

OKREŚLANIE JAKOŚCI DOPASOWANIA RÓWNANIA REGRESJI LINIOWEJ

Dawid Słomian

testowanie jakości dopasowania równania regresji

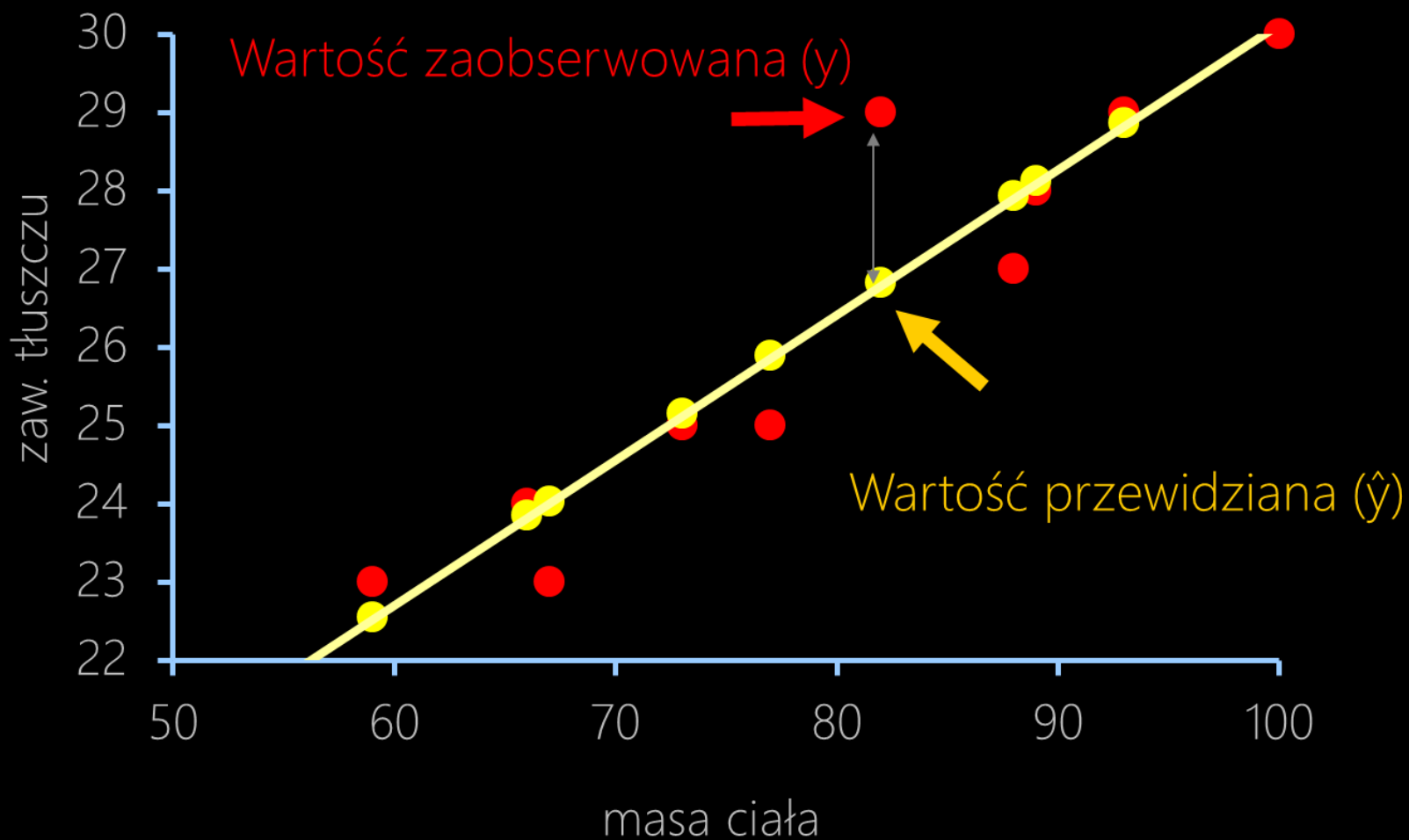
statystyki:

- R^2
- D

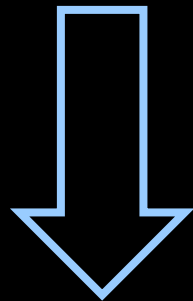
wykresy
diagnostyczne

RÓWNANIE REGRESJI

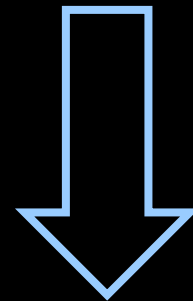
elementy równania regresji: błąd



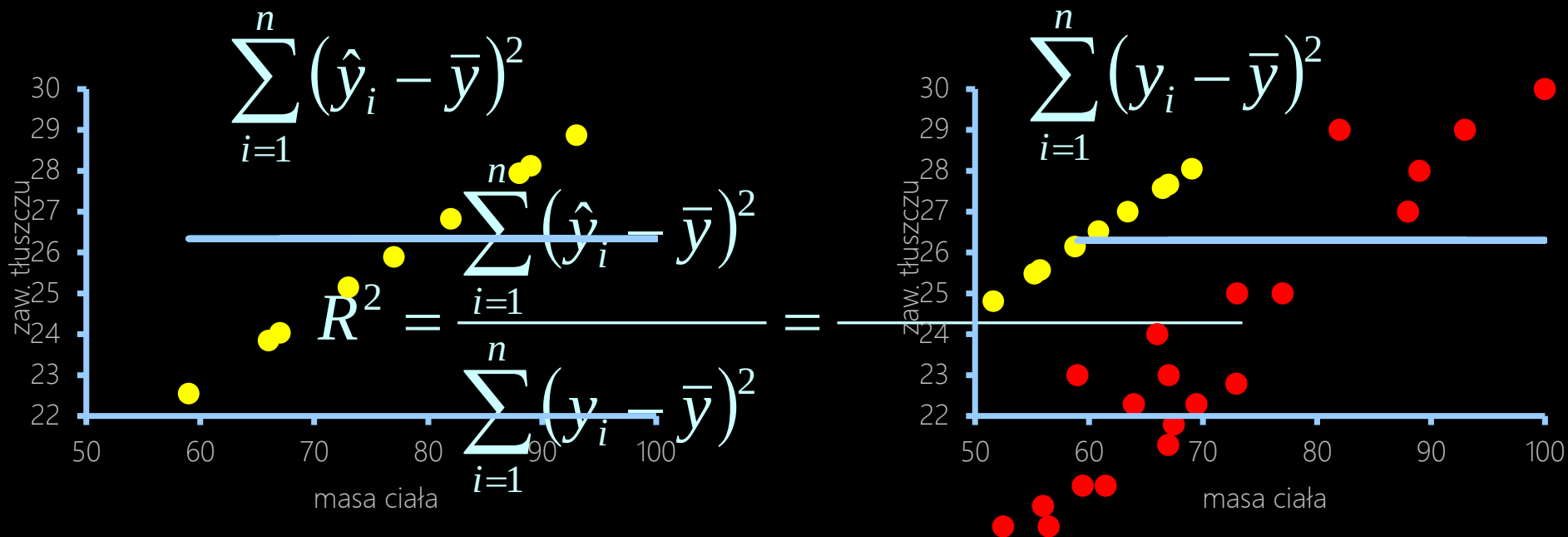
zmienna "y"

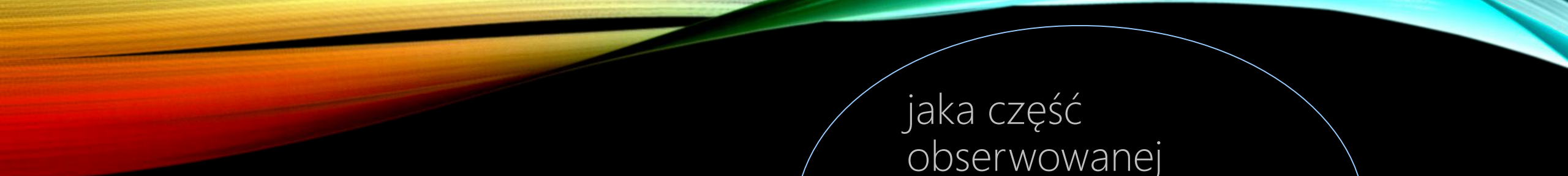


wyjaśniona przez równanie regresji



zaobserwowana





jaka część
obserwowanej
zmienności została
wyjaśniona przez
równanie regresji

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

PRÓBA DANYCH

1. Zmienna niezależna
2. Zmienna zależna, rozkład ciągły

$$\text{tuszcz} = 11.57 + 0.19\text{masa_ciała}$$

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



$R^2 0.94$



=

94% zmienności zawartości tłuszczu
jest wyjaśnione przez zmienność masy
ciała

RÓWNANIE REGRESJI

Co należy sprawdzić w równaniu regresji?

kryterium LINE:

- Linearity → zależność $E(Y | X)$ jest liniowa w stosunku do X
- Independence → y_i są niezależne od siebie
- Normality → Y pochodzi z rozkładu normalnego
- Equal variance → wariancja Y jest niezależna od X

BŁĘDY

Jak sprawdzić kryterium LINE?

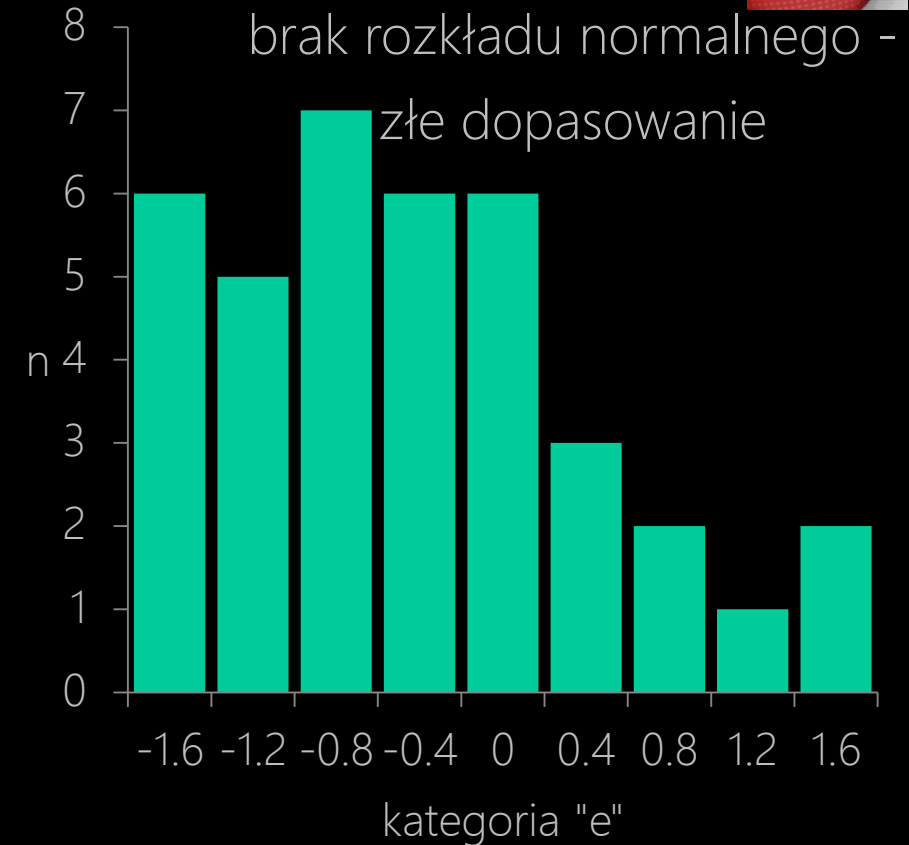
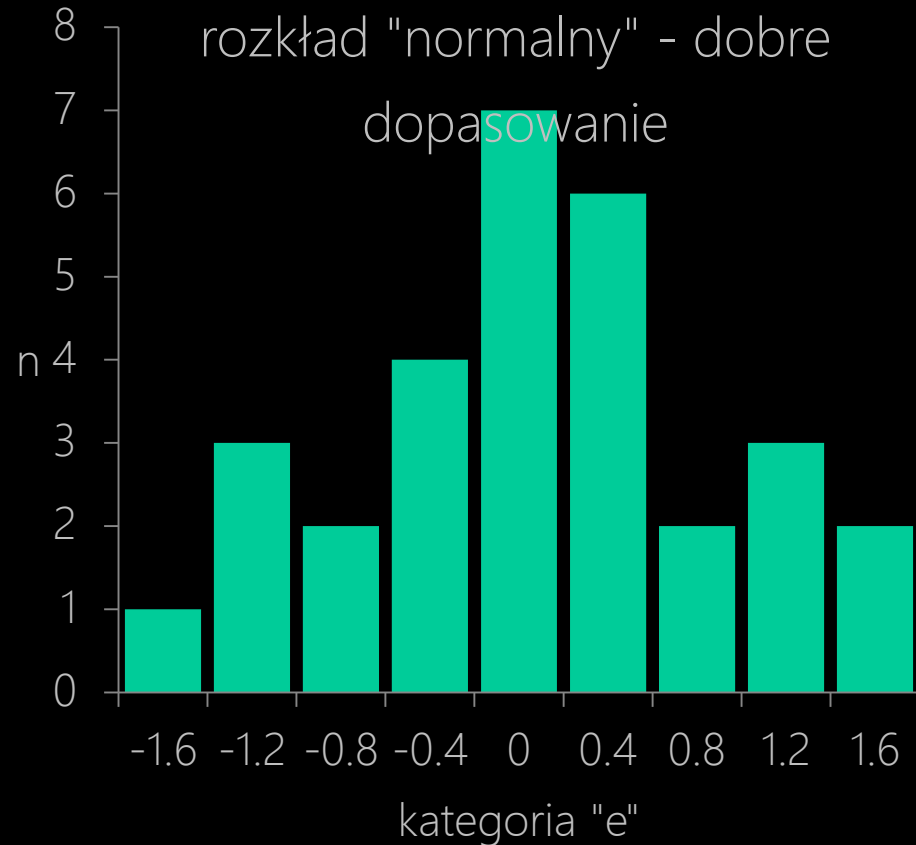
$$(y_i - \hat{y}_i) = e_i$$

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

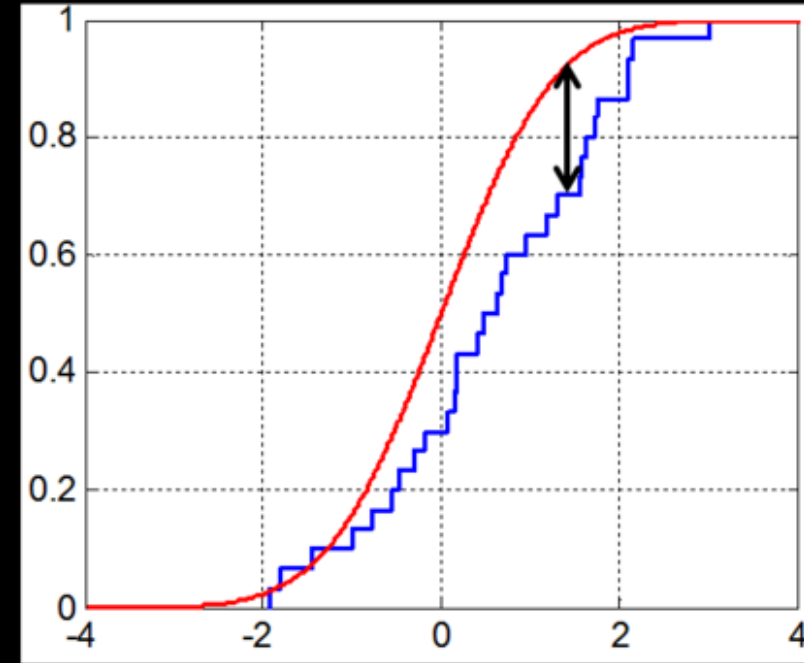
- np. test Kolmogorova-Smirnova
- H_0 : rozkłady są zgodne, H_1 : rozkłady nie są zgodne
- ...

- $D = \max_x |F_{exp}(x) - F_{obs}(x)|$

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

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

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



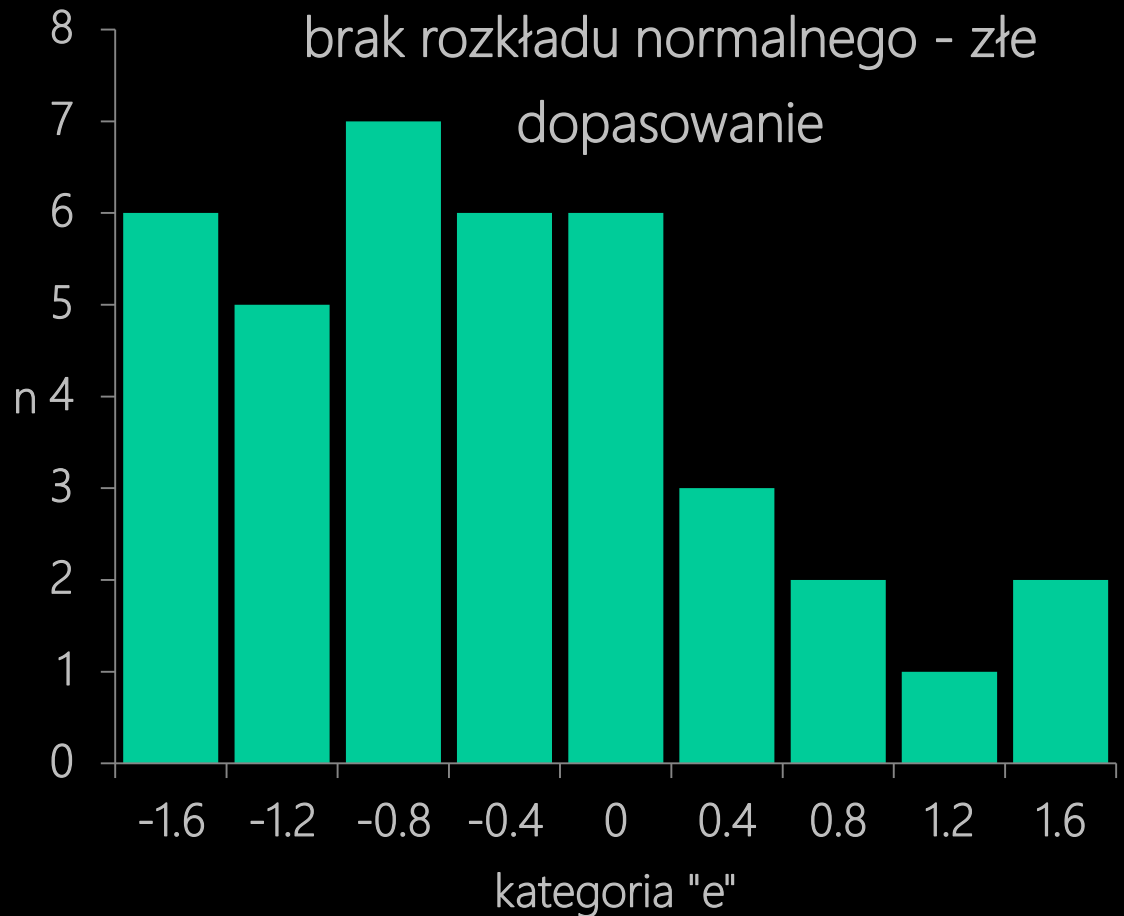
$$(y_i - \hat{y}_i) = e_i \sim N(0, \sigma_e^2)$$



ZŁE DOPASOWANIE

Zastosować transformację „y”:

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



QQPLOT – 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

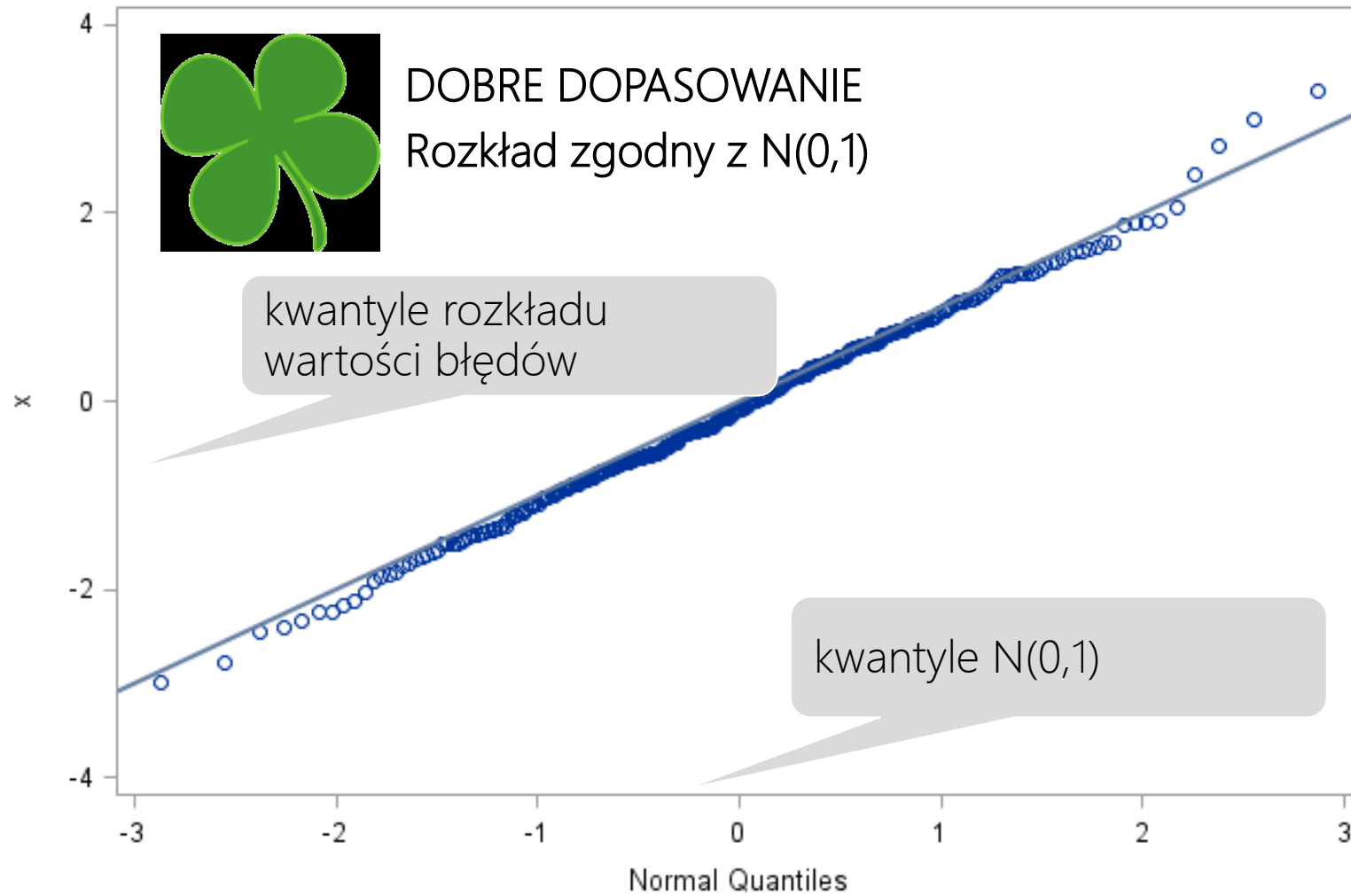
Q-Q Plot for x



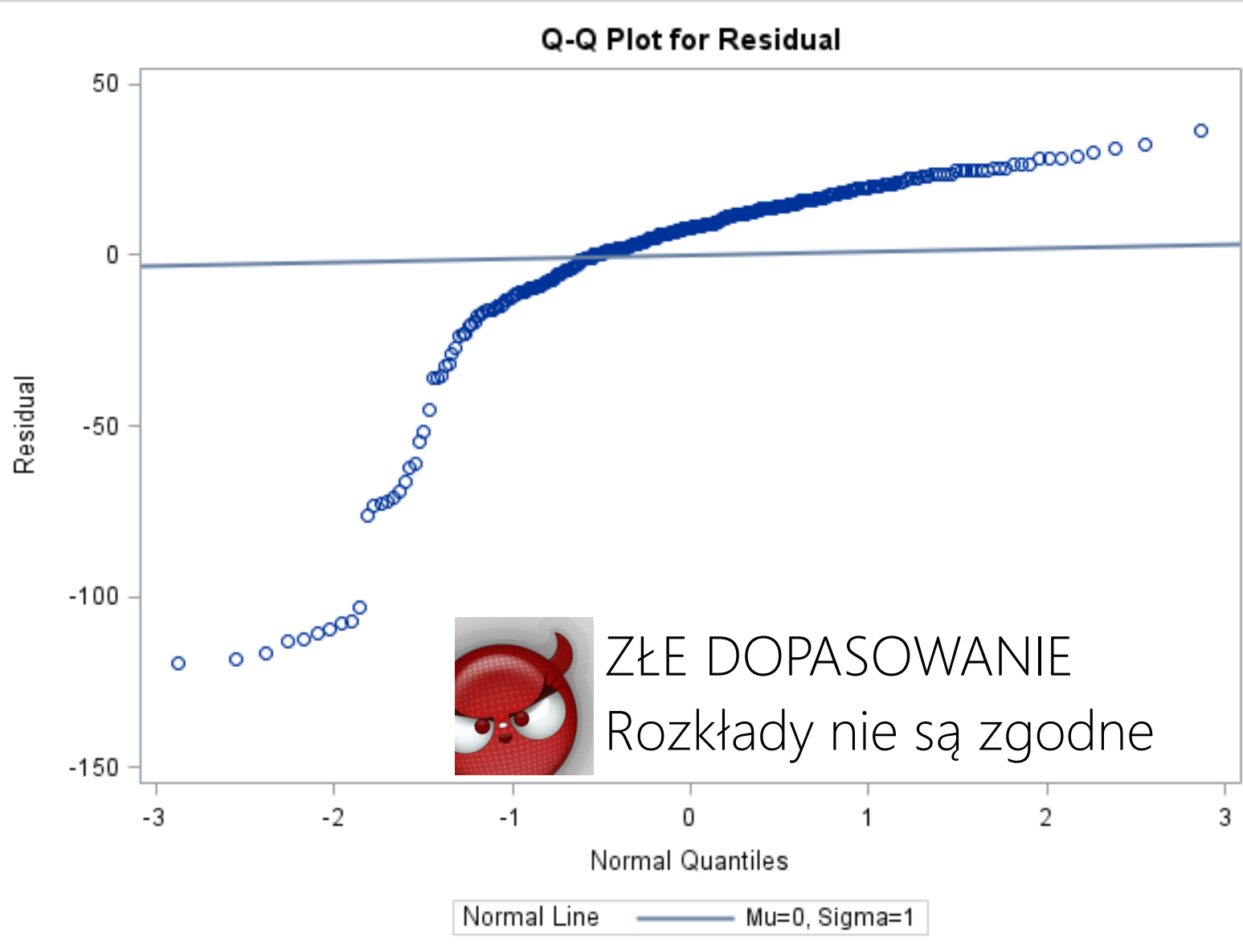
DOBRE DOPASOWANIE
Rozkład zgodny z $N(0,1)$

kwantyle rozkładu
wartości błędów

kwantyle $N(0,1)$



Normal Line — Mu=0, Sigma=1



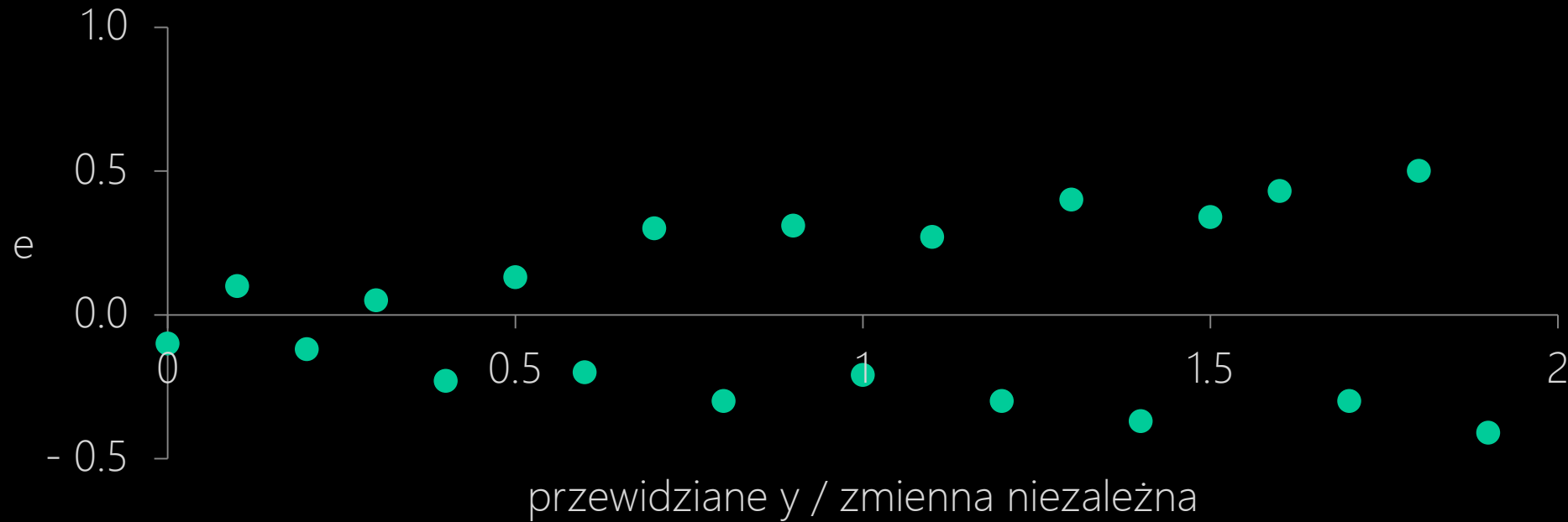
BŁĄD X PRZEWIDZIANY „Y” → JAKOŚĆ DOPASOWANIA REGRESJI



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

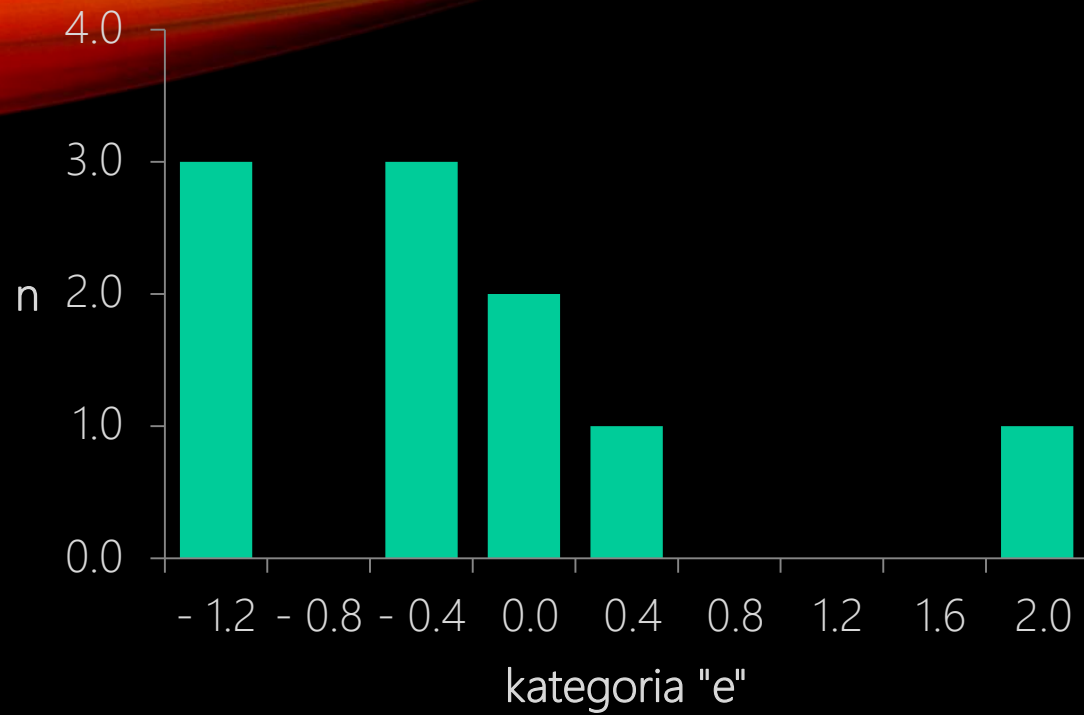
BŁĄD X PRZEWIDZIANY „Y” → JAKOŚĆ DOPASOWANIA REGRESJI



ODSTAJĄCE OBSERWACJE

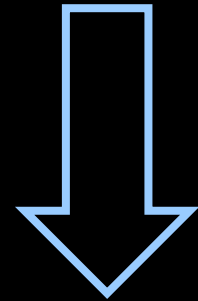
- Mają duży wpływ na estymatory wsp. równania regresji
- Sprawdzić dane - błędna wartość

HISTOGRAM

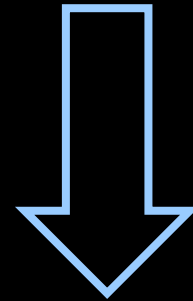


REGRESJA LOGISTYCZNA

zmienna "y"



wyjaśniona przez równanie
regresji

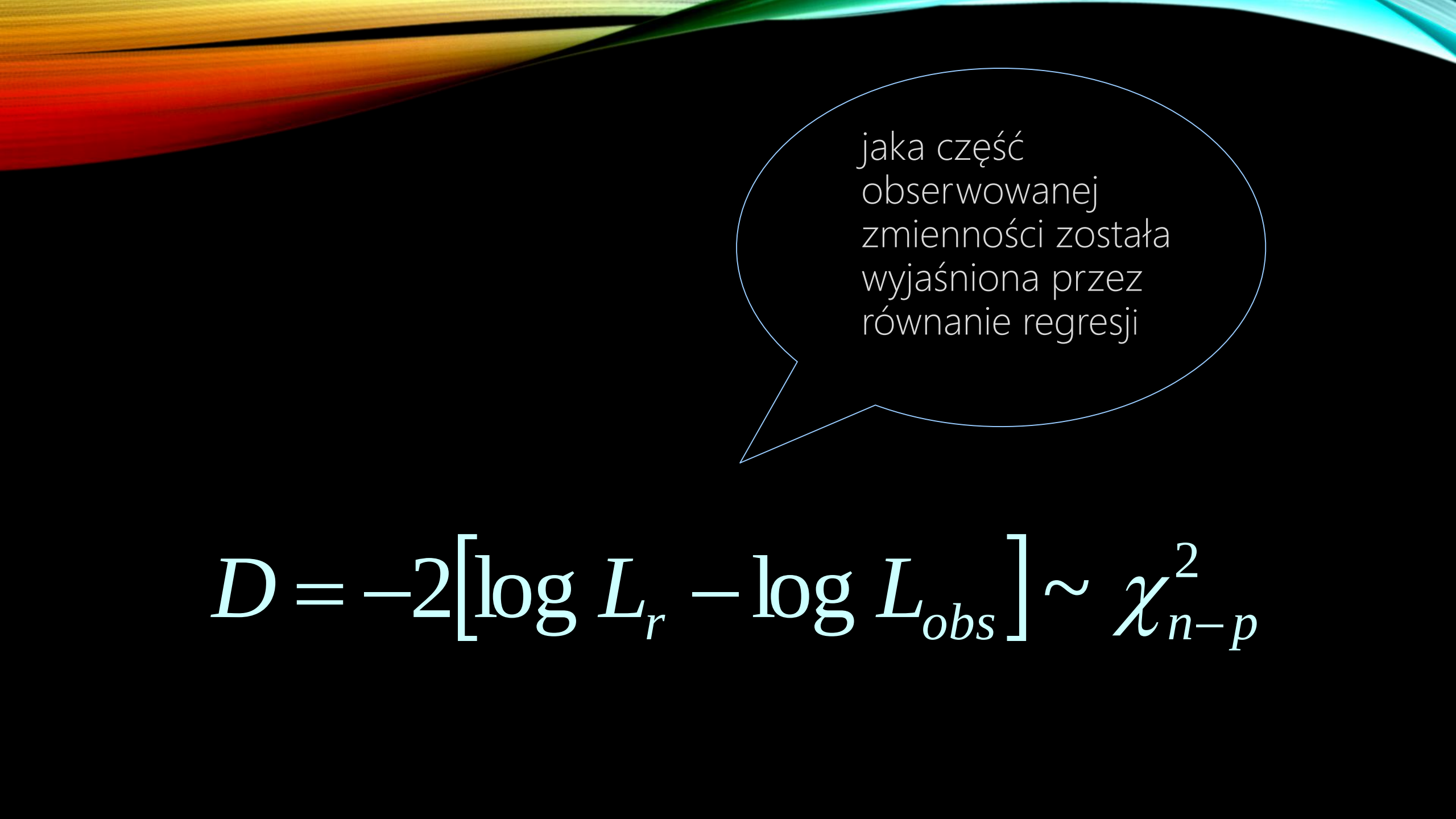


zaobserwowana

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

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

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



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

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

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

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

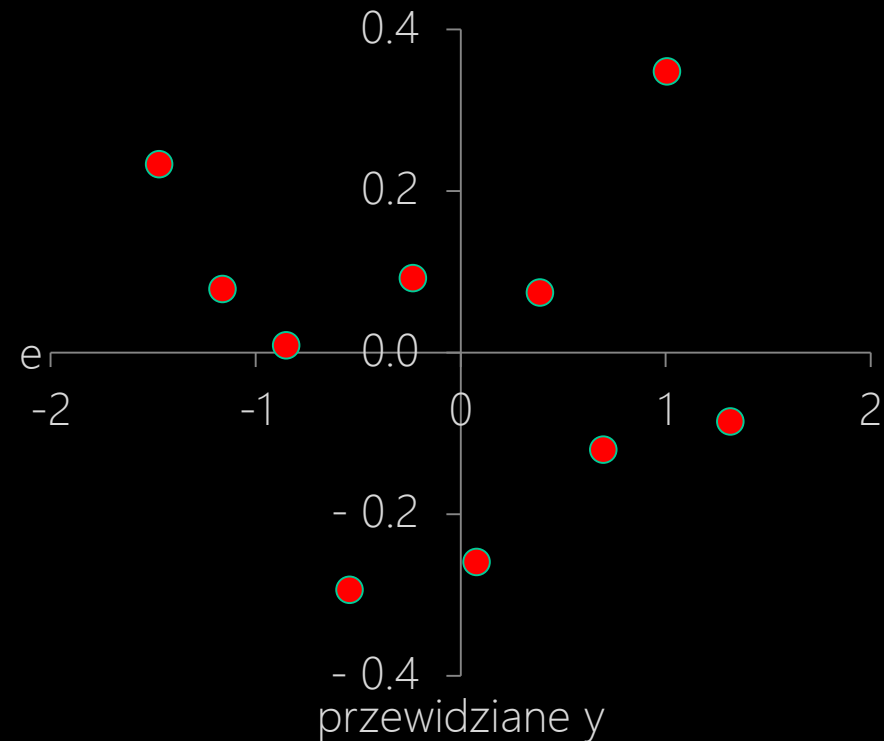
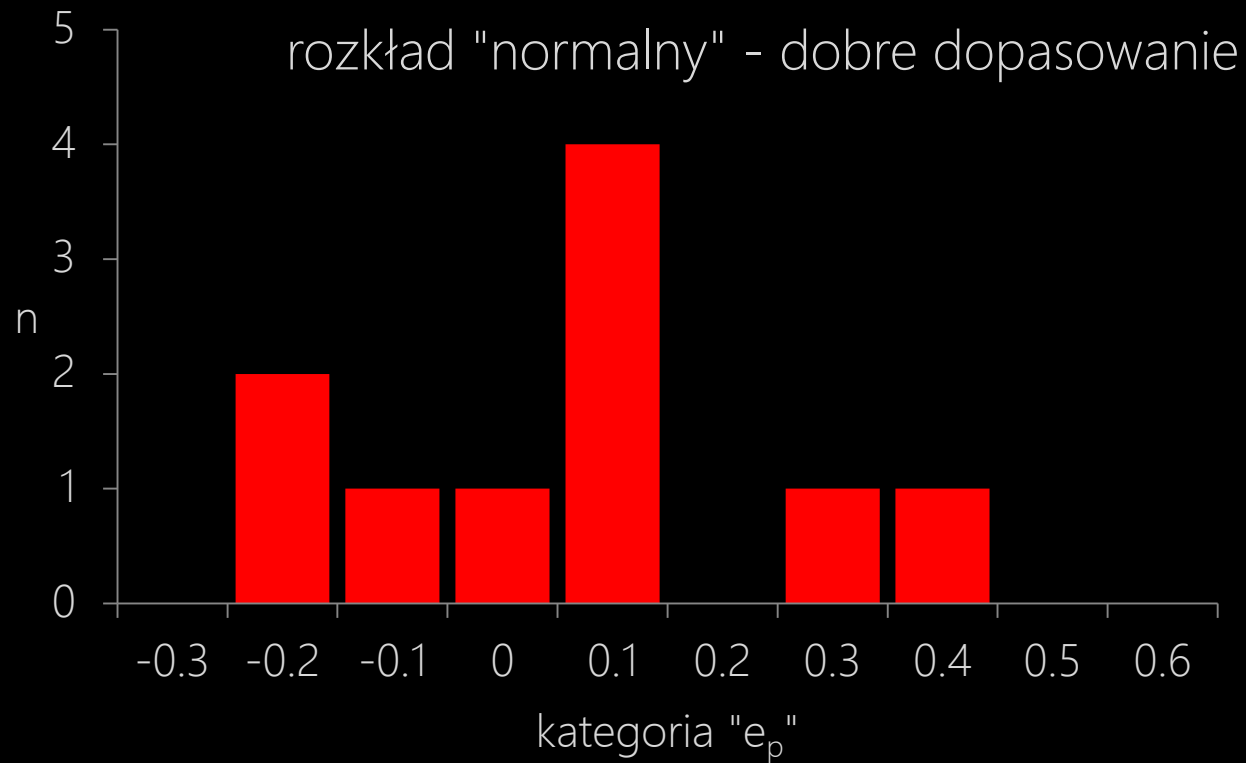


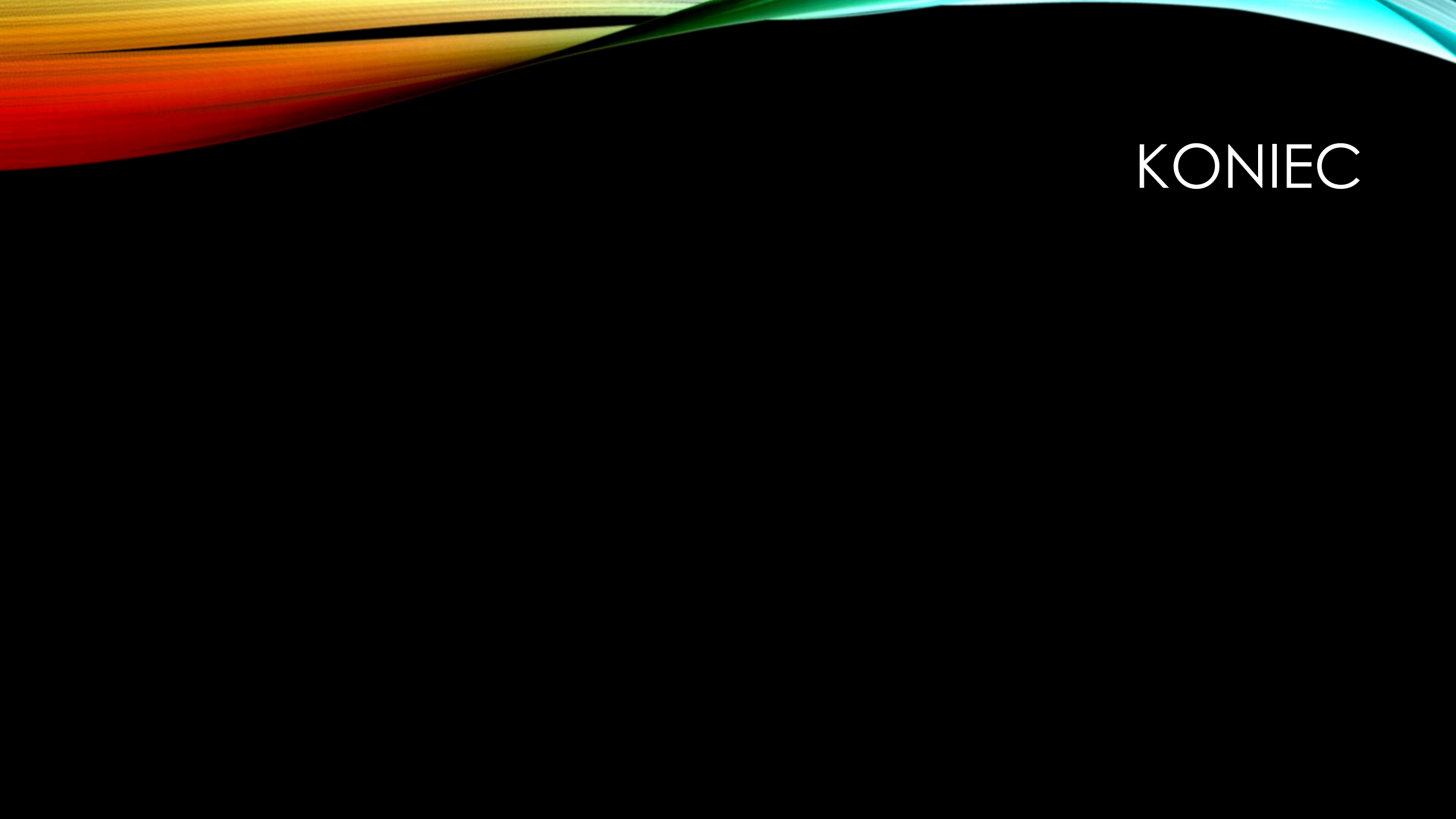
DOPASOWANIE REGRESJI NIELINIOWEJ - BŁĘDY

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

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

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





KONIEC