

Statystyka i modelowanie w naukach o środowisku

Wykład 10

Analiza regresji prostej

Regresja prosta liniowa

Regresja prosta jest metodą statystyczną, w której określamy zależność jednej zmiennej (Y) od drugiej (X), czyli zależność ta jest **między tylko dwiema zmiennymi**.

Regresja prosta liniowa

Regresja liniowa to metoda szacowania wartości oczekiwanej jednej zmiennej (Y) znając wartości innej zmiennej (X) na podstawie funkcji liniowej. Szukana zmienna Y jest nazywana zmienną zależną, zmienna X nazywa się zmienną niezależną.

Model regresji prostej liniowej

$$Y=a+bX+e_i$$

gdzie:

b – współczynnik regresji

a – stała regresji

e_i – błędy losowe o rozkładzie $N(0;\sigma_e^2)$

Stać regresji (a) jest zatem szacowaną średnią wartością zmiennej Y w przypadku gdy $X=0$, natomiast wartość współczynnika regresji (b) oznacza średnią zmianę wartości Y w przypadku gdy X zwiększymy o jedną jednostkę.

Ujemna wartość współczynnika regresji (b) świadczy o ujemnej zależności, a dodatnia wartość wskazuje na dodatnią zależność

Model regresji prostej liniowej

Estymację (szacowanie wartości) współczynników równania regresji prowadzi się zwykle metodą najmniejszych kwadratów, która polega na minimalizacji następującej sumy kwadratów:

$$\sum_{i=1}^n (y_i - a - bx_i)^2$$

Estymatory wartości współczynników a i b oblicza się ze wzorów:

$$b = \frac{s_{xy}}{s_x^2} \qquad a = \bar{y} - b\bar{x}$$

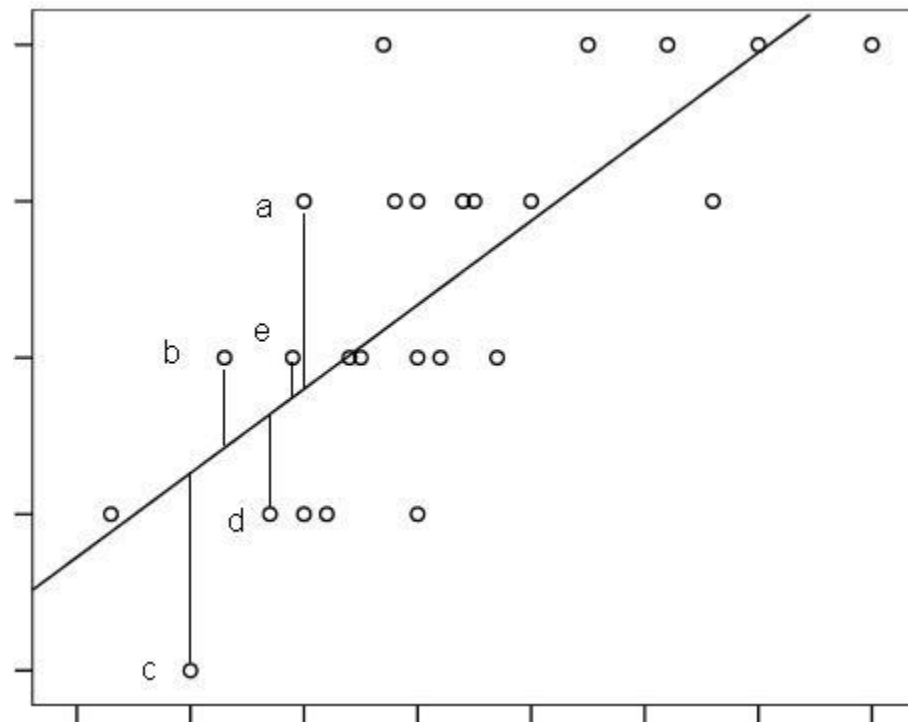
Analiza regresji prostej

Estymatory uzyskane metodą najmniejszych kwadratów

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

Metoda najmniejszych kwadratów



R^2 – współczynnik determinacji

Określa stosunek zmienności wyjaśnianej przez model regresji do zmienności całkowitej. W przypadku regresji prostej liniowej $R^2 = r_{xy}^2$

Czym wartość R^2 jest bliższa 100% (czyli 1) to zależność Y od X jest silniejsza, i na odwrót gdy wartość R^2 jest bliższa 0% (czyli 0) to zależność Y od X jest słabsza. Wartość współczynnika determinacji jest równa w przypadku regresji prostej liniowej kwadratowi współczynnika korelacji prostej Pearsona.

Testowanie hipotezy $H_0: \beta=0$ (współczynnik regresji dla całej populacji jest równy 0) pozwala na ocenę, czy występuje istotna zależność Y od X . Jeśli tę hipotezę odrzucimy to uznajemy, że Y istotnie zależy od X (powyższą hipotezę odrzucamy jeśli $p < \alpha$)

Analiza regresji prostej przykład

Przeprowadzono doświadczenie polowe, w którym oceniano zawartość białka w ziarnie pewnej odmiany pszenicy ozimej w zależności od stosowanej dawki nawożenia azotem. Zbadaj z użyciem regresji prostej liniowej wpływ nawożenia azotem na zawartość białka.

nawożenie azotem (kg/ha)	zawartość białka (%)
0	11,49
10	11,55
20	11,74
30	11,88
40	11,64
50	11,62
60	11,55
70	11,64
80	12,02
90	12,08
100	12,10
110	12,27
120	12,24
130	12,26
140	12,23
150	12,10
160	12,29
170	12,44
180	12,72
190	12,45
200	12,56
210	12,54
220	12,73
230	12,82
240	12,90
250	12,86

Analiza regresji prostej - przykład

STATISTICA

Statystyka → Regresja

wieloraka → podsumowanie
wyniki regresji

The screenshot shows the STATISTICA software interface. On the left is a data table with columns: nawożenie azotem, zawartość białka (%), Zmn3, Zmn4, and Zmn5. The data rows show values for nitrogen fertilization (0 to 250 kg/ha) and protein content (11.49 to 12.86%). Overlaid on the right are two dialog boxes. The top one is 'Regresja wieloraka: Arkusz1' with 'Zmienne' set to 'Zależna: brak, Niezależne: brak'. The bottom one is 'Wybierz listy zmiennych zależnych i niezależnych:' with 'Zmienne zależne' set to '2' and 'Lista zmiennych niezależnych' set to '1'. A checkbox 'Pokaż tylko zmienne o odpowiedniej skali' is checked.

Podsumowanie regresji zmiennej zależnej: zawartość białka (%) (Arkusz1)						
R= .95542186 R ² = .91283092 Popraw. R ² = .90919888						
F(1,24)=251.33 p<.00000 Błąd std. estymacji: .13382						
N=26	b*	Bł. std. z b*	b	Bł. std. z b	t(24)	p
W. wolny			11.48812	0.051009	225.2182	0.000000
nawożenie azotem (kg/ha)	0.955422	0.060266	0.00555	0.000350	15.8533	0.000000

$$t=15,85, p<\alpha$$

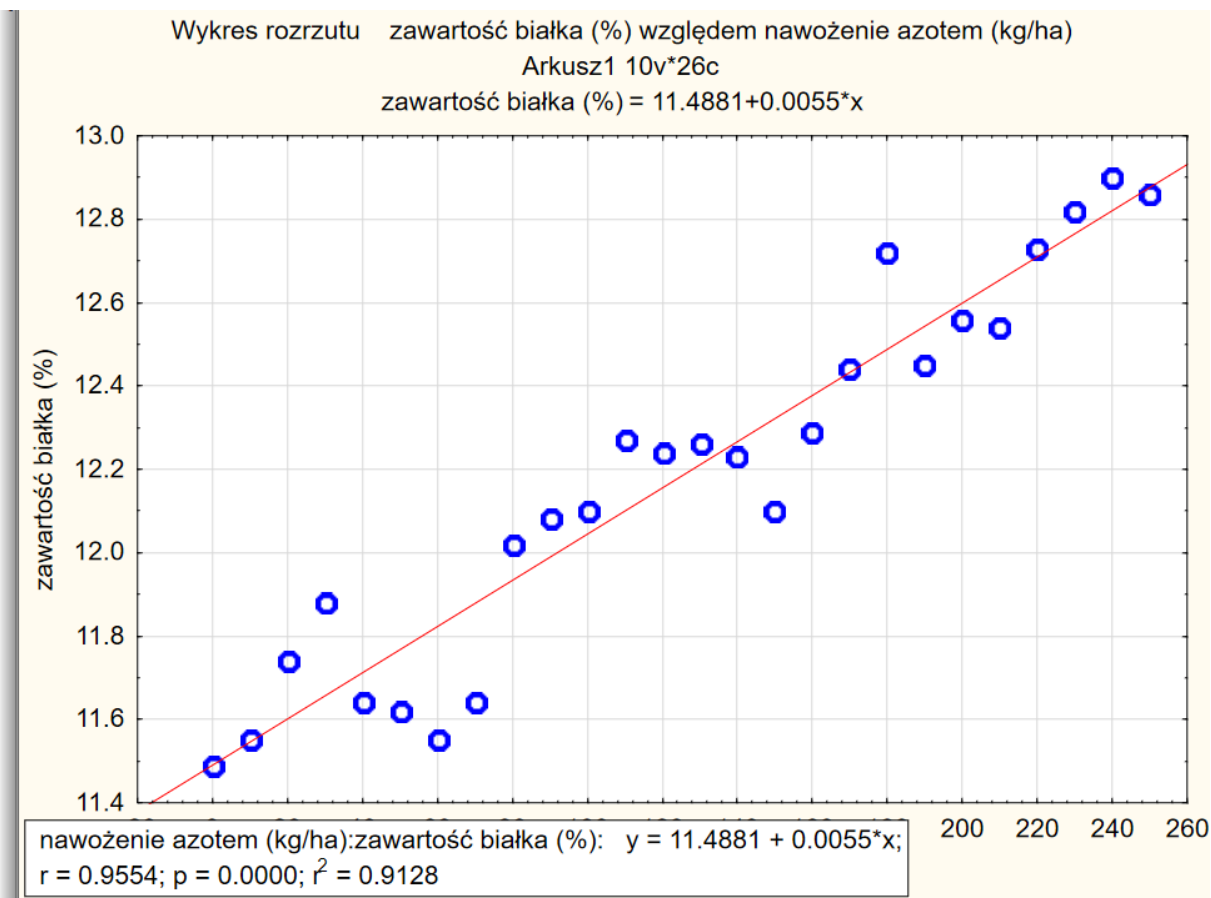
H_0 o braku wpływu zmiennej X na zmienną Y odrzucamy

Analiza regresji prostej - przykład

STATISTICA

Wykresy → Wykres

rozrzutu → zakładka „więcej”



Wykres rozrzutu 2W

Podstawowe Więcej Wygląd Skategoryzowane Opcje 1 Opcje 2

Zmienne: X: 1 Y: 2

Rodzaj wykresu:

- ☒ Zwykły
- ☐ Wielokrotny
- ☐ Podwójny Y
- ☐ Liczność
- ☐ Bąble
- ☐ Kwantyl
- ☐ Voronoia

Dopasuj

- ☐ Wylączone
- ☒ Liniowa
- ☐ Wielomian
- ☐ Logarytmiczna
- ☐ Wykładniczy
- ☐ NKWO
- ☐ NKWW
- ☐ Sklejana
- ☐ Lowess
- ☐ Ortogonalny

Definicja funkcji użytkownika

Statystyki

- ☒ R kwadrat (dopasow. lin.)
- ☒ Wsp. korel. i p (dop. lin.)
- ☒ Równanie regresji

Elipsa

- ☒ Nie
- ☐ Normalne Współcz.: .95
- ☐ Rozstęp

Pas regresji

- ☒ Nie
- ☐ Ufność Poziom: .95
- ☐ Prognoza

Zaznacz wybrane podzbiory: Nie

OK Anuluj Opcje Grupami Warunki selekcji Wagi przypadków Galeria wykresów Aktualizuj: Auto

$$Y = 0,0055 * X + 11,4881$$

EXCEL

Wykres XY → Dodaj linię trendu

Regresja prosta nieliniowa

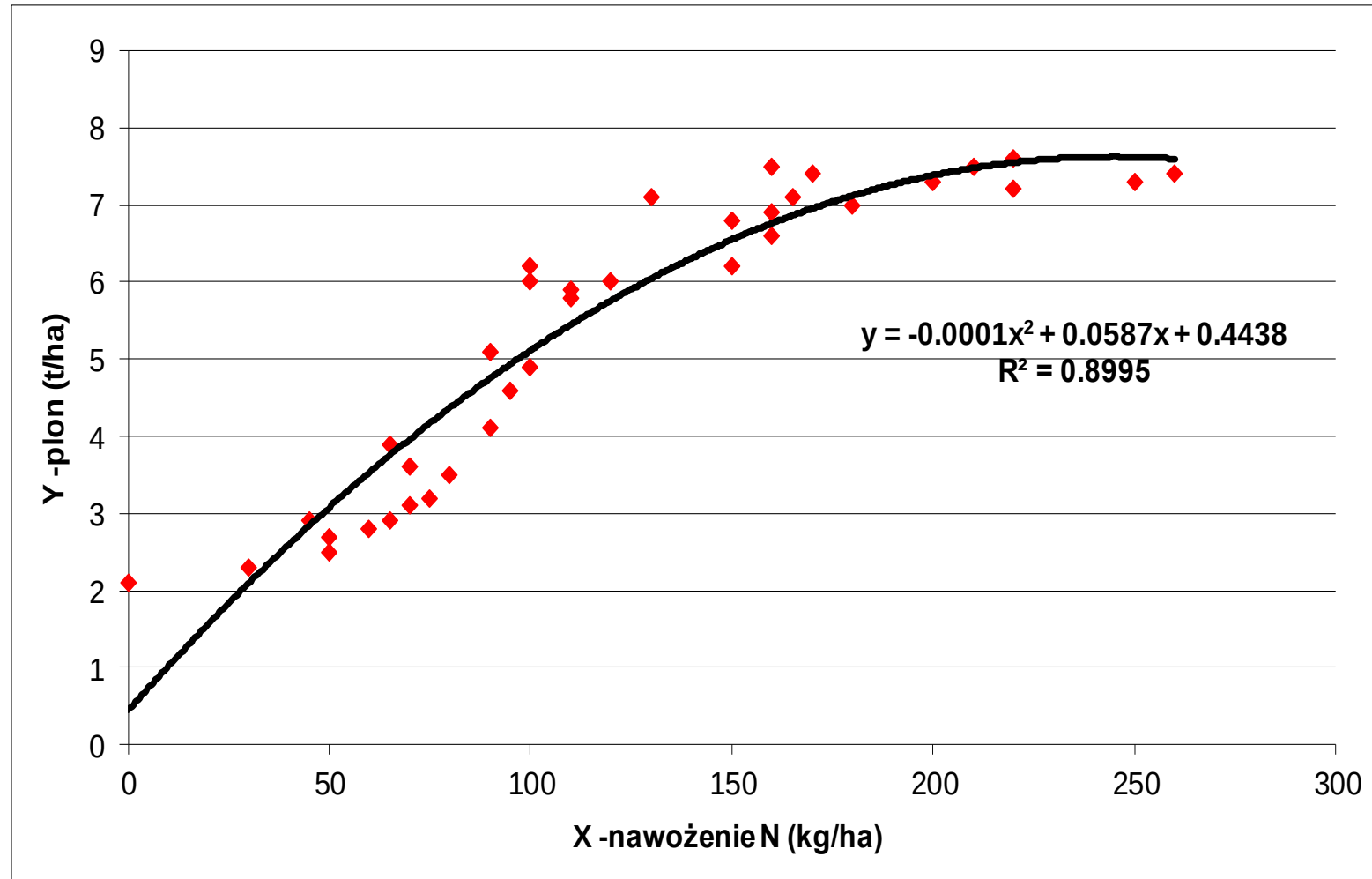
Nie wszystkie zależności między dwiema zmiennymi są liniowe, dlatego też czasami uzasadnione jest stosowanie innego niż liniowy modelu regresji. Stosowane są w tym celu różne inne modele regresji. Zamiast funkcji liniowej można użyć funkcji np.:

- kwadratowej
- pierwiastkowej
- logarytmicznej lub innych.

Dobór modelu regresji dokonuje się najczęściej na podstawie wartości współczynnika determinacji (R^2), większa wartość R^2 oznacza lepiej dopasowany model regresji, a tym samym lepiej opisujący zmiany Y w zależności od X .

Szczególnym przykładem regresji prostej jest regresja prosta wielomianowa, czyli wykorzystanie funkcji wielomianowej, w której zmienna niezależna (X) występuje w kolejnych potęgach. Najprostszym modelem regresji wielomianowej jest funkcja kwadratowa (X występuje w pierwszej i drugiej potęgach).

Regresja prosta nieliniowa



Regresja wielokrotna

Regresja wielokrotna jest metodą statystyczną, w której oceniamy wpływ wielu zmiennych niezależnych ($X_1, X_2, X_3, \dots$) na zmienną zależną (Y).

Najprostszym modelem regresji wielokrotnej, a jednocześnie najczęściej stosowanym w praktyce jest regresja wielokrotna liniowa.

Regresja wielokrotna liniowa

Jeżeli zmienna zależna (Y) jest zależna od więcej niż jednej zmiennej niezależnej (X_i) to estymowany model regresji możemy zapisać równaniem:

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_k \cdot X_k + e_i$$

a - stała regresji,

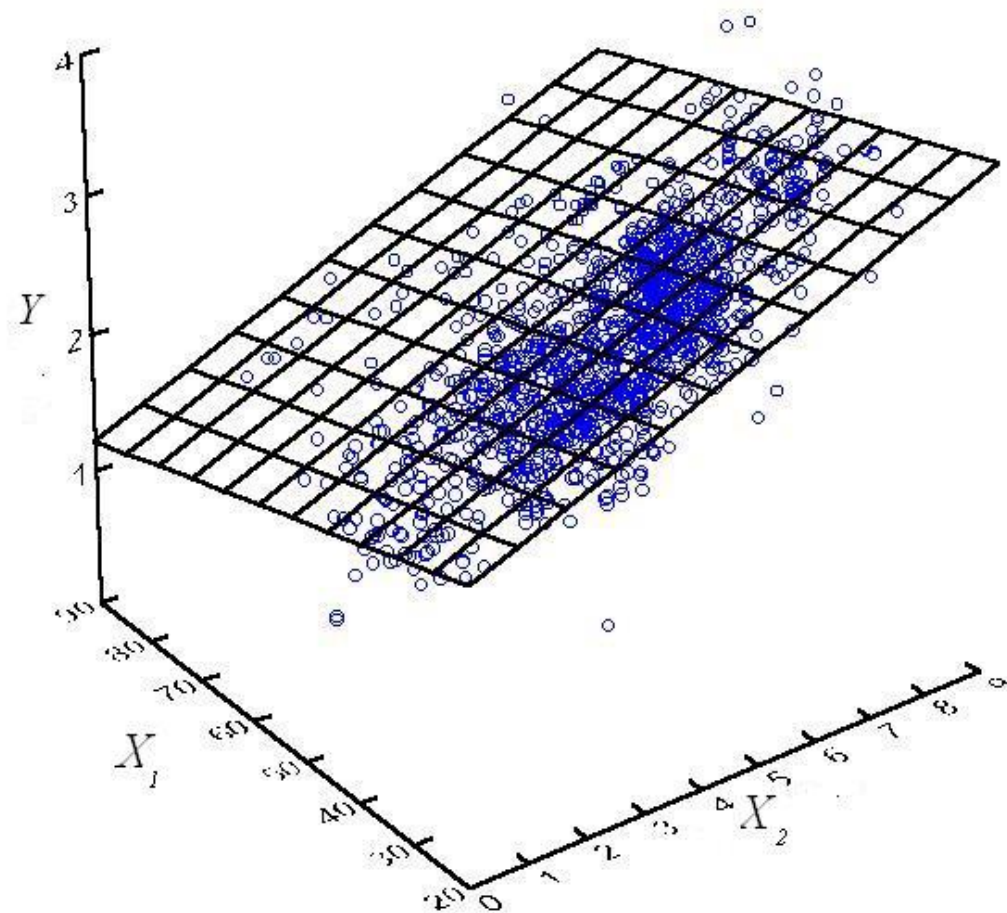
$b_1, b_2, \dots$ – cząstkowe współczynniki regresji

e_i – błąd losowy

Interpretacja wartości stałej regresji i cząstkowych współczynników regresji jest podobna jak w przypadku regresji prostej. Stała regresji jest to szacowana średnia wartość Y , gdy wszystkie zmienne niezależne (X_i) są równe 0. Wartość każdego cząstkowego współczynnika regresji oznacza szacowaną średnią zmianę wartości Y , gdy dana wartość zmiennej niezależnej (X_i) zwiększy się o jedną jednostkę.

W przypadku regresji wielokrotnej zastosowanie metody najmniejszych kwadratów to minimalizowanie sumy:

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



Graficzne przedstawienie regresji z 2 zmiennymi niezależnymi (X_1 , X_2)

Regresja wielokrotna liniowa

Hipoteza zerowa globalna:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad (\text{żadna z zmiennych niezależnych nie wpływa na zmienną zależną})$$

Weryfikację tej hipotezy wykonujemy na podstawie testu F w analizie wariancji

Jeśli odrzucimy hipotezę H_0 to stwierdzamy, że co najmniej jeden współczynnik regresji jest różny od zera, czyli istnieje związek liniowy między co najmniej jedną zmienną niezależną a zmienną zależną.

Jeśli odrzucimy globalną H_0 , to następnie powinniśmy przeprowadzić weryfikację k hipotez szczegółowych, które pozwolą odpowiedzieć na pytanie, które zmienne niezależne (czy wszystkie, czy też tylko wybrane) wpływają na zmienną zależną, a więc które zmienne niezależne powinny pozostać w modelu regresji.

Regresja wielokrotna liniowa

Hipotezy szczegółowe:

$$H_0 : \beta_i = 0$$

Weryfikację hipotez szczegółowych wykonujemy na podstawie testu t

Jeśli odrzucimy i -tą hipotezę zerową, to stwierdzamy, że dana zmienna niezależna (przyczynowa) ma istotny statystycznie wpływ na zmienną zależną

Zmienne niezależne, które nie wpływają istotnie na zmienną zależną powinny być usunięte z modelu regresji. Przy usuwaniu zmiennych z modelu należy uwzględniać ewentualne wzajemne współzależności między poszczególnymi zmiennymi niezależnymi.

Regresja wielokrotna liniowa

Nie wszystkie zmienne niezależne (X_i) które bierzemy do analizy regresji wielokrotnej mają wpływ na zmienną zależną (Y), a więc uzasadnione jest usunięcie tych zmiennych i pozostawienie tylko tych zmiennych niezależnych, które mają istotny wpływ.

W tym celu stosuje się różne metody pozwalające na usunięcie z modelu regresji nieistotnie wpływających zmiennych niezależnych i pozostawienie tylko tych, których wpływ udowodnimy. Jedną z metod, które są dość często stosowane jest **regresja krokowa**, która pozwala na dobór modelu z pominięciem zmiennych słabo lub nie wpływających na zmienną zależną.

Dobierając model funkcji regresji powinniśmy dążyć do uzyskania jak największego współczynnika determinacji

Regresja wielokrotna nieliniowa

Przykład modelu nieliniowego z dwoma zmiennymi niezależnymi:

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_2 + b_4x_2^2 + b_5x_1x_2$$

Możliwe zalety modeli nieliniowych regresji:

- uwzględnienie zależności nieliniowych
- uwzględnienie współdziałań zmiennych niezależnych

Regresja wielokrotna

Trudności w doborze najlepszego modelu regresji wielokrotnej

- Dla różnych modeli regresji wielokrotnej można uzyskać podobne parametry dopasowania (np. bardzo zbliżone wartości współczynnika determinacji)
- Wskazanie, jaka funkcja nieliniowa dla danej zmiennej jest optymalna nie jest łatwe ze względu na występowanie współdziałań i współzależności między cechami
- Ograniczone możliwości graficznego przedstawienia funkcji, a tym samym ocena tych funkcji odbywa się zazwyczaj na podstawie estymowanych parametrów.
- Dysponowanie ograniczonym zbiorem zmiennych niezależnych, który może nie zawierać zmiennych mających silny związek z zmienną zależną

Regresja wielokrotna

Ze względu na to, że związki między zmienną zależną (Y) a zmienną niezależną (X) w szerokim zakresie zmienności są często nieliniowe, to w analizie regresji prostej do adekwatnego opisu takich zależności wykorzystywane są nieliniowe funkcje regresji.

Najczęściej w tym celu wykorzystuje się następujące modele regresji prostej:

- linearyzowalne
- wielomianowe
- logistyczne

Regresja wielokrotna

Najczęściej wykorzystywane linearyzowalne funkcje regresji prostej:

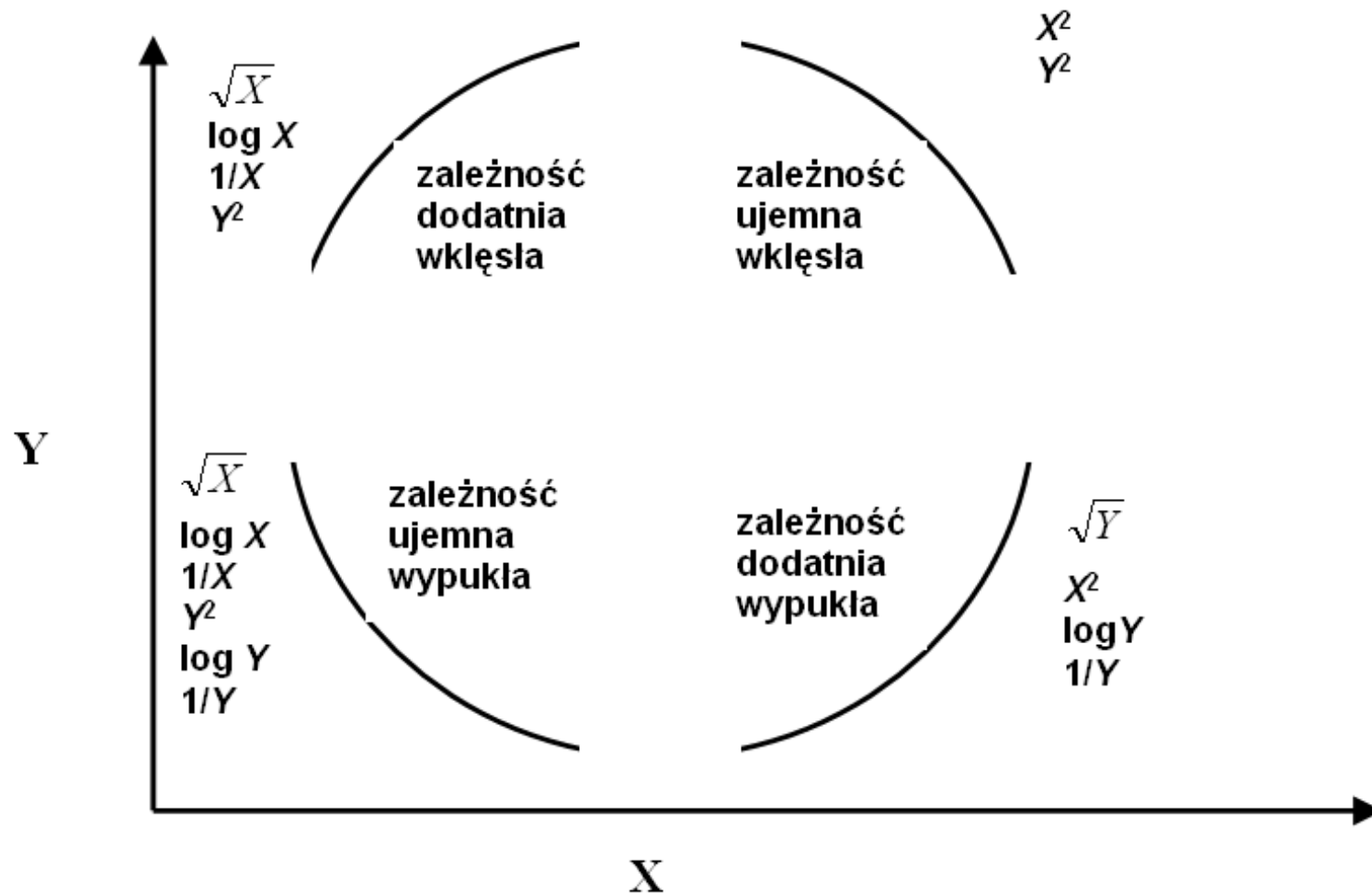
$Y=e^{(a+bx)}$, gdzie e jest podstawą logarytmu naturalnego $e \approx 2,718..$

$$Y=a+b*\log(X)$$

$$Y=a+b*X^2$$

$$Y=a+b*X^{(1/2)}$$

Przykłady typowych transformacji zmiennych stosowane w funkcjach nieliniowych



Interpretacja wyników analiz regresji z użyciem funkcji nieliniowych

W odróżnieniu od funkcji liniowych regresji parametry funkcji nieliniowych interpretuje się inaczej.

Interpretacja ta zależy od stosowanej funkcji regresji. Główną różnicą jest to, że współczynnik regresji (b) nie jest przeciętną zmianą wartości Y przy wzroście X o jedną jednostkę. Również dla wielu funkcji nieliniowych stała regresji (a) nie jest szacowaną wartością Y dla $X=0$.

Aby oszacować ile przeciętnie zmieni się Y , jeśli X wzrośnie o 1 należy obliczyć wartość pochodnej funkcji regresji. Oczywiście wartość ta będzie różna dla różnych wartości X .

Testowanie hipotezy H_0 : X nie wpływa na zmienność Y odbywa się podobnie jak w przypadku regresji prostej liniowej testem F .

Analiza regresji prostej nieliniowej – przykład

dane do analiz
liczba bakterii w 1 ml mleka
po 8 godz. przechowywania
w różnych temperaturach

temperatura przechowywania (°C)	liczba bakterii po 8 godz. przechowywania (mln)
4	0,10
4	0,14
4	0,09
8	0,23
8	0,26
8	0,29
12	0,48
12	0,49
12	0,58
16	0,67
16	0,72
16	0,59
20	1,4
20	1,7
20	1,3
25	2,8
25	3,1
25	2,5
30	8,7
30	9,5
30	10,2

Analiza regresji prostej nieliniowej – przykład

Dobór funkcji regresji na podstawie wartości współczynnika determinacji

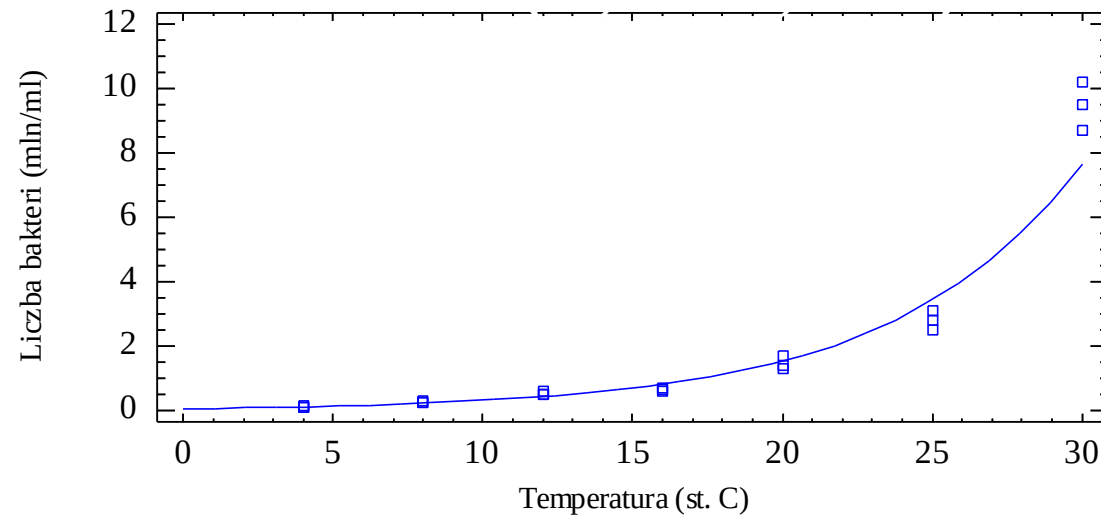
Comparison of Alternative Models

Model	Correlation	R-Squared
Exponential	0,9896	97,93%
Double reciprocal	0,9767	95,39%
Multiplicative	0,9477	89,81%
Square root-Y	0,9201	84,65%
S-curve	-0,8413	70,77%
Linear	0,8124	66,00%
Square root-X	0,7444	55,41%
Logarithmic-X	0,6629	43,94%
Reciprocal-X	-0,4957	24,57%
Reciprocal-Y	<no fit>	
Logistic	<no fit>	
Log probit	<no fit>	

Analiza regresji prostej nieliniowej – przykład

Oszacowanie funkcji regresji wykładniczej

Równanie regresji: $Y=e^{(-2,78 + 0,16 \cdot X)}$



Regresja wielomianowa

Analiza regresji wielomianowej pozwala na ocenę związków nieliniowych, które można opisywać z użyciem funkcji wielomianowych o ogólnej postaci:

$$\hat{Y} = b_0 + b_1X + b_2X^2 + \dots + b_kX^k$$

gdzie:

b_0 – stała regresji

$b_1, b_2, \dots$ – cząstkowe współczynniki regresji

Regresja wielomianowa – przykład

Dane doświadczalne:

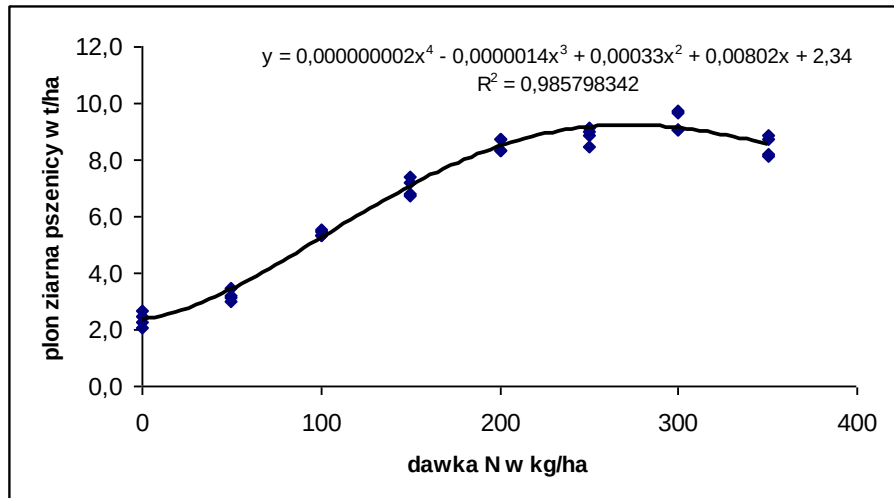
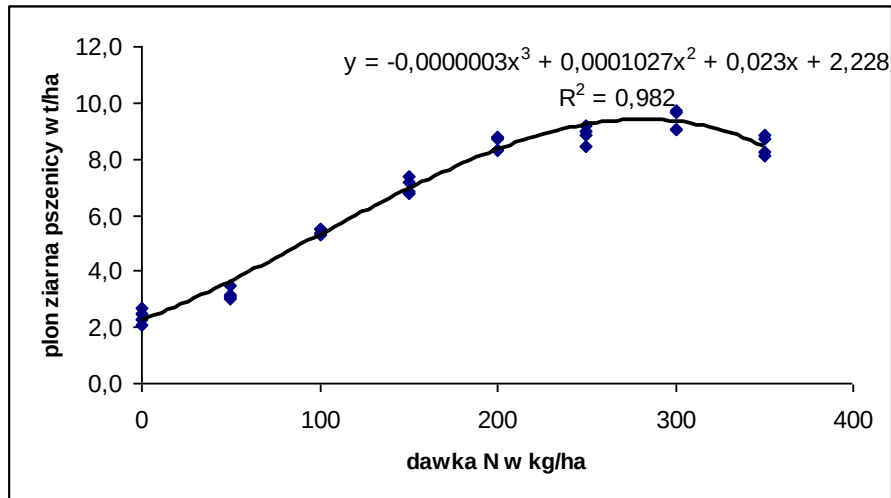
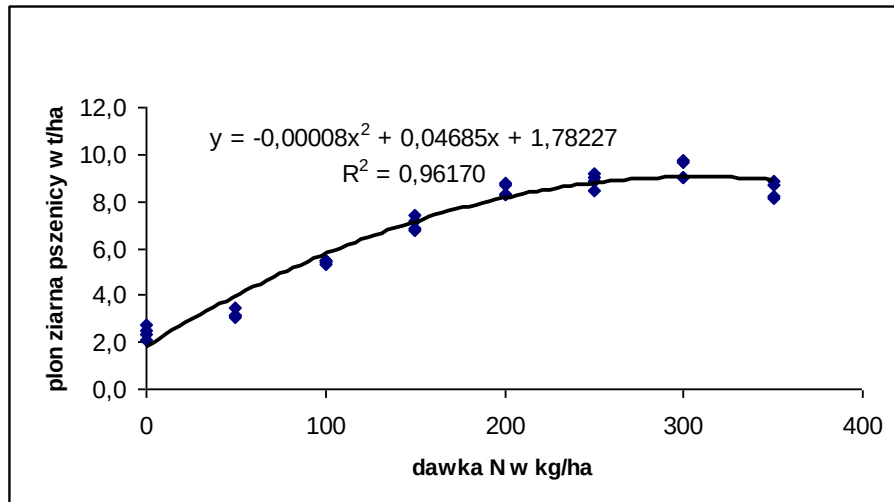
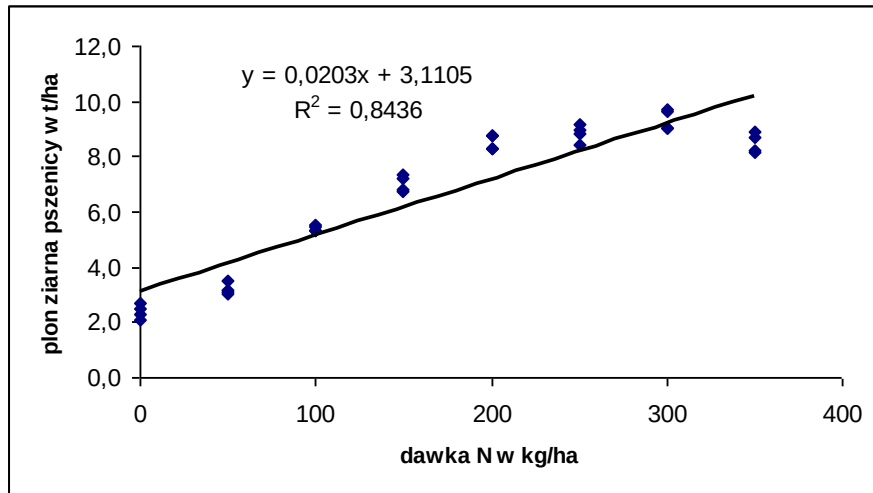
-plon ziarna (t/ha) jednej z odmian pszenicy ozimej badany dla dawek N w zakresie od 0 do 350 kg/ha

Cel analiz: określenie wpływu dawki nawożenia N na plonowanie pszenicy ozimej (m.in. określenie dawki N dla której uzyskujemy plon maksymalny)

nawożenie N	plon
0	2,1
0	2,3
0	2,7
0	2,5
50	3,4
50	3,1
50	3,2
50	3,3
100	5,2
100	5,5
100	5,1
100	5,5
150	6,7
150	7,4
150	7,2
150	7,4

nawożenie N	plon
200	8,0
200	8,8
200	8,0
200	8,1
250	9,3
250	9,5
250	8,7
250	8,9
300	9,2
300	9,2
300	9,3
300	9,3
350	8,5
350	8,2
350	9,4
350	9,4

Regresja wielomianowa - przykład



Regresja wielomianowa – przykład

testowanie hipotez i interpretacja wyników

```
Polynomial Regression Analysis
-----
Dependent variable: Y
-----
```

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	2,22877	0,168339	13,2397	0,0000
X	0,0234644	0,00448819	5,22802	0,0000
X^2	0,000102674	0,0000311703	3,29396	0,0027
X^3	-3,40146E-7	5,84423E-8	-5,82021	0,0000

```
-----
Analysis of Variance
-----
```

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	201,287	3	67,0958	529,14	0,0000
Residual	3,55042	28	0,126801		
Total (Corr.)	204,838	31			

```
-----
R-squared = 98,2667 percent
R-squared (adjusted for d.f.) = 98,081 percent
Standard Error of Est. = 0,356091
Mean absolute error = 0,281837
Durbin-Watson statistic = 1,35709
```


Regresja logistyczna

Analiza regresji logistycznej jest szczególnym rodzajem regresji ze względu na to, że zmienna zależna jest zmienną zero-jedynkową.

Analizę regresji logistycznej stosujemy zwykle do prognozowania wystąpienia pewnych zjawisk (zwykle przyjmujemy: 0- zjawisko nie wystąpiło, 1-zjawisko wystąpiło).

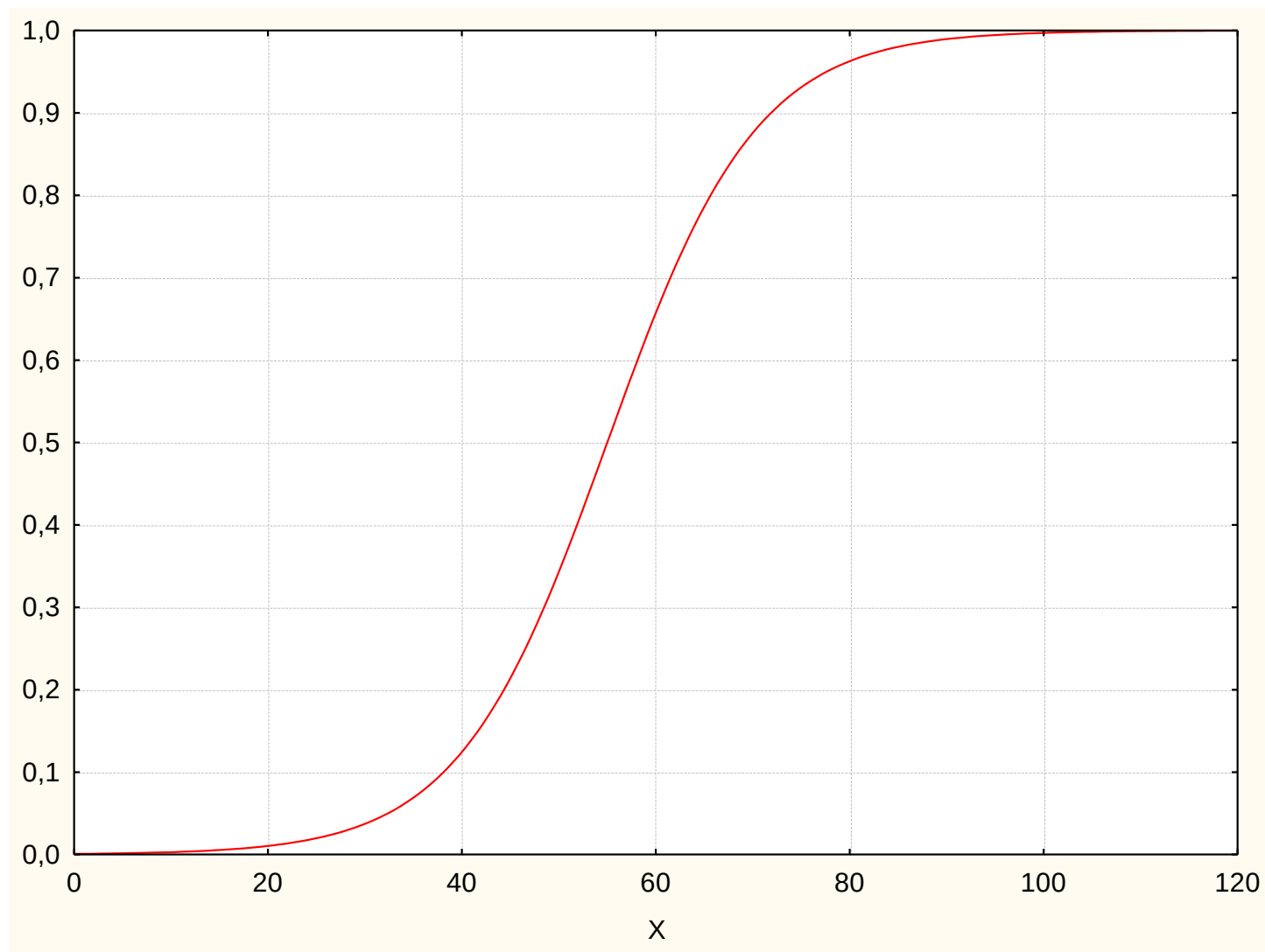
Regresja logistyczna - model

Model regresji logistycznej jest następujący:

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}} = \frac{1}{1 + e^{-(a+bX)}}$$

Na podstawie powyższej funkcji logistycznej szacujemy prawdopodobieństwo (P) wystąpienia pewnego zjawiska (czyli dla zmiennej zero-jedynkowej jest to prawdopodobieństwo, tego, że $Y=1$)

Regresja logistyczna – graficzne przedstawienie funkcji



Regresja logistyczna – testowanie wpływu zmiennej

Istotność wpływu zmiennej niezależnej (X) na zmienną zależną (Y) oceniana jest na podstawie testu chi-kwadrat.

Jeżeli: $\chi^2_{emp} > \chi^2_k$ gdzie k - liczba zmiennych
niezależnych

To odrzucamy hipotezę zerową (H_0 : zmienna X nie wpływa istotnie na zmienną Y),
czyli stwierdzamy statystycznie istotny wpływ X na Y.

Regresja logistyczna

- przykład

Dane pochodzą z doświadczenia w którym oceniano wpływ przechowywania pędów jabłoni w niskich temperaturach (w zakresie od -60 do -20 stopni C).

W każdej temperaturze umieszczono po 10 pędów a następnie po 24 godzinach określono ile z nich zamarło z powodu niskiej temperatury.

W tabeli obok znajdują się dane dotyczące każdego z pędów (1- pęd zmarł, 0- pęd przeżył).

Celem badań jest ustalenie poziomu temperatury, przy której zamiera 50% pędów.

Temperatura	Zamieranie	Temperatura	Zamieranie
-20	0	-40	0
-20	0	-40	1
-20	0	-40	1
-20	0	-40	1
-20	0	-40	0
-20	0	-50	1
-20	0	-50	1
-20	0	-50	1
-20	0	-50	1
-20	0	-50	1
-20	0	-50	1
-30	1	-50	1
-30	0	-50	1
-30	0	-50	0
-30	0	-50	0
-30	0	-50	1
-30	1	-60	1
-30	0	-60	1
-30	0	-60	1
-30	0	-60	1
-30	1	-60	1
-40	1	-60	1
-40	0	-60	1
-40	1	-60	1
-40	1	-60	1
-40	1	-60	1

Regresja logistyczna - przykład

Obliczona funkcja regresji logistycznej

$$y = e^{(-6,1235 + (-0,1659) \cdot X)} / (1 + e^{(-6,1235 + (-0,1659) \cdot X)})$$

Testowanie hipotezy o istotności wpływu X na Y

H_0 : zmienna X nie wpływa na zmienną Y

$$\chi^2_{emp} = 31,28; P < 0,001$$

Na podstawie testu chi-kwadrat odrzucamy hipotezę zerową, czyli stwierdzamy, że X wpływa statystycznie istotnie na Y (temperatura wpływa na zamieranie pędów)

Regresja logistyczna – przykład

