Regresja

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:

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

gdzie

$f(x_1, x_2, \dots, x_m)$ - funkcja regresji
ε_i - błędy (reszty)

Założenia analizy regresji

- Niezależność obserwacji dla poszczególnych jednostek eksperymentalnych
- Brak błędu systematycznego
- Jednakowa i stała wariancja błędów
- 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

Regresja

Metody estymacji funkcji regresji:

- **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:

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

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

Regresja liniowa

Regresja prosta liniowa

- X – zmienna niezależna (objaśniająca)
- Y – zmienna zależna (objaśniana)

- **Model:**

$$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1, \dots, n$$

gdzie
a, b – **parametry** liniowej funkcji regresji
ε_i – błędy (reszty)

Regresja prosta liniowa

- **Fakt**:
  Estymatorami najmniejszych kwadratów (ENK) parametrów a i b liniowej funkcji regresji są statystyki:
  $$\hat{a} = \bar{Y} - \hat{b}\bar{x}$$
  $$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- **Fakt**:
  W modelu prostej regresji liniowej, statystyki $\hat{a}$ i $\hat{b}$ są nieobciążonymi estymatorami parametrów a i b. Ponadto, statystyka
  $$\hat{\sigma}^2 = S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}x_i\right)^2$$
  jest nieobciążonym estymatorem parametru σ^2

Regresja prosta liniowa

- **Twierdzenie:**
  Przy dodatkowym założeniu normalności rozkładu błędów, w modelu prostej regresji liniowej, statystyki
  $$\hat{a} \text{ i } S^2 \quad \text{oraz} \quad \hat{b} \text{ i } S^2$$
  są niezależnymi zmiennymi losowymi
  Ponadto:
  $$\hat{a} \sim N\left(a, \sigma^2\left(\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$
  $$\hat{b} \sim N\left(b, \sigma^2\left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$
  $$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$$

Regresja prosta liniowa

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

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

gdzie

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2$$

$$\hat{Y}_i = \hat{a} + \hat{b} x_i$$

Prognozowanie (predykcja)

- Niech x_p oznacza wartość zmiennej niezależnej X dla której chcemy uzyskać prognozę zmiennej zależnej Y równą Y_p
- Przyjmujemy

$$\hat{Y}_p = \hat{a} + \hat{b}x_p$$

- $(1-\alpha)\cdot 100\%$ przedział predykcji dla Y_p, przy założeniu normalności rozkładów błędów, ma postać

$$\left(\hat{Y}_p - t\left(1-\frac{\alpha}{2}, n-2\right)S_p, \hat{Y}_p + t\left(1-\frac{\alpha}{2}, n-2\right)S_p\right)$$

gdzie

$$S_p = S\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Testy dla parametrów funkcji regresji

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

$$H_0: \quad a = 0$$
$$H_1: \quad a \neq 0$$

- Statystyka testowa:

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

- Rozkład statystyki testowej (przy założeniu normalności rozkładu błędów):

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

Testy dla parametrów funkcji regresji

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

$$H_0: \quad b = 0$$
$$H_1: \quad b \neq 0$$

- Statystyka testowa:

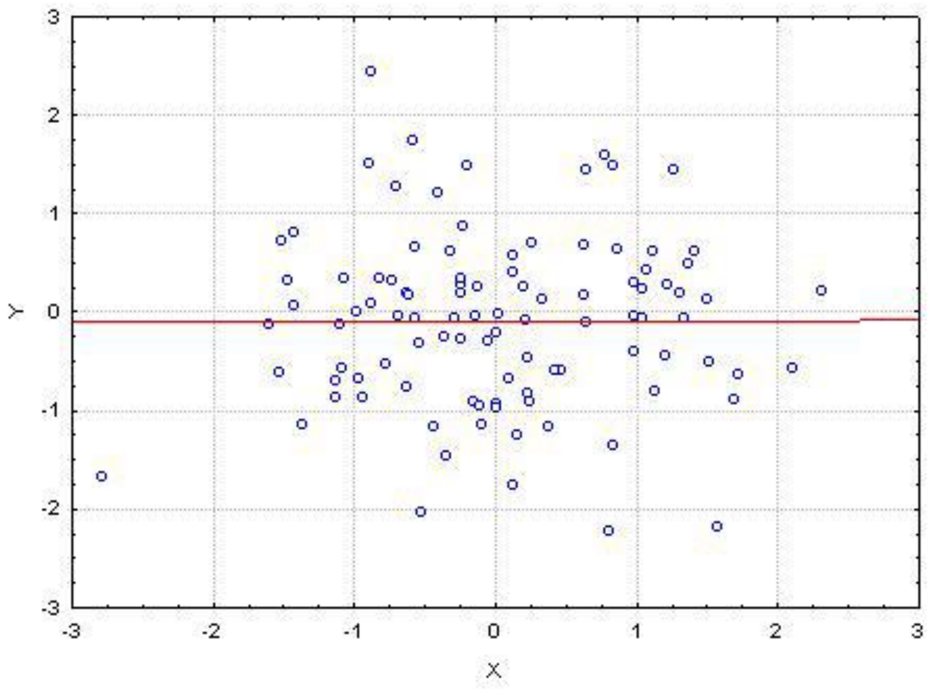
$$t = \frac{\hat{b}}{S_b}, \qquad S_b = S\sqrt{\frac{1}{\sum_{k=1}^{n}(x_k - \bar{x})^2}}$$

- Rozkład statystyki testowej (przy założeniu normalności rozkładu błędów):

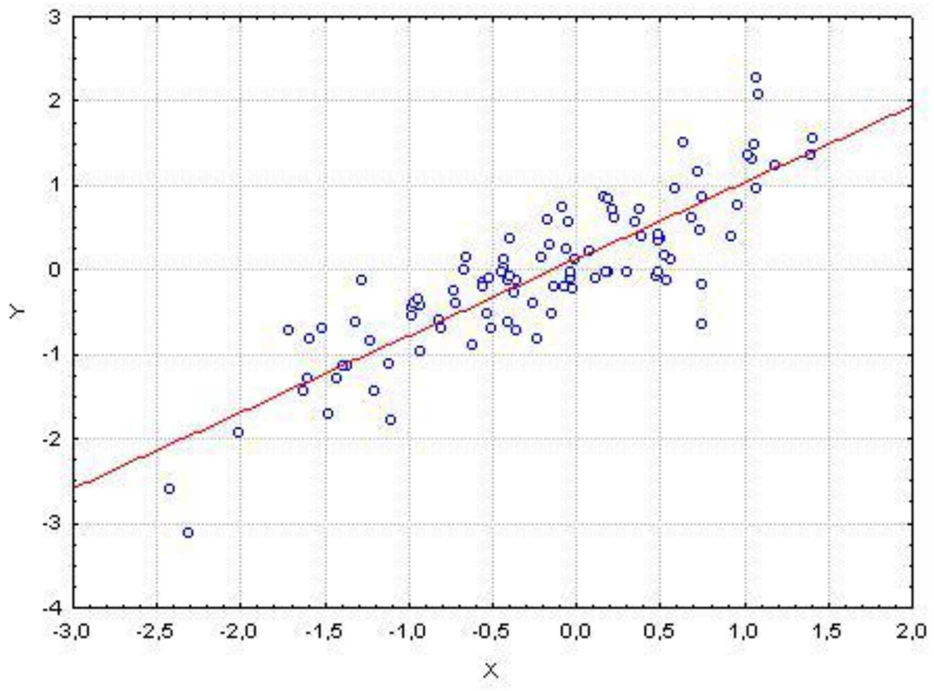
$$t\Big|_{H_0} \sim t(n-2)$$

Wpływ zmiennej niezależnej X na zmienną zależną Y

Brak istotnego wpływu, $b = 0$



Istotny wpływ, $b \neq 0$



Regresja wielokrotna (wieloraka) liniowa

- $X_1, X_2, \dots, X_m$ - zmienne niezależne (objaśniające)
- Y – zmienna zależna (objaśniana)

Model:

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

gdzie

$a_0, a_1, \dots, a_m$ - **parametry** liniowej funkcji regresji
ε_i - błędy (reszty)

Regresja wielokrotna liniowa

- **Zapis macierzowy**

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- **Model liniowy:**

$$Y = Xa + \varepsilon$$

Dodatkowe założenia

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

- Liczebność próby jest większa od liczby szacowanych parametrów:

$$n > m + 1$$

- Pomiędzy wektorami obserwacji zmiennych objaśniających **nie istnieje** zależność liniowa. Warunek ten oznacza, że

$$\text{rząd}(X) = m + 1$$

Estymatory parametrów modelu

- W modelu wielokrotnej regresji liniowej, statystyka
$$\hat{a} = (X'X)^{-1}X'Y$$
jest nieobciążonym estymatorem parametru a
- Ponadto, statystyka
$$\hat{\sigma}^2 = S^2 = \frac{1}{n-m-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$
gdzie
$$\hat{Y}_i = \hat{a}_0 + \hat{a}_1 x_{i1} + \cdots + \hat{a}_m x_{im}, \qquad i = 1, \dots, n$$
jest nieobciążonym estymatorem parametru σ^2

Współczynnik determinacji

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

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

gdzie

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2$$

gdzie

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

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

$$R_{pop}^2 = 1 - \frac{\frac{SSE}{n-m-1}}{\frac{SST}{n-1}}$$

Prognozowanie (predykcja)

- Niech X_p oznacza wektor wartości zmiennych objaśniających, dla której uzyskać chcemy prognozę zmiennej objaśnianej Y_p

$$X_p = \begin{bmatrix} 1 \\ x_1^p \\ \vdots \\ x_m^p \end{bmatrix}$$

- Przyjmujemy:

$$\hat{Y}_p = X_p' \hat{a}$$

Prognozowanie (predykcja)

- $(1-\alpha)\cdot 100\%$ **przedział ufności** dla Y_p, przy założeniu normalności rozkładu błędów:

$$\left(\hat{Y}_p - t\left(1-\frac{\alpha}{2}, n-m-1\right)S_p, \hat{Y}_p + t\left(1-\frac{\alpha}{2}, n-m-1\right)S_p\right)$$

gdzie

$$S_p^2 = S^2(1 + X_p'(X'X)^{-1}X_p)$$

Regresja nieliniowa

Regresja nieliniowa

Metody szacowania parametrów modelu:

- 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:
$$y' = \ln y, \qquad x_1' = \ln x_1, \qquad x_2' = \ln x_2, \qquad a_0' = \ln a_0$$
wtedy
$$y' = a_0' + a_1 x_1' + a_2 x_2'$$
- Numeryczne rozwiązanie zagadnienia minimalizacji sumy kwadratów błędów

Regresja logistyczna

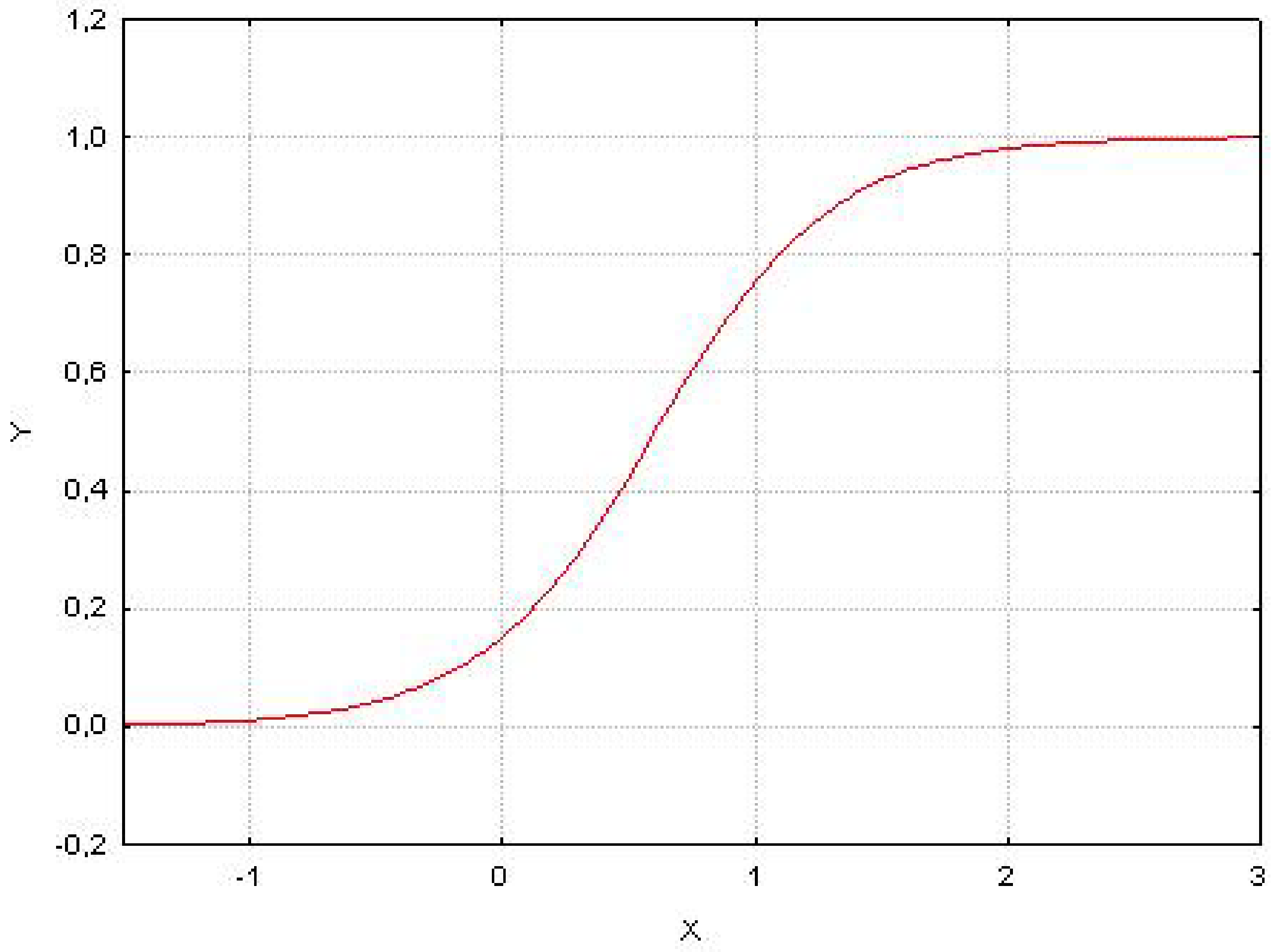
Regresja logistyczna

- W regresji logistycznej badamy wpływ m niezależnych zmiennych $X_1, X_2, \dots, X_m$ (ilościowych) na zależna zmienną Y mającą charakter zero-jedynkowy (dychotomiczny)
- **Model**:

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

gdzie p – prawdopodobieństwo sukcesu
$a_0, a_1, \dots, a_m$ - współczynniki regresji

Krzywa logistyczna



Regresja logistyczna

- Współczynniki regresji $a_0, a_1, \dots, a_m$ estymujemy metodą **największej wiarygodności** wykorzystując iteracyjny algorytm **IWLS** (algorytm iteracyjnie ważonych najmniejszych kwadratów)
- Wielkość
$$\ln \frac{p}{1-p} = a_0 + a_1 x_1 + \cdots + a_m x_m$$
nazywamy **logitem (funkcją logitową)**
- Wielkość
$$\frac{p}{1-p} = \exp(a_0 + a_1 x_1 + \cdots + a_m x_m)$$
nazywamy **ilorazem szans**

Regresja w R

Funkcje związane z analizą regresji:

- **lm** – regresja liniowa (procedura główna)
- **nls** – regresja nieliniowa (procedura główna)
- **glm** – regresja logistyczna (procedura główna)
- **predict** – prognozowanie