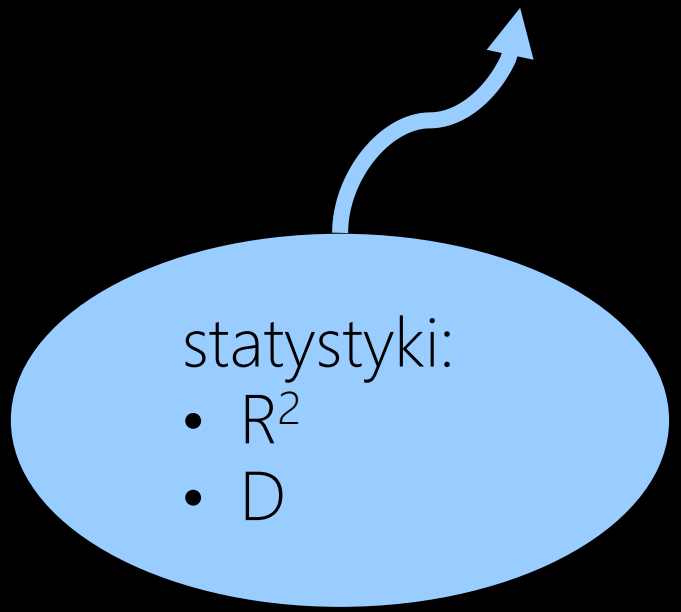
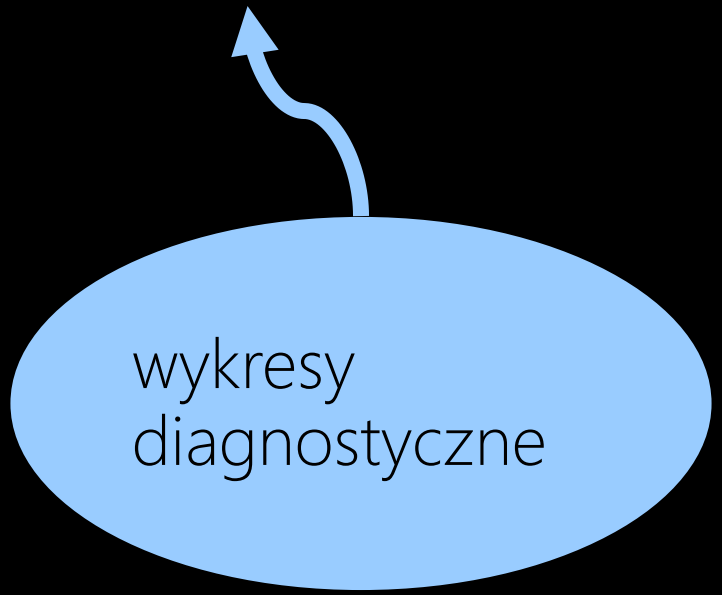
METODY STATYSTYCZNE W BIOLOGII

1. Wykład wstępny
2. Populacje i próby danych
3. Testowanie hipotez i estymacja parametrów
4. Planowanie eksperymentów biologicznych
5. Najczęściej wykorzystywane testy statystyczne I
6. Najczęściej wykorzystywane testy statystyczne II
7. Regresja liniowa
8. Regresja nieliniowa
9. Określenie jakości dopasowania równania regresji liniowej i nieliniowej
10. Korelacja
11. Elementy statystycznego modelowania danych
12. Porównywanie modeli
13. Analiza wariancji
14. Analiza kowariancji
15. Podsumowanie materiału, wspólna analiza przykładów, dyskusja

WSTĘP

testowanie jakości dopasowania równania regresji

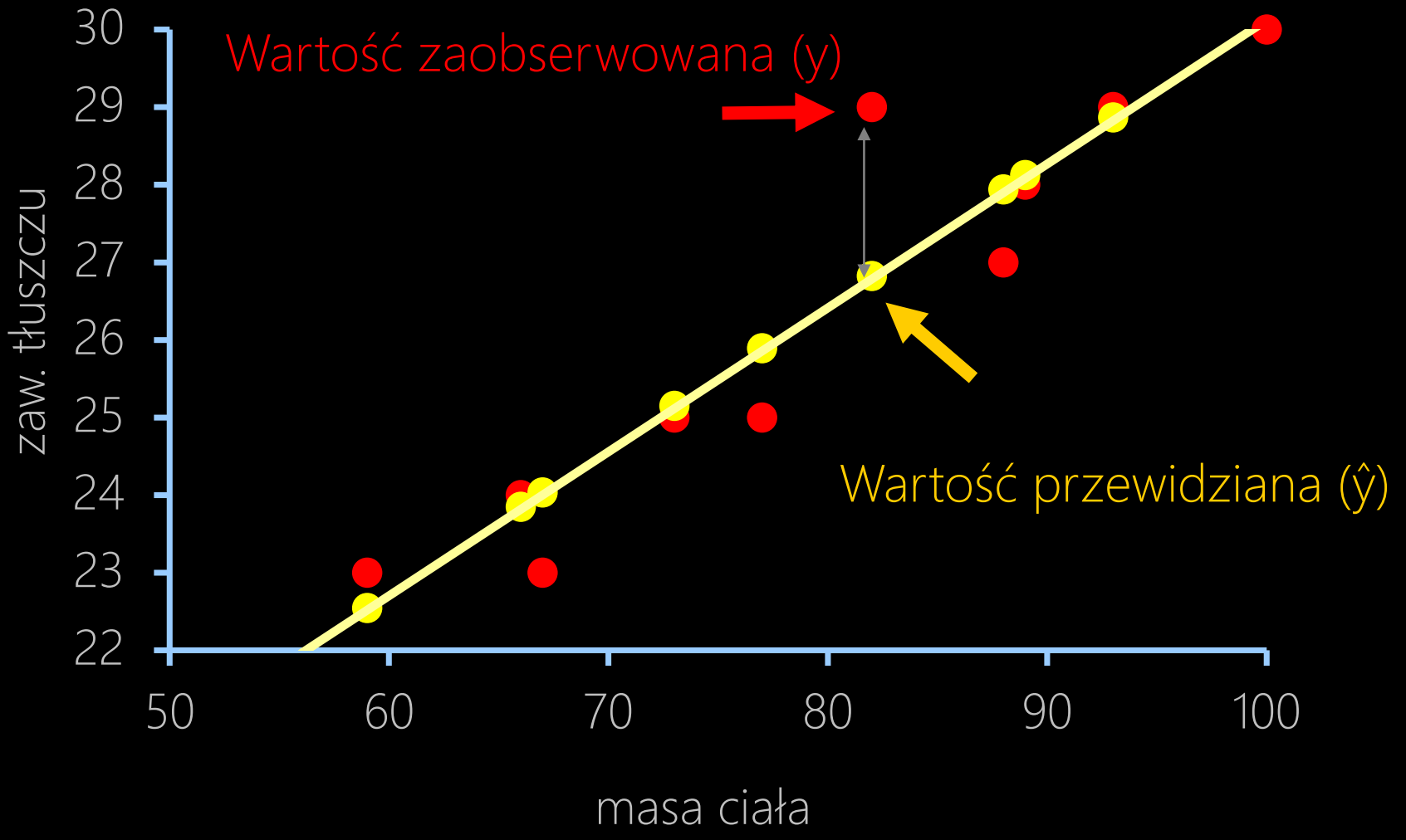




REGRESJA LINIOWA

Równanie regresji

elementy równania regresji: błąd



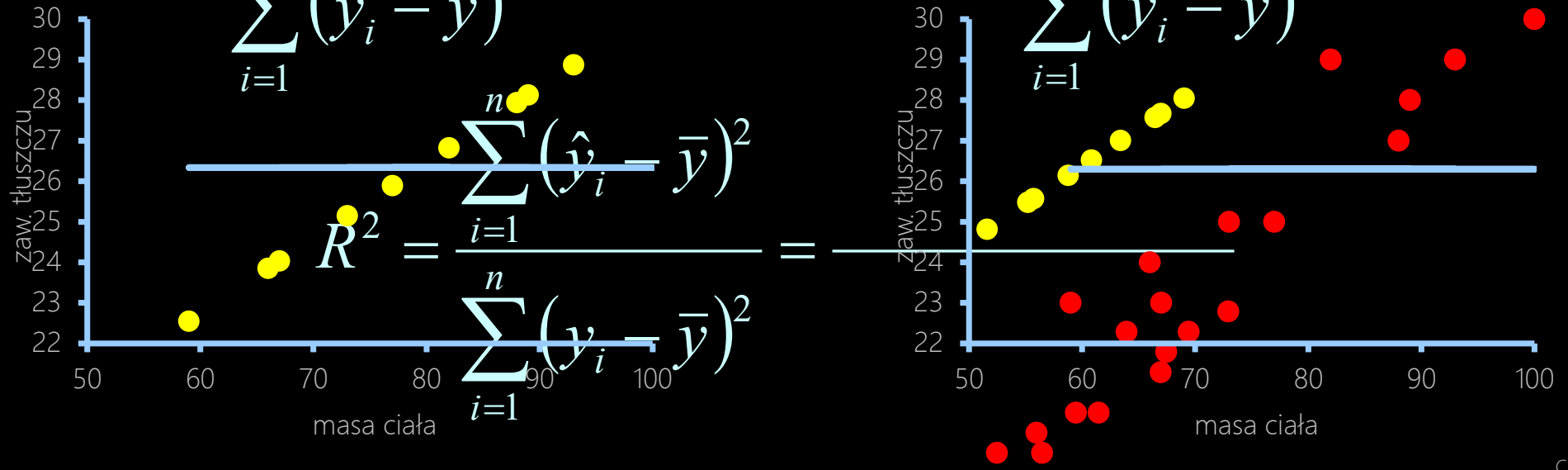
DOPASOWANIE REGRESJI LINIOWEJ - zmienność

zmienność "y"

wyjaśniona przez równanie regresji

zaobserwowana

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



Dopasowanie regresji liniowej - zmienność

jaka część obserwowanej zmienności została wyjaśniona przez równanie regresji

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Dopasowanie regresji liniowej - przykład

PRÓBA DANYCH

MASA CIAŁA	ZAW. TŁUSZCZU
89	28
88	27
66	24
59	23
93	29
73	25
82	29
77	25
100	30
67	23

1. Zmienna niezależna
2. Zmienna zależna, rozkład ciągły

$$tluszcz = 11.57 + 0.19 masa\_ciała$$

$R^2 = 0.94$

94% zmienności zawartości tłuszczu jest wyjaśnione przez zmienność masy ciała

Równanie regresji

Co należy sprawdzić w równaniu regresji?

kryterium **LINE**:

- **L**inearity → zależność E (Y | X) jest liniowa w stosunku do X
- **I**ndependence → y_i są niezależne od siebie
- **N**ormality → Y pochodzi z rozkładu normalnego
- **E**qual variance → wariancja Y jest niezależna od X

Błędy

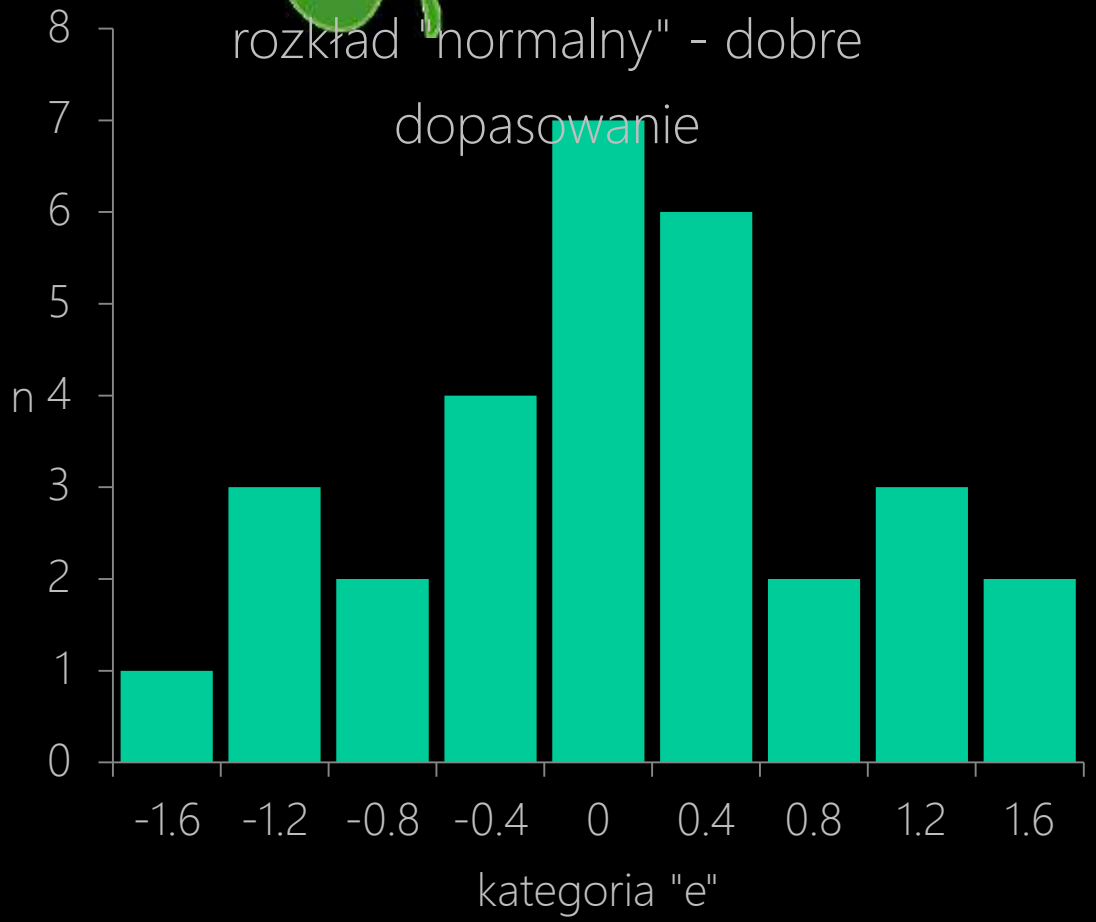
Jak sprawdzić kryterium LINE?

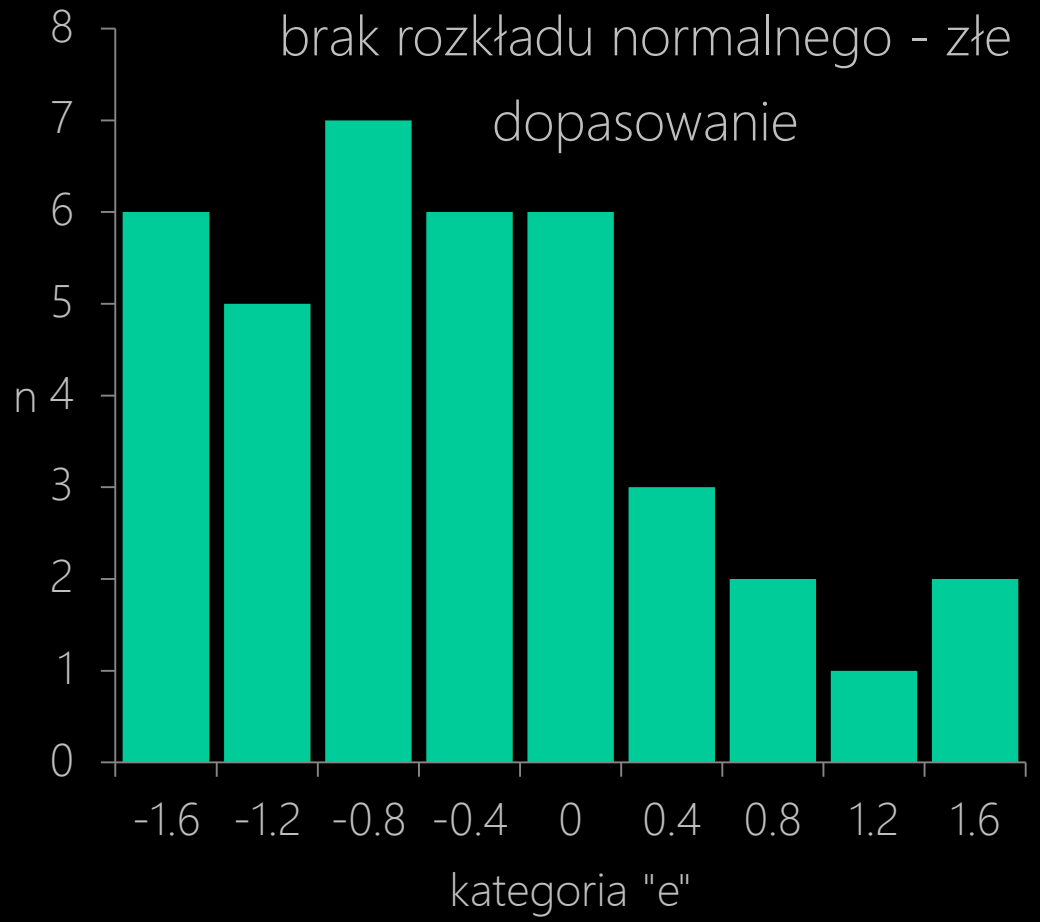
$$(y_i - \hat{y}_i) = e_i$$

1. Wartości błędów
2. Wartości reszt
3. Residuals
4. https://stattrek.com/regression/residual-analysis.aspx?Tutorial=reg

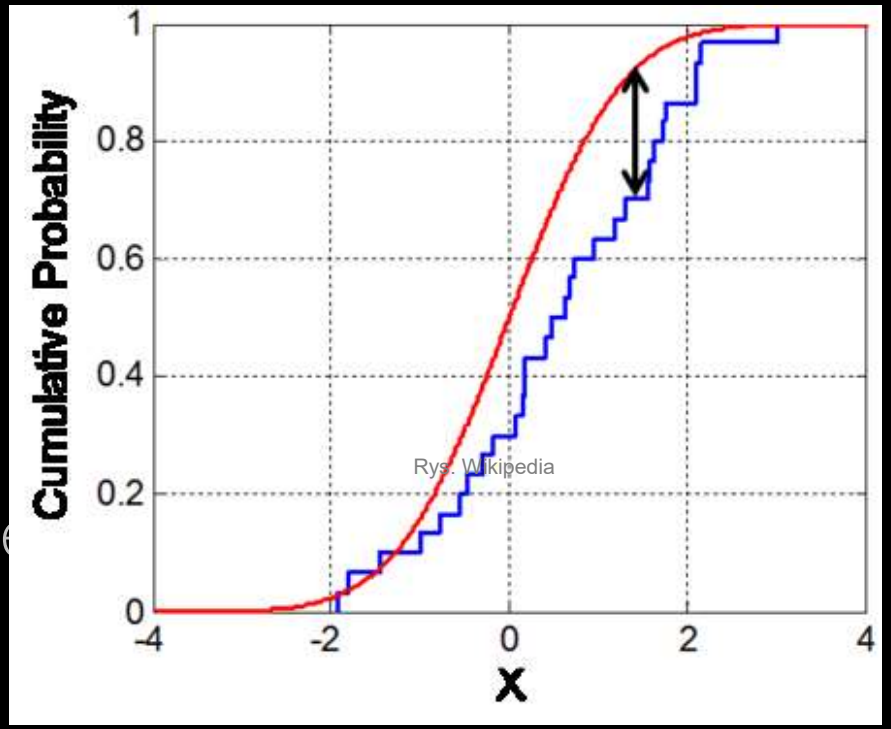
Histogram → sprawdzenie rozkładu wartości błędów

$$(y_i - \hat{y}_i) = e_i \sim N(0, \sigma_e^2)$$





Test zgodności z rozkładem normalnym



- np. test Kolmogorova-Smirnova
- H_0: rozkłady są zgodne, H_1: rozkłady nie są zgodne
- ...
- $D = max_x \left| F_{exp}(x) - F_{obs}(x) \right|$

$F_{exp}(x)$ - dystrybuanta rozkładu teoretycznego np. N(0,1)

$F_{obs}(x)$ - empiryczna dystrybuanta rozkładu porównywanego

- Wartość krytyczna dla $\alpha_{max} = \mathbf{0.05}$ wynosi $\frac{1.36}{\sqrt{n}}$

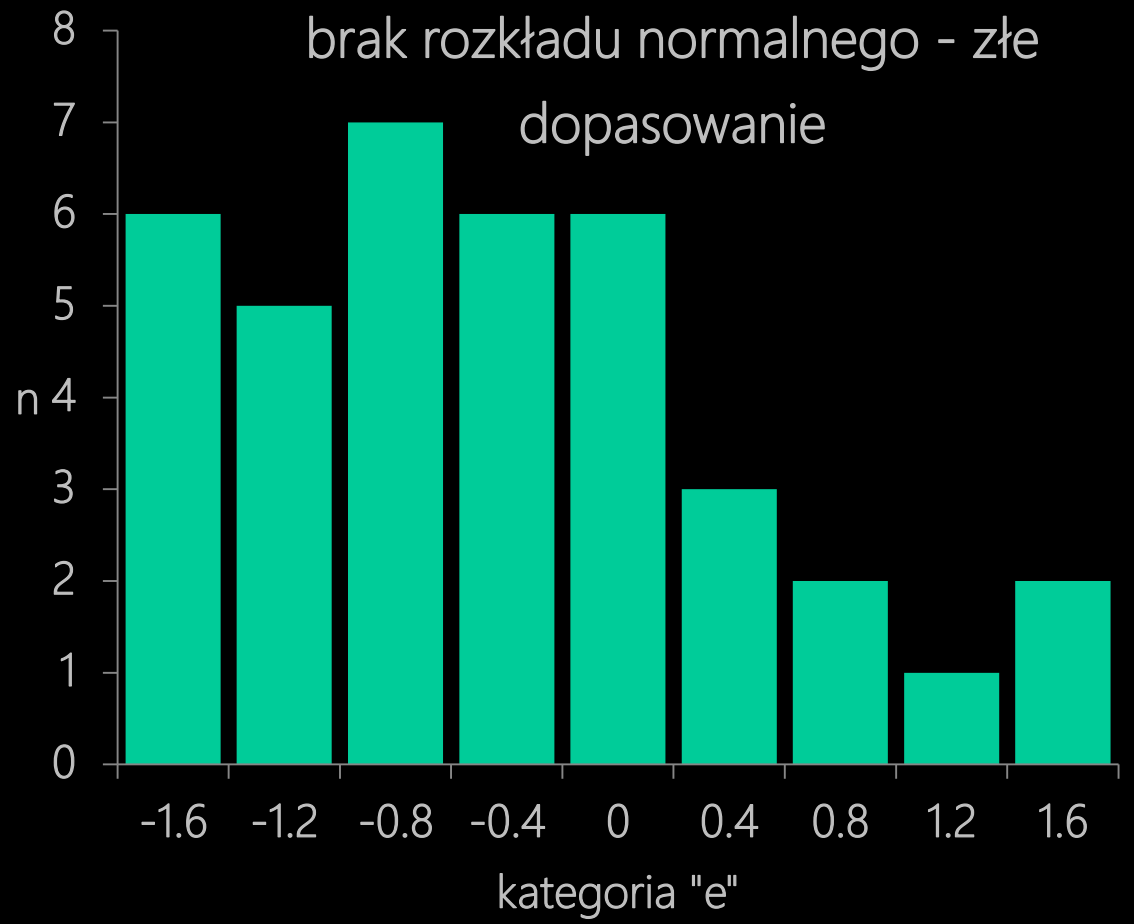
Histogram → sprawdzenie rozkładu wartości błędów

$$(y_i - \hat{y}_i) = e_i \sim N(0, \sigma_e^2)$$

ZŁE DOPASOWANIE

Zastosować transformację „y”:

- $\sqrt{y}$
- $\ln(y)$
- $\frac{1}{y}$

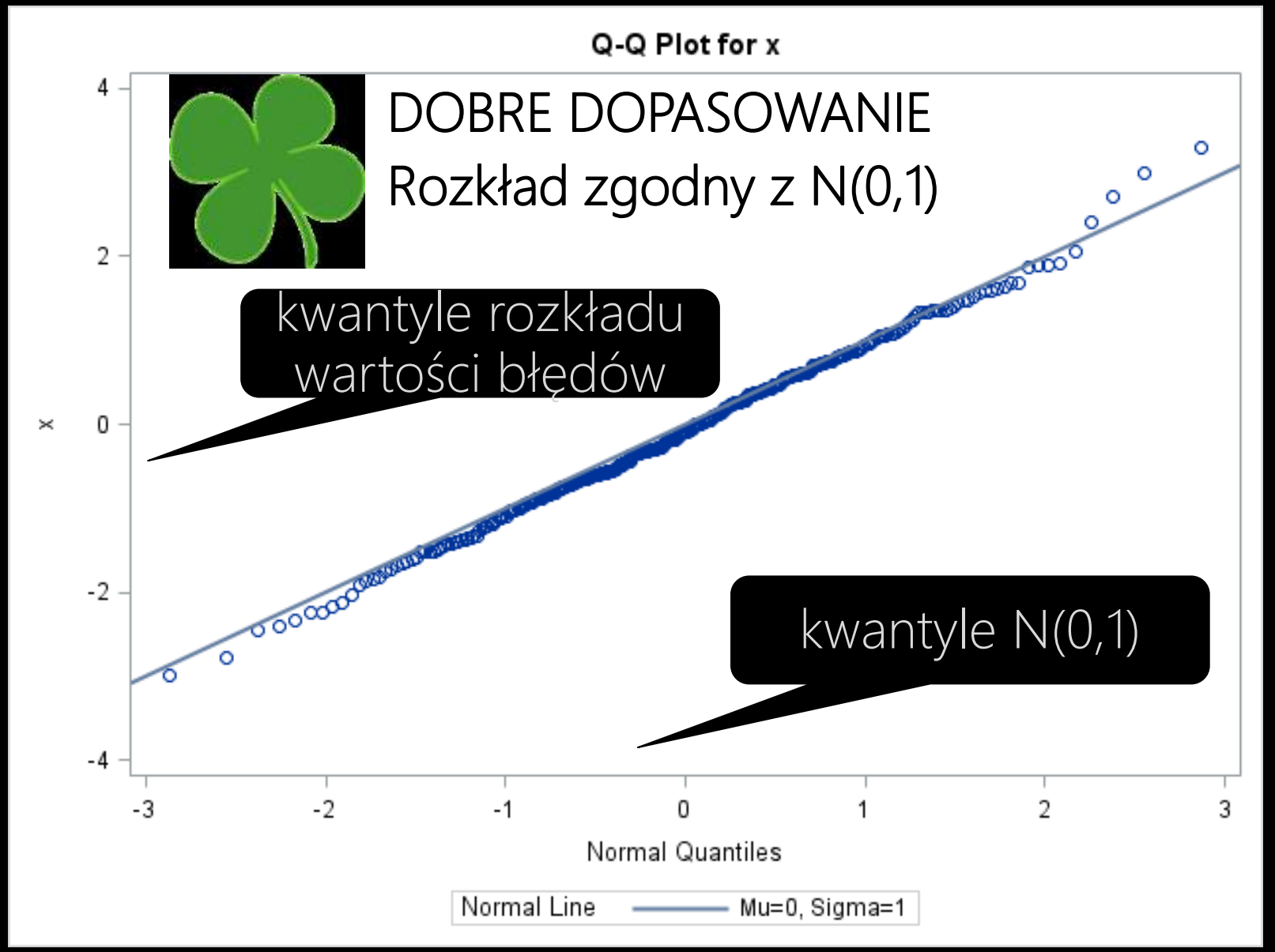


QQ plot → sprawdzenie rozkładu wartości błędów

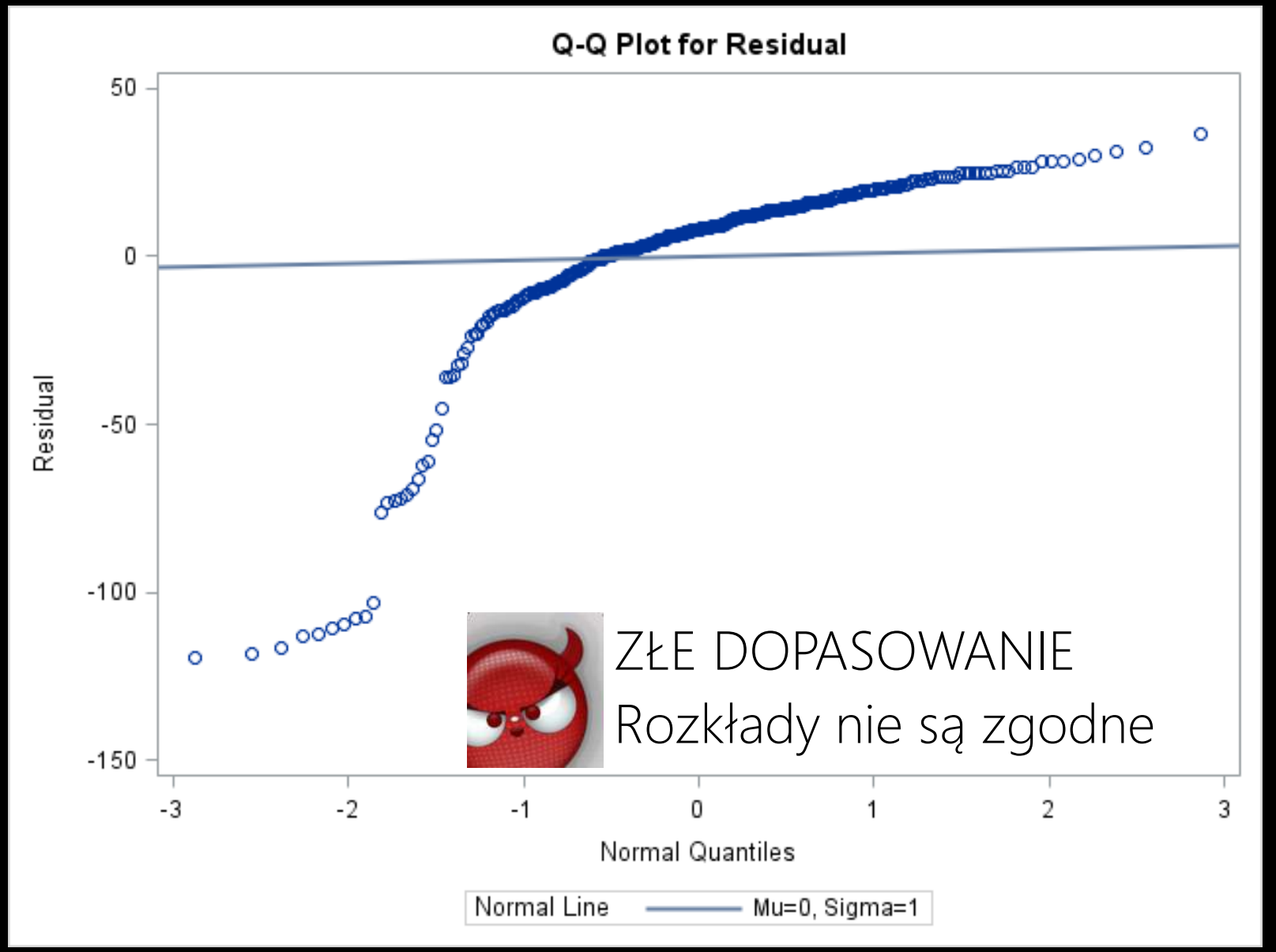
QQ plot → quantile-quantile plot

- porównanie kwantyli dwu rozkładów
- kwantyl → liczba dzieląca wykres gęstości rozkładu na części
- np. dla N(0,1) → 0.500 kwantyl = mediana = średnia = 0.000
  → 50% wartości ≤ 0.000 i 50% wartości > 0.000
- np. dla N(0,1) → 0.975 kwantyl = 1.960
  → 97.5% wartości ≤ 1.960 i 2.5% wartości > 1.960

QQ plot → sprawdzenie rozkładu wartości błędów



QQ plot → sprawdzenie rozkładu wartości błędów

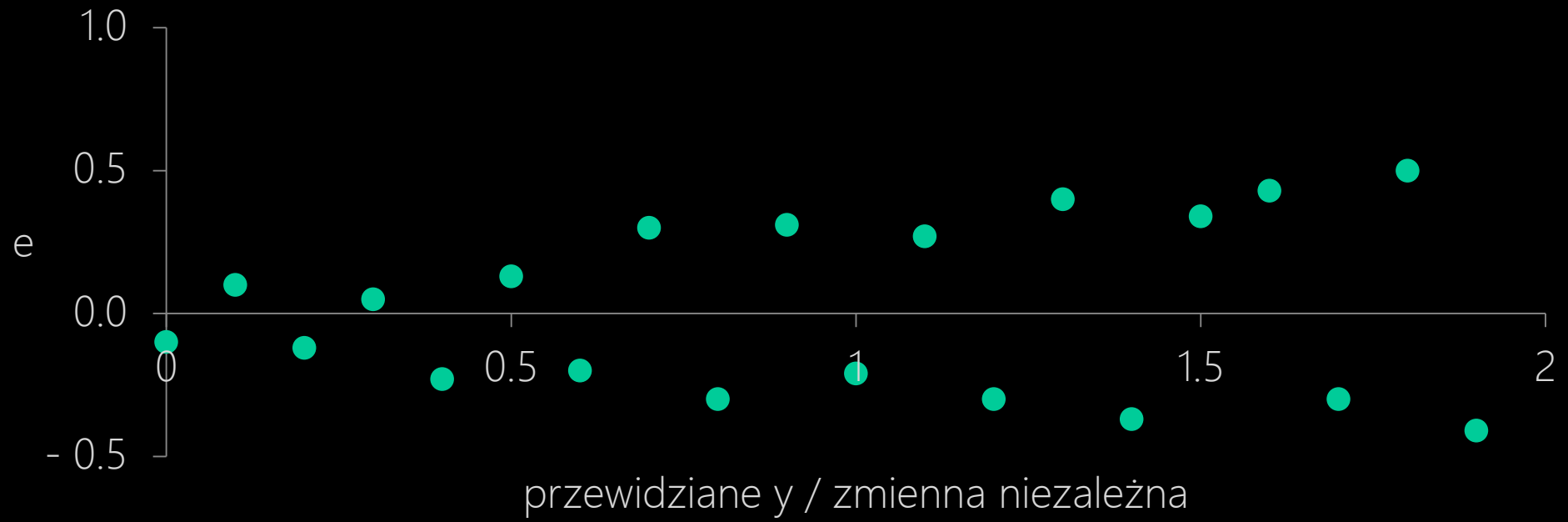


błąd x przewidziany „y” → jakość dopasowania regresji

DOBRE DOPASOWANIE

Brak trendu

błąd x przewidziany „y” → jakość dopasowania regresji



ZŁE DOPASOWANIE

Błąd nie jest losowy, wzrasta wraz ze wzrostem „y”

Zastosować: regresję ważoną, transformację "y"

błąd x przewidziany „y” → jakość dopasowania regresji

ZŁE DOPASOWANIE

Błąd nie jest losowy względem „y”

Zastosować inne / dodatkowe zmienne niezależne

błąd x przewidziany „y” → jakość dopasowania regresji

ZŁE DOPASOWANIE

Błąd nie jest losowy względem „y”

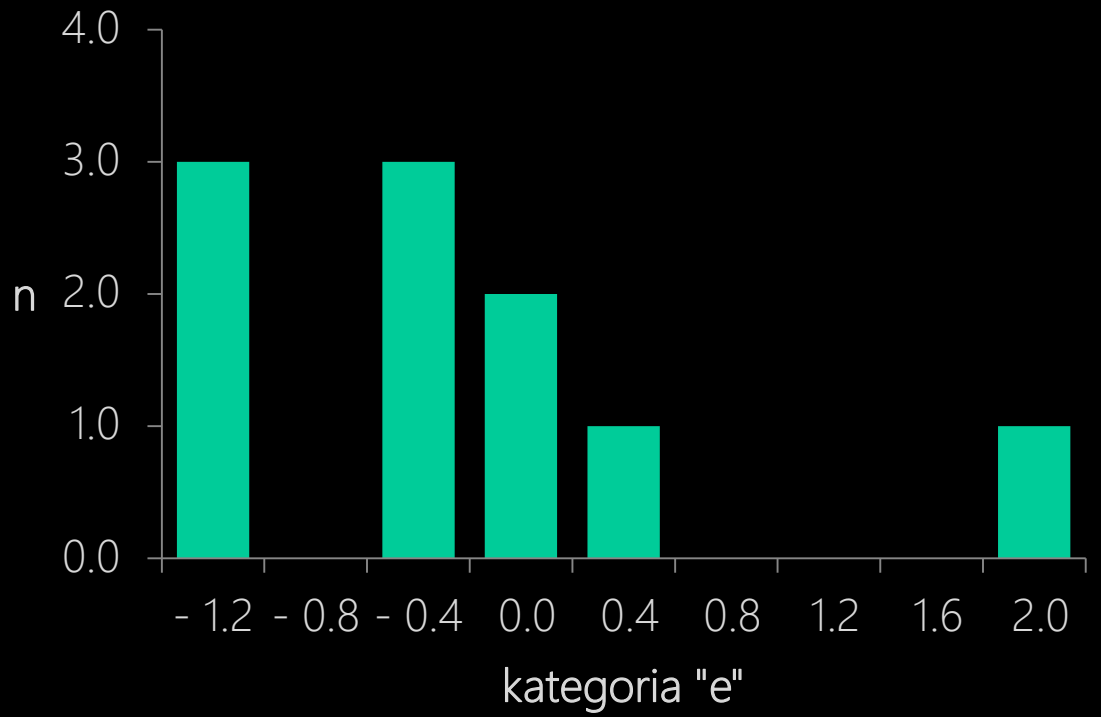
Zastosować inne / dodatkowe zmienne niezależne

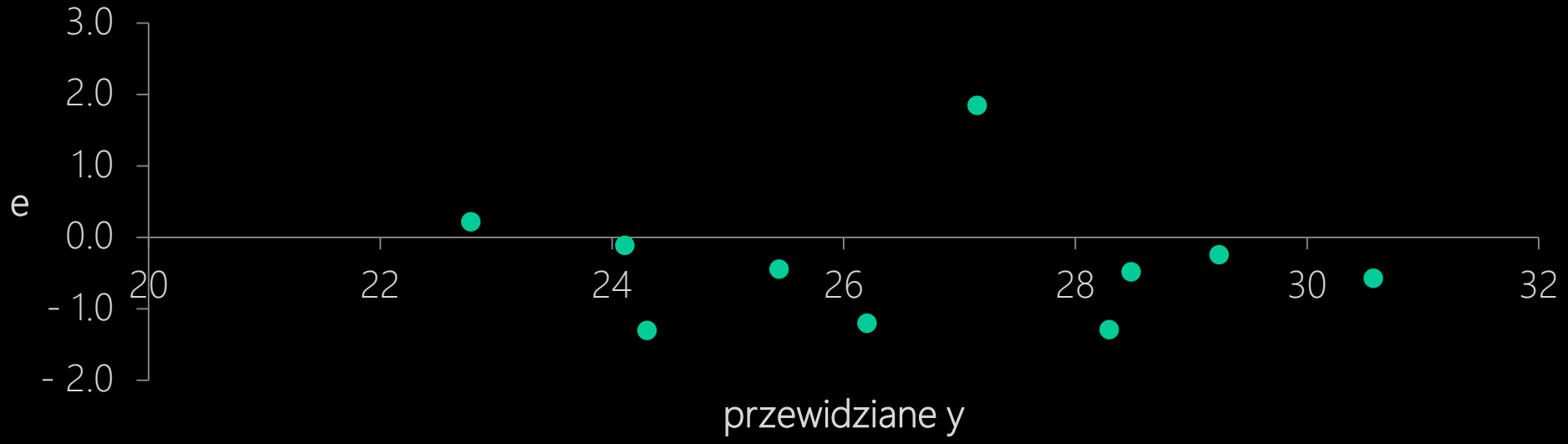
błąd x przewidziany „y” → jakość dopasowania regresji

ODSTAJĄCE OBSERWACJE

- Mają duży wpływ na estymatory wsp. równania regresji
- Sprawdzić dane - błędna wartość

Histogram





REGRESJA LOGISTYCZNA

Dopasowanie regresji nieliniowej - zmienność

zmienność "y"

wyjaśniona przez równanie regresji

zaobserwowana

$$\log L_r = \sum_{i=1}^{n} \left(\log \binom{n_i}{y_i} + y_i \log \square + (n_i - y_i) \log \left(1 - \square\right) \right)$$

$$\log L_{obs} = \sum_{i=1}^{n} \left(\log \binom{n_i}{y_i} + y_i \log \frac{y_i}{n_i} + (n_i - y_i) \log \left(1 - \square\right) \right)$$

$$D = -2\left[\log L_r - \log L_{obs}\right] \sim \chi^2_{n-p}$$

Dopasowanie regresji nieliniowej - zmienność

jaka część obserwowanej zmienności została wyjaśniona przez równanie regresji

$$D = -2[\log L_r - \log L_{obs}] \sim \chi^2_{n-p}$$

Dopasowanie regresji nieliniowej - zmienność

dane

$$\text{logit}(p) = \log\frac{p}{1-p} = -5.340 + 0.001548x$$

nacisk	ilość całkow.	ilość uszkod.
2500	50	10
2700	70	17
	...	
4300	65	51

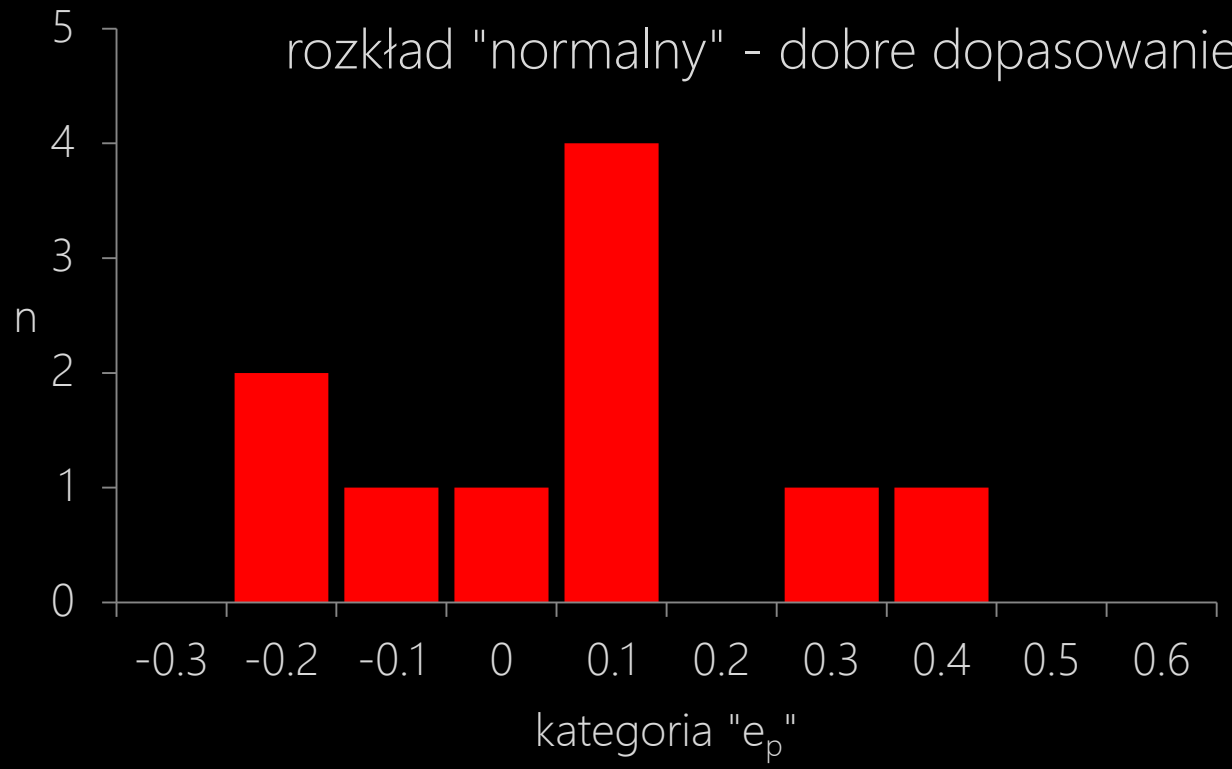
1. $\alpha_{MAX} = 0.05$
2. $D=0.3719 \sim \chi^2_{(10-2)st.sw.}$
3. $\alpha_T = 0.999957$
4. H_0
5. Dobre dopasowanie równania regresji

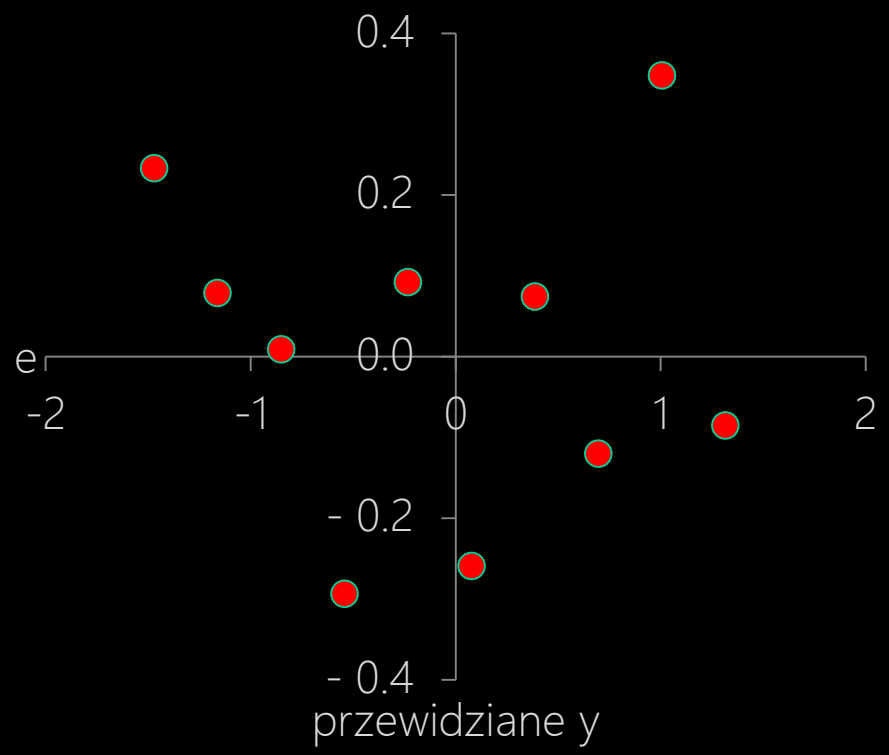
Dopasowanie regresji nieliniowej - błędy

$$(y_i - \hat{y}_i) = e_i \neq N(0, \sigma_e^2)$$

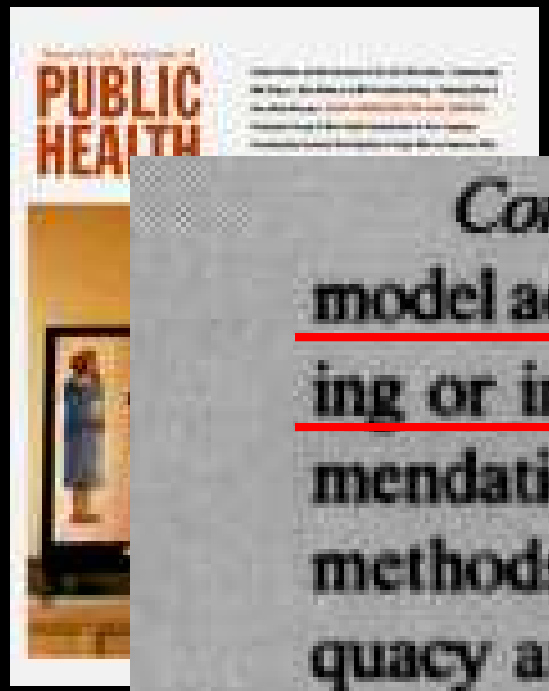
$$e_p = \frac{y_i - \hat{y}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$$

- Zależne od liczby obserwacji
- Błędy skorygowane na odchylenie stand.
- Błędy = reszty Pearson'a





Przykład



The Importance of Assessing the Fit

Conclusions. Failure to address model adequacy may lead to misleading or incorrect inferences. Recommendations are made for the use of methods for assessing model adequacy and for future editorial policy in regard to the review of articles using logistic regression. (*Am J Public Health.* 1991;81:1630–1635)

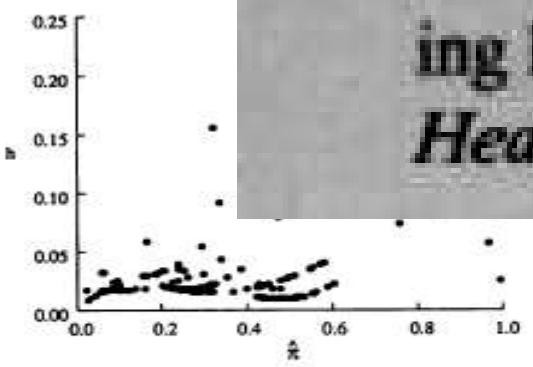


FIGURE 1—Plot of leverage vs the estimated logistic probabilities for the fitted model in Table 3.

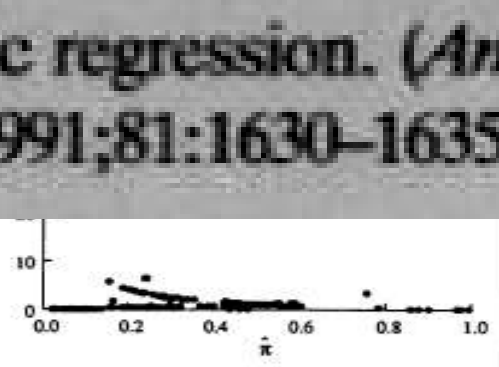


FIGURE 2—Plot of ΔX^2 vs the estimated logistic probabilities for the fitted model in Table 3.

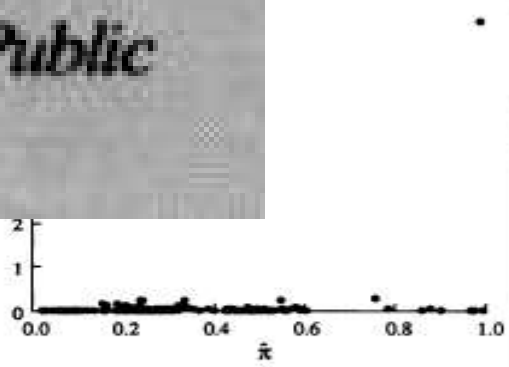


FIGURE 3—Plot of $\Delta\hat{\beta}$ vs the estimated logistic probabilities for the fitted model in Table 3.

testowanie jakości dopasowania równania regresji