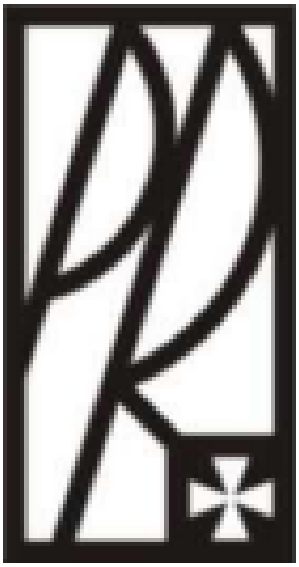


Politechnika Rzeszowska im. Ignacego Łukasiewicza
Wydział Matematyki i Fizyki Stosowanej

PRACA PROJEKTOWA

Projektowanie modeli łączenia źródeł danych

Kośniowska Katarzyna 173162
Kosiorowski Filip 173161
Inżynieria i Analiza Danych, semestr 5, grupa 2

Prowadzący: mgr Jarosław Napora

Rzeszów 2025

1. Opis zadania

Celem projektu jest zbudowanie modelu ekonometrycznego prognozującego współczynnik rozwoju w Bułgarii według HDI. Model ekonometryczny, wyrażony za pomocą równania lub układu równań, pozwala na zrozumienie mechanizmów zmian zachodzących w analizowanym obszarze. Opisuje on zależności między różnymi wielkościami ekonomicznymi, stanowiąc formalny matematyczny zapis występujących prawidłowości ekonomicznych.

Aby stworzyć model ekonometryczny, konieczne jest nie tylko solidne zrozumienie teorii ekonomii oraz wiedza matematyczno-ekonomiczna, ale także znajomość praktycznych aspektów funkcjonowania gospodarki. Model ten powinien mieć zarówno wartość teoretyczną, istotną z perspektywy nauki o ekonomii, jak i praktyczne zastosowanie, umożliwiając wykorzystanie go jako narzędzia do prognozowania w przyszłości.

2. Użyte dane

Wskaźnik Rozwoju Społecznego (HDI)

HDI to miernik opracowany przez Program Narodów Zjednoczonych ds. Rozwoju, który pokazuje średni poziom rozwoju społecznego w trzech kluczowych obszarach:

1. **Zdrowie** – oceniane na podstawie oczekiwanej długości życia przy narodzinach.
2. **Edukacja** – mierzone średnią liczbą lat edukacji dorosłych (25+) i oczekiwanym czasem edukacji dla dzieci.
3. **Standard życia** – mierzony dochodem narodowym brutto (GNI) na osobę (logarytmicznie, by uwzględnić mniejsze znaczenie wzrostu dochodu w wyższych wartościach).

HDI jest średnią geometryczną tych trzech wymiarów. Służy m.in. do porównywania polityk krajowych i analizowania, dlaczego kraje o podobnym dochodzie mogą mieć różne wyniki w rozwoju społecznym.

Oto zmienne, które użyjemy po to by prognozować HDI:

Y	HDI
X1	life_expectancy
X2	expec_yr_school
X3	mean_yr_school
X4	gross_inc_percap
X5	gender_development
X6	gender_inequality
X7	seats_in_parliament_f_%
X8	co2_emission_tons
X9	mat_footprint_percap_tons

Zebrane dane prezentują się następująco:

year	hdi	life_expe	expec_yr_	mean_yr_	gross_inc	gender_d	gender_in	seats_in_	co2_emis	mat_footprint
1990	0,698	71,357	12,29901	6,814899	14970,04	0,977	0,362	10,83333	8,740869	15,6424
1991	0,698	71,227	12,3546	7,088468	13790,48	0,974	0,363	10,83333	7,046399	9,1664
1992	0,7	71,215	12,12902	7,37302	14087,79	0,978	0,357	10,83333	6,610859	9,0196
1993	0,702	71,226	12,01152	7,668995	13951,33	0,981	0,355	10,83333	6,777421	8,9584
1994	0,704	70,901	11,88076	7,964969	14269,3	0,983	0,359	10,83333	6,593815	8,7761
1995	0,7	71,002	12,17437	8,260943	11111,07	0,984	0,354	10,83333	6,830842	9,2547
1996	0,709	70,874	12,51595	8,556918	11657,45	0,984	0,332	10,83333	6,958396	6,6734
1997	0,704	70,351	12,61157	8,852892	10111,59	0,981	0,337	10,83333	6,708658	7,522
1998	0,714	70,912	12,7276	9,148867	10677,56	0,983	0,327	10,83333	6,449978	9,0083
1999	0,718	71,596	12,89876	9,444841	9879,275	0,98	0,342	10,83333	5,674687	9,8746
2000	0,723	71,609	12,80523	9,740815	10279,6	0,978	0,333	10,83333	5,608204	10,2276
2001	0,732	71,915	12,79557	10,03679	11147,52	0,98	0,275	26,25	6,088706	11,3048
2002	0,743	72,162	13,19983	10,11272	12269,36	0,981	0,254	26,25	5,780627	11,2625
2003	0,753	72,408	13,66489	10,18865	12963,15	0,983	0,245	26,25	6,374762	11,2419
2004	0,76	72,609	13,87142	10,26458	13856,39	0,985	0,239	26,25	6,296734	12,4473
2005	0,765	72,592	14,02011	10,34051	14824,54	0,985	0,241	22,08333	6,474315	12,2874
2006	0,77	72,722	14,08424	10,41644	15543,42	0,989	0,231	22,08333	6,674782	13,7887
2007	0,776	73,037	14,25047	10,54241	15884,66	0,983	0,226	21,66667	7,213816	14,2533
2008	0,782	73,365	14,18345	10,58647	17416,56	0,987	0,227	21,66667	7,038842	16,7223
2009	0,784	73,705	14,23392	10,64169	17220,16	0,988	0,234	20,83333	5,99362	14,8956
2010	0,79	73,835	14,48367	10,72172	17799,12	0,991	0,229	20,83333	6,296981	10,8575
2011	0,794	74,189	14,58668	10,78185	18085,93	0,991	0,224	20,83333	7,042806	11,758
2012	0,798	74,357	14,5694	10,89634	18672,66	0,991	0,218	22,91667	6,448689	11,6906
2013	0,804	74,849	14,95896	11,0063	18511,18	0,992	0,213	24,58333	5,730823	14,2515
2014	0,807	74,482	15,17376	11,05153	19048,31	0,992	0,225	20	6,126009	15,1754
2015	0,809	74,632	15,07862	11,18832	19375,69	0,994	0,218	20,41667	6,583573	16,4428
2016	0,81	74,834	14,8293	11,23244	20093,46	0,994	0,216	20,41667	6,256839	14,2974
2017	0,81	74,799	14,6452	11,2427	20986,68	0,994	0,208	23,75	6,6037	14,9347
2018	0,811	74,898	14,49725	11,29953	21476,7	0,992	0,21	23,75	6,10889	14,6011
2019	0,813	75,062	14,25417	11,35635	22590,95	0,994	0,205	25,83333	5,988303	13,4022
2020	0,802	73,645	14,03667	11,41318	21652,93	0,997	0,204	26,66667	5,234659	14,1809
2021	0,796	71,798	13,86803	11,41318	23725,05	0,996	0,209	23,75	6,140388	15,9277
2022	0,799	71,528	13,86803	11,41318	25920,8	0,995	0,206	24,16667	6,140388	16,5341

Objaśnienie współczynników:

X1 – Przewidywana długość życia dla obu płci
X2 – przewidywana długość okresu szkolnego (w latach)
X3 – średnie lata nauki
X4 - Dochód narodowy brutto na mieszkańca
X5- Wskaźnik rozwoju/równości płci
X6 – Wskaźnik nierówności płci
X7 - Udział kobiet w parlamencie (% zajmowanych przez kobiety)
X8 - Emisje dwutlenku węgla na mieszkańca (produkcja) (w tonach)
X9 - Ilość surowców naturalnych zużywanych przez jedną osobę w danym kraju (w tonach)

3. Selekcja zmiennych

	X1	X2	X3	X4	X5	X6	X7	X8	X9
Średnia	72,71797	13,62309	9,971591	16177,29	0,986576	0,266	19,10354	6,443587	12,31458182
Odchylenie	1,474255	0,996575	1,371811	4245,924	0,006252	0,059706	6,125546	0,613181	2,805239801
Współczynnik zmienności	2,03%	7,32%	13,76%	26,25%	0,63%	22,45%	32,06%	9,52%	22,78%
Wartość krytyczna	20%								

Używając metody quasi-stałych obliczamy współczynnik zmienności. Pozwoli nam wybrać wstępnie, które ze zmiennych będą uwzględnione w modelu. Dla wartości krytycznej równej 20% wybrane zostaną zmienne:

X4 - Dochód narodowy brutto na mieszkańca
X6 - Wskaźnik nierówności płci
X7 - Udział kobiet w parlamencie (% zajmowanych przez kobiety)
X9- Ilość surowców naturalnych zużywanych przez jedną osobę w danym kraju (w tonach)

Po selekcji zmiennych model ma postać:

$$Y = a_0 + a_1X_4 + a_2X_6 + a_3X_7 + a_4X_9$$

4. Metoda Hellwiga

	Combination <chr>	H_Value <dbl>
2	X6	0.9298366
3	X4, X6	0.9251260
7	X4, X6, X7	0.8994955
15	X4, X6, X7, X9	0.8862236
11	X4, X6, X9	0.8825799
10	X6, X9	0.8692413
14	X6, X7, X9	0.8467592
5	X4, X7	0.8372582
13	X4, X7, X9	0.8266553
6	X6, X7	0.8203877

Metoda ta pozwala wyznaczyć syntetyczny miernik rozwoju, który umożliwia porównanie różnych obiektów. Opiera się na obliczeniu odległości każdego obiektu od wzorca idealnego – czyli hipotetycznego obiektu o najlepszych możliwych wartościach wszystkich analizowanych zmiennych.

Korzystając z metody Hellwiga możemy wybrać zmienne objaśniające, które powinny cechować się silną korelacją ze zmienną objaśnianą, a także słabą korelacją między sobą. W wierszu numer 15 mamy wzięte pod uwagę wszystkie 4 zmienne co daje nam prawie najwyższy wynik. Większa wartość ma jeden wiersz wyżej (7), jednakże różnica jest na tyle niewielka, że zdecydowaliśmy się wziąć tą z większą ilością zmiennych. Pomimo faktu, że wiersz 2 i 3 mają tą wartość jeszcze wyższą nie braliśmy akurat tych wartości, by model miał więcej zmiennych.

5. Estymacja parametrów

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotı	0,98713
R kwadrat	0,97442
Dopasowa	0,97076
Błąd stand	0,0074
Obserwac	33

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	4	0,05844	0,01461	266,62	7,5E-22
Resztkowy	28	0,00153	5,5E-05		
Razem	32	0,05997			

	Współczynnił	*standardc*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*olne 95,0%*	*órne 95,0%*
Przecięcie	1,00092	0,03401	29,4303	1,3E-22	0,93125	1,07058	0,93125	1,07058
X4	1,6E-06	5,5E-07	2,88897	0,00738	4,6E-07	2,7E-06	4,6E-07	2,7E-06
X6	-0,82962	0,06944	-11,948	1,7E-12	-0,97186	-0,68739	-0,97186	-0,68739
X7	-0,00248	0,00054	-4,61433	8E-05	-0,00358	-0,00138	-0,00358	-0,00138
X9	9,8E-05	0,00077	0,12763	0,89936	-0,00148	0,00167	-0,00148	0,00167

6. Testy i weryfikacja modelu

```
> cor(dane, Y, method = "pearson")
          [,1]
X4  0.8352585
X6 -0.9642804
X7  0.7970524
X9  0.7657708
```

Badanie istotności korelacji to test statystyczny, który sprawdza, czy związek (korelacja) między wszystkimi zmiennymi objaśniającymi w próbie jest statystycznie istotny, czyli czy można go uogólnić na całą populację, czy może wynika on jedynie z przypadku. Współczynnik X7 ma zły znak, dlatego wykluczamy go z modelu.

```
> Y <- as.vector(bulgaria$Y)
> wynik <- lm( Y ~ X4 + X6 + X9)
> wsp_wyr <- sd(wynik$residuals)/mean(Y) * 100
> wsp_wyr
[1] 1.208959
```

Badanie wyrazistości modelu (inaczej analiza dopasowania modelu) to proces oceny, jak dobrze dany model statystyczny lub ekonometryczny opisuje rzeczywistość i jak trafnie wyjaśnia zmienność badanych danych. Współczynnik zmienności losowej wynosi < 5%, to znaczy, że dane są bardzo stabilne.

```
> summary(wynik)

Call:
lm(formula = Y ~ X4 + X6 + X9)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018159 -0.005719 -0.001550  0.006022  0.017190

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.681e-01  2.360e-02  36.782  < 2e-16 ***
X4           2.437e-06  6.798e-07   3.584  0.00122 **
X6          -5.560e-01  4.708e-02 -11.808 1.34e-12 ***
X9           2.812e-05  1.003e-03   0.028  0.97782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009651 on 29 degrees of freedom
Multiple R-squared:  0.955,     Adjusted R-squared:  0.9503
F-statistic:   205 on 3 and 29 DF,  p-value: < 2.2e-16
```

Badanie istotności parametrów modelu to proces oceny, czy zmienne niezależne w modelu regresji mają rzeczywisty wpływ na zmienną zależną, czy ich wpływ jest przypadkowy. Każda zmienna objaśniająca została poddana testowi istotności statystycznej, który ocenia, czy jej wpływ na Y jest istotny.

- **Zmienna X4** ma współczynnik **2.437e-6** oraz wartość p równą **0.00122**, co oznacza, że jest **statystycznie istotna** ($p<0.05$). Wskazuje to, że X4 ma istotny wpływ na Y.
- **Zmienna X6** ma współczynnik **-0.556** i wartość p wynoszącą **1.34e-12**, co również wskazuje na jej **bardzo wysoką istotność statystyczną** ($p<0.05$). Oznacza to, że X6 jest silnym współczynnikiem zmiennej zależnej.
- **Zmienna X9** ma współczynnik **2.812e-5** i wartość p wynoszącą **0.97782**, co sugeruje, że **nie jest istotna statystycznie** ($p>0.05$). W związku z tym jej wpływ na Y jest znikomy stąd wyrzucamy ją w modelu.
- **Współczynnik determinacji R^2** wynosi **0.955**, co oznacza, że model wyjaśnia **95,5%** wariancji zmiennej zależnej Y. Jest to bardzo wysoka wartość, co wskazuje na dobre dopasowanie modelu do danych
- **Statystyka F**: Model osiągnął wartość **205**, co oznacza, że jako całość jest bardzo dobrze dopasowany.
- **Wartość p dla modelu** wynosi **< 2.2e-16**, co oznacza, że model jako całość jest statystycznie istotny

Po usunięciu X9 powtarzamy badanie:

```
Call:
lm(formula = Y ~ X4 + X6)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018154 -0.005564 -0.001596  0.006021  0.017224

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.684e-01  1.975e-02  43.968  < 2e-16 ***
X4           2.445e-06  5.977e-07   4.091 0.000297 ***
X6          -5.565e-01  4.251e-02 -13.092 6.15e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009489 on 30 degrees of freedom
Multiple R-squared:  0.955,     Adjusted R-squared:  0.952
F-statistic: 318.1 on 2 and 30 DF,  p-value: < 2.2e-16
```

Możemy zauważyć, że statystyka F zwiększyła się do 318.1, co możemy rozumieć jako poprawę istotności.

Nowy model ma postać:

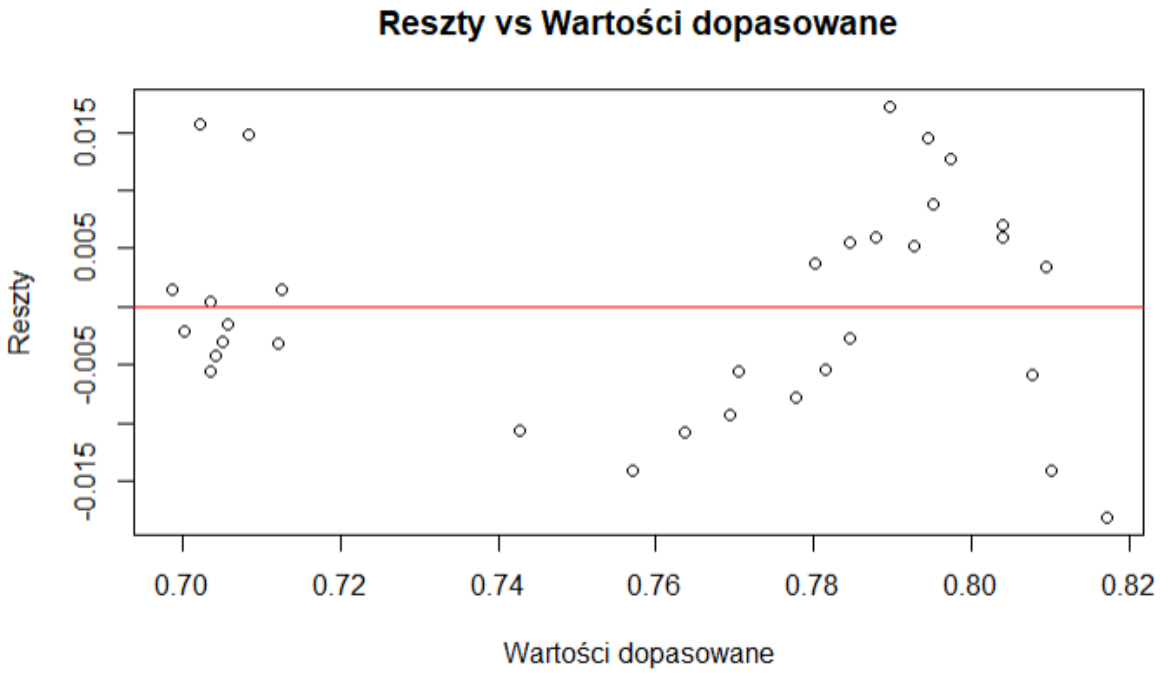
$$Y = 0{,}868406898662379 + 2{,}44529098829757E - 06X_4 + - \\ - 0{,}556487583052453X_6$$

Badanie autokorelacji składnika losowego polega na analizie zależności między wartościami tego składnika w różnych okresach czasu. Jest to ważne w kontekście modelowania szeregów czasowych, ponieważ pozwala określić, czy występuje systematyczna zależność pomiędzy błędami prognoz w różnych momentach. Wysoka autokorelacja może sugerować, że model nie uchwycił wszystkich istotnych zależności, co może wymagać jego poprawy. Badanie to jest kluczowe w celu oceny jakości modelu ekonometrycznego i wykrywania ewentualnych problemów, takich jak heteroskedastyczność czy nieodpowiednia struktura modelu.

Aby sprawdzić hipotezę o braku autokorelacji reszt w modelu, przeprowadzono test Durbina-Watsona.
H0: brak autokorelacji w resztach modelu
H1: obecność istotnej autokorelacji w resztach modelu

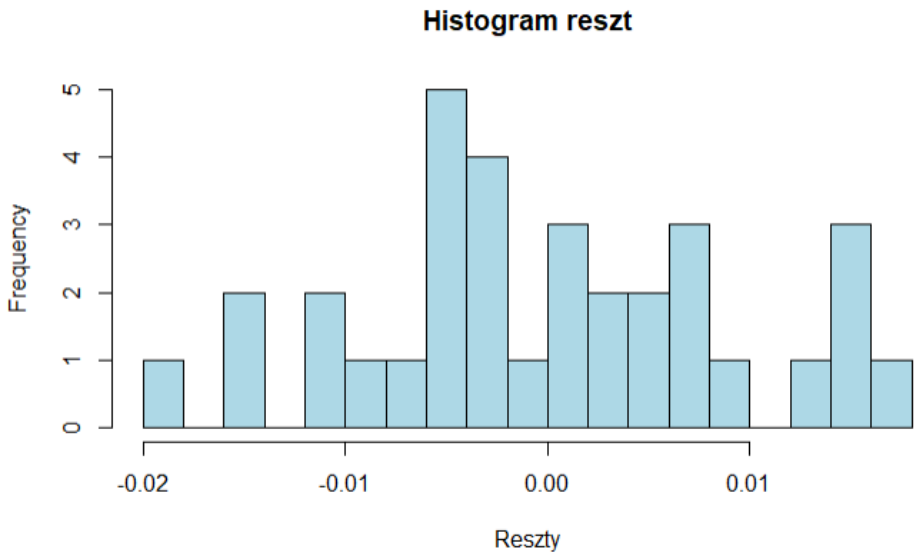
```
        Durbin-Watson test

data:  nowy_model
DW = 0.49531, p-value = 3.015e-09
alternative hypothesis: true autocorrelation is greater than 0
```

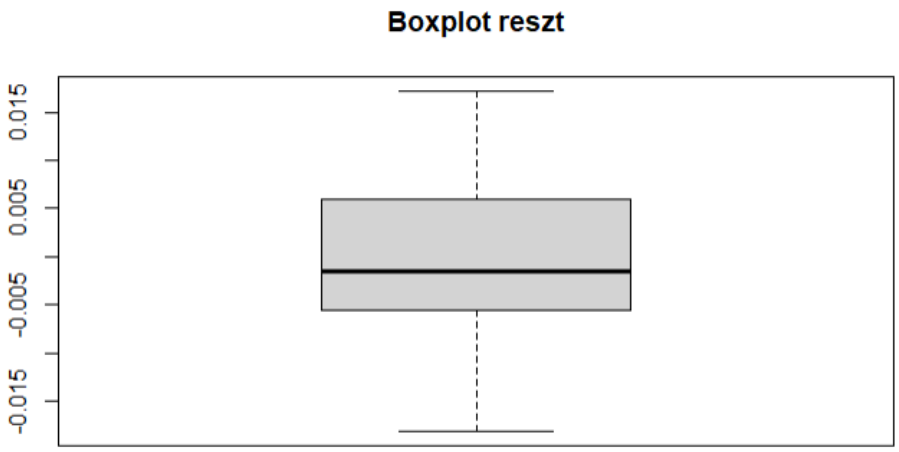


Badanie składnika losowego w modelu regresji liniowej odnosi się do analizy reszt (błędów) modelu, czyli różnicy między wartościami rzeczywistymi a przewidywanymi przez model. Jest to kluczowy krok w diagnostyce modelu, ponieważ pozwala ocenić, czy założenia klasycznej regresji liniowej są spełnione.

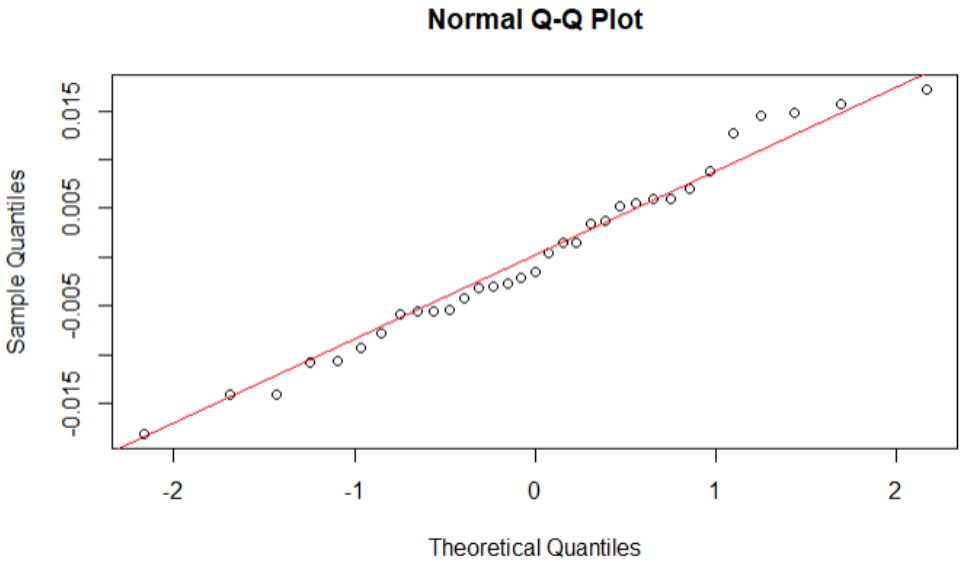
W dobrze dopasowanym modelu wartości reszt powinny oscylować wokół zera, co można zauważyć na powyższym i poniższym wykresie.



Średnia jest minimalnie poniżej zera dzięki czemu wiemy, że model jest prawidłowy.



Wykres pokazuje rozkład reszt modelu. Większość reszt znajduje się w wąskim przedziale blisko zera, a nie widać żadnych wartości odstających. Świadczy to o stabilnym rozkładzie reszt, co jest korzystne dla modelu.



Wykres Q-Q porównuje kwantyle teoretyczne rozkładu normalnego z kwantylami reszt. Punkty układają się wzdłuż linii prostej, co sugeruje, że reszty są zgodne z rozkładem normalnym. Jedynie na końcach widać niewielkie odchylenia, które nie powinny znacząco wpłynąć na założenia modelu regresji.

Test Shapiro-Wilka jest statystycznym testem służącym do weryfikacji, czy dane pochodzą z rozkładu normalnego. Jest to jeden z najczęściej stosowanych testów normalności, szczególnie dla małych i średnich prób. Hipoteza zerowa (H0) w tym teście zakłada, że dane są rozkładane normalnie, natomiast hipoteza alternatywna (H1) sugeruje, że dane nie mają rozkładu normalnego. Wynik testu (statystyka W) jest porównywany z wartością krytyczną. Jeśli wynik jest niski, odrzucamy hipotezę o normalności rozkładu.

```
        Shapiro-Wilk normality test

data:  residuals(nowy_model)
W = 0.97719, p-value = 0.6986
```

Przyjęta wartość krytyczna z tablicy rozkładu Shapiro-Wilka to 0.9869, wartość funkcji wynosi 0.97719 także odrzucamy hipotezę H0 na rzecz hipotezy H1.

Badanie symetrii reszt w modelu ekonometrycznym jest istotnym elementem oceny jakości i poprawności modelu. Składnik losowy odzwierciedla różnicę między rzeczywistą wartością a wartością prognozowaną przez model, uwzględniającą zmienne objaśniające w danym okresie. Idealnie, nasz model powinien jak najlepiej odwzorować rzeczywistość, co oznacza, że różnica między wartością rzeczywistą a prognozowaną powinna być minimalna i zbliżona do zera. W modelu składniki losowe powinny mieć rozkład normalny, charakteryzujący się symetrią. Celem testu jest sprawdzenie, czy stosunek dodatnich i ujemnych reszt wynosi 0,5.

$$H_0: \frac{m}{n} = \frac{1}{2}$$

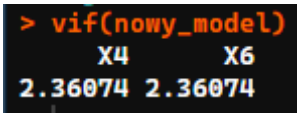
$$H_1: \frac{m}{n} \neq \frac{1}{2}$$

```
> nowy_model$residuals
           1            2            3            4            5            6            7            8            9           10
-0.0055644854 -0.0021236428 -0.0041895812 -0.0029688726  0.0004795535  0.0014198986 -0.0031588724 -0.0015963567  0.0014547959  0.0157541520
          11           12           13           14           15           16           17           18           19           20
 0.0147668589 -0.0106317411 -0.0140612050 -0.0107661029 -0.0092892824 -0.0055437140 -0.0078664593 -0.0054833249 -0.0026727863  0.0037028822
          21           22           23           24           25           26           27           28           29           30
 0.0055047411  0.0060209501  0.0052473176  0.0088597292  0.0172241431  0.0145281952  0.0126600718  0.0060239803  0.0069387143  0.0034316114
          31           32           33
-0.0058311564 -0.0141156473 -0.0181543644
>
> m <- 16
> n <- 33
>
> t <- (abs(m/n - 1/2))/(sqrt((m/n * (1 - m/n)) / (n - 1)))
> t
[1] 0.1714986
>
```

Statystyka T wynosi 2,0369. Obszar krytyczny testu jest obustronny $(-\infty, -2{,}0369] \cup [2{,}0369, +\infty)$, więc nie ma podstaw do odrzucenia hipotezy zerowej.

Współczynnik VIF (Variance Inflation Factor) to miara stosowana w analizie regresji w celu wykrycia wielokrotnej kolinearity między zmiennymi objaśniającymi (niezależnymi) w modelu. Określa on, jak bardzo wariancja oszacowanego współczynnika regresji jest "zwiększona" przez korelację między zmiennymi.

Wartości VIF wskazują na to, jak bardzo dana zmienna jest skorelowana z innymi zmiennymi w modelu. Jeśli VIF dla zmiennej jest wysoki (zwykle powyżej 10), oznacza to, że zmienna ta jest silnie skorelowana z innymi zmiennymi, co może prowadzić do problemów z wielokrotną kolinearną (np. zawyżenia oszacowanych współczynników regresji i trudności w ich interpretacji). Jeśli VIF jest niski (bliższy 1), oznacza to, że zmienna jest w dużym stopniu niezależna od pozostałych zmiennych w modelu. Wartości VIF mieszczące się w przedziale $1 < VIF < 5$ uznawane są za akceptowalne, co oznacza, że problem współliniowości zazwyczaj nie występuje w takim przypadku.



Wartości współczynnika VIF dla zmiennych w modelu mieszczą się w akceptowalnym przedziale poniżej 5, co oznacza, że problem wielokrotnej kolinearności nie występuje i zmienne są wystarczająco niezależne od siebie.

Test serii to statystyczna metoda służąca do sprawdzania losowości szeregu danych, np. reszt modelu regresji. Pomaga ocenić, czy obserwowane wartości są niezależne od siebie, czyli czy nie ma w nich ukrytego wzorca.

Rekord	reszta	seria
1	-0,005564485	1
2	-0,002123643	1
3	-0,004189581	1
4	-0,002968873	1
5	0,000479554	2
6	0,001419899	2
7	-0,003158872	3
8	-0,001596357	3
9	0,001454796	4
10	0,015754152	4
11	0,014766859	4
12	-0,010631741	5
13	-0,014061205	5
14	-0,010766103	5
15	-0,009289282	5
16	-0,005543714	5
17	-0,007866459	5
18	-0,005483325	5
19	-0,002672786	5
20	0,003702882	6
21	0,005504741	6
22	0,00602095	6
23	0,005247318	6
24	0,008859729	6
25	0,017224143	6
26	0,014528195	6
27	0,012660072	6
28	0,00602398	6
29	0,006938714	6
30	0,003431611	6
31	-0,005831156	7
32	-0,014115647	7
33	-0,018154364	7

```
        Runs Test

data:  as.factor(sign(residuals(nowy_model)))
Standard Normal = -3.7123, p-value = 0.0002054
alternative hypothesis: two.sided
```

Hipotezy testu serii

- **H_0:** Kolejność wartości jest losowa.
- **H_1:** Występuje pewien wzorzec (autokorelacja, trend itp.).

Według wyników powyżej, wartość p < 0,05 to odrzucamy hipotezę H0 na rzecz hipotezy H1, występuje więc tu pewien trend.

Badanie autokorelacji składnika losowego polega na analizie zależności między wartościami tego składnika w różnych okresach czasu. Jest to ważne w kontekście modelowania szeregów czasowych, ponieważ pozwala określić, czy występuje systematyczna zależność pomiędzy błędami prognoz w różnych momentach. Wysoka autokorelacja może sugerować, że model nie uchwycił wszystkich istotnych zależności, co może wymagać jego poprawy. Badanie to jest kluczowe w celu oceny jakości modelu ekonometrycznego i wykrywania ewentualnych problemów, takich jak heteroskedastyczność czy nieodpowiednia struktura modelu.

Aby sprawdzić hipotezę o braku autokorelacji reszt w modelu, przeprowadzono test Durbina-Watsona.

- **H_0:** brak autokorelacji w resztach modelu
- **H_1:** obecność istotnej autokorelacji w resztach modelu

```
        Durbin-Watson test

data:  nowy_model
DW = 0.49531, p-value = 3.015e-09
alternative hypothesis: true autocorrelation is greater than 0
```

Heteroskedastyczność to zjawisko w analizie regresji, w którym zmienność składnika losowego (reszt) modelu nie jest stała w różnych punktach danych. Oznacza to, że rozrzut błędów modelu zmienia się w zależności od wartości zmiennych niezależnych. W przypadku heteroskedastyczności wyniki modelu regresji mogą być zniekształcone, co prowadzi do błędnych wniosków dotyczących istotności współczynników oraz nieefektywności oszacowań. W takim przypadku często stosuje się testy diagnostyczne (np. test Breuscha-Pagana) oraz metody korygujące, jak heteroskedastyczność-robust standard errors, aby poprawić wyniki analizy.

H_0: występuje homoskedastyczność

H_1: występuje heteroskedastyczność

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

          Data
-----------------------------
Response : Y
Variables: fitted values of Y

       Test Summary
-----------------------------
DF            =    1
Chi2          =    1.479049
Prob > Chi2   =    0.2239234
>
```

Wartość prób jest większa od Chi2, tak więc nie ma podstaw do odrzucenia hipotezy zerowej.

Mierniki dopasowania modelu:

```
> r2 <- summary(nowy_model)$r.squared
> r2_adj <- summary(nowy_model)$adj.r.squared
> rse <- summary(nowy_model)$sigma
> reszty <- residuals(nowy_model)
> mae <- mean(abs(reszty))
> mse <- mean(reszty^2)
> rmse <- sqrt(mse)
>
> cat("Mierniki dopasowania modelu:\n")
Mierniki dopasowania modelu:
> cat("R²:", r2, "\n")
R²: 0.9549618
> cat("Skorygowany R²:", r2_adj, "\n")
Skorygowany R²: 0.9519593
> cat("Błąd standardowy reszt (RSE):", rse, "\n")
Błąd standardowy reszt (RSE): 0.009488792
> cat("Średni błąd absolutny (MAE):", mae, "\n")
Średni błąd absolutny (MAE): 0.007516218
> cat("Średni błąd kwadratowy (MSE):", mse, "\n")
Średni błąd kwadratowy (MSE): 8.185197e-05
> cat("RMSE:", rmse, "\n")
RMSE: 0.009047208
```

Model regresji wykazuje bardzo dobre dopasowanie do danych, o czym świadczy wysokie R^2 na poziomie 0,95 oraz skorygowany R^2 wynoszący 0,95. Błąd standardowy reszt (RSE) jest niski i wynosi 0,0095, co oznacza niewielkie odchylenie wartości rzeczywistych od przewidywanych. Ponadto, pozostałe miary błędu, takie jak MAE (0,0075) i RMSE (0,0090), również wskazują na wysoką precyzję modelu. Model można uznać za dobrze dopasowany i efektywny w prognozowaniu zmiennej zależnej.

Test istotności całego modelu (Test F) ocenia, czy co najmniej jedna zmienna objaśniająca w modelu regresji ma istotny wpływ na zmienną zależną. Hipoteza zerowa (H_o) zakłada, że wszystkie współczynniki regresji są równe zero, co oznacza brak wpływu zmiennych objaśniających na zmienną zależną.

```
> test_F <- (wsp_r)/(1-wsp_r)*(33-2)/2
> test_F
[1] 328.6524
```

Wyniki:

- **Obliczona wartość F:** 328,65
- **Wartość krytyczna F*** (z tablic dla poziomu istotności 0,05, $df_1 = 2$, $df_2 = 30$): 3,316

df_1 = liczba zmiennych objaśniających, df_2 = stopnie swobody reszt

Ponieważ obliczona wartość F jest znacznie większa od wartości krytycznej F*, odrzucamy hipotezę zerową. Oznacza to, że model regresji jako całość jest statystycznie istotny i przynajmniej jedna zmienna objaśniająca wpływa na zmienną zależną.

Testowanie pojedynczych zmiennych objaśniających to proces oceny, czy dana zmienna objaśniająca ma istotny wpływ na zmienną objaśnianą w modelu ekonometrycznym. Weryfikacja ta odbywa się za pomocą testu istotności współczynnika regresji, zwykle z wykorzystaniem statystyki t-Studenta. Celem jest sprawdzenie, czy współczynnik przy danej zmiennej objaśniającej różni się istotnie od zera. Jeśli współczynnik jest bliski zeru i nieistotny statystycznie, oznacza to, że dana zmienna może nie mieć znaczącego wpływu na zmienną zależną i może zostać usunięta z modelu.

- H_o: $a_i = 0$ - dana zmienna nie wpływa istotnie na zmienną objaśnianą.
- H_1: $a_i \neq 0$ - dana zmienna wykazuje statystycznie istotny wpływ.

	Współczynniki	*Błąd standardowy*
Przecięcie	0,868406899	0,019750739
X4	2,45E-06	5,97731E-07
X6	-0,556487583	0,042507067

n	k	r
33	2	33-(2+1) = 30
	X4	X6
statystyka t	4,09E+00	13,09164854
T = 2,0423		

Wyniki:

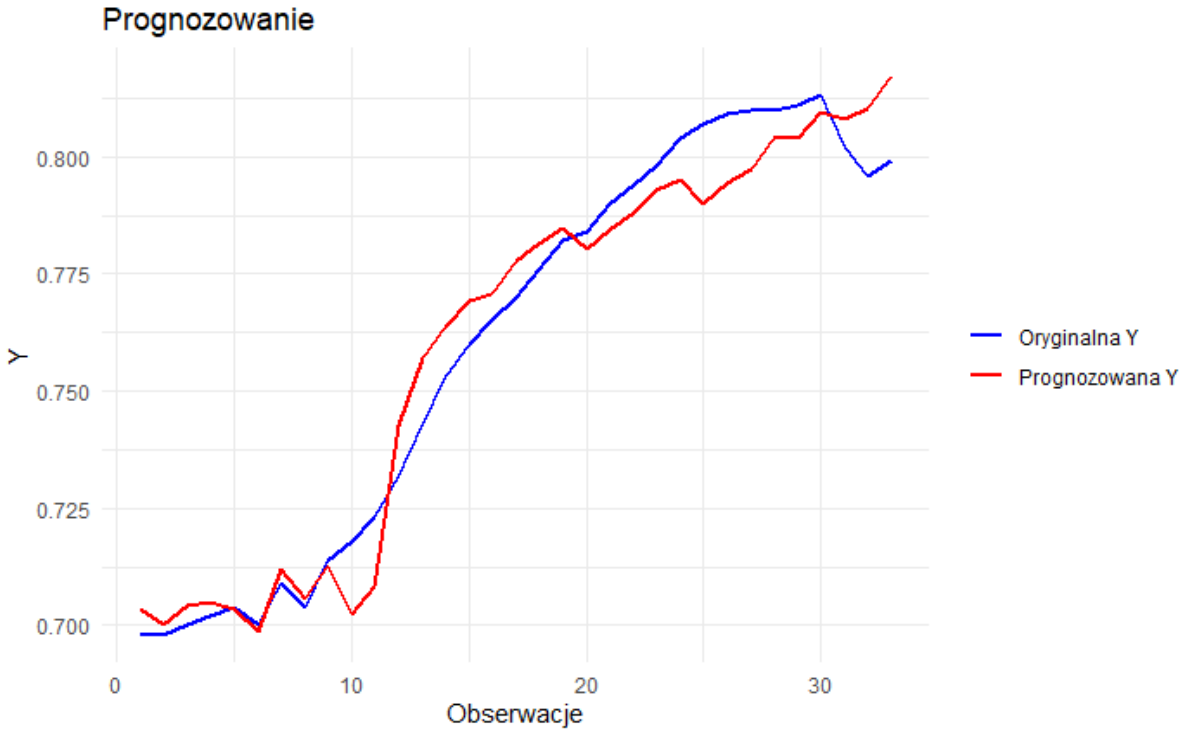
- **Liczba obserwacji (n):** 33
- **Liczba zmiennych objaśniających (k):** 2
- **Stopnie swobody reszt (r):** 30
- **Wartość krytyczna t (dla poziomu istotności 0,05):** 2,0423

Dla zmiennych X4 i X6 uzyskano następujące wartości statystyki t:

- **X4:** t = 4,09
- **X6:** t = 13,09

Statystyka t dla zmiennych X4 i X6 jest większa od wartości krytycznej 2,0423. W związku z tym odrzucamy hipotezę zerową (H_0) dla obu zmiennych. Zarówno zmienna X4, jak i X6 są istotne statystycznie na poziomie istotności 0,05. Obie zmienne mają znaczący wpływ na zmienną zależną i powinny pozostać w modelu.

7. Pełna prognoza



Według danych ze strony https://bti-project.org/en/reports/country-report/BGR wskaźnik HDI w roku 2024 w Bułgarii wynosi 7.42, przez co widać, że model przeszacował wartość (Wykres numer 2, część z czerwonymi kropkami to prognozy). Jednakże patrząc na wykres 1 można wysnuć wniosek, że model jest całościowo spójny (przeszacowania i niedoszacowania w większości się zerują).