MODEL REGRESJI LINIOWEJ

1. Dla 10 losowo wybranych studentów zanotowano dane o czasie nauki (X-w godz.) w okresie sesji i otrzymanych ocenach z egzaminu z filozofii (Y).
   a) Na podstawie poniższych danych oszacować liniową funkcję regresji osiąganych przez studentów ocen od czasu nauki i zinterpretować,
   b) Oszacować średnie błędy szacunki,
   c) Zbadać istotność współczynnika regresji (γ=0,05)
   d) Ocenić dopasowanie modelu regresji liniowej do danych empirycznych,
   e) Jakiej oceny powinien oczekiwać student uczący się 23 godziny (predykcja punktowa i przedziałowa dla γ=0,05)

i	x_i	y_i	y_i*x_i	x_i^2	$\hat{y}_i$	$e_i=y_i-\hat{y}_i$	e_i^2	$\sum_{i=1}^{n}(y_i-\bar{y}_i)^2$	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2$
1	3	2	6	9	2,270	-0,270	0,073	1,690	1,060
2	5	3	15	25	2,489	0,511	0,261	0,090	0,658
3	15	3,5	52,5	225	3,585	-0,085	0,007	0,040	0,080
4	10	3	30	100	3,037	-0,037	0,001	0,090	0,069
5	25	4,5	112,5	625	4,681	-0,181	0,033	1,440	1,907
6	17	3	51	289	3,804	-0,804	0,646	0,090	0,254
7	13	4	52	169	3,366	0,634	0,402	0,490	0,004
8	7	2	14	49	2,708	-0,708	0,501	1,690	0,350
9	23	5	115	529	4,462	0,538	0,289	2,890	1,350
10	6	3	18	36	2,598	0,402	0,162	0,090	0,493
∑	**124**	**33**	**466**	**2056**	**33,000**	**0,000**	**2,375**	**8,600**	**6,225**

$\bar{x} = 12,4;\ \bar{y} = 3,3$

a) Współczynnik regresji

$$\hat{\alpha} = \frac{n\cdot\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\cdot\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{c_{xy}}{s_x^2} = r_{xy}\cdot\frac{s_y}{s_x} = \frac{10\cdot 466 - 124\cdot 33}{10\cdot 2056-(124)^2} = 0,1096$$

Wydłużenie czasu nauki o godzinę podnosi **średnią ocenę** o 0,1096

Wyraz wolny

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i - \hat{\alpha}\cdot\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\alpha}\cdot\bar{x} = 3,3 - 0,1096\cdot 12,4 = 1,9410$$

Przy zerowym czasie nauki **średnia (**teoretyczna) **ocena** wynosi 1,94≈2

Funkcja regresji liniowej z oszacowanymi parametrami strukturalnymi

$$\hat{y}_i = 0,1096\cdot x_i + 1,9410$$

---

b) $$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{2,375}{10-2}} = 0,5449$$

Wszystkie czynniki nie uwzględnione w modelu zmieniają **średnio** ocenę o 0,5449

$$s_{\hat{\alpha}} = \sqrt{\frac{S_e^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \sqrt{\frac{S_e^2}{\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2}} = \sqrt{\frac{S_e^2}{(n-1)\cdot s_x^2}} = \sqrt{\frac{0,2969}{2056-10\cdot(12,4)^2}} =$$

0,0239

Przy szacowaniu wartości współczynnika regresji średnia wartość popełnianego błędu wynosi 0,0239.

$$s_{\hat{\beta}} = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot (\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2)}} = s_{\hat{\alpha}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = 0{,}0239 \cdot \sqrt{\frac{2056}{10}} =$$
$$= 0{,}3427$$

Przy szacowaniu wartości wyrazu wolnego średnia wartość popełnianego błędu wynosi 0,3427.

Model regresji liniowej z oszacowanymi parametrami strukturalnymi i stochastycznymi

$y_i = 0{,}1096 \cdot x_i + 1{,}9410 + \boldsymbol{e_i}$
[0,0239] [0,3427] [0,5449]

Zbadać istotność współczynnika regresji (γ=0,05)

c) $H_0: \alpha = 0$, nie istnieje regresja liniowa ocen względem czasu nauki w populacji generalnej, współczynnik regresji jest statystycznie nieistotny

$H_1: \alpha \neq 0$, występuje regresja liniowa ocen względem czasu nauki w populacji generalnej, współczynnik regresji jest statystycznie istotny

$$t = \frac{\hat{\alpha}}{s_{\hat{\alpha}}} = \frac{0{,}1096}{0{,}0239} = 4{,}586$$

γ=0,05 ; $v=n-2=10-2=8$

$\lambda = \{ t: t \in (-\infty; -t_{\gamma,v}\rangle \cup \langle t_{\gamma,v}; +\infty)\}$

$\lambda = \{t: t \in (-\infty; -2{,}306\rangle \cup \langle 2{,}306; +\infty)\}$

$t \in \lambda$

Ponieważ obliczona wartość testu znalazła się w obszarze odrzuceń to istnieją statystyczne podstawy do odrzucenia hipotezy H_0 i przyjęcia hipotezy H_1. Uznajemy, że istnieje regresja liniowa ocen względem czasu nauki. Decyzję podjęto w przy poziomie istotności γ=0,05

Ocenić dopasowanie modelu regresji liniowej do danych empirycznych,

d) $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 8,600=6,225+2,375

Współczynnik determinacji

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{c_{xy}}{s_x^2 \cdot s_y^2} = \frac{c_{xy}\cdot\hat{\alpha}}{s_y^2} = \frac{6{,}225}{8{,}600} = 0{,}724$$

$$R^2 \approx 1 - \frac{s_e^2}{s_y^2}$$

$R^2 = r_{xy}^2$

Zmienność czasu nauki wyjaśnia zmienność ocen w 72,4%.

Współczynnik indeterminacji $\varphi^2 = 1 - R^2$

$R^2 + \varphi^2 = 1$

$$\phi^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{c_{xy}^2}{s_x^2 \cdot s_y^2} = 1 - \frac{c_{xy}\cdot\hat{\alpha}}{s_y^2} = \frac{2{,}375}{8{,}600} = 0{,}276$$

$$\phi^2 \approx \frac{s_e^2}{s_y^2}$$

Zmienność czasu nauki nie wyjaśnia zmienności ocen w 27,6%.

Jakiej oceny powinien oczekiwać student uczący się 23 godziny (predykcja punktowa i przedziałowa dla $\gamma=0{,}05$)

e) $x_p=23$; $\hat{y}_p = \hat{\alpha}\cdot x_p + \hat{\beta}$

Predykcja punktowa:

$$\hat{y}_p = 0{,}1096 \cdot 23 + 1{,}9410 = 4{,}4618 \approx 4{,}5$$

Prognozowana ocena dla studenta uczącego się 23 godziny wynosi 4,5.

$$s_{(\hat{y}_p)} = s_e\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = 0{,}5449\sqrt{1+\frac{1}{10}+\frac{(23-12{,}4)^2}{518{,}4}} = 0{,}6253$$

Prognozując ocenę dla studenta uczącego się 23 godziny standardowy błąd prognozy wynosi 0,6253.

Predykcja przedziałowa:

$$P(\hat{y}_p - t_{\gamma,n-2}\cdot s_{(\hat{y}_p)} < y_p < \hat{y}_p + t_{\gamma,n-2}\cdot s_{(\hat{y}_p)}) = 1-\gamma$$

$4{,}4618 - 2{,}306\cdot 0{,}6253 < y_p < 4{,}4618 + 2{,}306\cdot 0{,}6253 \quad \Rightarrow \quad 3{,}0 < y_p < 5{,}9$

Przedział o końcówka (3,0 ; 5,9) jest jednym z przedziałów, które z prawdopodobieństwem 0,95 obejmują prognozowaną ocenę studenta uczącego się 23 godziny.