Zaawansowana Ekonometria II - model II

Autorzy	Bartłomiej Kuźma, Maciej Odziemczyk
Który temat	Regresja kwantylowa

Informacje o artykule będącym inspiracją dla minimodelu

Tytuł	Determinants of house prices in Istanbul: a quantile regression approach
Autorzy	Çaglayan Ebru, Arikan Eban
Miejsce publikacji	Quality & Quantity, 45 (2011)
Rok	2009
Zakres stron	305-317
Tematyka, problemy i cele badawcze	Tematyka analizowanego artykułu obejmuje problem modelowania wpływu różnych czynników, opisujących nieruchomość, na jej cenę. Głównym celem badania jest wyestymowanie hedonicznego modelu cen nieruchomości i sprawdzenie, które czynniki mają na nią wpływ. Autorzy sprawdzają też jaki jest wpływ poszczególnych czynników w różnych kwantylach rozkładu zmiennej zależnej.
Główne wnioski	Autorom udało się ustalić, że wiek nieruchomości ma pozytywby wpływ na jej cenę, co nie jest zgodne z intuicją i wcześniejszymi badaniami. Tłumaczą oni to faktem, iż w nowe budynki są budowane w azjatyckiej części Stambułu, gdzie przenoszą się ludzie chcący uciec z zatłoczonej europejskiej części (gdzie znajdują się między innymi centra biznesowe). Wśród czynników mających dodatni wpływ na cenę są też m.in. to czy w nieruchomości zamontowana jest telewizja kablowa, czy jest garaż oraz rodzaj systemu grzewczego. Posiadanie ochrony również ma dodatni wpływ na cenę, jak i kolejne pokoje i większa powierzchnia kuchni. Negatywny wpływ ma za to znajdowanie się nieruchomości przy głównej ulicy, co jest tłumaczone chęcią ucieknięcia od miejskiego zgiełku.
Metodyka badawcza	Wyestymowano model MNK, następnie przeprowadzono diagnostykę (test White'a i test Jarque'a Bera), a także przeprowadzono analizę outlierów (studentyzowane i standaryzowane reszty, odległość Cooka i dźwignia). W kolejnym kroku wyestymowano regresję kwantylową dla decyli od 1 do 9, przy użyciu bootstrapowanej macierzy warinacji-kowariancji.
Dane	Dane pochodzą z ankiety przeporwadzonej w okresie od października do grudnia 2007 na grupie agentów nieruchomości w Stambule. Spośród 1013 zebranych ankiet, wyelimoniowano te obserwacje, które zawierały braki danych, na skutek czego uzyskano zbiór liczący 992 obserwacje. Zmienną zależną był logarytm cen nieruchomości, a wśród regresorów były takie zmienne jak orientacja okien, wiek nieruchomości, liczba pokoi, liczba łazienek, system ogrzewania, itd. Uwzględniono również zmienne określające czy nieruchomość znajduje się po azjatyckiej czy europejskiej stronie Stambułu, czy znajduje się przy głównej ulicy oraz określające posiadanie lub nie takich "udogodnień" jak np. garaż, ochrona, telewizja kablowa itd.
Dlaczego wybrano ten artykuł	Kondycja rynku nieruchomości jest ważnym wskaźnikiem stanu gospodarki danego kraju. Dlatego analiza czynników wpływających na ceny mieszkań, może powiedzieć trochę o stanie w jakim znajduje się społeczeństwo, co jest dla niego ważne czy za co mogą decydować się zapłacić więcej (jak na przykład szczęśliwe numery w Chinach). Zastosowanie regresji kwantylowej, pozwala zauważyć różnice w różnych grupach społecznych (przyjmując niekontrowersyjne założenie, że biedniejsi ludzie kupują tańsze mieszkania).

Podstawowe informacje o minimodelu

Tytuł minimodelu	Determinanty ceny nieruchomości w Niemczech - estymacja regresji kwantylowej
Tematyka, problemy i cele badawcze	Modelowanie ceny nieruchomości jest często tematem prac naukowych. Odkrycie, które czynniki w jakim stopniu wpływają na nią, jest niewątpliwie wartościowe. Regresja kwantylowa pozwala rzucić na ten temat dodatkowe światło, dzięki któremu można dowiedzieć się jak poszczególne czynniki wpływają na cenę nieruchomości w poszczególnych kwantylach jej rozkładu. Praca ma na celu estymację modelu MNK jak i regresji kwantylowej, by sprawdzić jakie czynniki wpływają na cenę nieruchomości i jak ten wpływ zmienia się dla poszczególnych części rozkładu zmiennej objaśnianej.
Metodyka badawcza	Oszacowanie hedonicznego modelu MNK i przeprowadzenie jego diagnostyki. Następnie przeprowadzenie regresji kwantylowej, dla decyli 1-9 (bootstrapowana macierz wariancji-kowariancji, 999 replikacji). Graficzna ocena różnicy dla poszczególnych kwantyli, a następnie weryfikacja ich rzeczywistego istnienia przy użyciu odpowiedniego testu statystycznego.
Dane	Baza danych pochodzi ze serwisu Kaggle, a same dane zostały zescrappowane z niemieckiego rynku nieruchomości. Modele estymowane są dla trzech wybranych, sąsiadujących ze sobą landów (Północna Westfalia, Hesja, Dolna Saksonia), by różnice w położeniu geograficznym nie odgrywały znaczącej roli, równocześnie zapewniając wystarczającą liczbę obserwacji. Dane pochodzą z czerwca 2020 roku. W badaniu użyto zlogarytmowanych zmiennych takich jak, cena mieszkania, powierzchnia użytkowa, wielkość działki czy wiek nieruchomości. Użyto także binarnych zmiennych, jak posiadanie lub nie garażu, przynależność do danego Landu, a także, po zdekodowaniu, liczbę pokoi, łazienek, typ ogrzewania czy kondycja mieszkania.
Główne wnioski	Uzasadnione jest korzystanie z regresji kwantylowej w modelowaniu rynku nieruchomości, przeprowadzona analiza graficzna i testy wykazały niejednokrotnie zmienność oszacowań parametrów w zależności od rozkładu zmiennej zależnej. Często estymacje okazywły się zgodne z intuicją, np. czynnikami podwyższającymi cenę mieszkania okazały się powierzchnia użytkowa, powierzchnia działki czy liczba pokoi i łazienek. Starsze mieszkania są również tańsze. Okazało się również, że Hesja jest najdroższym z badanych landów, a Dolna Saksonia najtańszym. Zmienne mówiące o posiadaniu garażu, konkretnego systemu grzewczego czy stanu utrzymania nieruchomości okazywały się albo nieistotne statystycznie albo sprzeczne z oczekiwaniami.

1 Wstęp

Inspiracją do powstania prezentowanego badania był artykuł Ç.Ebru oraz A.Eban pt. "*Determinants of house prices in Istanbul: a quantile regression approach*" [2]. Badanie referencyjne polegało na estymacji hedonicznego równania cen mieszkań w Stalmbule za pomocą Modelu Regresji Liniowej MNK oraz Regresji kwantylowych dla decyli 1-9. Zaproponowane podejście wydaje się być niezwykle trafne, bowiem wpływ analizowanych czynników na cenę nieruchomości zmienia się w zależności od pozycji w rozkładzie zmiennej objaśnianej. W niniejszym projekcie wzięto pod lupę niemiecki rynek nieruchomości, a konkretnie trzy sąsiadujące ze sobą landy: Nadrenię Północną-Westfalię, Hesję oraz Dolną Saksonię. W toku analiz okazało się, że zaproponowane przez Ebru i Eban rozwiązanie modelowania zlogarytmowanej wersji równania jest znacznie wygodniejsze niż jego wersji standardowej. Model ten ma jednak swoje bolączki, kroki jakie podjęto w celu ich zniwelowania zaprezentowano w części poświęconej Metodologii. Wnioski nie zawsze okazywały się tożsame z badaniem referencyjnym, co jednak nie dziwi, a wręcz potwierdza różnice występujące między Turcją a Niemcami na analizowanej płaszczyźnie. Wynikami tego rodzaju badań mogą być zainteresowani zarówno deweloperzy, decydenci państwowi jak również konsumenci, właściwie wszyscy uczestnicy rynku nieruchomości, dlatego tak ważne jest modelowanie nie tyle wartości oczekiwanej, co całego rozkładu, na co pozwala regresja kwantylowa. W świetle uzyskanych wyników można podtrzymać postawioną hipotezę o występowaniu różnic międzykwantylowych w składowych cen nieruchomości.

2 Dane

Dane, którymi posłużono się w niniejszym badaniu, jak donosi ich autor, zostały *zescrappowane* ze strony: immobilienscout24 - link tj. niemieckiego rynku nieruchomości. Kompletny zbiór danych można znaleźć na kaggle.com - link (stan na 13.01.2021). W celu uzyskania możliwie wiarygodnych wyników przy jednoczesnym zachowaniu sensownej liczby obserwacji zdecydowano się na modelowanie cen mieszkań dla trzech sąsiadujących landów - Nadrenii Północnej-Westfalii, Hesji oraz Dolnej Saksonii. Bardzo ważnym elementem obróbki danych było ich oczyszczenie, m. in. usunięcie/uzupełnienie braków. W większości przypadków usuwano rekordy, w których występowały braki dla pożądanych zmiennych. Specjalna procedura została natomiast wdrożona w przypadku obróbki zmiennej mówiącej o tym czy posesja posiada garaż (Garage). W pełnym zbiorze danych znajdują się zmienne mówiące o typie garażu i ich liczbie. Zauważono, że indeksy braków danych w obydwu seriach są tożsame, ponadto dla zmiennej mówiącej o liczbie garaży nie występowały obserwacje zerowe, na tej podstawie wyciągnięto wniosek o braku garażu w przypadku braku tejże informacji w danych. Finalnie stworzono zmienną binarną, mówiącą o tym czy garaż jest czy go nie ma. Zdecydowano się również na usunięcie obserwacji, dla których ceny przyjmowały kontrowersyjne wartości (2 obserwacje o cenach 0 i 1 - nawet jeżeli jest to prawda to obserwacje te nie pasują do generalnego modelu bowiem ceny nie wydają się wówczas zależeć od charakterystyk nieruchomości), dodatkowo usunięto również obserwację, która charakteryzowała się powierzchnią użytkową równą 0 (sytuacja niemożliwa do realizacji, zwłaszcza jeżeli takowa nieruchomość posiada informacje o np. liczbie pokoi, czy łazienek). Z uwagi na znaczącą różnicę skali ciągłych zmiennych objaśniających zarówno między sobą, jak i ze zmienną objaśnianą, fakt występowania wielu zmiennych binarnych (dekodowanie poziomów) oraz niepożądane własności rozkładu zmiennej zależnej (kurtoza 25.97, skośność 4.048), zdecydowano się na zlogarytmowanie cen (zmiennej zależnej) oraz ciągłych zmniennych objaśniających: powierzchni użytkowej, wielkości działki, i wieku nieruchomości (ta ostatnia zmienna powstała poprzez odjęcie występującej w oryginalnej bazie danych zmiennej Year_built od liczby 2021 tj. roku bieżącego). Zmienne zlogarytmowane opatrzone są przedrostkiem "log_".
Jak donosi autor, dane zostały one pozyskane w czerwcu 2020 roku, co oznacza, że fluktuacje rynku nie są istotnym czynnikiem w estymowanym modelu. Warto w tym miescu również wyjaśnić pewną nieścisłość - wiek nieruchomości został obliczony w odniesieniu do 2021 roku, podczas gdy dane pochodzą z 2020, powodem takiego postępowania jest fakt występowania w bazie budynków oddanych do użytku w 2020 roku przez co wiek takiej nieruchomości, zakroąglony do lat, wynosiłby zero, a jak wiemy zero nie należy do dziedziny funkcji logarytm, jednocześnie przesunięcie zmiennej o stałą nie wpływa na oszacowania modelu. Tabela 1 zawiera szczegółowy opis wykorzystanych zmiennych, wiele z nich pokrywa się z badaniem referencyjnym, pojawiły się jednak też nowe regresory, czego uzasadnieniem jest ograniczenie wynikające z dostępności danych oraz specyfika badanej próby.

Tablica 1: Zmienne uwzględnione w modelach

log_Price	logarytm cen nieruchomości - zmienna zależna
garage	zmienna binarna 1 jeżeli jest garaż, 0 w p.p.
log_usable_area	logarytm powierzchni użytkowej (jednostka nie została podana w bazie danych)
log_lot	logarytm powierzchni działki (jednostka nie została podana w bazie danych)
bathrooms	zmienna poziomowa: _1 (poziom bazowy), _2, _3, _more_than_3 - liczbia łazienek
heating_central_heating	ogrzewanie centralne, poziom bazowy: ogrzewanie olejne
heating_heat_pump	pompa grzewcza, poziom bazowy: ogrzewanie olejne
heating_stove_heating	ogrzewanie za pomocą pieca, poziom bazowy: ogrzewanie olejne
heating_other	inne, niewymienione wyżej systemy ogrzewania, poziom bazowy: ogrzewanie olejne
bedrooms	zmienna poziomowa _1 (poziom bazowy), _2, _3, _4, _more_than_4 - liczba sypialni
condition_dilapidated	stan zniszczony, poziom bazowy: do lekkich poprawek
condition_maintained	stan utrzymany (niewyróżniający się), poziom bazowy: do lekkich poprawek
condition_modernized	stan lekko unowocześniony, poziom bazowy: do lekkich poprawek
condition_refurbished	stan po drobnych naprawach, poziom bazowy: do lekkich poprawek
condition_renovated	stan odświeżony, poziom bazowy: do lekkich poprawek
condition_other	stan inny niż wymienione, poziom bazowy: do lekkich poprawek
state_hessen	land położenia nieruchomości - Hesja, poziom bazowy Północna Westfalia
state_niedersachsen	land położenia nieruchomości - Dolna Saksonia, poziom bazowy Północna Westfalia
log_age	logarytm wieku (liczony od 2021 roku)

Źródło: Opracowanie własne.

3 Metodologia

Zadaniem niniejszego projektu jest oszacowanie hedonicznego równania cen mieszkań. Model hedoniczny w sensie ekonomicznym wynika z teorii ujawnionych preferencji i opiera się na założeniu możliwości dekompozycji danego dobra na jego składowe (charakterystyki), które wraz ze swoimi konsekwencjami znane są konsumentom *a priori* [1]. Jest to sposób na modelowanie popytu na dobra heterogeniczne, dlatego też znaduje szerokie zastosowanie w badaniu rynku mieszkań. W skrócie można powiedzieć, że teoria ta pozwala na estymację wartości charakterystyk w oczach konsumentów. W rzeczywistości prawdziwość wspomnianych założeń jest kontrowersyjna, bowiem na rynku mieszkań można doszukiwać się endogeniczności pewnych cech, np. występowanie garażu czy ogrodu wydaje się być zależne od położenia nieruchomości - w centrum miasta szansa na ich występowanie wydaje się niższa niż na przedmieściach; metraż dostępnych mieszkań może zależeć od gęstości zaludnienia - w Japonii, Chinach czy Indiach dostępna jest zapewne dużo niższa powierzchnia deweloperska niż w krajach Skandynawskich. Endogeniczność stoi w sprzeczności z możliwości pełnej wyceny konkretnych charakterystyk, podobnie założenie o znajomości charekterystyk *a priori* wydaje się być kontrowersyjne, bowiem jeżeli model uwzględnia czynniki nieobserwowalne, takie jak np. zanieczyszczenie powietrza wówczas założenie to jest również wątpliwe. Pomimo swoich wad model ten jest szeroko stosowany ze względu na swoją prostotę, elastyczność, dostęp do danych i możliwość skwantyfikowania pewnych cech. W kontekście ekonometrycznym jest to model regresji liniowej w postaci:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

lub log-liniowej:

$$\ln \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

gdzie:
$\boldsymbol{Y}$ jest wektorem objaśnianym np. ceny mieszkań,
$\boldsymbol{X}$ jest macierzą obserwacji, której kolumny stanowią kolejne charakterystyki danego dobra,
$\boldsymbol{\beta}$ jest wektorem "wartości" charakterystyk - parametrami strukturalnymi modelu,
$\boldsymbol{\varepsilon}$ jest wektorem błędów losowych.

W celu ominięcia pewnych niedognodości związanych z modelem hedonicznym zdecydowano się na:

- wyestymowanie modelu dla trzech sąsiadujących ze soba landów - aby wyeliminować efekt położenia geograficznego (landy stanowią zmienne, zatem pewne różnice cenowe między nimi zostaną wyjaśnione),
- uwzględnienie jedynie zmiennych obserwowalnych - w celu ograniczenia naruszenia założenia o znajmości charakterystyk *a priori* (stan mieszkania, mimo że subiektywny to jest obserwowalny przed dokonaniem transakcji, wyjątek stanowią wady ukryte),
- wykorzystanie danych pochodzących z jednego okresu - w celu wykluczenia efektu fluktuacji rynku nieruchomości.

Standardową analizę regresji zdecydowano się rozszerzyć o modele kwantylowe, co wydaje się być uzasadnione, bowiem niekontrowersyjne założenie o różnicowaniu dochodu konsumenta za pomocą jego preferencji i zachowań na rynku nieruchomości popycha do postawienia hipotezy o warunkowej względem rozkładu cen nieruchomości ważności ich charakterystyk.

Ponadto model MNK przetestowano pod kątem występowania heteroskedastyczności (test White'a), normalności reszt (test Jarque-Bera) oraz zbadano wartości odstające (odległość Cooka $> \frac{4}{N}$, reszty standaryzwowane i studentyzowane $|\hat{e}_i| > 2$ to obserwacje nietypowe oraz dźwignia).

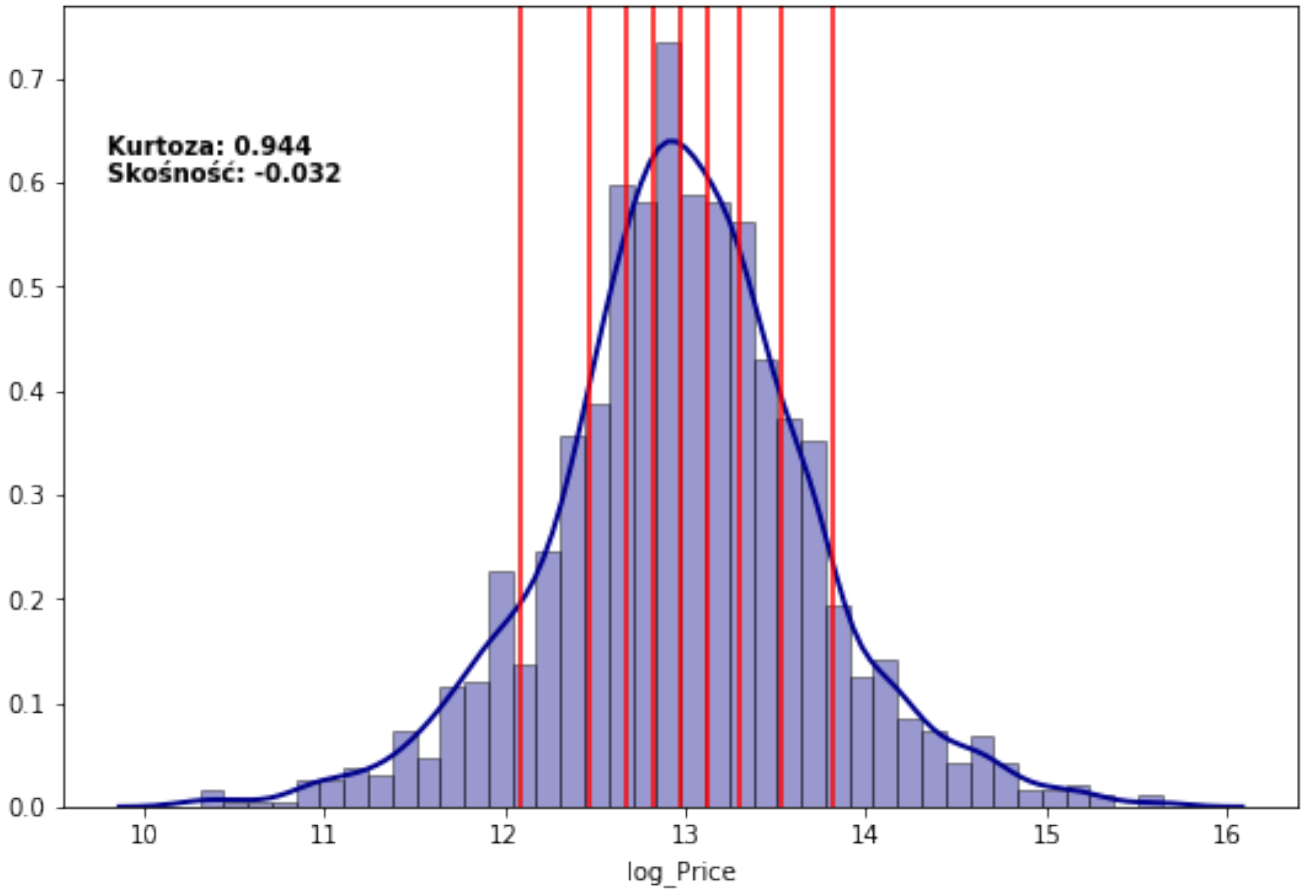
W ramach regresji kwantylowej oszacowano modele dla decyli $1-9$ z bootstrapową macierzą wariancji-kowariancji (999 replikacji). W celu uchwycenia trendów zmian przeprowadzono analizę graficzną oszacowań parametrów w zależności od kwantyla, a widoczne różnice przetestowano statystycznie - seria niezależnych hipotez prostych.

4 Wyniki

W celu zapoznania się z analizowanym problemem zdecydowano się na Rysunku 1 przedstawić histogram zmiennej zależnej - logarytm cen mieszkań wraz z jądrową estymacją funkcji gęstości, ponadto czerwone, wertykalne linie

oznaczają analizowane kwantyle rozkładu (decyle $1-9$). Na wykresie zamieszczono również informacje o kurtozie w sensie Fishera ($K-3$) oraz skośności.

Rysunek 1: Histogram, estymacja jądrowa gęstości i decyle logarytmu cen.



Źródło: Opracowanie własne.

Wykres pozwala na zauważenie leptokurtyczności rozkładu analizowanej zmiennej (kurtoza większa od rozkładu Normalnego - większa koncentracja wokół średniej, wysmukłość - warto podkreślić, że kurtoza obliczona została na podstawie obserwacji, różnica w wysokości wykresu z wynikiem jest efektem liczby słupków histogramu) oraz lekkiej asymetrii lewostronnej. W Tabeli 2 przedstawiono wyniki oszacowań dla modelu MNK i regresji kwantylowych. Test White'a dla MNK pozwolił na odrzucenie hipotezy zerowej o homoskedastyczności reszt (statystyka testowa: 123, p-value: 0.000), podobnie jak test Jarque-Bera pozwolił na odrzucenie hipotezy zerowej o normalności reszt (statystyka testowa: 38.962, p-value: 0.000), W załączniku znajduje się tabela wartości odstających - 90 obserwacji przekraczających statystykę odległości Cooka na poziomie $\frac{4}{N}$, można tam zaobserwować również dużą liczbę obserwacji ze standaryzowanymi i studentyzowanymi resztami przekraczającymi 2 co do wartości bezwzględnej. Otrzymane wyniki testów diagnostycznych MNK jakościowo pokrywają się z wynikami autorów badania referencyjnego.

Warto zauważyć występujące różnice oszacowań między regresją MNK (OLS), która modeluje wartość oczekiwaną, a modelem dla mediany (Q50), nie są one duże, co jest spójne z symetrią rozkładu logarytmu cen (skośność: -0.032). Interesujące są natomiast różnice w istotności liczby sypialni (Bedrooms) - dla decyli 5-10 są to zmienne nieistotne, w przeciwieństwie do modelu MNK, co może wynikać technicznie z występowania wartości odstających, a jakościowo z mniejszej wartości liczby sypialni w droższych mieszkaniach, być może bardziej zaczyna liczyć się ich jakość lub większa liczba sypialni jest swojego rodzaju standardem (poziomem bazowym jest 1 sypialnia) lub też mniejsza liczba członków rodziny objawia się średnio wyższym dochodem w przeliczeniu na jednostkę - dlatego potrzebnych jest mniej pokoi. Zauważyć można również nieistotność występowania zmiennej Garage we wszystkich modelach, ogrzewania pompą ciepła (Heating_heat_pump) (poziom bazowy to orzewanie olejne), czy zmiennych mówiących o stanie (zniszczony, zmodernizowany, renowacja i inne, gdzie poziom bazowy to do lekkich

poprawek). Jedynymi istotnymi w niektórych przypadkach stanami był utrzymany oraz po drobnych naprawach (condition_maintained, condition_refurbished). Warto zauważyć również wycenę Hesji jako miejsca zamieszkania - istotny, dodatni współczynnik powyżej mediany oraz zawsze istotny, ujemny współczynnik dla Dolnej Saksonii w stosunku do poziomu bazowego (Nadrenii Północnej-Westfalii). Na tym etapie należy mieć również na uwadze, że zmienna zależna jest zlogarytmowana co oznacza konieczność interpretacji parametrów jako semi-elastyczności w przypadku zmiennych binarnych (ciągłe, nominalne tutaj nie występują) tj. przemnożyć współczynnik przez 100 i interpretować jako zmiany procentowe, lub jako elastyczność w przypadku ciągłych zmiennych zlogarytmowanych (opatrzone przedostkiem log_, interpretacja w procentach). Generalnie można zauważyć pewne trendy rosnące w wadze powierzchni działki (log_lot) i ujemnej zależności wieku (log_age), to drugie jest wnioskiem przeciwnym do wyciągniętego w badaniu referencyjnym. Analiza trendów jest jednak znacznie łatwiejsza przy pomocy wykresów, dlatego też na rysunkach 2-7 zdecydowano się przedstawić oszacowania parametrów w zależności od kwantyla na tle oszacowań MNK, zobrazowanie przedziałów ufności (szare tło) pozwoliło na wytypowanie różnic do formalnego przetestowania.

4.1 Analiza różnic międzykwantylowych

Hipoteza zerowa przeprowadzanego testu na różnice międzykwantylowe mówi o braku różnicy oszacowań, hipoteza alternatywna jest oczywiście jej przeciwieństwem. Poniżej przedstawione zostały analizy dla poszczególnych zmiennych:

- Garage - można zauważyć trend spadkowy, jednakże przedziały ufności są na tyle szerokie, że różnica międzykwantylowa nie wydaje się istotna w żadnym wpadku, ponadto zmienna ta nie jest istotna w żadnym modelu, nie testowano istotności różnic,

- log_Usable_Area - obserwowalny trend wzrostowy od 1 do 3 decyla, później stabilizacja, szerokie przedziały ufności pozwalają jednak wysnuć hipotezę o niezmienności statystycznej oszacowania. Przetestowano decyl 1 z 3 - dwa ekstrema (stat. 3.23, p-value 0.07) oraz decyl 1 z 9 (stat. 0.99, p-value 0.32). Wnioskuje się zatem o stałej na przestrzeni kwantyli elastyczności cen względem powierzchni użytkowej, która wynosi między 0.05%, a 0.11% na 1% zmiany powierzchni,

- log_Age - silny trend rosnący, oszacowania dla decyli 1-4 i 6-9 wykraczają poza przedziały ufności MNK, przetestowano decyle 1 z 9 (stat 4.14, p-value 0.0422) oraz 4 z 7 (stat 8.96, p-value 0.0028) - silne różnice między ogonami rozkładu jak również w okolicy środka. Z uwagi na ujemne wartości oszacowania można wnioskować o malejącym, negatywnym wpływie wieku nieruchomości na jej cenę, co wydaje się zgodne z oczekiwaniami, bowiem centra miast zazwyczaj są zabudowywane w pierwszej kolejności, ponadto zamożni klienci mogą wyżej cenić sobie walory historyczne mieszkania,

- log_Lot - oszacowania powierzchni działki wyglądają nad wyraz ciekawie, widoczny trend spadkowy od 1 do 3 decyla, po czym widoczny wzrost. Statystycznie przetestowano różnice między kwantylem 30 i 60 oraz 30 i 90 - odrzucenie hipotez zerowych (stat. odpowiednio 4.94 i 6.05, p-value 0.026, 0.014). Hipoteza zerowa nie została odrzucona natomiast w przypadku decyla 2 i 7, 2 i 9 oraz 3 i 4 (stat. 1.17, 2.25, 1.03, p-value 0.28, 0.134, 0.311). Ciekawa zależność, na podstawie, której można domniemywać, że obszary wiejskie lub przedmieścia, cechujące się niższymi cenami mogą przyciągać ludność, dla której działka ma istotne znaczenie, potem obserwujemy klasę średnią, dla której wygoda i dostęp do pracy mogą być ważniejsze, a na koniec najzamożniejsi mieszkańcy, którzy również mogą być zainteresowani większym ogrodem czy działką użytkową.

- Bathrooms_2, _3, more_than_3, w tych wypadkach parametry wydają się oscylować wokół oszacowań MNK i nie różnić się między sobą statystycznie, przetestowano zatem tylko ekstrema nie uzyskując podstaw do odrzucenia hipotezy zerowej o braku różnic, odpowiednio stat. 2.97, 2.02 i 1.1, p-value 0.085, 0.155, 0.295. Można zaobserwować natomiast efekt "Momentum" ceny względem liczby łazienek - każda następna ma co raz wyższe oszacowanie, co pozwala wnioskować o rosnącej kontrybucji w cenie względem poziomu bazowego tj. jednej łazienki, co jest zgodne z intuicją, bowiem dodatkowe łazienki to dodatkowa powierzchnia, ponadto w bardziej luksusowych apartamentach często pokoje wyposażane są w osobne łazienki. Obserwujemy zatem, że dodatkowa łazienka podwyższa cenę mieszkania o około 15%, dwie dodatkowe o około 40%, a trzy i więcej to około 50%, podwyżka oczywiście *ceteris paribus*.

- Heating_central_heating, _heat_pump, _stove_heating niemalże pokrywają się z oszacowaniami MNK, nie testowano ich. Intrygujące są natomiast ujemne oszacowania parametrów tychże zmiennych, co z ogrzewaniem

olejnym jako poziomem bazowym wydaje się być dość nieintuicyjnym wynikiem, bowiem wobec wiedzy autorów niniejszego projektu to właśnie ogrzewanie olejne jest jednym z gorszych systemów, a prezentowane wyniki są sprzeczne z tym założeniem. Przetestowano natomiast heating_other - inne systemy ogrzewania, w tej grupie znalazły się np. ogrzewanie panelami słonecznymi czy też podłogowe, ale również palety - w żadnym wypadku nie było podstaw do odrzucenia hipotezy zerowej, a testowano decyle 1 z 7, 1 z 9 i 2 z 7 (stat. 3.79, 2, 1.38, p-value 0.0519, 0.1577, 0.24). Można próbować wnioskować o obniżaniu cen nieruchomości jeżeli system ogrzewania jest inny niż olejny, są to też dość duże obniżki (semielastyczność - współczynnik należy przemnożyć przez 100), wnioski te są więc wysoce kontrowersyjne. Na uwadze należy mieć fakt nieistotności niektórych poziomów w niektórych kwantylach co sugeruje, że mają one wówczas taki sam udział w kształtowaniu ceny nieruchomości co poziom bazowy,

- Bedrooms_2, _3, _4 oraz more_than_4 - wykresy wyglądają identycznie, zarówno jeżeli chodzi o kształt jak i skalę. Można zauważyć, że dodatkowe sypialnie ponad jedną wiążą się ze wzrostem ceny nieruchomości, jednak to jak dużo będzie tych dodatkowych sypialni, nie ma większego znaczenia - istotne różnice między kwantylami 20 i 60 (stat. 7.12, p-value 0.0077), ze względu na kształ wykresów przetestowano tylko przypadek dwóch sypialni - dodatkowy pokój to wzrost wartości mieszkania, jest on jednak dużo większy w przypadku drugiego decyla niż 6, od 50% do 20%, jednak powyżej mediany, oszacowania te są nieistotne statystycznie,

- condition_dilapidated - stan zniszczony, współczynniki regresji kwantylowych oscylują wokół MNK, nie testowano różnic, zmienna ta jest również nieistotna statystycznie w każdym z wyestymowanych modeli,

- condition_maintainted - stan utrzymany (niewyróżniający się) - trend potwierdzony testem dla 3 i 8 decyla (stat 6.04, p-value 0.014) - malejący negatywny wpływ na cenę, co w świetle poziomu bazowego tj. do lekkich poprawek jest wnioskiem sprzeczym z oczekiwaniami - zmienna istotna do mediany,

- condition_modernized - lekko unowocześniony, tutaj również kontrowersyjne oszacowanie, sugerujące spadek wartości. Decyle 3 i 8 różnią się statystycznie (stat 6.04, p-value 0.014), jednak zmienna ta nie jest istotna w tych modelach,

- condition_other - stan inny (wszystkie inne opisy nieuwzględnione w poziomach badanej zmiennej), zmienna nie jest istotna w żadnym modelu,

- condition_refurbished - stan po lekkich naprawach - również kontrowersyjne oszacowanie, poniżej zera. Występują istotne statystycznie różnice między 3 i 9 decylem (stat. 5.27, p-value 0.0219),

- condition_renovated - stan odświeżony, ponownie jak w przypadku pozostałych zmiennych mówiących o stanach oszacowanie jest sprzeczne z intuicją i w stosunku do stanu do lekkich naprawek, odświeżenie obniża cenę nawet do 20%. Brak różnic istotnych statystycznie między decylami, przetestowane ekstrema tj. D1 i D9 (stat. 2.81, p-value 0.094),

- bardzo ciekawe wnioski można natomiast wysnuć odnosząc się do oszacowań dla zmiennych mówiących o tym, w którym landzie znajduje się nieruchomość. W stosunku do poziomu bazowego - Północna Westfalia, Saksonia Dolna wydaje się być tańsza o ok 35% niezależnie od kwantyla (istotne w każdym modelu), natomiast Hesja nie dość, że jest droższa niż poziom bazowy niemal dla całego rozkładu, to jeszcze wzrost ten rośnie wraz ze wzrostem ceny, co zostało przetestowane statystycznie decyl 2 z 4 i 4 z 6 (stat 7.48 i 7.23, p-value 0.0063 i 0.0073), co pozwala wnioskować o efekcie "momentum" dla Hesji, obecość nieruchomości w tym landzie przyczynia się do wzrostu cen, który jest tym silniejszy im wyższa jest cena, nawet do 33% - warto tutaj też zauważyć istotność statystyczną dopiero od mediany w górę.

Tablica 2: Wyniki oszacowań

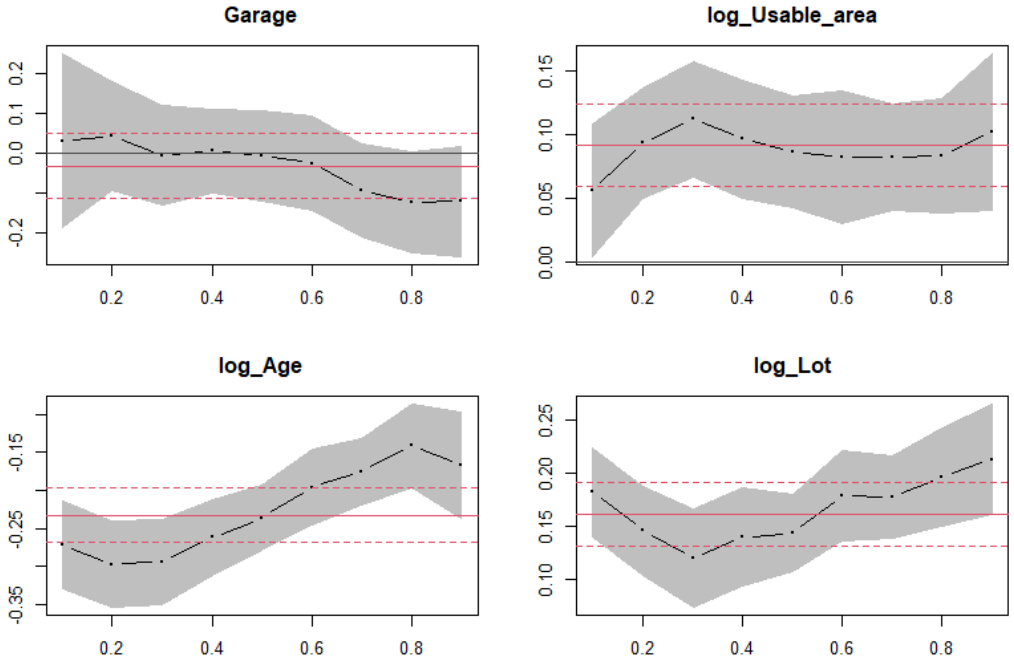
	OLS	Q10	Q20	Q30	Q40	Q50	Q60	Q70	Q80	Q90
garage	-0.0322	0.0312	0.0435	-0.00493	0.00550	-0.00636	-0.0248	-0.0919	-0.123	-0.120
	(0.049)	(0.126)	(0.0886)	(0.0775)	(0.0666)	(0.0667)	(0.0729)	(0.0704)	(0.0806)	(0.0838)
log_usable_area	0.0916***	0.0564	0.0933***	0.112***	0.0964***	0.0864***	0.0825**	0.0821**	0.0837**	0.102**
	(0.0196)	(0.0323)	(0.0277)	(0.0265)	(0.0278)	(0.0261)	(0.0308)	(0.0259)	(0.0284)	(0.0371)
log_lot	0.1616***	0.182***	0.146***	0.120***	0.140***	0.144***	0.179***	0.178***	0.196***	0.213***
	(0.0183)	(0.0276)	(0.0265)	(0.0274)	(0.0274)	(0.0223)	(0.0258)	(0.0228)	(0.0273)	(0.0316)
bathrooms_2	0.1931***	0.179*	0.210***	0.210***	0.165***	0.156***	0.142**	0.124**	0.129*	0.207***
	(0.0397)	(0.0724)	(0.0519)	(0.0471)	(0.0457)	(0.0417)	(0.0446)	(0.0449)	(0.0519)	(0.0587)
bathrooms_3	0.4028***	0.257**	0.384***	0.409***	0.390***	0.395***	0.378***	0.343***	0.390***	0.437***
	(0.0504)	(0.0901)	(0.0860)	(0.0683)	(0.0642)	(0.0548)	(0.0592)	(0.0603)	(0.0664)	(0.0961)
bathrooms_more_than_3	0.5708***	0.473***	0.496***	0.552***	0.497***	0.549***	0.581***	0.545***	0.576***	0.605***
	(0.0607)	(0.0831)	(0.101)	(0.0888)	(0.0858)	(0.0883)	(0.0811)	(0.0714)	(0.0873)	(0.103)
heating_central_heating	-0.236**	-0.215	-0.255*	-0.188	-0.304***	-0.292***	-0.268**	-0.194*	-0.318*	-0.221
	(0.0845)	(0.155)	(0.122)	(0.0987)	(0.0854)	(0.0809)	(0.0981)	(0.0907)	(0.145)	(0.152)
heating_heat_pump	-0.0871	-0.0583	-0.0553	-0.0710	-0.109	-0.0990	-0.117	-0.132	-0.151	-0.0439
	(0.0752)	(0.118)	(0.108)	(0.0981)	(0.0874)	(0.0724)	(0.0684)	(0.0708)	(0.116)	(0.155)
heating_other	-0.3224***	-0.466***	-0.339**	-0.378***	-0.309**	-0.258**	-0.215*	-0.191*	-0.207*	-0.227*
	(0.0756)	(0.132)	(0.121)	(0.109)	(0.102)	(0.0830)	(0.0908)	(0.0813)	(0.0961)	(0.115)
heating_stove_heating	-0.193**	-0.159	-0.153	-0.164*	-0.203**	-0.196**	0.207***	-0.170**	-0.216**	-0.193*
	(0.0644)	(0.0882)	(0.0948)	(0.0812)	(0.0759)	(0.0629)	(0.0614)	(0.0596)	(0.0832)	(0.0981)
bedrooms_2	0.4874***	0.709***	0.848***	0.807***	0.709*	0.198	0.249	0.230	0.259	0.257
	(0.117)	(0.197)	(0.184)	(0.230)	(0.331)	(0.311)	(0.212)	(0.171)	(0.182)	(0.182)
bedrooms_3	0.456***	0.655***	0.773***	0.709**	0.666*	0.183	0.233	0.182	0.247	0.169
	(0.111)	(0.176)	(0.172)	(0.227)	(0.328)	(0.308)	(0.207)	(0.162)	(0.173)	(0.160)
bedrooms_4	0.4639***	0.704***	0.761***	0.707**	0.688*	0.191	0.245	0.253	0.317	0.231
	(0.1124)	(0.176)	(0.168)	(0.227)	(0.327)	(0.308)	(0.206)	(0.165)	(0.178)	(0.158)
bedrooms_more_than_4	0.489***	0.652***	0.744***	0.720**	0.731*	0.229	0.256	0.252	0.293	0.212
	(0.114)	(0.176)	(0.177)	(0.231)	(0.330)	(0.309)	(0.207)	(0.164)	(0.178)	(0.169)

	OLS	Q10	Q20	Q30	Q40	Q50	Q60	Q70	Q80	Q90
condition_dilapidated	0.0805	0.0861	0.00630	-0.118	-0.106	-0.0142	0.0777	0.0873	0.205	0.109
	(0.086)	(0.186)	(0.150)	(0.131)	(0.0955)	(0.0754)	(0.102)	(0.111)	(0.120)	(0.129)
condition_maintained	-0.292***	-0.423*	-0.347*	-0.452***	-0.398***	-0.276***	-0.229*	-0.207	-0.0736	-0.140
	(0.082)	(0.205)	(0.147)	(0.128)	(0.0967)	(0.0711)	(0.0894)	(0.111)	(0.130)	(0.145)
condition_modernized	-0.156*	-0.0849	-0.123	-0.271*	-0.260***	-0.169***	-0.0900	-0.137	-0.126	-0.117
	(0.0705)	(0.169)	(0.128)	(0.109)	(0.0713)	(0.0500)	(0.0745)	(0.0887)	(0.103)	(0.103)
condition_other	-0.201	-0.353	-0.380	-0.394	-0.295	-0.238	-0.222	-0.286	0.0501	0.144
	(0.146)	(0.330)	(0.299)	(0.274)	(0.216)	(0.180)	(0.201)	(0.279)	(0.351)	(0.362)
condition_refurbished	-0.2932***	-0.423*	-0.389**	-0.489	-0.403***	-0.280***	-0.187*	-0.210*	-0.218	-0.125
	(0.0764)	(0.180)	(0.133)	(0.121)	(0.0863)	(0.0694)	(0.0876)	(0.0911)	(0.112)	(0.124)
condition_renovated	-0.0863	-0.0478	-0.0652	-0.244*	-0.188*	-0.112	-0.0556	-0.0463	-0.0244	0.0100
	(0.0798)	(0.179)	(0.132)	(0.118)	(0.0873)	(0.0660)	(0.0918)	(0.103)	(0.118)	(0.124)
state_hessen	0.1461***	0.0103	-0.0238	0.0374	0.115*	0.184***	0.228***	0.240***	0.274***	0.330***
	(0.0355)	(0.0688)	(0.0611)	(0.0573)	(0.0551)	(0.0458)	(0.0484)	(0.0411)	(0.0488)	(0.0621)
state_niedersachsen	-0.363***	-0.356***	-0.348***	-0.361***	-0.350***	-0.389***	-0.397***	-0.359***	-0.341***	-0.364***
	(0.0365)	(0.0638)	(0.0508)	(0.0468)	(0.0439)	(0.0377)	(0.0421)	(0.0388)	(0.0490)	(0.0603)
log_age	-0.233***	-0.271***	-0.297***	-0.294***	-0.262***	-0.236***	-0.196***	-0.176***	-0.142***	-0.167***
	(0.0217)	(0.0347)	(0.0354)	(0.0332)	(0.0303)	(0.0269)	(0.0314)	(0.0271)	(0.0340)	(0.0410)
_cons	12.14***	11.45***	11.72***	12.16***	12.15***	12.59***	12.24***	12.38***	12.20***	12.32***
	(0.195)	(0.324)	(0.289)	(0.319)	(0.386)	(0.355)	(0.307)	(0.257)	(0.301)	(0.307)
N	1431									
adj. R^2										

Standard errors in parentheses

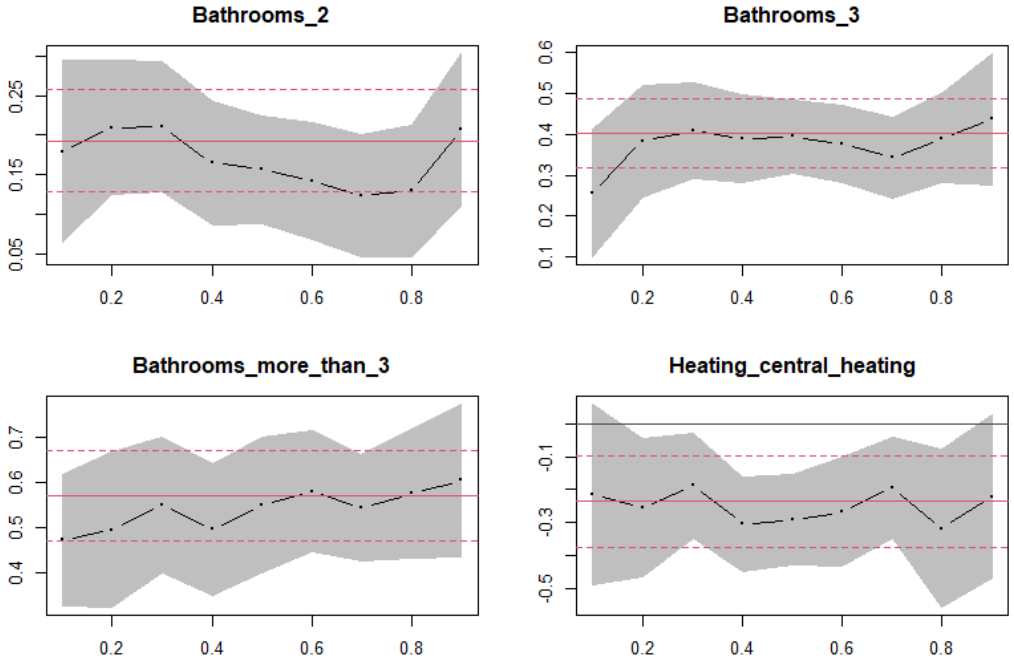
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Rysunek 2: Estymatory dla kwantyli



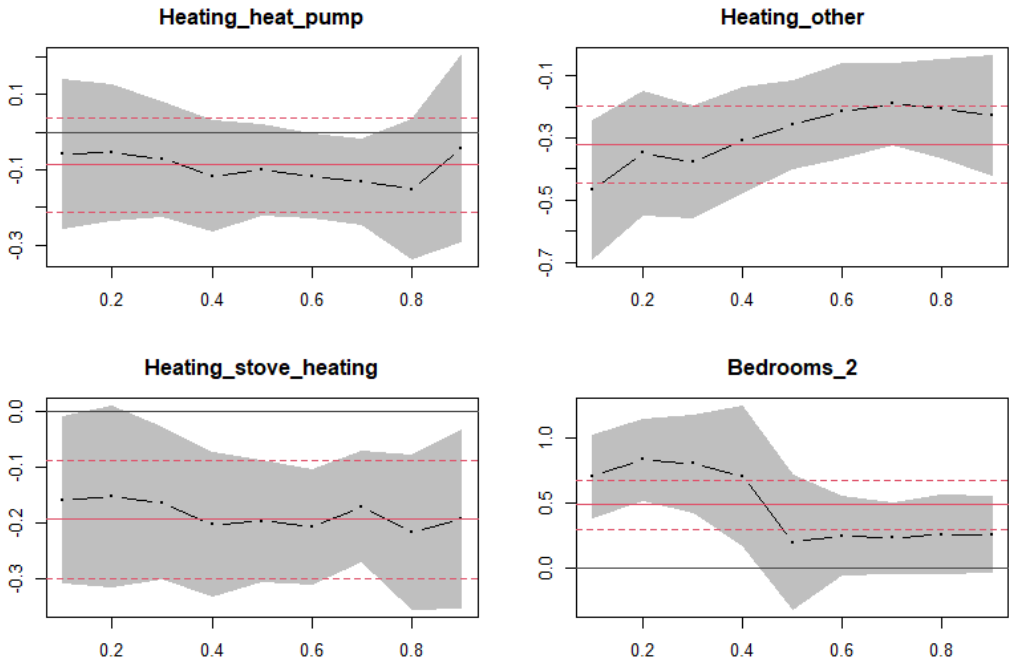
Źródło: Opracowanie własne.

Rysunek 3: Estymatory dla kwantyli



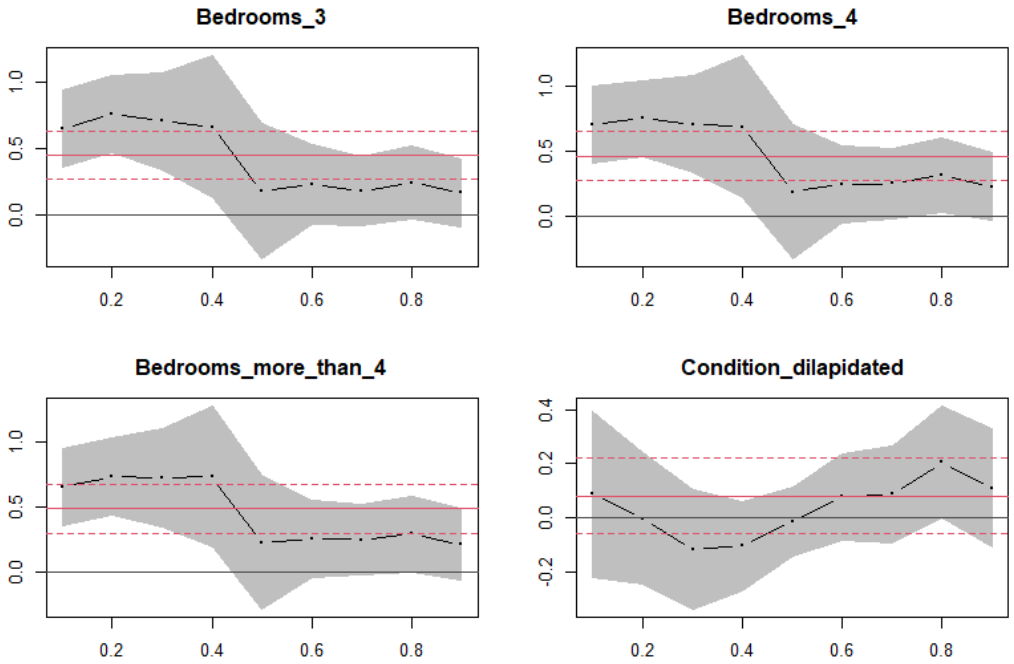
Źródło: Opracowanie własne.

Rysunek 4: Estymatory dla kwantyli



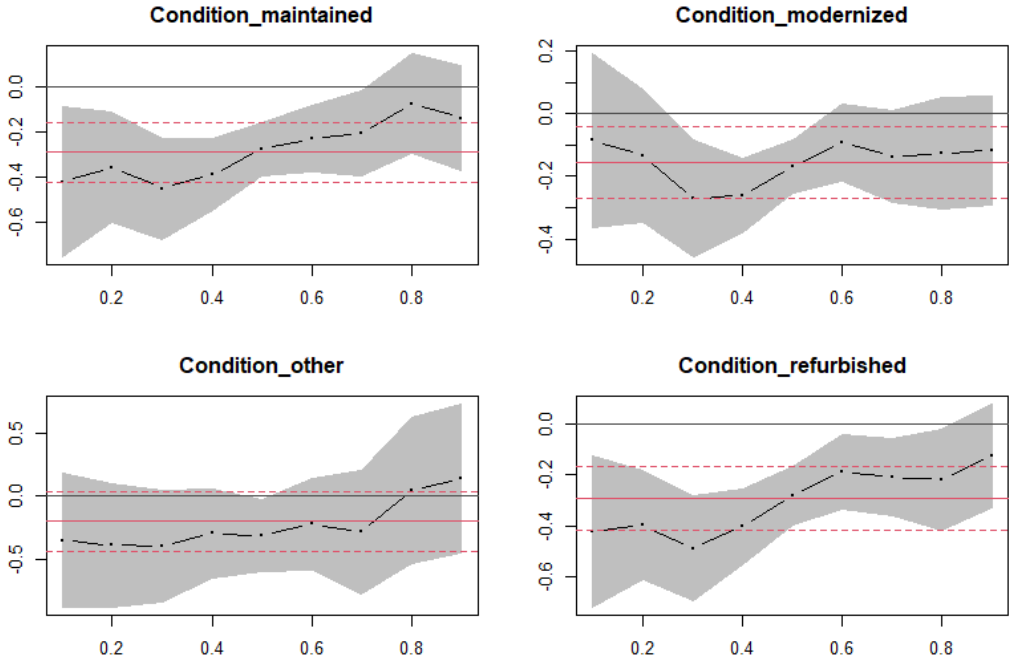
Źródło: Opracowanie własne.

Rysunek 5: Estymatory dla kwantyli



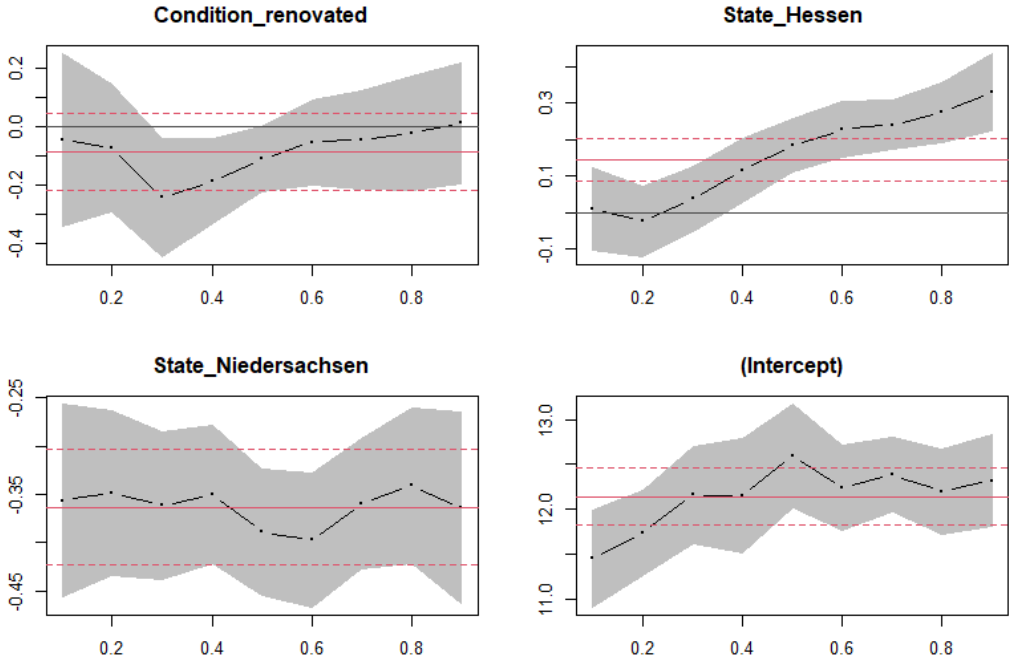
Źródło: Opracowanie własne.

Rysunek 6: Estymatory dla kwantyli



Źródło: Opracowanie własne.

Rysunek 7: Estymatory dla kwantyli



Źródło: Opracowanie własne.

5 Podsumowanie

W niniejszym badaniu oszacowano hedoniczne równanie logarytmu cen nieruchomości w trzech sąsiadujących Landach niemieckich - Nadrenii Północnej-Westfalii, Hesji oraz Dolnej Saksonii. Metoda Najmniejszych Kwadratów nie przeszła pozytywnie podstawowej diagnostyki, heteroskedastyczność i brak normalności reszt nie jest jednak problemem fundamentalnym, podobnie jak obserwacje odstające. Mając jednak na uwadze pewne bolączki estymatora MNK występujące w niniejszym problemie oraz chcąc zbadać rozkład zmiennej zależnej nieco dokładniej oszacowano regresje kwantylowe dla decyli z zakresu 1-9. Atutem zaproponowanego modelu jest możliwość zbadania warunkowej zależności charakterystyk nieruchomości względem rozkładu cen (kwantyla) i możliwość formalnego przetestowania występujących różnic. W świetle otrzymanych wyników wyciągnięto wnioski o nieistotności występowania garażu w wycenie nieruchomości w całym rozkładzie cen; niezależności statystycznej różnic międzykwantylowych cen od powierzchni użytkowej - udział między 0.05% a 0.11%; malejącym względem kwantyli, negatywnym wpływie wieku nieruchomości na jej cenę; nieliniową zależność cen od powierzchni działki - dla klasy średniej jest ona najmniej ważna jednak i tak podwyższa cenę co jest spójne z oczekiwaniami; braku różnic statystycznych dla liczby łazienek, jednocześnie każda kolejna łazienka zwiększa cenę; ogrzewaniu olejnym jako najbardziej kosztownym systemie; malejącym pozytywnym wpływie dodatkowej w stosunku do jednej sypialni i nieistotności tego czynnika powyżej mediany oraz międzylandowym zróżnicowaniu rynku nieruchomości - w Hesji jest najdrożej i różnice te postępują wraz z ceną, a Saksonia Dolna jest zdecydowanie najtańszym z trzech porównywanych landów (ok 35% niższe ceny). Otrzymano również często sprzeczne z intuicją oszacowania parametrów mówiących o stanie nieruchomości, należy mieć jednak na uwadze, że jest to dość subiektywna ocena sprzedawcy i pojęcie np. "do drobnych napraw" może znacząco się różnić między stronami transakcji. Należy mieć również na uwadze, że testowane były jednorównaniowe hipotezy proste, co oznacza, że rzeczywisty poziom istotności w przypadku wnioskowania jednoczesnego o wszystkich różnicach może się różnić od założonego. Celem projektu było natomiast pokazanie, że hedoniczne równanie cen jest warunkowo zależne od rozkładu zmiennej zależnej, co zostało dokonane. Wykorzystanie regresji kwantylowej w tego rodzaju analizach wydaje się zatem być uzasadnione.

Literatura

[1] Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.

[2] Arikan Eban Çaglayan Ebru. Determinants of house prices in istanbul: a quantile regression approach. *Quality and Quantity*, 45:305—-317, 2011.

6 Załącznik

Tablica 3: Outliery

id	Odległość Cooka	Standaryzowane reszty	Studentyzowane reszty	Statystyka dźwigni
26	0.0029	-2.3257	-2.3293	0.0127
51	0.0031	2.8597	2.867	0.009
66	0.0065	2.5796	2.5848	0.0228
91	0.0048	-1.4441	-1.4446	0.0523
92	0.0068	-2.2165	-2.2195	0.0322
109	0.0058	-1.4484	-1.449	0.0626
124	0.0037	-1.6545	-1.6555	0.0312
126	0.0039	-1.8374	-1.839	0.0269
127	0.004	-2.0121	-2.0143	0.0229
135	0.0065	-3.9383	-3.9588	0.01
144	0.0044	2.6526	2.6583	0.0149
148	0.003	-1.7658	-1.7671	0.0228
151	0.0033	1.5537	1.5545	0.0316
161	0.0032	1.5742	1.575	0.0302
169	0.0037	-2.2673	-2.2707	0.0172

Tablica 3 – ciąg dalszy z poprzedniej strony

id	Odległość Cooka	Standaryzowane reszty	Studentyzowane reszty	Statystyka dźwigni
173	0.0054	-1.5792	-1.5801	0.0496
189	0.0033	2.2479	2.2512	0.0152
194	0.0042	1.237	1.2372	0.0614
203	0.0089	3.0992	3.1087	0.0217
206	0.0036	-1.1159	-1.116	0.0644
214	0.0035	2.8201	2.8271	0.0104
231	0.0088	1.9331	1.935	0.0537
242	0.0113	1.9573	1.9592	0.0663
248	0.0045	2.4391	2.4434	0.0179
295	0.0101	2.1589	2.1617	0.0495
304	0.0081	2.643	2.6487	0.0271
305	0.0062	1.5605	1.5613	0.0578
307	0.0028	-2.1839	-2.1868	0.0139
335	0.0055	2.5371	2.5421	0.0201
371	0.0058	-2.2267	-2.2298	0.0274
382	0.0038	1.9053	1.9071	0.0246
394	0.0168	-2.7005	-2.7066	0.0525
401	0.0068	-1.5223	-1.523	0.0658
403	0.0033	1.5665	1.5674	0.0312
406	0.0036	-1.3149	-1.3152	0.0476
409	0.008	2.0393	2.0416	0.0439
418	0.0035	-1.6813	-1.6824	0.0288
436	0.0055	-2.6743	-2.6802	0.0182
438	0.011	3.6133	3.6289	0.0198
441	0.0262	4.5728	4.6055	0.0292
442	0.0028	-1.9648	-1.9668	0.0173
447	0.0051	-2.1106	-2.1132	0.0266
450	0.0121	-3.3797	-3.3923	0.0249
473	0.0117	3.4798	3.4937	0.0227
484	0.0068	2.6963	2.7023	0.0218
513	0.0064	-2.9013	-2.909	0.018
514	0.004	1.3401	1.3404	0.0503
547	0.0029	-1.7033	-1.7045	0.0232
571	0.0031	1.898	1.8998	0.0199
573	0.0031	2.5629	2.568	0.0113
597	0.0042	1.836	1.8376	0.0292
614	0.0031	2.4102	2.4144	0.0125
631	0.0103	-2.8596	-2.867	0.0294
637	0.004	-1.3226	-1.323	0.0525
640	0.0057	-2.5279	-2.5327	0.0209
642	0.0055	-2.7718	-2.7785	0.0168
643	0.0057	-2.5279	-2.5327	0.0209
647	0.0051	-1.8323	-1.8339	0.0353
654	0.0031	-1.7589	-1.7602	0.0234
676	0.0065	2.8386	2.8458	0.0191
737	0.0062	-2.0237	-2.0259	0.0348
760	0.0059	1.6325	1.6335	0.0501
774	0.032	3.1434	3.1534	0.072
799	0.0081	2.4493	2.4537	0.0314
834	0.0062	-1.6204	-1.6214	0.0536
855	0.0146	3.8969	3.9167	0.0226
860	0.0037	-1.9631	-1.9651	0.0223
866	0.004	-2.2198	-2.2229	0.0193
867	0.0114	1.7809	1.7823	0.0796

Tablica 3 – ciąg dalszy z poprzedniej strony

id	Odległość Cooka	Standaryzowane reszty	Studentyzowane reszty	Statystyka dźwigni
869	0.0067	-2.2239	-2.227	0.0316
879	0.0086	-2.1085	-2.1111	0.0443
912	0.0029	1.9205	1.9224	0.0185
923	0.0036	1.7569	1.7582	0.0274
952	0.003	-1.9651	-1.9671	0.0181
953	0.006	2.9265	2.9344	0.0165
977	0.0038	-2.5289	-2.5337	0.0141
981	0.0058	-1.5683	-1.5692	0.0539
1013	0.0208	-2.5087	-2.5134	0.0736
1018	0.0029	-2.6149	-2.6203	0.0102
1028	0.0033	1.5584	1.5592	0.0317
1061	0.0033	1.9624	1.9644	0.0199
1066	0.004	-2.2532	-2.2565	0.0187
1083	0.0108	-1.8143	-1.8157	0.0727
1145	0.0093	-2.9363	-2.9443	0.0252
1158	0.0066	2.5398	2.5447	0.0238
1167	0.0046	-1.7396	-1.7409	0.0353
1232	0.0051	2.304	2.3075	0.0225
1382	0.0032	-1.9856	-1.9877	0.0193
1392	0.0066	-2.4641	-2.4686	0.0253
1402	0.0068	-2.4818	-2.4864	0.0257

Źródło: Opracowanie własne.