WSPÓŁCZYNNIK DEZINFORMACJI R^2

dr inż. Piotr Cegielski
Rzeczoznawca Majątkowy Nr 829
Partner w kancelarii iCCS Spółka z o.o.

Streszczenie

Artykuł poświęcony jest analizie współczynnika determinacji R^2 jako miary stopnia dopasowania modelu regresji wielorakiej do obserwowanych cen rynkowych nieruchomości. W artykule skoncentrowano się na wybranych własnościach współczynnika R^2, w szczególności zaś omówiono najczęściej popełniane błędy, do których zaliczyć należy: nieprawidłową interpretację wzoru na współczynnik R^2, wyznaczanie współczynnika R^2 dla modeli liniowych bez wyrazu wolnego oraz niewłaściwą interpretację wartości liczbowej współczynnika R^2. W artykule poddano też krytyce przyjmowanie wartości liczbowej R^2, jako decydującego kryterium wyboru ekonometrycznego modelu wyceny.

Słowa kluczowe

metoda analizy statystycznej rynku, regresja liniowa, modele hedoniczne, współczynnik determinacji R^2

Wstęp

Od pewnego czasu polscy rzeczoznawcy majątkowi, dla potrzeb szacowania wartości rynkowej, coraz częściej wykorzystują metody analizy statystycznej rynku, a w szczególności modele regresji wielu zmiennych (regresji wielorakiej). Jednocześnie widoczny jest w środowisku rzeczoznawców majątkowych podział na zwolenników metod statystycznych oraz na zwolenników tzw. klasycznych metod wyceny, w tym w szczególności metody porównywania parami. Obydwie grupy przerzucają się argumentami wskazując zalety jednej, a wady drugiej grupy metod.

Zwolennicy metod statystycznych twierdzą, że wycena oparta na trzech parach porównawczych obarczona jest dużym potencjalnym błędem oszacowania, a stosowane korekty cen transakcyjnych, często oparte na arbitralnie przyjętych „wagach cech", są raczej dowodem na brak umiejętności stosowania ilościowych metod analizy danych, niż na rzeczywistą „wiedzę ekspercką". Z kolei zwolennicy klasycznych metod porównawczych podnoszą argument, że przeciętny rzeczoznawca majątkowy stosujący metody statystyczne w rzeczywistości nie rozumie tych metod i nie jest w związku z tym w stanie merytorycznie „obronić" wyniku oszacowania; co najwyżej powołując się tajemniczy „model" oraz na obiektywność metod statystycznych.

Argumenty obydwu stron w bardzo dużej części są nietrafne. Po pierwsze żaden przepis nie wymaga, by liczba par porównawczych była ograniczona do trzech; nic też nie stoi na przeszkodzie, by stosowane poprawki kwotowe wynikały z przeprowadzonej wcześniej statystycznej analizy danych, w szczególności przy wykorzystaniu analizy korelacji i regresji. Z kolei trudno cokolwiek zarzucić samej idei wykorzystywania elementów statystyki matematycznej i ekonometrii dla potrzeb analizy rynku oraz procesu wyceny nieruchomości; a wręcz przeciwnie wydaje się, że jest to właściwy kierunek rozwoju metod wyceny.

Niezależnie od powyższych rozważań wydaje się, że postulat, by rzeczoznawca majątkowy **rozumiał stosowane przez siebie metody wyceny**, jest postulatem właściwym. W amerykańskim standardach wyceny (USPAP) wymóg taki jest sformułowany wprost[1] wymaga się, by rzeczoznawca znał, rozumiał oraz prawidłowo stosował uznane metody oraz techniki wyceny niezbędne do uzyskania wiarygodnego wyniku oszacowania. Być może tego typu wymóg znajdzie się również w polskich standardach wyceny.

1. Jak (nie)prawidłowo budować model regresji wielu zmiennych

Stosunkowo często stosowanym w Polsce algorytmem tworzenia i weryfikacji statystycznych modeli wyceny nieruchomości jest ciąg następujących czynności:

1. Zebranie i selekcja danych rynkowych, wraz z opisem podstawowych cech cenotwórczych.
2. Analiza różnych postaci funkcyjnych możliwych modeli regresji liniowej, w szczególności polegająca na porównywaniu każdego z modeli według różnych kryteriów.
3. Wybór „najlepszego" modelu, według przyjętej miary dopasowania modelu do obserwacji rynkowych. Z reguły tą miarą jest współczynnik determinacji R^2 (*determination coefficient*).

[1] USPAP; Standard 1; Standard Rule 1 1: *In developing a real property appraisal, an appraiser must be aware of, understand and correctly employ those recognized methods and techniques that are necessary to produce a credible appraisal [...].*

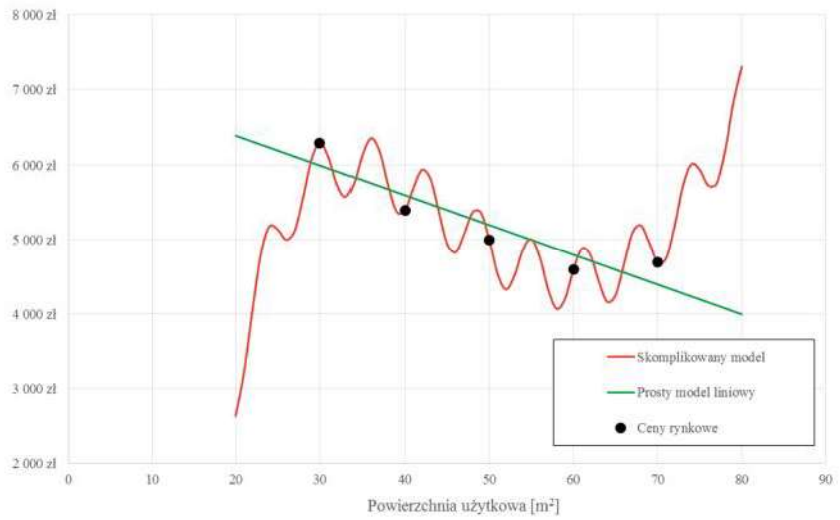
Wydawać by się mogło, że jest to procedura prawidłowa tak jednak nie jest. Dobrym modelem wyceny nieruchomości nie jest ten, który dobrze „dopasowuje się" do obserwacji, lecz ten model, który prawidłowo szacuje (estymuje) ceny. Przykład wyobraźmy sobie dane zobrazowane na Rysunku 1.

Mamy tutaj do czynienia z rynkowymi cenami jednostkowymi pięciu lokali mieszkalnych, różniących się powierzchnią użytkową. Wykres przedstawia też dwa modele opisujące tę zależność; prosty model liniowy, oraz bardziej złożony model charakteryzujący się m.in. tym, że uzyskane przy jego zastosowaniu oszacowania w pełni pokrywają się z obserwowanymi cenami rynkowymi. Wyniki uzyskane przy zastosowaniu modelu wielomianowego w 100% pokrywają się z obserwowanymi cenami rynkowymi, natomiast współczynnik determinacji R^2 dla modelu liniowego wynosi „zaledwie" 0,84. Tym niemniej jest oczywiste, iż z punktu widzenia szacowania wartości rynkowej zdecydowanie lepszy jest prosty model liniowy (a już zwłaszcza w przypadku wyceny lokali o powierzchniach spoza zakresu obserwacji, czyli mieszkań o powierzchni poniżej 30 m^2 lub powyżej 70 m^2).

Wydaje się, że nie sposób popełnić błędu polegającego na przyjęciu modelu o błędnej postaci funkcyjnej. Niestety, ten błąd jest popełniany częściej, niż by się to mogło wydawać. Przyczyn jest kilka; do najważniejszych zaliczyć należy:

- Duża liczba przyjętych cech porównawczych (zmiennych objaśniających), często ze sobą istotnie skorelowanych, lub też pozostających ze sobą w interakcji, przy jednoczesnym braku możliwości graficznego przedstawienia i analizy modelowanej zależności.
- Preferowanie podejścia statystycznego (zwłaszcza podejścia typu *data mining*), zamiast podejścia ekonometrycznego, we wstępnym doborze możliwych do przyjęcia zależności funkcyjnych.
- Wybór postaci funkcyjnych nieposiadających, z punktu widzenia modelowanej zależności, wygodnej interpretacji, umożliwiającej ocenę poprawności parametrów modelu.
- Ostateczny wybór modelu wyceny w oparciu o błędnie przyjęte kryterium (np. tylko i wyłącznie według kryterium maksymalnej wartości współczynnika determinacji R^2).

Rysunek 1
Porównanie dwóch modeli ceny rynkowej



Źródło: opracowanie własne.

Niniejszy artykuł koncentruje się wokół czwartej z wyżej wymienionych przyczyn, w szczególności na omówieniu zasadności stosowania współczynnika R^2, jako kryterium wyboru modelu wyceny. Należy przy tym wyraźnie zaznaczyć, że celem niniejszego artykułu nie jest krytyka tego kryterium jako taka, lecz wskazanie pułapek czyhających na rzeczoznawców, którzy bezkrytycznie zawierzą współczynnikowi R^2, jako mierze „jakości" modelu wyceny nieruchomości.

2. Czym jest współczynnik determinacji R^2 – fakty i mity

Spośród różnych miar jakości regresyjnych modeli wyceny najbardziej popularną wydaje się być współczynnik determinacji R^2. Jego właściwości oraz interpretacja wydają się być oczywiste:

a) współczynnik R^2 przyjmuje wartości liczbowe z przedziału od 0 do 1 (od 0 do 100%);
b) wartość liczbowa współczynnika R^2 przedstawia, jaki procent zmienności obserwowanych cen rynkowych wyjaśnia dany model;
c) spośród różnych modeli wyceny należy wybrać ten, dla którego wartość współczynnika determinacji R^2 jest najwyższa.

Wszystko powyższe wydaje się tak czytelne, proste, wygodne w interpretacji oraz w praktycznym wykorzystywaniu. Niestety; wszystko to nie jest prawdą (lub łagodniej rzecz ujmując nie do końca jest prawdą), a mianowicie:

d) współczynnik R^2 **może** przyjmować wartości mniejsze od zera lub większe od jedności;
e) wartość liczbowa współczynnika R^2 **nie przedstawia**, wyrażonego w ujęciu procentowym, stopnia „wytłumaczenia" przez model zróżnicowania obserwowanych cen rynkowych;
f) poziom wartości liczbowej współczynnika R^2 **nie jest** najwłaściwszym kryterium wyboru modelu wyceny.

Ponieważ powyższe stwierdzenia (od (d) do (f)) mogą wydawać się **herezją**, najpierw wyjaśnijmy sobie, czym w istocie jest współczynnik determinacji R^2.

Jedną z metod oceny stopnia dopasowania danego modelu do obserwacji jest ocena relatywna, polegająca na porównaniu poziomu błędu związanego z analizowanym modelem oraz błędu związanego z tzw. **modelem referencyjnym** (inaczej: modelem wzorcowym), stanowiącym punkt odniesienia. Przez błąd rozumie się tutaj niedopasowanie oszacowań modelu do obserwacji rynkowych, mierzone kwadratem różnic. Innymi słowy, ocena analizowanego modelu następuje poprzez porównanie z innym, referencyjnym modelem. Jak łatwo się domyślić, możliwe są zarówno sytuacje, w których analizowany model jest lepszy od modelu referencyjnego, jak również sytuacje, w których jest od niego gorszy.

W przypadku klasycznego współczynnika determinacji R^2 modelem referencyjnym jest **model wartości średniej**. Model wartości średniej jest to model, który każdej zmiennej niezależnej x_i przypisuje wartość y_i równą wartości średniej $\bar{y}$.

Przykład nr 1

Wyobraźmy sobie, że na rynku są następujące dane dotyczące cen jednostkowych mieszkań jak w Tabeli 1.

Przyjmijmy, że powierzchnia użytkowa jest jedyną cechą różnicującą te mieszkania. Przyjmijmy też, że mamy dwóch rzeczoznawców majątkowych. Pierwszy rzeczoznawca doszedł do wniosku, że wyznaczy średnią jednostkową cenę rynkową i każde z mieszkań będzie wyceniał w ten sposób, iż będzie mnożył cenę średnią przez powierzchnię użytkową. Innymi słowy przyjął do wyceny **model średniej ceny jednostkowej**, który nie uwzględnia ewentualnej zależności pomiędzy jednostkową ceną rynkową a powierzchnią użytkową mieszkania. Wyniki oszacowań uzyskane przy zastosowaniu tego modelu oraz wielkość różnic pomiędzy rzeczywistymi cenami rynkowymi, a wynikami oszacowań są przedstawione w Tabeli 2.

Sumę kwadratów różnic w przypadku tego modelu oznaczymy symbolem SST:

$$SST = \sum (y_i - \bar{y})^2 = 1\,900\,000 \quad .$$

Drugi rzeczoznawca, po przeanalizowaniu danych rynkowych doszedł do wniosku, że możliwa jest zależność pomiędzy powierzchnią użytkową a ceną jednostkową, i że zależność tę można modelować przy użyciu funkcji liniowej postaci:

$$y = a + b \cdot x$$

gdzie:

- y — cena jednostkowa lokalu mieszkalnego;
- x — powierzchnia użytkowa lokalu mieszkalnego;
- a — wyraz wolny równania (hipotetyczna cena jednostkowa mieszkania o zerowej powierzchni użytkowej);
- b — współczynnik kierunkowy linii regresji (zmiana ceny jednostkowej na skutek jednostkowego wzrostu powierzchni użytkowej).

Tabela 1

Lp.	Pow. użytkowa x_i	Cena jednostkowa y_i
1	30,00	6 300 zł
2	40,00	5 400 zł
3	50,00	5 000 zł
4	60,00	4 600 zł
5	70,00	4 700 zł
Średnia	**50,00**	**5 200 zł**

Źródło: opracowanie własne.

A następnie wyznaczył parametry funkcji liniowej stosując tzw. metodę najmniejszych kwadratów (dalej: MNK):

a = 7 200;

b = 400;

uzyskując model, zgodnie z którym liniowe tempo spadku cen jednostkowych mieszkań na skutek jednostkowego wzrostu powierzchni użytkowej wynosi 40 złotych, zaś punkt przecięcia prostej z osią Y jest na poziomie 7 200 złotych. W oparciu o ten model rzeczoznawca dokonał oszacowania ceny rynkowej każdego z lokali, uzyskując wyniki przedstawione w Tabeli 3.

Tabela 2

Lp.	Pow. użytkowa x_i	Cena jednostkowa y_i	Oszacowanie $y_{i(model)}$	Różnica $(y_i - y_{i(model)})$	Kwadrat różnic $(y_i - y_{i(model)})^2$
1	30,00	6 300 zł	5 200 zł	1 100 zł	1 210 000
2	40,00	5 400 zł	5 200 zł	200 zł	40 000
3	50,00	5 000 zł	5 200 zł	-200 zł	40 000
4	60,00	4 600 zł	5 200 zł	-600 zł	360 000
5	70,00	4 700 zł	5 200 zł	-500 zł	250 000
Średnia	**50,00**	**5 200 zł**		**Suma**	**1 900 000**

Źródło: opracowanie własne.

Tabela 3

Lp.	Pow. użytkowa x_i	Cena jednostkowa y_i	Oszacowanie $y_{i(model)}$	Różnica $(y_i - y_{i(model)})$	Kwadrat różnic $(y_i - y_{i(model)})^2$
1	30,00	6 300 zł	6 000 zł	300 zł	90 000
2	40,00	5 400 zł	5 600 zł	-200 zł	40 000
3	50,00	5 000 zł	5 200 zł	-200 zł	40 000
4	60,00	4 600 zł	4 800 zł	-200 zł	40 000
5	70,00	4 700 zł	4 400 zł	300 zł	90 000
Średnia	**50,00**	**5 200 zł**		**Suma**	**300 000**

Źródło: opracowanie własne.

W przypadku tego modelu sumę kwadratów różnic oznaczymy symbolem SSE:

$$SSE = \sum(\hat{y}_i - \bar{y})^2 = 300\,000 \quad .$$

Ocena modelu liniowego następuje poprzez podzielenie sumy kwadratów różnic obliczonych dla tego modelu SSE przez sumę kwadratów różnic modelu wartości średniej SST, który pełni tutaj rolę omówionego wcześniej modelu referencyjnego:

$$\varphi^2 = \frac{SSE}{SST} \qquad (1).$$

Powyższy iloraz nazywany jest **współczynnikiem zbieżności** i oznaczany jest symbolem φ^2. W literaturze można przeczytać, że współczynnik ten informuje o tym, jaka część zmienności zmiennej objaśnianej (tutaj: jednostkowej ceny rynkowej) **nie jest wyjaśniona przez model**.

W naszym przypadku wartość liczbowa współczynnika zbieżności wynosi:

$$\varphi^2 = \frac{SSE}{SST} = \frac{300\,000}{1\,900\,000} = 0{,}1579$$

i oznacza po prostu, że suma kwadratów różnic SSE jest równa około 15,8% sumy kwadratów różnic SST. Tylko tyle, lub aż tyle.

Współczynnik determinacji R^2 jest dopełnieniem współczynnika zbieżności do jedności i wyraża się wzorem:

$$R^2 = 1 - \varphi^2 = 1 - \frac{SSE}{SST} \qquad (2)$$

i dla analizowanego przykładu wynosi 0,8421. Znowu, w literaturze możemy przeczytać, że wartość liczbowa współczynnika R^2 informuje o tym, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez model (tutaj: około 84,2%).

Wróćmy ponownie do przykładu i przeanalizujmy wyniki przedstawione w postaci graficznej (Rysunek 2)

Na Rysunku 2 zaznaczono ceny rynkowe, poziom ceny średniej oraz przebieg wyznaczonej linii regresji. Rozpatrzmy wartości dla mieszkania o powierzchni 30 m^2, którego jednostkowa cena rynkowa wynosi 6 300 zł, zaś wynik oszacowania jest równy 6 000 zł.

Dla tej obserwacji wyznaczono kwadraty różnic pomiędzy ceną rynkową a ceną średnią, między ceną rynkową a ceną z modelu oraz pomiędzy ceną z modelu a ceną średnią:

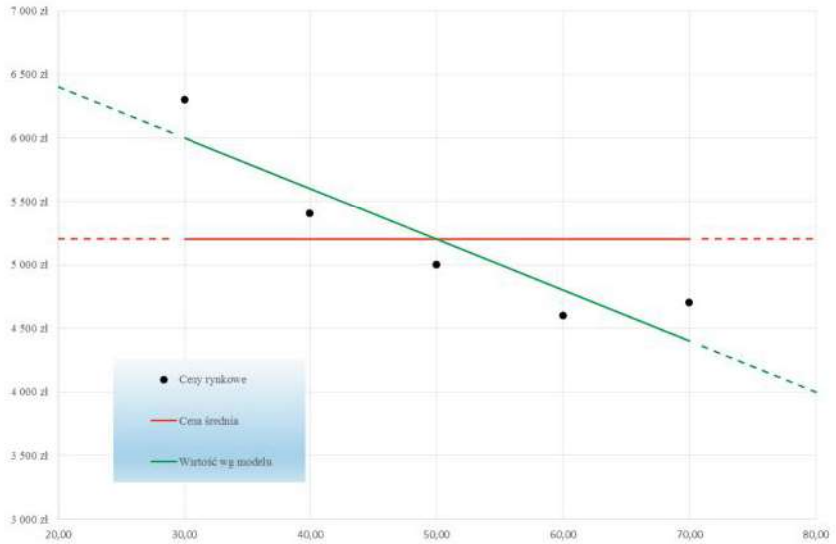
$$(y_1 - \bar{y})^2 = (6300 - 5200)^2 = 1\,210\,000 \;;$$

$$(y_1 - \hat{y}_1)^2 = (6300 - 6000)^2 = 90\,000 \;;$$

$$(\hat{y}_1 - \bar{y})^2 = (6000 - 5200)^2 = 640\,000 \;.$$

Rysunek 2

Ceny rynkowe, cena średnia oraz przebieg funkcji regresji liniowej dla przykładu nr 1



Źródło: opracowanie własne.

Postępując w ten sposób dla wszystkich obserwacji uzyskuje się następujące wyniki:

$$SST = \sum (y_i - \bar{y})^2 = 1\,900\,000 \quad ;$$

$$SSE = \sum (\hat{y}_i - \bar{y})^2 = 300\,000 \quad ;$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = 1\,600\,000 \quad .$$

Powyżej pojawiła się jeszcze trzecia suma oznaczona symbolem *SSR*, która jest równa sumie kwadratów różnic pomiędzy wynikami oszacowań według modelu, a wartością średnią zmiennej zależnej.

Dzieląc SSR przez SST również otrzymuje się wartość współczynnika determinacji:

$$R^2 = \frac{SSR}{SST} \qquad (3).$$

Jest tak dlatego, gdyż w sytuacji, gdy funkcja regresji liniowej jest wyznaczona przy użyciu metody najmniejszych kwadratów, spełniona jest równość[2]:

$$SSR + SSE = SST \qquad (4).$$

W naszym przypadku oczywiście otrzymujemy wynik równy:

$$R^2 = \frac{SSR}{SST} = \frac{1\,600\,000}{1\,900\,000} = 0{,}8421 \quad .$$

Reasumując współczynnik determinacji R^2 jest relatywną miarą stopnia dopasowania modelu do obserwacji, względem referencyjnego modelu wartości średniej.

[2] Dowód na to można znaleźć w wielu podręcznikach z zakresu statystyki matematycznej.

3. Dlaczego współczynnik R^2 może przyjmować wartości spoza przedziału od 0 do 1

Jak już zostało to zasygnalizowane na początku artykułu, wartość współczynnika determinacji R^2 może, w określonych sytuacjach, przyjmować wartości liczbowe poniżej zera lub powyżej jedności. Przeanalizujmy kolejny przykład liczbowy.

Przykład nr 2 (dane obliczeniowe z przykładu 1)

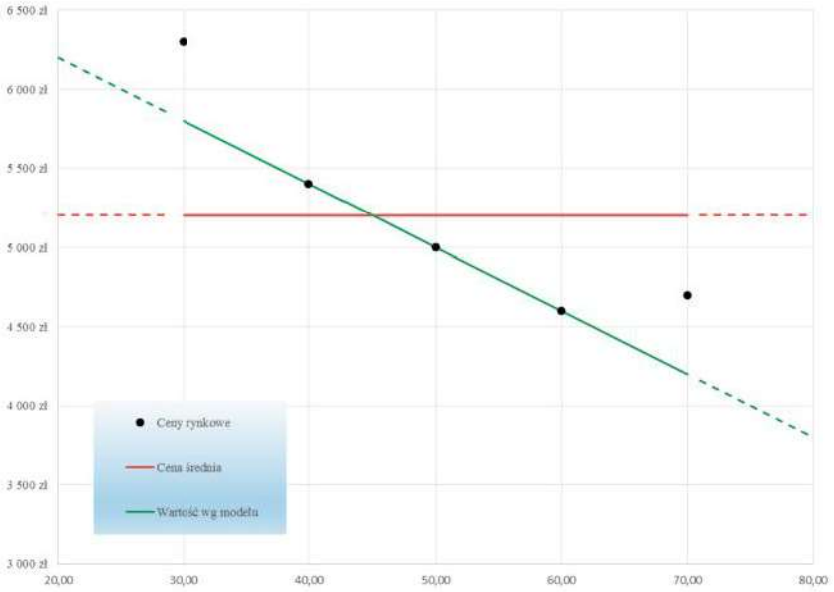
Dla danych przedstawionych w przykładzie nr 1 kolejny rzeczoznawca majątkowy, z sobie tylko wiadomych względów przyjął, iż oszacowania linii regresji dokona tylko w oparciu o ceny lokali o powierzchni 40, 50 oraz 60 m^2. W oparciu o te dane wyznaczył funkcję liniową postaci:

$$y = a + b \cdot x = 7000 - 40 \cdot x \quad .$$

Jak widzimy (zobacz Rysunek 3), przy zastosowaniu tego modelu wyniki oszacowań dokładnie pokrywają się z cenami rynkowymi dla mieszkań o powierzchni 40, 50 i 60 m^2.

Rysunek 3

Ceny rynkowe, cena średnia oraz przebieg funkcji regresji liniowej dla przykładu nr 2



Źródło: opracowanie własne.

Tabela 4

Lp.	x_i	y_i	$y_{i(model)}$	SST	SSE	SSR
1	30,00	6 300 zł	5 800 zł	1 210 000	250 000	360 000
2	40,00	5 400 zł	5 400 zł	40 000	0	40 000
3	50,00	5 000 zł	5 000 zł	40 000	0	40 000
4	60,00	4 600 zł	4 600 zł	360 000	0	360 000
5	70,00	4 700 zł	4 200 zł	250 000	250 000	1 000 000
Średnia	**50,00**	**5 200 zł**	**Suma**	**1 900 000**	**500 000**	**1 800 000**

Źródło: opracowanie własne.

Jednocześnie model ten, dla mieszkań o powierzchni 20 oraz 70 m^2, zwraca wartości istotnie niższe od cen rynkowych. Porównując ten model z referencyjnym modelem ceny średniej otrzymujemy wartości przedstawione w Tabeli 4.

Stosując wzór (2) na współczynnik determinacji otrzymujemy wartość:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{500000}{1900000} = 0{,}7368\ ;$$

ale stosując wzór (3) otrzymujemy wartość:

$$R^2 = \frac{SSR}{SST} = \frac{1800000}{1900000} = 0{,}9473\ .$$

Jak widać, w zależności od tego, jaki został zastosowany wzór na R^2, uzyskaliśmy różne wartości współczynnika determinacji. Stało się tak, gdyż zależność (4) jest prawdziwa tylko w przypadku, gdy funkcja regresji jest wyznaczona metoda najmniejszych kwadratów (MNK). W naszym przykładzie rzeczoznawca zastosował inną procedurę doboru funkcji regresji, a w związku z tym zależność (4) nie jest spełniona, co skutkuje różnymi wartościami liczbowymi współczynnika R^2.

Kontynuując, już czysto hipotetycznie, nasze rozważania zobaczmy, co się stanie, jeśli analizowany model będzie „gorszy" od modelu ceny średniej. Przez gorszy rozumiemy, że suma kwadraty różnic pomiędzy zaobserwowanymi cenami rynkowymi a oszacowaniami modelu będzie większa od sumy kwadratów różnic pomiędzy cenami rynkowymi a ceną średnią (ujmując rzecz inaczej: analizowany model będzie gorzej szacowała wartości, niż model polegający na przypisaniu każdemu z mieszkań, bez względu na jego powierzchnię, wartości równej średniej cenie jednostkowej).

Przykład nr 3

Przyjmijmy, że analizowany model ma następującą postać:

$$y = a + b \cdot x = 3500 + 40 \cdot x\ .$$

Zgodnie z tym modelem ceny jednostkowe rosną wraz ze wzrostem powierzchni. Dopasowanie modelu do cen rynkowych przedstawia Rysunek 4.

Wartość współczynnika determinacji R^2 obliczamy, zgodnie z przyjętymi wzorami, w sposób przedstawiony w Tabeli 5.

Stosując wzór (2) otrzymujemy wartość:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{7150000}{1900000} = -2{,}7632\ ;$$

ale stosując wzór (3) na współczynnik determinacji otrzymujemy wartość:

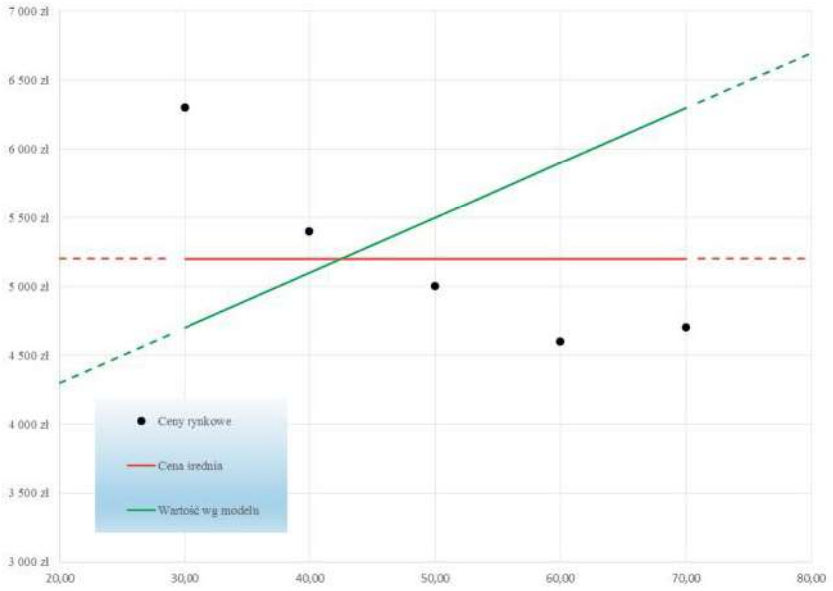
$$R^2 = \frac{SSR}{SST} = \frac{2050000}{1900000} = 1{,}0789\ .$$

Jak widać, w zależności od zastosowanego wzoru otrzymaliśmy niejako jednocześnie wartość współczynnika determinacji poniżej zera oraz powyżej jedności. (!!!)

Oznacza to, że współczynnik determinacji R^2 powinien stosowany tylko w przypadku analizy funkcji regresji liniowej, której parametry zostały wyznaczone przy użyciu MNK. Natomiast nie powinien być stosowany w przypadku analizy funkcji nieliniowych (oczywiste), jak również funkcji liniowych, których parametry zostały wyznaczone przy użyciu metody innej, niż MNK.

Rysunek 4

Ceny rynkowe, cena średnia oraz przebieg funkcji regresji liniowej dla przykładu nr 3



Źródło: opracowanie własne.

Tabela 5

Lp.	x_i	y_i	$y_{i(model)}$	SST	SSE	SSR
1	30,00	6 300 zł	4 700 zł	1 210 000	2 560 000	250 000
2	40,00	5 400 zł	5 100 zł	40 000	90 000	10 000
3	50,00	5 000 zł	5 500 zł	40 000	250 000	90 000
4	60,00	4 600 zł	5 900 zł	360 000	1 690 000	490 000
5	70,00	4 700 zł	6 300 zł	250 000	2 560 000	1 210 000
Średnia	**50,00**	**5 200 zł**	**Suma**	**1 900 000**	**7 150 000**	**2 050 000**

Źródło: opracowanie własne.

4. Wartość współczynnika R^2 w analizie funkcji regresji bez wyrazu wolnego

Wydawać by się mogło, że powyższe rozważania mają charakter czysto teoretyczny, gdyż z reguły stosuje się pewne pakiety statystyczne, które, niejako automatycznie, wyznaczają parametry funkcji liniowej przy użyciu MNK, a co za tym idzie równanie (4) zawsze jest spełnione. Można by tak było założyć, gdyby nie pewien szczególny rodzaj funkcji regresji liniowej, a mianowicie funkcja liniowa bez wyrazu wolnego.

Przykład nr 4

Przyjmijmy, że z sobie tylko wiadomych przyczyn, kolejny rzeczoznawca, wyznaczając funkcję regresji liniowej dla danych z przykładu nr 1 wybrał opcję „Stała wynosi Zero”[3] (zobacz Rysunek 5).

W wyniku wykorzystania modułu *„Regresja”* w pakiecie Excel uzyskujemy funkcję regresji liniowej bez wyrazu wolnego postaci:

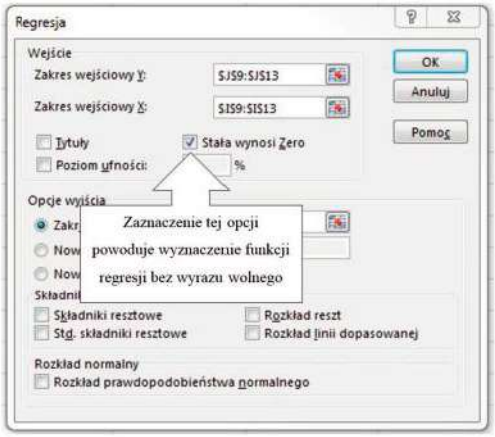
$$y = a + b \cdot x = 0 + 93{,}3(3) \cdot x \quad .$$

Jednocześnie arkusz Excel podaje wartość współczynnika R^2 równą 0,8578, podczas gdy analiza tej zależności prowadzi do oczywistego wniosku, że model jest bardzo źle dopasowany do obserwacji (zobacz Rysunek 6).

Przeprowadzone obliczenia dają wartości przedstawione w Tabeli 6

[3] Oznacza przyjęcie modelu bez wyrazu wolnego, a co za tym idzie linia regresji będzie przechodzić przez początek układu współrzędnych.

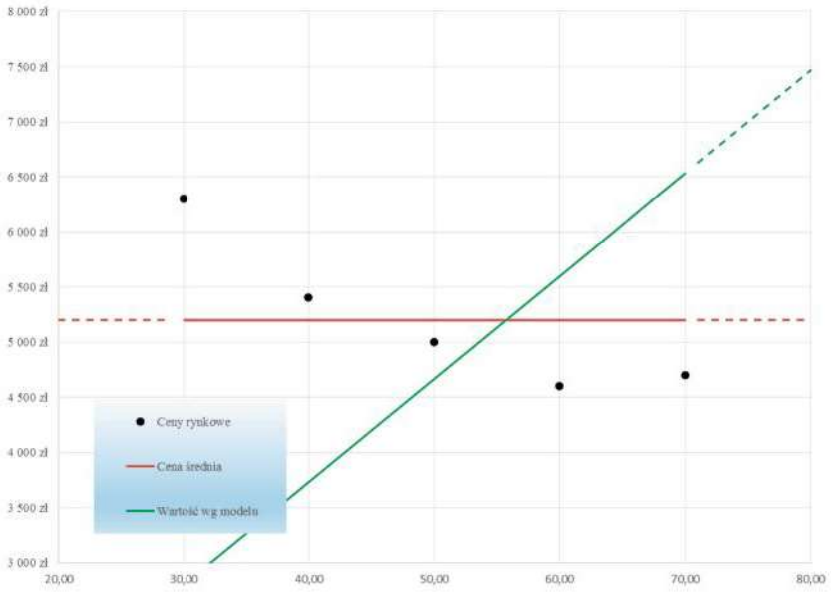
Rysunek 5



Źródło: opracowanie własne.

Rysunek 6

Ceny rynkowe, cena średnia oraz przebieg funkcji regresji liniowej dla przykładu nr 4



Źródło: opracowanie własne.

Tabela 6

Lp.	x_i	y_i	$y_{i(model)}$	SST	SSE	SSR
1	30,00	6 300 zł	2 800 zł	1 210 000	12 250 000	5 760 000
2	40,00	5 400 zł	3 733 zł	40 000	2 777 778	2 151 111
3	50,00	5 000 zł	4 667 zł	40 000	111 111	284 444
4	60,00	4 600 zł	5 600 zł	360 000	1 000 000	160 000
5	70,00	4 700 zł	6 533 zł	250 000	3 361 111	1 777 778
Średnia	**50,00**	**5 200 zł**	**Suma**	**1 900 000**	**19 500 000**	**10 133 333**

Źródło: opracowanie własne.

z których wynika, że wartość współczynnika determinacji, w zależności od zastosowanego wzoru, wynosi:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{19\,500\,000}{1\,900\,000} = -9{,}2632\,;$$

albo:

$$R^2 = \frac{SSR}{SST} = \frac{10\,133\,333}{1\,900\,000} = 5{,}3333 \quad .$$

Dlaczego tak się stało? Otóż w przypadku modeli liniowych bez wyrazu wolnego modelem referencyjnym nie powinien być model ceny średniej[4].

Jest to bardzo ważne oprócz sytuacji, gdzie brak wyrazu wolnego wynika wprost z charakteru modelowanego zjawiska, nie powinno się wyznaczać funkcji regresji bez wyrazu wolnego, gdyż wartość współczynnika R^2 wprowadzi nas w błąd. Natomiast w sytuacji, gdy modelowane zjawisko z natury rzeczy wymaga funkcji regresji bez wyrazu wolnego, ocena stopnia dopasowania modelu do obserwacji powinna być dokonana przy użyciu innej miary, niż współczynnik determinacji R^2 w jego klasycznej postaci.

Zobaczmy jeszcze, żeby uzmysłowić sobie skalę możliwego do popełnienia błędu, co się stanie, gdy do modelowania jakiejś zależności zastosujemy funkcję liniową bez wyrazu wolnego, a miarą stopnia dopasowania modelu będzie klasyczny współczynnik R^2.

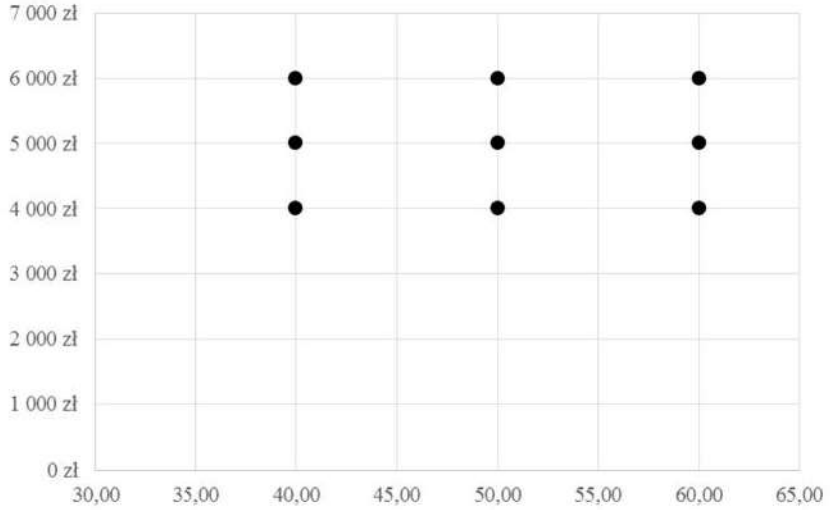
Tabela 7

Pow. użytkowa	Cena nr 1	Cena nr 2	Cena nr 3
40,00	4 000 zł	5 000 zł	6 000 zł
50,00	4 000 zł	5 000 zł	6 000 zł
60,00	4 000 zł	5 000 zł	6 000 zł

Źródło: opracowanie własne.

Rysunek 7

Cena jednostkowa mieszkań o różnej powierzchni



Źródło: opracowanie własne.

Przykład nr 5

Przyjmijmy, że dla mieszkań o powierzchni użytkowej 40, 50 oraz 60 m^2 stwierdzono ceny rynkowe jak w Tabeli 7.

Jak łatwo zauważyć brak jest tutaj jakiejkolwiek zależności pomiędzy powierzchnią użytkową, a ceną jednostkową (zobacz Rysunek 7).

Stosując dla powyższych danych moduł *„Regresja”* z pakietu Excel i nie zaznaczając opcji *„Stała wynosi Zero”* uzyskamy na wyjściu model wartości średniej, czyli model postaci:

$$y = 5000 \quad ;$$

przy czym wartość współczynnika R^2 będzie równa zero. Taki wynik jest zgodny z oczekiwaniami. Jeśli natomiast zaznaczymy opcję *„Stała wynosi Zero”* to uzyskamy[5] model postaci:

$$y = 97{,}40 \cdot x \quad ;$$

przy współczynniku determinacji R^2 równym 0,9487 (!!!).

5. Dlaczego współczynnik R^2 nie przedstawia, wyrażonego w ujęciu procentowym, stopnia dopasowania modelu do danych

Przyjrzyjmy się jeszcze raz wartościom liczbowym z przykładu nr 1 ceny rynkowe różnią się od ceny średniej odpowiednio o 1 100, 200, 200, 500 oraz 600 zł (patrz Tabela 1). Jednocześnie ceny rynkowe różnią się od oszacowań uzyskanych przy wykorzystaniu liniowego modelu regresji odpowiednio o 300, 200, 200, 200 oraz 300 zł (patrz Tabela 2). Analizując tak zestawione wartości liczbowe trudno się zgodzić z tym, że model wyjaśnił około 84% zmienności.

Problem polega na tym, że wyznaczając współczynnik R^2 porównujemy **sumy kwadratów różnic**, podczas gdy z „naszego” punktu widzenia[6] bardziej wartościową miarą byłyby wielkości wyrażone wprost w analizowanych jednostkach (tutaj w złotych). Dlatego lepszą miarą opisującą, jaka część zmienności zmiennej zależnej jest wytłumaczona modelem, byłaby następująca miara:

$$1 - \sqrt{1 - R^2} = 1 - \sqrt{1 - \frac{SSR}{SST}} = 1 - \sqrt{\frac{SSE}{SST}} \quad (5).$$

Można to wytłumaczyć w ten sposób wróćmy do Rysunku 1 i przeanalizujmy jeszcze raz dane dla mieszkania o powierzchni użytkowej 30 m^2. Dla tego mieszkania różnica pomiędzy rzeczywistą ceną rynkową, a ceną średnią wynosi 6 300 5 200 = 1 100 zł, co

[4] Szerzej: Robert Bartels (University of Sydney Business School): *Re-interpreting R-squared, regression through the origin, and weighted least squares*. November 2015.

[5] Sprawdzenie obliczeń pozostawiamy Czytelnikowi.

[6] Przez „nasz punkt widzenia” rozumiemy tutaj intuicyjną interpretację sformułowania „procent wytłumaczonej zmienności”.

podniesione do kwadratu daje liczbę równą 1 210 000. Jednocześnie różnica pomiędzy rzeczywistą ceną rynkową, a ceną wynikającą z modelu regresji jest równą 6 300 6 000 = 300 zł, co podniesione do kwadratu daje liczbę równą 90 000. Nas bardziej interesuje iloraz:

$$\frac{300}{1100} = 0{,}273 \quad ,$$

który można w uproszczeniu interpretować w ten sposób, że model wytłumaczył około 72,7% procent różnicy pomiędzy ceną rynkową mieszkania o powierzchni 30 m^2, a ceną średnią, zaś 27,3% tej różnicy pozostało niewytłumaczone.

Natomiast, wyznaczając poziom współczynnika determinacji R^2, pracujemy na **kwadratach różnic**, czyli w tym przypadku dokonujemy (w pewnym uproszczeniu) następującego porównania:

$$\frac{300^2}{1100^2} = 0{,}074 \quad .$$

Interpretacja powyższego wyniku jest następująca model „tłumaczy" 92,6% procent **kwadratu** różnicy pomiędzy ceną rynkową mieszkania o powierzchni 30 m^2 a ceną średnią, zaś 7,4% **kwadratu** tej różnicy pozostało niewytłumaczone.

Co to oznacza? Z jednej strony rzeczywiście, jeśli wartość współczynnika R^2 wynosi zero to oznacza, że analizowany model nic nie wniósł w próbę wyjaśnienia obserwowanej zmienności[7], zaś w przypadku, gdy jest on równy 1 oznacza, że 100% zmienności jest wytłumaczone modelem. Z drugiej strony taki sam efekt uzyskamy stosując miarę opisaną wzorem (5). Zysk zaś jest taki, że będziemy operować na wielkościach liczbowych wyrażonych w interpretowalnych jednostkach (tutaj: w złotych). Ujmując rzecz jeszcze trochę inaczej, jeśli potocznie rozumiemy pojęcie *„procent wytłumaczonej zmienności"*, jako przeciętną różnicę wartości wynikających z modelu od wartości średniej, to współczynnik R^2 zawyża tak rozumianą miarę i powinniśmy stosować wzór (5). Jeżeli natomiast jesteśmy w pełni świadomi, że w przypadku współczynnika R^2 pomiar dotyczy kwadratów różnic, to możemy poprzestać na samym współczynniku determinacji.

W Tabeli 8, żeby zobrazować różnicę pomiędzy obydwiema miarami stopnia dopasowania modelu do danych, przedstawiamy zestawienie wartości.

Tabela 8

R^2	$1 - \sqrt{1 - R^2}$
0,99	0,90
0,95	0,78
0,90	0,68
0,85	0,61
0,80	0,55
0,75	0,50
0,50	0,29

Źródło: opracowanie własne.

6. Dlaczego porównywanie współczynników R^2 może skutkować wyborem złego modelu

Kolejnym mitem jest, że spośród różnych modeli wybrać należy ten, dla którego wartość liczbowa współczynnika determinacji R^2 jest najwyższa. Otóż nie jest to do końca prawdą; istotne jest bowiem, w jakim celu budujemy dany model. Jeżeli model ma służyć szacowaniu wartości nieruchomości, to podejście bazujące tylko i wyłącznie na kryterium współczynnika R^2 może być błędne. Posłużmy się kolejnym przykładem.

Przykład nr 6

Rozpatrzmy działania dwóch rzeczoznawców, którzy podjęli się dokonać wyceny mieszkań na jednym z wrocławskich osiedli, zabudowanego 11-kondygnacyjnymi budynkami wykonanymi w technologii wielkopłytowej:

a) Pierwszy rzeczoznawca (a) dokonał analizy całego wrocławskiego rynku nieruchomości. Przeanalizował ogółem 10 tys. transakcji i ustalił, że ceny zawierają się w przedziale od 3 do 20 tys. zł (cena średnia równa 6 000 zł, odchylenie standardowe 2 000 zł). Na podstawie tych danych zbudował model charakteryzujący się współczynnikiem $R^2 = 0{,}90$.

b) Drugi rzeczoznawca (b) dokonał analizy cen wyłącznie w 11-kondygnacyjnych budynkach położonych na terenie analizowanego osiedla. Łączna liczba transakcji wyniosła 50. Ceny jednostkowe mieszkań zawierały się w granicach od 4 do 6 tys. zł, przy cenie średniej 5 000 zł i odchyleniu standardowym 500 zł. W oparciu o te dane zbudował model regresji dwóch zmiennych objaśniających (powierzchnia użytkowa mieszkania i położenie na kondygnacji), dla którego wartość współczynnika R^2 wyniosła „zaledwie" 0,10.

Pytanie brzmi, który z modeli powinno się stosować do szacowania wartości rynkowej lokali na tym konkretnym osiedlu?

Z punktu widzenia sposobu wykorzystywania ekonometrycznego modelu wyceny nieruchomości alternatywną miarą dopasowania wyników oszacowań do obserwacji rynkowych jest tzw. ***standardowy błąd estymacji***, który wyraża się wzorem:

$$\sqrt{\frac{SSE}{N-2}} \qquad (6),$$

gdzie N jest to liczba obserwacji rynkowych, a występująca w mianowniku różnica *(N 2)* jest korektą liczby obserwacji ze względu na tzw. **liczbę stopni swobody**. Warto przy tym zaznaczyć, że w przypadku stosunkowo dużej liczby obserwacji N, zbliżone wartości błędu standardowego estymacji uzyska się również przy zastosowaniu następującej zależności[8]:

$$\sqrt{\frac{SSE}{N}} \qquad (7).$$

[7] Względem referencyjnego modelu wartości średniej.

[8] Ze względu na to, że celem niniejszego artykułu nie jest tłumaczenie pewnych założeń statystyki matematycznej, w tym tzw. liczby stopni swobody, w dalszej części artykułu będziemy używać wzoru (7), co będzie bardziej czytelne dla odbiorców artykułu.

Interpretacja tej miary jest następująca standardowy błąd estymacji pokazuje, o ile, w przybliżeniu, wyniki obserwacji (tutaj: jednostkowe ceny rynkowe), będą się różnić od wartości oszacowanych przy zastosowaniu danego modelu. Na przykład, jeśli przy użyciu modelu regresji jednej zmiennej (powierzchnia użytkowa) wartość rynkowa 1 m^2 mieszkania o powierzchni użytkowej 50 m^2 wynosi 5 000 zł, a standardowy błąd estymacji wynosi 300 zł to oznacza, w uproszczeniu, że przeciętnie ceny jednostkowe mieszkań o tej powierzchni będą się różnić od oszacowanej wartości o około 300 zł.

W przypadku, gdy znane jest odchylenie standardowe cen rynkowych od ceny średniej, wyrażone wzorem[9]:

$$\sigma = \sqrt{\frac{SST}{N}} \qquad (8)$$

to, biorąc pod uwagę zależności wynikające ze wzorów (1 4), można dokonać następującego przekształcenia:

$$\sqrt{\frac{SSE}{N}} = \sqrt{\frac{SST \cdot (1-R^2)}{N}} = \sqrt{\frac{\sigma^2}{N} \cdot \varphi^2} = \sqrt{\sigma^2 \cdot \varphi^2} \quad (9),$$

co oznacza, że błąd standardowy estymacji jest równy pierwiastkowi kwadratowemu z iloczynu kwadratu odchylenia standardowego obserwacji rynkowych oraz współczynnika zbieżności.

Powróćmy teraz do naszego przykładu pierwszy rzeczoznawca zbudował model, dla którego współczynnik determinacji R^2 wynosi 0,90, przy odchyleniu standardowym cen rynkowych od ceny średniej równym 2 000 zł. Ponieważ model przeciętnie tłumaczy 90% kwadratów różnic, to niewyjaśnione pozostało 10%, co daje następujący standardowy błąd estymacji:

$$\sqrt{\frac{SSE}{N}} = [...] = \sqrt{\sigma^2 \cdot \varphi^2} = \sqrt{2000^2 \cdot 0{,}1} = 632 .$$

Powyższy wynik można interpretować w ten sposób, że przeciętny błąd oszacowania uzyskiwany przy użyciu tego modelu będzie wynosił około 632 zł za 1 m^2 powierzchni użytkowej.

Drugi rzeczoznawca zbudował model, dla którego współczynnik determinacji R^2 wynosi 0,10. Biorąc pod uwagę, że odchylenie standardowe wynosi 500 zł uzyskujemy następujący wynik:

$$\sqrt{\frac{SSE}{N}} = [...] = \sqrt{\sigma^2 \cdot \varphi^2} = \sqrt{500^2 \cdot 0{,}9} = 474 .$$

Oznacza to, że z punktu widzenia odchyleń cen rynkowych od oszacowań modelu, model drugiego rzeczoznawcy charakteryzuje się mniejszym błędem estymacji, pomimo bardzo niskiej wartości współczynnika R^2.

Reasumując z punktu widzenia wykorzystania modelu regresji do szacowania wartości rynkowej nieruchomości (np. lokali mieszkalnych), gdzie przez wartość rynkową rozumie się cenę, wokół której oscylują rzeczywiste ceny nieruchomości o analogicznej charakterystyce rynkowej, lepszym może być model, który tłumaczy mały procent stosunkowo małej zmienności cen, niż model, który tłumaczy duży procent relatywnie dużej zmienności cen. W tym zakresie współczynnik determinacji niekoniecznie jest najwłaściwszą miarą dopasowania modelu do obserwacji rynkowych, zwłaszcza, że jest ona ilorazem sum kwadratów różnic, co może prowadzić do błędnej interpretacji jego wartości liczbowej.

Niska wartość współczynnika R^2 nie oznacza automatycznie, że model jest zły, lecz może oznaczać, iż dobór danych do bazy jest tak trafny, iż zmienność cen w zdecydowanej większości ma charakter odchyleń losowych, których nie da się wytłumaczyć żadnymi cechami rynkowymi.

Podsumowanie

Jak już zostało to zasygnalizowane na samym wstępie, celem artykułu nie była krytyka przyjmowania współczynnika determinacji R^2, jako kryterium wyboru regresyjnego modelu wyceny nieruchomości, lecz uzmysłowienie własności tej statystyki, a w szczególności uświadomienie pułapek, jakie czyhają na rzeczoznawców majątkowych, którzy bezkrytycznie zawierzą tej mierze stopnia dopasowania modelu do obserwacji rynkowych; zwłaszcza biorąc pod uwagę, że model ma być wykorzystywany do szacowania wartości rynkowych. Z tego punktu widzenia analizą należy również objąć inne statystyki, a w szczególności standardowy błąd estymacji.

Jednocześnie warto zasygnalizować, że omówiona w niniejszym artykule problematyka jest pewnym wycinkiem bardziej ogólnego zagadnienia, które dotyczy zasad doboru postaci funkcyjnej modelu. Tej tematyce poświęcone zostaną kolejne artykuły.

[9] Z analogicznych przyczyn, formalnie w mianowniku wzoru na odchylenie standardowe występuje (N 1). Z tych samych przyczyn, co w przypadku wzoru na błąd standardowy estymacji, będziemy stosować uproszczony wzór na odchylenie standardowe.

Bibliografia

1. *Bartels R. [2015] Re-interpreting R^2, regression through the origin, and weighted least squares. University of Sydney Business School.*
2. *Maddala G.S. [2008] Ekonometria, Wydawnictwo Naukowe PWN.*
3. *Wolferton M.L. [2009] An Introduction to Statistics for Appraisers, Appraisal Institute.*
4. *Spiess A.N., Neumeyer N. [2010] An Evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo Approach, BMC Pharmacol.*
5. *Sobczyk M. [2013] Ekonometria, Wydawnictwo C.H. Beck.*
6. *Uniform Standards of Professional Appraisal Practice, 2016 2017 Edition, The Appraisal Foundation.*

DISINFORMATION FACTOR R^2

Summary

The article treats about the analysis of the coefficient of determination R^2 as a measure of the degree of matching multiple regression model to the observed market prices of the property. The article focuses on selected characteristics of R^2 coefficient, in particular the common errors that are made, which include: incorrect interpretation of the formula of R^2, determination of R^2 coefficient for linear regression through the origin and the wrong interpretation of the numerical value of the R^2 coefficient. The article also criticizes the accepting a numerical value R^2, as the decisive criterion for the selection of an econometric model of valuation.

Key words

statistical method in market research, linear regression, hedonic models, coefficient of determination R^2

PRAWO

ZMIANA USTAWY O LASACH

30 kwietnia 2016r. weszła w życie Ustawa z dnia 13 kwietnia 2016r. o zmianie ustawy o lasach (Dz.U. 2016r. poz. 586), zgodnie z którą Skarb Państwa (Lasy Państwowe) uzyskał m.in. prawo pierwokupu gruntów zalesionych sprzedawanych przez osoby fizyczne, osoby prawne lub jednostki organizacyjne nieposiadające osobowości prawnej.

Podpisana nowela wprowadza do ustawy o lasach w art. 37a ust. 1 mechanizm stanowiący, iż w przypadku sprzedaży przez osoby fizyczne, osoby prawne lub jednostki organizacyjne nieposiadające osobowości prawnej, gruntów:

- oznaczonych jako las w ewidencji gruntów i budynków,
- przeznaczonych do zalesienia określonego w miejscowym planie zagospodarowania przestrzennego albo w decyzji o warunkach zabudowy i zagospodarowania terenu,
- lasów objętych uproszczonym planem urządzenia lasu lub decyzją określającą zadania z zakresu gospodarki leśnej,
- Skarbowi Państwa, reprezentowanemu przez Lasy Państwowe, przysługuje z mocy ustawy prawo pierwokupu tych gruntów.

W ramach procedury skorzystania z tych uprawnień nadleśniczy za pośrednictwem właściwego dyrektora regionalnej dyrekcji Lasów Państwowych kieruje pisemny wniosek do Dyrektora Generalnego Lasów Państwowych o wyrażenie zgody na nabycie gruntu.

Nadleśniczy jest uprawniony do zlecenia określenia przez rzeczoznawcę majątkowego wartości gruntu, którego dotyczy wniosek. Ponadto zgodnie z przyjętą ustawą, jeżeli nadleśniczy uzna, że cena określona w dotyczących gruntu umowie lub jednostronnej czynności prawnej rażąco odbiega od wartości rynkowej gruntu, może wystąpić do sądu o ustalenie ceny tego gruntu. W takim przypadku Sąd ustala cenę w oparciu o wartość nieruchomości określoną zgodnie z przepisami ustawy o gospodarce nieruchomościami.

Ustawa wprowadza również zasadę, iż czynności prawne dokonane niezgodnie z przyjętymi regulacjami, w szczególności bez zawiadomienia nadleśniczego, są nieważne.

Ponadto ustawa wprowadziła obowiązek ogłaszania w Biuletynie Informacji Publicznej informacji o wykonaniu przez Lasy Państwowe prawa pierwokupu lub prawa wykupu wraz z podaniem danych dotyczących nabytego gruntu (województwo, powiat, gmina, nazwa oraz numer obrębu ewidencyjnego, a także numer działki ewidencyjnej) oraz ceny, za jaką nastąpiło nabycie.

Opracowanie na podstawie: Dz.U. 2016r. poz. 586 oraz bip.lasy.gov.pl.

Opr. *Wojciech Gryglaszewski*

AKTUALNOŚCI