

Raport końcowy

Sara Łuba
Grupa 21MAT-SD

W raporcie wykorzystamy dane z gretla o nazwie „Annual US labor-market data”, co możemy przetłumaczyć jako roczne dane z amerykańskiego rynku pracy. Dane pochodzą z lat 1947-1962. W pliku mamy 6 zmiennych:

Nazwa zmiennej	Objaśnienie
<i>employ</i>	Zatrudnienie ogółem
<i>prdefl</i>	Deflator PKB, miernik poziomu cen.
<i>gnp</i>	Produkt Krajowy Brutto
<i>unemp</i>	Bezrobocie
<i>armfrc</i>	Wielkość sił zbrojnych
<i>pop</i>	Ludność nieinstytucjonalna w wieku 14 lat i więcej

1. Model liniowy

Chcemy przedstawić następujący model liniowy:

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \dots + \beta_k X_{t,k} + \epsilon_t$$

gdzie Y_t jest zmienną objaśnianą, ϵ_t jest składnikiem losowym, a $X_{t,i}$ dla $1 \leq i \leq k$ są zmiennymi objaśniającymi. W naszym przypadku chcemy rozpatrzyć model:

$$unemp = \beta_0 + \beta_1(employ) + \beta_2(prdefl) + \beta_3(gnp) + \beta_4(armfrc) + \beta_5(pop) + \epsilon,$$

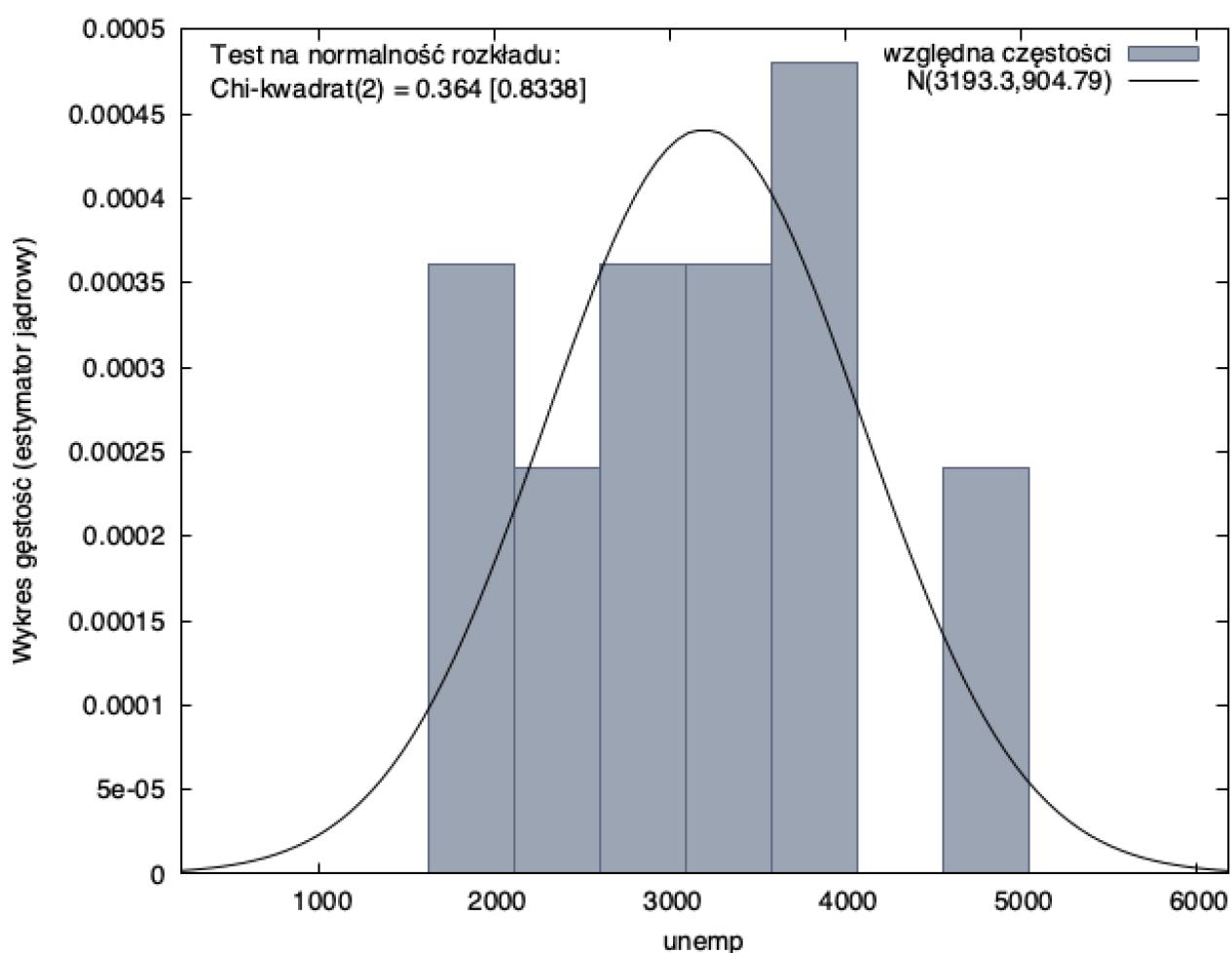
co oznacza, że rolę $X_{i,t}$ pełnią *employ*, *prdefl*, *gnp*, *armfrc* oraz *pop*, gdzie $i \in [1,5]$, $t = 1, \dots, n$, gdzie $n = 15$ (liczba lat).

Aby uznać zmienną za objaśnianą musimy sprawdzić, czy pochodzi z rozkładu normalnego.

Weryfikujemy normalność zmiennej *unemp*, korzystając z testu Doornika-Hansena, Shapiro-Wilka, Lillieforsa oraz Jarque'a-Bera.

Nazwa testu	Wynik
Test Doornika-Hansena	0.625881, z wartością p 0.731293
Test Shapiro-Wilka	0.949721, z wartością p 0.485273
Test Lillieforsa	0.127383, z wartością p ~ 0.69
Test Jarque'a-Bera	0.823229, z wartością p 0.66258

Wyniki testu wskazują na brak podstaw do odrzucenia hipotezy, że dane dotyczące bezrobocia pochodzą z rozkładu normalnego ($p \geq 0.05$ dla każdego testu).



Test Chi-kwadrat również nie wyklucza że $unemp$ pochodzi z rozkładu normalnego. Skoro wszystkie dostępne testy: Doornika-Hansena, Shapiro-Wilka, Lillieforsa, Jarque'a-Bera oraz test Chi-kwadrat przyjęły hipotezę o normalności, możemy przyjąć za zmienną objaśnianą zmienną $unemp$.

2. METODA WYBORU ZMIENNYCH

METODA WSTECZ

Metoda wstecz polega na skonstruowaniu modelu zawierającego wszystkie potencjalne zmienne objaśniające. Następnie stopniowo eliminujemy te zmienne, które nie są statystycznie istotne dla naszego modelu, czyli te, dla których wartość p jest większa od 0.05. Kończymy wykonywanie metody dopiero, gdy wszystkie zmienne, które zostały, są istotne.

Zaczynamy od następującego modelu:

$$unemp = \beta_0 + \beta_1(employ) + \beta_2(prdefl) + \beta_3(gnp) + \beta_4(armfrc) + \beta_5(pop) + \epsilon$$

Model 1: Estymacja KMNK, wykorzystane obserwacje 1947–1962 (N = 16)
Zmienna zależna (Y): unemp

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	-45695.1	28175.4	-1.622	0.1359	
employ	-0.193588	0.210203	-0.9210	0.3788	
prdefl	191.161	69.5812	2.747	0.0206	**
gnp	-0.0421968	0.0235193	-1.794	0.1030	
armfrc	-0.317528	0.208772	-1.521	0.1592	
pop	0.504861	0.186007	2.714	0.0218	**

Usuwamy zmienną *employ*, ponieważ ma najwyższą wartość p, która jest większa od 0.05. Otrzymujemy następujący model:

Model 2: Estymacja KMNK, wykorzystane obserwacje 1947–1962 (N = 16)
Zmienna zależna (Y): unemp

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	-68988.3	12329.5	-5.595	0.0002	***
prdefl	217.552	62.9672	3.455	0.0054	***
gnp	-0.0608971	0.0117859	-5.167	0.0003	***
armfrc	-0.226751	0.182764	-1.241	0.2405	
pop	0.632421	0.123293	5.129	0.0003	***

Eliminujemy zmienną *armfrc*. Otrzymujemy:

Model 3: Estymacja KMNK, wykorzystane obserwacje 1947–1962 (N = 16)
Zmienna zależna (Y): unemp

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	-77834.0	10282.5	-7.570	6.59e-06	***
prdefl	211.834	64.1939	3.300	0.0063	***
gnp	-0.0677876	0.0106264	-6.379	3.51e-05	***
pop	0.730421	0.0967716	7.548	6.79e-06	***

Możemy przyjąć, że zmienne, które pozostały mają istotny wpływ na zmienną objaśnianą, zatem jest to model statystycznie istotny. Współczynnik determinacji wynosi 0.894322 , zatem model został wyjaśniony w ok. 89.43 %.

Otrzymujemy:

$$unemp = - 77834 + 211.834(prdefl) - 0.068(gnp) + 0.73(pop) + \epsilon$$

Według danego modelu bezrobocie rośnie wraz ze wzrostem *prdefl* i *pop*, a maleje wraz ze wzrostem *gnp*.

METODA GRAFÓW

Metoda ta polega na obliczeniu korelacji Pearsona dla każdej pary zmiennych objaśniających, oraz zmiennej objaśnianej. Następnie tworzymy graf powiązań taki, że krawędź należy do grafu jeśli korelacja jest większa od wartości krytycznej r^* . Wybieramy następnie zmienne, które mają najwięcej powiązań. Do modelu dołączamy zmienne z największą ilością powiązań oraz takie, które są najsilniej skorelowane z naszą zmienną objaśnianą.

Współczynniki korelacji Pearsona dla zmiennych objaśnianych zostały podane w poniższej tabeli:

	employ	prdefl	gnp	armfrc	pop
employ	1	0,9526055573	0,9835516112	0,4573074	0,9603905716
prdefl	0,9526055573	1	0,9977170738	0,9896527827	0,9976870858
gnp	0,9835516112	0,9977170738	1	0,4464367919	0,9910900695
armfrc	0,4573074	0,9896527827	0,4464367919	1	0,3644162672
pop	0,9603905716	0,9976870858	0,9910900695	0,3644162672	1

Wartość krytyczna r^* współczynnika korelacji:

$$r^* = \sqrt{\frac{2.145^2}{2.145 + 14}} = 0.533837...$$

Zmienne dla których:

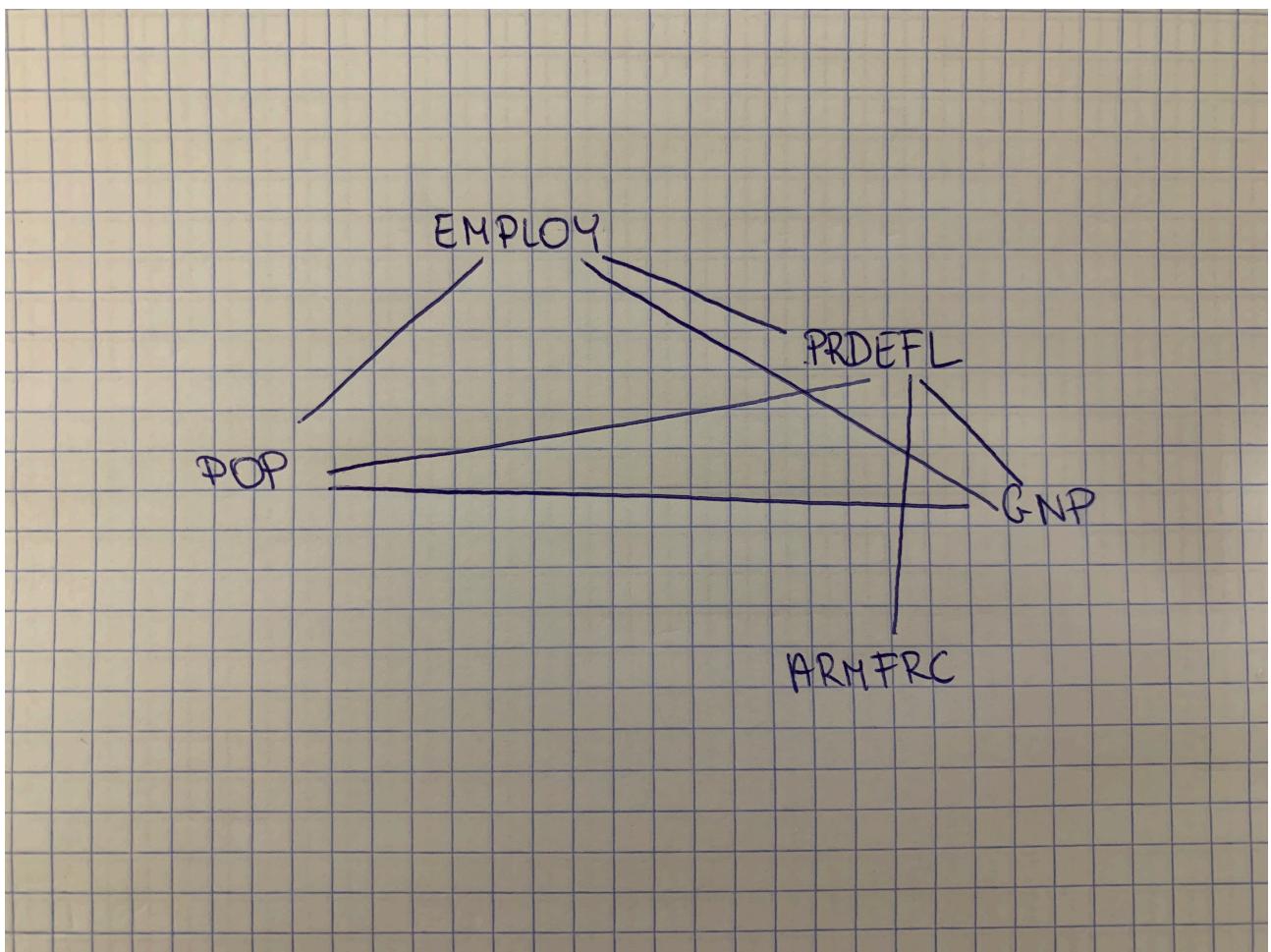
$$| r_j | \leq r^*$$

zostaną zastąpione zerem.

Otrzymujemy następującą macierz:

	employ	prdefl	gnp	armfrc	pop
employ	1	0,9526055573	0,9835516112	0	0,9603905716
prdefl	0,9526055573	1	0,9977170738	0,9896527827	0,9976870858
gnp	0,9835516112	0,9977170738	1	0	0,9910900695
armfrc	0	0,9896527827	0	1	0
pop	0,9603905716	0,9976870858	0,9910900695	0	1

Budujemy graf powiązań:



Nie mamy punktów odizolowanych, wybieramy zmienne, które mają największą ilość powiązań, czyli zmienne *employ*, *prdefl*, *pop* i *gnp*.

	unemp
employ	0,5024980839
prdefl	0,2919603441
gnp	0,6042609399
armfrc	-0,1774206295
pop	0,6865515164

Najsiłniej skorelowana ze zmienną *unemp* jest zmienna *pop*.
Otrzymujemy następujący model:

$$unemp = -7636.65 + 0.092(pop) + \epsilon$$

Według danego modelu bezrobocie rośnie wraz ze wzrostem *pop*.

METODA WPRZÓD

Przy metodzie wprzód zaczynamy od modelu bez zmiennych objaśniających. Następnie, stopniowo dodajemy zmienne, których wartość p jest najmniejsza i $p < 0.05$. Kończymy wykonywanie metody w momencie, gdy wszystkie zmienne które pozostały do sprawdzenia mają wartość p większą od 0.05, czyli nie są statystycznie istotne dla naszego modelu.

unemp	employ	prdefl	gnp	armfrc	pop
współczynnik	0.133705	53.7420	0.00568098	-0.238236	0.0922295
t-Studenta	2.175	2.962	2.838	-0.6745	3.533
wartość p	0.0473	0.0103	0.0132	0.5109	0.0033

Największą wartość t-Studenta i najmniejszą wartość p ma *pop*, więc jako pierwszy zostanie dołączony do modelu.

unemp	employ	prdefl	gnp	armfrc
współczynnik	-0.537504	-108.375	-0.0403680	-0.662114
t-Studenta	-4.410	-1.345	-4.593	-2.937
wartość p	0.0007	0.2016	0.0005	0.0116

gnp ma najwyższą wartość t-Studenta i najmniejszą wartość p. Zatem jest dodany do modelu. W naszym modelu mamy już dwie zmienne, *pop* oraz *gnp*.

unemp	employ	prdefl	armfrc
współczynnik	-0.262625	211.834	-0.180533
t-Studenta	-1.162	3.300	-0.7164
wartość p	0.2677	0.0063	0.4875

Dodajemy *prdefl* do modelu.

unemp	employ	armfrc
współczynnik	-0.0426454	-0.226751
t-Studenta	-0.2175	-1.241
wartość p	0.8318	0.2405

Ani *employ*, ani *armfc* nie może zostać dodana do modelu, ponieważ odrzucamy hipotezę, że odpowiedni współczynnik jest nieistotny. Zatem ostateczny model wygląda następująco:

$$unemp = -77834 + 0.73(pop) - 0.068(gnp) + 211.834(prdefl) + \epsilon$$

Model 30: Estymacja KMNK, wykorzystane obserwacje 1947–1962 (N = 16)
Zmienna zależna (Y): unemp

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	-77834.0	10282.5	-7.570	6.59e-06	***
pop	0.730421	0.0967716	7.548	6.79e-06	***
gnp	-0.0677876	0.0106264	-6.379	3.51e-05	***
prdefl	211.834	64.1939	3.300	0.0063	***

Jak widzimy, otrzymaliśmy identyczny model jak w przypadku metody wstecz.

3. ANALIZA RESZT

METODA WSTE CZ

Dla modelu

$$unemp = -77834 + 211.834(prdefl) - 0.068(gnp) + 0.73(pop) + \epsilon$$

obliczam reszty

$$\epsilon_t = unemp - (-77834 + 211.834(prdefl) - 0.068(gnp) + 0.73(pop))$$

W celu zweryfikowania hipotezy normalności zastosuję testy Doornika-Hansena, Shapiro-Wilka, Lillieforsa oraz Jarque'a-Bera. Zastosuję również test chi-kwadrat.

Test na normalność rozkładu uhat38:

Test Doornika-Hansena = 0.120769, z wartością p 0.941403

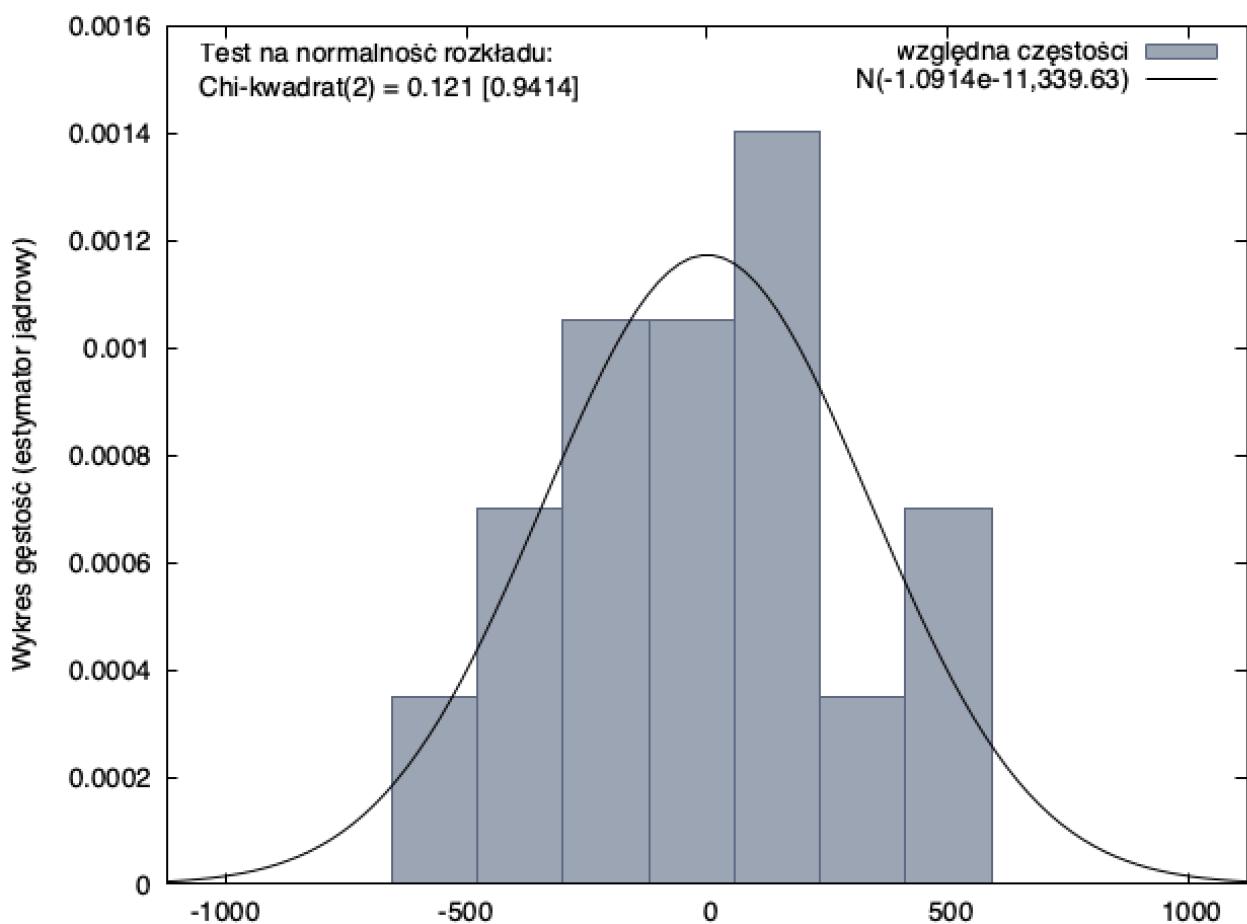
Test Shapiro-Wilka = 0.970567, z wartością p 0.847708

Test Lillieforsa = 0.0972162, z wartością p ~ 1

Test Jarque'a-Bera = 0.304649, z wartością p 0.85871

Wszystkie testy przyjmują hipotezę o normalności reszt.

Test chi-kwadrat prowadzi do potwierdzenia hipotezy o normalności reszt z $p = 0.9414$



Statystyka Durbina-Watsona:

test Durbina-Watsona pozwala ocenić, czy występuje autokorelacja wśród reszt.

Określając wartość testu musimy skorzystać z tablic rozkładu Durbina-Watsona. Dla liczby predytorów w modelu oraz liczby obserwacji otrzymujemy dwie wartości d_L oraz d_U .

- $DW = 2$ oznacza brak autokorelacji
- jeżeli $DW > 2$:
 - dla $DW > 4-dL$ korelacja jest ujemna
 - dla $4-dU < DW < 4-dL$ problem jest nierożstrzygnięty
 - dla $DW < 4 - dU$ otrzymujemy brak autokorelacji
- jeżeli $DW < 2$:
 - dla $DW < dL$ korelacja jest dodatnia
 - dla $dL < DW < dU$ problem jest nierożstrzygnięty
 - dla $DW > dU$ otrzymujemy brak autokorelacji.

Wartość DW: 1.42287

Statystyka testu Durbina–Watsona dla 5% poziomu istotności, n = 16, k = 3

$$\begin{aligned} dL &= 0.8572 \\ dU &= 1.7277 \end{aligned}$$

$$dL \leq DW \leq dU$$

Zatem problem jest nierożstrzygnięty.

Sprawdzamy homoskedastyczność rozkładu reszt. Do sprawdzenia wykorzystamy test White'a, który pozwala zbadać, czy wariancja reszt w modelu jest stała.

Test White'a na heteroskedastyczność reszt (zmienna wariancji resztowej)
Estymacja KMNK, wykorzystane obserwacje 1947–1962 (N = 16)
Zmienna zależna (Y): uhat²

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	-2.10942e+08	6.44120e+08	-0.3275	0.7544
pop	6463.07	11959.8	0.5404	0.6084
gnp	-182.836	1243.56	-0.1470	0.8879
prdefl	-2.56846e+06	5.39092e+06	-0.4764	0.6506
sq_pop	-0.0463544	0.0594860	-0.7792	0.4654
X2_X3	0.00333102	0.0112646	0.2957	0.7774
X2_X4	30.6255	53.6663	0.5707	0.5889
sq_gnp	-2.21257e-05	0.000616508	-0.03589	0.9725
X3_X4	-1.81746	5.13135	-0.3542	0.7353
sq_prdefl	-1833.30	13546.7	-0.1353	0.8968

Wsp. determ. R-kwadrat = 0.427913

Statystyka testu: TR² = 6.846600,
z wartością p = P(Chi-kwadrat(9) > 6.846600) = 0.653087

Test White'a przyjmuje hipotezę że reszty są homoskedastyczne, zatem mają jednakową wariancję.

METODA GRAFÓW

Dla modelu

$$unemp = -7636.65 + 0.092(pop) + \epsilon$$

obliczam reszty:

$$\epsilon_t = unemp - (-7636.65 + 0.092(pop))$$

Podobnie jak w przypadku analizy reszt dla metody wstecz, korzystam z testów Doornika-Hansena, Shapiro-Wilka, Lillieforsa, Jarque'a-Bera oraz testu chi-kwadrat.

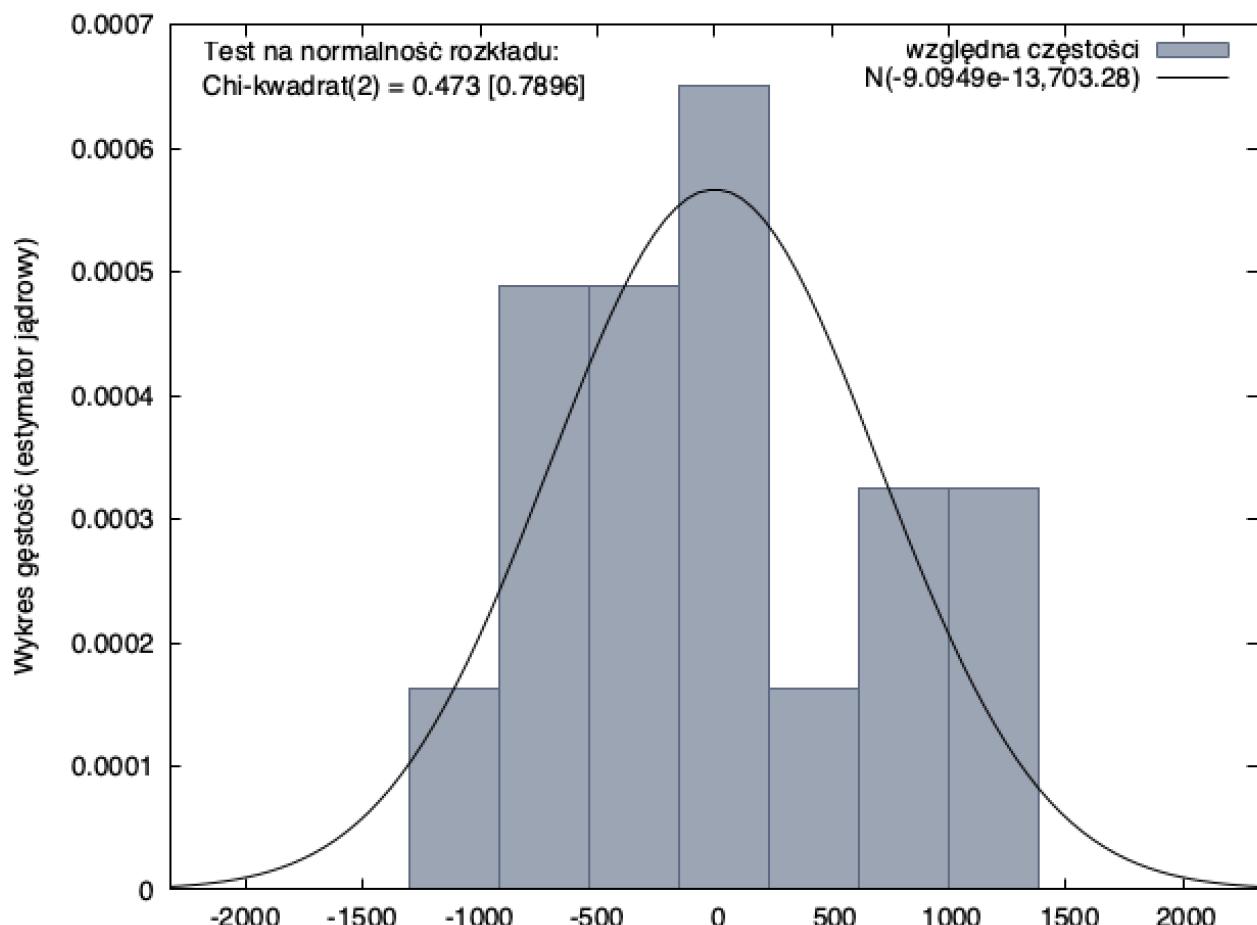
Test na normalność rozkładu uhat40:

Test Doornika-Hansena = 0.472532, z wartością p 0.789571

Test Shapiro-Wilka = 0.965641, z wartością p 0.764019

Test Lillieforsa = 0.145501, z wartością p ~ 0.48

Test Jarque'a-Bera = 0.706295, z wartością p 0.702474



Testy przyjmują hipotezę o normalności reszt.

Test chi-kwadrat potwierdza hipotezę o normalności reszt z $p = 0.7896$.

Podsumowując, ponieważ wymienione testy przyjmują hipotezę o normalności rozkładu reszt, my również przyjmujemy tę hipotezę.

Statystyka Durbina-Watsona:

Statystyka testu Durbina-Watsona dla 5% poziomu istotności, $n = 16$, $k = 1$

$$\begin{aligned} dL &= 1.1062 \\ dU &= 1.3709 \end{aligned}$$

wartość DW = 1.73076 < 2

$$DW > dU$$

Zatem przyjmujemy hipotezę o braku korelacji.

Sprawdzamy homoskedastyczność:

Test White'a na heteroskedastyczność reszt (zmienna wariancji resztowej)

Estymacja KMNK, wykorzystane obserwacje 1947–1962 ($N = 16$)

Zmienna zależna (Y): $uhat^2$

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	-2.73999e+07	3.95475e+07	-0.6928	0.5006
pop	486.192	669.565	0.7261	0.4806
sq_pop	-0.00211496	0.00282667	-0.7482	0.4676

Wsp. determ. R-kwadrat = 0.086069

Statystyka testu: $TR^2 = 1.377098$,
z wartością $p = P(\text{Chi-kwadrat}(2) > 1.377098) = 0.502304$

Przyjmujemy hipotezę, że reszty są homoskedastyczne.

METODA WPRZÓD

Model wyszedł taki sam jak dla metody wstecz, zatem wyniki również będą takie same.

4. PODSUMOWANIE

Model uzyskany metodą wprzód i wstecz oraz model uzyskany metodą grafów mają różną ilość zmiennych, dlatego do ich porównania wykorzystamy metodę Akaike'a. Wybieramy model dla którego otrzymamy najmniejszą wartość:

$$AIC = -2 \sum_j \ln(\pi_j) + 2q.$$

gdzie:

π_j oznacza estymowane prawdopodobieństwo uzyskania wartości obserwacji j jako była naprawdę uzyskana,

q oznacza liczbę parametrów modelu.

Do obliczenia kryterium informacyjnego Akaike'a wykorzystamy program gretl.

Dla modelu:

$$unemp = -77834 + 211.834(prdefl) - 0.068(gnp) + 0.73(pop) + \epsilon$$

uzyskaliśmy wartość kryterium Akaike'a równą 235.2948.

Dla modelu:

$$unemp = -7636.65 + 0.092(pop) + \epsilon$$

uzyskaliśmy wartość kryterium Akaike'a równą 257.0536.

Zatem według kryterium Akaike'a model uzyskany metodami wprzód i wstecz jest lepszy od modelu uzyskanego metodą grafów.