Wpływ czynników socjodemograficznych na procent ważnych głosów w wyborach parlamentarnych we Włoszech z 2018 roku

Anna Wieczorek

May 2020

Spis treści

1 Wstęp

Projekt ten jest poświęcony tematyce ważnych głosów oddanych w wyborach. Analizie został poddany odsetek głosów ważnych w wyborach parlamentarnych w poszczególnych regionach we Włoszech w roku 2018. Podział na regiony został dokonany wg. NUTS 3, czyli wyodrębnienia jednostek terytorialnych np. prowincji we Włoszech, na obszarze Unii Europejskiej na potrzeby statystyki.
Omawiane zagadnienie jest ciekawe z punktu widzenia nauk socjologicznych, jak również może wydawać się interesujące dla politologów. Obie te grupy zawodowe może zastanawiać czy oddawanie kart do głosowania błędnie wypełnionych lub pustych jest czysto przypadkowym zjawiskiem. W tej pracy postaram się przeanalizować to, czy ilość nieważnych głosów zależy od pewnych czynników socjodemograficznych jak zagęszczenie ludności lub wzrost PKB, czy może jest to dzieło jedynie przypadku, na przykład nieuwagi wyborców.

2 Opis danych

Dane użyte do analizy pochodzą z witryn internetowych
data.europa.eu/euodp/en/data oraz elezionistorico.interno.gov.it, czyli odpowiednio z portalu zawierającego dane statystyczne Unii Europejskiej i strony rządowej Włoch z danymi na temat wyborów.

2.1 Opis zmiennych

Wyjściowy zbiór danych zawiera następujące zmienne:

- REG - numer prowincji wg. podziału NUTS 3,
- POPULATION - wielkość populacji w prowincji,
- DENSITY - zagęszczenie ludności na 5 km^2,
- VOTES VALID - liczba ważnych głosów w prowincji,
- INVALID - liczba nieważnych (lub pustych) głosów w prowincji,
- AROP EST - procent społeczeństwa zagrożony ubóstwem w prowincji, wartość szacowana na rok 2011, źródła: Eurostat, World Bank, ESPON, DG REGIO,
- N3 GDPPC - PKB per capita w prowincji, w umownej walucie PPS,
- N2 EMP RATE - poziom zatrudnienia w regionie wg. podziału NUTS 2, wyrażony jako procent spośród populacji regionu w wieku 20-64 lat,
- N2 TERT EDUC - procent populacji w wieku 25-64 lat w regionie, z wyższym wykształceniem,
- N3 GDP PC GROWTH - średni roczny wzrost procentowy PKB per capita w prowincji w latach 2000-2014,
- N3 EMP GROWTH - średni roczny wzrost procentowy poziomu zatrudnienia w prowincji w latach 2000-2014,
- NAT GROWTH - średni roczny przyrost naturalny w prowincji w latach 2000-2016,
- NET MIGRATION - średnia roczna migracja netto w prowincji, jako procent populacji, w latach 2000-2016,
- SH BORN REP CNTR - część populacji prowincji pochodząca z Włoch, jako procent całej populacji prowincji,
- SH BORN OUTSIDE EU - część populacji prowincji pochodząca spoza obszaru Unii Europejskiej, jako procent całej populacji prowincji,
- SH RES 1Y SAME - część populacji, która nie zmieniła miejsca zamieszkania na rok przed zbieraniem danych, jako procent całej populacji prowincji.

Początkowo zmienną objaśnianą miała być ilość ważnych głosów oddanych w prowincji, zatem zmienna VOTES VALID. Jednakże, ze względu na, wynikającą w naturalny sposób, bardzo dużą korelację z populacją prowincji, zdecydowano się na zmianę zmiennej objaśnianej na procent ważnych głosów, spośród wszystkich oddanych kart. W ten sposób powstała zmienna SH VALID, wg. wzoru:

$$SH\ VALID = \frac{VOTES\ VALID}{VOTES\ VALID + INVALID} * 100.$$

Jako, że celem analizy jest wyjaśnienie zjawiska powstawania ważnych i nieważnych głosów, jest to naturalny kandydat na zmienną objaśnianą. Oczywiście, mogłaby to być również zmienna wyjaśniająca procent nieważnych głosów w wyborach, byłby to wektor dany jako $100 - SH\ VALID$.

2.2 Podstawowe statystyki danych

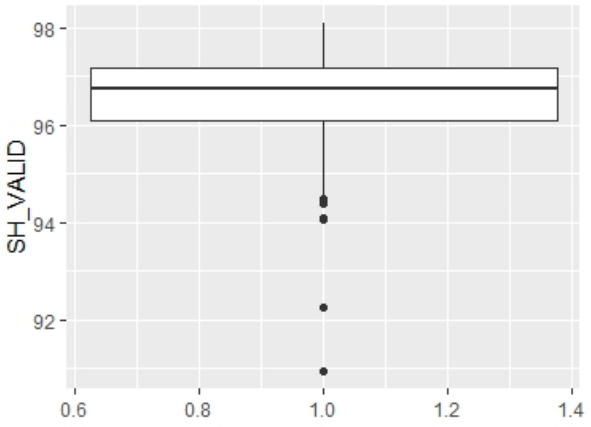
Przy użyciu funkcji `summary` na pierwotnym zbiorze danych (czyli zbiorze bez zmiennej SH VALID), wyświetlone zostały podstawowe statystyki każdej zmiennej, co pozwoliło zidentyfikować już podstawowe problemy.
W zmiennej REG widać, że jeden kod prowincji: ITG27, pojawia się dwa razy w zbiorze danych. Nie powinno to mieć miejsca, ponieważ wszystkie regiony mają indywidualne kody. Oznacza to, że albo zbiór zawiera dwa razy ten sam wiersz, albo występuje tam inny problem, na przykład jeden region został podzielony na dwa mniejsze. Sytuacja ta byłaby niepożądana, ponieważ analiza opiera się na podziale NUTS 3.
Widać również braki danych w zmiennych POPULATION (jedna brakująca obserwacja) oraz INVALID (dwa braki danych).

3 Obróbka danych i dobór zmiennych

3.1 Czyszczenie danych oraz uzupełnianie brakujących informacji

Na początku poprawiony został błąd w zmiennej REG. Gdyby dwa wiersze były takie same, to jeden trzeba byłoby odrzucić, ponieważ dwie takie same obserwacje w macierzy danych spowodują singularność macierzy $X^T X$, wykorzystywanej w liczeniu estymatora OLS. Jeżeli natomiast wartości są różne, to najprawdopodobniej region został sztucznie podzielony na dwie części podczas zbierania danych.
W tym wypadku mamy do czynienia z drugim przypadkiem, czyli z dwoma różnymi obserwacjami. Może być to także błąd w kodzie regionu, ale liczba nieważnych głosów jest taka sama, co jest mało prawdopodobne. Założyłam zatem, iż te wartości to jeden region podzielony na dwa. Aby uporać się z tym problemem, obserwacje zostały połączone w jedno, poprzez wzięcie sumy, tam gdzie miało to sens (wielkość populacji i ilość ważnych głosów), średniej w przypadku gęstości zaludnienia oraz poprzez pozostawienie pozostałych wartości (poza zmienną INVALID) skopiowanych z jednego z wierszy, ponieważ były one zbierane dla regionu NUTS 3 lub NUTS 2, zatem były takie same w obu wierszach. Została jednak jeszcze wartość zmiennej INVALID. Ponieważ w obu przypadkach była taka sama, co wydawało się być podejrzane, zdecydowałam się na zostawienie tam braku wartości, aby możliwie nie zaburzyć obrazu zmiennej, poprzez wstawienie niepasującej wartości.

Postanowiłam nie uzupełniać braków danych w zmiennej INVALID, ze względu na dużą zależność wartości liczbowej ilości głosów nieważnych od ilości głosów ważnych i populacji. Takie uzupełnienie mogłoby dać nienajlepsze efekty, np. mogłoby skutkować nienaturalnie dużą liczbą głosów w porównaniu do populacji, czy też dużym procentem głosów nieważnych. Z tych względów, braki uzupełniłam już po stworzeniu zmiennej SH VALID, poprzez wstawienie w odpowiednie miejsca mediany tej zmiennej. Mediana wydawała się być lepszym kandydatem niż średnia, ze względu na obserwacje odstające (Rysunek 1.), które są



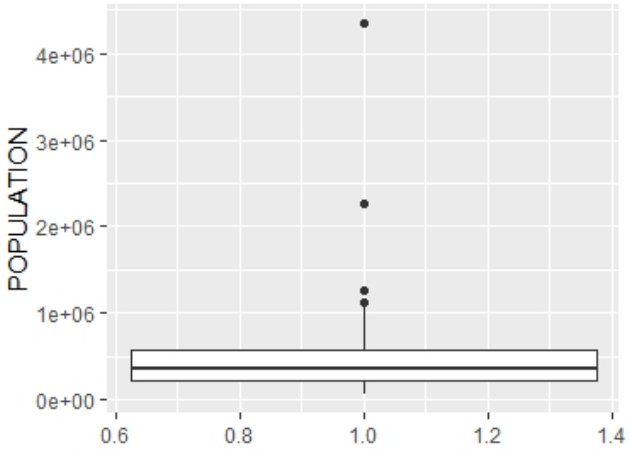
Rysunek 1: Wykres pudełkowy zmiennej SH VALID. Jest kilka wartości odstających w dół, mogących zaniżać średnią.

widoczne i niektóre z nich odstają dosyć mocno w dół, co
może istotnie zaniżać średnią, natomiast mediana, jako statystyka pozycyjna pozwala uniknąć tego problemu.

Pozostały wartości brakujące w zmiennej POPULATION. Ponownie, ze względu na charakter zmiennej, mamy do czynienia z wartościami odstającymi, tym razem w górę (Rysunek 2.). Nie powinno to dziwić, ponieważ we Włoszech są prowincje, gdzie jest dużo wiosek i małych miast, ale jest też kilka prowincji, w których mieszczą się wielkie, milionowe miasta, jak Rzym czy Mediolan. Z tego powodu uzupełnienie braków danych medianą z populacji wydaje się być lepszym wyborem niż średnia, która może być mocno zawyżona (średnia wynosi 497192, podczas gdy mediana to 360125).

3.2 Wybór zmiennych do modelu

W pierwotnym zbiorze danych mamy 15 zmiennych numerycznych oraz 1 zmimenną kategoryczną (kod regionu). Skoro każda prowincja ma swój indywidualny kod, to zmienna z owym kodem nie niesie żadnych informacji oprócz pewnego identyfikatora. Spodziewamy się zatem, że nie jest to zmienna istotna statystycznie i na tym etapie analizy można z niej zrezygnować. Z kolei zmienne VOTES VALID i INVALID nie będą potrzebne, ponieważ zostały już wykorzystane do stworzenia nowej zmiennej SH VALID. Mogłyby natomiast powodować problem ze współliniowością, w połączeniu z nową zmienną. Zatem tych trzech zmiennych nie uwzględniałam w dalszej analizie, dlatego nowy zbiór danych zawiera zmienne ze zbioru pierwotnego, poza tymi trzema kolumnami, natomiast dodajemy zmienną SH VALID.



Rysunek 2: Wykres pudełkowy zmiennej POPULATION. Wyraźnie widać wartości odstające - kilkumilionowe prowincje.

Po krótce zostaną teraz omówione wszystkie zmienne ze zbioru danych.
Zacznę od zmiennej objaśnianej SH VALID. Zgodnie z tym, co zostało zauważone już wcześniej, mamy do czynienia tu z kilkoma obserwacjami odstającymi - widoczny na Rysunku 1. wykres pudełkowy ujawnia 5. Istnienie wartości nietypowo małych widać również w podstawowych statystykach zmiennej, gdzie najmniejsza wartość wynosi 90.95, pierwszy kwartyl to 96.10, a trzeci 97.17. Średnia jest również zaniżona w porównaniu do mediany - średnia to 96.47, a mediana 96.75. Różnica nie jest duża, jako wartość, jednakże w odniesieniu do rozstępu kwartylowego wynoszącego 1.07, wydaje się to być znaczącą różnicą.
Przejdę teraz do zmiennych objaśniających.

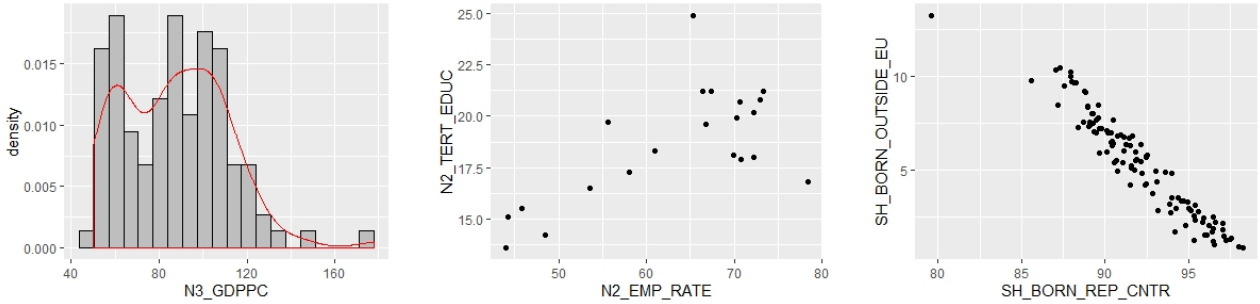
- POPULATION - zgodnie z tym, co było zaznaczone wcześniej, są regiony, które istotnie się wyróżniają pod względem liczby ludności. Ciekawą informacją (również w kontekście gęstości zaludnienia) mogłaby się okazać wielkość prowincji, ale tych danych tu nie mamy.
- DENSITY - w przypadku gęstości zaludnienia obserwujemy podobną sytuację jak w przypadku populacji. Spodziewamy się, iż może tu istnieć duża korelacja, ponieważ najczęściej w dużych miastach, czyli tam gdzie mieszka duża ilość ludzi, jest również duże zagęszczenie ludności, głównie ze względu na blokowiska, podczas gdy w małych miejscowościach zagęszczenie jest nieduże.
- AROP EST - nie ma wartości odstających, jest to spowodowane dużym rozstępem kwartylowym: pierwszy kwartyl na poziomie 10.83, a trzeci na poziomie 26.45. Różne regiony są inaczej narażone na ubóstwo, w dużych miastach często bywają dzielnice, gdzie poziom biedy jest dosyć duży, natomiast małe miejscowości, gdzie nie ma turystyki, ani przemysłu również często cechują się wysokim poziomem niezamożnych mieszkańców.

- N3 GDPPC - na histogramie z gęstością (Rysunek 3a.) widać dwa skupiska: w okolicy wartości 50 oraz w okolicy 100. Uwagę zwraca również jedna wartość odstająca w górę.
- N2 EMP RATE - mniej różnych wartości, ponieważ mamy tu podział NUTS 2, zatem kilka prowincji ma te same wartości zmiennej. Nie ma wartości odstających, większość prowincji ma zatrudnienie pomiędzy 55% i 70%.
- N2 TERT EDUC - mediana i średnia są na poziomie ok. 18%, nie ma wartości odstających, ale jest jedna wartość nietypowa ok. 25%, może to być duże miasto, gdzie przyjeżdżają ludzie na studia lub już wykształceni i zostają. Na wykresie zmiennych N2 EMP RATE vs N2 TERT EDUC (Rysunek 3b.) widać zależność liniową między tymi zmiennymi, duża korelacja została też potwierdzona poprzez obliczenie współczynnika korelacji Pearsona, który wynosi tu ok. 0.79.
- N3 GDP PC GROWTH - wg. obliczeń występują dwie wartości odstające w zmiennej, jednakże nie jest to widoczne bez obliczeń, rozkład jest dosyć symetryczny. Średnia wynosi -0.7006, może to wskazywać na ogólną tendencję spadkową, jeśli chodzi o gospodarkę Włoch, choć może się również okazać, że w ogólnym rozrachunku wzrost PKB kraju jest dodatni.
- N3 EMP GROWTH - tu rozkład jest relatywnie symetryczny, jest jedna wartość odstająca, ale tak jak poprzednio nie jest ona wyraźnie widoczna. Korelacja pomiędzy tą zmienną, a N3 GDP PC GROWTH nie jest bardzo duża, ponieważ jest to ok. 0.48, warto było to jednak sprawdzić, ponieważ można się spodziewać, że tam, gdzie wzrost gospodarczy jest większy, zwiększa się poziom zatrudnienia ludności.
- NAT GROWTH - rozkład dość symetryczny, bez wartości odstających. Wszytskie trzy kwartyle, to wartości poniżej zera, co nie dziwi, ponieważ taka jest tendencja w krajach rozwiniętych, a Włochy do tejże grupy należą.
- NET MIGRATION - tu z kolei wszystkike trzy kwartyle są dodatnie, co sugeruje, że więcej osób osiedla się we Włoszech, niż emigruje.
- SH BORN REP CNTR - jedna wartość dość mocno odstaje w dół, co oznacza, że stosunkowo mało mieszka Włochów w tej prowincji. Można się spodziewać, że będzie to Rzym. Jest duża (na moduł) korelacja (-0.7600736) ze zmienną NET MIGRATION, czego można było się spodziewać, gdyż niewielki odsetek Włochów zamieszkujących prowincję oznacza, że dużo osób spoza kraju musiało się tam osiedlić.
- SH BORN OUTSIDE EU - jedna wartość nietypowa, ale nie jest ona obserwacją odstającą. Bardzo duża (na moduł) korelacja ze zmienną SH BORN REP CNTR (niemalże -1), ponieważ te zmienne opisują bardzo podobne zjawisko, czyli osadzanie się w karju imigrantów. Zależnośc widoczna na Rysunku 3c.
- SH RES 1Y SAME - poza jedną wartością odstającą nie widać anomalii.

Z powodu bardzo dużej korelacji zmiennej SH BORN REP CNTR ze zmienną SH BORN OUTSIDE EU oraz dużej korelacji z NET MIGRATION (SH BORN OUTSIDE EU skorelowane z NET MIGRATION niewiele mniej), zmienna ta nie zostanie przeze mnie dopuszczona do modelu, ponieważ bardzo możliwe, że spowoduje współliniowość. Ponadto, zmienne POPULATION i DENSITY zdecydowałam zamienić na zmienne kategoryczne, ponieważ zawierają one wartości mocno odstające, a przypuszczam, że wielkość populacji (tak samo zagęszczenie zaludnienia) od pewnej wartości może nie mieć takiego znaczenia - duże miasta mogą się różnić nawet kilkukrotnie ilością mieszkańców, ale struktury społeczne są bardzo podobne, dlatego też wprowadzam w obu zmiennych trzy poziomy: "small", "medium" i "large", przy czym używam proporcji 20/60/20, czyli 20% najmniejszych wartości ma poziom "small", 20% największych poziom "large", a pozostałe 60% - poziom "medium".

4 Budowa i kalibracja modelu

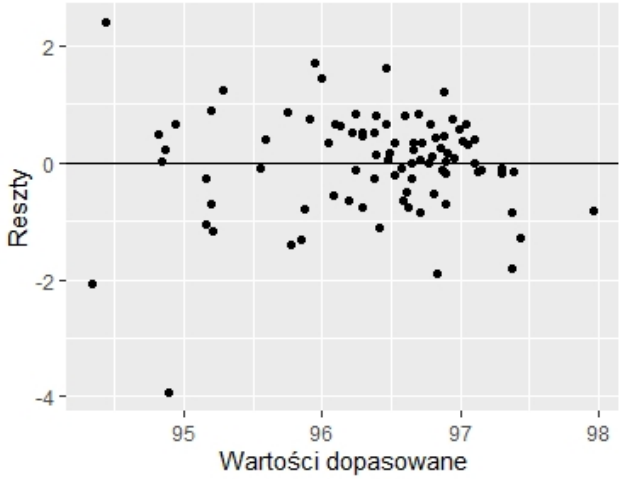
Skorzystałam z walidacji krzyżowej. W tym celu zbiór danych podzieliłam na zbiór uczący i zbiór testowy w proporcji 80/20. Mój model ma na celu wyjaśnić pewne zjawisko, a nie dokonywac predykcji, zatem metody tej używam



(a) Histogram N3 GDPPC (b) N2 EMP RATE vs N2 TERT EDUC (c) SH BORN REP CNTR vs SH BORN OUTSIDE EU

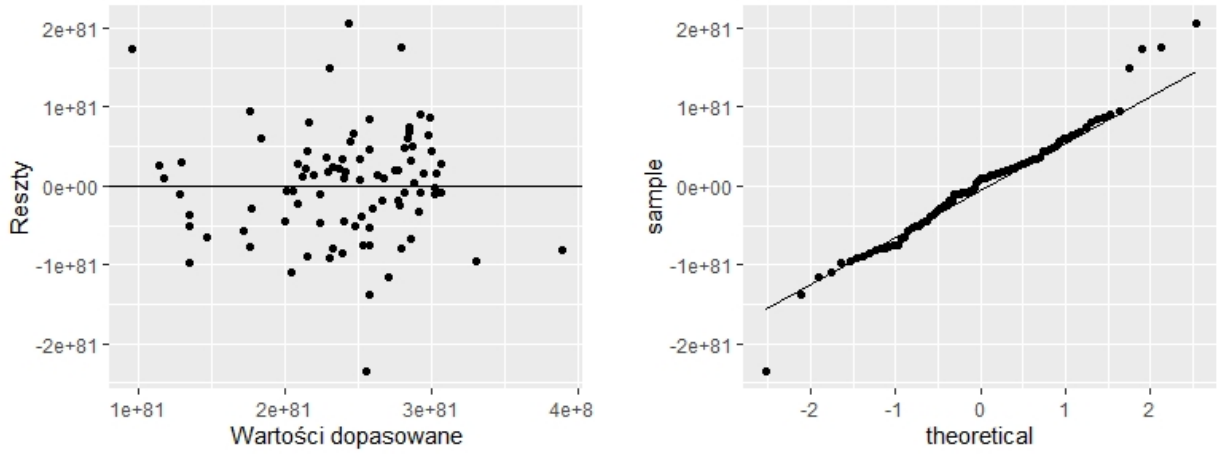
Rysunek 3: Z lewej: na histogramie zmiennej N3 GDPPC wraz z estymatorem jądrowym gęstości (na czerwono) widać dwa skupiska oraz wartość odstającą w górę. Po środku: na wykresie zależności zmiennych N2 EMP RATE vs N2 TERT EDUC widać wyraźnią korelację. Z prawej: na wykresie zależności zmiennych SH BORN REP CNTR vs SH BORN OUTSIDE EU widać bardzo mocną korelację.

do sprawdzenia dopasowania modelu.
Pierwszy model tworzę ze wszystkimi zmiennymi ze zbioru danych. W takim modelu jedynie SH RES 1Y SAME okazała się być zmienną istotną statystycznie. Dopasowane R^2 było na poziomie ok. 0.29, co wydaje się nie być złą wartością dla pierwszego modelu dla tej zależności. P-value testu F kazało odrzucić hipotezę o braku zależności liniowej między zmienną objaśnianą, a zmiennymi objaśniającymi, choć nie jest to bardzo mała wartość p-value (0.0001735). Dla wielu zmiennych błąd standardowy SE jest duży, np. kilkukrotnie większy niż wartość wyestymowanego współczynnika. Wartości VIF dla poszczególnych zmiennych sugerują problem ze współliniowością. Szczególnie dużą wartość obserwujemy dla zmiennej N2 EMP RATE (ok. 13.44), dla zmiennych NET MIGRATION, AROP EST, N3 GDPPC, SH BORN OUTSIDE EU, N3 EMP GROWTH jest to ponad 5. Niektóre zmienne są silnie skorelowane, jak AROP EST i N2 EMP RATE.
Model ten nie spełnia założenia homoskedastyczności reszt, co wynika z analizy wykresu residuów w zależności od wartości dopasowanych (Rysunek 4.), gdzie widać malejącą wariancję czynnika losowego. Niespełnione jest także założenie o normalności reszt, co widać po odstępstwach punktów od prostej kwantyli teoretycznych rozkładu normalnego na wykresie kwantyl-kwantyl (Rysunek 5.). Przy użyciu funkcji `Box.test` przeprowadzony został test Boxa-Pierce'a badający autokorelację czynnika losowego poprzez sumowanie kwadratów współczynników korelacji między błędami oddalonymi o P, gdzie P to maksymalny rząd autokorelacji, jaki chcemy badać, następnie sumę mnożąc razy liczbę obserwacji. Zatem, skoro chcemy, aby nie było autokorelacji, to statystyka testowa powinna być możliwie mała. W tym przypadku statystyka testowa wynosiła 0.23039, co dało p-value równe 0.6312 i nie dało podstaw do odrzucenia hipotezy o braku autokorelacji składnika losowego modelu.



Rysunek 4: Wykres reszt w zależności od wartości dopasowanych modelu pierwszego.

Dobierając zmienne do następnych modeli posługiwałam się funkcją `step`, która wykonuje algorytm krokowy w tył, za kryterium przyjmując kryterium informacyjne Akaike (AIC), które mówi o tym, jak dużą część informacji tracimy w stosunku do prawdziwej macierzy β, opierając się na logarytmie naturalnym z RSS przeskalowanym przez $\frac{1}{n}$. Zatem algorytm ten usuwa z modelu zmienne tak, aby osiągnąć jak największą poprawę kryterium. W tym wypadku, funkcja została użyta dla modelu z już usuniętymi zmiennymi: N2 EMP RATE (ze względu na dużą korelację z innymi zmiennymi), NET MIGRATION (duża korelacja ze zmienna SH BORN OUTSIDE EU, a SH BORN OUTSIDE EU



(a) Wartości dopasowane vs reszty

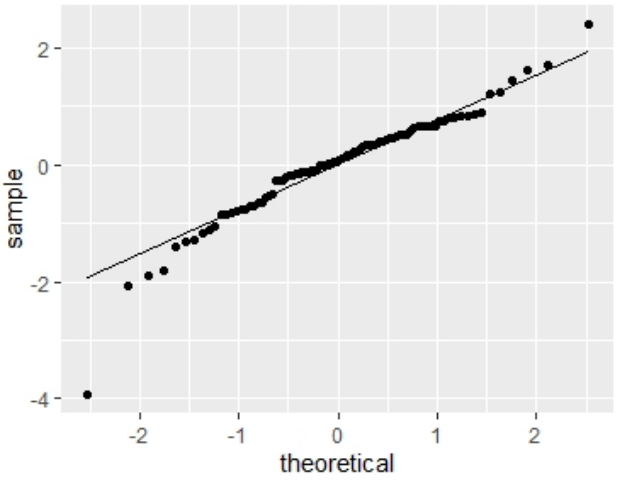
(b) Wykres Q-Q

Rysunek 6: Z lewej: wykres zależności reszt modelu od wartości dopasowanych. Nie widać tu wyraźnego trendu, co sugeruje spełnianie założenia o homoskedastyczności residuów. Z prawej: wykres kwanty-kwantyl. Widać odstępstwa od kwantyli teoretycznych w ogonach rozkładu, choć w środku rozkład wygląda na zgodny z normalnym.

może być istotna statystycznie), N3 EMP GROWTH (ze względu na bardzo duży błąd standardowy w porownaniu do wartości współczynnika oraz duże p-value testu t). Dodatkowo, algorytm sugerował największą poprawę przy pozostawieniu jedynie zmiennych POPULATION, AROP EST, SH BORN OUTSIDE EU, N3 GDPPC i SH RES 1Y SAME.
Po dalszej eliminacji zmiennych, z powodu dużego ich skorelowania i możliwego problemu ze współliniowością w modelu, ostatecznie wybrane zostały trzy zmienne objaśniające: POPULATION, SH BORN OUTSIDE EU i SH RES 1Y SAME. Oznacza to, że na procent ważnych głosów może wpływać wielkość prowincji, procent osób, które pochodzą spoza Unii Europejskiej oraz procent populacji, które nie zmieniły miejsca zamieszkania na rok przed zbieraniem danych. Istotne jest to, że w przypadku mniejszych i średnio dużych prowincji znaki współczynników modelu są ujemne, przy czym wartość jest mniejsza dla małych regionów. Wg. naszej analizy im mniejsza (pod względem liczby ludności) jest prowincja, tym stosunkowo więcej w niej nieważnych głosów. Pozostałe współczynniki są dodatnie, zatem im więcej osób pochodzi zza granic UE, tym większy jest odsetek ważnych głosów w wyborach. Większy jest jeszcze współczynnik dla SH RES 1Y SAME, zatem można przypuszczać, iż stateczność (w sensie miejsca zamieszkania) pociąga za sobą większą ilość ważnych głosów.
Wszytskie modele od tego momentu tworzone były z tymi trzema zmiennymi objaśniającymi.

Wariancja błędów zdawała się maleć wraz ze wzrostem wartości dopasowanych (Rysunek 4.), dlatego też postanowiłam zastosować transformację zmiennej objaśnianej z y na $log(y)$, ze względu na podejrzenie błędu multiplikatywnego, jednakże nie poprawiło to modelu. Następną transformacją zmiennej SH VALID, którą zdecydowałam się zastosować, była transformacja Boxa-Coxa. Przy użyciu funkcji `boxcox`, która rysuje wykres funkcji log-największej wiarygodności dla transformacji potęgowej, zaznaczając maksimum oraz 95% przedział ufności, najlepszą wartością okazało się być 41.

Oznacza to, iż zmienną objaśnianą należało podnieść do potęgi 41. Takie przekształcenie dało istotnie lepsze dopasowane R^2 (na poziomie 0.34), wszytskie trzy zmienne okazały się być istotne statystycznie, p-value testu F znacznie się



Rysunek 5: Wykres kwantyl-kwantyl dla reszt modelu pierwszego. Widoczne są odstępstwa od linii kwantyli teoretycznych rozkładu normalnego.

zmniejszyło, tym bardziej powodując odrzucenie o braku liniowej zależności.

Ponadto, wykres zależności reszt od wartości dopasowanych nie sugerował już heteroskedastyczności reszt (Rysunek 6a.), taki sam wniosek sugerował test Breuscha-Pagana, opierający się na dopasowaniu nowego modelu do kwadratów residuów w zależności od zmiennych objaśniających i sprawdzeniu poziomu R^2 przeskalowanego. P-value testu wyniosło 0.4466, co nie dało podstaw do odrzucenia hipotezy zerowej o homoskedastyczności. Co więcej, widać również poprawę w założeniu o normalności błędów. Wykres kwantyl-kwantyl (Rysunek 6b.) wygląda znacznie lepiej niż poprzednio, jednakże nadal widać grube ogony rozkładu.

Jednakże, model ten jest dosyć skomplikowany, potęga 41 zmiennej objaśnianej utrudnia interpretację niematematyczną modelu: operujemy na ogromnych wielkościach, w prawdzie możemy analizować znaki współczynników, ale same liczby niewiele mówią w tej skali.

Jako, że zmienna objaśniana jest procentem, możemy (po podzieleniu jej na 100) zastosować transformację logitową zmiennej objaśnianej, czyli zamienić prawdopodobieństwo oddania ważnego głosu, jak możemy również w przybliżeniu interpretować zmienną objaśnianą, na logarytm z szans na oddanie takiego głosu. Używamy zatem funkcji $logit(p) = \frac{p}{1-p}$. Model po dokonaniu takiej transformacji zyska również na prostocie interpretacji - wartości nie będą duże, a znaki współczynników można od razu czytać jako wzrost lub spadek szans.

Po zbudowaniu takiego modelu dopasowanie nieznacznie się pogorszyło (adj. R^2 spadło do 0.326), natomiast wszystkie zmienne były znów istotne statystycznie, odrzuciliśmy także hipotezę o braku zależności liniowej (p-value testu F wynosiło 1.728e-07). Po wstępnej analizie można przypuszczać, iż założenia modelu są spełnione, poza normalnością błędów. Test RESET Ramsey'a sugerował poprawę modelu przy dodaniu sześcianu którejś ze zmiennych objaśniających, jednakże po dalszej analizie nie okazało się to dobrym pomysłem (problem ze współliniowością, gorsze dopasowanie po zastąpieniu zmiennej SH BORN OUTSIDE EU jej sześcianem, dla SH RES 1Y SAME praktycznie brak zmiany, ze względu na prawie jednostkową korelację między nią i jej sześcianem).

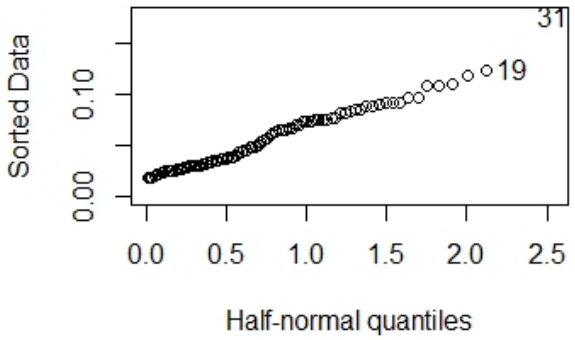
Porównałam modele po transformacji Boxa-Coxa i po transformacji logitowej za pomocą dwóch statystyk: RMSE liczonym po dokonaniu predykcji dla zbioru testowego oraz predictive R^2. Ze względu na różne skale, w jakich wyszło RMSE (ze względu na różne skale zastosowane w modelach), zdecydowałam się podzielić obie wartości przez średnie z odpowiednich wartości rzeczywistych (wartości SH VALID[41] oraz logit z SH VALID*0.01). Po pomnożeniu razy 100, otrzymałam wartości odpowiednio ok. 18 i ok. 5. Następnie porównałam predictive R^2. Jest to statystyka mówiąca o dopasowaniu modelu do obserwacji po usunięciu jej ze zbioru. Różni się ona od R^2 tym, że RSS w definicji zastępujemy przez PRESS, dane wzorem $\sum_{i=1}^{n}(\frac{\varepsilon_i}{1-h_i})^2$, gdzie h_i to dźwignia i-tej obserwacji. Sprawdzamy w ten sposób, jak dużą część wariancji wyjaśniamy dla nowych obserwacji, czyli takich, na których model nie był uczony. Pomaga to pokazać ewentualny problem przetrenowania modelu. Dla naszego pierwszego modelu wartość ta wynosiła ok. 0.30, a dla drugiego ok. 0.29. Jest to porównywalna strata w stosunku do adj. R^2 obu modeli (ok. 4 punkty procentowe).

5 Diagnostyka i weryfikacja modelu finalnego

Ze względu na lepszą interpretowalność i dobre własności modelu z transformacją logitową, na model finalny wybrałam ten właśnie model. Pod względem dopasowania wydaje się być on podobny do modelu po transformacji Boxa-Coxa, jednkaże jego prostota wydaje się przeważać.

5.1 Identyfikacja obserwacji nietypowych

Po policzeniu dźwigni dla każdej z obserwacji za pomocą funkcji `hatvalues` oraz narysowaniu wykresu pół-normalnego dla wektora dźwigni (Rysunek 7.) zidentyfikowane zostały dwie obserwacje wpływowe: jedna, która niewiele się różni od pozostałych oraz jedna różniąca się wyraźnie. Jednak obie dźwignie były większe niż wartość $2\frac{k}{n}$, zatem wymagały zwrócenia na nie uwagi.



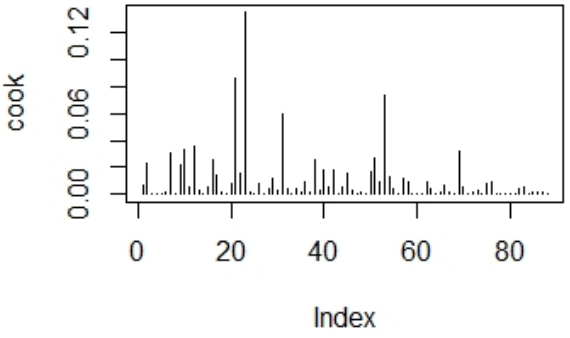
Rysunek 7: Wykres pół-normalny dla dźwigni. Wyróżniające się indeksy 31 i 19 wskazują na duże wartości.

Następnie, poprzez analizę wykresu odległości Cooka dla każdej obserwacji (Rysunek 8.) wyznaczone zostały trzy obserwacje o największych wartościach tej statystyki - żadna z nich nie pokrywała się z obserwacjami o najwyższej dźwigni. Największa odległość Cooka jest prawie dwa razy większa od drugiej największej, ale nie jest to też duża wartość (ok. 0.13), chociaż warto się jej przyjżeć. Przeprowadzony test polegający na znalezieniu największych reszt studentyzowanych i decyzji czy mamy do czynienia z obserwacją odstającą na podstawie poprawki Bonferroni'ego sugeruje, iż mamy do czynienia z jedną wartością odstającą. Jest to też jedna z obserwacji o największej odległości Cooka. Była to obserwacja odstająca względem zmiennej objaśnianej. Zdecydowałam się usunąć tę obserwację. Jeszcze jedna obserwacja wydała się niepokojąca. Miała największą wartość odległości Cooka i niemałą dźwignię. Ten sam test nie potwierdził, że jest to obserwacja odstająca, ale zwrócił na nią uwagę, jako na największą resztę studentyzowaną. Zdecydowałam się mimo wszytsko tę obserwację usunąć.
Po odrzuceniu tych dwóch obserwacji model znacznie się poprawił - dopasowane R^2 na poziomie 0.46, AIC także znacznie się zmniejszyło.

5.2 Weryfikacja założeń

5.2.1 Liniowa zależność

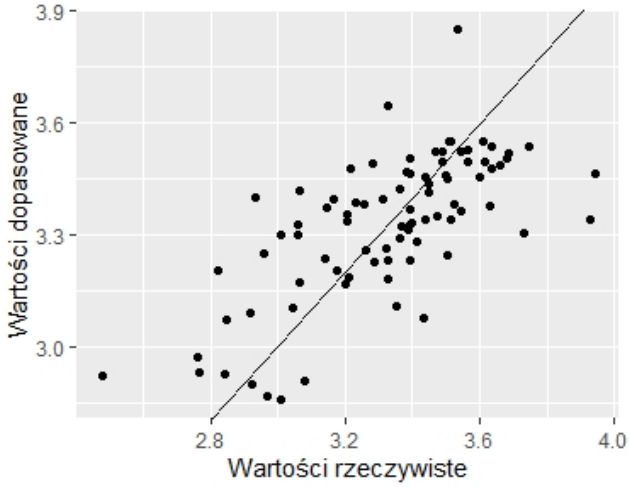
P-value testu F przeprowadzanego przez funkcję `summary` każe odrzucić hipotezę o braku liniowej zależności między logarytmem z szans na oddanie ważnego głosu, a wielkością populacji prowincji, ilością cudzoziemców spoza UE zamieszkujących region oraz ilością ludzi, którzy nie zmienili miejsca zamieszkania na rok przez zbieraniem danych.



Rysunek 8: Wykres odległości Cooka dla poszczególnych obserwacji. Wysoki słupek (duża wartość) w okolicach indeksu 20.

Dodatkowo przeprowadzony został test Rainbow badający, czy skoro (przy założeniu prawdziwości hipotezy zerowej, czyli liniowej zależności) zależność jest liniowa, to na mniejszej grupie obserwacji również jest to prawda. Został przeprowadzony przy użyciu funkcji `raintest` i z p-value na poziomie ok. 0.19, nie dał podstaw do odrzucenia hipotezy o liniowej zależności.
Narysowałam także wykres zależności wartości dopasowanych od prawdziwych wartości zmiennej objaśnianej (po transformacji), widoczny na rysunku nr 9. Niestety, punkty nie układają się dokładnie wzdłuż diagonali, co może świadczyć o istnieniu innego typu zależności, niż wyjaśniana za pomocą modelu.

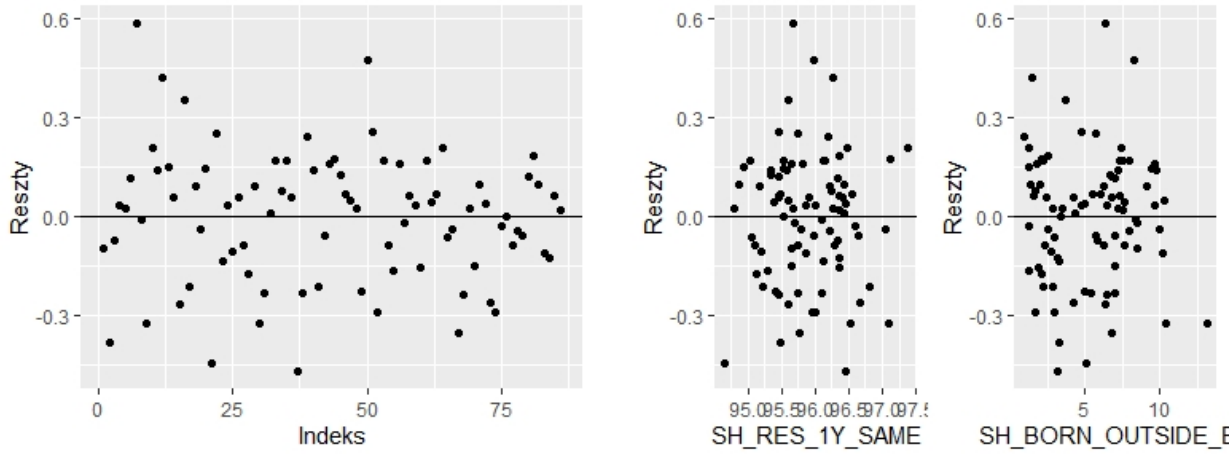
Zdecydowałam się jednak ten model zostawić, ponieważ ani dołączanie potęg, ani transformacje logarytmiczne czy eksponenty zmiennych objaśniających nie pomagały. Jedyną transformacją, która sprawiała wrażenie lepszej, wydawała się transformacja zmiennej objaśnianej z y na $arcsin(\sqrt{y})$, dawała niewiele lepsze dopasowanie (pod względem R^2 i wykresu), ale bardzo cierpiała na tym przekształceniu interpretacja niematematyczna modelu.
Ostatecznie zdecydowałam się na zostanie przy modelu z transformacją logitową, ze względu na prostotę i tylko nieznacznie gorsze dopasowanie do danych.



Rysunek 9: Wykres wartości rzeczywistych zmiennej objaśnianej (po transformacji) vs wartości dopasowanych. Punkty układają się w kształcie pewnej krzywej.

5.2.2 Brak współliniowości

Badanie braku współliniowości danych zaczęłam od zbadania macierzy korelacji zmiennych objaśniających. Wszystkie wartości mieściły się w przedziale (-0.5, 0.5), z wyjątkiem poziomów zmiennej POPULATION, ale to jest ta sama zmienna,



(a) Reszty w zależności od numeru obserwacji

(b) Reszty w zależności od zmiennych objaśniających

Rysunek 11: Z lewej: wykres zależności reszt modelu od indeksów obserwacji. Nie widać tu wyraźnego trendu. Z prawej: wykresy zależności reszt od zmiennych objaśniających. Również nie widać tu zależności. Na każdym z wykresów residua są równomiernie rozproszone wokół zera. Wszystko to sugeruje spełnianie przez model założenia o egzogeniczności składnika losowego.

więc większa korelacja nie martwi. Wartości korelacji nie są bardzo małe, ale nie są też duże i nie powinny sprawiać problemów. Takie skorelowanie zmiennych wynika z ich natury: wiele czynników socjodemograficznych zależy od siebie w pewien sposób, ponieważ np. duża populacja oznacza najczęściej duże miasto, gdzie jest więcej imigrantów niż w małych miejscowościach, ze względu na rozpoznawalność miasta. Odpowiada temu dodatnia korelacja między zmiennymi POPULATION medium i SH BORN OUTSIDE EU oraz ujemny współczynnik, gdzie pierwszą zmienną zastąpimy poziomem small. Natomiast z małych miast wiele osób wyjeżdża (na studia, szukając pracy itp.), więc korelacja między POPULATION small i SH RES 1Y SAME jest ujemna.
Po policzeniu wartości własnych macierzy danych oraz zbadaniu wartości pierwiastka stosunku największej wartości do najmniejszej, okazało się, że wartość ta jest bardzo duża. Zbadałam dlatego także wartości VIF. Dla każdej zmiennej wartość znajdowała się poniżej 1.5, nie wzbudzało to podejrzeń. Dodatkowo, rząd macierzy danych wyniósł 5, był zatem pełny. Założenie uznałam za spełnione. Wartości własne macierzy danych w prawdzie wydawały się niepokojące, ale wszystkie inne statystyki wyglądały dobrze.

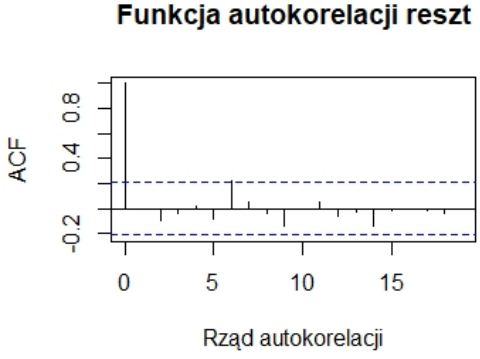
5.2.3 Egzogeniczność czynnika losowego

Założenie to sprawdzałam głównie za pomocą metod graficznych. Na początku narysowałam wykres reszt w zależności od numeru obserwacji (Rysunek 10a.), następnie wykonałam rysunki residuów w zależności od zmiennych objaśniających SH RES 1Y SAME i SH BORN OUTSIDE EU (Rysunek 10b.).

Na żadnym z wykresów nie widać zależności, reszty są rozproszone wokół prostej w zerze, co świadczy o tym, że założenie o niezależności czynnika losowego jest spełnione.

5.2.4 Homoskedastyczność i brak autokorelacji błędów

Zaczełam od analizy ewentualnej autokorelacji składnika losowego. Ze względu na to, iż obserwacje są prowincjami Włoch, z jednej strony autokorelacja reszt nie powinna mieć miejsca,



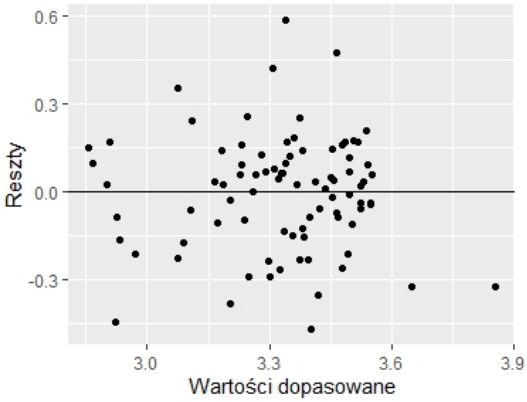
Rysunek 10: Wykres funkcji autokorelacji składnika losowego. Wartości są nieduże, jedynie przy szóstym rzędzie widać wyższy słupek, choć nadal nie jest on bardzo duży.

ponieważ np. wielkość jednej prowincji nie ma wpływu na sąsiednie regiony. Z drugiej jednak strony, we Włoszech widoczny jest przekrój kulturowy oraz różna mentalność mieszkańców, jeśli spojrzymy na północną i południową część kraju. Może to mieć wpływ na zachowanie ludzi również w kwestii głosowania w wyborach.
Sprawdzam to na wykresie funkcji autokorelacji reszt (Rysunek 11.). Słupek przy rzędzie szóstym jest większy niż pozostałe, chociaż nie jest duży. Warto to jednak sprawdzić za pomocą testu.

Wybrałam do tego test Boxa-Pierce'a, ponieważ bada on autokorelację wszystkich rzędów do ustalonego, za pomocą sumy kwadratów współczynników korelacji Pearsona pomiędzy odpowiednio oddalonymi resztami, przemnożonej następnie przez liczbę obserwacji. Naturalnie, chcemy, aby statystyka testowa była jak najmniejsza - będzie to oznaczać małe wartości korelacji. Test przeprowadziłam za pomocą funkcji `Box.test` dla korelacji do rzędu szóstego (ze względu na analizę wykresu funkcji autokorelacji), gdzie jako parametr typu ustawiłam `"Box-Pierce"`. Przy wartości statystyki testowej 5.8657, mającej asymtotycznie rozkład χ^2, p-value wyniosło 0.4384, nie dając podstaw do odrzucenia hipotezy o braku autokorelacji reszt.

Przechodząc do założenia o homoskedastyczności, znów najpierw analizowałam wykres, tym razem reszt w zależności od wartości dopasowanych (Rysunek 12.).

Nie widać tam trendów, ani grupowania się reszt, chociaż nie widać także idealnego losowego rozproszenia reszt. Warto to sprawdzić za pomocą testu Breuscha-Pagana. Wybrałam ten test, ponieważ nie widzę dwóch grup o różnej wariancji, a jednynie zastanawiam się nad zależnością reszt od zmiennych objaśniających. Test ten za hipotezę zerową przyjmuje homoskedastyczność błędów, następnie sprawdza to dopasowując nowy model liniowy do ich kwadratów, za zmienne objaśniające biorąc zmienne objaśniające modelu wyjściowego. Następnie liczymy R^2. Jeśli statystyka jest dostatecznie duża, hipoteza zerowa jest odrzucana. W R funkcja przeprowadzająca test Breuscha-Pagana to `bptest`. P-value zwrócone przez tę funkcję wynosiło ok. 0.82, zatem nie ma podstaw do odrzucenia hipotezy o homoskedastyczności.

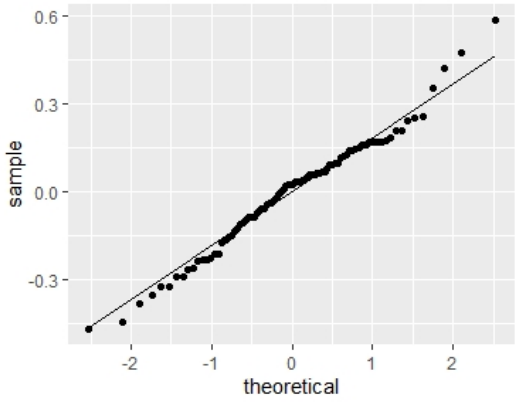


Rysunek 12: Wykres zależności reszt od wartości dopasowanych. Nie widać grupowania się punktów ani trendów, ale mimo wszystko warto przeprowadzić test statystyczny, w celu sprawdzenia homoskedastyczności.

5.2.5 Normalny rozkład składnika losowego

Ostatnim sprawdzanym założeniem jest normalność residuów. Z analizy wykresu kwantyl-kwantyl (Rysunek 13.) możemy podejrzewać grubszy prawy ogon niż w przypadku rozkładu normalnego. Oznacza to, że może być więcej dużych obserwacji, niż powinno mieć to miejsce, jeśli próba (tutaj reszty) pochodziłaby z rozkładu normalnego.
Sprawdziłam odstępstwa w ogodnach za pomocą testu Andersona-Darlinga. Jest to test oparty na odległości Cramera von Misesa między dystrybuantą empiryczną, a dystrybuantą teoretyczną z parametrami estymowanymi z próby. Autorzy testu dokonali poprawki tej odległości tak, aby test o nią oparty był bardziej czuły na odstępstwa w ogonach rozkładu. Dlatego wybrałam właśnie ten test.

Jest on przeprowadzany przy pomocy funkcji `ad.test`, p-value zwrócone przez tę funkcję wyniosło ok. 0.18, nie dając w ten sposób powodów do odrzucenia hipotezy o normalności reszt w modelu.



Rysunek 13: Wykres kwantyl-kwantyl dla reszt modelu. Prawy ogon rozkładu wydaje się być grubszy niż powinien być w przypadku rozkładu normalnego.

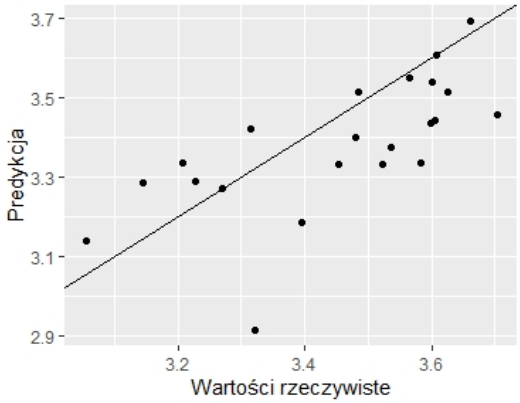
5.3 Badanie mocy prognostycznej modelu

Badanie mocy prognostycznej w przypadku mojego modelu ma bardziej na celu ogólną ocenę jego stabilności, ponieważ nie jest on przeznaczony do prognozowania, a jednynie ma na celu wyjaśnienie zjawiska.

Korzystałam tu z tego, iż na początku analizy zbiór danych został podzielony na zbiór treningowy i testowy. Za pomocą funkcji `predict` dokonuję predykcji wartości transformowanej zmiennej SH VALID na zbiorze testowym. Efekt dopasowania widoczy jest na rysunku nr 14. Wydaje się, że prosta mogłaby być bardziej nachylona, jednakże punkty rozkładają się dosyć dobrze wdłuż diagonali. Następnie liczyłam predictive R^2, wynosi ono ok. 0.42, co jest wartością o 4 punkty procentowe mniejszą niż dopasowane R^2, co wydaje się być dobrą wartością, gdyż strata nie jest duża. Błąd RMSE predykcji w stosunku do wartości rzeczywistych wynosi ok. 0.16, co po podzieleniu przez średnią wartość rzeczywistą na zbiorze testowym daje ok. 4.57. Oznacza to, iż błąd RMSE średnio wynosi ok. 4.57% rzeczywistej wartości zmiennej objaśnianej. Wydaje się to być niedużą wartością, szczególnie porównując do wartości dla poprzednich modeli, jest to poprawa.

5.4 Interpretacja modelu

Patrząc na wyestymowane współczynniki modelu możemy powiedzieć, że mniejsze prowincje (pod względem liczby ludności) sprzyjają obniżaniu się szans na to, iż losowo wybrany głosujący odda ważny głos. Wynika to z ujemnych znaków współczynników dla poziomów small i medium zmiennej POPULATION, przy czym dla poziomu small współczynnik jest około 5 razy mniejszy. Możemy także powiedzieć, iż większa część populacji nie pochodząca z Unii Europejskiej, powoduje zwiększenie procentu ważnych głosów. Zwiększenie zmiennej SH BORN OUTSIDE EU o 1 jednostkę, spowoduje wzrost szans na oddanie ważnego głosu $e^0.03$ razy. Choć wpływ tego czynnika na zwiększenie się szans na oddanie poprawnie wypełnionej karty do głosowania jest zdecydowanie mniejszy, niż wpływ odstetka populacji, który nie zmienił miejsca zamieszkania w ciągu roku przed przeprowadzeniem badań. Zwiększenie zmiennej SH RES 1Y SAME o 1 jednostkę, spowoduje wzrost szans na oddanie ważnego głosu aż $e^0.21$ razy. Może to oznaczać, iż stabilizacja życiowa (tutaj objawiająca się stałym miejscem zamieszkania) sprzyja podjęciu decyzji o zagłosowaniu na któregoś z kandydatów, a także może wpływać na zmniejszenie rozproszenia podczas głosowania, co skutkuje poprawnym oddaniem głosu.



Rysunek 14: Wartości przewidywane przez model vs wartości rzeczywiste. Prosta mogłaby być bardziej nachylona, ale punkty rozkładają się dosyć dobrze wdłuż diagonali.

6 Podsumowanie

Z całą pewnością model ten nie jest idealny. Z wykresu na Rysunku 9. widać, że może istnieć inna zależność, niż ta, którą przedstawia model, jednakże może nie być ona oczywista.
Finalny model spełnia wszytskie założenia klasycznej regresji liniowej. Wszystkie zmienne w tymże modelu są istotne statystycznie, błędy standardowe estymatowanych parametrów nie wydają się być duże.
Udało się wyjaśnić około 46% zmienności zjawiska, nie jest to bardzo duża wartość, jednakże biorąc pod uwagę naturę zjawiska występowania nieważnych głosów, osobiście uznaję to za zadowalający wynik, ponieważ nie spodziewałam się, iż w takiej części nie jest to zjawisko dziełem przypadku.
Mając na uwadze to wszystko, przyjmuję ten model, jako najlepszy spośród zaprezentowanych.