ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

13/11/2014

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Regresja liniowa

Korelacja

Modelowanie

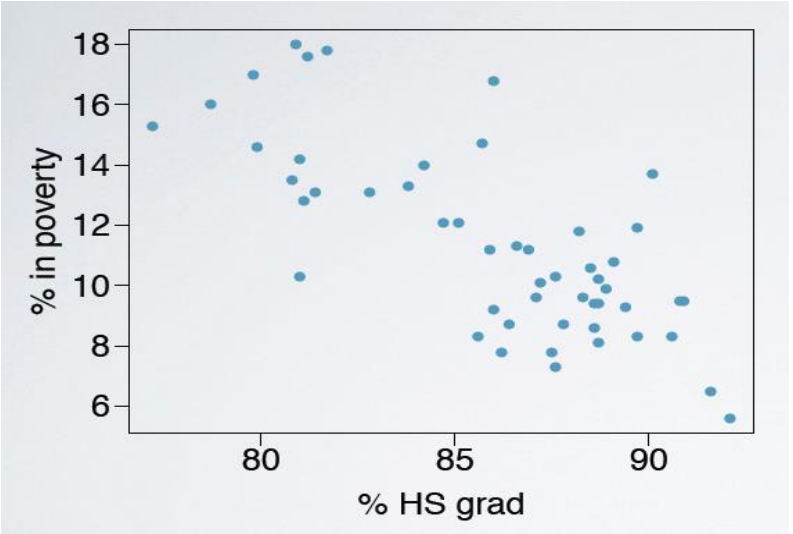
Analiza modelu

Wnioskowanie

Korelacja

Korelacja R:
charakteryzuje linową zależność pomiędzy dwoma zmiennymi.

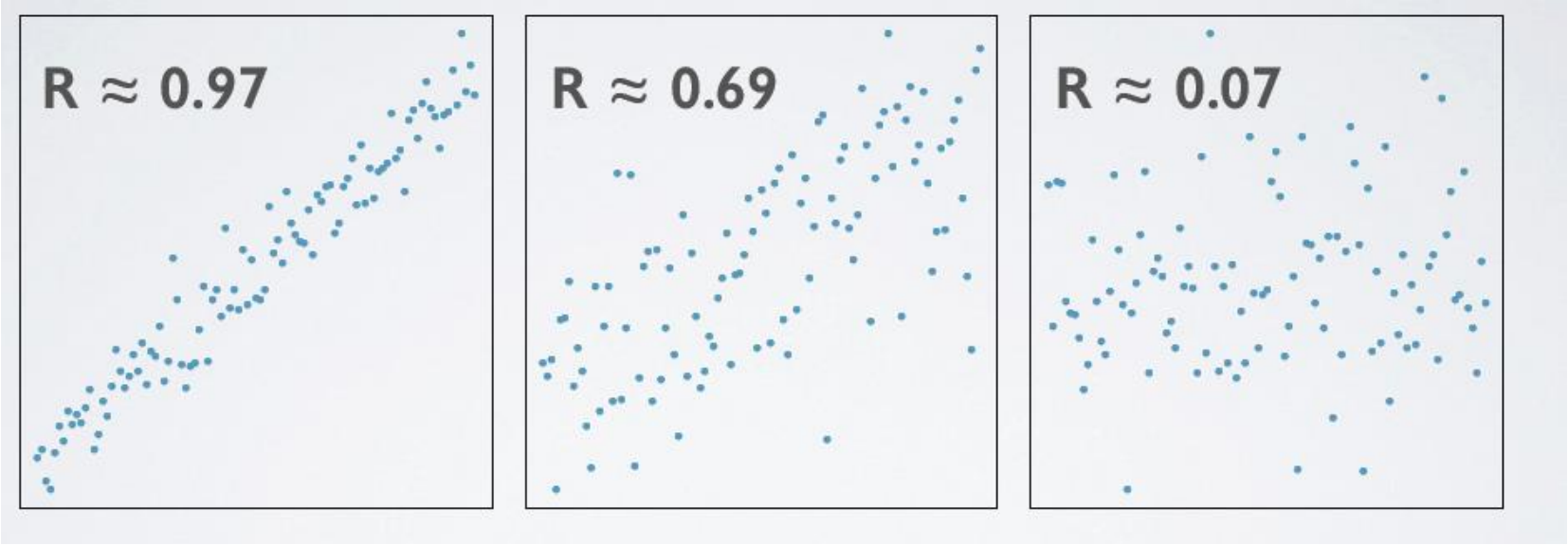
- Skutek: % z dochodem poniżej minimum
- Przyczyna: % z wyższym wykształceniem
- Relacja: liniowa, ujemny współczynik, średnio-silna



% HS grad: % z wyższym wykształceniem
% in poverty: dochód na rodzinę poniżej minimum
Dane: 50 stanów USA,
Minimum: 23 tys $/rok na rodzine 4 osobową

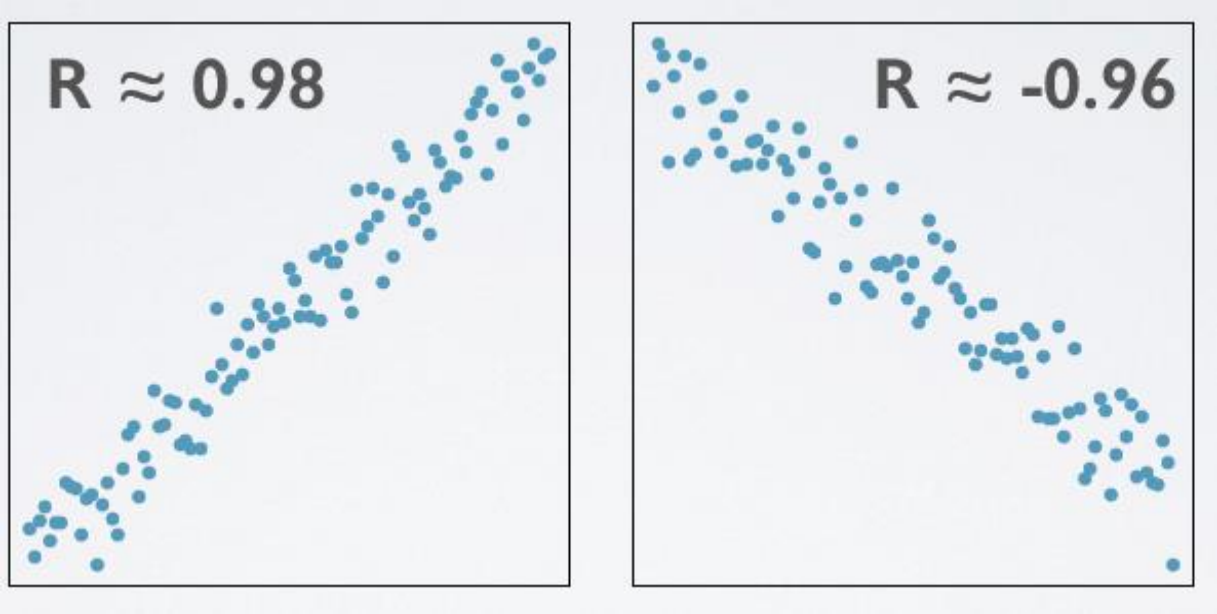
Korelacja: własności

- Wielkość (absolutna) współczynnika korelacji mierzy siłę liniowej asocjacji pomiędzy dwoma zmiennymi numerycznymi



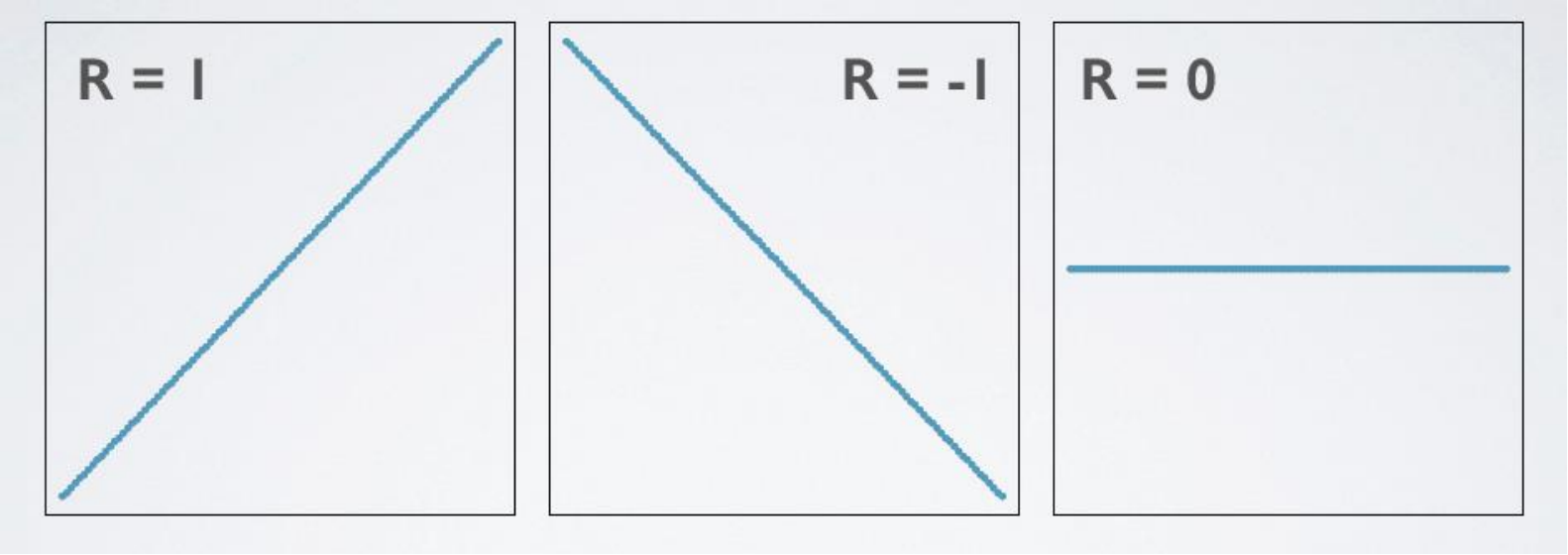
Korelacja: własności

- Znak korelacji wskazuje na kierunek asocjacji



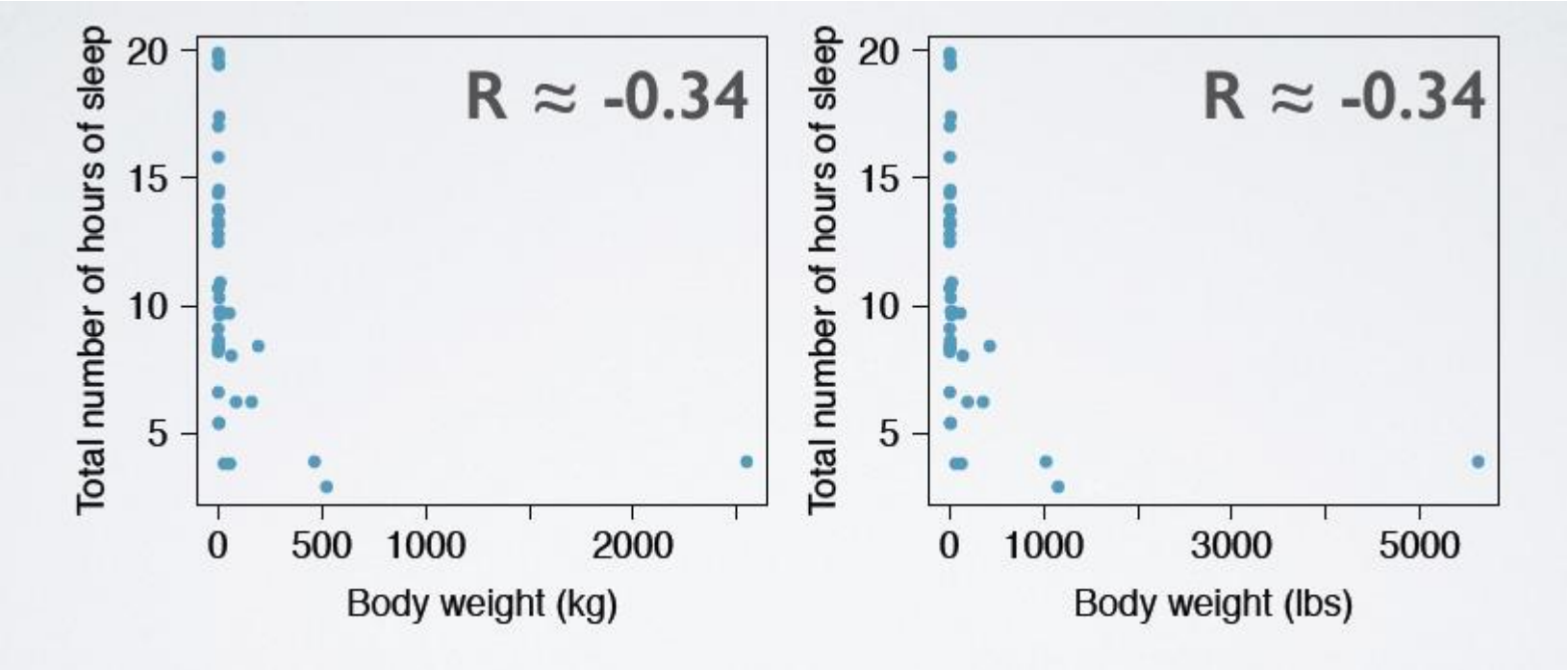
Korelacja: własności

- Z definicji, wartość R jest pomiędzy -1 (idealna negatywna asocjacja) do 1 (idealna pozytywna asocjacja)
- R = 0 wskazuje na brak zależności pomiędzy zmiennymi



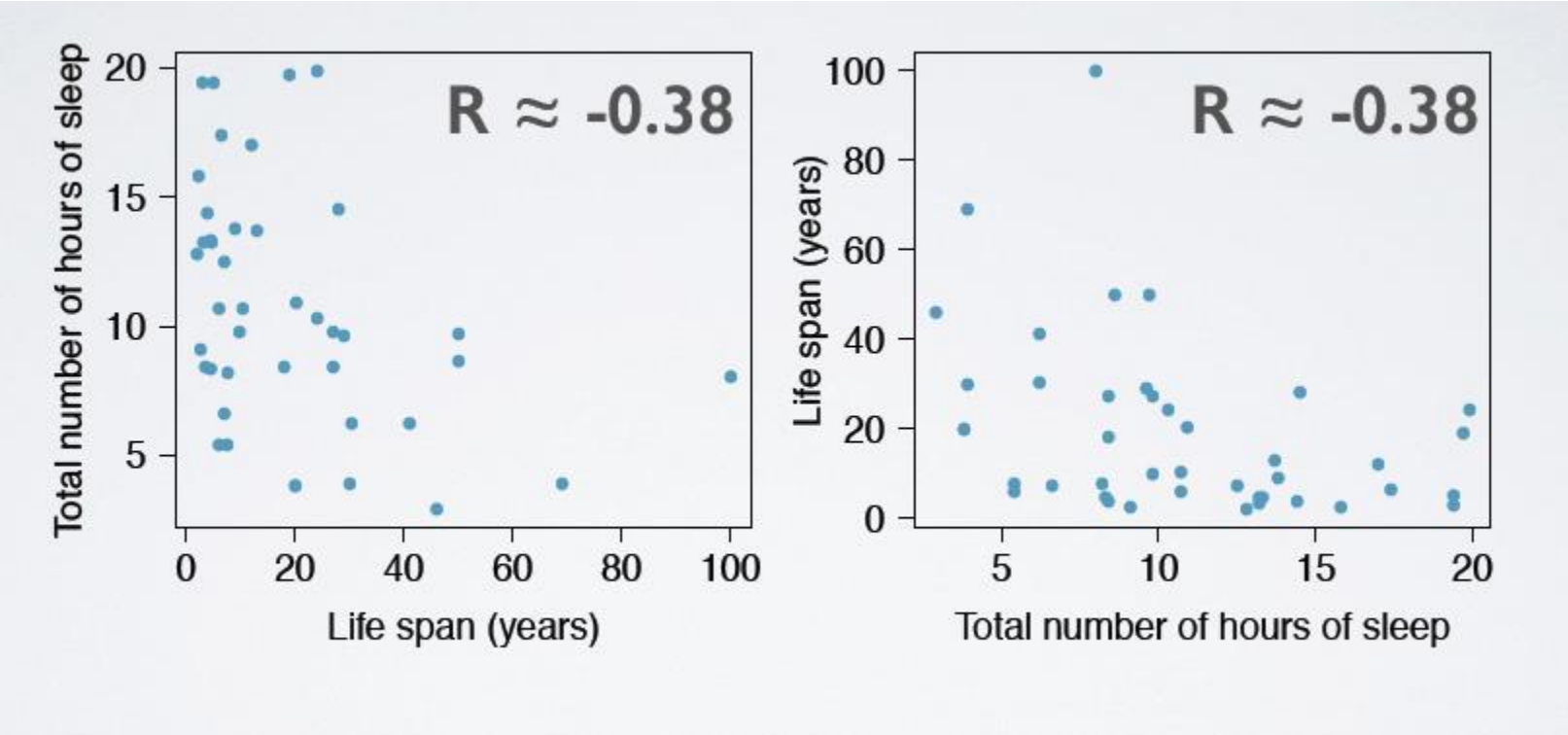
Korelacja: własności

- Współczynnik jest bezwymiarowy, nie zależy od skali żadnej ze zmiennych lub przesunięcia



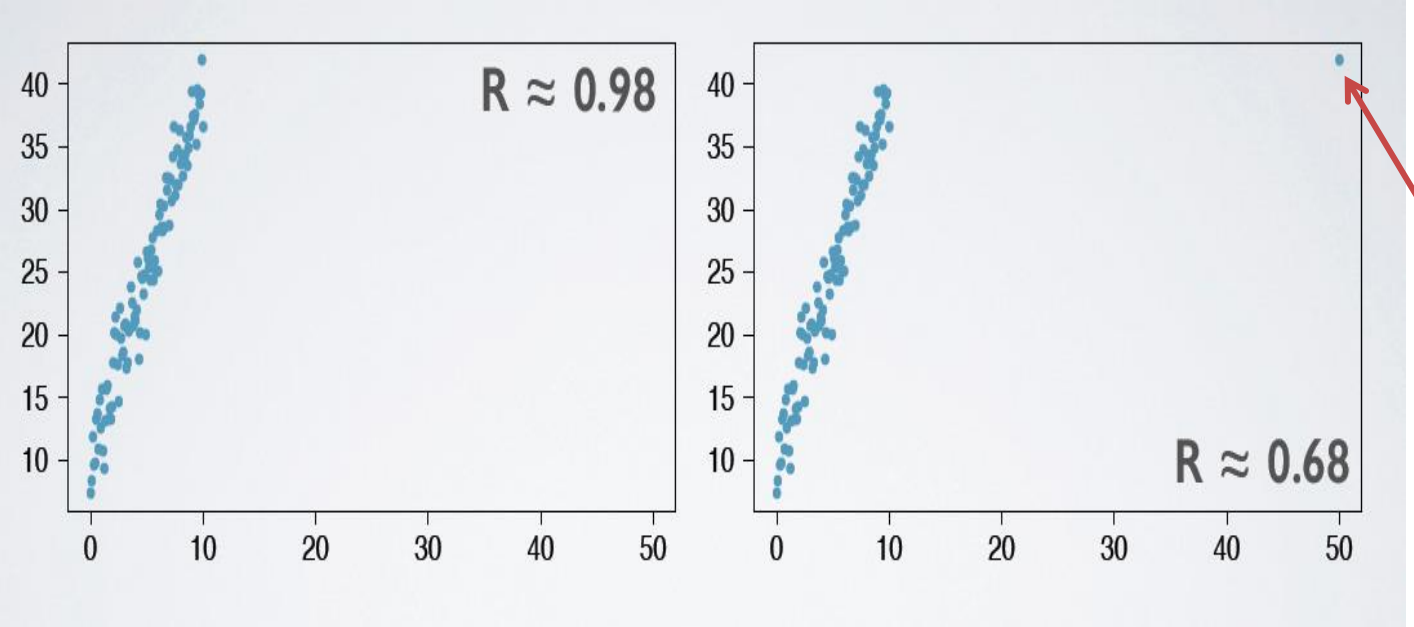
Korelacja: własności

- Korelacja pomiędzy X i Y jest taka sama jak pomiędzy Y i X



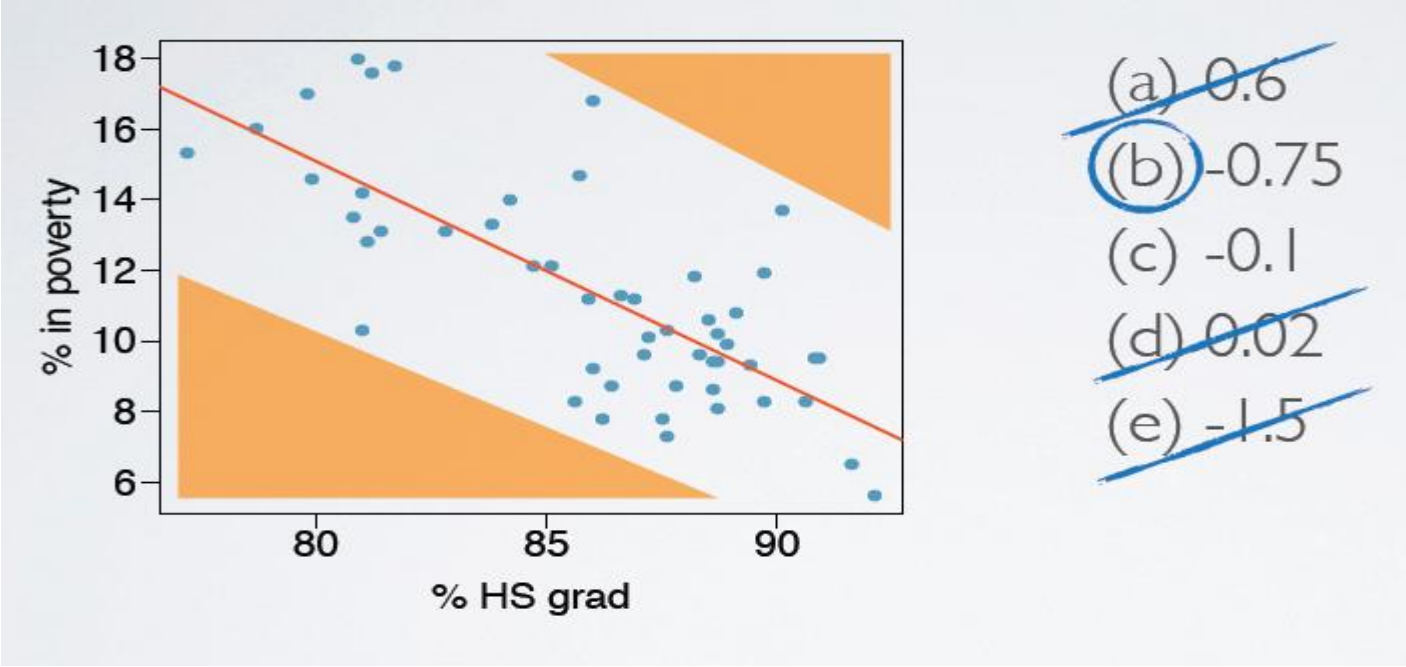
Korelacja: własności

- Korelacja R jest zależna od występowania tzw. „outliers”



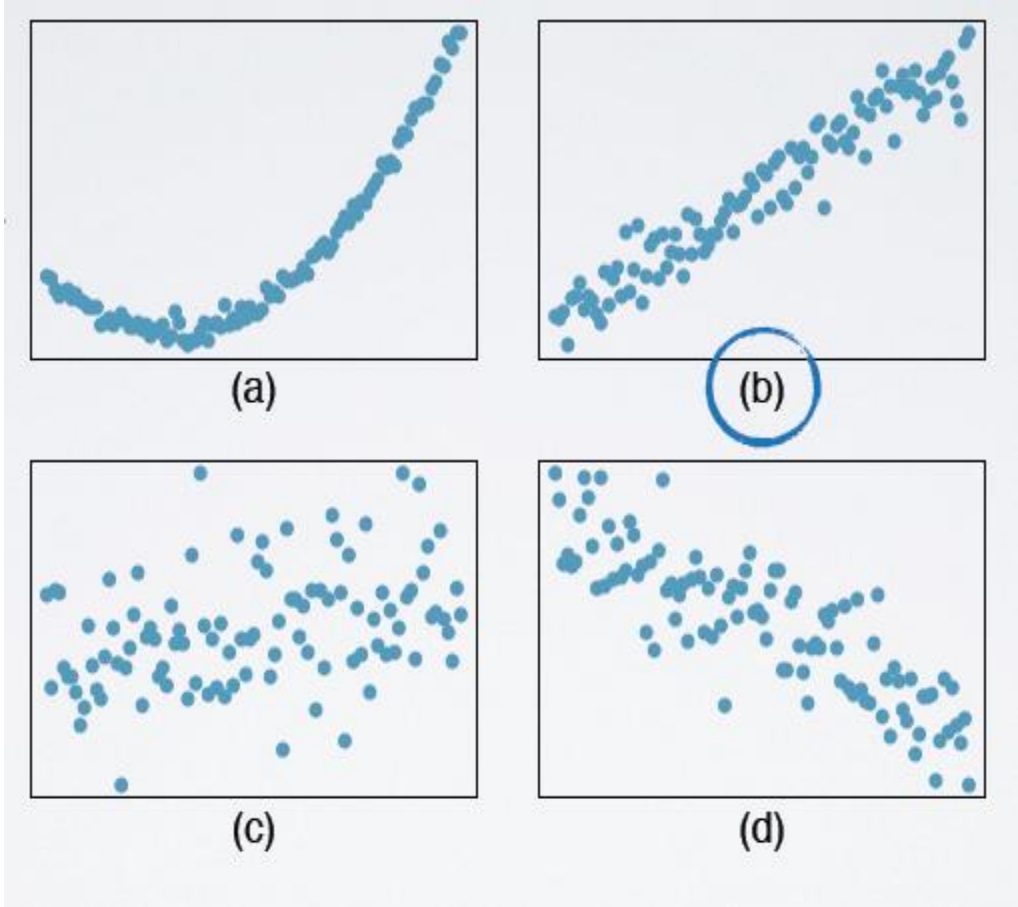
Korelacja

- Jaki jest R w naszym przykładzie?



Korelacja

Bardzo silna nieliniowa



Silna liniowa

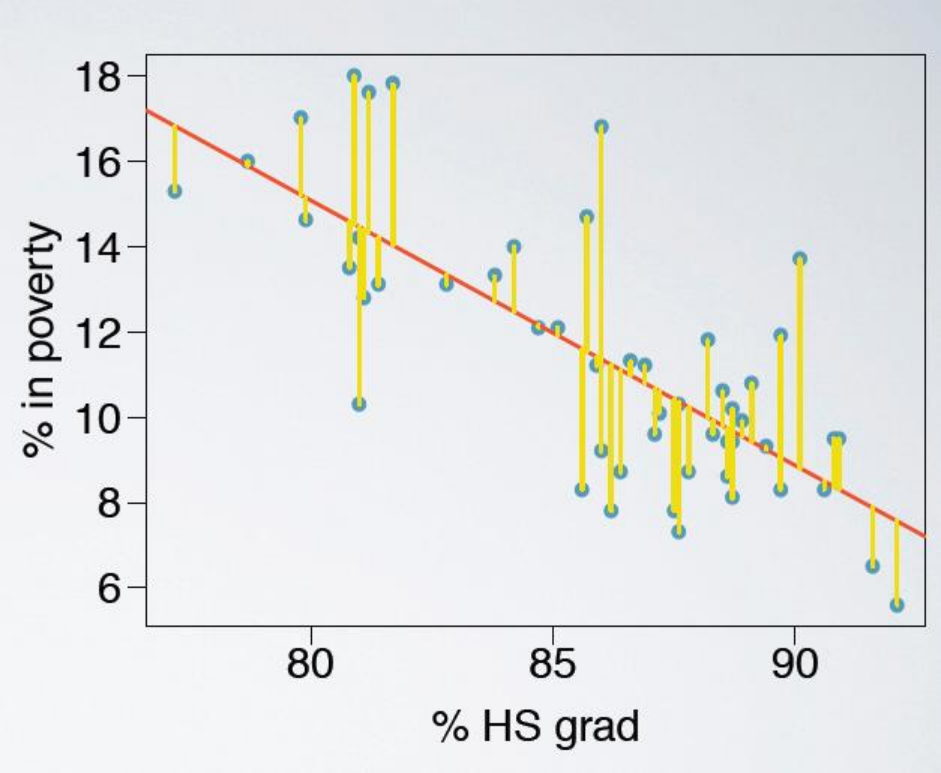
Bardzo słaba

słaba

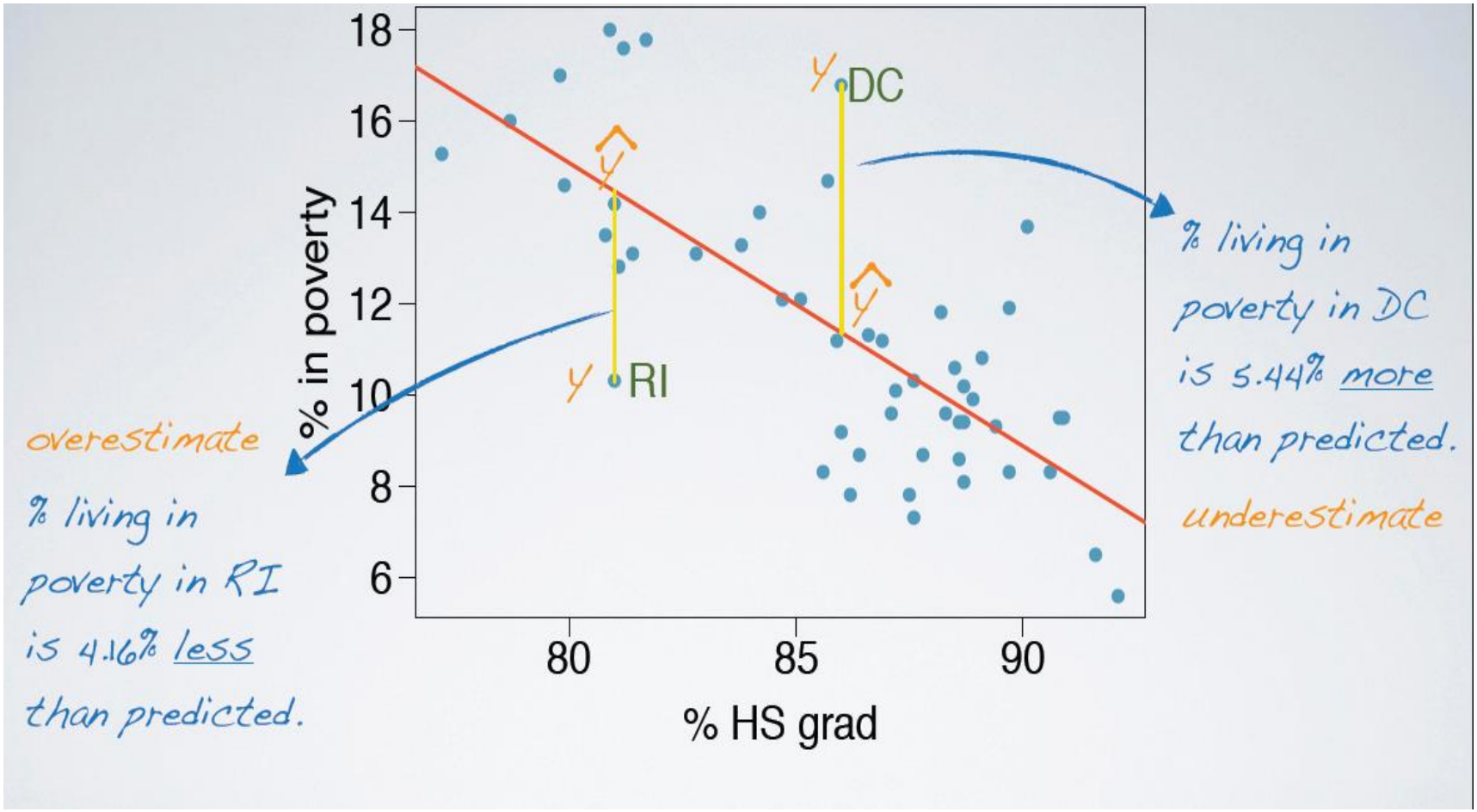
Residuals

- Odchylenia od wartości modelowej
- Dane = fit + residuals
- Różnica pomiędzy przewidywane i obserwowana wartością y

$$e_i = y_i - \hat{y}_i$$



Residuals


18
16
14
12
10
8
6
% in poverty
80
85
90
% HS grad
DC
RI
overestimate
% living in poverty in RI is 4.16% less than predicted.
% living in poverty in DC is 5.44% more than predicted.
underestimate

Residuals

- W jaki sposób wyznaczmy model:
  - Minimalizacja sumy wartości odchyleń?

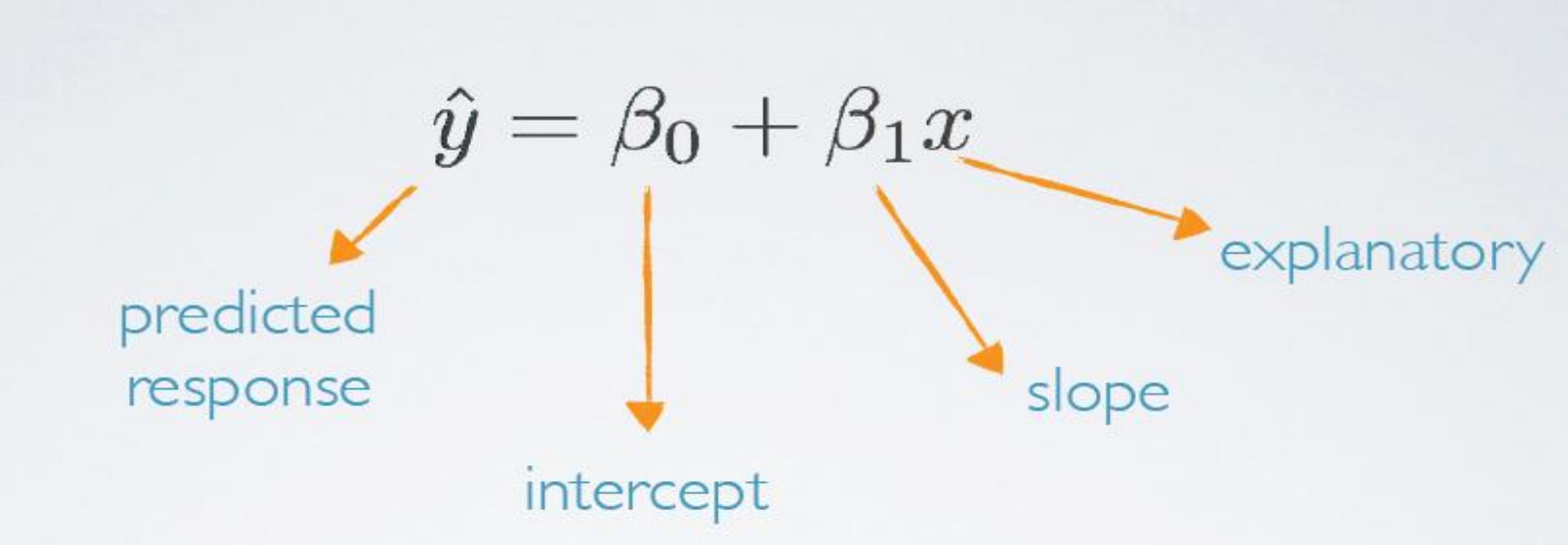
$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Minimalizacja sumy kwadratów odchyleń?

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Metoda najmniejszych kwadratów

Model liniowy regresji

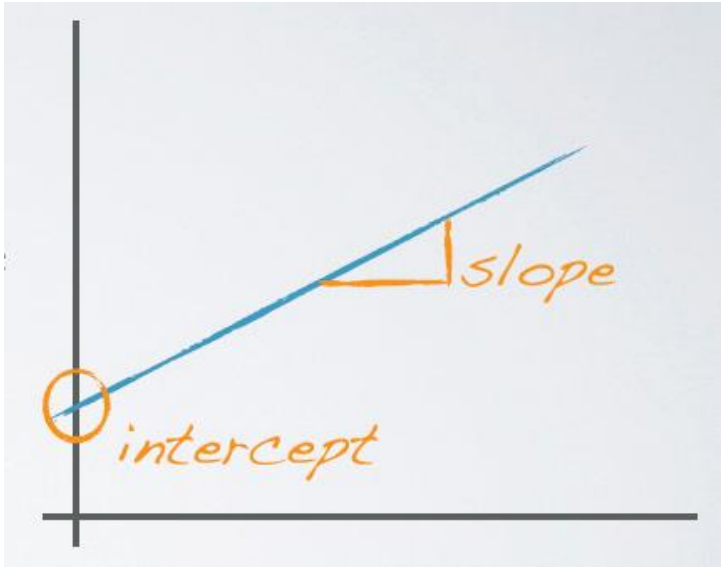


- Notacja:

	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

Parametry modelu

- Punkt przecięcia (intercept): przewidywana wartość y dla x=0
- Współczynnik nachylenia (slope): dla zwiększenia x,y o jednostkę przewidywany wzrost/zmniejszenie x,y



Parametry modelu

- Slope:

slope: $b_1 = \frac{s_y}{s_x} R$

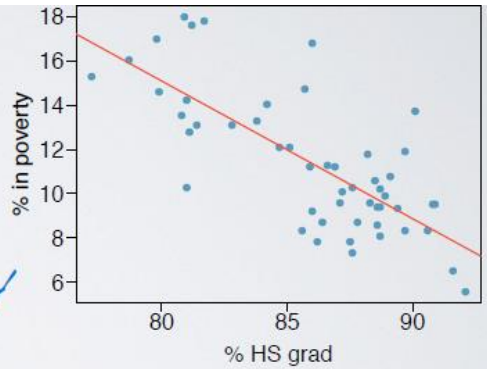
s_x : SD of x
s_y : SD of y
$R = cor(x, y)$

Dane:

$s_y = 3.1\%$
$s_x = 3.73\%$
$R = -0.75$

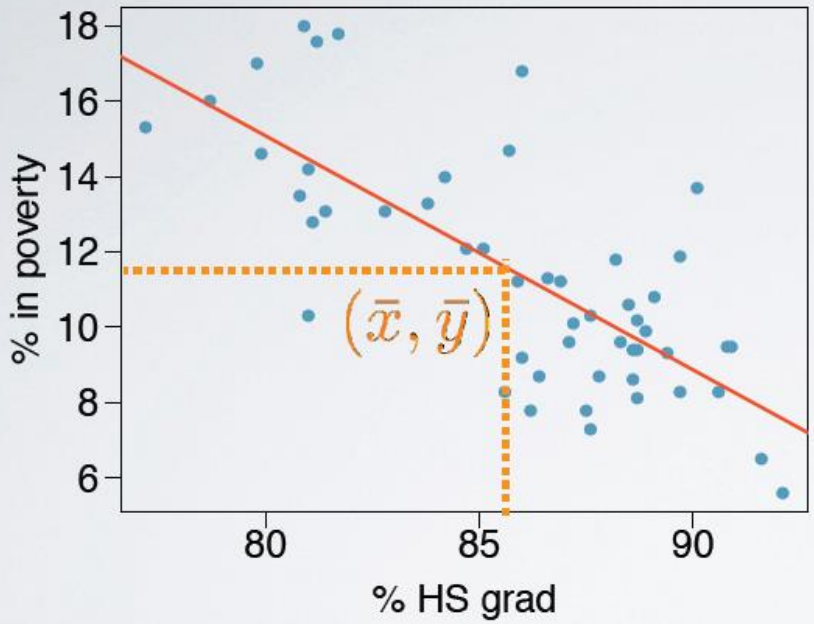
$b_1 = \frac{s_y}{s_x} R = \frac{3.1}{3.73} \times -0.75 \approx -0.62$

For each % point increase in H/S graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.



Parametry modelu

- Punkt przecięcia:



the least squares line always goes through $(\bar{x}, \bar{y})$

$$\bar{y} \quad \hat{y} = b_0 + b_1 x \quad \bar{x}$$

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

Parametry modelu

- Punkt przecięcia:
  - Często nie ma praktycznego znaczenie

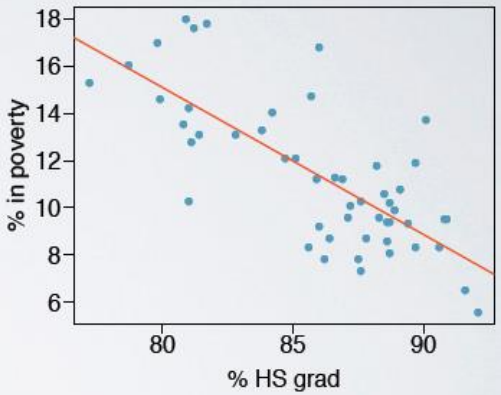
Dane:

$\bar{y} = 11.35\%$

$\bar{x} = 86.01\%$

$b_0 = \bar{y} - b_1\bar{x} = 11.35 - (-0.62)\ 86.01 = 64.68$

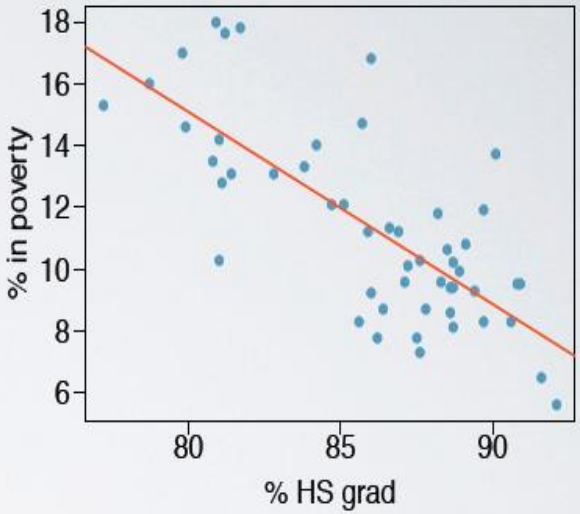
States with no H/S graduates are expected on average to have 64.68% of their residents living below the poverty line.



Parametry modelu liniowego

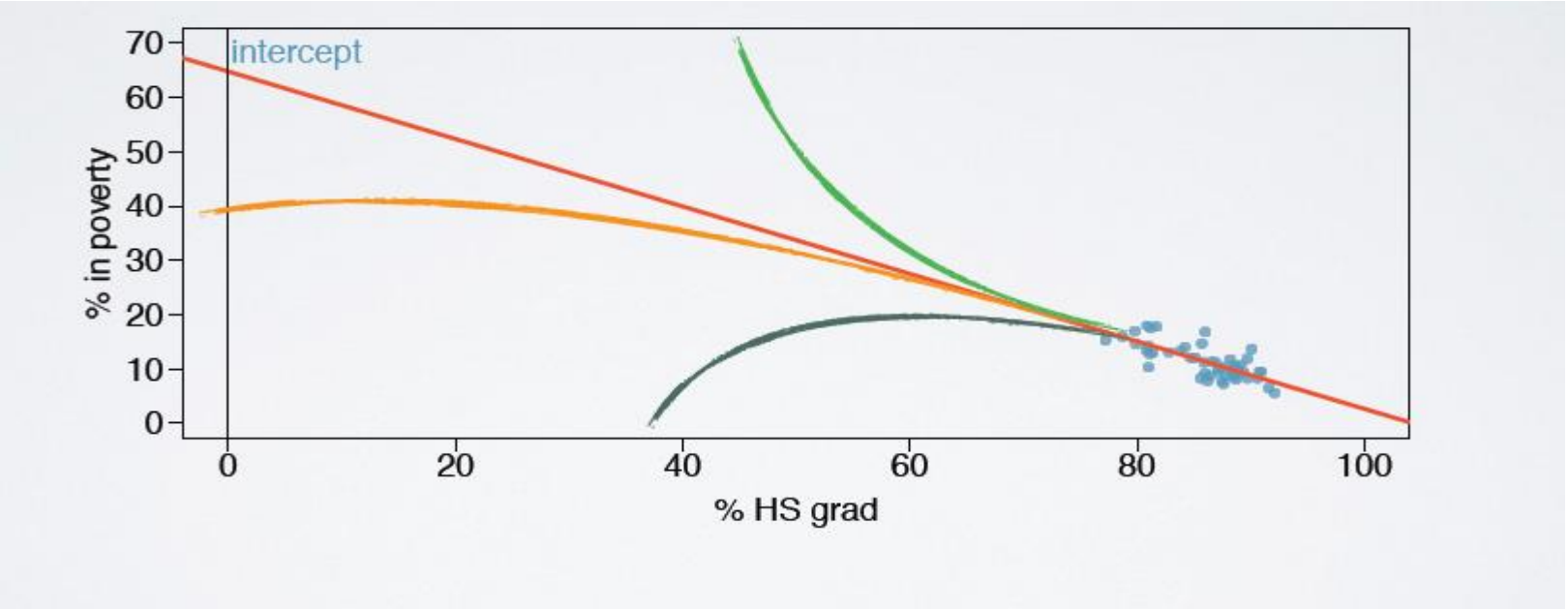
$\widehat{\%\ \text{in poverty}} = 64.68 - 0.62\ \%\ \text{HS grad}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.78	6.80	9.52	0.00
hsgrad	-0.62	0.08	-7.86	0.00



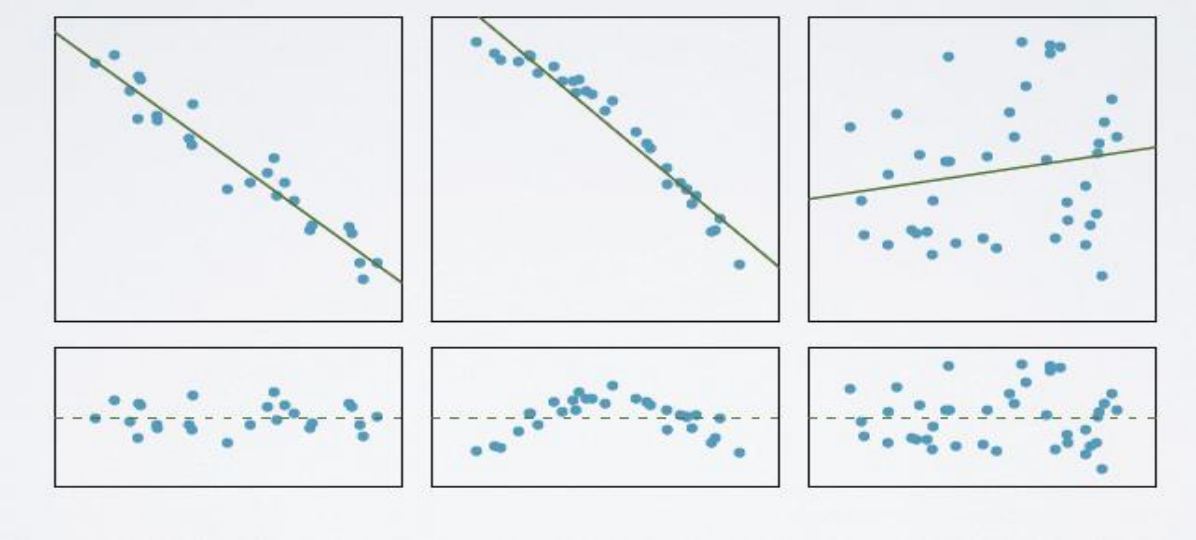
Extrapolacja

- Używamy model aby przewidzieć wartość x poza obszarem dla którego mamy punkty obserwacyjne

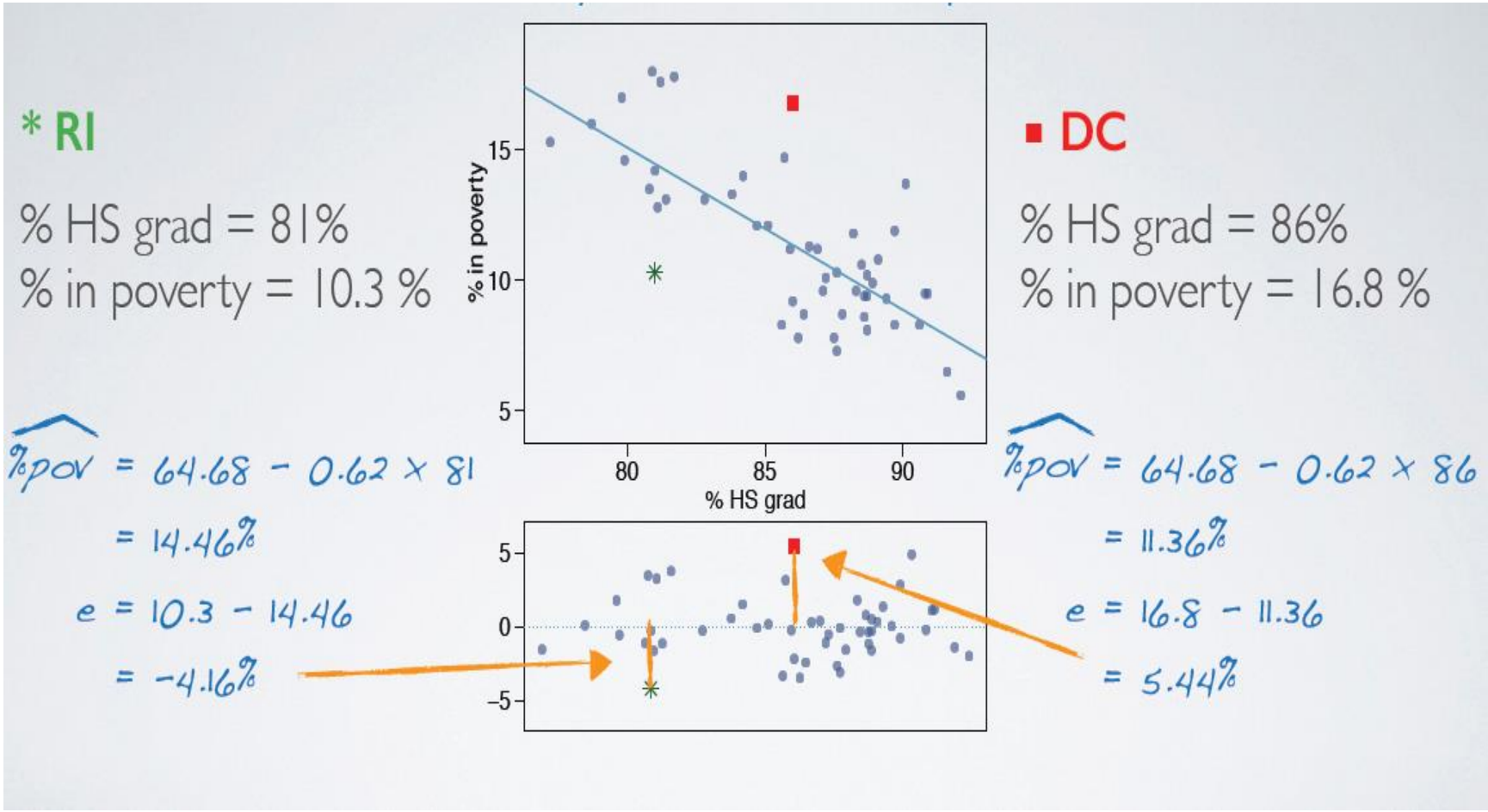


Warunki dla liniowej regresji

- Liniowość relacji pomiędzy x i y
  - Istnieją też metody aby fitować nieliniowe modele
  - Sprawdzamy używając scatter-plotu i residual plotu

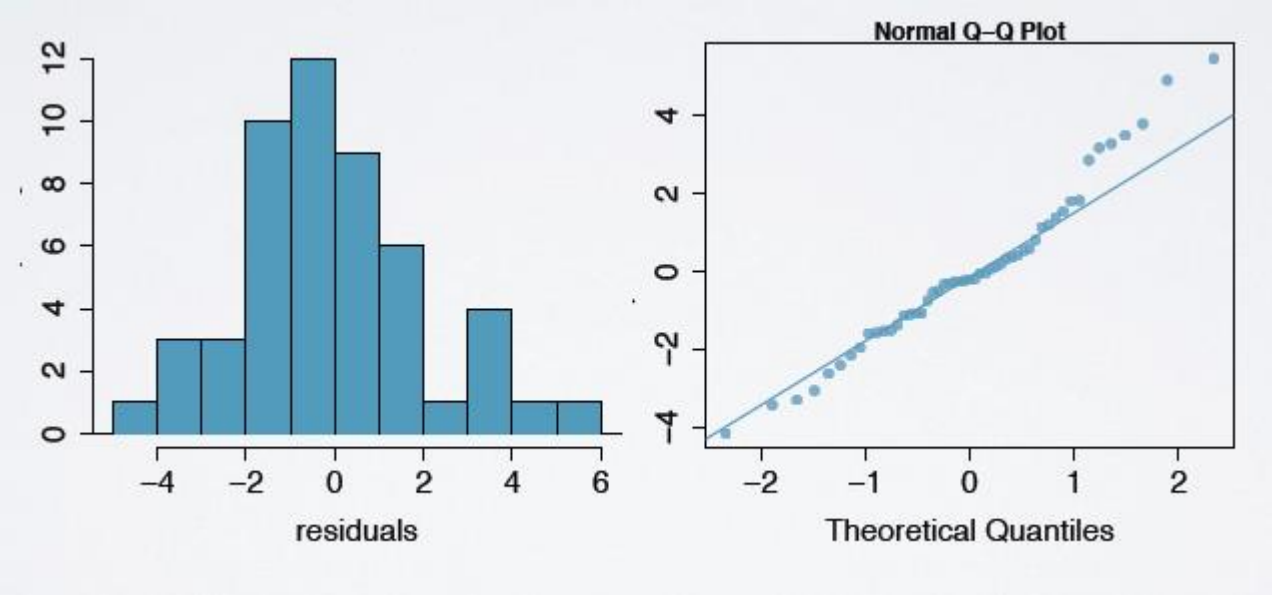


Przykład:


* RI
% HS grad = 81%
% in poverty = 10.3 %
■ DC
% HS grad = 86%
% in poverty = 16.8 %
% in poverty
15
10
5
80
85
90
% HS grad
5
0
−5
%pov = 64.68 − 0.62 × 81
= 14.46%
e = 10.3 − 14.46
= −4.16%
%pov = 64.68 − 0.62 × 86
= 11.36%
e = 16.8 − 11.36
= 5.44%

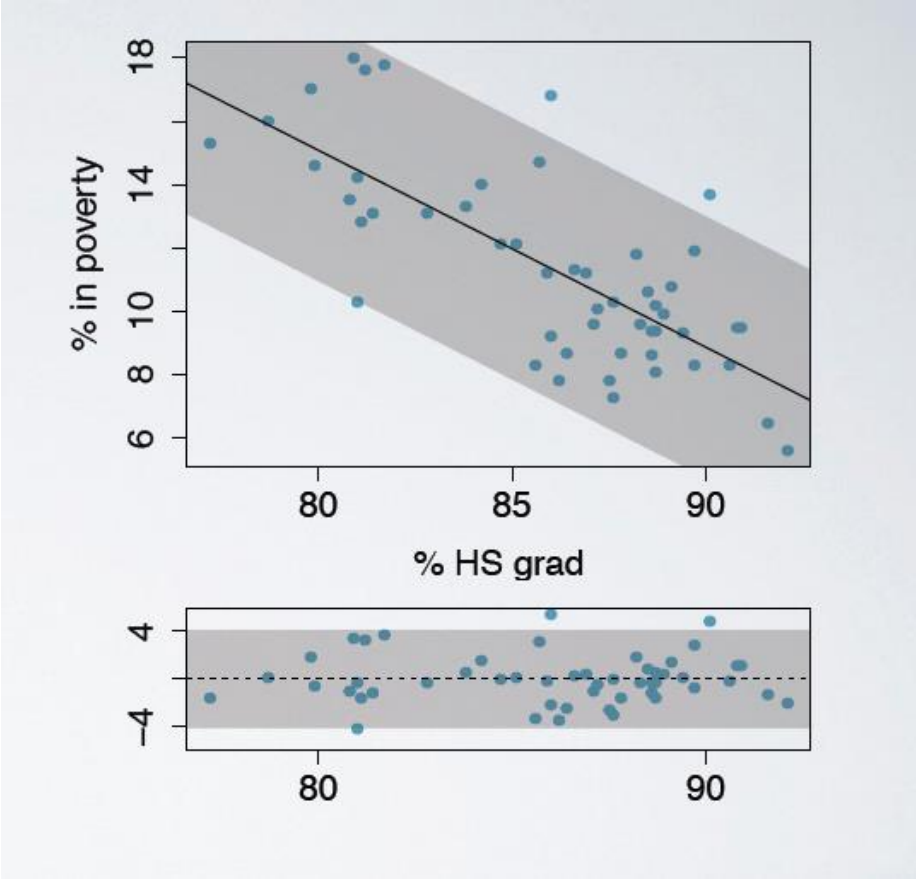
Rozkład punktów residuals

- Powinien być zbliżony do normalnego
- Może nie być spełniony jeżeli są obserwacje które maja inny trend niż reszta danych
- Możemy sprawdzić rysując histogram rozkładu percentili.

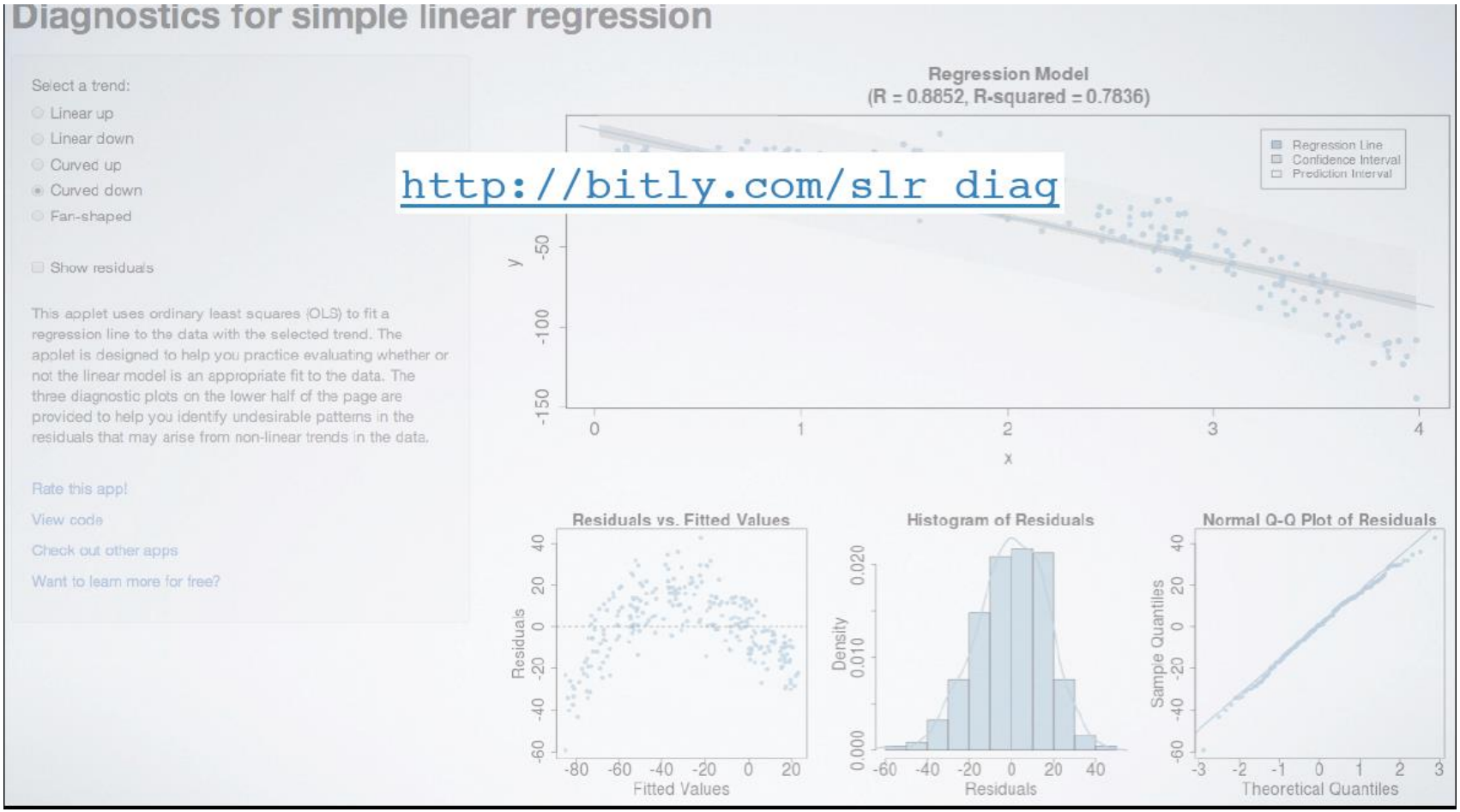


Stała zmienność

- Rozrzut punktów wokół lini modelu powinien być w miarę równomierny
- Sprawdzamy na plocie residuals



Możemy też użyć apletu

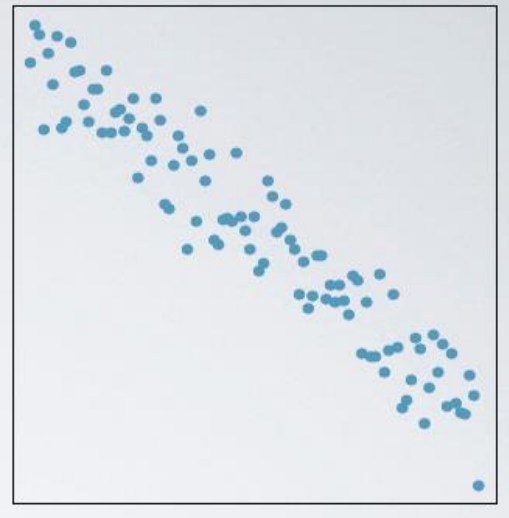


Miara zgodności modelu

- Jako miarę zgodności modelu często używamy R^2
- Ilościowa ocena jaki procent rozrzutu danych może być wyjaśnione przez model

$R^2 = 92.16\%$

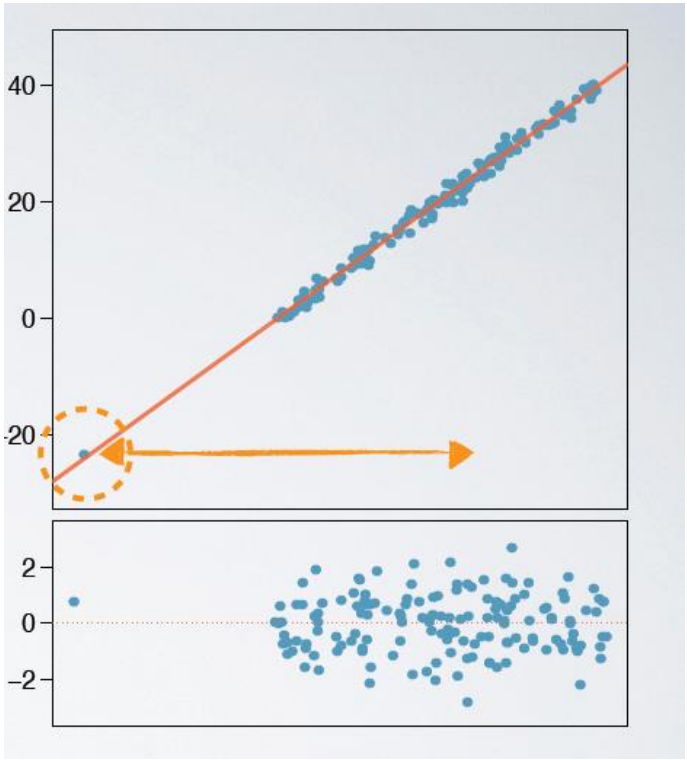
$\sqrt{0.9216} = 0.96 \rightarrow R = -0.96$



Outliers

- Outliers: punkty który leżą daleko od skupisk punktów
- Rozróżniemy dwie kategorie:
  - „leverage”: leżą wzdłuż osi poziomej na plocie residuals
  - „influential”: uwzględnienie silnie zmienia krzywą nachylenia lini

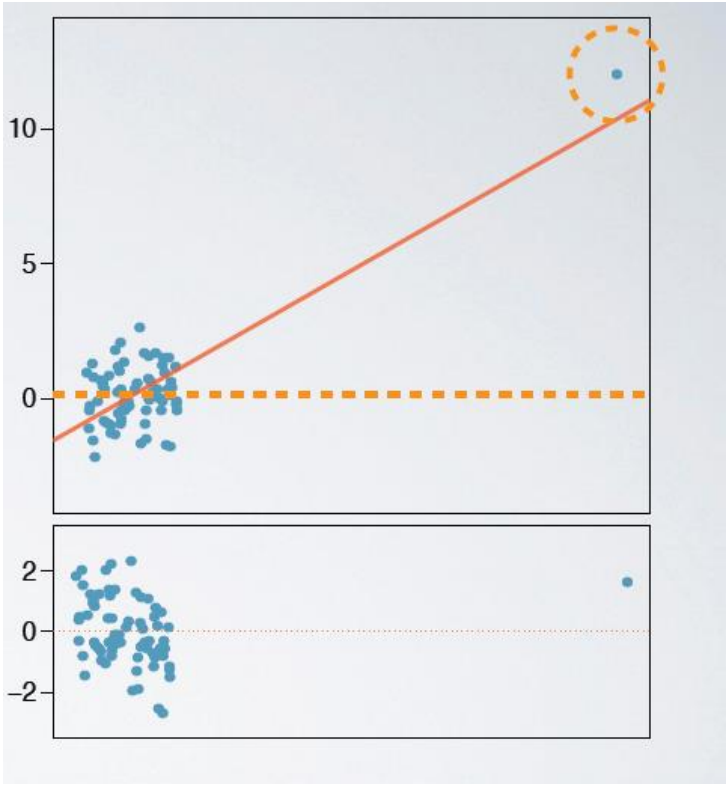
Leverage outlier



Outliers

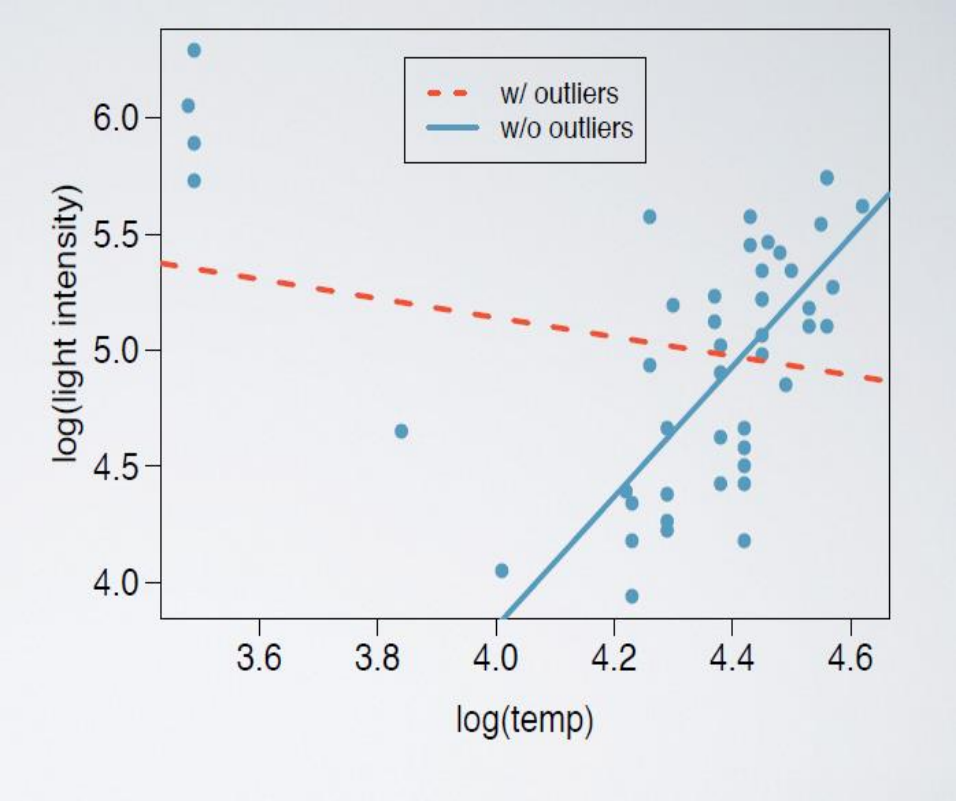
- Outliers: punkty który leżą daleko od skupisk punktów
- Rozróżniemy dwie kategorie:
  - „leverage points”: leżą wzdłuż osi poziomej na plocie residuals
  - „influential points”: uwzględnienie silnie zmienia krzywą nachylenia lini
- W tym przykładzie bez punktu „outliers” nie byłoby relacji pomiedzy zmienną x i y.

Influential outlier



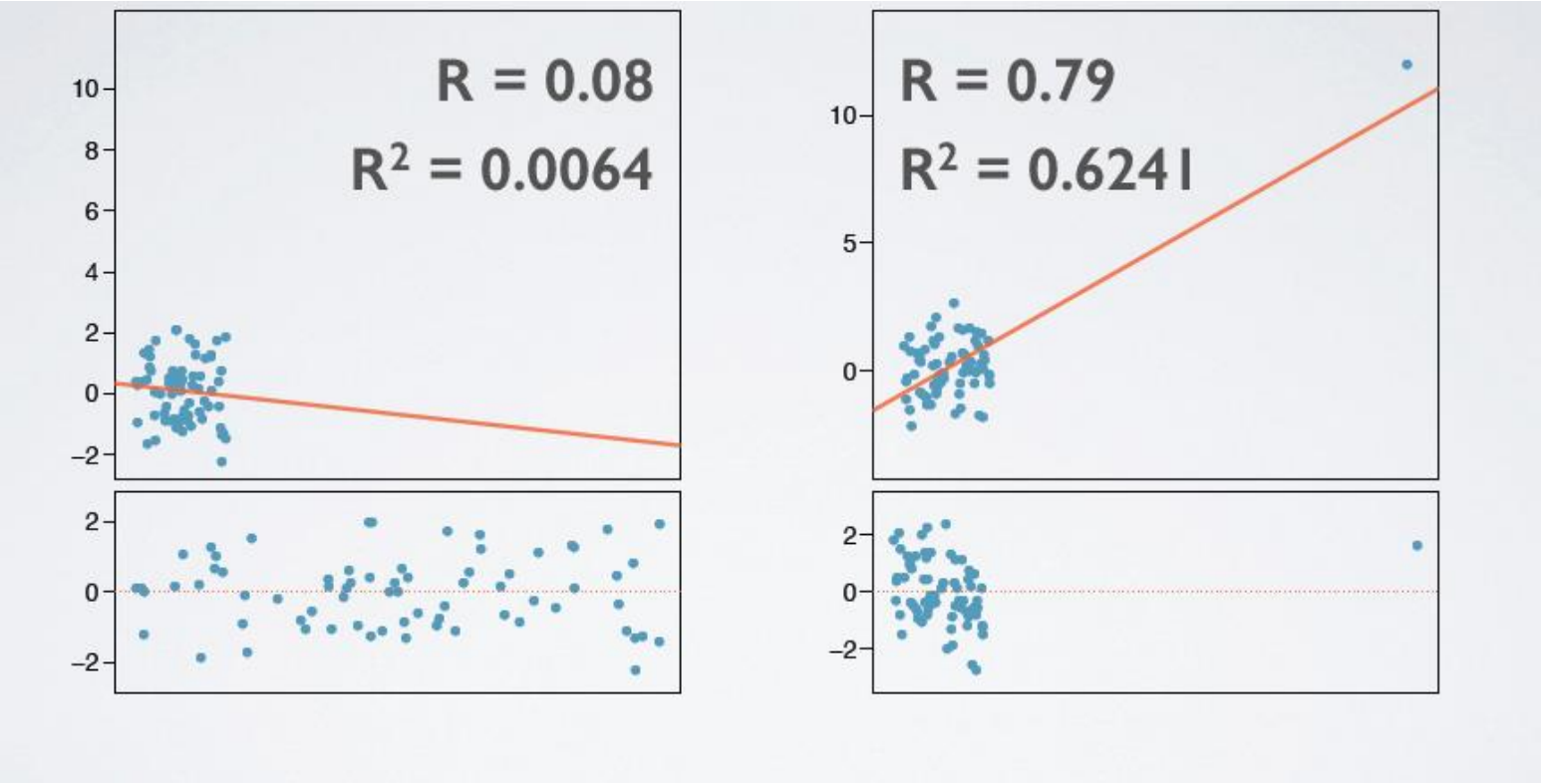
Influential outliers

- Relacja pomiędzy intensywnością promieniowania i temperaturą powierzchni dla 47 gwiazd w clustrze CYG OBI



Influential outliers

- Nie muszą zmniejszać R^2



Wnioskowanie na podstawie regresji liniowej

- Wyniki:

regression output:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

linear model: $\widehat{fosterIQ} = 9.2076 + 0.9014\ bioIQ$

R^2: $R^2 = 0.78$

- Wnioskowanie statystyczne: czy istnieje jakiś związek? Testujemy kierunek nachylenia krzywej
- $H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$

Wnioskowanie statystyczne

- Używamy t-statystykę dla wnioskowania

b_1

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

SE_{b_1}

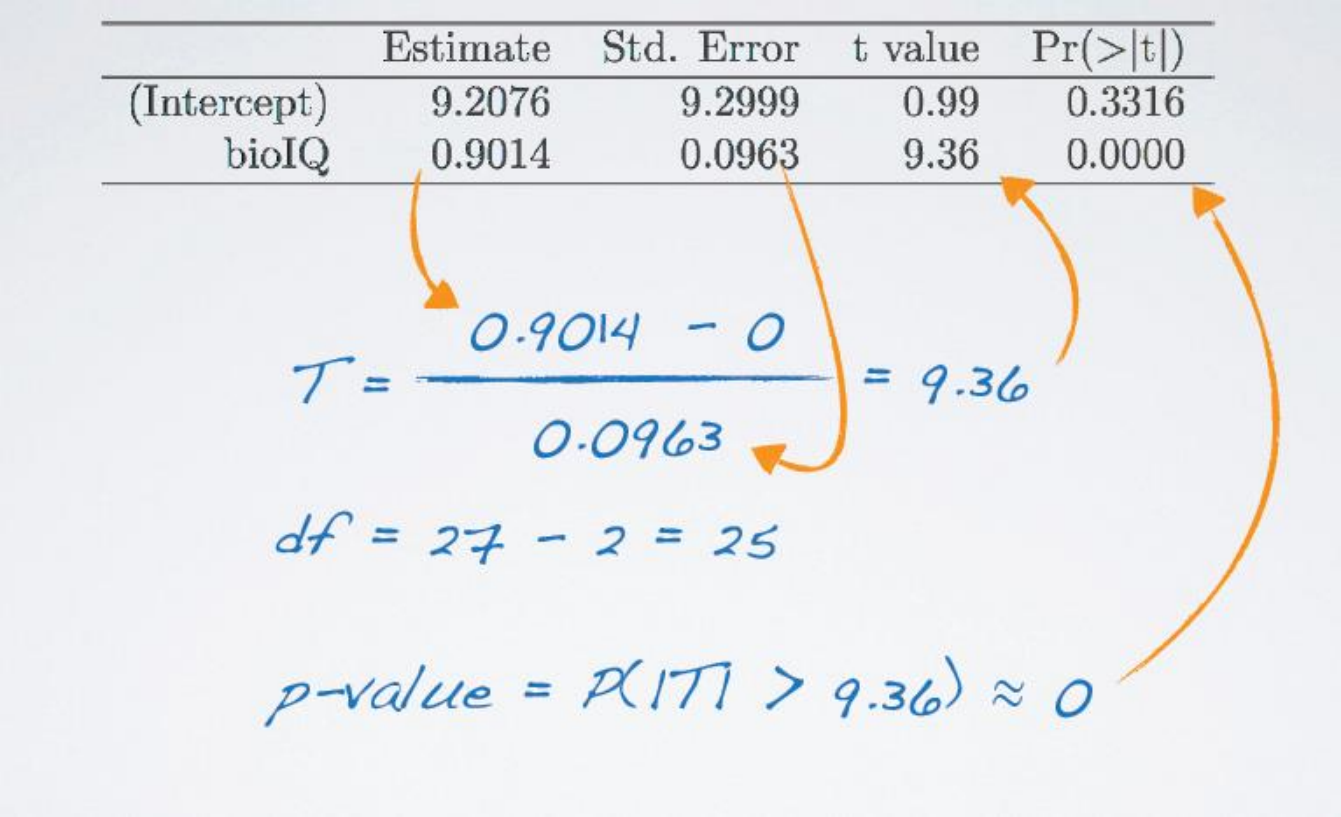
t-statistic for the slope:

$$T = \frac{b_1 - 0}{SE_{b_1}} \qquad df = n - 2$$

Odejmujemy 1 stopień swobody dla każdego parametru: β_o i β_1

Wnioskowanie statystyczne

- Używamy t-statystykę dla wnioskowania



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

$df = 27 - 2 = 25$

$p\text{-value} = P(|T| > 9.36) \approx 0$

Wnioskowanie statystyczne

- Używamy t-statystykę dla przedziału ufności

point estimate ± margin of error

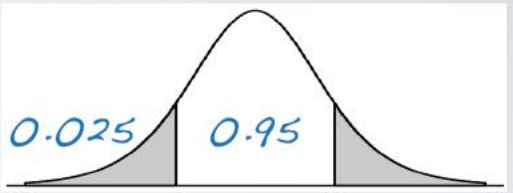
$$b_1 \pm t^\star_{df} SE_{b_1}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$df = 27 - 2 = 25$

$t^*_{25} = 2.06$

$0.9014 \pm 2.06 \times 0.0963 = (0.7, 1.1)$



R

```
> qt(0.025, df = 25)
[1] -2.059539
```

Wnioskowanie statystyczne

hypothesis test:

$$T = \frac{b_1 - \text{null value}}{SE_{b_1}} \qquad df = n - 2$$

confidence interval:

$$b_1 \pm t^{\star}_{df} SE_{b_1}$$

- Hipoteza H_0 jest często $\beta_1 = 0$
- Wynik wnioskowania: b_1, SE_{b1}, oraz dwu-stronne przedział ufności for t-test.
- Wnioskowanie dla punktu przecięcia jest rzadko stosowane.

Analiza wariancji

- T-test pozwalał ocenić jak silna jest hipoteza o liniowej relacji pomiędzy x i y
- Alternatywy: sprawdzenie jak bardzo zmienność y jest wyjaśniona przez model: analiza wariancji.

Analiza wariancji

anova output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

sum of squares

total variability in y: $SS_{Tot} = \sum (y - \bar{y})^2 = 6724.66$

unexplained variability in y (residuals): $SS_{Res} = \sum (y - \hat{y})^2 = \sum e_i^2 = 1493.53$

explained variability in y: $SS_{Reg} = 6724.66 - 1493.53 = 5231.13$

Analiza wariancji

anova output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

degrees of freedom

total degrees of freedom: $df_{Tot} = 27 - 1 = 26$

regression degrees of freedom: $df_{Reg} = 1$ only 1 predictor

residual degrees of freedom: $df_{Res} = 26 - 1 = 25$

Analiza wariancji

anova output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

mean squares

MS regression: $MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$

MS residual: $MS_{Res} = \frac{SS_{Res}}{df_{Res}} = \frac{1493.53}{25} = 59.74$

F statistic

ratio of explained to unexplained variability

$$F_{(1,25)} = \frac{MS_{Reg}}{MS_{Res}} = 87.56$$

Analiza wariancji

anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

small p-value → reject H_0

$$H_0 : \beta_1 = 0$$

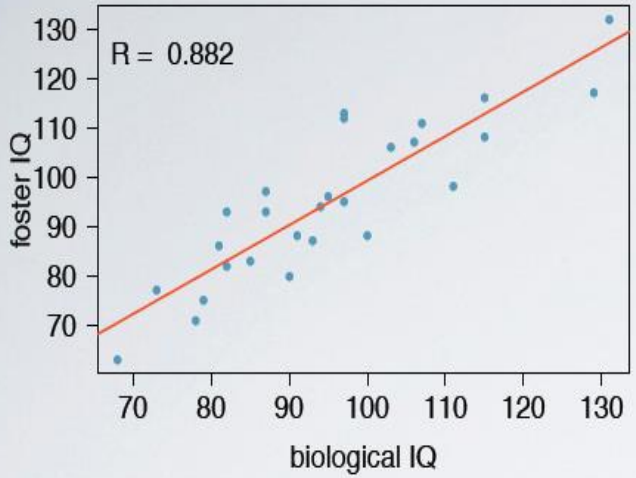
$$H_A : \beta_1 \neq 0$$

The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

Jeszcze raz na temat R^2

- R^2 – proporcja zmienności wyjaśniona przez model
  - Duże: istnieje liniowy związek pomiędzy x i y
  - Małe: nie można wyciągnąć takiego wniosku
- Dwa sposoby aby to policzyć
  - Na podstawie korelacji: policzyć współczynnik korelacji
  - Z definicji: policzyć procent wyjaśnionej zmienności

Analiza wariancji



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bioIQ	1	5231.13	5231.13	87.56	0.0000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

(1) $R^2 = \text{square of correlation coefficient} = 0.882^2 \approx 0.78$

(2) $R^2 = \dfrac{\text{explained variability}}{\text{total variability}} = \dfrac{SS_{Reg}}{SS_{Tot}} = \dfrac{5231.13}{6724.66} \approx 0.78$