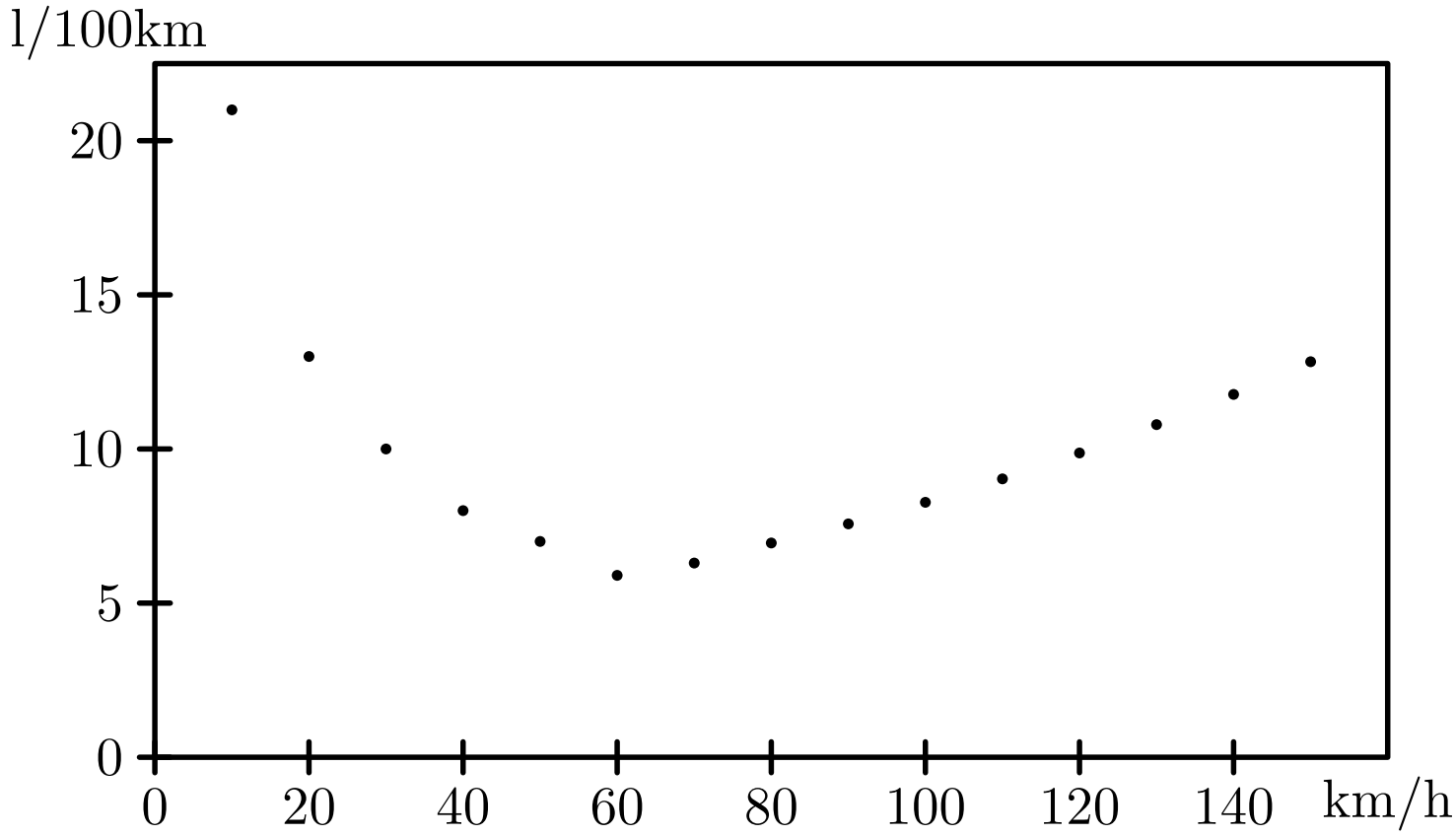
Elementarna statystyka
Podstawowa analiza zależności w danych
(Examining relationships)

Alexander Bendikov

6 kwietnia 2016

Czy jadąc szybko marnujemy paliwo? W tabeli są dane dotyczące zużycia paliwa (brytyjska wersja Forda Escorta)

Prędkość (km/h)	Zużycie paliwa (l/100 km)	Prędkość (km/h)	Zużycie paliwa (l/100 km)
10	21,00	90	7,57
20	13,00	100	8,27
30	10,00	110	9,03
40	8,00	120	9,87
50	7,00	130	10,79
60	5,90	140	11,77
70	6,30	150	12,83
80	6,95		



Rysunek: Wykres punktowy zmiennych „Prędkość” i „Zużycie paliwa”

W przykładzie są dwie zmienne zależne:
X (prędkość), *zmienna objaśniająca*, która jest zmienną decydującą w tej zależności,
Y (zużycie paliwa), *zmienna zależna*, która jest zmienną reagującą.

W przykładzie są dwie zmienne zależne:
X (prędkość), *zmienna objaśniająca*, która jest zmienną decydującą w tej zależności,
Y (zużycie paliwa), *zmienna zależna*, która jest zmienną reagującą.
Główne zadanie to objaśnienie rodzaju zależności

$$X \longleftrightarrow Y.$$

W przykładzie są dwie zmienne zależne:
X (prędkość), *zmienna objaśniająca*, która jest zmienną decydującą w tej zależności,
Y (zużycie paliwa), *zmienna zależna*, która jest zmienną reagującą.
Główne zadanie to objaśnienie rodzaju zależności

$$X \longleftrightarrow Y.$$

Wykres punktowy pokazuje zależność pomiędzy dwoma zmiennymi ilościowymi X i Y. Poszczególne obserwacje zbioru danych odpowiadają punktom wykresu.

Współczynnik korelacji $R_{X,Y}$

Niech X i Y będą zmiennymi losowymi, ze średnimi i odchyleniami standardowymi odpowiednio $\mu_X, \sigma_X, \mu_Y, \sigma_Y$. Jeżeli X i Y są *niezależne* to

$$Var(X + Y) = Var(X) + Var(y) = \sigma_X^2 + \sigma_Y^2.$$

Współczynnik korelacji $R_{X,Y}$

Niech X i Y będą zmiennymi losowymi, ze średnimi i odchyleniami standardowymi odpowiednio $\mu_X, \sigma_X, \mu_Y, \sigma_Y$. Jeżeli X i Y są *niezależne* to

$$Var(X+Y) = Var(X) + Var(y) = \sigma_X^2 + \sigma_Y^2.$$

Jeżeli X i Y nie są niezależne, to

$$Var(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2 \cdot \sigma_X \cdot \sigma_Y \cdot R_{X,Y},$$

gdzie

$$R_{X,Y} = E\Big(\frac{X-\mu_X}{\sigma_X}\Big)\Big(\frac{Y-\mu_Y}{\sigma_Y}\Big).$$

Współczynnik korelacji $R_{X,Y}$

Niech X i Y będą zmiennymi losowymi, ze średnimi i odchyleniami standardowymi odpowiednio $\mu_X, \sigma_X, \mu_Y, \sigma_Y$. Jeżeli X i Y są *niezależne* to

$$Var(X+Y) = Var(X) + Var(y) = \sigma_X^2 + \sigma_Y^2.$$

Jeżeli X i Y nie są niezależne, to

$$Var(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2 \cdot \sigma_X \cdot \sigma_Y \cdot R_{X,Y},$$

gdzie

$$R_{X,Y} = E\Big(\frac{X-\mu_X}{\sigma_X}\Big)\Big(\frac{Y-\mu_Y}{\sigma_Y}\Big).$$

Wielkość $R_{X,Y}$ nazywamy *współczynnikiem korelacji* zmiennych X i Y

Własności $R_{X,Y}$

1. $-1 \leq R_{X,Y} \leq 1$,

Własności $R_{X,Y}$

1. $-1 \leq R_{X,Y} \leq 1$,
2. $R_{X,Y} = \pm 1 \Leftrightarrow X, Y$ są *liniowo zależne*, to znaczy

$$Y = k\,X + b, \quad \text{lub} \quad X = k\,Y + b.$$

W takim przypadku mamy dodatkowo $R_{X,Y} = 1$ jeżeli $k > 0$ i $R_{X,Y} = -1$ jeżeli $k < 0$.

Własności $R_{X,Y}$

1. $-1 \leq R_{X,Y} \leq 1$,
2. $R_{X,Y} = \pm 1 \Leftrightarrow X, Y$ są *liniowo zależne*, to znaczy

$$Y = k\,X + b, \quad \text{lub} \quad X = k\,Y + b.$$

   W takim przypadku mamy dodatkowo $R_{X,Y} = 1$ jeżeli $k > 0$ i $R_{X,Y} = -1$ jeżeli $k < 0$.

3. Współczynnik korelacji mierzy siłę współzależności typu *liniowego*. Nie opisuje dobrze zależności krzywoliniowych.

Współczynnik korelacji w próbie $r_{X,Y}$

Załóżmy, że mamy próbki $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$ pobrane z populacji o rozkładach X i Y odpowiednio. Możemy korzystać z przybliżeń $\overline{x} \approx \mu_x$, $s_x \approx \sigma_x$, $\overline{y} \approx \mu_y$, $s_y \approx \sigma_y$.

Współczynnik korelacji w próbie $r_{X,Y}$

Załóżmy, że mamy próbki $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$ pobrane z populacji o rozkładach X i Y odpowiednio. Możemy korzystać z przybliżeń $\overline{x} \approx \mu_x$, $s_x \approx \sigma_x$, $\overline{y} \approx \mu_y$, $s_y \approx \sigma_y$.
A w jaki sposób możemy przybliżyć współczynnik korelacji $R_{X,Y}$?

Współczynnik korelacji w próbie $r_{X,Y}$

Załóżmy, że mamy próbki $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$ pobrane z populacji o rozkładach X i Y odpowiednio. Możemy korzystać z przybliżeń $\overline{x} \approx \mu_x$, $s_x \approx \sigma_x$, $\overline{y} \approx \mu_y$, $s_y \approx \sigma_y$.
A w jaki sposób możemy przybliżyć współczynnik korelacji $R_{X,Y}$?

- $r_{X,Y} = \frac{1}{n-1} \sum_i (\frac{x_i - \overline{x}}{s}_x)(\frac{y_i - \overline{y}}{s}_y)$,
- $r_{X,Y} \approx R_{X,Y}$ dla $n \gg 1$.

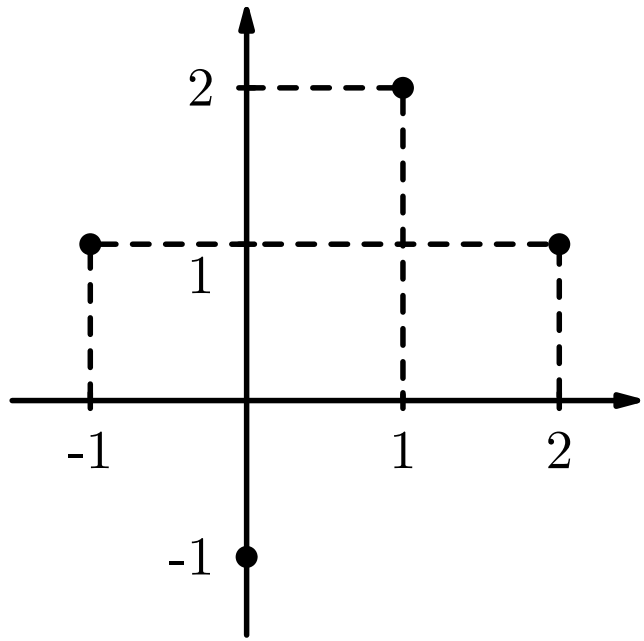
Współczynnik korelacji w próbie $r_{X,Y}$

Załóżmy, że mamy próbki $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$ pobrane z populacji o rozkładach X i Y odpowiednio. Możemy korzystać z przybliżeń $\overline{x} \approx \mu_x$, $s_x \approx \sigma_x$, $\overline{y} \approx \mu_y$, $s_y \approx \sigma_y$.
A w jaki sposób możemy przybliżyć współczynnik korelacji $R_{X,Y}$?

- $r_{X,Y} = \frac{1}{n-1} \sum_i (\frac{x_i - \overline{x}}{s}_x)(\frac{y_i - \overline{y}}{s}_y)$,
- $r_{X,Y} \approx R_{X,Y}$ dla $n \gg 1$.

Przykład:

$x.$	$y.$	$x. - \overline{x}$	$y. - \overline{y}$
-1	1	-1,5	0,25
0	-1	-0,5	-1,75
1	2	0,5	1,25
2	1	1,5	0,25
$\overline{x} = 0,5$	$\overline{y} = 0,75$	$s_x = 1,3$	$s_y = 1,2$



Rysunek: Wykres punktowy zmiennych X i Y

$$\sum(x. - \overline{x})(y. - \overline{y}) = 1,5$$

$$r_{X,Y} = 0,32.$$

Własności $r_{X,Y}$

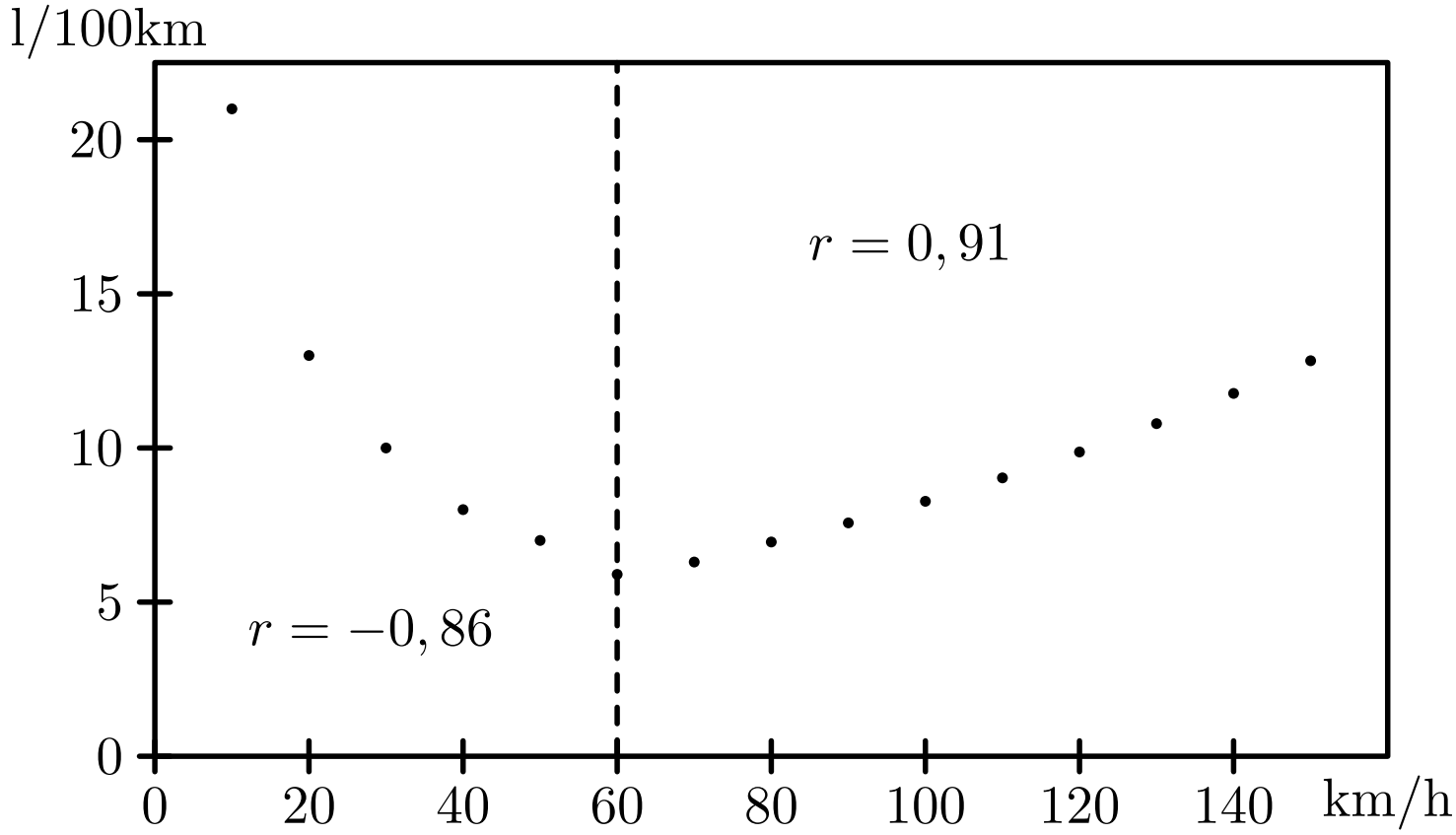
1. $-1 \leq r \leq 1$,
2. $r = \pm 1$ wtedy i tylko wtedy, gdy wszystkie obserwacje leżą na jednej prostej. Czyli $r = \pm 1$ tylko w przypadku idealnie liniowej zależności.
3. $r \approx 0$ oznacza bardzo słaba zależność liniową.

Własności $r_{X,Y}$

1. $-1 \leq r \leq 1$,
2. $r = \pm 1$ wtedy i tylko wtedy, gdy wszystkie obserwacje leżą na jednej prostej. Czyli $r = \pm 1$ tylko w przypadku idealnie liniowej zależności.
3. $r \approx 0$ oznacza bardzo słaba zależność liniową.

Przykład: W przypadku zużycia paliwa mamy:

- 1 zakres, $10 - 60$ km/h $r = -0,86$
- 2 zakres, $60 = 150$ km/h $r = 0,91$
- W całym zakresie prędkości $10 - 150$km/h mamy $r = -0,15$ - bardzo słaba zależność liniowa



Rysunek: Wykres punktowy, 2 zakresy

Przykłady związane z korelacją

1. Galton (1857) 1078 par pomiarów wzrostów:
   - Ojcowie i synowie: $r \approx 0,5$,
   - Matki i synowie: $r \approx 0,494$.
2. Badania związane z ochroną zdrowia (1960-62)
   - Wzrosty i wagi 411 mężczyzn w wieku 18-24 lat:

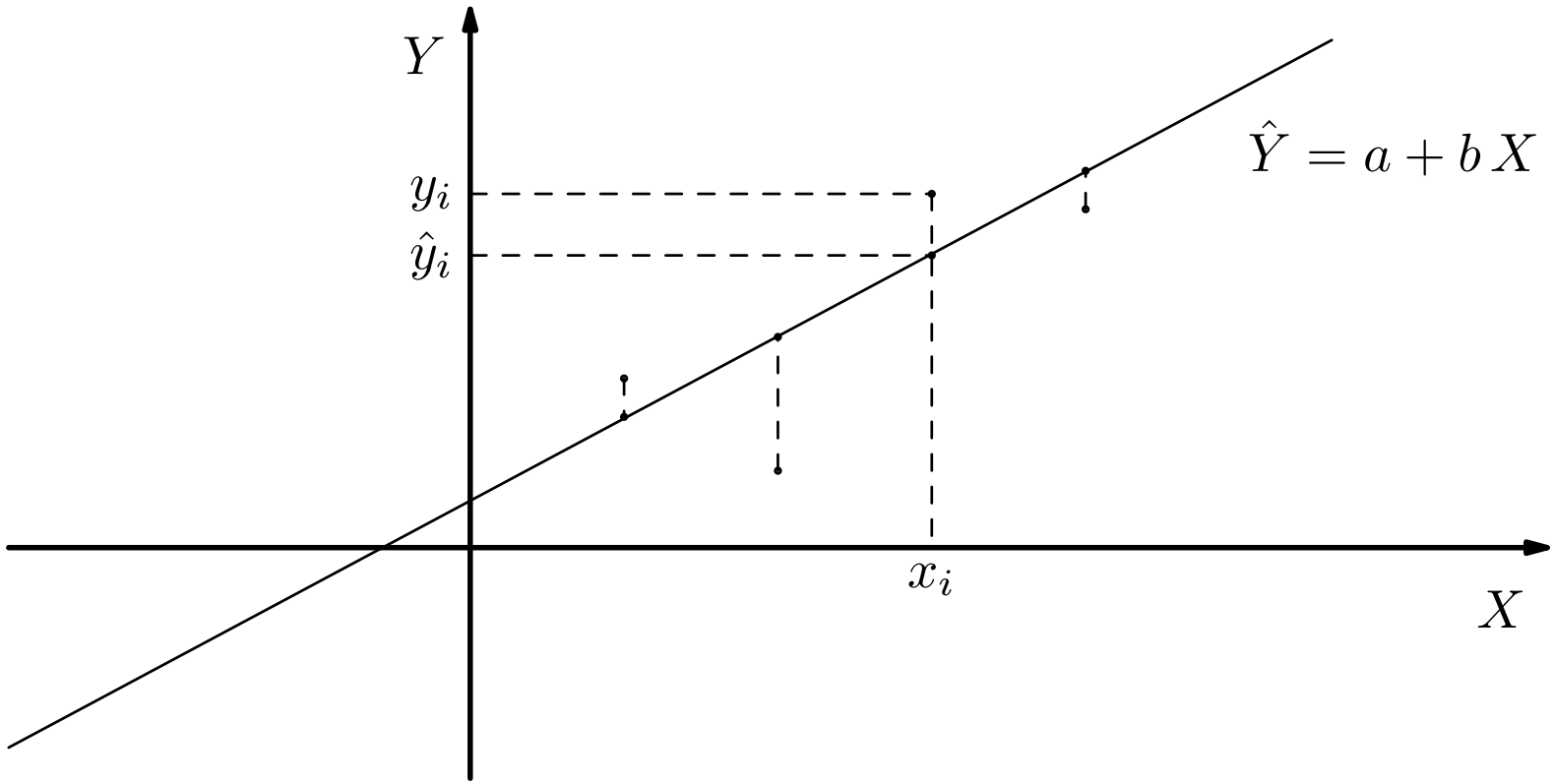
     $$r \approx 0,36.$$

   - Wykształcenie i dochód:
     (a) dla mężczyzn w wieku 25-34: $r \approx 0,4$,
     (b) dla mężczyzn w wieku 35-44: $r \approx 0,6$.
3. Iloraz inteligencji identycznych bliźniaków:

   $$r \approx 0,95.$$

Linia regresji najmniejszych kwadratów

Zasada: linia regresji najmniejszych kwadratów zmiennych X i Y jest prostą o równaniu $\hat{Y} = a + b\,X$ dla której suma kwadratów $\sum(y_\cdot - \hat{y}_\cdot)^2$ jest najmniejsza



Rysunek: Linia regresji najmniejszych kwadratów

Musimy więc znaleźć a, b takie, że

$$\sum (y_i - \hat{y}_i)^2 \longrightarrow \min,$$

gdzie

- y_i jest obserwacją zmiennej Y,
- $\hat{y}_i = a + b\ x_i$ jest przewidywaną wartością zmiennej Y, odpowiadającą obserwacji x_i zmiennej X
- $y_i - \hat{y}_i$ jest resztą.

Rozwiązanie problemu minimalizacji: Linia regresji najmniejszych kwadratów ma równanie

$$\hat{Y} = a + bX,$$

gdzie:

1. Współczynnik kierunkowy $b = r \cdot \frac{s_Y}{s_X}$
2. Odsunięcie $a = \overline{y} - b\,\overline{x}$

Równoważnie:

$$\frac{\hat{Y} - \overline{y}}{s_Y} = r \cdot \frac{\hat{X} - \overline{x}}{s_X}.$$

Przykład: Dla naszego Forda Escorta mamy:

1. W zakresie $10 - 60$ km/h mamy

$$\hat{Y} = -0,3X + 21,5,$$

a więc następujące prognozy: $x = 25, \hat{y} = 14$, $x = 40, \hat{y} = 9,5$, $x = 50, \hat{y} = 6,5$, $x = 70, \hat{y} = 0,5$?!

Przykład: Dla naszego Forda Escorta mamy:

1. W zakresie $10 - 60$ km/h mamy

$$\hat{Y} = -0,3X + 21,5,$$

a więc następujące prognozy: $x = 25, \hat{y} = 14$, $x = 40, \hat{y} = 9,5$, $x = 50, \hat{y} = 6,5$, $x = 70, \hat{y} = 0,5$?!

2. W całym zakresie $10 - 150$ km/h mamy

$$\hat{y} = -0,01466\ X + 11,058,$$

a więc następujące prognozy: $x = 25, \hat{y} = 10,65$, $x = 40, \hat{y} = 9,32$, $x = 70, \hat{y} = 10,03$!

r^2 jako ułamek zmienności

$$r^2 = \frac{\text{Całkowita zmienność (wariancja) wartości prognozowanych } \hat{Y}}{\text{Całkowita zmienność (wariancja) wartości obserwowanych } Y}.$$

r^2 jako ułamek zmienności

$$r^2 = \frac{\text{Całkowita zmienność (wariancja) wartości prognozowanych } \hat{Y}}{\text{Całkowita zmienność (wariancja) wartości obserwowanych } Y}.$$

Powyższy wzór łatwo jest uzasadnić korzystając z równania regresji

$$\frac{\hat{y}_. - \overline{y}}{s_Y} = r \cdot \frac{\hat{x}_. - \overline{x}}{s_X}.$$

r^2 jako ułamek zmienności

$$r^2 = \frac{\text{Całkowita zmienność (wariancja) wartości prognozowanych } \hat{Y}}{\text{Całkowita zmienność (wariancja) wartości obserwowanych } Y}.$$

Powyższy wzór łatwo jest uzasadnić korzystając z równania regresji

$$\frac{\hat{y}_. - \overline{y}}{s_Y} = r \cdot \frac{\hat{x}_. - \overline{x}}{s_X}.$$

1) $\sum \frac{\hat{y}_. - \overline{y}}{s_Y} = r \cdot \sum \frac{\hat{x}_. - \overline{x}}{s_X} = 0 \;\Rightarrow\; \overline{\hat{y}} = \overline{y}$

r^2 jako ułamek zmienności

$$r^2 = \frac{\text{Całkowita zmienność (wariancja) wartości prognozowanych } \hat{Y}}{\text{Całkowita zmienność (wariancja) wartości obserwowanych } Y}.$$

Powyższy wzór łatwo jest uzasadnić korzystając z równania regresji

$$\frac{\hat{y}_. - \overline{y}}{s_Y} = r \cdot \frac{\hat{x}_. - \overline{x}}{s_X}.$$

1) $\sum \frac{\hat{y}_. - \overline{y}}{s_Y} = r \cdot \sum \frac{\hat{x}_. - \overline{x}}{s_X} = 0 \;\Rightarrow\; \overline{\hat{y}} = \overline{y}$

2) $\frac{1}{n-1} \sum \left(\frac{\hat{y}_. - \overline{\hat{y}}}{s_Y} \right)^2 = r^2 \frac{1}{n-1} \sum \left(\frac{\hat{x}_. - \overline{x}}{s_X} \right)^2 = r^2$

r^2 jako ułamek zmienności

$$r^2 = \frac{\text{Całkowita zmienność (wariancja) wartości prognozowanych } \hat{Y}}{\text{Całkowita zmienność (wariancja) wartości obserwowanych } Y}.$$

Powyższy wzór łatwo jest uzasadnić korzystając z równania regresji

$$\frac{\hat{y}_\cdot - \overline{y}}{s_Y} = r \cdot \frac{\hat{x}_\cdot - \overline{x}}{s_X}.$$

1) $\sum \frac{\hat{y}_\cdot - \overline{y}}{s_Y} = r \cdot \sum \frac{\hat{x}_\cdot - \overline{x}}{s_X} = 0 \;\Rightarrow\; \overline{\hat{y}} = \overline{y}$

2) $\frac{1}{n-1} \sum \left(\frac{\hat{y}_\cdot - \overline{\hat{y}}}{s_Y} \right)^2 = r^2 \frac{1}{n-1} \sum \left(\frac{\hat{x}_\cdot - \overline{x}}{s_X} \right)^2 = r^2$

W końcu,

$$r^2 = \frac{\sum (\hat{y}_\cdot - \overline{\hat{y}})^2}{\sum (y_\cdot - \overline{y})^2}.$$

Innymi słowy, r^2 to procent zmienności Y, który można uzasadnić linią regresji.