8 Regresja

Głównym celem **analizy regresji** jest wyznaczenie funkcji opisującej (w przybliżeniu) zależność pomiędzy **zmienną niezależną** - objaśniającą (lub wieloma zmiennymi niezależnymi – objaśniającymi, a **zmienną zależną** - objaśnianą.

Przyjmujemy następujący model:

$$Y_i = f(x_{i1}, x_{i2}, \ldots, x_{im}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdzie

$f(x_1, x_2, \ldots, x_m) f(x_1, x_2, \ldots, x_m)$ - funkcja regresji,

$\varepsilon_i \varepsilon_i$ - błędy (reszty).

Założenia analizy regresji

1. Niezależność obserwacji dla poszczególnych jednostek eksperymentalnych.
2. Brak błędu systematycznego.
3. Jednakowa i stała wariancja błędów.
4. Brak korelacji błędów.

Uwaga: W procedurach testowych oraz w przypadku wykorzystywania przedziału predykcji, potrzebne jest dodatkowe założenie normalności błędów. Powoduje ono, że brak korelacji błędów oznacza ich niezależność.

Metody estymacji funkcji regresji:

1. Parametryczne - zakładamy znajomość postaci funkcji regresji z dokładnością do skończonej (zazwyczaj małej) liczby parametrów. W tym przypadku, do estymacji funkcji regresji używamy najczęściej metody **najmniejszych kwadratów** polegającej na minimalizacji wyrażenia:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_{i1}, x_{i2}, \ldots, x_{im})]^2.$$

2. Nieparametryczne - nie zakładamy żadnej konkretnej postaci funkcji regresji, a do jej estymacji wykorzystujemy np. metodę jądrową.

Regresja prosta liniowa

XX - zmienna niezależna (objaśniająca),

YY - zmienna zależna (objaśniana).

Model:

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdzie

a, ba, b - parametry liniowej funkcji regresji,

$\varepsilon_i \varepsilon_i$ - błędy (reszty).

FAKT

Estymatory parametrów aa i bb funkcji regresji uzyskane metodą najmniejszych kwadratów mają postać:

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

Liczbową miarą dopasowania prostej regresji do danych empirycznych jest **współczynnik determinacji** (podawany w %)

$$R^2 = 1 - \frac{SSE}{SST},$$

gdzie $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \ SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \ \hat{y}_i = \hat{a} + \hat{b}x_i.$
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \ SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \ \hat{y}_i = \hat{a} + \hat{b}x_i.$

Testy dla parametrów funkcji regresji

Hipoteza zerowa aa: wyraz wolny nie jest istotnie różny od zera (brak możliwości odrzucenia tej hipotezy skutkuje czasami przyjęciem modelu regresji bez wyrazu wolnego).

$$H_0: \ a = 0, \ H_1: \ a \neq 0.$$

Statystyka testowa:

$$t = \frac{\hat{a}}{S_a}, \ S_a = S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}}, \ S_e^2 = SSE/(n-2).$$

Rozkład statystyki testowej (przy założeniu normalności rozkładu błędów): $t|_{H_0} \sim t(n-2)$
$t|_{H_0} \sim t(n-2)$

Hipoteza zerowa b: współczynnik kierunkowy nie jest istotnie różny od zera, tzn. zmienna niezależna X nie ma istotnego wpływu na zmienną zależną Y.

$$H_0: b = 0,\ H_1: b \neq 0.$$

Statystyka testowa:

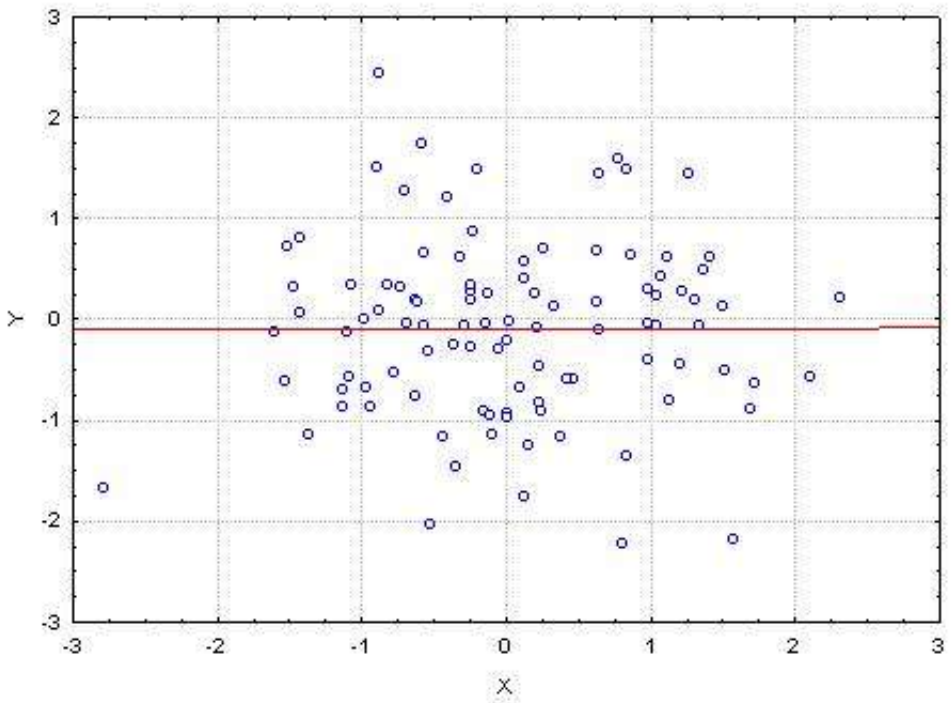
$$t = \frac{\hat{b}}{S_b},\ S_b = S_e\sqrt{\frac{1}{\sum_{k=1}^{n}(x_k - \bar{x})^2}}.$$

Rozkład statystyki testowej (przy założeniu normalności rozkładu błędów): $t\,|_{H_0} \sim t(n-2)$

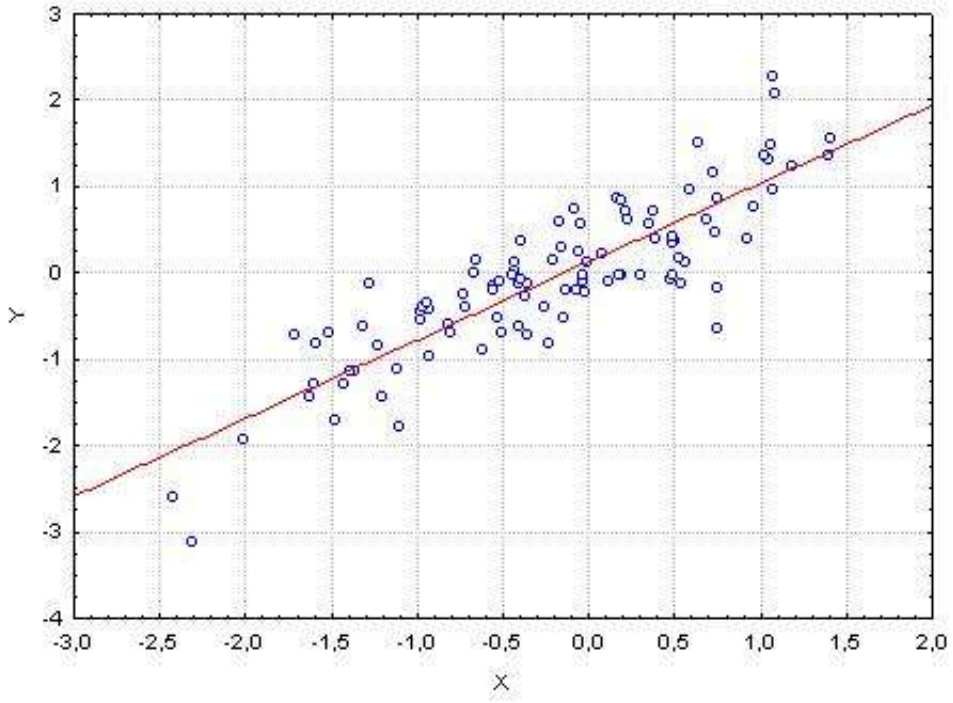
$t|_{H_0} \sim t(n-2)$

Wpływ zmiennej niezależnej X na zmienną zależną Y

1. Brak istotnego wpływu, $b = 0$ $b = 0$.



2. Istotny wpływ, $b \neq 0$ $b \neq 0$.



Prognozowanie (predykcja)

Niech $x_p x_p$ oznacza wartość zmiennej niezależnej XX dla której uzyskać chcemy prognozę zmiennej zależnej YY równą $y_p y_p$.

Przyjmujemy:

$$y_p = \hat{a} + \hat{b}x_p.$$

Regresja wielokrotna (wieloraka) liniowa

$X_1, X_2, \ldots, X_m X_1, X_2, \ldots, X_m$ - zmienne niezależne (objaśniające),

YY - zmienna zależna (objaśniana).

Model:

$$Y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_m x_{im} + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdzie

$a_0, a_1, \ldots, a_m a_0, a_1, \ldots, a_m$ - parametry liniowej funkcji regresji,

$\varepsilon_i \varepsilon_i$ - błędy (reszty).

Zapis macierzowy.

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}. \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Model liniowy:

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \varepsilon.\ \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon}.$

Dodatkowe założenia i estymatory parametrów

Dodatkowe założenia wynikające z używania $m > 1 m > 1$ zmiennych niezależnych:

1. Liczebność próby jest większa od liczby szacowanych parametrów, tzn. $n > m + 1$ $n > m + 1.$
2. Pomiędzy wektorami obserwacji zmiennych objaśniających nie istnieje zależność liniowa. Warunek ten oznacza, że

$$\text{rząd}(\boldsymbol{X}) = m + 1.$$

Estymatory parametrów funkcji regresji} uzyskane metodą najmniejszych kwadratów maja postać:

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Liczbową miarą dopasowania hiperpłaszczyzny regresji do danych empirycznych jest **współczynnik determinacji** (podawany w %)

$$R^2 = 1 - \frac{SSE}{SST},$$

gdzie

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

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{i1} + \cdots + \hat{a}_m x_{im}, \quad i = 1, \dots, n.$$

W przypadku wielu zmiennych niezależnych stosujemy **poprawiony współczynnik determinacji**

$$R^2_{pop} = 1 - \frac{SSE/(n-m-1)}{SST/(n-1)}.$$

Testowanie istotności parametrów modelu

Hipotezy: $H_0: a_j = 0,\ H_1: a_j \neq 0,\ j = 1, 2, \ldots, m.$

Statystyka testowa:

$$t_j = \frac{\hat{a}_j}{S_{a_j}},$$

gdzie

$$S^2_{a_j} = MSE \cdot d_{jj}$$

oraz d_{jj} jest j-tym elementem głównej przekątnej macierzy $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Rozkład statystyki testowej (przy założeniu normalności rozkładu błędów): $t_j | _{H_0} \sim t(n - m - 1).$

Prognozowanie (predykcja)

Niech X_p oznacza wektor wartości zmiennych objaśniających dla której uzyskać chcemy prognozę zmiennej objaśnianej y_p.

Dokładnie:

$$\boldsymbol{X}_p = \begin{bmatrix} 1 \\ x_1^p \\ \vdots \\ x_m^p \end{bmatrix}.$$

Przyjmujemy:

$$y_p = \boldsymbol{X}_p' \hat{\boldsymbol{a}}.$$

Regresja nieliniowa

Metody szacowania parametrów modelu:

1. linearyzacja - polega na przekształceniu modelu nieliniowego do modelu liniowego, poprzez transformację zmiennych niezależnych lub/i zmiennej zależnej. Przykładowo, model Cobba-Douglasa postaci

$$y = a_0 x_1^{a_1} x_2^{a_2},$$

można przekształcić do modelu liniowego poprzez transformację $y' = \ln(y)$, $x_1' = \ln(x_1)$, $x_2' = \ln(x_2)$ oraz $a_0' = \ln(a_0)$.

Wtedy

$$y' = a_0' + a_1 x_1' + a_2 x_2'.$$

2. numeryczne rozwiązanie zagadnienia minimalizacji sumy kwadratów błędów.

Regresja logistyczna

W regresji logistycznej badamy wpływ m niezależnych zmiennych $X_1, \ldots, X_m$ (ilościowych) na zależną zmienną Y mającą charakter dychotomiczny (zero-jedynkowy).

Model

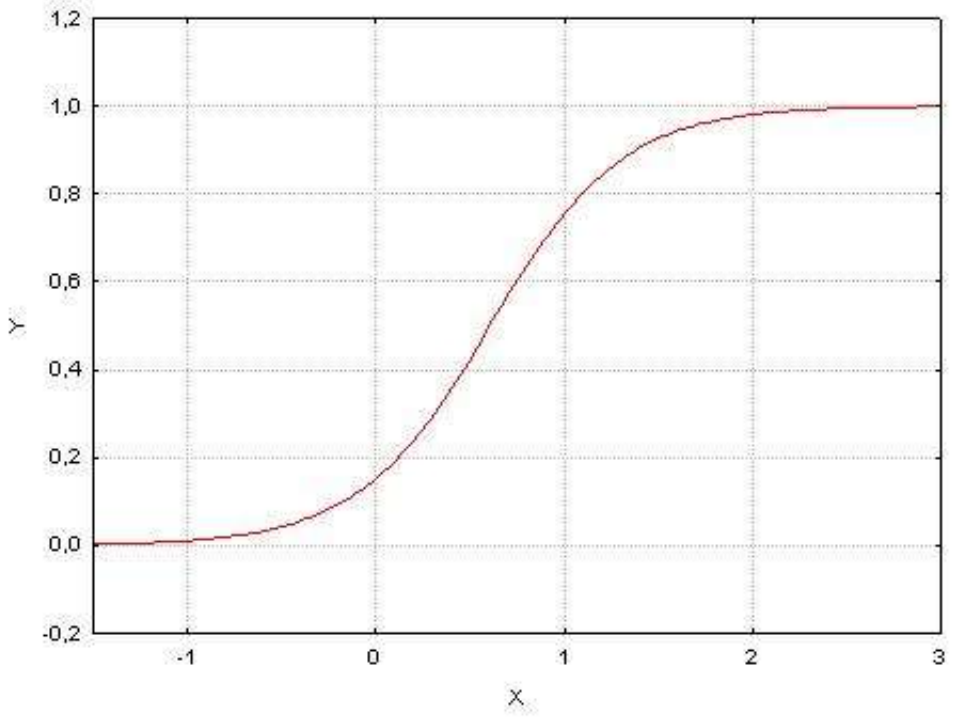
$$p = E(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(a_0 + a_1 x_1 + \cdots + a_m x_m)}{1 + \exp(a_0 + a_1 x_1 + \cdots + a_m x_m)},$$

gdzie

p - prawdopodobieństwo sukcesu,

$a_0, a_1, \ldots, a_m$ - współczynniki regresji.

Krzywa logistyczna



1. Współczynniki regresji $a_0, a_1, \ldots, a_m$ estymujemy metodą największej wiarogodności wykorzystując iteracyjny algorytm **IWLS** (algorytm iteracyjnie ważonych najmniejszych kwadratów).

2. Wielkość

$$\ln\frac{p}{1-p} = a_0 + a_1 x_1 + \cdots + a_m x_m$$

nazywamy **logitem**.

3. Wielkość

$$\frac{p}{1-p} = \exp(a_0 + a_1 x_1 + \cdots + a_m x_m)$$

nazywamy **ilorazem szans**.

Funkcje związane z analizą regresji:

lm - regresja liniowa, procedura główna,

nls - regresja nieliniowa, procedura główna,

glm - regresja logistyczna, procedura główna,

predict - prognozowanie.