Zadania z wykładu 11

Krystian Baran 145000

25 maja 2021

Spis treści

1 Zadanie 1

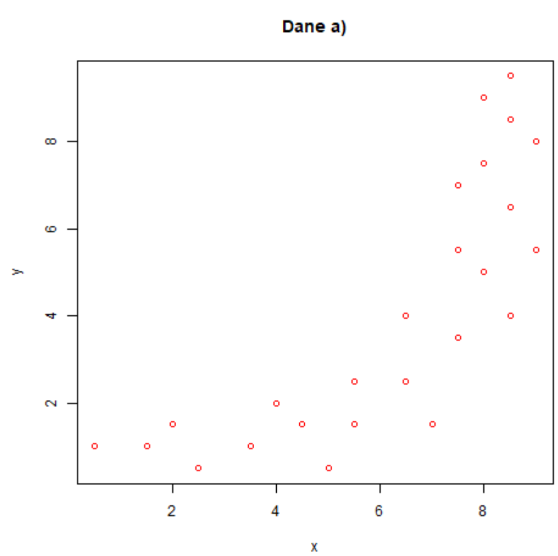
Sporządzić diagram rozrzutu, wyznaczyć oceny współczynników korelacji i determinacji, wyznaczyć równania prostych regresji (Y względem X, X względem Y), błędy standardowe estymacji oraz wykreślić równanie regresji dla podanych prób:

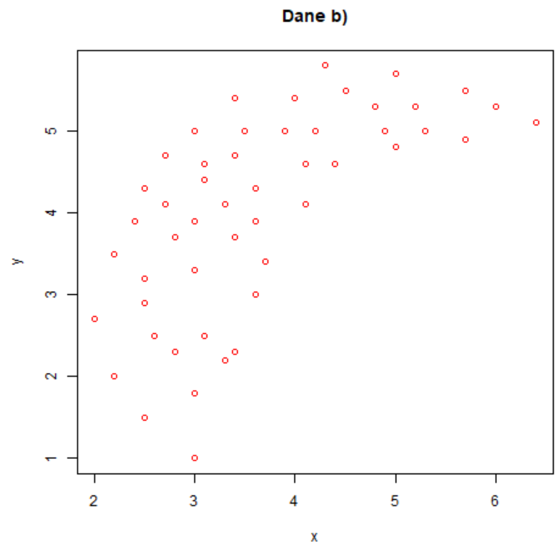
a) $[x; y] =$
$\{[5.5, 1.5], [8.5, 4.0], [4.0, 2.0], [8.0, 7.5], [2.5, 0.5], [8.0, 5.0], [8.5, 8.5], [3.5, 1.0],$
$[6.5, 2.5], [9.0, 8.0], [0.5, 1.0], [8.5, 6.5], [7.5, 3.5], [1.5, 1.0], [8.5, 9.5],$
$[2.0, 1.5], [8.0, 9.0], [7.5, 5.5], [9.0, 5.5], [7.0, 1.5], [7.5, 7.0], [5.0, 0.5],$
$[4.5, 1.5], [5.5, 2.5], [6.5, 4.0]\}$.

b) $[x; y] =$
$\{[3.4, 3.7], [2.7, 4.7], [4.4, 4.6], [2.6, 2.5], [5.2, 5.3], [3.1, 4.6], [2.2, 3.5], [3.3, 4.1],$
$[6.0, 5.3], [4.0, 5.4], [2.0, 2.7], [3.9, 5.0], [2.5, 1.5], [2.5, 4.3], [3.6, 3.0], [6.4, 5.1],$
$[2.8, 3.7], [4.3, 5.8], [5.7, 5.5], [2.5, 3.2], [4.9, 5.0], [3.0, 1.8], [3.6, 4.3], [5.7, 4.9],$
$[3.0, 1.0], [4.1, 4.1], [5.0, 4.8], [2.2, 2.0], [3.7, 3.4], [5.0, 5.7], [3.1, 4.4], [3.4, 5.4],$
$[3.4, 2.3], [2.5, 2.9], [5.3, 5.0], [4.1, 4.6], [3.0, 5.0], [2.8, 2.3], [3.0, 3.9], [2.4, 3.9],$
$[4.5, 5.5], [3.5, 5.0], [4.8, 5.3], [3.1, 2.5], [2.7, 4.1], [3.0, 3.3], [4.2, 5.0], [3.3, 2.2],$
$[3.6, 3.9], [3.4, 4.7]\}$.

1.1 Diagram rozrzutu

Jako pierwsze sporządzono diagram rozrzutu gdzie na osi X są wartości x, a na osi Y wartości y. Wykres sporządzono w R.





1.2 Współczynnik korelacji

Obliczymy teraz współczynnik korelacji zgodnie ze wzorem:

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

Gdzie:

$$SS_{xx} = \sum(x_i^2) - \frac{(\sum x_i)^2}{n} \overset{R}{=} sum(x^2) - sum(x)^2/length(x)$$

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n} \overset{R}{=} sum(x * y) - sum(x) * sum(y)/length(x)$$

Otrzymano następujące wartości:

	SS_{xx}	SS_{yy}	SS_{xy}	r
a)	155.64	209.74	142.44	0.7883711
b)	57.4248	74.1122	42.3684	0.6494526

Ponieważ oba r są dodatnie możemy sformułować hipotezę że współczynnik korelacji pomiędzy x i y jest dodatni.

H_0	$\rho \leqslant 0$
H_1	$\rho > 0$

Aby sprawdzić tę hipotezę zastosujemy następującą statystykę:

$$Z = (U - u_0) \cdot \sqrt{n-3}$$

Która, dla $n > 7$ ma w przybliżeniu standardowy rozkład normalny. Poniżej przedstawiono obliczenia dla:

a)

$$U = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln(\frac{1.7883711}{0.2116289} \approx 1.067113$$

$$u_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = 0$$

$$Z_0 = 2.51587 \cdot \sqrt{10-3} \approx 5.005205$$

b)

$$U = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln(\frac{1.6494526}{0.3505474} \approx 0.7743514$$

$$u_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = 0$$

$$Z_0 = 2.51587 \cdot \sqrt{10-3} \approx 5.308686$$

Wtedy można obliczyć *p-value* dla oby prób, zgodnie ze wzorem:

$$\text{p-value}_a = 1 - \Phi(5.005205) \stackrel{R}{=} 1 - pnorm(5.005205, 0, 1) \approx 2.790134e-07$$

$$\text{p-value}_b = 1 - \Phi(5.308686) \stackrel{R}{=} 1 - pnorm(5.308686, 0, 1) \approx 5.520921e-08$$

Przyjmując $\alpha = 0.05$ oba *p-value* są mniejsze od α; zatem odrzucamy hipotezę zerową i wnioskujemy że korelacja pomiędzy x i y jest typu dodatniego, co widać na wykresach i z obliczonych wartości r.

1.3 Współczynnik determinacji i równania regresji

Aby wyznaczyć współczynnik determinacji potrzebne jest wyliczenie SSE, które wyraża się następująco:

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

lub

$$\text{SSE} = \sum (x_i - \hat{\text{x}}_i)^2$$

Zatem potrzebujemy najpierw wyznaczyć równanie regresji. Do tego równania potrzebujemy dwa współczynniki:

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \stackrel{R}{=} mean(y) - b1 * mean(x)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = b1$$

Natomiast dla X zależne od Y parametry są następujące:

$$\beta_0 = \overline{x} - \beta_1 \overline{y} \overset{R}{=} mean(x) - b1 * mean(y)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{yy}} = b1$$

Rozważymy najpierw Y zależne od X. Równania wyglądają następująco:

$$y = \beta_0 + \beta_1 \cdot x$$
$$y_a = -1.580956 + 0.9151889 \cdot x$$
$$y_b = 1.342481 + 0.7378067 \cdot x$$

Wtedy parametry SSE wynoszą:

	SSE
a)	79.38049
b)	42.85251

Możemy teraz wyznaczyć współczynnik determinacji, który wynosi:

$$r^2 = 1 + \frac{SSE}{SS_{yy}}$$

	r^2
a)	0.6215291
b)	0.4217887

Następnie rozważymy dla X zależnego od Y. Wtedy, analogicznie do wcześniej:

$$x = \beta_0 + \beta_1 \cdot y$$
$$x_a = 3.389911 + 0.6791265 \cdot y$$
$$x_b = 1.341846 + 0.5716792 \cdot y$$

Wtedy parametry SSE wynoszą:

	SSE
a)	58.90522
b)	33.20367

Możemy teraz wyznaczyć współczynnik determinacji, który wynosi:

$$r^2 = 1 + \frac{SSE}{SS_{yy}}$$

	r^2
a)	0.6215291
b)	0.4217887

Wartości r^2 nie zmieniają się, zatem, równanie regresji zmniejsza całkowitą sumę kwadratów o 62% dla próby a) i 42% dla próby b) od średniej arytmetycznej.

1.4 Błąd modelu

Następnie obliczymy błędy modelu zgodnie ze wzorem:

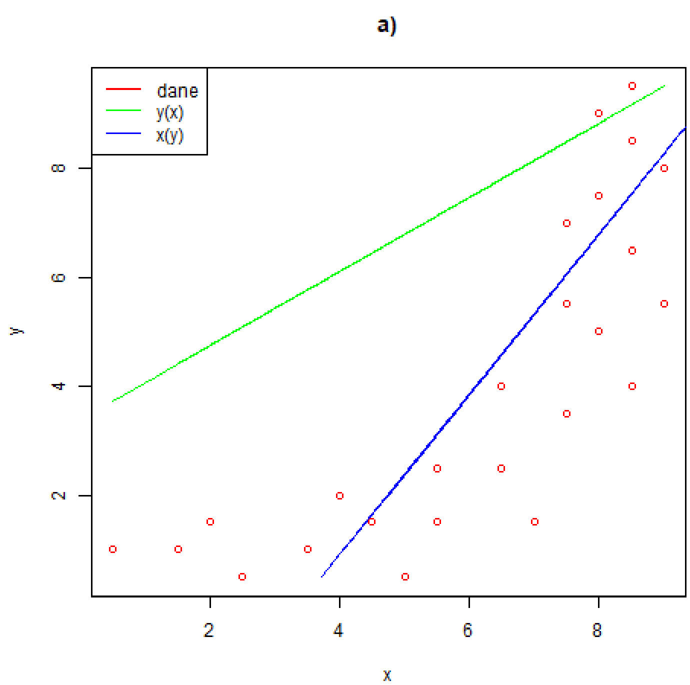
$$S^2 = \frac{SSE}{n-2}$$

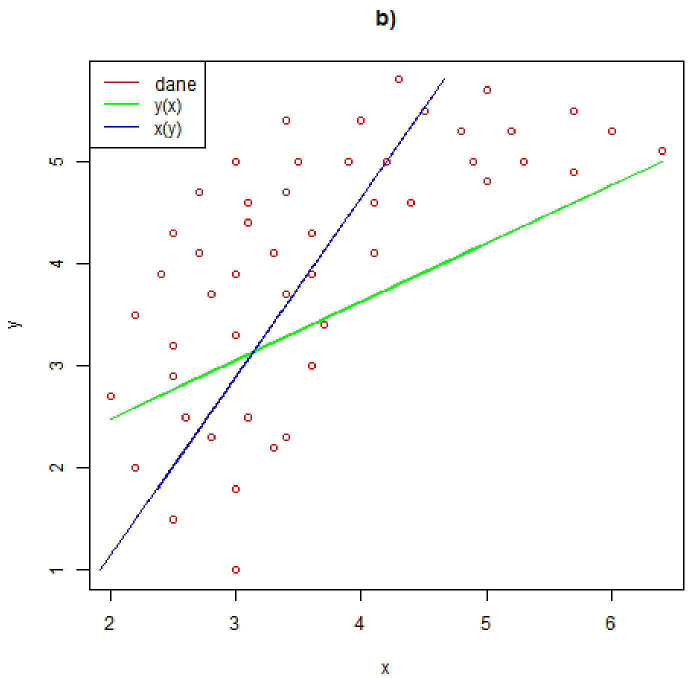
Liczebności prób są, dla a) $n = 25$, dla b) $n = 50$. Wtedy błędy modelu są następujące:

	Y od X	X od Y
a)	3.451326	2.561096
b)	0.8927607	0.6917431

1.5 Wykresy regresji

Jako ostatnie przedstawiono wykresy prostych regresji na wykresach z danymi.





Dla danych a) prosta x(y) lepiej obrazuje przebieg danych, natomiast dla danych b) nie ma znacznej różnicy pomiędzy prostymi.