Regresja prosta liniowa

Regresja liniowa to metoda estymowania wartości oczekiwanej jednej zmiennej (*Y*) znając wartości innej zmiennej (*X*). 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 $N(0;\sigma_e^2)$

Estymację 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:

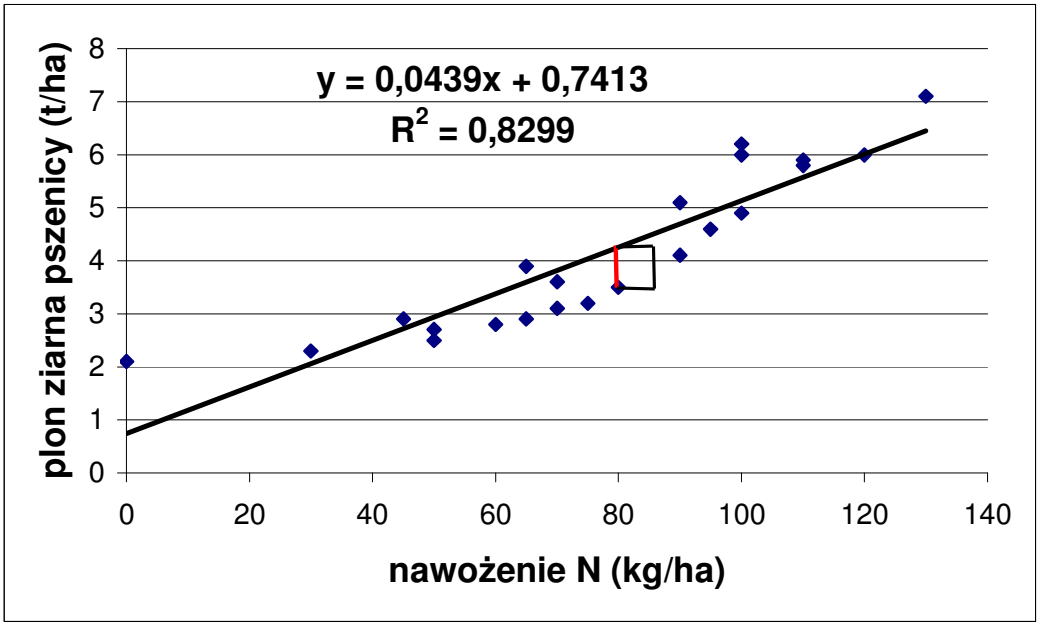
$$b = \frac{\mathrm{cov}(X,Y)}{s_x^2} \qquad a = \overline{y} - b\overline{x}$$

Przedział ufności dla współczynnika regresji:

$$(b - t_{\alpha;n-2} \cdot S_b;\ b + t_{\alpha;n-2} \cdot S_b)$$

gdzie wariancja estymatora b $S_b = \frac{S^2}{\text{var } X}$

Testowanie hipotezy H_0: b=0 jest równoważne z testowaniem hipotezy o istotności korelacji



***R^2* – współczynnik determinacji** , który określa stosunek zmienności wyjaśnianej przez model regresji do zmienności całkowitej. W przypadku regresji prostej liniowej $R^2=r_{xy}^2$

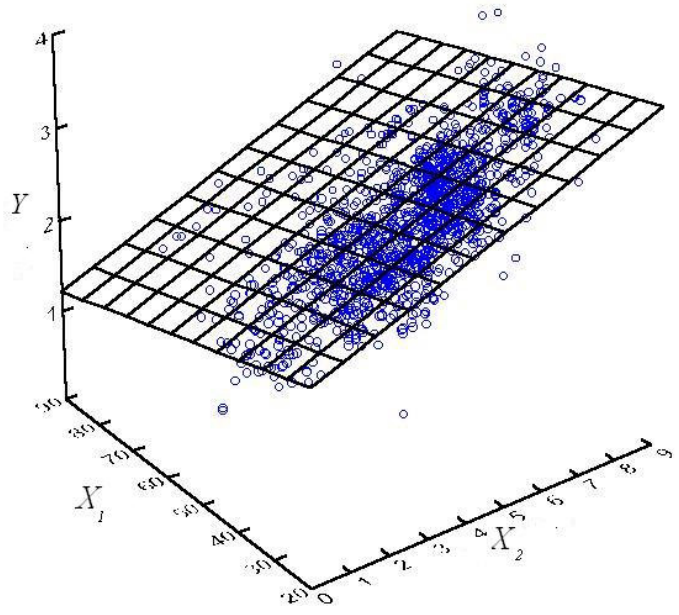
Regresja wielokrotna liniowa

Jeżeli zmienna zależna (Y) jest determinowana przez więcej niż jedną zmienną niezależną (X_i) to estymowany model regresji możemy zapisać równaniem:

$$\boldsymbol{Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_k \cdot X_k}$$

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

$$\sum_{i=1}^{n} (y_i - a - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2$$



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

Test niezależności cech jakościowych - Test χ^2

Rozważając liczbę obserwacji sklasyfikowanych wg dwóch kryteriów, np. ludzi wg koloru oczu i koloru włosów (kolory oczu: brązowy, niebieski; kolory włosów: blondyni, szatyni, bruneci) lub np. rośliny pszenicy wg odmiany i stopnia porażenia chorobą (odmiany: Olimpia, Eta, Kontesa; stopień porażenia: brak, słaby, średni, duży, bardzo duży) w każdej z klas liczymy liczbę osobników i przedstawiamy w postaci tablicy dwudzielnej zwanej tablica kontyngencji

Tablica kontyngencji

Klasy cechy *Y*	**Klasy cechy *X***						
	A_1	A_2	A_3	A_4	…	A_m	**razem**
B_1	n_{11}	n_{21}	n_{31}	n_{41}	…	n_{m1}	Σn_{i1}
B_2	n_{12}	n_{22}	n_{32}	n_{42}	…	n_{m2}	Σn_{i2}
B_3	n_{13}	n_{23}	n_{33}	n_{43}	…	n_{m3}	Σn_{i3}
…					…		
B_k	n_{1k}	n_{2k}	n_{3k}	n_{4k}	…	n_{mk}	
razem	Σn_{1j}	Σn_{2j}	Σn_{3j}	Σn_{4j}	…		Σn_{ij}

n- liczebności osobników zaliczonych do określonej klasy

H_o: Cechy X i Y są niezależne

Statystyka testowa

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}}$$

$$n^t_{ij} = \frac{n_{i.}n_{.j}}{N},\ N = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}$$

$$n_{i.} = \sum_{j=1}^{m} n_{ij},\ n_{.j} = \sum_{i=1}^{k} n_{ij}$$

Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; (k-1)(m-1))$, to hipotezę H_0 odrzucamy