Imię i nazwisko: Paulina Busz.
Nr. Indeksu: 186022.
Kierunek: Analityka gospodarcza.

Statystyczna analiza wielowymiarowa

Analiza głównych składowych

Dane: Analizę przeprowadzono na zbiorze danych „Dane transport" zawierającym 7 zmiennych oraz 17 obserwacji.

Zmienne:

- Rail accidents: liczba wypadków kolejowych w 2010 roku, podane w liczbach.
- Persons killed in road accidents: liczba zabitych osób na drogach w 2010 roku, podane w liczbach.
- Length of motorway and e-roads: długość autostrad oraz tras europejskich w 2010 roku, podane w kilometrach.
- Length of tracks: długość torów kolejowych w 2010 roku, podane w kilometrach.
- Goods transport by rail: ilość przewiezionych towarów drogą kolejową w 2010 roku, podane w tysiącach ton.
- Goods transport by road: ilość przewiezionych towarów przy wykorzystaniu dróg w 2010 roku, podane w tysiącach ton.
- Passenger cars: liczba zarejestrowanych samochodów osobowych w 2010 roku, podane w liczbach.

Obserwacje: 17 państw Unii Europejskiej.

Cel: Celem badania było sprawdzenie, czy można zmniejszyć wymiar badanego modelu (model 1) za pomocą analizy głównych składowych. Jako iż metoda analizy głównych składowych jest metodą nieparametryczną, nie jest konieczne sprawdzanie założeń odnośnie rozkładu danych.

Wykonano regresję liniową za pomocą metody najmniejszych kwadratów dla danych rzeczywistych:

	Estimate	Std. Error	t value	Pr(>|t|)	
Rail accidents	60,200	30,000	2,006	0,073	.
Persons killed in road accidents	-13,410	2,875	-4,664	0,001	***
Length of motorways and e-roads	-0,164	0,355	-0,462	0,654	
Length of tracks	0,616	0,205	3,007	0,013	*
Goods transport by rail	-0,066	0,026	-2,500	0,031	*
Goods transport by road	-0,011	0,006	-2,032	0,070	.
Passenger cars	0,003	0,000	7,080	0,000	***

Tabela 1. Regresja dla danych rzeczywistych.

Kryterium informacyjne Akaikego	317.3081

Tabela 2. Kryterium informacyjne.

Z uwagi na różne jednostki badanych zmiennych, zmienne te poddano standaryzacji. Metoda składowych głównych ma głównie charakter eksploracyjny i umożliwia redukcję danych w przypadku zbioru skorelowanych ze sobą zmiennych[1], w związku z czym zbadano korelację analizowanych zmiennych.

	Rail accidents	Persons killed in road accidents	Length of motorways and e-roads	Length of tracks	Goods transport by rail	Goods transport by road	Passenger cars
Rail accidents	1,000	0,660	-0,140	0,700	0,796	0,264	0,211
Persons killed in road accidents	0,660	1,000	0,481	0,838	0,605	0,804	0,830
Length of motorways and e-roads	-0,140	0,481	1,000	0,320	-0,084	0,728	0,662
Length of tracks	0,700	0,838	0,320	1,000	0,799	0,802	0,752
Goods transport by rail	0,796	0,605	-0,084	0,799	1,000	0,469	0,410
Goods transport by road	0,264	0,804	0,728	0,802	0,469	1,000	0,954
Passenger cars	0,211	0,830	0,662	0,752	0,410	0,954	1,000

Tabela 3. Macierz korelacji.

Silna korelacja pomiędzy zmiennymi objaśniającymi w modelu regresji powoduje powstawanie zależności w szeregach czasowych zmiennych objaśniających. Sytuacja ta może prowadzić do niedokładnego oszacowania parametrów modelu. Istnienie współliniowości zmiennych nie wpływa na zgodność i nieobciążoność estymatorów, wpływa natomiast na ich efektywność.

Obliczono macierz kowariancji, za pomocą której wyznaczono wartości oraz wektory własne. Wyniki przedstawiono w tabeli 4.

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Wartości własne	4,557	1,782	0,335	0,206	0,092	0,022	0,007

Tabela 4. Wartości własne.

Wyznaczono składowe oraz obliczono statystyki wyjaśnionej wariancji.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	2,071	1,295	0,561	0,440	0,294	0,144	0,081
Proportion of Variance	0,651	0,255	0,048	0,029	0,013	0,003	0,001
Cumulative Proportion	0,651	0,906	0,953	0,983	0,996	0,999	1,000

Tabela 5. Statystyki wyjaśnionej wariancji.

Z tabeli 5 wynika, iż dwie pierwsze składowe wyjaśniają 90,6% wariancji całkowitej. Procent ten jest na tyle duży, by móc pominąć 5 pozostałych składowych.

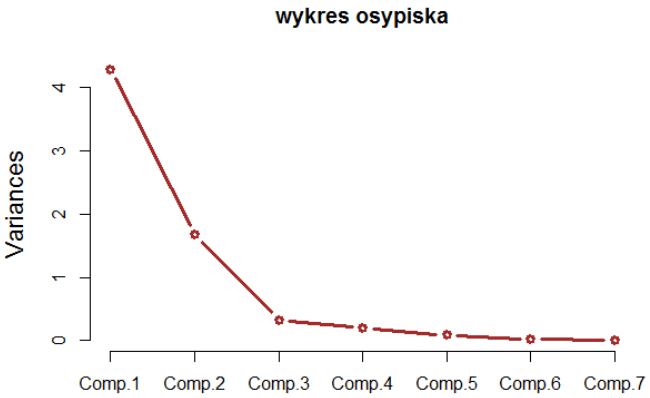
[1] „Prognozowanie ekonomiczne – teoria, przykłady, zadania", B. Pawełek, S. Wanat, A. Zeliaś, Wydawnictwo Naukowe PWN, Warszawa 2003

Aby wybrać odpowiednią liczbę składowych głównych posłużono się trzema następującymi kryteriami[2]:

1. Procent wariancji: Obliczono wskaźnik:

$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5+\lambda_6+\lambda_7}$, Którego wartość wynosiła: 90,55%. Jest on, większy od ustalonej wartości 80%[3], można więc pominąć składowe 3,4,5,6,7.

2. Wartość średnia z wartości własnych: Obliczona średnia arytmetyczna z wyznaczonych wartości własnych wynosi 1. Tylko dwie pierwsze wartości własne są większe od średniej. Oznacza to, iż w stopniu wystarczającym wyjaśniają całkowitą zmienność 7 zmiennych.



Wykres 1. Wykres osypiska.

3. Wykres osypiska: Z wykresu odczytano liczbę składowych głównych jako 2. Bazując na samym wykresie można się wahać co do liczby składowych między 2 a 3, ponieważ nie można jednoznacznie określić momentu spłaszczenia się wykresu, jednakże biorąc pod uwagę inne kryteria, wybrano 2 główne składowe.

Wyznaczono ładunki:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Rail accidents	-0,291	0,532	0,557	0,053	0,106	-0,310	-0,460
Persons killed in road accidents	-0,441	0,011	0,402	-0,484	-0,290	0,167	0,545
Length of motorways and e-roads	-0,244	-0,577	0,428	0,591	-0,111	0,227	-0,105
Length of tracks	-0,442	0,150	-0,183	0,116	0,749	0,387	0,157
Goods transport by rail	-0,337	0,449	-0,403	0,438	-0,549	0,156	0,064
Goods transport by road	-0,426	-0,276	-0,247	0,091	0,091	-0,787	0,213
Passenger cars	-0,412	-0,289	-0,296	-0,447	-0,146	0,177	-0,637

Tabela 6. Ładunki.

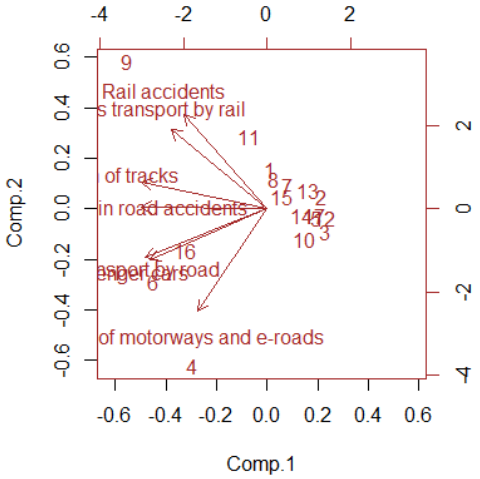
Z tabeli 6 odczytano, iż największy udział w budowie pierwszej składowej ma zmienna „Length of tracks" natomiast największy udział w budowie drugiej składowej, a zmienna „Rail accidents". Warto również zauważyć, iż najmniejszy udział w budowie drugiej składowej ma zmienna „Persons killed in road accidents".

[2] „Systemy uczące się", M. Krzyśko, W. Wołyński, T. Górecki, M. Skorzybut, WNT. Warszawa 2008
[3] „Systemy uczące się", M. Krzyśko, W. Wołyński, T. Górecki, M. Skorzybut, WNT. Warszawa 2008

Składowe główne:

	Comp.1	Comp.2
Czech Republic	0,117	0,844
Estonia	1,833	0,257
Ireland	1,970	-0,483
Spain	-2,499	-3,321
Croatia	1,674	-0,182
Italy	-3,819	-1,535
Hungary	0,723	0,507
Austria	0,256	0,638
Poland	-4,694	3,105
Portugal	1,237	-0,632
Romania	-0,621	1,522
Slovenia	1,931	-0,203
Slovakia	1,398	0,381
Finland	1,172	-0,137
Sweden	0,491	0,258
United Kingdom	-2,764	-0,875
Norway	1,595	-0,144

Tabela 7. Składowe główne.



Wykres 2. Biplot.

Z wykresu *biplot* wynika, iż między zmiennmi „Rail accidents" oraz „Goods transport by rail" zachodzi korelacja. Podobną zależność można odczytać również między następującymi parami zmiennych: „Length of tracks" – „Persons killed in road accidents" oraz „Goods transport by road" – „Passenger cars". Z wykresu można również odczytać między innymi ujemny wpływ zmiennej „Length of motorways and e-roads" na składową pierwszą.

Następnie wykonano regresję metodą najmniejszych kwadratów dla głównych składowych, gdzie zmienną objaśnianą (y) jest liczba pasażerokilometrów (w milionach) podróżująca koleją.

$$y = -0{,}369\ comp1 - 0{,}251\ comp2$$

	Estimate	Std. Error	t value	Pr(>|t|)	
1 Comp.	-0,369	0,063	-5,879	0,000030	***
2 Comp.	-0,251	0,100	-2,501	0.0245	*

Tabela 8. Statystyki.

Kryterium informacyjne Akaikego	30,87388

Tabela 9. Kryterium informacyjne.

W celu zweryfikowania założenia dotyczącego normalności rozkładu reszt w modelu wykonano test Shapiro-Wilka z następującymi hipotezami:

H_0: Reszty modelu regresji głównych składowych mają rozkład normalny.

H_1: Reszty modelu regresji głównych składowych nie mają rozkładu normalnego.

Uzyskane prawdopodobieństwo testowe (p-value) na poziomie istotności 0,01 wynoszące 0,00001 implikuje stwierdzenie o odrzuceniu hipotezy zerowej na rzecz hipotezy alternatywnej.

W przypadku braku normalności estymatory parametrów stukturlanych są nieobciążone, zgodne i najbardziej efektywne w klasie esytmatorów liniowych i nieobciążonych, jednakże licząc się z brakiem normalności składnika

resztowego należy uważać przy wyciąganiu wniosków na podstawie testów, które korzystają z założenia o normalności składnika resztowego takich jak np. test T-Studenta.

Zabadano także homoskedastyczność wariancji za pomocą testu Breuscha-Pagana z następującymi hipotezami:

H_0: Homoskedastyczność wariancji składnika resztowego.

H_1: Heteroskedastyczność wariancji składnika resztowego.

Analizując uzyskaną wartość prawdopodobieństwa testowego równego 0,009 na poziomie istotności 0,05 odrzucamy hipotezę zerową mówiącą o homoskedastyczności reszt składnika resztowego.

W przypadku występowania heteroskedastyczności w modelu należy uznać, iż estymator z niego uzyskany będzie zgodny oraz nieobciążony, jednakże nie będzie on efektywny. Natomiast estymatory wariancji składnika losowego mogą być obciążone, co może skutkować niedoszacowaniem średnich błędów estymatorów parametrów, nieprawidłowymi przedziałami ufności.

Powyższa analiza została wykonana w oparciu o dane standaryzowane zważywszy na różne jednostki analizowanych zmiennych. Mając jednak na uwadze, iż teoria mówi o scentrowaniu a nie standaryzacji zmiennych dokonano analizy głównych składowych dla danych scentrowanych.

W oparciu o macierz kowariancji wyznaczono jej wartości własne:

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Wartości własne dla danych scentrowancyh	118806100000000,000	28354880000,000	2058012000,000	15098660,000	2451071,000	272662,900	340,219

Tabela 10. Wartości własne dla danych scentrowanych.

Obliczono statystyki wyjaśnionej wariancji dla danych scentrowanych.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	10574380,000	163361,400	44010,830	3769,682	1518,845	506,581	17,894
Proportion of Variance	1,000	0,000	0,000	0,000	0,000	0,000	0,000
Cumulative Proportion	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Tabela 11. Statystyki wyjaśnionej wariancji dla danych scentrowanych.

Z wyników obliczeń przedstawionych w tabeli 11 wynika, iż pierwsza składowa wyjaśnia 100% wariancji całkowitej. Aby zbadać powód zaistniałej sytuacji obliczono ładunki, z których wynika, iż w skład pierwszej składowej wchodzi głównie zmienna „Passenger cars".

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Rail accidents	0.000	0.000	-0.002	0.010	-0.012	-0.089	0.996
Persons killed in road accidents	0.000	0.000	-0.008	0.059	-0.171	-0.979	-0.090
Length of motorways and e-roads	0.000	0.006	0.039	-0.081	-0.982	0.166	0.004
Length of tracks	-0.001	0.019	-0.114	0.988	-0.073	0.073	-0.004
Goods transport by rail	-0.002	0.086	-0.989	-0.118	-0.028	0.006	0.000
Goods transport by road	-0.049	0.995	0.087	-0.008	0.010	-0.003	0.000
Passenger cars	-0.999	-0.049	-0.002	0.000	0.000	0.000	0.000

Tabela 12. Ładunki.

Z powodu uzyskania tylko jednej składowej w analizie głównych składowych dla danych scentrowanych iż wykonywanie regresji dla składowych głównych może być niemiarodajne.