Zależność ceny gazu ziemnego od charakterystyk makroekonomicznych krajów Europy.

Seweryn Turula

Projekt Ekonometria

Czerwiec 2022

Spis treści

Spis rysunków

1 Wstęp

W poniższym projekcie postaram się zbudować model ekonometryczny opisujący zależność między ceną gazu ziemnego a charakterystykami makroekonomicznymi w krajach Europy.
Zmienność cen gazu ziemnego dotyka znaczną część mieszkańców Europy i bezpośrednio wpływa na jakość ich życia. Chciałbym dowiedzieć się, z czego wywodzą się różnice w cenach gazu ziemnego w poszczególnych krajach Europy oraz które zmienne mają na nie największy wpływ. Dane, na których będę pracować pochodzą z 2021 roku.

2 Opis danych

Zmienne, za pomocą których postaram się opisać cenę gazu ziemnego to:

1. **HDD-Heating days degree**- indeks oparty na warunkach pogodowych panujących w danym kraju, mający na celu opisanie zapotrzebowania na energię grzewczą budynków. Wyliczenie HDD opiera się na temperaturze bazowej, zdefiniowanej jako najniższa dobowa średnia temperatura powietrza, która nie prowadzi do ogrzewania pomieszczenia i zsumowaniu tych temperatur po każdym dniu badanego okresu.
   Po obróbce minimalna i maksymalna zaobserwowana wartość tej zmiennej to odpowiednio: 1065 i 5201, gdzie średnio wynosiła 2861.
   Źródła: `https://ec.europa.eu/eurostat/cache/metadata/en/nrg_chdd_esms.htm`
   `https://ec.europa.eu/eurostat/databrowser/view/nrg_chdd_a/default/table?lang=en`

2. **Eco_index**- zmienna typu factor przyjmująca trzy poziomy: "Catch-up", "Average" oraz "Leader". Opisuje na jakim poziomie względem średniej dla Unii Europejskiej znajduje sie konkretne państwo pod względem innowacji ekologicznych.
   Po obróbce liczba obserwacji danego poziomu wynosi:"Average"-16,"Catch-up"-7,"Leader"-8.
   Źródło: `https://ec.europa.eu/environment/ecoap/indicators/index_en`

3. **Nuclear_plant**- binarna zmienna typu factor przyjmująca poziomy "Yes" lub "No" i bezpośrednio wskazuje na to, czy na terenie danego państwa znajduje się elektrownia jądrowa.
   Po obróbce liczba obserwacji danego poziomu wynosi:"Yes"-12,"No"-19.
   Źródło: `https://www.euronuclear.org/glossary/nuclear-power-plants-in-europe/`

4. **Import_dependency**- udział całkowitych potrzeb energetycznych kraju pokrywanych przez import z innych krajów. Jest obliczany na podstawie bilansów energii jako import netto podzielony przez dostępną energię brutto i wyrażany w procentach.
   Po obróbce minimalna i maksymalna zaobserwowana wartość tej zmiennej to odpowiednio: 10.50% i 92.46%, gdzie średnio wynosiła 55.25%.
   Źródła: `https://ec.europa.eu/eurostat/web/products-datasets/-/nrg_ind_id`
   `https://ec.europa.eu/eurostat/cache/metadata/en/nrg_ind_id_esms.htm`

5. **GDP (Gross Domestic Product) per capita**-opisuje zagregowaną wartość dóbr i usług finalnych wytworzonych przez narodowe i zagraniczne czynniki produkcji na terenie danego kraju przypadającą na jednego obywatela, wyrażona w Euro.
   Po obróbce minimalna i maksymalna zaobserwowana wartość tej zmiennej to odpowiednio: 14150€ i 126570€, gdzie średnio wynosiła 43259€.

Zmienną objaśnianą w tym modelu jest:

6. **Natural_gas_price**- zmienna opisująca cenę gazu ziemnego, jaką musi zapłacić przeciętne gospodarstwo, aby ogrzać budynek mieszkalny. Dane o cenach gazu podawane są przez krajowe urzędy statystyczne, ministerstwa, agencje energetyczne lub w przypadku monopoli przez pojedyncze spółki gazowe. Ceny zawierają cenę podstawową gazu, usługi przesyłowe, systemowe i dystrybucyjne. Jednostka, w której podawana jest zmienna to Euro/kWh.
   Po obróbce minimalna i maksymalna zaobserwowana wartość tej zmiennej to odpowiednio: 3.262 i 34.275, gdzie średnio wynosiła 13.832.
   Źródła: `https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_202_esms.htm` `https://data.europa.eu/data/datasets/m836egkc7u1woixxzprv7a?locale=en`

3 Dobór i obróbka zmiennych

Dane zostały zebrane dla 32 państw europejskich, głównie krajów członkowskich Unii Europejskiej.
Praktycznie wszystkie braki danych obecne są w zmiennych HDD oraz ECO_index. Państwa, dla których danych brakuje, nie są krajami członkowskimi UE. Wyjątkiem jest zmienna Import_dependency dla państwa Liechtenstein.

Wstępnym rozwiązaniem tego problemu jest dokładniejsze przyjrzenie się zmiennej HDD. Jest to zmienna, którą odpowiada na pytanie *"Jak duża jest potrzeba ogrzewania budynków w danym państwie w okresie jednego roku?"*, zatem jest to zmienna zależna od tego, w jakiej strefie klimatycznej dane państwo się znajduje. Biorąc średnią z wartości HDD państw sąsiadujących z danym krajem, otrzymujemy dobre oszacowanie brakującej wartości współczynnika dla państwa, dlatego zdecydowałem się na taką metodę.

Zmienną Eco_index uzupełniam środkowym poziomem "Average" dla każdego państwa, dla którego brakuje danych.
Postanowiłem całkowicie pozbyć się obserwacji dla państwa Liechtenstein, głównie ze względu na brak wartości dla zmiennej Import_dependency, jednak niepokojąca była też wartość zmiennej GDP per capita, dla której Liechtenstein był wartością zdecydowanie odstającą.

Na tak wyczyszczonych i uzupełnionych danych starał się będę zbudować tytułowy model statystyczny.

4 Budowa i kalibracja modelu

4.1 Wstępny model

Wyjściowy model, który będę rozważał, wykorzystuje wszystkie dostępne dane i ma postać:

$$Natural\_gas\_price \sim \beta_0 + \beta_1 \cdot HDD + \beta_2 \cdot GDP + \beta_3 \cdot Nuclear\_plant + \beta_4 \cdot Import\_dependency + \beta_5 \cdot Eco\_index,$$

gdzie zmienna objaśniana (*Natural_gas_price*) jest zmienną numeryczną, trzy zmienne objaśniające (*HDD,GDP,Import_depedency*) również są numeryczne i dwie zmienne objaśniające (*Eco_index, Nuclear_plant*) są typu factor.

Podstawowe statystyki modelu otrzymuję z funkcji *summary* i przedstawiają się one w następujący sposób:

- t-test: żadna zmienna objaśniająca, oprócz wyrazu wolnego, nie jest istotna statystycznie na poziomie 5%. Ogólniej mówiąc, dla prawie żadnej zmiennej nie jesteśmy w stanie odrzucić hipotezy zerowej o tym, że współczynnik β przy tej zmiennej równa się zero. Wyniki te wskazują na poważne problemy w modelu, ponieważ nie można jednoznacznie stwierdzić, czy model bez danej zmiennej nie tłumaczyłby zależności między zmiennymi w lepszy sposób.
- F-test: standardowy test, mówiący o tym, czy model całościowo można zastąpić modelem zawierającym tylko wyraz wolny. W przypadku tego modelu *p-value* testu wynosi 0.00294, zatem odrzucimy hipotezę zerową, że zależność opisaną przez model da się całkowicie zastąpić wyrazem wolnym. Wynik ten jest minimalnym pozytywem w ocenie tego modelu.
- skorygowany współczynnik R^2: współczynnik informujący o tym, jaka część zmienności zmiennej objaśnianej została opisana przez model. Skorygowany współczynnik wprowadza karę za użycie zbyt dużej liczby zmiennych, które nie polepszają dopasowania (a obniżają liczby stopni swobody w modelu). W tym modelu współczynnik ten wynosi 0.4206, co jest stosunkowo niskim poziomem.

Na tym stadium zakończę ocenę tego modelu, ponieważ już tak podstawowe statystyki wskazują na niedociągnięcia i złe dopasowanie modelu. Tak prosto zbudowany model nie opisuje szukanej relacji w jakkolwiek zadowalający sposób i nie powinien być podstawą do wnioskowania na temat zależności między zmiennymi.

4.2 Modele pośrednie

Następnym krokiem, który podjąłem w poszukiwaniu odpowiedniego modelu, było stworzenie oraz ocena modeli pośrednich, zawierających tylko jedną zmienną objaśniającą. Modele te były postaci:

$$Natural\_gas\_price \sim \beta_0 + \beta_1 \cdot HDD,$$

odpowiednio dla każdej ze zmiennych objaśniających. Już na wstępnym etapie analizy wszystkie modele takiej postaci miały bardzo podobne problemy jak model wyjściowy.

Jedynym modelem wartym uwagi był model zależny od zmiennej GDP. Zmienna ta okazała się być istotna statystycznie z *p-value t-testu* na poziomie 0.01. Główny problem tego modelu polegał na bardzo małej wartości współczynnika R^2, wynoszącego tylko 0.1957, jednak pomimo tego postanowiłem dokładniej przeanalizować modele zawierające tą zmienną.

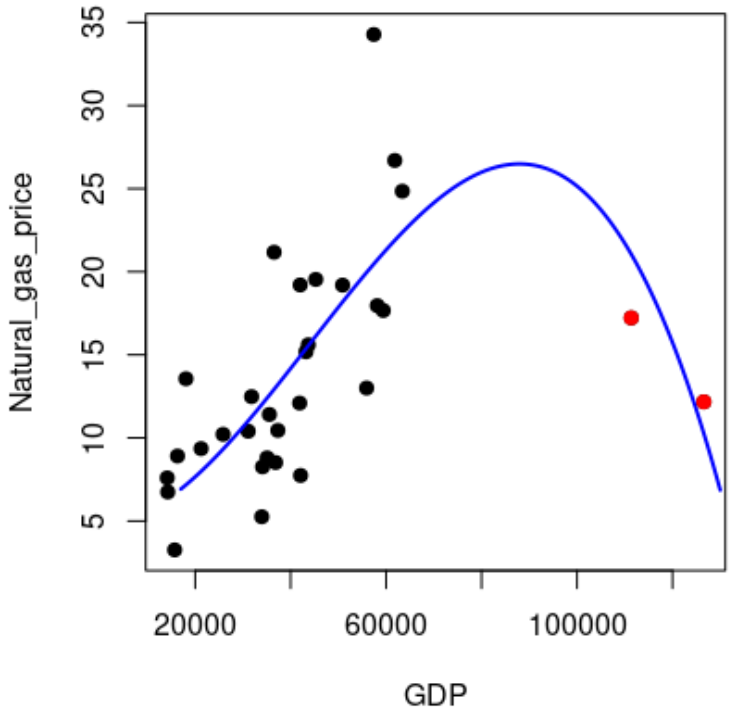
Kolejnymi modelami, które tworzyłem, były modele zawierające wyższe potęgi zmiennej GDP. Model, który na podstawie tych samych kryteriów okazał się najlepszy, miał postać:

$$Natural\_gas\_price \sim \beta_0 + \beta_1 \cdot GDP^3 + \beta_2 \cdot GDP^2,$$

gdzie potęgi zmiennej były istotne statystycznie oraz współczynnik R^2 wzrósł do poziomu 0.5458.

Takie wyniki, zważając na małą liczność obserwacji, na których zbudowano model, powinny budzić podejrzenia w sprawie zjawiska zwanego *overfit*. *Overfit* występuje, gdy dany model statystyczny doszukuje się zależności, których w rzeczywistości nie ma. Modele budowane na małej liczbie obserwacji są na to zjawisko niezwykle podatne, dlatego należy sprawdzić, czy nie występuje ono w tym modelu.

Najprostszym i jednocześnie najbardziej skutecznym podejściem do tego problemu jest narysowanie wykresu punktowego zmiennej objaśnianej i dopasowanie linii regresji wynikającej z modelu. Należy następnie ocenić, czy na wykresie nie występują grupowania punktów bądź wartości odstające.



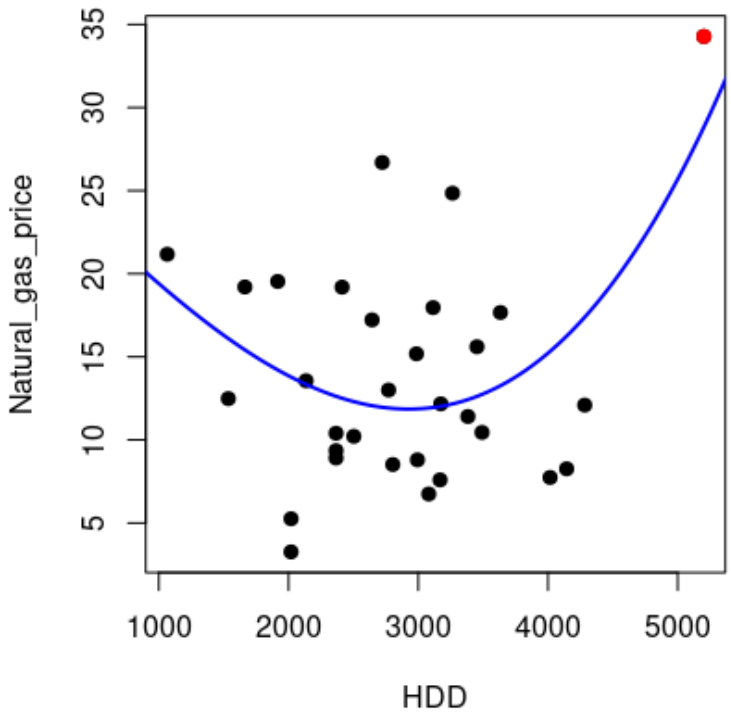
Rys. 1: Wykres zależności ceny gazu ziemnego od GDP per capita.
Na wykresie punkty odpowiadają poszczególnym obserwacjom, kolorem niebieskim zaznaczono dopasowaną przez model linię regresji, kolorem czerwonym zaznaczono dwa punkty odstające.

Jak widać na Rysunku 1, linia regresji dopasowana przez model próbuje dopasować się jak najlepiej do obserwacji zaznaczonych kolorem czerwonym, co daje złudne wrażenie, że model ten jest poprawny i odpowiednio opisuje zależności między zmiennymi. Obserwacje zaznaczone na czerwono to odpowiednio Irlandia oraz Luksemburg. Po ich usunięciu i ponownym zbudowaniu modelu otrzymujemy zupełnie przeciwne wnioski niż w przypadku uwzględniającym te dwie obserwacje. Zmienne stają się nieistotne statystycznie, z *p-value* standardowych *t-testów* na poziomie 0.65 oraz 0.62. Na podstawie ewidentnego zjawiska *overfit* zdecydowałem się nie przyjmować tego modelu jako finalnego.

Kolejny model, który postanowiłem przeanalizować, był podobny do poprzedniego, jednak zależny od zmiennej HDD i był postaci:

$$Natural\_gas\_price \sim \beta_0 + \beta_1 \cdot HDD^3 + \beta_2 \cdot HDD.$$

Skorygowany współczynnik R^2 tego modelu na poziomie 0.185 nie wskazywał na to, że model dobrze wyznacza szukaną zależność, jednak istotność statystyczna obydwu zmiennych skłoniła mnie do dalszej analizy. Sugerując się poprzednim modelem, postanowiłem sprawdzić, czy w modelu zależnym od zmiennej HDD znów nie występuje *overfit*. Dalsze postępowanie jest analogiczne jak w przypadku model zależnego od GDP.



Rys. 2: Wykres zależności ceny gazu ziemnego od zmiennej HDD
Na wykresie punkty odpowiadają poszczególnym obserwacjom, kolorem niebieskim zaznaczono dopasowaną przez model linię regresji, kolorem czerwonym zaznaczono punkt odstający.

Po raz kolejny można zauważyć obserwację szczególnie odstającą, do której linia regresji próbuje się nadmiernie dopasować. Obserwacją zaznaczoną na czerwono na Rysunku 2 jest Szwecja. Ponownie, po usunięciu tej obserwacji i zbudowaniu identycznego modelu okazało się, że zmienne są nieistotne statystycznie i modelu nie przyjąłem jako finalnego.

4.3 Wybór finalnego modelu

Po usunięciu trzech wartości odstających dla zmiennych HDD i GDP (Irlandia, Szwecja, Luksemburg) ponownie zacząłem budować modele, tym razem zawierające zarówno zmienną HDD jak i GDP wraz z ich potęgami. Zbudowałem w tym miejscu wiele modeli (pozwolę sobie ominąć opis każdego z nich), jednak ten, który przykuł moją uwagę, był postaci:

$$Natural\_gas\_price \sim \beta_0 + \beta_1 \cdot HDD + \beta_2 \cdot GDP^2.$$

Podstawowe statystyki tego modelu, na które do tej pory zwracałem główną uwagę, są zadowalające i mogą sugerować, że model dobrze opisuje zależności między zmiennymi.

P-value dla *t-testów* każdej zmiennej są na poziomie: $5.38 \cdot 10^{-6}$ dla wyrazu wolnego, 0.00302 dla zmiennej HDD oraz $4.61 \cdot 10^{-8}$ dla zmiennej GDP^2, zatem stanowczo poniżej 0.05 przyjętej za punkt graniczny istotności statystycznej. *P-value* dla *F-testu* również jest na bardzo niskim poziomie $1.841 \cdot 10^{-7}$, a skorygowany współczynnik R^2 wynosi 0.6877, co jest stosunkowo wysokim wynikiem.

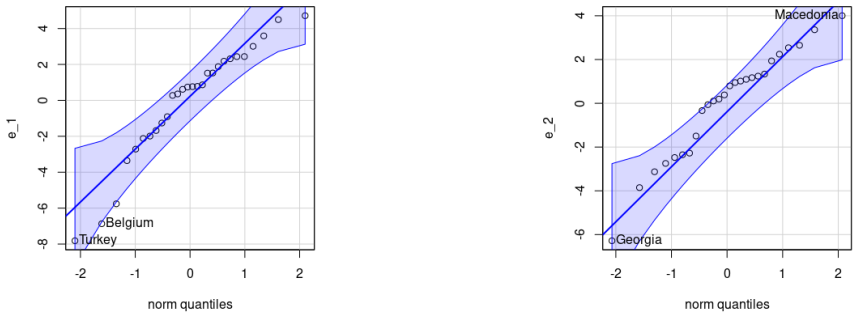
Wstępna analiza modelu daje zadowalające i obiecujące wyniki na to, że model będzie w stanie opisać szukaną zależność na dobrym poziomie. Z tego powodu, postanowiłem przyjąć ten model za model finalny, na którym przeprowadziłem dogłębniejszą diagnostykę oraz weryfikacje założeń.

5 Diagnostyka oraz weryfikacja finalnego modelu

5.1 Założenia modelu regresji liniowej

5.1.1 Normalność błędów

Pierwszym założeniem, które sprawdziłem, było założenie o normalności błędów modelu. Graficzną metodą, która pomogła dokonać ostatecznej korekty dla danych był wykres *QQ (quantile-quantile) plot* z biblioteki *car*.



(a) $qqPlot_1$ (before)

(b) $qqPlot_2$ (after)

Rys. 3: Quantile-quantile plot dla danych przed i po ostatecznej korekcie.
Niebieska linia odpowiada teoretycznym kwantylom rozkładu normalnego, punkty odpowiadają kwantylom z próbki, zacieniowany niebieski obszar to przedział ufności dla każdego z kwantyli.

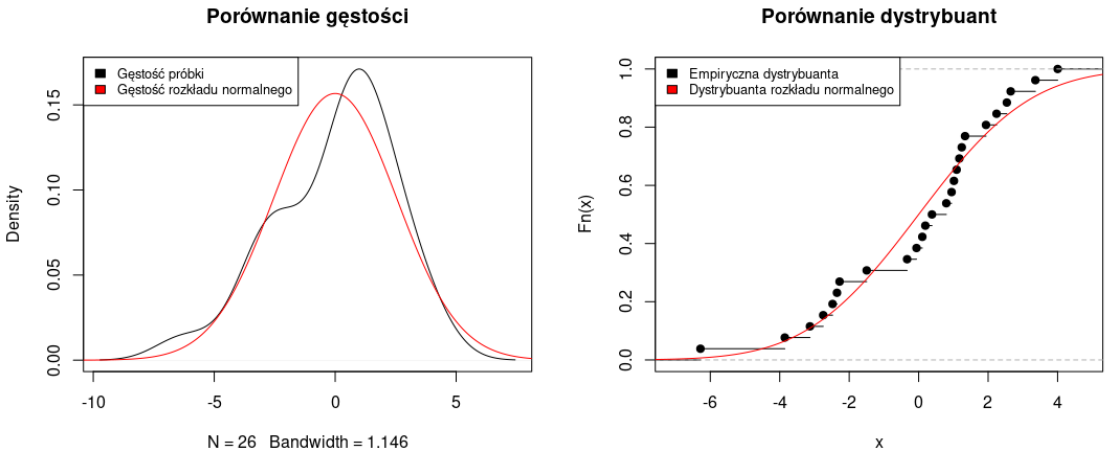
Jak można zauważyć, na Rysunku 3a kwantyle wyznaczone dla Turcji oraz Belgii nie należą do przedziału ufności wyznaczonego przez funkcję *qqPlot*. Sprawdziłem zatem, czy po usunięciu tych obserwacji sytuacja się poprawi. Ilustruje to Rysunek 3b, na którym wszystkie kwantyle z próbki leżą w wyznaczonym przedziale ufności.

Aby utwierdzić się w decyzji o usunięciu kolejnych dwóch obserwacji, zastosowałem test statystyczny *Lilliefors'a* dla danych przed i po wyrzuceniu. Test ten jest ogólniejszą odmianą testu *Kołmogorova-Smirnova*, sprawdzający, czy dane pochodzą z rozkładu normalnego, jednak w przeciwieństwie do testu *Kołmogorowa* nie precyzuje on średniej ani odchylenia standardowego badanego rozkładu normalnego. Funkcja, której użyłem, to *lillie.test* z pakietu *nortest*.

P-value w pierwszym przypadku wynosi 0.02428, gdy w drugim 0.1807. Wywnioskować stąd można, że dla danych przed usunięciem obserwacji na poziomie 5% odrzucimy hipotezę zerową o normalności błędów, a gdy je usuniemy, nie będziemy mogli tej hipotezy odrzucić. Wyniki tego testu jak i inspekcji graficznej Rysunku 3 potwierdziły decyzję o wyrzuceniu obserwacji dla Belgii i Turcji.

Po ostatecznej korekcie wyjściowych danych, wszystkie zmienne zostały na takim samym poziomie istotności statystycznej, jednak skorygowany współczynnik R^2 wzrósł do zadowalającego poziomu 0.8148, co jedynie bardziej poparło decyzję o korekcie danych. Od tej pory finalnym modelem nazywać będę model bez obserwacji dla Belgii i Turcji.

Kolejnym graficznym podejściem sprawdzenia normalności błędów będzie porównanie gęstości wyznaczonej za pomocą funkcji *density* dla standardowego jądra Gauss'owskiego z gęstością rozkładu normalnego o odchyleniu standardowym równemu błędowi standardowemu residuum (wyestymowanemu w standardowy sposób przez estymator OLS parametru σ^2), oraz empirycznej dystrybuanty residuów do dystrybuanty tego samego rozkładu normalnego.



(a) Wykres gęstości

(b) Wykres dystrybuant

Rys. 4: Wykresy obrazujące dystrybuantę empiryczną oraz gęstość błędów

Zarówno na Rysunku 4a jak i Rysunku 4b widać, że wykresy dystrybuanty empirycznej i gęstości pokrywają się z tymi wyznaczonymi w sposób teoretyczny. Dodatkowo postanowiłem wykorzystać również inne testy statystyczne badające normalność błędów w modelu.

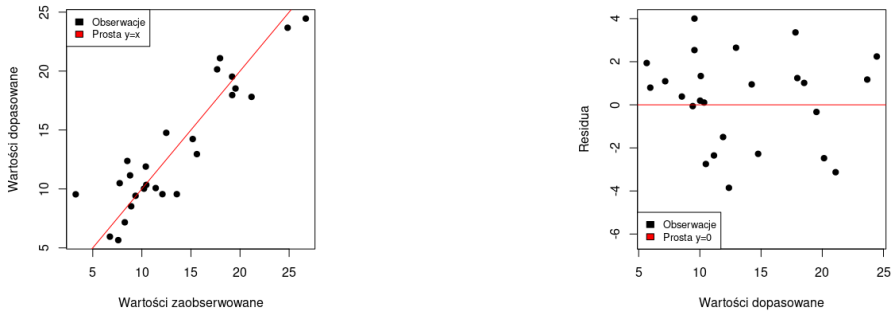
Oprócz wspomnianego wcześniej testu *Lilliefors'a* wykorzystałem test *Anderson'a-Darling'a, Jarque-Bera* oraz *Shapiro-Wilk'a*. Funkcje, których użyłem to odpowiednio *ad.test, jarque.bera.test, shapiro.test* pochodzące z pakietów *nortest, tseries, stats*. Hipoteza zerowa dla każdego z tych testów zakłada, że błędy modelu pochodzą z rozkładu normalnego.

- test *Anderson'a-Darling'a*- test statystyczny badający średni ważony kwadrat odległości między dystrybuantą empiryczną a teoretyczną. *P-value* wynosi 0.1688, zatem nie odrzucamy hipotezy zerowej. Wynik tego testu pokrywa się z inspekcją wizualną Rysunku 4b.
- test *Jarque Bera*- test statystyczny sprawdzający "grubość" ogona rozkładu w oparciu o empiryczny trzeci i czwarty moment z próbki. *P-value* wynosi 0.3918, zatem nie odrzucamy hipotezy zerowej. Wynik tego testu również można porównać z inspekcją wizualną Rysunku 4a.
- test *Shapiro-Wilk'a*- test statystyczny sprawdzający, czy empiryczny drugi moment jest porównywalny do oczekiwanej ważonej średniej podniesionej do kwadratu. *P-value* wynosi 0.2715, zatem znów nie mamy podstaw do odrzucenia hipotezy zerowej.

Posiadając szeroki pogląd na temat pochodzenia błędów modelu, stwierdzam, że finalny model po korekcie zbioru danych spełnia założenie o normalności błędów, co potwierdzają powyższe wyniki.

5.1.2 Liniowa zależność

Kolejnym założeniem ważnym dla oceny modelu jest sprawdzenie, czy zależność między zmiennymi w modelu faktycznie jest liniowa. Ze względu na obecność składnika GDP^2 bardziej adekwatnym nazwaniem tego modelu byłaby *Regresja wielomianowa* bądź *Regresja nieliniowa względem zmiennych*. Jednak pomimo użycia takiego przekształcenia jednej ze zmiennych, natura tego modelu w istocie dalej jest liniowa, za wyjątkiem tego, że przy interpretacji modelu zmienną *GDP* podnosimy do kwadratu.



(a) Wartości dopasowane względem zaobserwowanych (b) Residua względem wartości dopasowanych

Rys. 5: Wykresy obrazujące zależności między wartościami dopasowanymi a residuami i wartościami zaobserwowanymi

Główną metodą sprawdzenia liniowości, którą wykorzystałem, jest analiza wykresów dopasowanego modelu.

Pierwszy wykres, przedstawiony na Rysunku 5a, przedstawia zależność między wartościami dopasowanymi przez model a wartościami zaobserwowanymi. Punkty rozproszone są symetrycznie wokół prostej $y = x$, co wskazuje na dobre dopasowanie modelu.

Podobna sytuacja ma miejsca na Rysunku 5b, który przedstawia zależność residuum do wartości dopasowanych przez model. Tym razem punkty równomiernie rozproszone są wokół prostej $y = 0$ oraz nie widać żadnych trendów.

Test statystyczny, którego wynik pokrywa się z analizą wizualną wykresów jest *Rainbow-test*. Test ten bada, czy dla średnich wartości zmiennej objaśniającej dopasowanie na podpróbkach nie różni się od pełnego dopasowania. Funkcja, której użyłem do przeprowadzenia testu to *raintest* z pakietu *lmtest*. *P-value* na poziomie 0.9741 nie daje podstaw do odrzucenia hipotezy zerowej o tym, że zależność, którą bada model, jest liniowa.

Mając na uwadze wyniki przedstawione na Rysunku 5 oraz *Rainbow-test* wnioskuję, że model spełnia warunek o liniowej postaci modelu.

5.1.3 Współliniowość

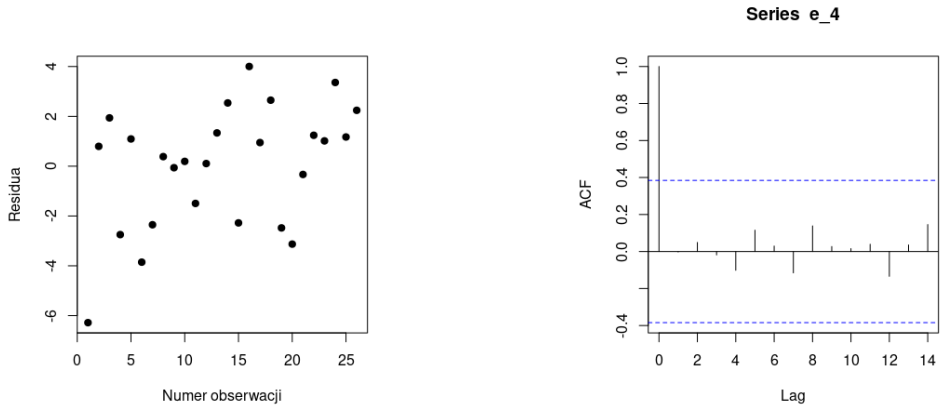
Aby przeanalizować, czy w modelu nie występuje współliniowość, wykorzystam macierz korelacji pomiędzy zmiennymi, współczynnik *VIF (Variance-Inflation-Factor)* i współczynnik tolerancji *T*.

Korelacja zmiennej HDD ze zmienną GDP^2 odczytana z macierzy korelacji wynosi 0.2474, co nie wskazuje na współliniowość między tymi zmiennymi.

Potwierdza to również współczynnik tolerancji T obliczany za pomocą współczynnika R^2 modelu, gdzie zmienną objaśnianą jest HDD a zmienną objaśniająca GDP^2. Współczynnik T wynosi w tym przypadku 0.9388, a współczynnik *VIF*, obliczany jako odwrotność T, 1.06517. Obydwie wartości współczynników interpretujemy jako brak współliniowości.

5.1.4 Niezależność błędów

Analizę niezależności błędów rozpocznę od posortowania danych względem zmiennej objaśnianej *Natural_gas_price*. Dane, na których zbudowany jest model, nie są w żaden sposób uporządkowane ani nie są szeregiem czasowym. Z tego względu metody sprawdzenia tego założenia takie jak analiza autokorelacji bądź testy statystyczne nie dają adekwatnych wyników. Po posortowaniu danych względem zmiennej objaśnianej problem ten częściowo zanika, ponieważ sprawdzamy wtedy, czy istnieje korelacja między "kolejnością w szeregu" obserwacji, a wartością residuum. Takie podejście fragmentarycznie odzwierciedla sytuację, kiedy to dane tworzą szereg czasowy i sprawdzane jest istnienie korelacji między obserwacjami, a czasem gdy zostały zebrane.
Następnym krokiem jest faktyczne sprawdzenie braku istnienia takowej zależności. Posłuży do tego analiza graficzna wykresu przedstawiającego residua wraz z indeksem odpowiadającej obserwacji (Rysunek 6a) oraz autocorrelation function *(acf)* (Rysunek 6b) z podstawowego pakietu *stats*.
Z Rysunku 6a nie da się wywnioskować żadnego trendu ani grupowania residuów, co sugeruje brak autokorelacji. Potwierdza to również Rysunek 6b, na którym widać, że wszystkie współczynniki autokorelacji dla danych opóźnień leżą w wyznaczonym przedziale ufności i są na niskim poziomie.



(a) Numer obserwacji względem residuum

(b) Wykres autokorelacji

Rys. 6: Wykresy obrazujące autokorelację

Problem autokorelacji w danych nie tworzących szeregu czasowego ogólnie mówiąc jest problemem źle zdefiniowanym. W przypadku danych tworzących ten model autokorelacja nie powinna mieć miejsca, ponieważ dane nie są uporządkowane względem czasu. Bardziej odpowiednim podejściem do sprawdzenia tego założenia byłoby sprawdzenie tak zwanej *Spatial correlation*, która wskazuje na zależność między miejscem zebrania danych, a ich wartościami. W przypadku obecnych tu danych, gdzie każda obserwacja odpowiada danemu państwu w Europie, niestety obliczenie tej statystyki również nie przynosi żadnych interpretowalnych wyników. Należałoby odpowiednie obserwacje uporządkować w deterministycznie dobrane "grupy" państw, co niesie za sobą oczywiste problemy i nieścisłości w doborze kryteriów jak każde z państw zakwalifikować do odpowiedniej grupy. Podsumowując, uważam, że problem niezależności błędów nie występuje w danych użytych do budowy finalnego modelu powołując się głównie na analizę graficzną Wykresu 6a, gdyż inne tradycyjne metody wykorzystywane w analizie szeregów czasowych nie są formalnie zdefiniowane dla typu danych użytych w tym projekcie.

5.1.5 Homoskedastyczność błędów

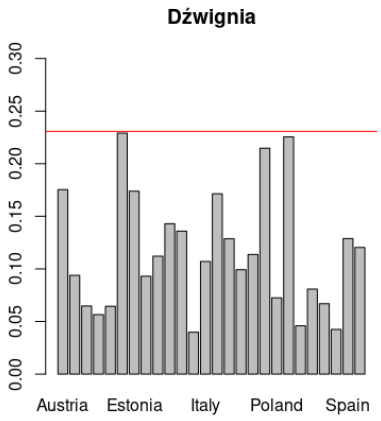
Kolejną istotną cechą modelu jest homoskedatyczność czynnika losowego. Homoskedastyczność oznacza stałą wariancję błędu modelu. Gdyby błędy były heteroskedastyczne, to znaczy nie miałyby stałej wariancji, mogłoby być to przesłanką do uznania modelu jako źle dopasowany. Z heteroskedastyczności można wywnioskować, że w szumie modelu ukryte są jeszcze jakieś informacje, których nasz model nie wychwytuje. Należałoby w takiej sytuacji model poprawić przez na przykład wyszukanie nowych zmiennych objaśniających związanych z opisywaną zależnością.

Do zbadania tego założenia wykorzystam już wcześniej użyty wykres przedstawiony na Rysunku 5b. Pokazuje on zależność między residuum a wartościami dopasowanymi przez model. Na co należy zwrócić uwagę przy badaniu homoskedastyczności jest rozproszenie punktów wokół prostej $y = 0$. Na wykresie tym widać, że punkty są równomiernie rozproszone i nie da się z niego odczytać żadnego trendu, co wskazuje na brak heteroskedastyczności i stałą wariancję wśród błędów badanego modelu.

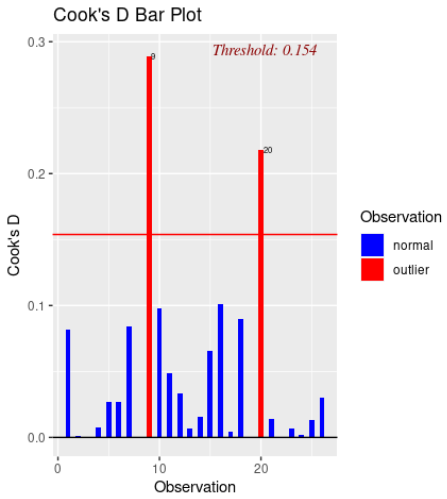
Testy statystyczne badające spełnienie tego założenia to *Breusch-Pagan* test oraz *Goldfeld-Quandt* test. Funkcje, które wykorzystałem do przeprowadzenia testów to *bptest* i *gqtest* z pakietu *lmtest*. *P-value* otrzymane z funkcji to odpowiednio 0.4769 oraz 0.893, zatem nie jesteśmy w stanie odrzucić hipotezy zerowej o homoskedastyczności błędów.
Podsumowując wyniki statystyczne jak i graficzne, badany model nie ma problemu z heteroskedastycznością czynnika losowego.

5.2 Obserwacje odstające

Ważnym elementem w ocenie modelu liniowego jest identyfikacja obserwacji odstających oraz wpływowych dla modelu. Charakterystyki, na których skupiam się w analizie danych, to dźwignia oraz *Cook's distance*.



(a) Wartość dźwigni dla każdej obserwacji. Czerwona linia zaznaczona jest na wysokości $y = \frac{6}{26}$



(b) Wartość odległości Cook'a

Rys. 7: Wykresy dla wartości odstających

Jak widać na Rysunku 7a wartość dźwigni dla żadnej z obserwacji nie przekracza granicznego poziomu równego $\frac{6}{26}$. Oznacza to, że dane rozdystrybuowane są względem wartości zmiennych objaśniających równomiernie. Jest to wynik, którego powinniśmy się spodziewać, zważając na to, że w sekcji 4.2 usunięte zostały problematyczne obserwacje po graficznej analizie wykresów linii regresji. Analizując wykres z Rysunku 7b łatwo zauważyć dwie obserwacje znacznie wykraczające poza threshold wyznaczony przez funkcję *ols_plot_cooksd_bar* z pakietu *olsrr*. Tymi obserwacjami są odpowiednio Gruzja i Portugalia, jednak wartości tego współczynnika nie powinny prowadzić do usunięcia tych obserwacji. Współczynnik mówi o tym, jak bardzo istotne dla modelu są dane obserwacje, jednak sytuacja, kiedy wartość tego współczynnika wykracza poza wyznaczoną granicę, nie jest sytuacją negatywną.

5.3 Analiza stabilności oraz mocy prognostycznej modelu

Przeznaczeniem modelu ekonometrycznego opisywanego w tym projekcie było znalezienie zależności między cenami gazu ziemnego a charakterystykami makroekonomicznymi państw Europy. Oprócz tego ciekawym i wartym uwagi zagadnieniem do analizy jest sprawdzenie mocy prognostycznej modelu. Innymi słowy, chcemy sprawdzić, jak dobrze dany model poradzi sobie z "przewidzeniem" wartości zmiennej objaśnianej dla nowo podanych obserwacji. W tym celu wprowadza się określenia zbioru uczącego (*training data*) służącego do kalibracji modelu oraz zbioru testowego (*test data*), na którym sprawdzamy, na jakim poziomie model zdołał dopasować wartości do tych rzeczywiście zaobserwowanych.

Aby przetestować moc prognostyczną finalnego modelu, zastosuję tak zwany sprawdzian k-krzyżowy (k-*fold cross-validation*). Polega on na rozdzieleniu badanego zbioru danych na k podzbiorów o podobnej liczności. W drugiej kolejności tworzy się k modeli statystycznych, w których po kolei ustala się wybrany podzbiór jako zbiór testowy aż do momentu, kiedy każdy z podzbiorów pełnił funkcję zbioru testowego. Dla każdego ze stworzonych modeli wylicza się testowy błąd średniokwadratowy (*test mean-squared error*) za pomocą zsumowania kwadratów różnic wartości dopasowanych i zaobserwowanych dla danych, na których model nie był kalibrowany (czyli na zbiorze testowym). Następnie wystarczy wyliczyć średnią z k otrzymanych błędów średniokwadratowych, którą nazywa się ogólnym testowym błędem średniokwadratowym.

Przed interpretacją wyników należy wspomnieć, że schemat postępowania przedstawiony w tym podrozdziale jest możliwy dlatego, że finalny model spełnia warunek o niezależności błędów oraz dane nie są szeregiem czasowym. Gdyby dane nie spełniały tego założenia bądź tworzyłyby szereg czasowy, należałoby z większą ostrożnością podejść do rozdzielenia zbioru danych na podzbiory. Stosowniejszą metodą ustalenia zbioru uczącego i testowego w przypadku szeregu czasowego mogłoby być rozdzielenie danych na tylko dwa podzbiory, zachowując przy tym uporządkowanie obserwacji w czasie. Temat ten dla konkretnego zbioru danych z państw europejskich omówiony został w podrozdziale 5.1.4.

Przechodząc do wyników otrzymanych dla badanego modelu, *test MSE* wynosi 8.746. Lepiej interpretowalną miarą będzie podanie *test root mean-squared error*, czyli pierwiastek z *MSE*. Współczynnik *RMSE* mierzy średnią różnicę między przewidywaniami dokonanymi przez model a rzeczywistymi obserwacjami. Im niższy *RMSE*, tym dokładniej model może przewidywać rzeczywiste obserwacje. W przypadku sprawdzianu k-krzyżowego należy najpierw wziąć pierwiastki z każdego z *MSE* a dopiero potem wyliczyć średnią. Wartość dla tego współczynnika wynosi 2.762. Porównując wynik z *training RMSE* wyliczonym dla modelu wykorzystującego wszystkie dane, który wynosi 2.395, stwierdzam, że model gorzej radzi sobie z predykcją niż z tłumaczeniem zależności.

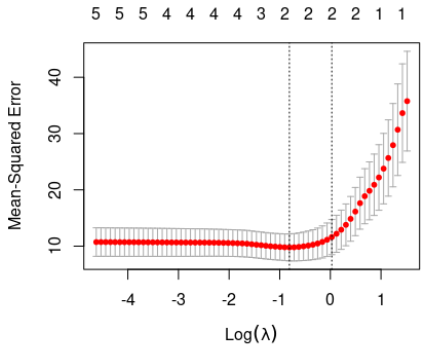
Pomimo stosunkowo niewielkiej różnicy między współczynnikami uczącymi i testowymi stwierdzam, że model nie powinien być stosowany do przewidywań na podstawie nowych obserwacji. Należy mieć również na uwadze fakt, że zbiór, na którym zbudowany jest model, posiada końcowo tylko 26 obserwacji, zatem wykorzystanie go jako podstawę do predykcji przyszłych wartości może okazać się niestabilne, szczególnie jeśli nowe obserwacje będą znacząco odbiegać od tych użytych do kalibracji modelu. Przykładem takich obserwacji mogłyby być dane dla państw z innego kontynentu, np. Afryki lub Ameryki Południowej, dla których najprawdopodobniej model nie dałby adekwatnych wyników.

5.4 Metoda LASSO

Końcowym etapem oceny modelu będzie porównanie finalnego modelu z modelem otrzymanym przez zastosowanie metody LASSO dla wszystkich wyjściowych zmiennych objaśniających. Metoda LASSO (*least absolute shrinkage and selection operator*) jest metodą analizy regresji, która równocześnie dokonuje doboru zmiennych oraz ich regularyzacji w celu zwiększenia dokładności przewidywania i interpretacji modelu statystycznego. Nadaje ona karę równą sumie wartości bezwzględnej wielkości współczynników β. Ten rodzaj regularyzacji może finalnie skutkować modelem z małą ilością współczynników. Niektóre z nich mogą zostać zupełnie wyzerowane i wyeliminowane z modelu. Większe kary skutkują wartościami współczynników zbliżonymi do zera, co jest korzystne dla tworzenia prostszych i lepiej interpretowalnych modeli. Bardziej matematyczny zapis wygląda następująco:

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

gdzie celem jest minimalizacja powyższego wyrażenia.
Do zastosowania metody LASSO użyłem funkcji *cv.glmnet* z pakietu *glmnet*. λ minimalizująca *test MSE* wynosi 0.4452581, co przedstawia Rysunek 8.



Rys. 8: Wykres porównujący $log(\lambda)$ do testowego błędu średniokwadratowego

Współczynniki modelu LASSO otrzymane z funkcji *glmnet* wynoszą: dla wyrazu wolnego 8.4345, dla zmiennej HDD -0.002429, dla zmiennej GDP 0.0003187. Dla pozostałych zmiennych współczynniki zostały wyzerowane.

Miary, na podstawie których porównywać będę dwa modele to współczynnik R^2 oraz *test MSE*, które dla modelu otrzymanego z metody LASSO wynoszą odpowiednio: 0.78 dla R^2 i 9.774 dla *test MSE*. Dla finalnego modelu współczynniki te wynosiły R^2 na poziomie 0.8296 oraz wyliczone w podrozdziale 5.3 *test MSE* równe 8.74. Z tak krótkiej i podstawowej analizy wynika, iż model finalny, który uwzględnia zmienną GDP podniesioną do kwadratu lepiej minimalizuje testowy błąd średniokwadratowy oraz posiada większy współczynnik R^2.

6 Komentarz końcowy

6.1 Parametry finalnego modelu

Podsumowując, finalny model opisuje szukaną zależność w następujący sposób:

$$Natural\_gas\_price = 14.8 - 0.00319 \cdot HDD + 4.798 \cdot 10^{-9} \cdot GDP^2 + \epsilon$$

Podstawowe statystyki tego modelu wynoszą:

- Skorygowany współczynnik R^2 równy 0.8148
- *Training MSE* równy 5.73
- *Residual standard error* równy 2.546

Wszystkie zmienne łącznie z wyrazem wolnym są istotne statystycznie oraz model spełnia założenia modelu liniowego, jak zostało to pokazane w rozdziale 5.1.

6.2 Interpretacja

Głównym celem projektu było znalezienie modelu, który potrafi wytłumaczyć zmienność ceny gazu ziemnego w zależności od państwa.

Zważając na pośrednie modele testujące pozostałe zmienne, można stwierdzić, że cena gazu w danym państwie zależy głównie od indeksu HDD oraz $GDPpercapita$. Takie wnioski pokrywają się z intuicją, ponieważ patrząc na konstrukcję indeksu HDD widać, że ze względu na sposób, w jaki się go wylicza, jest on bezpośrednio związany z "systemem grzewczym" danego państwa. Ujemny znak współczynnika β sugeruje, że im więcej dni grzewczych w danym państwie, tym niższa powinna być w nim cena gazu ziemnego. Dodatni znak współczynnika β przy zmiennej GDP^2 wskazuje jednak, że im bogatsze jest państwo, tym cena gazu ziemnego jest w nim wyższa, co jest zaskakującym wnioskiem i należałoby zwrócić na niego szczególną uwagę. Brak istotności statystycznej zmiennej *Dependency_import* również jest wynikiem nieoczekiwanym i warto byłoby pochylić się nad głębszą analizą tej zmiennej i jej wpływu na badane modele.

Analiza założeń oraz współczynników tego modelu napawa optymizmem i nie wykazuje większych problemów z badanym modelem. Należy jednak zwrócić szczególną uwagę na bardzo specyficzny i minimalistyczny zbiór danych, na których model został zbudowany. Przypominając rozważania z rozdziału 5.3, nie ma pewności i nic nie wskazuje na to, że model ten zachowa swoje właściwości i dobre cechy, kiedy to analizie poddamy państwa o innych charakterystykach niż te europejskie.

Podsumowując, model poradził sobie zadowalająco z postawionym mu zadaniem wytłumaczenia zmienności cen gazu ziemnego w krajach Europy. Co należy jeszcze raz podkreślić, model ten nie ma zadowalającej mocy prognostycznej i nie należy uznawać wyników dla innych krajów niż europejskie jako poprawne. Głównym problemem modelu jest zdecydowanie zbyt skromny i zbyt mało różnorodny zbiór dostępnych danych. Gdyby udało się go wyeliminować i posiadać dostęp do obszernej bazy danych, analiza i wnioski przedstawione w tej pracy mogłyby być dogłębniejsze i nieść za sobą inne ciekawe wyniki.