Kraków
Maj 2020

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE

„Badanie czynników wpływających na wysokość dziennego kosztu zakwaterowania w Warszawie”

WERONIKA OWSIANKA

Spis treści

1. Wstęp ... 3
1.1. Opis problemu ... 3
1.2. Cel projektu ... 3
1.3. Źródło danych ... 3
1.4. Gretl ... 3
2.Specyfikacja modelu ... 4
2.1. Postać analityczna modelu ... 4
2.2. Wstępny model ... 4
2.3. Opis zmiennych ... 5
2.3. Hipotezy badawcze ... 5
2.4. Poziom istotności ... 6
2.5. Statystyki opisowe danych ... 6
2.6. Outliery ... 7
2.7. Weryfikacja współczynnika zmienności ... 7
2.8. Macierz korelacji ... 7
2.9. Klasyczna metoda najmniejszych kwadratów ... 8
2.11. Oszacowany model klasyczną metodą najmniejszych kwadratów ... 9
2.12. Test normalności reszt ... 10
2.13. Test t-Studenta – badanie istotności pojedynczych zmiennych objaśniających. ... 10
2.13. Współczynnik determinacji ... 11
3.Metody doboru zmiennych ... 12
3.1. Metoda Hellwiga ... 12
3.1.1 Postać modelu ... 13
3.2. Metoda krokowa wstecz ... 13
3.3. Wybór postaci modelu ... 14
3.4. Badanie koincydencji ... 15
3.5. Badanie współliniowości ... 16
3.6. Postać modelu i interpretacja parametrów ... 16
4. Weryfikacja modelu ... 17
4.1. Badanie efektu katalizy ... 17
4.1.1. Postać modelu po usunięciu katalizatora ... 18
4.2. Test liczby serii- testowanie losowości reszt modelu ... 19
4.3. Test Ramseya RESET- testowanie liniowości modelu ... 20
4.4. Test Chowa - testowanie stabilności funkcyjnej modelu ... 21
4.5. Badanie heteroskedastyczności – test White'a ... 21

4.6. Badanie heteroskedastyczności – test Breuscha-Pagan'a 22
5. Prognozy 23
5.1. Prognoza ex-ante 23
5.2. Prognoza ex-post 24
7. Weryfikacja hipotez 24
8. Ostateczny model 25
9. Bibliografia 26

1. Wstęp

1.1. Opis problemu

Rozwój nowoczesnej gospodarki turystycznej, z uwagi na perspektywy rozwoju polskiej turystyki w gospodarce XXI wieku determinowany jest w dużym stopniu przez rozwój hotelarstwa. Wciąż kluczowymi miejscami branży hotelarskiej są największe miasta m.in. Warszawa.

W trakcie poszukiwań miejsca noclegu w stolicy Polski uwzględniamy wiele czynników. Rozwój Internetu znacząco zmienił warunki funkcjonowania i konkurowania branży hotelarskiej. Jednym z elementów tej nowej rzeczywistości są elektroniczne oceny konsumenckie i możliwość wystawiania opinii, który bierzemy pod uwagę przy wyborze hotelu. Zazwyczaj oferty noclegów w samym centrum Warszawy są nieco droższe, ale dogodna lokalizacja sprawia, że masz do dyspozycji największe atrakcje w okolicy. Decydując się na konkretny hotel rozpatrujemy też dostęp do bezpłatnego parkingu czy też wielkość apartamentu. Co tak naprawdę wpływa na wysokość ceny noclegu w Warszawie?

1.2. Cel projektu

Wiele osób, które chciałyby odwiedzić Warszawę, jest przekonanych, że ceny noclegu w Warszawie są wysokie – w tym apartamentów i pokoi na wynajem krótkoterminowy i zależy im na najtańszym rozwiązaniu. Czy zawsze wysoka cena jest równoznaczna z dobrą lokalizacją bądź hotel mający dobrą ocenę użytkowników na stronie znaczy duży koszt zakwaterowania?

Celem przedstawionego projektu jest zbadanie wpływu 9 czynników na cenę dziennego zakwaterowania w Warszawie oraz określenie siły ich oddziaływania.

1.3. Źródło danych

Wszystkie dane wykorzystane w projekcie pochodzą ze serwisu „Booking.com” (50 obserwacji), zostały zebrane z wystawionych ogłoszeń na nocleg w dniu 01.09.2020r.

1.4. Gretl

Do przeprowadzenia badań zostanie użyty program „Gretl”, autorstwa Allina Cottrella z Uniwersytetu Wake Forest w Północnej Karolinie w Stanach Zjednoczonych, jest rozwijanym od kilku lat pakietem ekonometrycznym. Program ten jest niezwykle popularny i powszechnie stosowany na świecie, zawiera podstawowe procedury i metody ekonometryczne.

2.Specyfikacja modelu

2.1. Postać analityczna modelu

Budowany model regresji wielorakiej pozwala na zbadanie wpływu wielu zmiennych niezależnych ($X_1, X_2, \ldots, X_k$) na jedną zmienną zależną (Y). Najczęściej wykorzystywaną odmianą regresji wielorakiej jest Liniowa Regresja Wieloraka. Liniowy model regresji wielorakiej przyjmuje postać:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + \varepsilon$$

Równanie 1: Liniowy model ekonometryczny

,gdzie:

Y- zmienna objaśniana,

X_j - zmiennie objaśniające j=1,2,3,…,k,

α_j - nieznane parametry strukturalne modelu j=0,1,…,k

ε –składnik losowy

Rozpatrujemy liniową zależność zmiennej objaśnianej od zmiennych objaśniających i składnika losowego

2.2. Wstępny model

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \ldots + \alpha_9 x_9 + \varepsilon$$

,gdzie:
Zmienna objaśniana:

y- cena noclegu dla 2 osób

Zmienne objaśniające:

- ilościowe:

x_1- centrum [m]
x_2 - zamek [m]
x_3 - łazienki [m]
x_4 - opinie
x_5 - rating
x_6 – powierzchnia [m^2]

- binarne:

x_7 - parking
x_8 - toaleta
x_9 – zwierzęta

2.3. Opis zmiennych

1. y- cena noclegu dla 2 osób

Zmienna objaśniana „Y" w modelu. Jest to koszt (w złotówkach) dziennego noclegu w Warszawie.

2. x_1- centrum [m]

Zmienna ilościowa mówiąca ile wynosi odległość od miejsca zakwaterowania do dworca PKP Warszawa Centralna.

3. x_2 - zamek [m]

Zmienna ilościowa wyrażająca odległość od miejsca zakwaterowania w metrach do Zamku Królewskiego w Warszawie, w jego pobliżu znajdują się też inne najpopularniejsze atrakcje turystyczne (np. kolumna Zygmunta, Starówka, rynek Starego Miasta).

4. x_3 - łazienki [m]

Zmienna ilościowa mówiąca ile wynosi odległość od miejsca zakwaterowania do Łazienek Królewskich.

5. x_4 - opinie

Zmienna ilościowa wyrażająca liczbę wystawionych opinii dla danego ośrodka.

6. x_5 - rating

Zmienna ilościowa wyrażająca średnią ocenę wystawioną na stronie booking.com przez gości po ich pobycie w danym ośrodku.

7. x_6 – powierzchnia [m^2]

Zmienna ilościowa opisująca, ile metrów kwadratowych ma powierzchnia apartamentu.

8. x_7 – parking

Zmienna binarna mówiąca czy ośrodek oferuję bezpłatny parking dla swoich klientów.

$$\begin{cases} 1 & , gdy\ możliwość\ bezpłatnego\ parkingu \\ 0 & , gdy\ brak\ możliwości\ bezpłatnego\ parkingu \end{cases}$$

9. x_8 - toaleta

Zmienna binarna mówiąca czy zakwaterowanie jest z łazienką w pokoju.

$$\begin{cases} 1 & , gdy\ łazienka\ w\ pokoju \\ 0 & , gdy\ brak\ łazienki\ w\ pokoju \end{cases}$$

10. x_9 - zwierzęta

Zmienna binarna mówiąca czy zwierzęta domowe są akceptowane.

$$\begin{cases} 1 & , gdy\ zwierzęta\ są\ akceptowane \\ 0 & , gdy\ zwierzęta\ nie\ są\ akceptowane \end{cases}$$

2.3. Hipotezy badawcze

Przed rozpoczęciem analizy tematu należy postawić zestaw badawczych hipotez, które zostaną zweryfikowane po przeprowadzeniu badań:

1) Na wysokość ceny noclegu najsilniej wpływa odległość od Zamku Królewskiego.

2) Wielkość apartamentu nie ma znaczenia na cenę noclegu.

3) Im lepsza ocena użytkowników na stronie tym droższy nocleg.

4) Im bliżej Zamku Królewskiego tym droższy nocleg.

2.4. Poziom istotności

Do przeprowadzania badań przyjęto poziom istotności na poziomie 5%, jest to dopuszczalne ryzyko popełnienia błędu rodzaju I, czyli uznania prawdziwej hipotezy zerowej za fałszywą.

2.5. Statystyki opisowe danych

Zmienne	*Średnia*	*Mediana*	*Odchylenie Standardowe*	*Skośność*	*Kurtoza*	*Maksimum*	*Minimum*
cena	208,38	200	59,67	-0,18	-0,63	330	85
centrum	3503,3	2800	2962,0	0,68	-0,79	10100	250
zamek	4395,6	3000	3254,4	0,81	-0,74	11100	1100
łazienki	4175,6	3800	1915,1	0,45	-0,66	8300	1100
opinie	1884,6	1283	1608,6	0,99	-0,17	5369	5
rating	8,08	8,3	0,89	-0,41	-1,03	9,3	6,3
powierzchnia	26,3	25	8,24	0,41	-1,02	43	15
parking	0,67	1	0,48	-0,71	-1,5	1	0
toaleta	0,89	1	0,32	-2,48	4,1	1	0
zwierzęta	0,54	1	0,5	-0,13	-1,9	1	0

Tabela 1: Statystyki opisowe dla obserwacji z próby 1-45

Interpretacja:

- Cena- średnia cena noclegu wynosi około 208 złotych, a wyniki odchylają się średnio o 60 złotych. Lewostronna skośność pokazuję, że większość cen zakwaterowania ma cenę wyższą od średniej.
- Centrum – średnia odległość od Dworca Centralnego wynosi 3503 metrów, odchylenie od tej odległość wynosi około 2962 metrów. Większość badanych ośrodków jest bliżej centrum niż średnia wartość. Najbliższy ośrodek znajduję się jedynie 250 metrów od dworca.
- Zamek- średnia odległość od Zamku Królewskiego wynosi 4295 metrów, odchylenie od tej odległość wynosi około 3254 metrów. Tak we wcześniejszym punkcie, większość badanych ośrodków jest bliżej zamku niż średnia wartość.
- Łazienki- średnia odległość od Łazienek Królewskich wynosi 4175 metrów, odchylenie od tej odległość wynosi około 1915 metrów. Tak we wcześniejszych punktach, większość badanych ośrodków jest bliżej zamku niż średnia wartość.
- Opinie- średnia liczba opinii wynosi 1885 opinii. Najmniejsza liczba wystawionych opinii dla danego ośrodka wynosi 5.
- Rating- średnia ocena to 8,08 z średnim odchyleniem 0,89. Najgorsza ocena wynosi 6,3 a najlepsza 9,3.
- Powierzchnia- średnia wielkość apartamentu wynosi 26m^2 z odchyleniem około 8,2m^2. Najmniejszy apartament ma 15 metrów kwadratowych a największy 43 m^2.
- Parking- w większości ośrodków znajduje się bezpłatny parking
- Toaleta- większość apartamentów ma własną łazienkę w pokoju
- Zwierzęta- zauważamy, że jedynie połowa ośrodków zezwala na pobyt ze zwierzętami

2.6. Outliery

Obserwacja odstająca (outlier) – obserwacja znacząco różniąca się od pozostałych. Obserwacje odstające mogą mieć znaczący wpływ na wynik testu statystycznego, dlatego powinny zostać usunięte lub przekształcone. Outlier może zmienić nachylenie linii korelacji, co zwiększa prawdopodobieństwo popełniania błędu pierwszego lub błędu drugiego rodzaju. W badanym zbiorze danych nie ma wartości odstających.

Nasza próbka składa się z 45 obserwacji (nie występują żadne braki w danych).

2.7. Weryfikacja współczynnika zmienności

Współczynnik zmienności –nam jak bardzo grupa obserwacji jest zróżnicowana względem pewniej cechy. Umownie przyjmuje się, że jeżeli współczynnik V nie przekracza 10%, to cechy wykazują niewielkie zróżnicowanie.

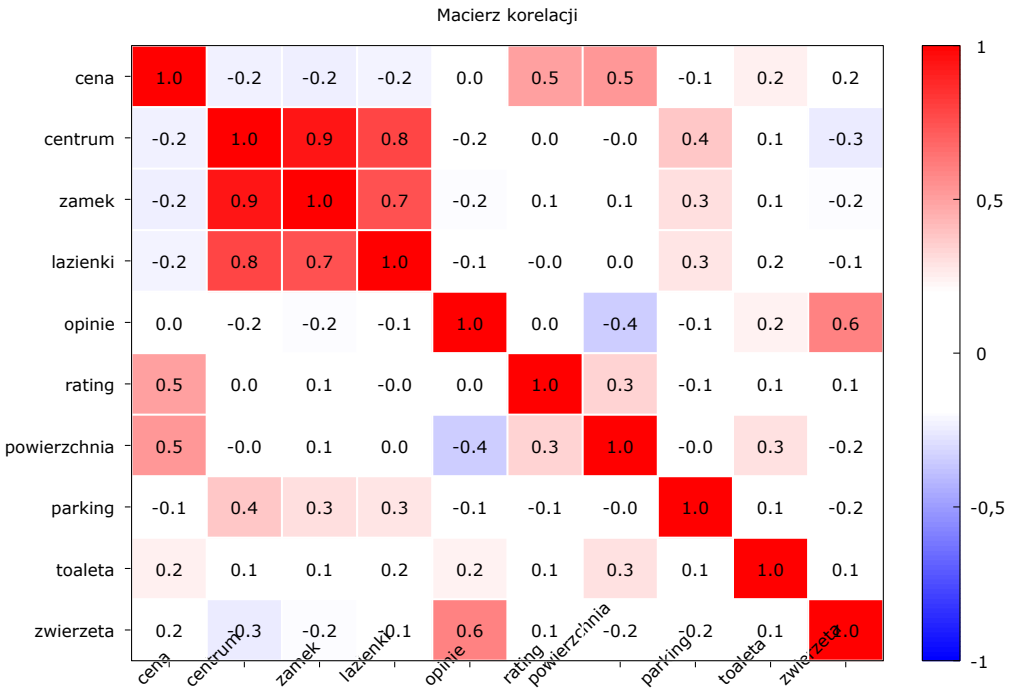
Zmienna	*Współczynnik Zmienności*
cena	0,28637
centrum	0,84549
zamek	0,74039
łazienki	0,45866
opinie	0,85355
rating	0,11023
lóżka	0,31388
parking	0,71510
Toaleta	0,35755
zwierzęta	0,94598

Tabela 2: Współczynnik zmienności

Pierwszy model będzie posiadał wszystkie zmienne, ze względu na poziom współczynnika zmienności powyżej 10%.

2.8. Macierz korelacji

Badając korelację liniową zmiennych, sprawdzamy, jak zachowują się one względem siebie np. gdy jedna wzrasta to druga maleje. Współczynnik korelacji liniowej, mieści się w przedziale <1,-1>, gdzie 1 oznacza silną korelację dodatnią (zmienne reagują w tym samym kierunku) a -1 silna korelację ujemną (zmienne reagują w przeciwnym kierunku). Jeżeli współczynnik wynosi 0 mówimy o braku korelacji.



Rysunek 1: Współczynniki korelacji

Przypuszczamy, że do modelu najprawdopodobniej będą wzięte pod uwagę zmienne „powierzchnia" (0,52) i „rating" (0,49), ponieważ są one najsilniej skorelowane ze zmienną objaśnianą a nie są silnie skorelowane ze sobą (0,33). Nieco słabiej jest skorelowana jest zmienna „toaleta" ze zmienną objaśnianą, ale nadal możliwe będzie, że będzie brana do modelu. Zmienne „centrum", „zamek" i „łazienki" wszystkie mają podobną korelacje ze zmienną objaśnianą „cena" na poziomie -0,24. Jednak są one też silnie skorelowane ze sobą, więc prawdopodobnie wszystkie nie wejdą do modelu jednocześnie. Najsłabiej skorelowana jest zmienna „opinie" (0,0449) i za pewne nie będzie brana pod uwagę.

Po otrzymanych wynikach możemy przypuszczać, że cena noclegu w Warszawie zależy głownie od powierzchni pokoju i oceny użytkowników na stronie, odległości od centrum/Zamku/Łazienek. Dobór zmiennych wymaga jednak dalszej analizy i znajomość korelacji nie niewystarczająca do podjęcia ostatecznej decyzji.

2.9. Klasyczna metoda najmniejszych kwadratów

Najczęściej stosowaną metodą estymacji modelu ekonometrycznego jest klasyczna metoda najmniejszych kwadratów. Polega ona na wyznaczeniu takich oszacowań parametrów, dla których suma kwadratów reszt jest najmniejsza:

$$\chi = \sum_{i=1}^{n} e_i^2 \rightarrow min.$$

Równanie 2: Minimalizacja sumy kwadratów reszt

w wyniku czego, otrzymujemy następujący estymator oszacowań parametrów liniowego modelu regresyjnego:

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Równanie 3: Estymator parametrów strukturalnych

Twierdzenie Gaussa – Markowa

Estymator $\hat{\beta}$ uzyskany Klasyczną Metodą Najmniejszych Kwadratów jest estymatorem BLUE [best linear unbiased estimator], tj. zgodnym, nieobciążonym i najefektywniejszy w klasie liniowych estymatorów wektora β.

Założenia Gaussa – Markowa:

- Zmienne objaśniające są niezależne, nielosowe i nieskorelowane ze składnikiem losowym
- Wartości oczekiwane składników losowych są równe zeru: $E(\varepsilon)$= 0.
- Wariancje składników losowych są stałe, tzn. $D^2(\varepsilon) = E(\varepsilon\varepsilon^T) = \sigma^2 I$
- $\mathrm{r(X)} = \mathrm{k} + 1 \leq \mathrm{n}$, gdzie r() oznacza rząd macierzy
- Składniki losowe $\varepsilon_i, i = 1,2, \dots, n$ mają rozkład losowy

2.11. Oszacowany model klasyczną metodą najmniejszych kwadratów

Model 2: Estymacja KMNK, wykorzystane obserwacje 1-45
Zmienna zależna (Y): cena

	Współczynnik	*Błąd stand.*	*t-Studenta*	*wartość p*	
const	−58,0631	69,7795	−0,8321	0,4110	
centrum	0,00846275	0,00774468	1,093	0,2820	
zamek	−0,00995063	0,00630322	−1,579	0,1234	
lazienki	−0,00499668	0,00604141	−0,8271	0,4138	
opinie	0,00134557	0,00615383	0,2187	0,8282	
rating	22,7435	8,61612	2,640	0,0123	**
powierzchnia	3,31525	1,06361	3,117	0,0036	***
parking	0,409977	16,1007	0,02546	0,9798	
toaleta	19,7221	25,5076	0,7732	0,4446	
zwierzeta	19,3285	18,0379	1,072	0,2913	

Średn.aryt.zm.zależnej	208,3778	Odch.stand.zm.zależnej	59,67232
Suma kwadratów reszt	73741,40	Błąd standardowy reszt	45,90095
Wsp. determ. R-kwadrat	0,529334	Skorygowany R-kwadrat	0,408306
F(9, 35)	4,373635	Wartość p dla testu F	0,000704
Logarytm wiarygodności	−230,3895	Kryt. inform. Akaike'a	480,7790
Kryt. bayes. Schwarza	498,8457	Kryt. Hannana-Quinna	487,5141

Rysunek 2: Model estymowany KMNK

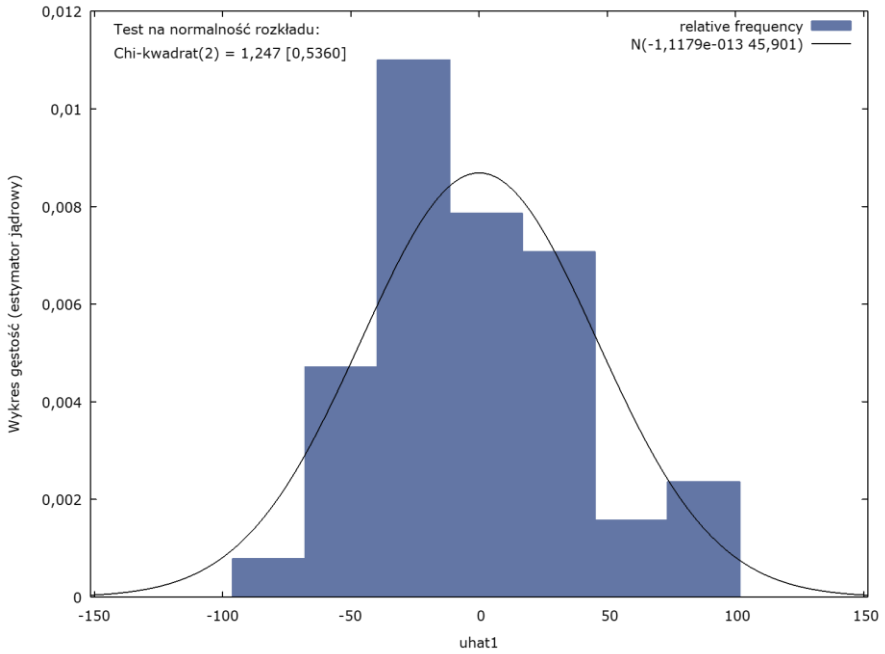
Przed sprawdzeniem istotności każdej zmiennej objaśniającej należy sprawdzić czy założenie testu t-Studenta, z którego skorzystamy jest prawdziwe. W celu tego przeprowadzony zostanie test normalności rozkładu reszt modelu.

2.12. Test normalności reszt

Jednym z warunków poprawności modelu ekonometrycznego jest rozkład normalny reszt. Istnieje wiele testów do jego weryfikacji jednym z nich jest test Doornika - Hansena.

$$H_0 \sim \textit{reszty mają rozkład normalny}$$
$$H_1 \sim \textit{reszty nie mają rozkładu normalnego}$$

Rysunek 3: Hipotezy testu Doornika-Hansena



Rysunek 4: Wykres reszt modelu

Wartość p-value testu Doornika - Hansena wynosi 0.54 , czyli nie mamy postaw do odrzucenia hipotezy zerowej. Rozkład reszt w tym modelu jest rozkładem normalnym.

2.13. Test t-Studenta – badanie istotności pojedynczych zmiennych objaśniających.

W modelu regresji zwykle weryfikujemy istotność wszystkich współczynników (poza stałą). Dokonuje się tego testując, czy oszacowanie parametru jest statystycznie różne od zera, korzystając z testu t-Studenta Jeśli tak jest, to określona zmienna objaśniająca X ma istotny wpływ na zmienną objaśnianą Y i mówimy, że zmienna X jest istotna statystycznie. Zmienne o najniższym p-value są najbardziej istotne i wiążą się z najmniejszym prawdopodobieństwem popełniania błędu.

Hipoteza zerowa

$H_0: \beta_k = 0$ (zmienna k jest nieistotna statystycznie)

Hipoteza alternatywna

$H_1: \beta_k \neq 0$ (zmienna k jest istotna statystycznie)

Rysunek 5: Hipotezy testu t-Studenta

Zmienna	*p-value*
const	0,4110
centrum	0,2820
zamek	0,1234
lazienki	0,4138
opinie	0,8282
rating	0,0123
powierzchnia	0,0036
parking	0,9798
toaleta	0,4446
zwierzeta	0,2913

Tabela 3: Wartości p-value zmienncyh dla testu t-Studenta

Najniższe p -value ma zmienna powierzchnia i rating, tak jak przypuszczaliśmy po analizie korelacji zmiennych. Pozostałe czynniki mają wartość p -value powyżej 5% zatem są nieistotne w modelu. Najmniej istotną zmienną jest parking, pomimo po macierzy korelacji spodziewaliśmy się, że to zmienna opinie będzie najmniej istotna.

2.13. Współczynnik determinacji

Współczynnik determinacji informuje o tym, jak część zmian zmiennej objaśnianej jest wyjaśniona przez zmiany zmiennej objaśniającej. Inaczej mówiąc, pokazuje jaki procent zmiennej zależnej (objaśnianej) jest wyjaśniany za pomocą zmiennej niezależnej. Informuje nas, ile nasz model, nasz badany czynnik wyjaśnia zgromadzone dane pomiarowe. Można go przedstawić za pomocą poniższego wzoru:

$$R^2 = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y}\backslash)^2}{\sum_{t=1}^{n}(y_t - \bar{y}\backslash)^2}$$

Równanie 4: Współczynnik determinacji

gdzie:

R^2 - współczynnik determinacji

y_t - rzeczywista wartość zmiennej zależnej

$\hat{y}_t$ - przewidywana wartość zmiennej zależnej

$\bar{y}$ - średnia wartość rzeczywistej zmiennej zależnej

Dla wyestymowanego modelu współczynnik determinacji wynosi:

Wsp. determ. R-kwadrat	0,529334

Zbudowane model umożliwia objaśnienie zmienności zmiennej Y w 53%. Należy pamiętać, że współczynnik determinacji jest miarą dopasowania modelu ekonometrycznego do danych empirycznych, lecz informacja jaką niesie o modelu może być fałszywa, jeśli w modelu występują zmienne, które nazywamy katalizatorami.

3.Metody doboru zmiennych

3.1. Metoda Hellwiga

Jest to metoda doboru zmiennych objaśniających do modelu statystycznego, a w szczególności ekonometrycznego. Zmienne wybrane do liniowego modelu ekonometrycznego powinny być silnie skorelowane ze zmienną objaśnianą, a słabo między sobą. Nie jest to jednak ścisłe kryterium doboru zmiennych. Dlatego też potrzebne jest kryterium liczbowe, które pozwoli wybrać tę spośród branych pod uwagę kombinacji potencjalnych zmiennych objaśniających, która je spełnia. Na tej idei jest oparta metoda pojemności nośników informacji, czyli metoda Hellwiga. W obliczeniach wykorzystuje się współczynniki korelacji między zmiennymi, w tym wektor współczynników korelacji między zmienną objaśnianą a zmienną objaśniającymi:

$$R_o = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$$

macierz współczynników korelacji między zmiennymi objaśniającymi:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{12} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \cdots & 1 \end{bmatrix}$$

Wyszukana zostaje najlepsza kombinacja zmiennych objaśniających o największym integralnym wskaźniku pojemności informacyjnej. Do wyboru jest $L = 2^n - 1$ kombinacji zmiennych objaśniających. Dla każdej zmiennej zawartej w kombinacji definiuję się tzw. Indywidualną pojemność nośników informacji. Indywidualnie wskaźniki pojemności informacyjnej zmiennych dla rozpatrywanej l-tej kombinacji są zdefiniowane następująco:

$$h_{kj} = \frac{r_j^2}{1 + \sum_{l=1}^{m_k} |r_{lj}|}, (j=1, 2, \ldots, m_k),$$

gdzie:
h_{kj} – indywidualna pojemność informacyjna j-tej zmiennej w l-tej kombinacji,
r_j – wartość wektora korelacji R_0,
r_{lj} – wartość z macierzy korelacji R,
l – numer kombinacji,
j – numer zmiennej w kombinacji ($j = 1, 2 \ldots, m_k$),
m_k – liczba zmiennych w k-tej kombinacji.

Po obliczeniu wartości indywidualnych pojemności nośników informacji dla wszystkich zmiennych zawartych w kombinacji oblicza się pojemność integralną kombinacji nośników informacji według wzoru:

$$H_k = \sum_{j=1}^{mk} h_{kj} (k = 1, 2, \ldots, 2^m - 1).$$

Wybiera się tę kombinację, dla której H_k jest najwyższa.

Przeprowadzono metodę Hellwiga w badanej próbie biorąc pod uwagę wszystkie zmienne objaśniające, aby dobrać optymalny podzbiór. Otrzymano:

```
? H_max
0,44055638
? najlepszalista
centrum rating powierzchnia
```

Metoda Hellwiga wskazuje, że najlepszy model objaśniający wpływ czynników na wysokość ceny zakwaterowania w Warszawie można zbudować przy użyciu zmiennych: odległość od Dworca Centralnego w Warszawie, oceny użytkowników na stronie booking.com i powierzchni pokoju -podobnie jak się spodziewaliśmy po macierzy korelacji.

3.1.1 Postać modelu

Oszacowano model:

Model 5: Estymacja KMNK, wykorzystane obserwacje 1-45
Zmienna zależna (Y): cena

	Współczynnik	*Błąd stand.*	*t-Studenta*	*wartość p*	
const	−51,1026	63,1607	−0,8091	0,4231	
centrum	−0,00507178	0,00232713	−2,179	0,0351	**
powierzchnia	2,89075	0,885900	3,263	0,0022	***
rating	24,9293	8,21022	3,036	0,0041	***

Średn.aryt.zm.zależnej	208,3778	Odch.stand.zm.zależnej	59,67232
Suma kwadratów reszt	85595,75	Błąd standardowy reszt	45,69137
Wsp. determ. R-kwadrat	0,453672	Skorygowany R-kwadrat	0,413696
F(3, 41)	11,34882	Wartość p dla testu F	0,000015
Logarytm wiarygodności	−233,7436	Kryt. inform. Akaike'a	475,4872
Kryt. bayes. Schwarza	482,7139	Kryt. Hannana-Quinna	478,1813

Dla tej postaci modelu wykonując test normalności reszt Doornika – Hansena p -value wynosi 0,4407 , co oznacza że reszty pochodzą z rozkładu normalnego. A wykonując test t-Studenta otrzymujemy wynik, że wszystkie zmienne są statystycznie istotne (p-value<5%).

3.2. Metoda krokowa wstecz

Wychodzimy od modelu ze wszystkimi potencjalnymi zmiennymi objaśniającymi. Następnie w kolejnych etapach usuwane są kolejne najmniej istotne zmienne, aż do uzyskanie modelu uwzględniającego tylko zmienne istotne.

Za pomocą metody krokowej wstecz otrzymano model:

METODA KROKOWA WSTECZ
Model 24: Estymacja KMNK, wykorzystane obserwacje 1-45
Zmienna zależna (Y): cena

	Współczynnik	Błąd stand.	t-Studenta	wartość p	
const	−52,6278	61,6582	−0,8535	0,3983	
zamek	−0,00542444	0,00207723	−2,611	0,0125	**
rating	25,5624	8,04043	3,179	0,0028	***
powierzchnia	2,98549	0,866836	3,444	0,0013	***

Średn.aryt.zm.zależnej	208,3778	Odch.stand.zm.zależnej	59,67232
Suma kwadratów reszt	81891,49	Błąd standardowy reszt	44,69176
Wsp. determ. R-kwadrat	0,477315	Skorygowany R-kwadrat	0,439070
F(3, 41)	12,48036	Wartość p dla testu F	6,17e-06
Logarytm wiarygodności	−232,7482	Kryt. inform. Akaike'a	473,4964
Kryt. bayes. Schwarza	480,7231	Kryt. Hannana-Quinna	476,1904

Zmienne usuwane w kolejności:

1)parking
2)opinie
3)łazienki
4)toaleta
5)centrum
6)zwierzęta

Po każdym wyeliminowaniu zmiennej zostaje sprawdzany poziom p-vaulue testu normalności rozkładu reszt – zawsze był powyżej 5%. Dla końcowej postaci modelu wykonując test normalności reszt Doornika – Hansena p -value wynosi 0,5595 , co oznacza że reszty pochodzą z rozkładu normalnego. Zauważamy, że otrzymany model różni się od modelu otrzymanego metodą Hellwiga. Konieczne jest wybranie lepszego z nich.

3.3. Wybór postaci modelu

Z powodu, że wyniku obu metod otrzymano różne modele należy wybrać ten, który lepiej opisuje badane zjawisko- zmienność wysokości ceny noclegu.

Porównanie otrzymanych modeli, bierzemy pod uwagę współczynnik determinacji oraz kryteria informacyjne:

	METODA HELLWIGA	METODA KROKOWA WSTECZ
Wsp. determ. R-kwadrat	0,453672	0,477315
Kryt. bayes. Schwarza	482,7139	480,7231
Kryt. inform. Akaike'a	475,4872	473,4964
Kryt. Hannana-Quinna	478,1813	476,1904

Tabela 4:Porównanie metody Hellwiga z krokową wstecz

Bardziej wiarygodny jest model uzyskany metodą krokową wstecz, ze względu na wyższa wartość współczynnika determinacji. Model ten objaśnia zmienną „cena" w 48% , czyli lepiej niż model z metody Hellwiga, w którym R- kwadrat jest na poziomie 45%. Fryterium Bayes-Schwarza, Akaike-a oraz Hannana-Quinna są niższe po wykorzystaniu metody krokowej-wstecz, co mówi o tym, że ten model jest dokładniejszy. Przy weryfikacji macierzy korelacji zmienne z modelu metody krokowej były podejrzewane o wprowadzenie do modelu. Korelacje ze zmienną „Cena" wynosiły kolejno 0.2 , 0.5 , 0.5.

3.4. Badanie koincydencji

Model ekonometryczny posiada własność koincydencji, jeśli dla każdej zmiennej objaśniającej znak współczynnika stojącego przy zmiennej w modelu jest równy współczynnikowi korelacji ze zmienną objaśnianą. Brak koincydencji często świadczy o współliniowości zmiennych objaśniających.

$$\operatorname{sign}\left(r\left(x_j, y\right)\right) = \operatorname{sign}\left(a_j\right),$$

gdzie: $\operatorname{sign}(r(x_j, y))$ – znak współczynnika korelacji pomiędzy zmienną objaśniającą x_j a zmienną objaśnianą y,
$\operatorname{sign}(a_j)$ – znak współczynnika a_j w modelu ekonometrycznym przy zmiennej x_j.

Sprawdzono model otrzymany metodą Hellwiga i metodą krokową wstecz:

Metoda Hellwiga

	Znak ai	Znak ri
centrum	-	-
rating	+	+
powierzchnia	+	+

Metoda krokowa wstecz

	Znak ai	Zank ri
zamek	-	-
rating	+	+
powierzchnia	+	+

Oba modele są koincydentne, więc do dalszych weryfikacji bierzemy model z metody krokowej wstecz.

3.5. Badanie współliniowości

Współliniowość innymi słowy oznacza sytuację, w której zmienne wprowadzane do modelu regresji są ze sobą silnie skorelowane i taka sytuacja powoduje „pogorszenie się" parametrów modelu. O występowaniu współliniowości mówi właśnie współczynnik VIF. Dla każdego ze zmiennych można go obliczyć ze wzoru:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Dla współczynnika współliniowości przyjęto następujące interpretacje:

- VIF=1 : brak współliniowości zmiennych
- 1<VIF<10 : występuje nieznaczna współliniowość zmiennych
- VIF>10 : występuje silna współliniowość zmiennych, należy usunąć z modelu zmienną

Wyniki:

zamek	1,007
rating	1,128
powierzchnia	1,125

Tabela 5: Wartości VIF

Wartości są niższe od 10, czyli zjawisko współliniowości nie występuje.

3.6. Postać modelu i interpretacja parametrów

$$y = -52{,}63 - 0{,}005x_1 + 25{,}56x_2 + 2{,}99x_3 + \varepsilon$$

,gdzie:

x_1-odległość od Zamku Królewskiego
x_2- ocena użytkowników
x_3- powierzchnia

Interpretacja parametrów:

- Każde zwiększenie odległości od Zamku Królewskiego o kilometr powoduje o spadek wysokości ceny noclegu o 5 zł (przy zasadzie ceteris paribus)
- Gdy ocena zwiększa się o 1, to cena noclegu rośnie o 25,56 zł (przy zasadzie ceteris paribus)
- Wraz ze wzrostem powierzchni o 1 metr kwadratowy cena zwiększa się o około 3 złotych(przy zasadzie ceteris paribus)

Przykład 1 : Pokój o powierzchni 20 m^2 z oceną 7,5, oddalony o 800 metrów od Zamku Królewskiego
Cena=-52.63-0.005*800+25.56*7.5+2.99*20= 194,87 zł

Przykład 2 : Pokój o powierzchni 26 m^2 z oceną 8, oddalony o 500 metrów od Zamku Królewskiego
Cena=-52.63-0.005*500+25.56*8+2.99*26= 227,09 zł

4. Weryfikacja modelu

4.1. Badanie efektu katalizy

Po sprawdzeniu, że współliniowość zmiennych nie występuje przechodzimy do sprawdzenia czy w modelu nie występują katalizatory, które zawyżają sztucznie współczynnik determinacji. Konieczne jest wtedy usunięcie wszystkich występujących w modelu katalizatorów.

Efekt katalizy zachodzi, gdy dla regularnej pary korelacyjnej zmienia X_i z pary (X_i,X_j) jest katalizatorem, jeżeli:

$$r_{ij} < 0 \text{ lub } r_{ij} > \begin{matrix} r_i \\ r_j \end{matrix}$$

Przy założeniu:

Regularna para korelacyjna: para $(\mathbf{R},\mathbf{R}_0)$, gdy współczynniki korelacji spełniają warunek :

$0 < |r_1| \leq |r_2| \leq \ldots \leq |r_k|$

Pomiar zjawiska katalizy :

-natężenie zjawiska katalizy:

$$\eta = R^2 - H,$$

gdzie H - integralna pojemność informacyjna zestawu zmiennych objaśniających;

-względne natężenie efektu katalizy:

$$W_\eta = \frac{\eta}{R^2}\ 100\%.$$

Zbadano czy w modelu są katalizatory:

```
KATALIZATOR:
zamek
W PARZE:
zamek rating
KATALIZATOR:
zamek
W PARZE:
zamek powierzchnia
#nat◆enie efektu katalizy
? ols Y const xlist --quiet
? H=helwig(Y,xlist)
Wygenerowano skalar H = 0,425144
? scalar natezenie_efektu_katalizy=$rsq-H
Wygenerowano skalar natezenie_efektu_katalizy = 0,0521705
? natezenie_efektu_katalizy
0,052170526
```

Katalizatorem w modelu okazała się zmienna „zamek", czyli zakłamuje ona stopień dopasowania modelu i należy ją z niego usunąć. Natężenie efektu katalizy wynosi 0,05. Dzieląc je przez współczynnik determinacji w ujęciu procentowym względne natężenie efektu katalizy wynosi 10,9%.

4.1.1. Postać modelu po usunięciu katalizatora

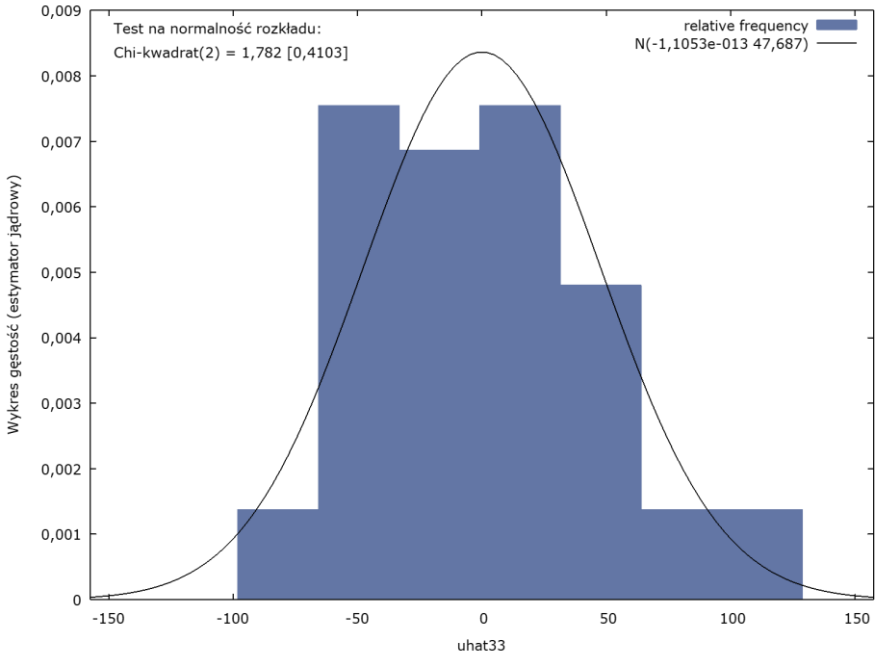
Model 33: Estymacja KMNK, wykorzystane obserwacje 1-45
Zmienna zależna (Y): cena

	Współczynnik	*Błąd stand.*	*t-Studenta*	*wartość p*	
const	−64,1925	65,6212	−0,9782	0,3336	
rating	24,2648	8,56299	2,834	0,0070	***
powierzchnia	2,91695	0,924517	3,155	0,0030	***

Średn.aryt.zm.zależnej	208,3778	Odch.stand.zm.zależnej	59,67232
Suma kwadratów reszt	95512,07	Błąd standardowy reszt	47,68749
Wsp. determ. R-kwadrat	0,390379	Skorygowany R-kwadrat	0,361350
F(2, 42)	13,44765	Wartość p dla testu F	0,000031
Logarytm wiarygodności	−236,2100	Kryt. inform. Akaike'a	478,4200
Kryt. bayes. Schwarza	483,8400	Kryt. Hannana-Quinna	480,4405

Rysunek 6: Model po usunięciu katalizatorów

Wartości p-value są mniejsze niż 5%, więc zmienne są istotne statystycznie. Model ten objaśnia zmienność ceny noclegu w 39%, wartość ta spadła z 48%, co potwierdza przypuszczenie, że zmienna objaśniająca zamek sztucznie zawyżała współczynnik determinacji. Po przeprowadzeniu test normalności reszt Doornika-Hansena przyjmujemy hipotezę zerową, reszty mają rozkład normalny p-value 0.41).



Rysunek 7: Reszty modelu

Zjawisko współliniowości nie występuję:

```
Ocena współliniowości VIF(j) - czynnik rozdęcia wariancji
VIF (Variance Inflation Factors) - minimalna możliwa wartość = 1.0
Wartości > 10.0 mogą wskazywać na problem współliniowości - rozdęcia wariancji

      rating    1,124
powierzchnia    1,124

VIF(j) = 1/(1 - R(j)^2), gdzie R(j) jest współczynnikiem korelacji wielorakiej
pomiędzy zmienną 'j' a pozostałymi zmiennymi niezależnymi modelu.
```

Do ostatecznego modelu dobrane zostały dwie zmienne objaśniające „rating" i „powierzchnia". W tym modelu nie występuje katalizator.

4.2. Test liczby serii- testowanie losowości reszt modelu

Nieparametryczny test służący do sprawdzenia czy wyniki spełniają postulat losowości próby. Serią nazywamy ciąg takich samych elementów uporządkowanych w określony sposób.

$H_0 \sim$ *próba jest losowa*
$H_1 \sim$ *próba nie jest losowa*

Rysunek 8: Hipotezy testu serii

Dla posortowanych reszt przeprowadzono test dla dodatnich i ujemnych serii:

```
Test serii

Liczba serii (R) dla zmiennej 'reszty' = 20
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(23,5, 3,31662),
test z-score = -1,05529, przy dwustronym obszarze krytycznym p = 0,291293
```

Wartośc p-value wynosi 0,29, co oznacza że nie ma podstaw do odrzucenia hipotezy zerowej o losowości reszt w modelu.

4.3. Test Ramseya RESET- testowanie liniowości modelu

Test poprawnej specyfikacji modelu Ramseya RESET jest ogólnym testem, który pozwala zidentyfikować niepoprawną postać funkcyjną modelu ekonometrycznego. TEST RESET oparty jest na regresji rozszerzonej o zbiór zmiennych powstałych z oszacowania zmiennej objaśnianej y w postaci jej potęg. Podobnie jak test liczby serii hipoteza zerowa zakłada liniowość modelu i losowość reszt.

$H_0 \sim$ *próba jest losowa*
$H_1 \sim$ *próba nie jest losowa*

Rysunek 9:Hipotezy testu Ramseya RESET

Przeprowadzono test Ramseya RESET:

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 2,735642,
z wartością p = P(F(2,40) > 2,73564) = 0,077

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 4,363693,
z wartością p = P(F(1,41) > 4,36369) = 0,043

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 4,064121,
z wartością p = P(F(1,41) > 4,06412) = 0,0504
```

Z drugie przypadku można wnioskować , że wybrana postać analityczna nie jest dobrze dobrana. Z tego powodu postawiono sprawdzić jak zachowywał by się model po dodaniu logarytmów i kwadratów zmiennych. Lecz dla nich wartości p-value także były poniżej lub bardzo bliskich wartości 5%.

Przykład :

```
 Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 2,703191,
z wartością p = P(F(2,40) > 2,70319) = 0,0792

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 4,586583,
z wartością p = P(F(1,41) > 4,58658) = 0,0382

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 4,321526,
z wartością p = P(F(1,41) > 4,32153) = 0,0439
```

Dlatego postanowiono pozostać przy wejściowym modelu.

4.4. Test Chowa - testowanie stabilności funkcyjnej modelu

Pozwala na zweryfikowanie czy oszacowane parametry modelu są stabilne, czyli czy w momencie zmiany próbki ich wartości nie zmienią się istotnie. Należy zweryfikować tą cechę, ponieważ chcemy, aby model działał dla całej wybranej populacji a nie jedynie jej skrawka. Dzięki temu możliwe jest dokonywanie prognoz za pomocą modelu jak i poprawna analiza zależności łączących wybrane cechy.

$H_0 \sim$ *parametry stabilne*
$H_1 \sim$ *parametry niestabilne*

Rysunek 10: Hipotezy testu Chowa

W przypadku szeregów czasowych lepsza jest opcja podziału próby do testu dni ileś początkowych i końcowych obserwacji, gdyż sprawdzamy jak na oszacowanie parametrów wpływa czas. Natomiast w naszym przypadku danych przekrojowych bardziej sensowna wydaje się podział danych według zmiennej binarnej- u nas „parking":

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	6,17546	142,423	0,04336	0,9656
rating	18,1554	17,6486	1,029	0,3100
powierzchnia	2,38874	1,44235	1,656	0,1057
parking	−85,2079	161,973	−0,5261	0,6018
pa_rating	6,17217	20,4995	0,3011	0,7649
pa_powierzchnia	0,958998	1,91792	0,5000	0,6199

Średn.aryt.zm.zależnej	208,3778	Odch.stand.zm.zależnej	59,67232
Suma kwadratów reszt	93560,27	Błąd standardowy reszt	48,97940
Wsp. determ. R-kwadrat	0,402837	Skorygowany R-kwadrat	0,326278
F(5, 39)	5,261760	Wartość p dla testu F	0,000870
Logarytm wiarygodności	−235,7455	Kryt. inform. Akaike'a	483,4909
Kryt. bayes. Schwarza	494,3309	Kryt. Hannana-Quinna	487,5319

Test Chowa na strukturalne różnice poziomów ze względu na zmienną: parking
F(3, 39) = 0,271199 z wartością p 0,8458

Zauważamy, że p -value wyższa niż 5%, więc parametry modelu są stabilne.

4.5. Badanie heteroskedastyczności – test White'a

Zjawisko to związane jest bezpośrednio z założeniem stałości wariancji składnika losowego w modelu oszacowanym MNK. Jeżeli wariancja ε nie jest stała, mamy do czynienia z heteroskedastycznością, a oszacowane w ten sposób parametry modelu nie są efektywne, a co za tym idzie nie są też BLUE. Heteroskedastyczność może być skutkiem nieliniowej zależności zmiennych lub niewłaściwie dobranej postaci analitycznej modelu. Brak istotności stałej modelu świadczy o braku liniowej zależności zmiennej objaśnianej od zmiennych objaśniających lub występowania współzależności liniowej zmiennych objaśniających.

W związku z tym, że heteroskedastyczność jest zjawiskiem niepożądanym, w momencie kiedy występuje musimy sobie z nią poradzić. Istnieje wiele sposobów korekty heteroskedastyczności m.in.:

- korekta heteroskedastyczności,
- ważona MNK,
- estymatory robust,
- ponowne próbkowanie/dobór danych

Jednym z testów do wykrywania heteroskedastyczności jest test White'a .

$H_0 \sim$ *reszty modelu są homoskedastyczne,*

$H_1 \sim$ *reszty modelu są heteroskedastyczne*

Rysunek 11:Hipotezy testu White'a

Wynik testu White'a:

```
Test White'a na heteroskedastyczność reszt (zmienność wariancji resztowej)
Estymacja KMNK, wykorzystane obserwacje 1-45
Zmienna zależna (Y): uhat^2

                  współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const            -1902,27        36629,0          -0,05193    0,9588
  rating           -2397,82         9209,57         -0,2604     0,7960
  powierzchnia      1263,58          547,518         2,308      0,0264   **
  sq_rating          265,410         616,682         0,4304     0,6693
  X2_X3             -106,595          67,4589       -1,580      0,1222
  sq_powierzchnia     -5,88437         6,95303      -0,8463     0,4025

  Wsp. determ. R-kwadrat = 0,212961

Statystyka testu: TR^2 = 9,583258,
z wartością p = P(Chi-kwadrat(5) > 9,583258) = 0,087942
```

Wartość p-value wyższa niż przyjęty poziom istotności 5%, daje to podstawy do przyjęcia hipotezy zerowej -nie występuje zatem zjawisko heteroskedastyczności.

4.6. Badanie heteroskedastyczności – test Breuscha-Pagan'a

Innym testem weryfikującym heteroskedastyczność jest test Breuscha-Pagana, którego hipotezy są takie same jak testu White'a.

$H_0 \sim$ *reszty modelu są homoskedastyczne,*

$H_1 \sim$ *reszty modelu są heteroskedastyczne*

Rysunek 12:Hipotezy testu Breuscha-Pagana

Wyniki testu Breuscha-Pagana:

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-45
Zmienna zależna (Y): standaryzowane uhat^2

                współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const           4,15301       1,69350          2,452     0,0184   **
  rating         -0,496662      0,220987        -2,247     0,0299   **
  powierzchnia    0,0326577     0,0238592         1,369     0,1783

  Wyjaśniona suma kwadr. = 8,30898

Statystyka testu: LM = 4,154491,
z wartością p = P(Chi-kwadrat(2) > 4,154491) = 0,125275
```

Potwierdza to homoskedastyczność w modelu, ponieważ wartość p wyższa niż 5%, przyjęto h0.

5. Prognozy

Etapem wieńczącym budowę modelu ekonometrycznego jest praktyczne wykorzystanie modelu, najczęściej w procesie predykcji. Predykcją ekonometryczną nazywamy proces wnioskowania w przyszłość na podstawie modelu ekonometrycznego. Efektem predykcji jest oszacowanie nieznanej wartości zmiennej objaśnianej w okresie prognozowanym, zwane często prognozą

5.1. Prognoza ex-ante

Ponieważ w chwili wyznaczania prognozy nie jest znana wartość rzeczywista zmiennej prognozowanej błąd prognozy ex ante może być tylko oszacowany. Wartość błędu ex ante przynosi informacje o oczekiwanych przeciętnych odchyleniach realizacji zmiennej prognozowanej od prognoz w czasie t>n.

Przeprowadzono prognozę punktową:

```
? scalar me_rating=mean(rating)
Wygenerowano skalar me_rating = 8,07556
? scalar me_powierzchnia=mean(powierzchnia)
Wygenerowano skalar me_powierzchnia = 26,2667
#predykcja punktowa
? scalar y_pred=stala +a1*me_rating+a2*me_powierzchnia
Wygenerowano skalar y_pred = 208,378
```

Rysunek 13: Prognoza punktowa

Wysokość ceny noclegu prognozowana dla takiej obserwacji to 208,4 zł. Sporządzono także prognozę przedziałową, która dała następujący wynik:

```
? scalar L=y_pred-critical(t,$df,0.025)*blad_prognozy
Wygenerowano skalar L = 111,077
? scalar U=y_pred+critical(t,$df,0.025)*blad_prognozy
Wygenerowano skalar U = 305,678
```

Rysunek 14: Prognoza przedziałowa

Oznacza to , że na 95% ta wysokość ceny noclegu znajduję się w przedziale: **(111,1zł ; 305,7zł)**

5.2. Prognoza ex-post

Dokładność prognozy przeprowadzono prognozowanie ex-post, polegające na przeprowadzeniu tego procesu dla znanych już prawdziwych wartości. Podzielono zbiór obserwacji na zbiór uczący i zbiór testowy w proporcjach 90% i 10%. W celu zbadania dokładności prognozy wybrano 2 rodzaje błędów , które niosą takie informacje:

1. Średni absolutny błąd predykcji MAE

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_\tau - y_\tau^P|$$

Wyniósł on w przybliżeniu 39,3zł, co mówi, że prognoza może odbiegać od rzeczywistej wartości o taką wielkość, czyli dla obliczonego wcześniej punktu z wynikiem 208,4zł rzeczywista wartość może znajdować się w przedziale (169,1; 247,7)

2. Średni procentowy błąd prognozy MAPE

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_\tau - y_\tau^P}{y_\tau}|$$

Błąd prognozy MAPE wyniósł około 22%, co oznacza, że taki błąd procentowo popełniany jest przy prognozowaniu za pomocą analizowanego modelu.

7. Weryfikacja hipotez

Po przeprowadzeniu badań możemy zweryfikować prawdziwość hipotez badawczych:

1) Na wysokość ceny noclegu najsilniej wpływa odległość od Zamku Królewskiego. FAŁSZ
2) Wielkość apartamentu nie ma znaczenia na cenę noclegu. FAŁSZ
3) Im lepsza ocena użytkowników na stronie tym droższy nocleg. PRAWDA
4) Im bliżej Zamku Królewskiego tym droższy nocleg. PRAWDA

8. Ostateczny model

$$y = -64{,}19 + 24{,}26x_1 + 2{,}92x_2 + \varepsilon$$

,gdzie:
x_1- ocena użytkowników
x_2- powierzchnia apartamentu

Oznacza to, że wzrost oceny użytkowników na stronie bookin.com o 1 jednostkę ciągnie za sobą zwiększenie ceny noclegu o 24,3 złotego a z każdym m^2powierzchni cena noclegu w Warszawie wzrasta o 3 zł. Zaskakujące jest, że nieistotna okazała się jest dostępność łazienki w pokoju ani bezpłatny parking.

Przeprowadzone badanie ukazuje zależności, jakie zachodzą między czynnikami, które bierzemy pod uwagę przy wyborze miejsca noclegu, a kosztem zakwaterowania.

9. Bibliografia

- Metoda Najmniejszych Kwadratów [dostęp 21.05.2020], http://www.woiz.polsl.pl/~asojda/orig_mod_ekon.pdf
- Mariusz Doszyń, Testowanie Egzogeniczności Zmiennych W Modelach Ekonometrycznych, [dostęp 23.05.2020] http://www.wneiz.pl/nauka_wneiz/sip/sip15-2009/SiP-15-33.pdf
- Jakub Mućk , Prognozowanie ekonometryczne, ocena stabilności oszacowań parametrów strukturalnych [dostęp 21.05.2020], https://web.sgh.waw.pl/~jmuck/Ekonometria/EkonometriaPrezentacja2018Z_4.pdf
- Sebastian Skoczypiec, Jakość prognoz, [dostęp 24.05.2020], https://m6.pk.edu.pl/materialy/mp/MP_02_bledy_prognozy.pdf
- Homoscedastyczność , [dostęp 21.05.2020] , https://www.naukowiec.org/wiedza/statystyka/homoscedastycznosc_533.html
- Barbara Gładysz, Jacek Mercik Koincydencja , „Modelowanie ekonometryczne Studium przypadku", [dostęp 22.05.2020] ,https://dbc.wroc.pl/Content/2182/PDF/Gladysz_modelowanie_ekonometryczne.pdf
- Badanie współczynników determinacji i zbieżności ,[dostęp 22.05.2020] , http://www.ekonometria.4me.pl/prognozowanie1.htm
- Współczynnik współliniowości, [dostęp 22.05.2020] , http://statystycznie-istotne.pl/slownik-statystyczny/vif-wspolczynnik-wspolliniowosci/
- Adam Kopiński, Dariusz Porębski, Metoda Hellwiga ,[dostęp 22.05.2020] https://journals.umcs.pl/h/article/viewFile/121/118
- Marcin Kurpas, „Wybrane zagadnienia ekonometrii z wykorzystaniem programu Statistica", [dostęp 22.05.2020], https://studylibpl.com/doc/651405/rozdzia%C5%82-2-klasyczna-metoda-najmniejszych-kwadrat%C3%B3w
- Barbara Jasiulis-Gołdyn," Specyfikacja i weryfikacja modelu liniowego dobór zmiennych objaśniających część 1" ,[dostęp 24.05.2020] , https://docplayer.pl/65754250-Ekonometria-wykaad-8.html