Czwarte zadanie z kolokwium RPiS - sprawozdanie

Kacper Bajkiewicz

24 maja 2020

1 Słowem wstępu

Zadanie polega na znalezieniu regresji liniowej zachorowań (i regresji zgonów) względem piramidy wiekowej, posiłkując się danymi o kilkunastu krajach. W moim rozwiązaniu wziąłem państwa od Półwyspu Iberyjskiego po Morze Czarne, jednak wszystkie z nich znajdjują się w Europie (i ich struktura wiekowa nie jest od siebie bardzo rozbieżna). Dane dla 17 państw wyglądają tak:

		0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	100- –	TOTAL	ZAKAŻENIA	ZGONY
FRANCJA	1	3907487	7883477	7371029	8011050	8325667	8635056	7764785	5727704	3142141	865684	19443	61653523	140959	27972
FINLANDIA	1	568679	606660	672353	706580	659560	720555	712325	582988	255524	54519	975	5540718	6399	301
AUSTRIA	1	874182	876208	1115377	1243416	1178265	1407000	1033320	790437	402701	84339	1155	9006400	16309	640
CZECHY	1	1113646	1061331	1105339	1478210	1798354	1351270	1324617	1031099	379276	65453	387	10708982	8647	302
CHORWACJA	1	385776	408106	472200	539290	557002	581383	565378	361955	205097	28844	236	4105267	2232	96
DANIA	1	606029	675948	775962	672290	738207	810126	654530	586138	225319	46328	1326	5792203	11044	551
HOLANDIA	1	1752763	1953692	2097480	2097530	2151436	2524072	2129501	1591524	700180	133731	2963	17134872	44249	5715
NORWEGIA	1	614323	643031	726383	738724	724930	712508	586156	446448	184086	43572	1080	5421241	8257	233
HISZPANIA	1	4234486	4736077	4617599	5901993	7938499	7046327	5340654	4015306	2327453	583305	13083	46754782	232037	27778
SŁOWACJA	1	566591	545494	655785	847603	861400	709217	695211	398711	157104	22362	164	5459642	1495	28
SZWECJA	1	1193951	1127127	1276930	1320299	1264123	1296647	1093868	994244	431499	98312	2270	10099270	30799	3743
UKRAINA	1	4589511	4383517	4898491	7208958	6406404	5906737	5539856	2983430	1599359	215577	1919	43733759	19230	564
BELGIA	1	1302703	1312251	1386085	1503997	1518439	1598741	1369876	939413	537528	118698	1885	11589616	55791	9108
WĘGRY	1	908566	969207	1151690	1257142	1588213	1201388	1306432	846576	371819	58699	618	9660350	3598	470
NIEMCY	1	7880903	7930616	9377361	10872020	10243351	13488393	10644142	7471414	4894143	962305	19295	83783943	176107	8090
LITWA	1	300675	242841	316877	337565	355909	420558	352811	223283	139169	32030	571	2722289	1636	60
SZWAJCARIA	1	884945	834867	1039729	1219224	1166588	1320622	977435	751993	373598	83921	1695	8654617	30535	1613

Chcemy zbadać regresję względem zgonów i osobno regresję względem zakażeń. Niech Y będzie wektorem zgonów (zakażeń), β wektorem współczynników $\beta_0, \beta_1, ..., \beta k$ a X taką macierzą, że na miejscu x_{ij} znajduje się liczba ludności dla kraju $i <= n$ a w grupie wiekowej $j <= k$. Wtedy, sprowadzając problem do rachunku macierzowego, otrzymujemy następujące quasi-równanie (musimy po prostu znaleźć taki wektor β, żeby to *dopasowanie* było jak najlepsze.

$$\begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ 1 & ... & ... & \ddots & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \approx \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{1}$$

Pomnóżmy więc obie strony równania przez X^T. Wtedy $X^T X\beta = X^T Y$ Mnożąc zaś obustronnie przez $(X^T X)^{-1}$ sprowadza się to do postaci $Id_k\beta = (X^T X)^{-1} X^T Y$, czyli

$$\beta = (X^T X)^{-1} X^T Y \tag{2}$$

Na początku W rozwiązaniu tego problemu będę korzystać z Excelowych funkcji znajdujących transpozycję, odwrotność i iloczyn macierzy (liczenie tego na piechotę będzie co najmniej karkołomne), toteż wszystko zostanie zapisane w arkuszu kalkulacyjnym. Policzmy najpierw $(X^T X)^{-1} X^T$. Będzie to macierz wyglądająca mniej więcej tak:

264999,201	146122,81	-463964,22	342550,167	2672640,32	854546,3	-1386254,95	983036,55	559444,567	441865,5713	-357512,734621	-267365,5	-1041362,9	-2267287	-1486052	1051249	-46654,942
-13,006385	2,6798292	-9,000806	2,8308202	-10,420377	-4,803489	8,241882369	0,1013121	1,22331496	-2,811135647	6,29618319055	2,7023153	12,6638979	-0,587148	1,9693869	6,914009	-3,9936097
7,89383424	-5,0554317	-5,3948826	8,42311438	21,4672877	4,33275	10,99011973	9,134904	6,22092379	-4,132737111	-9,01581334047	3,5906326	16,7043495	-27,106524	-14,88285	-8,6119	-13,55778
-0,1996497	-9,9749867	18,511698	-20,6907708	-8,5682621	2,582437	-16,2995756	4,8902913	-13,672	-0,495782743	8,51103339462	-10,77192	2,55536282	37,68371	6,2292468	-4,44857	5,15773875
5,65020771	8,213104	-2,0866175	1,48247254	4,20828431	-4,106033	-15,1377226	1,622713	-1,0471704	5,293261708	7,91441485397	12,92222	-13,828707	-15,317589	0,8327382	-3,87934	8,26376007
-2,3119061	-9,8231462	0,6946587	0,55442736	3,87474752	3,365895	-9,7126803	3,1995231	10,29516	-0,727904275	1,57686799676	-4,764386	0,23883092	10,733045	-2,401674	-2,64411	-1,1473464
-2,812215	-7,0094588	11,032321	6,01130063	-0,653916	4,591264	20,95718976	-2,8874239	3,01153201	1,694033214	-17,594856843	-2,536349	-0,2892673	-16,330009	-2,658715	2,482514	3,99205655
4,32529401	14,302651	-18,141001	-3,89968144	-24,72211	-9,689019	24,02072442	-12,077617	-10,333067	9,239198597	-9,8656712706	3,912022	-2,4810752	15,843193	7,1409482	18,27457	-4,8493588
-0,7197385	15,93191	4,4434473	16,9276175	-0,0424504	5,277455	-1,17772295	-8,4307518	-1,7860069	-2,95176856	6,16219119642	-6,011116	-17,873924	-8,5901131	-0,850292	-6,93773	7,62899257
-13,151842	-24,078433	-10,318257	-23,3351106	78,3457543	19,36243	-56,8771723	13,782813	19,8532636	-26,10561876	31,5119524981	11,267779	2,91778612	0,7949677	22,578416	-32,2799	-13,268842
72,826428	62,704083	141,79467	71,2992947	-194,37408	-165,3005	29,91093052	-40,349261	-77,762036	-0,501454655	-41,8345608864	-89,09382	147,997202	51,289915	-82,1397	96,4498	18,0830562
-313,9774	446,74124	-5933,3523	-2811,50131	-713,17426	2822,444	-174,721824	682,30038	2304,12284	1266,127965	1658,76404632	2250,5059	-5279,2775	-251,41201	2574,7058	795,6057	677,098929

2 Właściwe obliczenia i wnioski

Wiadomo, że jeśli dla danej grupy wiekowej

2.1 Regresja zakażeń względem piramidy ludności

Po przemnożeniu macierzy powyżej przez wektor zakażeń otrzymujemy:

-5613,347	
0,0465978	0-10
0,0613187	10-20
-0,0213552	20-30
-0,0399683	30-40
0,0214606	40-50
0,0315774	50-60
-0,0721494	60-70
-0,0291789	70-80
0,046244	80-90
0,2760273	90-100
-7,0331774	100++

Widać, że współczynniki, które wyszły są... dość losowe. Z regresji powinniśmy wnioskować, że osoby najmłodsze mają dość wysoki udział w liczbie zakażeń (w przeciwieństwie do osób w wieku 40-80 lat), co jednak nie jest zgodne z tym, co mówią lekarze i epidemiolodzy. Dziwi też to, że zdecydowanie najmniej 'dokładają się' do zakażeń osoby najstarsze (100+). Prawdopodobnie inne czynniki niż wiek mają większy wpływ na zachorowalność i regresja została po prostu zniekształcona.

2.2 Regresja zgonów względem piramidy ludności

Zaś po przemnożeniu macierzy $(X^T X)^{-1} X^T$ przez wektor Y dostajemy:

-1030,473	
0,00834	0-10
0,016168	10-20
-0,004885	20-30
-0,009941	30-40
0,002633	40-50
0,001348	50-60
-0,003355	60-70
-0,01107	70-80
0,001443	80-90
0,090017	90-100
-2,145764	100++

Tutaj zaś widać bardzo sporą nieregularność w wyniku, spodziewalibyśmy się, że wśród osób w wieku 50-100+ będą miały duży wpływ na śmiertelność (według epidemiologów, ponadto te osoby też najczęściej chorują na choroby współistniejące). Tymczasem regresja wnioskuje, że im procentowo więcej jest ludzi najstarszych, tym maleje śmiertelność. Do tego osoby w wieku 0-10 mają większy udział niż w grupach 60-90 lat. Jedynie prawdopodobną sugestią jest wniosek dla grupy wiekowej 90-100. Tak jak w przypadku zakażeń, prawdopodobnie inne czynniki mają większy wpływ na zgony.

Kacper Bajkiewicz
nr indeksu 314438