum_{i=1}^n x_i \right) = 0$$

rozwiązując otrzymany układ równań (np. pierwsze równanie mnożymy przez $-\bar{x} = -\frac{1}{n} \sum_{i=1}^n x_i$ i dodając stronami obliczamy β_1) otrzymamy wzory na przybliżone wartości parametrów strukturalnych

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum x_i^2 - (\bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Można wykazać, że dla tych wartości spełniony jest warunek dostateczny minimum.

Prostą

$$\hat{Y} = b_0 + b_1 X$$

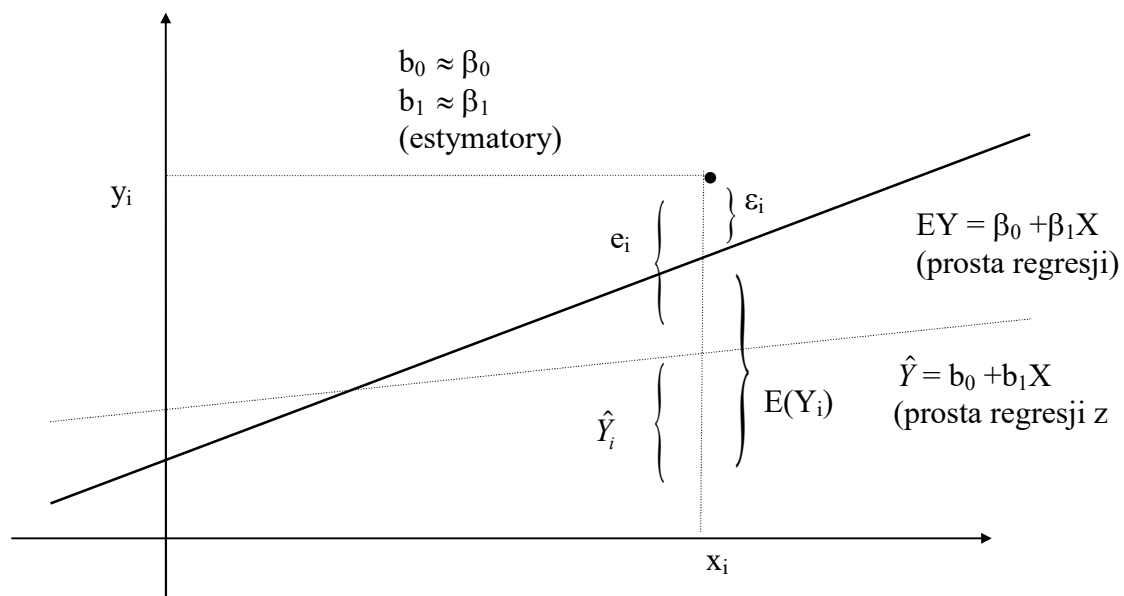
nazywamy **prostą regresji z próby**.

Uwaga:

$$\text{a) } \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\text{b) } \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

Model regresji liniowej:



Uwaga

Gdy X jest zmienną czasową $x_i = t$ tzn. model ma postać $Y = \beta_0 + \beta_1 t + \varepsilon$ wówczas taki model nazywamy **modelem tendencji rozwojowej** lub **modelem trendu liniowego**.

Wtedy korzystając z ustalonych wartości t wzory na b_1 i b_0 można uprościć (patrz odpowiedni temat).

Miary dopasowania.

Wariancja resztowa:

Wariancja resztowa to uśrednienie pól kwadratów zbudowanych na resztach i odzwierciedla stopień dopasowania prostej regresji do danych statystycznych.

Niech, $e_i = y_i - \hat{y}_i$, gdzie $\hat{y}_i = b_0 + b_1 x_i$ wtedy

$$S_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

czyli

$$S_e^2 = \frac{\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i}{n-2}$$

$S_e = \sqrt{S_e^2}$ oznacza średnie (standardowe) odchylenie od prostej regresji.

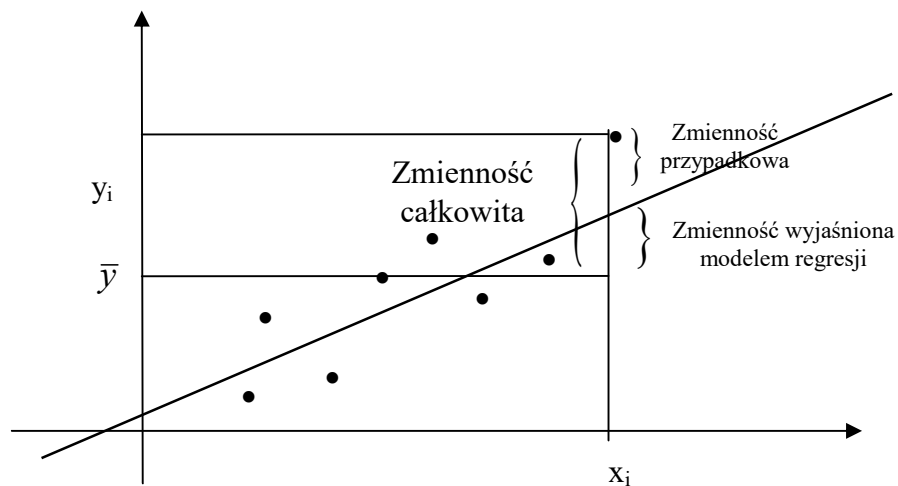
Dopasowanie modelu do danych empirycznych można oceniać odchyleniem standardowym reszt lecz jest to miara bezwzględna i nieunormowana, dlatego do porównań lepsze są miary względne lub unormowane.

Najprostszą względną miarą dopasowania jest **współczynnik zmienności losowej** :

$$V_e = \frac{S_e}{\bar{Y}} 100\%$$

Współczynnik ten informuje jaką część średniej wartości badanego zjawiska stanowi odchylenie standardowe reszt.

Mniejsze wartości tego współczynnika wskazują na lepsze dopasowanie modelu do danych empirycznych, niekiedy żąda się aby np. $V_e < 0,2$ (0,3).



Wprowadzamy oznaczenia:

Całkowita suma kwadratów (zmienność całkowita): $CSK = \sum (y_i - \bar{y})^2$

Wyjaśniona suma kwadratów (zmienność wyjaśniona): $WSK = \sum (\hat{y}_i - \bar{y})^2$

Niewyjaśniona suma kwadratów (zmienność przypadkowa): $NSK = \sum e_i^2$

gdzie: $\hat{y}_i = b_0 + b_1 x_i$

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

Czyli $CSK = WSK + NSK$

Miarą dopasowania modelu do rzeczywistości (wartości zaobserwowanych) jest również współczynnik determinacji R^2

Współczynnik determinacji: $R^2 = \frac{WSK}{CSK} \quad R^2 \in \langle 0, 1 \rangle$

Współczynnik ten określa jaka część całkowitej zmienności zmiennej objaśnianej została wyjaśniona przez model regresji liniowej.

Praktyczne sposoby obliczania współczynnika determinacji:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = \frac{b_0 \sum y_i + b_1 \sum x_i y_i - n \bar{y}^2}{\sum y_i^2 - n(\bar{y})^2} =$$

$$= \frac{b_1 (\sum x_i y_i - n \bar{x} \bar{y})}{\sum y_i^2 - n(\bar{y})^2} = 1 - \frac{S_e^2}{\frac{n}{n-2} S_Y^2} = \frac{\text{cov}^2(X, Y)}{S_X^2 S_Y^2} = r^2$$

Wniosek. Dla modelu liniowego wartość liczbową współczynnika determinacji jest równa kwadratowi współczynnika korelacji Pearsona.

Przykład

Badano zależności kosztów całkowitych (w tys. zł.) Y od wielkości produkcji (tys. szt.) X w 6-ciu zakładach produkcyjnych.

x_t	2	4	3	2	6	1
y_t	2	5	4	4	7	2

Dla modelu $Y = \beta_0 + \beta_1 x + \varepsilon$ wyznaczamy przybliżone wartości parametrów strukturalnych i współczynnik determinacji.

Obliczenia wykonamy w tabeli

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
2	2	-1	-2	2	1	4	3	-1	1
4	5	1	1	1	1	1	5	1	1
3	4	0	0	0	0	0	4	0	0
2	4	-1	0	0	1	0	3	-1	1
6	7	3	3	9	9	9	7	3	9
1	2	2	-2	4	4	4	2	-2	4
18	24	0	0	16	16	18	24	0	16

$$\bar{x} = \frac{18}{6} = 3; \quad \bar{y} = \frac{24}{6} = 4; \quad b_1 = \frac{16}{16} = 1; \quad b_0 = 4 - 1 \cdot 3 = 1$$

zatem związek pomiędzy kosztami całkowitymi a wielkością produkcji wyraża się zależnością liniową w postaci

$$\hat{Y} = 1 + X$$

Współczynnik determinacji

$$R^2 = \frac{16}{18} = 0,89$$

należy oczekiwać, że rozpatrywany model wyjaśnia 89% całkowitej zmienności kosztów całkowitych produkcji.

Prognoza na podstawie modelu liniowego. (jedna zmienna)

$\hat{Y} = b_0 + b_1 X$ oszacowany model ekonometryczny.

Prognoza punktowa.

Niech x_τ – przewidywana wartość cechy X w okresie prognozy.

Prognoza punktowa y_τ^* to przewidywana wartość cechy Y odpowiadająca wartości x_τ cechy X .

$$y_\tau^* = b_0 + b_1 x_\tau$$

Standardowy błąd prognozy

$$s_{\tau} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{\tau} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = s_e \sqrt{1 + \frac{\sum_{i=1}^n x_i^2 + nx_{\tau}^2 - 2x_{\tau} \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}}$$

Zatem należy traktować wartość prognozy jako $y_{\tau}^* \pm s_{\tau}$

Gdzie $s_e = \sqrt{s_e^2}$ to odchylenie resztowe.

Niech $e_i = y_i - \hat{y}_i$, gdzie $\hat{y}_i = b_0 + b_1 x_i$

wtedy

$$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

czyli

$$s_e^2 = \frac{\sum_{i=1}^n y_i^2 - b_0 \sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i y_i}{n-2}$$

Jakość prognozy punktowej możemy ocenić względnym błędem prognozy punktowej

$$\delta_{punkt} = \frac{s_{\tau}}{y_{\tau}^*} \cdot 100\%$$

Model tendencji rozwojowej

Gdy X jest zmienną czasową $x_i = t$ ($t = 1, 2, \dots, n$) tzn. model regresji ma postać $\hat{Y} = b_0 + b_1 t$ wówczas taki model nazywamy **modelem tendencji rozwojowej** lub **modelem trendu liniowego**

Wtedy korzystając z własności:

$$(*) \quad \sum_{t=1}^n t = \frac{n(n+1)}{2}, \quad \sum_{t=1}^n t^2 = \frac{n(n+1)(2n+1)}{6}, \quad \bar{t} = \frac{n+1}{2} \quad \sum (t - \bar{t})^2 = \sum t^2 - n(\bar{t})^2$$

Mamy

$$b_1 = \frac{n \sum t y_t - \sum t \sum y_t}{n \sum t^2 - (\sum t)^2} = \frac{12 \sum (t - \bar{t}) y_t}{n(n^2 - 1)}$$

$$b_0 = \bar{y} - b_1 \bar{t} = \bar{y} - b_1 \frac{n+1}{2}$$

Wariancja resztowa

Niech $e_i = y_i - \hat{y}_i$, (gdzie $\hat{y}_i = b_0 + b_1 x_i$) to reszty modelu, wtedy

$$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} \quad \text{czyli} \quad s_e^2 = \frac{\sum_{t=1}^n y_t^2 - b_0 \sum_{t=1}^n y_t - b_1 \sum_{t=1}^n t y_t}{n-2}$$

$s_e = \sqrt{s_e^2}$ oznacza średnie (standardowe) odchylenie od trendu liniowego.

Dopasowanie modelu do danych empirycznych oceniamy też współczynnikiem determinacji

$$\begin{aligned} R^2 &= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = \\ &= \frac{b_0 \sum y_i + b_1 \sum ty_t - n\bar{y}^2}{\sum y_t^2 - n(\bar{y})^2} = \frac{b_1 (\sum ty_t - n\bar{t} \bar{y})}{\sum y_t^2 - n(\bar{y})^2} = r^2 \end{aligned}$$

Proгноza dla modelu trendu

Niech t_τ – okres prognozy.

Proгноza punktowa y_τ^* to przewidywana wartość cechy Y w okresie t_τ .

$$y_\tau^* = b_0 + b_1 t_\tau$$

Standardowy błąd prognozy punktowej

$$s_\tau = s_e \sqrt{1 + \frac{1}{n} + \frac{(t_\tau - \bar{t})^2}{\sum_{t=1}^n (t - \bar{t})^2}} = s_e \sqrt{1 + \frac{\sum_{t=1}^n t^2 + n t_\tau^2 - 2 t_\tau \sum_{t=1}^n t}{n \sum_{t=1}^n t^2 - \left(\sum_{t=1}^n t \right)^2}}$$

Wzór ten można uprościć korzystając z własności (*).

$$\begin{aligned}
 s_{\tau} &= s_e \sqrt{1 + \frac{\frac{n(n+1)(2n+1)}{6} + nt_{\tau}^2 - 2t_{\tau} \frac{n(n+1)}{2}}{\frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}}} = \\
 &= s_e \sqrt{1 + \frac{2(2n+1) + \frac{12t_{\tau}^2}{n+1} - 12t_{\tau}}{n^2 - n}}
 \end{aligned}$$

Zatem należy traktować wartość prognozy jako $y_{\tau}^* \pm s_{\tau}$

Jakość prognozy punktowej możemy ocenić względnym błędem prognozy punktowej

$$\delta_{punkt} = \frac{s_{\tau}}{y_{\tau}^*} \cdot 100\%$$

Przykład. 1

Liczba maturzystów (tys. osób) w pewnym województwie w latach 1991-1999 wynosiła;

Rok	1991	1992	1993	1994	1995	1996	1997	1998	1999
Y	3,8	6	7,5	7,5	9,2	9,9	12,3	13,2	13,4

Obliczenia wykonamy w tabelce:

rok	t	y_t	$t \cdot y_t$	y_t^2
1991	1	3,8		
1992	2	6		
1993	3	7,5		
1994	4	7,5		
1995	5	9,2		
1996	6	9,9		
1997	7	12,3		
1998	8	13,2		
1999	9	13,4		

rok	t	y_t	t^*y_t	y_t^2
1991	1	3,8	3,8	14,44
1992	2	6	12	36
1993	3	7,5	22,5	56,25
1994	4	7,5	30	56,25
1995	5	9,2	46	84,64
1996	6	9,9	59,4	98,01
1997	7	12,3	86,1	151,29
1998	8	13,2	105,6	174,24
1999	9	13,4	120,6	179,56
	45	82,8	486	850,68

$$n = 9$$

Obliczamy wartości średnie:

$$\bar{t} = \frac{n+1}{2} = (9+1)/2 = 5$$

$$\bar{y} = \frac{\sum y_t}{n} = \frac{82,8}{9} = 9,2$$

Obliczamy oszacowane parametry modelu

$$\hat{Y} = b_0 + b_1 t$$

Mamy

$$b_1 = \frac{12(\sum ty_t - \bar{t} \sum y_t)}{n(n^2 - 1)} = \frac{12(486 - 5 \cdot 82,8)}{9(81 - 1)} = 1,2$$

$$b_0 = \bar{y} - b_1 \bar{t} = 9,2 - 1,2 \cdot 5 = 3,2$$

Otrzymane równanie trendu liniowego ma postać

$$\hat{Y} = 3,2 + 1,2t$$

Wniosek: W rozpatrywanym województwie przeciętnie z roku na rok przybywało ok. 1200 maturzystów.

Obliczamy wariancję resztową

$$s_e^2 = \frac{\sum_{t=1}^n y_t^2 - b_0 \sum_{t=1}^n y_t - b_1 \sum_{t=1}^n t y_t}{n-2} = \frac{850,68 - 3,2 \cdot 82,8 - 1,2 \cdot 486}{9-2} = 0,36$$

zatem

$$s_e = \sqrt{s_e^2} = \sqrt{0,36} = 0,6$$

Dopasowanie modelu do danych empirycznych oceniamy współczynnikiem zmienności resztowej V_e .

$$V_e = \frac{S_e}{\bar{y}} = \frac{0,6}{9,2} = 6,52\%$$

Wniosek: Mała wartość współczynnika zmienności świadczy o **dobrym** dopasowaniu modelu do danych empirycznych. Wynik jest **pozytywny** dla modelu.

Dopasowanie modelu do danych empirycznych oceniamy też współczynnikiem determinacji

$$R^2 = \frac{b_1 \left(\sum ty_t - n\bar{t} \bar{y} \right)}{\sum y_t^2 - n(\bar{y})^2} = \frac{1,2(486 - 9 \cdot 5 \cdot 9,2)}{850,68 - 9 \cdot (9,2)^2} = 0,97 = 97\%$$

Wniosek: Duża wartość współczynnika determinacji świadczy o **dobrym** dopasowaniu modelu do danych empirycznych. Rozpatrywany model wyjaśnia ok. 97% całkowitej zmienności liczby maturzystów. Wynik jest **pozytywny** dla modelu.

Wyznamy prognozę na rok 2000 i 2001 i ocenimy jej błąd.

Przewidywana liczba maturzystów w **roku 2000**:

$$y_{\tau}^* = b_0 + b_1 t_{\tau} = 3,2 + 1,2 \cdot 10 = 15,2$$

Zatem przewidujemy, że w roku 2000 będzie około 15200 maturzystów.

Standardowy błąd tej prognozy wynosi

$$s_{\tau} = s_e \sqrt{1 + \frac{2(2n+1) + \frac{12t_{\tau}^2}{n+1} - 12t_{\tau}}{n^2 - n}} =$$

$$= 0,6 \sqrt{1 + \frac{2(2 \cdot 9 + 1) + \frac{12 \cdot 10^2}{9+1} - 12 \cdot 10}{9^2 - 9}} = 0,742$$

Zatem w roku 2000 będzie około 15200 ± 742 maturzystów.

Jakość prognozy możemy ocenić błędem względnym

$$\delta_{\tau} = \frac{s_{\tau}}{y_{\tau}} \cdot 100\% = \frac{0,742}{15,2} \cdot 100\% = 4,9\%$$

Zatem prognoza **jest dopuszczalna** (ma mały błąd względny).

Przewidywana liczba maturzystów w roku 2001:

$$y_{\tau}^* = b_0 + b_1 t_{\tau} = 3,2 + 1,2 \cdot 11 = 16,4$$

Zatem przewidujemy, że w roku 2001 będzie około 16400 maturzystów.

Standardowy błąd tej prognozy wynosi

$$s_{\tau} = s_e \sqrt{1 + \frac{2(2n+1) + \frac{12t_{\tau}^2}{n+1} - 12t_{\tau}}{n^2 - n}} =$$

$$= 0,6 \sqrt{1 + \frac{2(2 \cdot 9 + 1) + \frac{12 \cdot 11^2}{9+1} - 12 \cdot 11}{9^2 - 9}} = 0,785$$

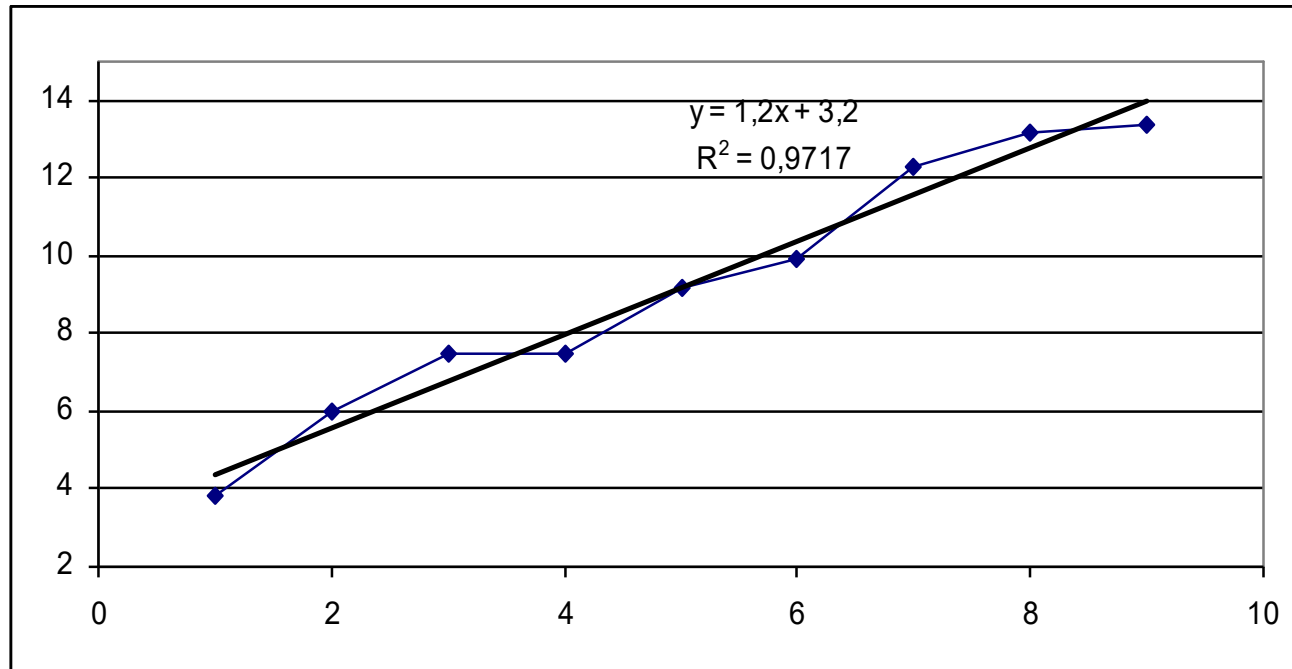
Zatem w roku 2001 będzie około 16400 ± 785 maturzystów.

Jakość prognozy możemy ocenić błędem względnym

$$\delta_{\tau} = \frac{s_{\tau}}{y_{\tau}} \cdot 100\% = \frac{0,785}{16,4} \cdot 100\% = 4,8\%$$

Zatem prognoza **jest dopuszczalna** (ma mały błąd względny).

Wykres danych i równania trendu (Excel).



Analiza danych → Regresja

1	3,8
2	6
3	7,5
4	7,5
5	9,2
6	9,9
7	12,3
8	13,2
9	13,4

PODSUMOWANIE - WYJŚCIE

Statystyki regresji

Wielokrotność R	0,985728116
R kwadrat	0,971659919
Dopasowany R kwadrat	0,967611336
Błąd standardowy	0,6
Obserwacje	9

← Se

ANALIZA WARIANCJI

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Istotność F</i>
Regresja	1	86,4	86,4	240	1,13E-06
Resztkowy	7	2,52	0,36		
Razem	8	88,92			

	<i>Współczynniki</i>	<i>Błąd standardowy</i>	<i>t Stat</i>	<i>Wartość-p</i>	<i>Dolne 95%</i>	<i>Górne 95%</i>	<i>Dolne 95,0%</i>	<i>Górne 95,0%</i>
Przecięcie	3,2	0,43589	7,341303	0,000157	2,169284	4,230716	2,169284	4,230716
Zmienna X 1	1,2	0,07746	15,49193	1,13E-06	1,016837	1,383163	1,016837	1,383163

b0

b1