Zadanie 4

Marko Golovko

21 maja 2020

Przygotowanie danych

Ze strony www.populationpyramid.net, pobrałem dane dla 20 krajów (dane z dnia 14.05.2020). Program, który zawiera się w csv_convert.py tworzy tabelę z 10-letnimi przedziałami czasowymi dla ludności ogółem.

Dla przykładu część tabeli. Wierszy odpowiadają za przedziały czasowe od 0-9 do 100+. Cała tablica znajduje się w pliku CountriesPopulation.csv.

RUS	ESP	UK	ITA	GER	TUR	FRA	BEL
$1.86 \cdot 10^7$	$4.23 \cdot 10^6$	$8.04 \cdot 10^6$	$4.99 \cdot 10^6$	$7.88 \cdot 10^6$	$1.34 \cdot 10^7$	$7.53 \cdot 10^6$	$1.3 \cdot 10^6$
$1.53 \cdot 10^7$	$4.74 \cdot 10^6$	$7.64 \cdot 10^6$	$5.73 \cdot 10^6$	$7.93 \cdot 10^6$	$1.36 \cdot 10^7$	$7.88 \cdot 10^6$	$1.31 \cdot 10^6$
$1.56 \cdot 10^7$	$4.62 \cdot 10^6$	$8.56 \cdot 10^6$	$6.1 \cdot 10^6$	$9.38 \cdot 10^6$	$1.32 \cdot 10^7$	$7.37 \cdot 10^6$	$1.39 \cdot 10^6$
$2.45 \cdot 10^7$	$5.9 \cdot 10^6$	$9.3 \cdot 10^6$	$7 \cdot 10^6$	$1.09 \cdot 10^7$	$1.28 \cdot 10^7$	$8.01 \cdot 10^6$	$1.5 \cdot 10^6$
$2.04 \cdot 10^7$	$7.94 \cdot 10^6$	$8.6 \cdot 10^6$	$9.02 \cdot 10^6$	$1.02 \cdot 10^7$	$1.14 \cdot 10^7$	$8.33 \cdot 10^6$	$1.52 \cdot 10^6$
$1.89 \cdot 10^7$	$7.05 \cdot 10^6$	$9.17 \cdot 10^6$	$9.57 \cdot 10^6$	$1.35 \cdot 10^7$	$8.91 \cdot 10^6$	$8.64 \cdot 10^6$	$1.6 \cdot 10^6$
$1.85 \cdot 10^7$	$5.34 \cdot 10^6$	$7.29 \cdot 10^6$	$7.48 \cdot 10^6$	$1.06 \cdot 10^7$	$6.19 \cdot 10^6$	$7.76 \cdot 10^6$	$1.37 \cdot 10^6$
$8.55 \cdot 10^6$	$4.02 \cdot 10^6$	$5.83 \cdot 10^6$	$6.03 \cdot 10^6$	$7.47 \cdot 10^6$	$3.35 \cdot 10^6$	$5.73 \cdot 10^6$	$9.39 \cdot 10^5$
$4.88 \cdot 10^6$	$2.33 \cdot 10^6$	$2.81 \cdot 10^6$	$3.7 \cdot 10^6$	$4.89 \cdot 10^6$	$1.33 \cdot 10^6$	$3.14 \cdot 10^6$	$5.38 \cdot 10^5$
$7.71 \cdot 10^5$	$5.83 \cdot 10^5$	$6.21 \cdot 10^5$	$8.12 \cdot 10^5$	$9.62 \cdot 10^5$	$1.43 \cdot 10^5$	$8.66 \cdot 10^5$	$1.19 \cdot 10^5$
9,407	13,083	15,834	16,517	19,295	1,277	19,443	1,885

W pliku dane0512.ods znajduję się łączna liczba zachorowań w tych krajach oraz liczba zgonów.

Teoria

Regresja (względem) wielu zmiennych.

Dane są niezależne obserwacje $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, dla $i = 1, \ldots, n$. Szukamy wektor $\beta = [\beta_0, \beta_1, \ldots, \beta_k]$ minimalizujący wartość funkcji

$$f(\beta) = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k - y_i)^2.$$

Gdzie n to ilość krajów. y_i łączna liczba zachorowań w i-tym kraju lub łączna liczba zgonów w zależności od przypadku. x_{ik} liczba ludzi w i-tym kraju i k-tym przedziale

(x_1 to $0-9, \ldots, x_{10}$ to 100+).
Przechodząc do wersji macierzowej.

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \quad \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \approx \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Mnożąć powyższą równość lewostronnie, przez X^T otrzymujemy

$$X^T X\beta = X^T Y$$

I obliczamy wektor

$$\beta = (X^T X)^{-1} X^T Y$$

Obliczenia i wnioski

Wyniki obliczeń

Obliczenia są zapisane w pliku dz4.py. Pokazuję tylko końcowy wynik ze względu na rozmiar macierzy otrzymanych w pośrednich obliczeniach.
Dla równania regresji zachorowań

$$\beta_c^T = [-4835 \quad 0.1084 \quad -0.1332 \quad 0.0378 \quad -0.0839 \quad 0.0679 \quad 0.0637 \quad -0.029 \quad 0.0182$$
$$-0.1395 \quad 0.2332 \quad 0.7787]$$

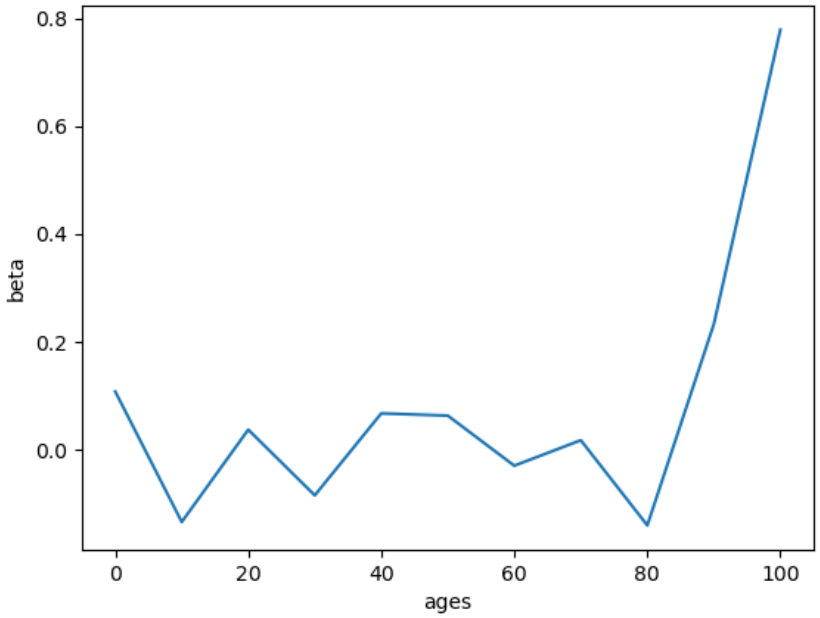
Dla równania regresji zgonów

$$\beta_d^T = [-2493 \quad 0.0102 \quad -0.0093 \quad -0.0026 \quad -0.0048 \quad 0.0057 \quad 0.0048 \quad -0.0102 \quad 0.0243$$
$$-0.0345 \quad 0.1064 \quad -2.5124]$$

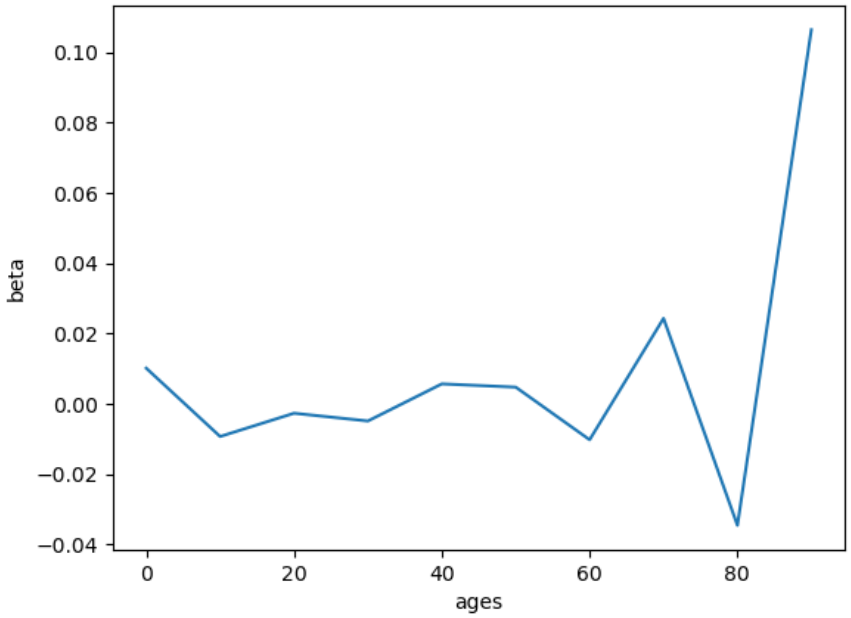
Wnoiski

Wizualizując obliczenia, otrzymałem takie dwa wykresy. zachorowania i zgony

Rysunek 1: Zachorowania



Rysunek 2: Zgony



Z wykresów wynika przypuszczenie, że dla osób w wieku więcej od 60, zwiększa się ryzyko.