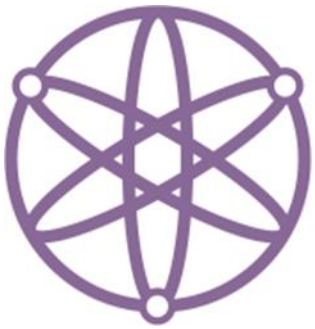


WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Projektowanie modeli łączenia źródeł danych

Dokumentacja projektu

Gabriel Lichacz

Rzeszów, 2022

Spis treści

1. Dane

Wszystkie dane zostały pobrane ze strony banku danych lokalnych. Jako wartość y przyjmuję liczbę lokali mieszkalnych ogółem w Polsce w latach 2008-2020. Chcę przewidzieć zmianę ich liczby w najbliższych latach.

Liczba lokali mieszkalnych ogółem w Polsce												
2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]
5 751 174	5 820 763	5 869 959	6 006 608	6 063 721	6 123 726	6 182 136	6 244 730	6 308 344	6 375 734	6 443 611	6 629 920	6 636 883

rys. 1-1 Dane

Jako zmienne objaśniające wybieram:

- Drogi ogółem – każdy lokal mieszkalny potrzebuje drogi dojazdowej
- Liczba nowo utworzonych miejsc pracy – więcej miejsc pracy oznacza mniejsze bezrobocie a więc w teorii zamożniejsze społeczeństwo
- Stopa bezrobocia – podobnie jak wyżej
- Małżeństwa zawarte – małżeństwa na ogół potrzebują miejsca do życia
- Wartość brutto środków trwałych na 1 mieszkańca – zamożność społeczeństwa
- Mieszkania rozpoczęta budowa – rozpoczęcie budowy niekoniecznie musi być jednoznaczne z jej zakończeniem
- Rynkowa sprzedaż lokali mieszkalnych – duża sprzedaż oznacza duży popyt na lokale mieszkalne
- Liczba i kwoty wypłaconych dodatków mieszkaniowych – pomoc państwa przy zakupu lokali mieszkaniowych
- Ilość miast – miasta są bardziej zabudowane niż tereny wiejskie

2. Odrzucenie danych

2.1. Wstępne odrzucenie

Na początku obliczam współczynnik zmienności i współczynnik korelacji Pearsona. Przyjmuję jako wartość krytyczną współczynnika zmienności 15%, a współczynnika korelacji Pearsona 75%.

W ten sposób odrzucam zmienne objaśniające, które nie spełniają powyższych kryteriów:

- Drogi ogółem – współczynnik zmienności mniejszy niż 15%
- Liczba nowo utworzonych miejsc pracy – oba kryteria niższe niż wartości krytyczne
- Małżeństwa zawarte – współczynnik zmienności mniejszy niż 15%
- Rynkowa sprzedaż lokali mieszkalnych – współczynnik zmienności niższy niż 15%
- Ilość miast – współczynnik zmienności mniejszy niż 15%

Wybieram do dalszej analizy:

- Wartość brutto środków trwałych na 1 mieszkańca
- Mieszkania rozpoczęta budowa
- Stopa bezrobocia
- Liczba i kwoty wypłaconych dodatków mieszkaniowych

Wartość brutto środków trwałych na 1 mieszkańca														Można wybrać			
2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020		Średnia	Odchylenie standardowe z próby	Współczynnik zmienności	Współczynnik korelacji Pearsona
[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]	[zł]		85668,51	19201,50062	0,224137214	0,993448428
58407,1	62276	65428,2	70088,7	74687,5	79597,2	84695,3	90323,9	95255,2	99703,2	104910,1	111079,3	117238,9				większy niż 15%	wieksza niż 75%

rys. 2-1 Przykład wybranej zmiennej objaśniającej

Liczba nowo utworzonych miejsc pracy													Odrzucam			
2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Średnia	Odchylenie standardowe z próby	Współczynnik zmienności	Współczynnik korelacji Pearsona
[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	[tys. msc.]	581,20	85,36917867	0,14688434	0,424143532
490,6	521,6	609,3	580,3	465	502,4	614,8	595,8	618,7	694,1	717,8	674,8	470,4			mniejszy niż 15%	mniejszy niż 75%

rys. 2-2 Przykład odrzuconej zmiennej objaśniającej

2.2. Regresja obliczona wstępnie

Następnie korzystając z pakietu Analiza danych w MS Excel obliczam regresję. Dzięki niej wiem, że wartość p dla jednej zmiennej jest zbyt wysoka. Zmiennej objaśniającej jeszcze nie odrzucam.

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*Dolne 95,0%*	*Górne 95,0%*
Przecięcie	3976974,426	595291,8925	6,680713236	0,000155776	2604228,86	5349719,991	2604228,86	5349719,991
Liczba i kwoty wypłaconych dodatków mieszkaniowych	0,031131166	0,047944719	0,649313762	0,534329353	-0,079429555	0,141691887	-0,079429555	0,141691887
Mieszkania rozpoczęta budowa	1,881176416	1,057315413	1,779200789	0,11308827	-0,556997297	4,31935013	-0,556997297	4,31935013
Stopa bezrobocia	27804,66635	11311,42823	2,458103944	0,039435184	1720,466075	53888,86662	1720,466075	53888,86662
Wartość brutto środków trwałych na 1 mieszkańca	17,24465453	2,092940354	8,239439071	3,52944E-05	12,41832542	22,07098364	12,41832542	22,07098364

rys. 2-3 Regresja obliczona wstępnie

2.3. Hellwig

Kolejny krok to użycie algorytmu Hellwiga. Z najlepszych wartości wybieram 10 najlepszą opcję. Dzięki uprzednio obliczonej regresji wiem, że jest to najbardziej optymalna do odrzucenia zmienna.

	Liczba i kwoty wypłaconych dodatków mieszkaniowych	Mieszkania rozpoczęta budowa	Stopa bezrobocia	Wartość brutto środków trwałych na 1 mieszkańca	hellwigP
1	0.0000000	0.0000000	0.0000000	0.9869398	0.9869398
2	0.4672716	0.0000000	0.0000000	0.5022594	0.9695310
3	0.9181888	0.0000000	0.0000000	0.0000000	0.9181888
4	0.3270654	0.0000000	0.2289227	0.3548408	0.9108289
5	0.3255606	0.2225328	0.0000000	0.3588754	0.9069688
6	0.0000000	0.3291599	0.0000000	0.5528795	0.8820395
7	0.0000000	0.0000000	0.3350879	0.5433615	0.8784494
8	0.2506873	0.1634304	0.1684311	0.2767289	0.8592777
9	0.4983778	0.0000000	0.3303598	0.0000000	0.8287376
10	0.0000000	0.2144483	0.2196281	0.3793807	0.8134572
11	0.4948922	0.3166983	0.0000000	0.0000000	0.8115905
12	0.3403616	0.2090882	0.2175870	0.0000000	0.7670369
13	0.0000000	0.3005723	0.3113454	0.0000000	0.6119177
14	0.0000000	0.0000000	0.6086401	0.0000000	0.6086401
15	0.0000000	0.5875801	0.0000000	0.0000000	0.5875801

rys. 2-4 Najlepsze wartości

2.4. Podsumowanie

Po odrzuceniu zmiennych obliczam regresję z pozostałych oraz tworzę model. Pozostałe zmienne to:

- Mieszkania rozpoczęta budowa
- Stopa bezrobocia
- Wartość brutto środków trwałych na 1 mieszkańca

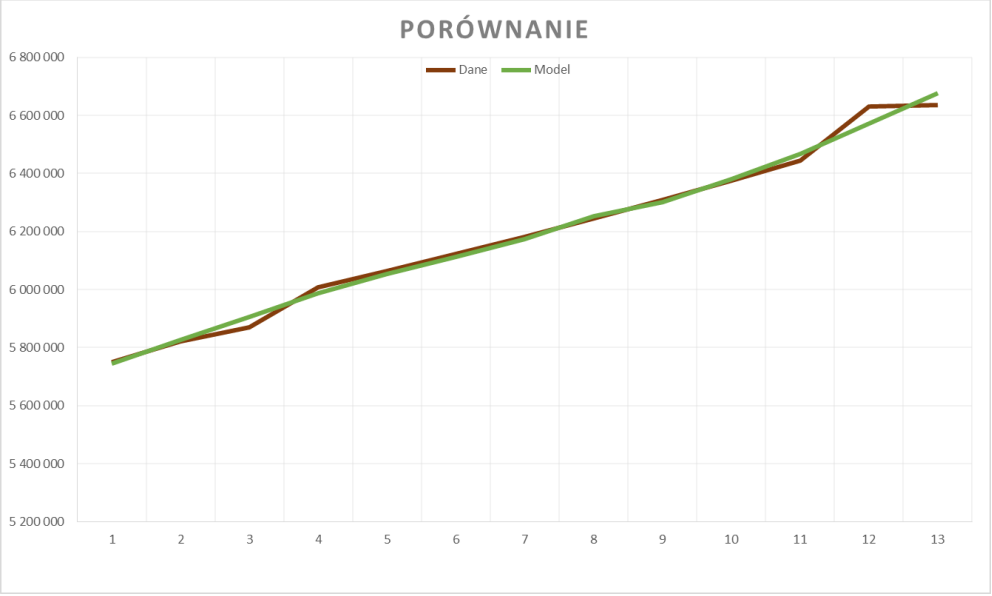
3. Model

Postać modelu:

$$Y = 1{,}44 * X_1 + 24587{,}25 * X_2 + 15{,}98 * X_3 + 4327063{,}59$$

Model	Liczba lokali mieszkalnych ogółem w Polsce
5746132	5 751 174
5826025	5 820 763
5905658	5 869 959
5988566	6 006 608
6054750	6 063 721
6112427	6 123 726
6174639	6 182 136
6252058	6 244 730
6301965	6 308 344
6379968	6 375 734
6466479	6 443 611
6572503	6 629 920
6676138	6 636 883

rys. 3-1 Wartości modelu oraz wartości oryginalne



rys. 3-2 Porównanie modelu i danych

4. Weryfikacja modelu

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotność R	0,99627713
R kwadrat	0,99256812
Dopasowany R kwadrat	0,990090827
Błąd standardowy	28655,56542
Obserwacje	13

wysokie

ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*	
Regresja	3	9,87011E+11	3,29004E+11	400,6663639	6,78265E-10	niskie
Resztkowy	9	7390272865	821141429,4			
Razem	12	9,94402E+11				

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*Dolne 95,0%*	*Górne 95,0%*
Przecięcie	4327063,586	244087,1818	17,72753306	2,62338E-08	3774900,019	4879227,152	3774900,019	4879227,152
Mieszkania rozpoczęta budowa	1,442951779	0,787305009	1,832773527	0,100053926	-0,338055887	3,223959445	-0,338055887	3,223959445
Stopa bezrobocia	24587,25191	9836,189003	2,499672577	0,033880024	2336,246505	46838,25732	2336,246505	46838,25732
Wartość brutto środków trwałych na 1 mieszkańca	15,98138525	0,74632353	21,4134817	4,9676E-09	14,29308413	17,66968637	14,29308413	17,66968637

< 0.1

rys. 4-1 Podsumowanie modelu w MS Excel

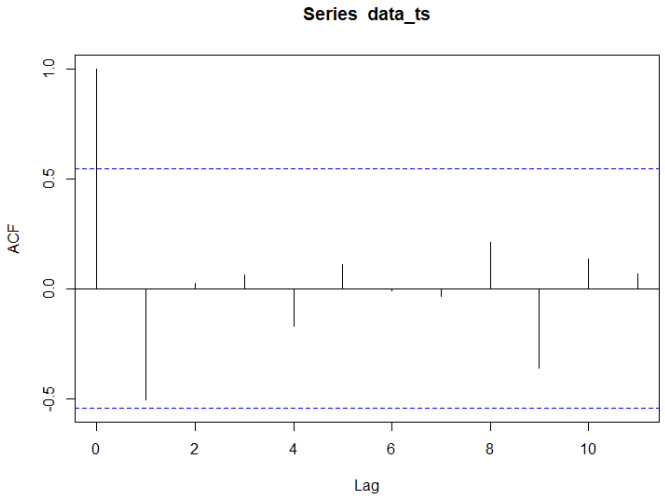
```
> summary(wynik)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-39255  -7328   5042   8971  57417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.327e+06  2.441e+05  17.728 2.62e-08 ***
x1          1.443e+00  7.873e-01   1.833   0.1001
x2          2.459e+04  9.836e+03   2.500   0.0339 *
x3          1.598e+01  7.463e-01  21.413 4.97e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28660 on 9 degrees of freedom
Multiple R-squared:  0.9926,    Adjusted R-squared:  0.9901
F-statistic: 400.7 on 3 and 9 DF,  p-value: 6.783e-10
```

rys. 4-2 Podsumowanie modelu w R

Niska wartość testu Fishera-Snedecora (F) oznacza, że model jest istotny statystycznie. Wysoka wartość R^2, czyli jakość dopasowania modelu jest wysoka. P-value mniejsze niż 0.1 dla wszystkich zmiennych, więc wszystkie są istotne statystycznie. Model jest koincydentny, tzn. znaki przy zmiennych objaśniających zgadzają się z tymi stojącymi przy obliczonych wartościach.

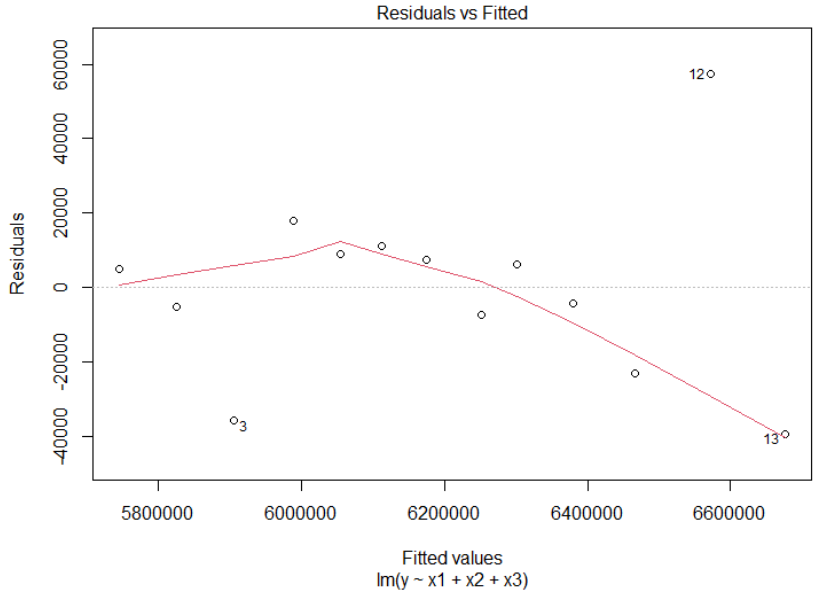
Reszty	Różnica %
5 042	0,087740414
-5 262	-0,090324596
-35 699	-0,604479853
18 042	0,301282008
8 971	0,148162207
11 299	0,184856326
7 497	0,121409858
-7 328	-0,117215491
6 379	0,101228932
-4 234	-0,06636922
-22 868	-0,353646402
57 417	0,873587626
-39 255	-0,587986471

rys. 4-3 Reszty modelu



rys. 4-4 Wykres ACF

Autokorelacja reszt nie występuje.



rys. 4-5 Reszty względem wartości dopasowanych

Reszty modelu są małe. Ich różnica w procentach również jest niska. Rozkład ujemnych reszt do dodatnich jest dość równy i wynosi 6:7.

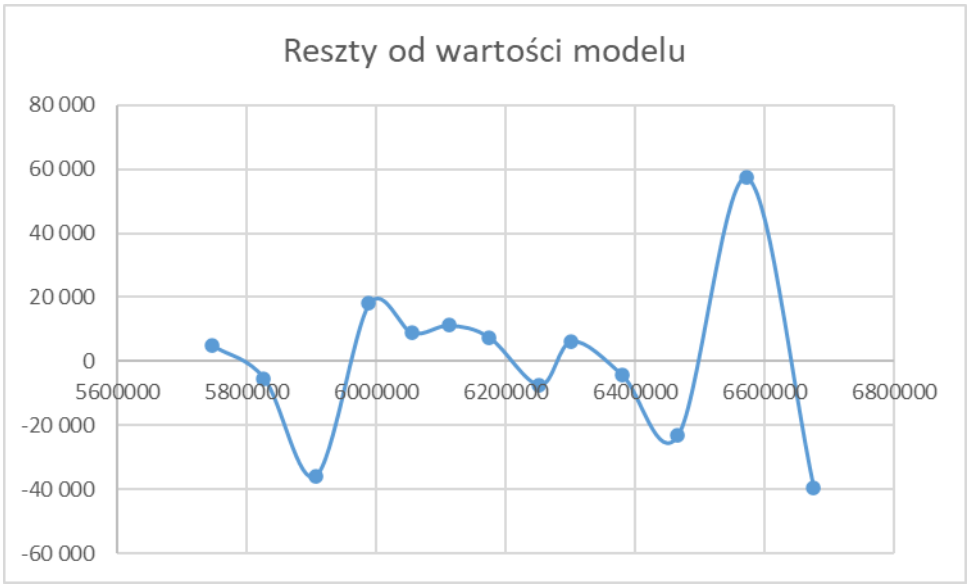
Hipoteza	Rozkład jest normalny
Test JB	2,003172254
Chi test	0,3672964
	Nie mamy podstaw by odrzucić hipotezy

rys. 4-6 Test na normalność rozkładu reszt w MS Excel

```
> JarqueBera.test(reszty)

        Jarque Bera Test

data:  reszty
X-squared = 0.66137, df = 2, p-value = 0.7184


        Skewness

data:  reszty
statistic = 0.45159, p-value = 0.5062


        Kurtosis

data:  reszty
statistic = 3.6366, p-value = 0.6394

> shapiro.test(reszty)

        Shapiro-Wilk normality test

data:  reszty
W = 0.92471, p-value = 0.2901
```

rys. 4-7 Test na normalność rozkładu reszt w R

Reszty mają rozkład normalny. Wartości p są większe niż 0.05.



rys. 4-8 Wykres reszt od wartości modelu

Homoskedastyczność rozkładu. Reszty nie są większe, im większe są wartości zmiennej objaśniającej. Wariancja jest stała.

mean(y)	odchylenie(reszt)	Wyrazistość modelu
6 189 024	24816,44761	0,004009752

rys. 4-9 Wyrazistość modelu

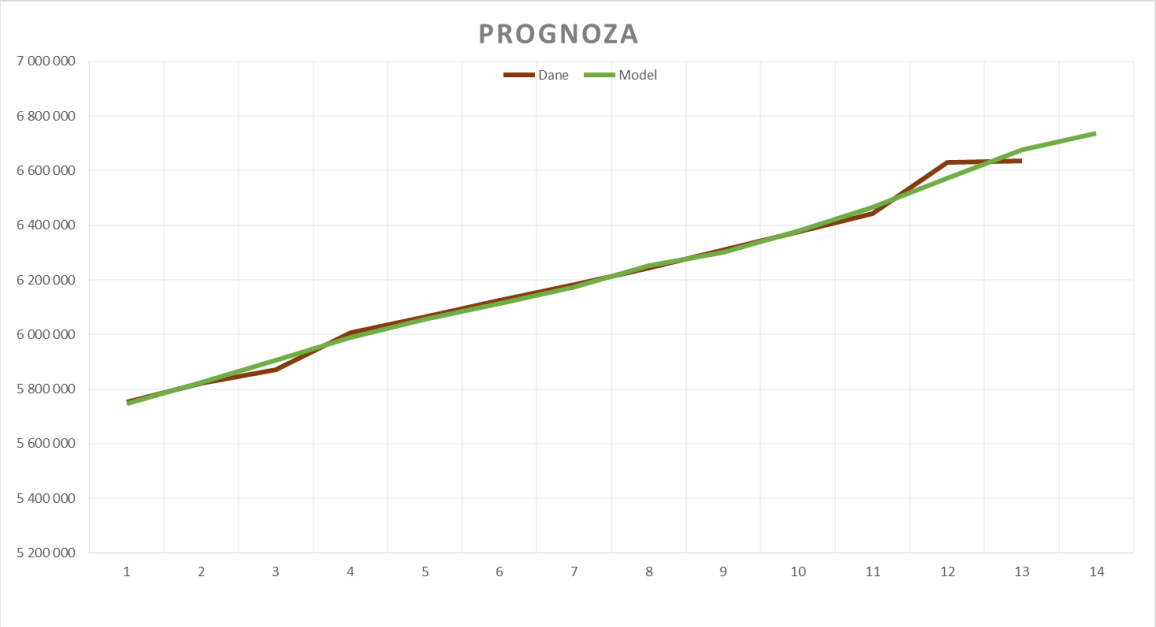
Wyrazistość modelu wynosi wiele poniżej 15%. Oznacza to dużą zgodność modelu z danymi empirycznymi

AIC	308.952656647616
AICC	317.524085219044
BIC	311.777403434923

rys. 4-10 Wartości kryteriów informacyjnych

Kryterium AIC jest dość niskie w porównaniu do wartości modelu.

5. Podsumowanie i prognoza teoretyczna



rys. 5-1 Model z prognozą na rok

Prognozę przeprowadziłem na podstawie dostępnych danych dla stopy bezrobocia, interpolacji liniowej wartość brutto środków trwałych na 1 mieszkańca (wartość rośnie prawie liniowo na przestrzeni lat) oraz prognozy liczby rozpoczętych budów mieszkań w 2021 roku według głównego urzędu statystycznego.

Według prognozy na kolejny rok (2021) liczba lokali mieszkalnych w Polsce będzie się zwiększać. Obserwując tendencje rynkowe i popyt na lokale mieszkalne prognoza może się sprawdzić.

6. Spis rysunków