

analiza danych

1. Wprowadzenie
2. Regresja liniowa – przypomnienie i najprostszy klasyfikator - regresja logistyczna
3. Metody wyboru zmiennych oparte na równaniu regresji
4. Metody preselekcji zmiennych
5. Transformacje – embedding ①
6. Transformacje – embedding ②
7. Prezentacja projektów

Plan wykładu

1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

- Równanie regresji logistycznej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście klasyfikacji

1. Predykcja → Regresja liniowa

- **Równanie regresji liniowej**
- Estymacja współczynników równania
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

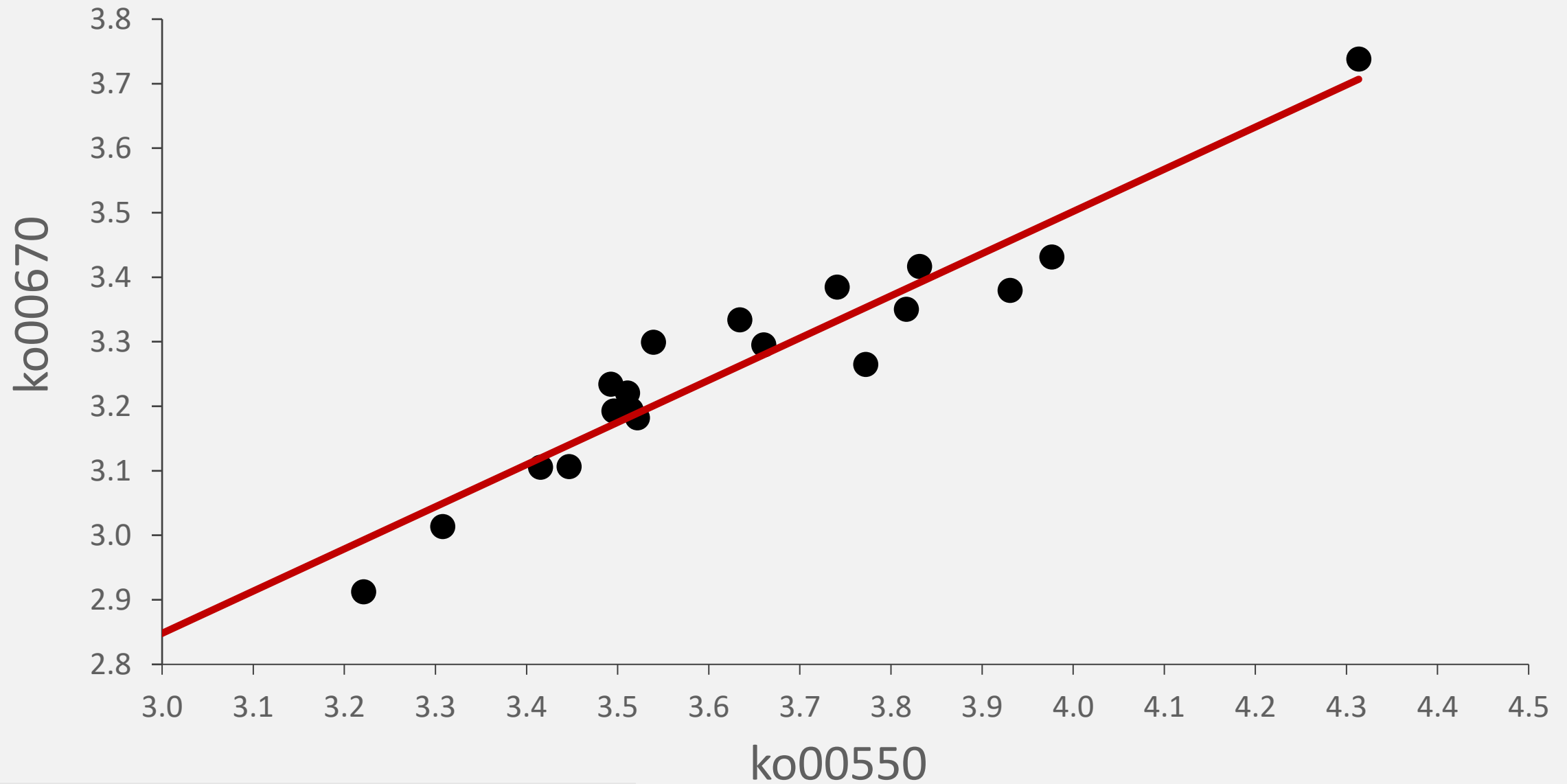
- Równanie regresji logistycznej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście klasyfikacji

Równanie regresji liniowej

ko00550 = X	ko00670 = Y
3.660441903	3.295163266
3.816934062	3.350180872
3.740988692	3.384414410
3.831473084	3.416426693
3.492409600	3.234108014
3.772208127	3.264723303
3.930720948	3.379615594
4.313441202	3.738029771
3.976612106	3.431232349
3.521691667	3.182005509
3.633996584	3.333867499
3.514567367	3.194138026
3.510741551	3.220545514
3.415240961	3.105287833
3.495769055	3.192413515
3.221046948	2.912379310
3.539335329	3.299250591
3.307837658	3.013511785
3.446605760	3.106521577

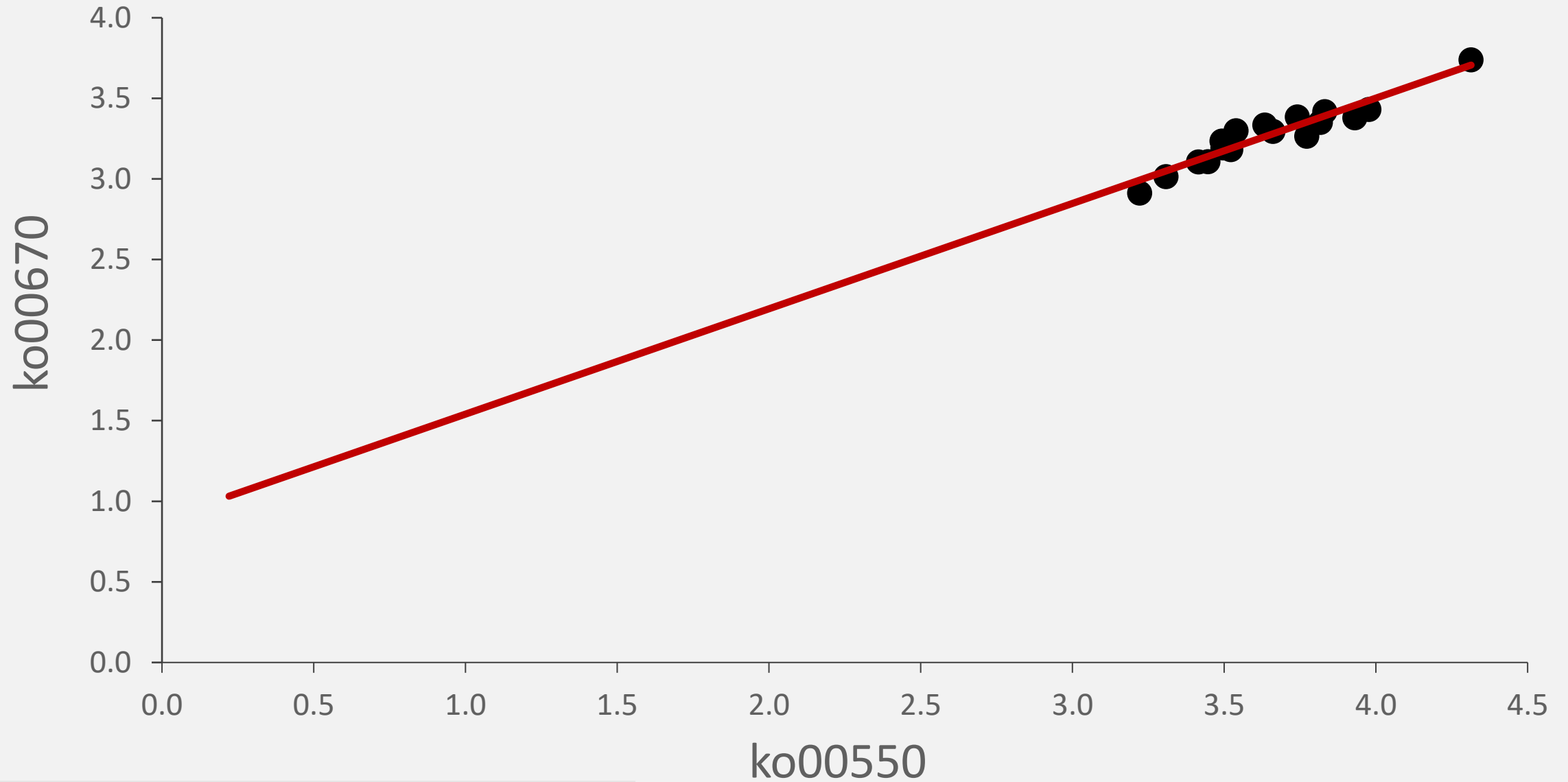
- Regresja liniowa → wyznaczenie prostej, która pozwala na predykcję wartości zmiennej Y na podstawie znajomości zmiennej X
- Zmienna niezależna (X)
Dokładnie znane wartości / wartości eksperymentalne
- Zmienna zależna (Y)
Zakładamy błąd pomiarowy

Równanie regresji liniowej



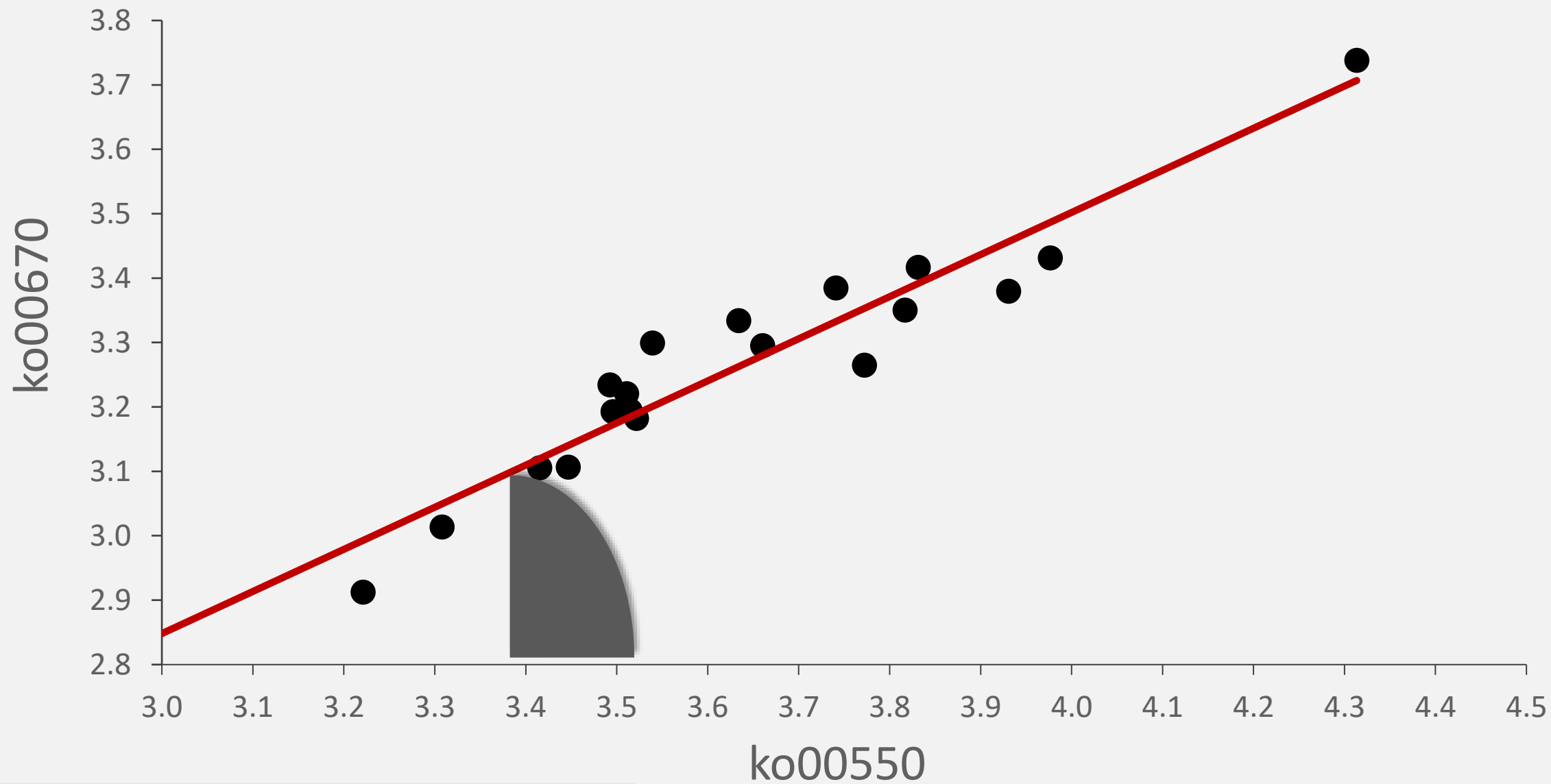
Wyraz wolny, intercept, b , β_0

Równanie regresji liniowej



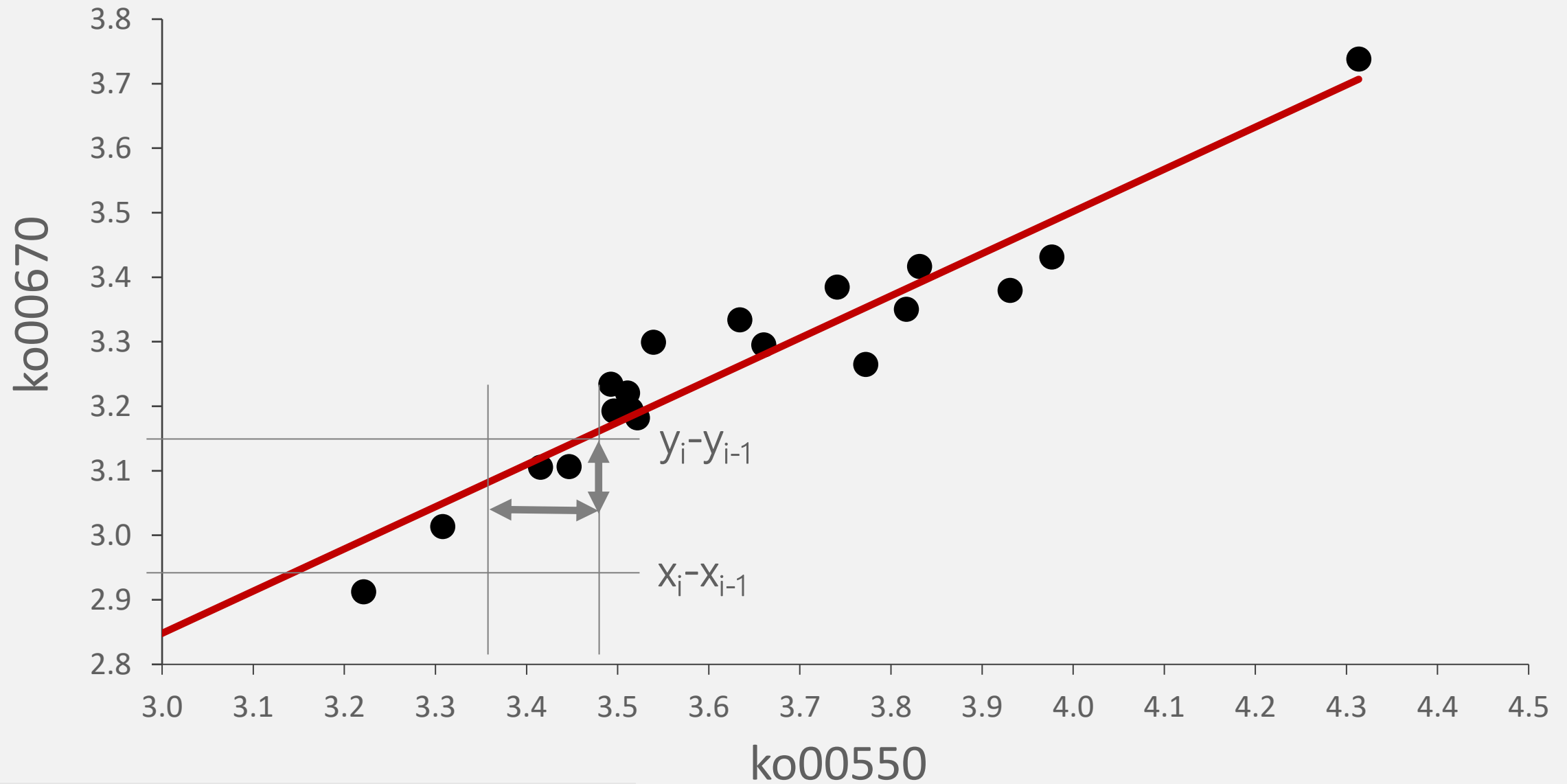
Wyraz wolny → interpretacja

Równanie regresji liniowej



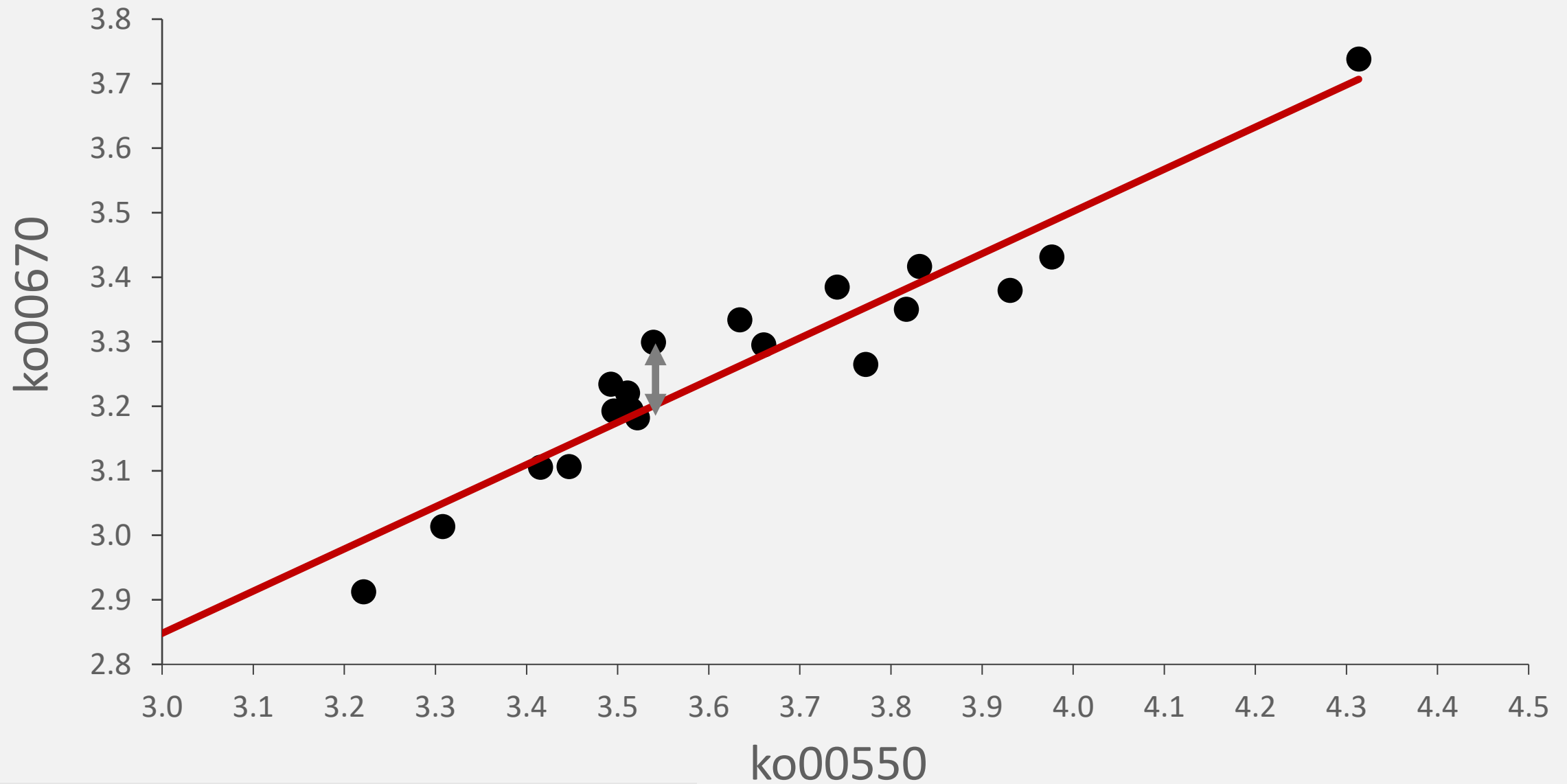
Gradient, slope, a , β_1

Równanie regresji liniowej



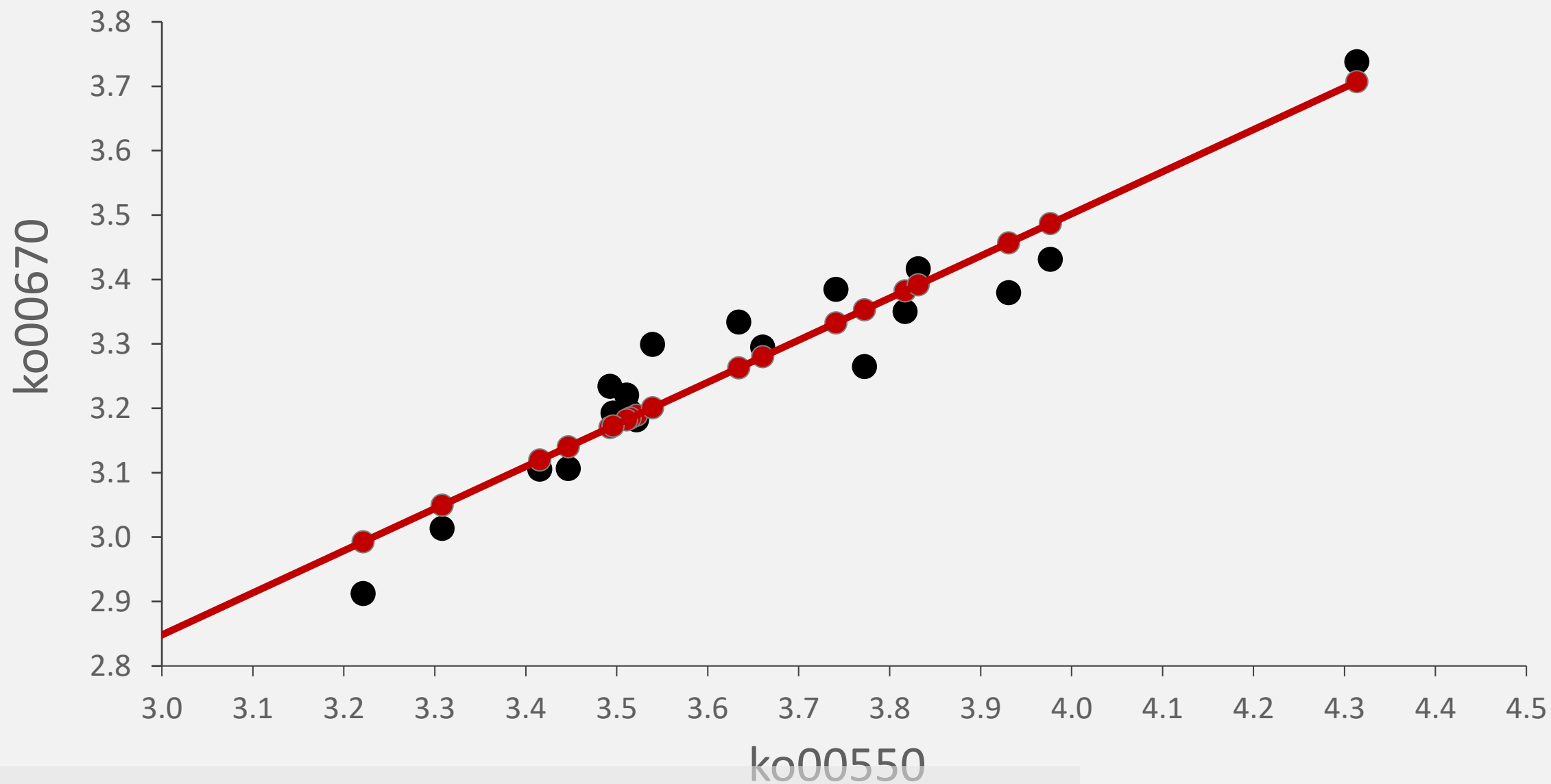
Gradient → interpretacja

Równanie regresji liniowej



Błąd, error, residual → interpretacja

Równanie regresji liniowej



Wartości zaobserwowane ●, wartości przewidziane ●

Równanie regresji liniowej

wyraz wolny

gradient

błąd

$$y = \beta_0 + \beta_1 x + e$$

zmienna zależna

zmienna niezależna

Równanie regresji liniowej

0.89 0.65 ?

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + e$$

ko00670 ko00550

The diagram illustrates the components of the linear regression equation $y = \hat{\beta}_0 + \hat{\beta}_1 x + e$. Arrows point from labels to each term: '0.89' points to $\hat{\beta}_0$, '0.65' points to $\hat{\beta}_1$, '?' points to x , 'ko00670' points to y , and 'ko00550' points to e .

Równanie regresji liniowej

Zmienna zależna

Zmienna niezależna

Left Hand Side

Right Hand Side

LHS

RHS

$$y = \beta_0 + \beta_1 x + e$$

Równanie regresji liniowej

Komponent systematyczny

$$\beta_0 + \beta_1 x$$

Komponent losowy

$$e \sim N(0, \sigma_e^2)$$

$$y = \beta_0 + \beta_1 x + e$$

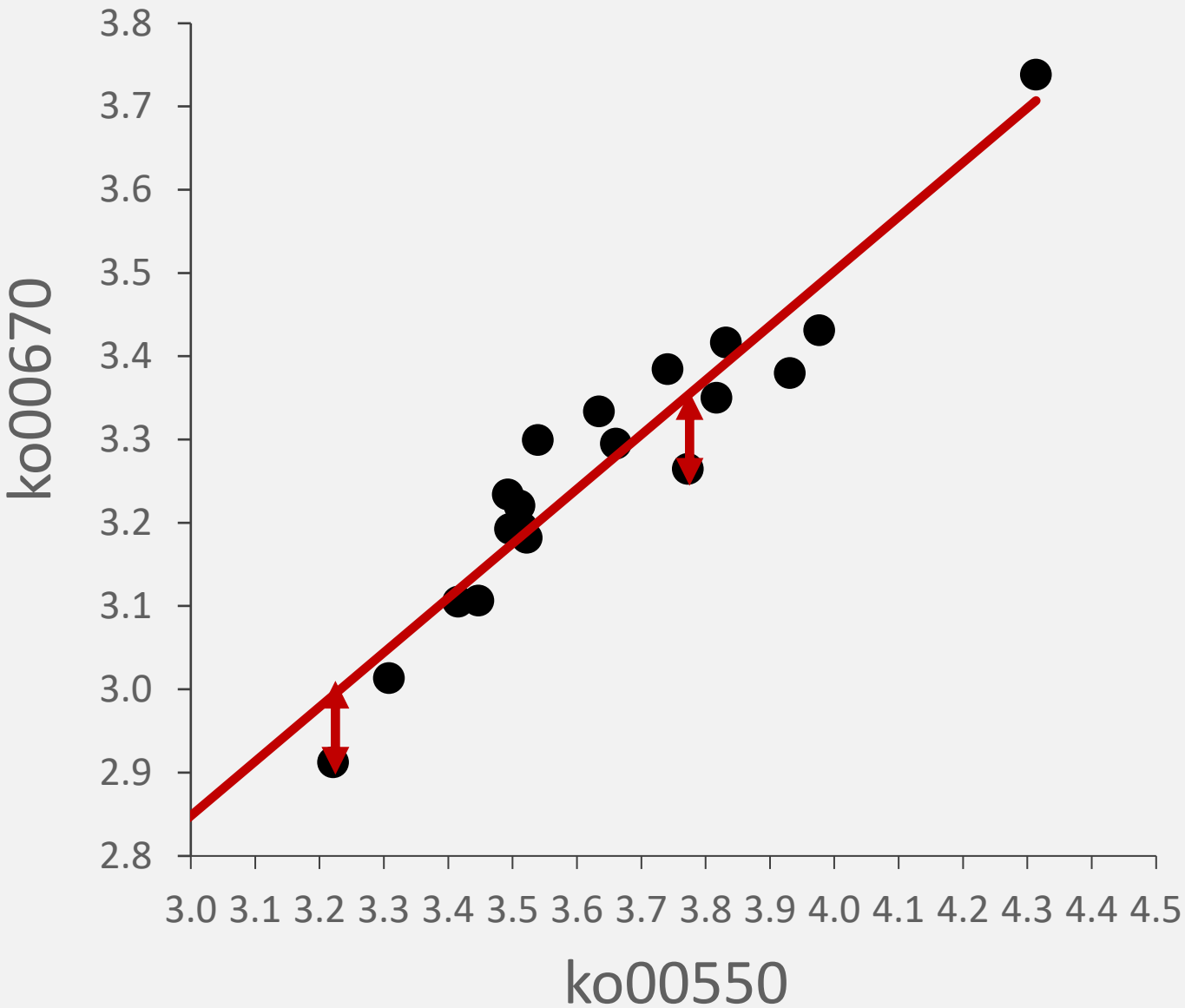
1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- **Estymacja współczynników równania**
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

- Równanie regresji logistycznej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście klasyfikacji

Estymacja współczynników równania regresji liniowej



- Jakie kryterium zastosować do uzyskania $\hat{\beta}_0$ i $\hat{\beta}_1$?
- $\Sigma(\updownarrow)^2 \rightarrow \text{minimum}$
- Metoda (ważonych) najmniejszych kwadratów

Estymacja współczynników równania regresji liniowej

- Wyznaczyć $\hat{\beta}_0$ i $\hat{\beta}_1$ gwarantujące minimum funkcji: $\sum_{i=1}^N (y_i - \hat{y}_i)^2$
- Wyznaczyć $\hat{\beta}_0$ i $\hat{\beta}_1$ gwarantujące minimum funkcji: $\sum_{i=1}^N (y_i - [\beta_0 + \beta_1 x_i])^2$
- $$\frac{\partial \left\{ \sum_{i=1}^N (y_i - [\beta_0 + \beta_1 x_i])^2 \right\}}{\partial (\beta_0)} = 0 \quad \rightarrow \quad -2 \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) = 0$$
- $$\frac{\partial \left\{ \sum_{i=1}^N (y_i - [\beta_0 + \beta_1 x_i])^2 \right\}}{\partial (\beta_1)} = 0 \quad \rightarrow \quad -2 \sum_{i=1}^N x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$
- $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- Estymacja współczynników równania
- **Interpretacja wyników w kontekście predykcji**

2. Klasyfikacja → Regresja logistyczna

- Równanie regresji logistycznej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście klasyfikacji

$$\text{ko00670} = 0.89 + 0.65 \cdot \text{ko00550} + e$$

- Co nam mówią wartości współczynników regresji?
- Czy w mikrobiomie jelitowym ryb ścieżki metaboliczne ko00670 oraz ko00550 są antagonistyczne względem siebie?
- Jaka będzie liczebność odczytów sekwencji korespondujących ze ścieżką ko00670 w mikrobiomie ryby, u której zidentyfikowano liczebność ścieżki ko00550 = 3.51?
- Jak można by ulepszyć ten model dla informacji o ko00670 wykorzystując nasze dane ? →

Interpretacja

$$\text{ko00670} = 0.89 + 0.65 \cdot \text{ko00550} + e$$

- Jak można by ulepszyć ten model dla informacji o ko00670 wykorzystując nasze dane ?
- Uwzględnić inne ścieżki

ko00550	ko00670	ko00524	ko04622	ko02010	ko03410	ko00280	ko05012	ko00603	ko04650	ko04930	ko00510	ko03050
3.660441903	3.295163266	0.7470717626	-4.393766742	5.25120709	3.019666163	3.2799314	1.481156786	1.271772319	-8.057328388	1.263554649	-3.996885378	-8.057328388
3.816934062	3.350180872	0.8388024265	-3.870926483	5.244817357	3.062652333	3.101492585	1.1762109	1.132378085	-8.090434188	1.227160205	-1.930282276	-8.090434188

- Ile / które ?

1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

- Równanie regresji logistycznej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście klasyfikacji

Równanie regresji liniowej – problem



Założenia regresji liniowej

- Liniowa zależność między Y a X 😞
- Niezależność między X_i → no collinearity 😞
- $p(e|X) \sim N(0, I\sigma_e^2)$
- Niezależność między e → brak kowariancji
- Jednakowa wariancja e dla każdej wartości x → homoscedascity

Równanie regresji liniowej

```
fit1 <- lm(y~x)  
p1 <- predict(fit1)
```

1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

- **Równanie regresji logistycznej**
- Estymacja współczynników równania
- Interpretacja wyników w kontekście klasyfikacji

Równanie regresji logistycznej

ko00550 = X	Staw = Y
3.660441903	0
3.816934062	0
3.740988692	0
3.831473084	0
3.492409600	0
3.772208127	0
3.930720948	0
4.313441202	0
3.976612106	0
3.521691667	1
3.633996584	1
3.514567367	1
3.510741551	1
3.415240961	1
3.495769055	1
3.221046948	1
3.539335329	1
3.307837658	1
3.446605760	1

- Regresja logistyczna → obliczenie prawdopodobieństwa wystąpienia zdarzenia na podstawie znajomości zmiennej X
- Np. obliczenie prawdopodobieństwa że osobnik pochodził ze stawu eksperymentalnego (Y=1) na podstawie liczebności występowania ko00550 (X)
- Klasyfikacja – 2 klasy
- Zmienna zależna = prawdopodobieństwo

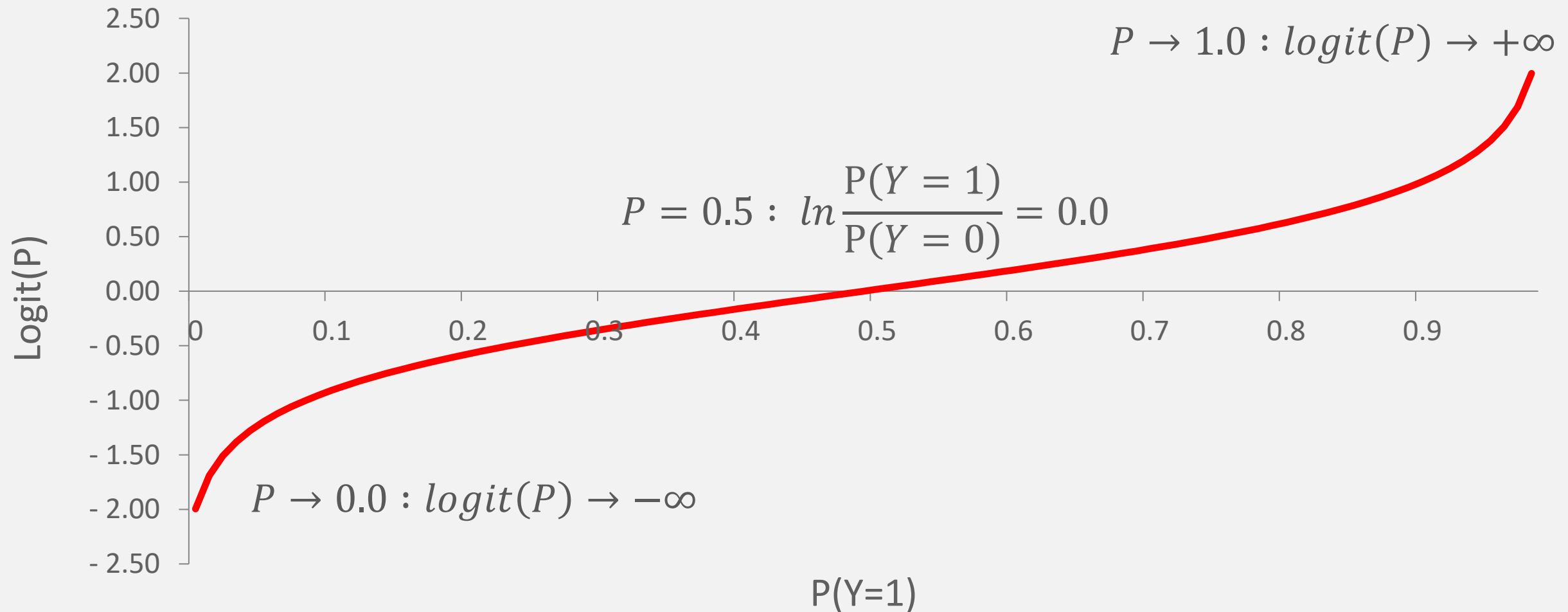
Równanie regresji logistycznej

$$P(Y = 1) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

- RHS 😞
- Powrót do $\beta_0 + \beta_1 x$ jak w regresji liniowej
- Jak ???
- Transformacja LHS $\rightarrow$ przejście ze skali $[0 : 1]$ na skalę $[-\infty : +\infty]$
- Jak ???
- Zamiast $P(Y = 1)$ zastosować $\ln \frac{P(Y=1)}{1-P(Y=1)} = \ln \frac{P(Y=1)}{P(Y=0)} = \textit{logit}(P)$

Równanie regresji logistycznej

$$\text{logit}(P) = \ln \frac{P(Y = 1)}{P(Y = 0)} = \beta_0 + \beta_1 X$$



1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

- Równanie regresji logistycznej
- **Estymacja współczynników równania**
- Interpretacja wyników w kontekście klasyfikacji

Estymacja współczynników równania regresji logistycznej

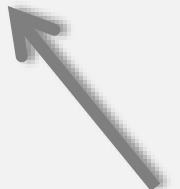
$$\text{logit}(P) = \ln \frac{P(Y = 1)}{P(Y = 0)} = \beta_0 + \beta_1 X$$

- Fisher's method of scoring
- Metoda iteracyjna

- $$\hat{\beta}^{t+1} = \hat{\beta}^t + \frac{\sum_{i=1}^N (x_i (y_i - p_i))}{-E \left(\frac{\partial^2 \left[\binom{n}{k} p^{(t)k} (1-p^{(t)})^{(n-k)} \right]}{\partial \beta_0 \partial \beta_1} \right)}$$



Pochodna funkcji regresji logistycznej
 y_i – klasa
 p_i – prawdopodobieństwo



Fisher's information matrix
Wartość oczekiwana macierzy drugich pochodnych
 $2 \times 2 (\beta_0 \beta_1)$

1. Predykcja → Regresja liniowa

- Równanie regresji liniowej
- Estymacja współczynników równania
- Interpretacja wyników w kontekście predykcji

2. Klasyfikacja → Regresja logistyczna

- Równanie regresji logistycznej
- Estymacja współczynników równania
- **Interpretacja wyników w kontekście klasyfikacji**

Interpretacja

$$\text{logit}(P) = \ln \frac{P(Y = 1)}{P(Y = 0)} = \beta_0 + \beta_1 X$$

- Interpretacja $\hat{\beta}_0$ i $\hat{\beta}_1$ jak w regresji liniowej

$$\text{logit}(P) = \ln \frac{P(Y = 1)}{P(Y = 0)} = 67.51 - 18.68 \cdot \text{ko00550}$$

- Wzrost liczebności sekwencji reprezentującej ko00550 o 1 powoduje spadek

$$\ln \frac{P(Y=1)}{P(Y=0)} \quad 18.60$$

- Wzrost liczebności sekwencji reprezentującej ko00550 o 1 powoduje spadek

$$P(Y = 1) \approx \frac{e^{-18.68}}{1 + e^{-18.68}} = 7.715766e-09$$

Interpretacja

$$\text{logit}(P) = \ln \frac{P(Y = 1)}{P(Y = 0)} = 67.51 - 18.68 \cdot \text{ko00550}$$

- Do którego stawu, 0 czy 1 będzie należała ryba, u której wykryto 3.12 odczytów reprezentujących ko00550 ?
- $P(Y = 1) = \frac{e^{67.51 - 18.68 \cdot 3.12}}{1 + e^{67.51 - 18.68 \cdot 3.12}} = 0.9999018$
- Klasyfikacja = staw 1

Równanie regresji logistycznej

```
# 2 class
```

```
yc2=c(0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1)
```

```
fit3 <- glm(yc2~x, family = binomial)
```

```
# probabilities of class 1
```

```
p3 <- predict(fit3, type = "response")
```

Równanie regresji logistycznej

```
library(nnet)
```

```
# 5 class
```

```
yc5=c(0,0,1,1,2,2,3,3,4,4,5,5,0,1,2,3,4,5,0)
```

```
fit4 <- multinom(as.factor(yc5)~x)
```

```
# probabilities of classes
```

```
p4probs <- predict(fit4, type = "probs")
```

1. Predykcja → Regresja liniowa
 - Równanie regresji liniowej
 - Estymacja współczynników równania
 - Interpretacja wyników w kontekście predykcji
2. Klasyfikacja → Regresja logistyczna
 - Równanie regresji logistycznej
 - Estymacja współczynników równania
 - Interpretacja wyników w kontekście klasyfikacji

1. Kiedy np. NIE można zastosować regresji liniowej ?
2. Jak dokonać predykcji wartości zmiennej zależnej w regresji ?
3. Kiedy NIE można zastosować regresji logistycznej ?
4. Jak uzyskać prawdopodobieństwo przynależności do klasy 1 ?