WYDZIAŁ ZARZĄDZANIA

KIERUNEK: INFORMATYKA I EKONOMETRIA

Studia stacjonarne

Rok II Semestr IV

Analiza wpływu wybranych czynników na ocenę jakości wina na przykładzie gatunku „Vinho verde”

Autor:

Wiktoria Skowron

Kraków, czerwiec 2022

Spis treści

1. Wprowadzenie ... 3
1.1 Cel projektu ... 3
1.2 Opis zbioru danych ... 3
1.3 Opis zmiennych ... 4
2. Budowa modelu ... 7
2.1 Analiza modelu ściśle liniowego ... 7
2.2 Dobór odpowiedniej postaci modelu ... 8
2.2.1 Model liniowy po doborze zmiennych ... 8
2.2.2 Model liniowy po transformacji z użyciem logarytmów ... 9
2.2.3 Model logitowy ... 9
2.3 Właściwości modelu ... 11
2.3.1 Współczynnik determinacji ... 11
2.3.2 Efekt katalizy ... 11
2.3.3 Normalność rozkładu składnika losowego ... 12
2.3.4 Istotność zmiennych ... 12
2.3.5 Testy dodanych oraz pominiętych zmiennych ... 13
2.3.6 Obserwacje odstające ... 14
2.3.7 Test liczby serii ... 14
2.3.8 Test RESET ... 14
2.3.9 Testowanie heteroskedastyczności ... 15
2.3.10 Test Chowa ... 16
2.3.11 Współliniowość ... 16
2.3.12 Koincydencja ... 16
2.3.13 Interpretacja parametrów modelu ... 17
2.3.14 Predykcja wraz z 95% przedziałem ufności ... 17
3. Podsumowanie ... 19
4. Bibliografia ... 19

1. Wprowadzenie

1.1 Cel projektu

Celem projektu jest analiza wpływu wybranych czynników na ocenę jakości wina. W badaniu skoncentrowano się na czynnikach, które podczas doboru zmiennych okazały się istotne i silnie skorelowane ze zmienną jakości.

Hipotezy badawcze:

Czy wyższą oceną jakości cechują się wina o wyższym czy niższym stopniu kwasowości?

Czy jakość wina zależy od zawartości alkoholu?

Czy większa zawartość siarczanów przyczyniła się do gorszej oceny jakości wina?

1.2 Opis zbioru danych

Zbiór danych pochodzi z 2009 roku i zawiera informacje dotyczące portugalskich czerwonych win „Vinho verde". Vinho verde to produkt z regionu Minho (północno-zachodnia Portugalia). Ma umiarkowaną zawartość alkoholu i jest cenione ze względu na orzeźwiający smak. Dane zostały zebrane przez portugalskich badaczy oraz odpowiednio skorygowane i dopasowane do przeprowadzanego badania.

Zmienna jakość została opracowana za pomocą sieci neuronowych. Na początku sieć wytrenowano przy użyciu wcześniej ocenionych pod względem jakości obserwacji. Następnie zastosowano ją do oceny obserwacji z analizowanego zbioru danych. [Cortez et al., 2009]

Zmienne:

Fixed acidity = Kwasowość stała – Wskazuje poziom kwasu winnego i jabłkowego w winie. Ich zawartość w winie zależy od odmiany winogron, rocznika, ale też sposobu fermentacji wina.

Volatile acidity = Kwasowość lotna – Odpowiada za poziom kwasu octowego w winie. Normalny poziom stężenia kwasu octowego w winie powinien wahać się od 0,03% do 0,06%.

Citric acid = Kwas cytrynowy – Opisuje poziom zawartości kwasu cytrynowego.

Residual sugar = Cukier resztkowy – Informuje o poziomie zawartości cukru resztkowego. Cukier resztkowy to cukier, który pozostaje w winie po zakończeniu procesu fermentacji. Przyjmuje się, że gdy cukier resztkowy występuje w ilości do 4 g/l to wino możemy nazywać winem wytrawnym.

Chlorides = Chlorki – Wskazuje na poziom zawartości ezymów dodawanych do wina.

Free sulfur dioxide = Wolny dwutlenek siarki – Odpisuje zawartość wolnych związków siarki. Siarka w winie odpowiada za przedłużenie trwałości produktu. Większą zawartością siarki cechują się wina czerwone, które są analizowane w zbiorze danych.

Total sulfur dioxide = Dwutlenek siarki ogółem – Wskazuje na łączną zawartość związków siarki, która odpowiada za trwałość produktu.

Density = Gęstość – Gęstość badanego nie jest zróżnicowana, a jej wartość jest zbliżona do gęstości wody.

pH = pondus Hydrogenii – Jest to rodzaj miary stopnia kwasowości danego roztworu. Odczyn pH w winie, powinien wynosić 2,8-4.

Sulphates = Siarczany – Opisuje poziom zawartości siarczanów w winie.

Alcohol = Zawartość alkoholu – Odpowiada za poziom zawartości alkoholu w winie. Powinna być wartością z przedziału 8,5% do 15%.

Quality = Jakość – Jest to zmienna oszacowana z pomocą wnioskowania sztucznej inteligencji. Może przyjmować wartości oceny od 0 do 10, gdzie 10 jest najlepszą oceną.

1.3 Opis zmiennych

Zbiór danych składa się 12 zmiennych, z których docelowo do modelu opisującego jakość czerwonego wina od innych czynników zostanie użye kilka najlepiej dopasowanych. Aby wybrać odpowiednio zróżnicowane i niosące możliwie najwięcej informacji zmienne przygotowano zestawienie statystyk (*Tabela 1*) dla każdej ze zmiennych.

Zmienna	Liczba obserwacji	Średnia	Odchylenie standardowe	Wartość minimum	25%	50%	75%	Wartość maximum	Współczynnik zmienności
fixed acidity	1599	8,32	1,74	4,60	7,10	7,90	9,20	15,90	0,21
volatile acidity	1599	0,53	0,18	0,12	0,39	0,52	0,64	1,58	0,34
citric acid	1599	0,27	0,19	0,00	0,09	0,26	0,42	1,00	0,70
residual sugar	1599	2,54	1,41	0,90	1,90	2,20	2,60	15,50	0,56
chlorides	1599	0,09	0,05	0,01	0,07	0,08	0,09	0,61	0,56
free sulfur dioxide	1599	15,87	10,46	1,00	7,00	14,00	21,00	72,00	0,66
total sulfur dioxide	1599	46,47	32,90	6,00	22,00	38,00	62,00	289,00	0,71
density	1599	1,00	0,00	0,99	1,00	1,00	1,00	1,00	0,00
pH	1599	3,31	0,15	2,74	3,21	3,31	3,40	4,01	0,05
sulphates	1599	0,66	0,17	0,33	0,55	0,62	0,73	2,00	0,26
alcohol	1599	10,42	1,07	8,40	9,50	10,20	11,10	14,90	0,10
quality	1599	5,64	0,81	3,00	5,00	6,00	6,00	8,00	0,14

Tabela 1. Zestawienie statystyk dla zmiennych ze zbioru danych

Zaprezentowane statystyki wskazują na obszary, w których można przeprowadzić analizę. Jednymi z najbardziej istotnych i miarodajnych statystyk powinny być zarówno średnia oraz odchylenie standardowe, które są składowymi potrzebnymi do obliczenia współczynnika zmienności, jak i również wartości minimalne i maksymalne osiągane przez zmienne. Odpowiednie wartości kwantyli mogą nieść informację odnośnie rozłożenia wartości w obserwacjach.

Statystyki dla zmiennych powiązanych z kwasowością (*fixed acidity, volatile acidity, citrcic acid*) wykazują się dużą zmiennością ze względu na wartość współczynnika zmienności oraz znaczącą rozbieżność w obserwacjach (różnica pomiędzy statystyką min i max).

Statystyka opisująca zawartość cukru resztkowego (zmienna *residual sugar*) wskazuje na średnią zawartość cukru równą 2,54 gramów o odchyleniu równym 1,14 gramów. Zawartość cukrów do 4 gramów oznacza, że zdecydowana większość badanych win jest wytrawna. Jednak wartość maksymalna równa 15,5 informuje, że wśród obserwacji znajdują się również dane dotyczące win półsłodkich lub słodkich.

Dane dotyczące zmiennej *chlorides* wskazują na zawartość enzymów od 1,2% aż do 61%. Jednak na podstawie wyliczonych kwantyli można zauważyć, że znaczna większość obserwacji sięga maksymalnie 9%. Wartości odstające mogą świadczyć o błędach pomiaru, pomyłkach w przygotowaniu danej partii wina lub specyfice danego gatunku wina.

Statystyki dla zmiennych związanych z zawartością siarki (*free sulfur dioxide, total sulfur dioxide*) również wykazują się bardzo wysoką zmiennością. Jednak w przypadku tych zmiennych dużą rolę może odgrywać występowanie wartości odstających. W obu zmiennych wartość maksymalna znacząco różni się od wartości kwantyla 75%.

Zmienna opisująca zawartość siarczanów *sulphates* wskazuje na umiarkowaną zmienność, a statystyka średniej i odchylenia wskazuje na stosunkowo małe wartości tej zmiennej w badanym zbiorze danych.

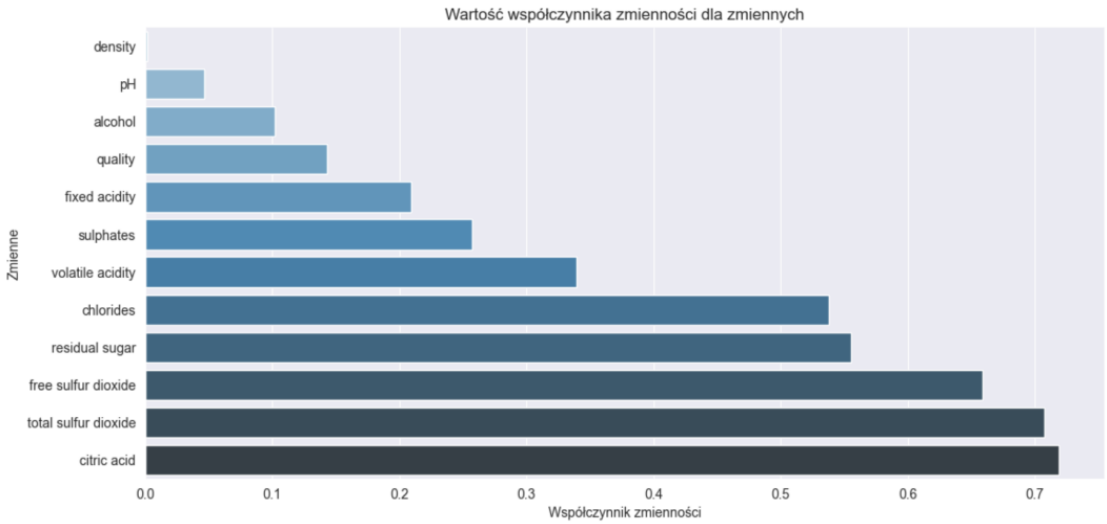
Statystyki dotyczące gęstości (*density*) jasno wskazują, na jednolitość w gęstości badanych win. Wartość zmiennej waha się między 0,9907 a 1,0037, a jej współczynnik zmienności wskazuje niemalże brak zmian. Zmienna nie wnosi istotnych informacji do modelu, więc nie będzie używana w dalszej analizie.

Zmienna *pH* opisująca stopień kwasowości przyjmuje wartości z przedziału od 2,74 do 4,01, co jest charakterystyczne dla wina. Wartości zmiennej nie są zróżnicowane, a sama zmienna nie wnosi istotnych zależności do badania, więc nie będzie używana do budowania modelu.

Zmienna *alcohol* wskazuje na zawartość alkoholu w zakresie 8,4%-14,9%. Wartości przyjmowane przez zmienna są standardowe dla win. Statystyki nie wskazują na wartości odstające, więc wszystkie badane napoje są klasyfikowane jako wina.

Quality jest zmienną zależną od pozostałych zmiennych. Mimo iż jej wartości mogłyby wynosić od 1 do 10, dla badanych obserwacji zmienna przyjmuje wartości od 3 do 8. Średnia ocena jakości wynosi 5,64, a jej średnie odchylenie jest mniejsze niż 1. Statystyki wskazują, że w badanym zbiorze danych wina są klasyfikowane jako przeciętnej jakości. Wśród obserwacji brakuje danych o winach złej oraz znakomitej jakości, co mogłoby być pomocne w opracowaniu modelu.

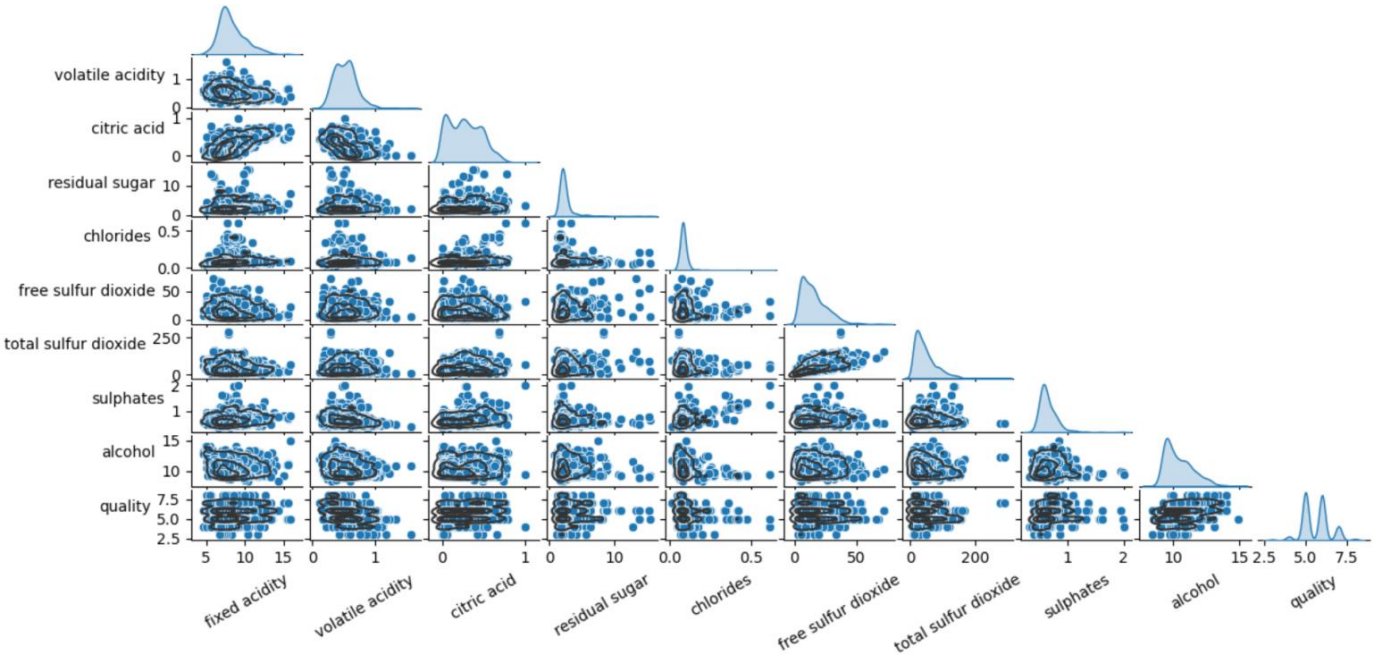
Z uwagi na fakt, że za wstępną miarę przydatności zmiennej w modelu uznano współczynnik zmienności do oceny i porównania wartości współczynnika użyto również metody graficznej. (*Rysunek 1*)



Rysunek 1. Wartość współczynnika zmienności dla zmiennych

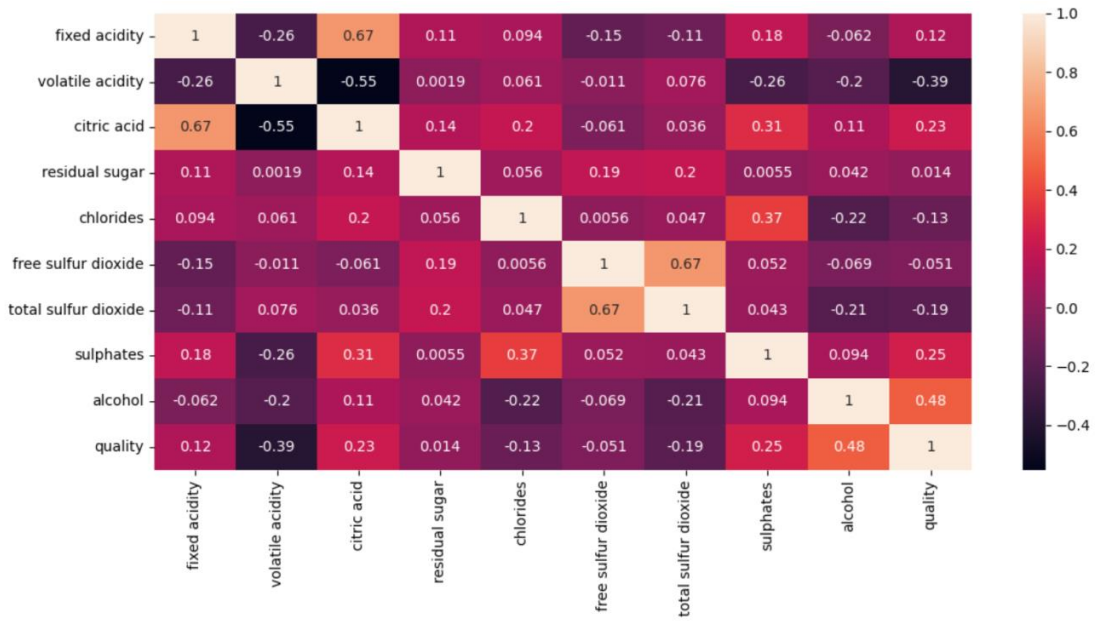
Za współczynnik zmienności wskazujący na odpowiednie zróżnicowanie modelu przyjęto 10%. Z tego powodu zmienne *density* oraz *pH* zostały uznane za niewystarczająco zróżnicowane.

Zestaw zmiennych poddano dalszej analizie i przygotowano dla nich wykresy zależności między sobą wzajemnie oraz macierz korelacji. (*Rysunek 2*)



Rysunek 2. Wykresy zależności między zmiennymi

Wykres wskazuje na niską lub umiarkowaną zależność między zmiennymi, co potwierdza również macierz korelacji. (*Rysunek 3*) Do stworzenia macierzy korelacji użyto także skali ciepła, który wizualnie wskazuje na pary, które mogą okazać się źródłem współliniowości w dalszej analizie modelu regresji.



Rysunek 3. Macierz korelacji między zmiennymi

Macierz wskazuje na wartość współczynnika korelacji poszczególnych czynników ze zmienną opisującą jakość wina. Niektóre z nich wykazują mały wpływ na zmienną zależną oraz wysokie korelacje między sobą, co zostanie wzięte pod uwagę przy dobieraniu zmiennych i odpowiedniej postaci modelu.

2. Budowa modelu

2.1 Analiza modelu ściśle liniowego

Dla zmiennych wybranych na podstawie wstępnej analizy za pomocą Metody Najmniejszych Kwadratów wyestymowano model regresji liniowej, a jego parametry zostały zaprezetowane w *Tabeli 2.*

zmienna	współczynnik	błąd standardowy	wartość p
const	2,64	0,24	3,40E-27
sulphates	0,89	0,11	2,65E-15
alcohol	0,28	0,02	8,82E-57
fixed acidity	0,03	0,01	9,50E-03
residual sugar	0,01	0,01	0,49
free sulfur dioxide	0,00	0,00	0,08
total sulfur dioxide	0,00	0,00	5,81E-05
citricacid	-0,17	0,15	0,26
volatile acidity	-1,14	0,12	3,89E-21
chlorides	-1,62	0,41	8,02E-05

Tabela 2. Parametry wyestymowane we wstępnym modelu MNK

Otrzymany model wymaga poprawy i ograniczenia liczby zmiennych tylko do tych, które wnoszą do modelu najwięcej informacji. W celu dobrania odpowiednich zmiennych zastosowano metodę Hellwiga i metodę krokową-wsteczną.

W metodzie Hellwiga obliczono informacyjne pojemności nośników informacji oraz integralne pojemności każdej z kombinacji. Następnie porównano integralne pojemności dla każdej z kombinacji i wybrano największą z nich równą $H_{max} = 0{,}328$ dla zestawu zmiennych *volatile acidity*, *alcohol* i *sulphates*.

W metodzie krokowej-wstecznej po kolei usuwano zmienne zaczynając od najmniej istotnych. Po każdorazowym usunięciu sprawdzano wpływ zmiany na pozostałe parametry modelu. Przy zastosowaniu metody wybrano 5 zmiennych, które wykazują się istotnością przy badaniu wpływu na jakość. Otrzymany model poddano wstępnej analizie, a także testom mającym na celu zweryfikować słuszność doboru zmiennych.

Wyniki z obu metod porównano (*Tabela 3*) i zdecydowano, że do modelu właściwego użyte zostaną zmienne dobrane podczas obu badań i niebędące katalizatorem. W dalszej analizie model będzie tworzony przy użyciu zmiennych *volatile acidity*, *alcohol* oraz *sulphates*.

metoda doboru zmiennych	liczba zmiennych	zmienne	współczynnik determinacji	uwagi
metoda Hellwiga	3	volatile acidity, alcohol, sulphates	0.335	Brak katalizatorów
metoda krokowa-wsteczna	5	chlorides, sulphates, volatile acidity, total sulfur dioxide, alcohol	0.351	W modelu występują dwa katalizatory: chlorides, total sulfur dioxde. Po usunięciu katalizatorów w modelu pozostają te same zmienne, które uzyskano w metodzie Hellwiga.

Tabela 3. Porównanie wyników metody Hellwiga i krokowej-wstecznej

2.2 Dobór odpowiedniej postaci modelu

2.2.1 Model liniowy po doborze zmiennych

We właściwym modelu zmienna jakość powinna być opisywana za pomocą stałej, kwasowości lotnej, zawartości alkoholu oraz ilości siarczanów zawartych w winie. Jego parametry wraz z oszacowaniem odchyleń i istotności zostały zaprezetowane w *Tabeli 4*.

zmienna	współczynnik	błąd standardowy	wartość p
const	2,61	0,20	1,42E-38
sulphates	-1,22	0,10	1,00E-34
volatile acidity	0,31	0,02	1,41E-76
alcohol	0,68	0,10	2,26E-11

Tabela 4. Parametry wyestymowane we właściwym modelu MNK

2.2.2 Model liniowy po transformacji z użyciem logarytmów

Otrzymany model poddano wstępnej analizie, a następnie dokonano licznych prób transformacji zmiennych, aby polepszyć stopień dopasowania modelu do danych i wyników testów statystycznych.

Uzyskane wyniki uległy niewielkiej poprawie przy zastosowaniu logarytmów. Parametry modelu po korekcie zostały przedstawione w *Tabeli 5*. Do budowy modelu użyto logarytmu zmiennej objaśnianej, a także logarytmów zmiennych zależnych.

$$\log(quality) = \alpha_1 \log(sulphates) + \alpha_2 \log(volatile\ acidity) + \alpha_3 \log(alcohol) + \varepsilon$$

Wartości wszystkich kryteriów informacyjnych świadczą, że wyestymowane parametry dokładniej opisują badane zjawisko. Model z użyciem logarytmów będzie również łatwiejszy w interpretacji, ponieważ będzie wskazywał na zmiany procentowe badanych zmiennych.

zmienna	współczynnik	błąd standardowy	wartość p
const	0,42	0,07	1,43E-08
l_sulphates	0,12	0,01	5,36E-16
l_volatile acidity	-0,10	0,01	2,17E-29
l_alcohol	0,55	0,03	6,13E-63

Tabela 5. Parametry modelu wyestymowane przy użyciu logarytmów

Uznano, że badanym modelem MNK będzie model po transformacji z użyciem logarytmów.

2.2.3 Model logitowy

Przy dokładniej weryfikacji modelu MNK zwrócono uwagę, że model liniowy zbudowany na podstawie logarytmów zawiera wady takie jak niski współczynnik determinacji, brak stabilności parametrów czy prognozy obarczone błędami.

Wzięto pod uwagę również fakt, że zmienna *quality* jest zmienną dyskretną. Uznano zatem, że zmienna *quality* powinna zostać opisana za pomocą modelu logitowego.

Model logitowy to model, który pozwala na oszacowanie prawdopodobieństwa, że przy wzroście danego czynnika wzrośnie wartość zmiennej objaśnianej. Parametry modelu estymowane są przy użyciu Metody Największej Wiarygodności.

zmienna	współczynnik	błąd standardowy	wartość p
residualsugar	0,0531	0,0382	0,1637
chlorides	-4,5010	1,3173	0,006
sulphates	2,7598	0,3539	6,3E-15
citricacid	-0,7790	0,4617	0,0916
freesulfurdioxide	0,0127	0,0067	0,059
volatileacidity	-3,5606	0,3982	3,83E-19
fixedacidity	0,1231	0,0429	0,0041
totalsulfurdioxide	-0,0104	0,0023	8,45E-06
alcohol	0,8773	0,0587	1,5E-50
cut1	3,4190	0,8463	5,34E-05
cut2	5,3366	0,7936	1,77E-11
cut3	9,0601	0,7940	3,68E-30
cut4	11,8990	0,8294	1,12E-46
cut5	14,8954	0,8760	7,63E-65

Tabela 6. Parametry modelu logitowego uporządkowanego

Estymacja logitowa jest innym rodzajem estymacji niż metoda MNK, więc na początku zbudowano model na podstawie wszystkich danych. Oszacowania zostały przedstawione w *Tabeli 6*. Poprawność predykcji dla wyestymowanych parametrów wynosi 59,1%. Następnie z modelu usunięto nieistotne zmienne, a wyniki porównano.

zmienna	współczynnik	błąd standardowy	wartość p
chlorides	-5,0374	1,2607	6,45E-05
sulphates	2,7349	0,3538	1,08E-14
volatileacidity	-3,2338	0,3298	1,08E-22
fixedacidity	0,0711	0,0312	0,0229
totalsulfurdioxide	-0,0077	0,0017	3,69E-06
alcohol	0,8693	0,0570	1,70E-52
cut1	3,0466	0,8248	2,00E-04
cut2	4,9606	0,7704	1,20E-10
cut3	8,6490	0,7704	2,06E-29
cut4	11,4828	0,8043	3,03E-46
cut5	14,4849	0,8527	1,01E-64

Tabela 7. Parametry modelu logitowego po usunięciu zmiennych nieistotnych

W modelu logitowym po korekcie zmiennych nieistotnych pozostały zmienne objaśniające używane w modelu MNK (zawartość alkoholu, siarczanów oraz kwasowość lotna), ale użyto również zmiennych opisujących kwasowość stałą, całkowitą zawartość siarki oraz zawartość enzymów (chlorków). Oszacowania zostały przedstawione w *Tabeli 7*. Poprawność predykcji dla otrzymanego modelu wynosi 60%.

Parametry wskazują, że na wzrost jakości wpływ ma wzrost zawartości siarczanów, stałej kwasowości oraz wzrost poziomu alkoholu. Natomiast wzrost zawartości enzymów

(chlorków), kwasowości lotnej oraz siarki wpływa na spadek oceny jakości wina. W modelu nie ma współliniowości, więc możliwe jest prognozowanie.

2.3 Właściwości modelu

Głównym zadaniem testowania właściwości wyestymowanego modelu jest weryfikacja, czy model spełnia założenia Klasycznej Metody Najmniejszych Kwadratów. W kolejnych punktach zostaną przetestowanie parametry modelu oszacowanego za pomocą MNK po transormacji za pomocą logarytmów.

2.3.1 Współczynnik determinacji

Współczynnik determinacji określa stopień dopasowania modelu do danych. Wartość współczynnika determinacji zawiera się w przedziale [0,1]. Im większa wartość współczynnika, tym wyższy stopień dopasowania modelu.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Skorygowany współczynnik determinacji również służy do oceny dopasowania modelu, jednak koryguje zwykły współczynnik o liczbę obserwacji i liczbę zmiennych w modelu.

$$\overline{R^2} = 1 - \frac{(1 - R^2)(n - 1)}{n - (k + 1)}$$

W zbudowanym modelu zwykły współczynnik determinacji wynosi 32.1%, a skorygowany 32%. Współczynnik determinacji wskazuje, że model wyjaśnia zmienną jakość wina w 32%.

2.3.2 Efekt katalizy

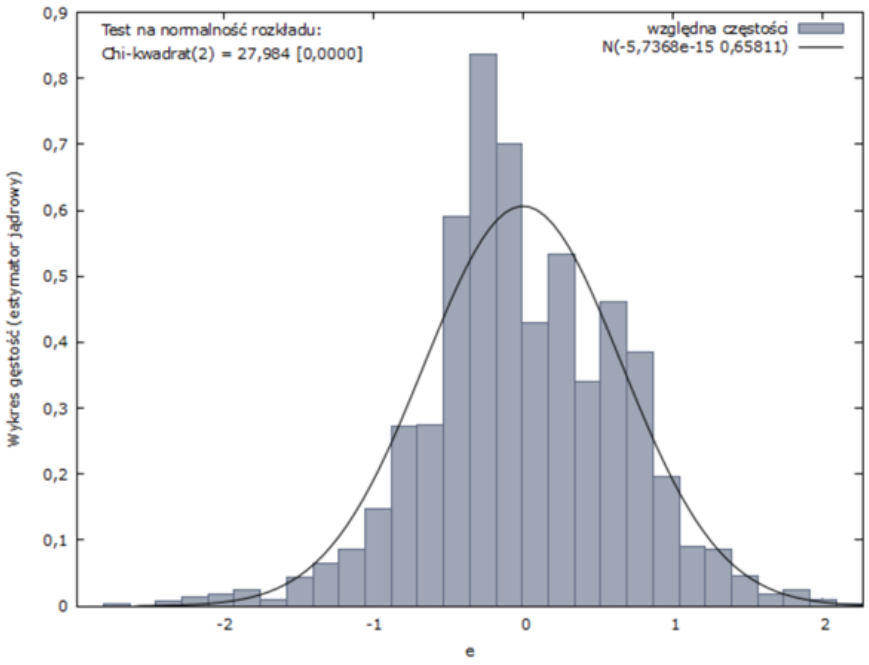
Wartość natężenia efektu katalizy wskazuje, czy współczynnik determinacji R^2 obliczony na podstawie zbudowanego modelu nie jest fałszywie zawyżany przez występowanie katalizatorów, a także wskazuje na prawdopodobieństwo, że ocena zbudowanego modelu jest zawyżona. [Gruszczyński, Pogórska, 2004] Natężenie efektu katalizy obliczane jest przy użyciu pojemności integralnej kombinacji nośników informacji (obliczonej przy użyciu metody Hellwiga) oraz współczynnika determinacji.

$$\mu = H - R^2$$

W badanym modelu nie ma katalizatorów, a natężenie efektu katalizy wyniosło 0.7%, więc prawdopodobieństwo fałszywego zawyżenia wartości współczynnika determinacji R^2 jest znikome.

2.3.3 Normalność rozkładu składnika losowego

Normalność rozkładu składnika losowego jest jednym z dodatkowych założeń weryfikacyjnych KMNK. Jeśli przy małej liczności próby nie ma rozkładu normalnego to miarodajność dalszych testów statystycznych jest obniżona.



Rysunek 4. Wykres częstości rozkładu składnika losowego

Hipotezy:

H_0: *składnik losowy ma rozkład normalny*

H_1: *składnik losowy nie ma rozkładu normalnego*

Wnioski: Dla analizowanych obserwacji wartość p=0, zatem odrzucono H_0. Według testu składnik losowy nie ma rozkładu normalnego.

Z uwagi na fakt, że w zbiorze danych znajduje się 1599 obserwacji można przyjąć, że estymatory parametrów MNK są asymptotycznie normalne i ich rozkład zbliżony jest do rozkładu normalnego. To oznacza, że można swobodnie interpretować wartości parametrow i testów oraz wykonywać proces weryfikacji. [Gładysz, Mercik, 2007]

2.3.4 Istotność zmiennych

Istotność zmiennych modelu wskazuje na powiązanie dobranych zmiennych objaśniających ze zmienną objaśnianą. Weryfikacja istotności zmiennych przeprowadzana jest za pomocą testu t-Studenta. [Gładysz, Mercik, 2007]

Hipotezy testu:

$H_0: \ \propto_j = 0 \ \rightarrow zmienna\ jest\ nieistotna$

$H_1: \ \propto_j \neq 0 \ \rightarrow zmienna\ jest\ istotna$

Do oceny słuszności hipotezy zerowej używana jest statystyka:

$$t = \frac{a_j}{S(\propto_j)}, gdzie\ a_j - estymator\ współczynnika\ \propto_j,$$

$$S(\propto_J) = estymator\ dyspersji\ współczynnika\ \propto_j$$

Wartości p dla analizowanych zmiennych zostały przestawione w *Tabeli 5*.

Wnioski: Odrzucono H_0. Każda z badanych zmiennych jest istotna w budowie modelu.

2.3.5 Testy dodanych oraz pominiętych zmiennych

Testy dodanych zmiennych wskazują, że dodanie do modelu niektórych zmiennych mogłoby mieć wpływ na niewielką poprawę dopasowania do danych . Jednak w procesie doboru zmiennych *chlorides, fixedacidity* i *totalsulfurdioxide* zostały uznane za nieisotne lub wpływające na efekt katalizy (*chlorides, totalsulfuracidity*). Wnioski zostały przedstawione w *Tabeli 8.*

dodane zmienne	skorygowany współczynnik determinacji	wpływ na kryteria
l_residualsugar	31,97%	brak poprawy kryteriów informacyjnych
l_chlorides	32,55%	poprawa wszystkich kryteriów informacyjnych
l_freesulfurdioxide	31,97%	brak poprawy kryteriów informacyjnych
l_fixedacidity	32,18%	poprawa 2 z 3 kryteriów informacyjnych
l_totalsulfurdioxide	32,19%	poprawa 2 z 3 kryteriów informacyjnych
l_residualsugar l_freesulfurdioxide	31,93%	brak poprawy kryteriów informacyjnych
l_fixedacidity l_totalsulfurdioxide	32,57%	poprawa wszystkich kryteriów informacyjnych

Tabela 8. Wnioski z testów dodanych zmiennych

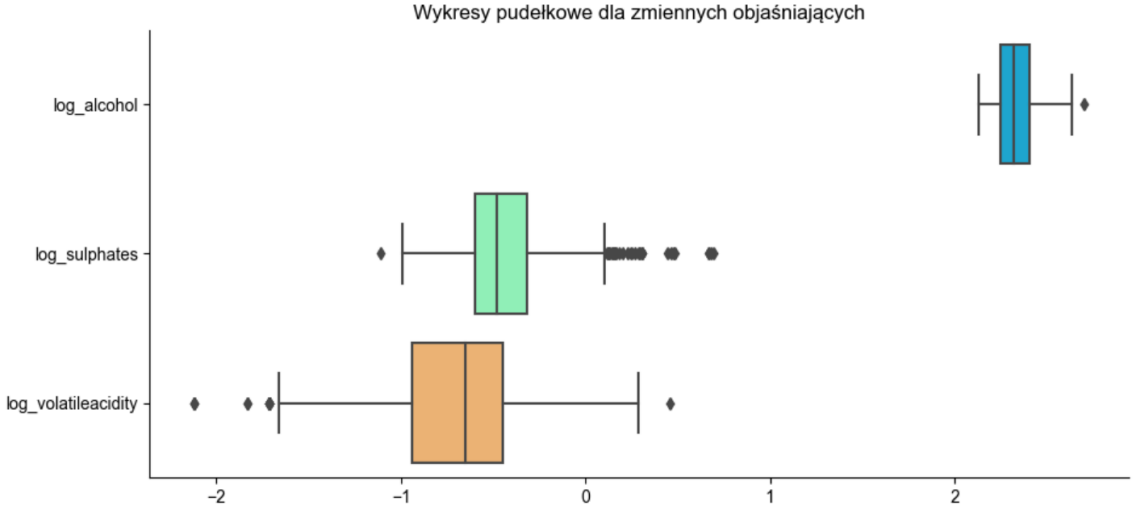
Przeprowadzone testy pominiętych zmiennych wskazują, że pominięcie dowolnej zmiennej bądź ich kombinacji nie spowoduje poprawienia dopasowania modelu. W każdym przypadku wpłynie na spadek współczynnika determinacji i wartość kryteriów wskazującą na mniej dokładne opisywanie zmiennej jakości przez parametry. Wnioski zostały przedstawione w *Tabeli 9.*

pominięte zmienne	skorygowany współczynnik determinacji	wpływ na kryteria
l_volatilacidity	26,43%	brak poprawy kryteriów informacyjnych
l_alcohol	18,98%	brak poprawy kryteriów informacyjnych
l_sulphates	29,19%	brak poprawy kryteriów informacyjnych
l_volatilacidity, l_alcohol	9,09%	brak poprawy kryteriów informacyjnych
l_volatilacidity, l_sulphates	20,57%	brak poprawy kryteriów informacyjnych
l_alcohol, l_sulphates	15,19%	brak poprawy kryteriów informacyjnych

Tabela 9. Wnioski z testów pominiętych zmiennych

2.3.6 Obserwacje odstające

Obserwacje odstające to ekstremalnie wysokie lub niskie wartości danych w odniesieniu do pozostałych wartości w zbiorze danych. Są to wartości skrajne, które znacznie wyróżniają się od ogólnego rozkładu danych. [Gruszczyński, Pogórska, 2004]



Rysunek 5. Wykresy pudełkowe dla zmiennych objaśniających

W zbiorze zmiennych zależnych można zaobserwować nieliczne wartości odstające. Obserwacje te nie odstają znacząco od pozostałych wartości. Wartości odstające zostały zaprezentowane graficznie na *Rysunku 5*. Ze względu na to, że jest ich mało nie będą miały wpływu na zakłócenie modelu.

2.3.7 Test liczby serii

Test liczby serii weyfikuje, czy istnieje liniowa zależność między zmienną objaśnianą a zmiennymi objaśniającymi. Test ma szczególne znaczenie dla prawidłowej interpretacji współczynnika determiacji. [Gruszczyński, Pogórska, 2004]. Przed wykonaniem testu serii dane zostały posortowane po wartościach zmiennej *quality*, czyli zmiennej objaśnianej.

Hipotezy testu

H_0: *postać modelu jest dobrze dobrana, próba jest losowa*

H_1: *postać modelu nie jest dobrze dobrana, próba nie jest losowa*

Wnioski: Obliczona liczba serii wynosi 340. Wartość jest mniejsza od wartości krytycznej, dlatego odrzucono H_0. Postać modelu nie jest dobrze dobrana.

2.3.8 Test RESET

Testowanie błędnej specyfikacji modelu to sposób na sprawdzenie, czy model jest odpowiedni lub czy można go udoskonalić. Model może być niewłaściwie zdefiniowany, jeśli pominięto istotne zmienne, włączono zmienne nieistotne, wybrano niewłaściwą formę funkcyjną lub opracowano model, który narusza założenia modelu regresji. Test RESET

(REgression Specification Error Test) służy do wykrywania zmiennych pominiętych i nieprawidłowej postaci analitycznej. [Hill, Griffiths, Lim, 2011].

Hipotezy testu:

H_0: *wybór postaci analitycznej modelu jest prawidłowy*

H_1: *wybór postaci analitycznej modelu nie jest prawidłowy*

Wnioski: Wartość p wynosi 0,122, więc nie ma podstaw do odrzucenia hipotezy zerowej. Wybór postaci analitycznej modelu jest prawidłowy.

2.3.9 Testowanie heteroskedastyczności

Heteroskedastyczność to zjawisko polegające na różnych wariancjach składnika losowego, które zgodnie z założeniami Klasycznej Metody Najmniejszych Kwadratów powinny być takie same.

Heteroskedastyczność można badać poprzez obserwacje wariancji składnika losowego na wykresie lub przy użyciu testów statystycznych takich jak test Lagrange'a, test White'a czy test Breusch'a-Pagan'a. [Hill, Griffiths, Lim, 2011] Dla modelu wykonano test Breusch'a-Pagan'a.

H_0: *Brak heteroskedastyczności, wariancje składnika losowego są równe*

H_1: *W modelu istnieje heteroskedastyczność, wariancje składnika losowego są różne*

Wnioski: Odrzucono H_0, ponieważ wartość p jest równa 0,000245. W modelu MNK zbudowanym na podstawie dobranych zmiennych można zaobserwować heteroskedastyczność składnika losowego. Zaobserowane zjawisko powoduje, że oszacowane parametry są nieefektywne, co oznacza, że nie nadają się do prognozowania. Należy zmienić metodę estymacji.

Na podstawie wyciągniętych wniosków, przed przystąpieniem do prognozowania, dokonano korekty heteroskedastyczności, czyli zastosowano uogólniony model KMNK.

Zgodnie z wcześniejszymi analizami i wnioskami badany model nie spełnia założenia MNK o homoskedastyczności dlatego należy dokonać korekty heteroskedastyczności, a następnie stworzyć prognozę na bazie otrzymanego modelu. Do uzyskania modelu zastosowano uogólnioną metodę MNK, która nie ma założenia o homoskedastyczności, a jej jedynym założeniem jest normalność rozkładu składnika losowego.

zmienna	współczynnik	błąd standardowy	wartość p
const	0,550	0,075	3,60E-13
l_volatileacidity	-0,110	0,009	2,24E-31
l_alcohol	0,493	0,032	5,09E-50
l_sulphates	0,140	0,015	3,63E-20

Tabela 10. Parametry modelu po korekcie heteroskedastyczności

W uogólnionym modelu MNK skorygowany współczynnik determinacji uległ nieznacznemu pogorszeniu – model wyjaśnia zmienną jakość wina w 30,7%. Współczynniki dla parametrów również uległy niewielkim zmianom. Parametry modelu po korekcie zostały zaprezentowane w *Tabeli 8.*

2.3.10 Test Chowa

Test Chowa jest testem F i spełnia swoją funkcję przy założeniu homoskedastyczności. Test polega na podziale zbiory danych na dwie próby, a nastepnie weryfikacji parametrów. Parametry ocenia się jako stabilne, jeżeli współczynniki przy parametrach mają podobną (nie rożne istotnie) wartość w obu próbach. [Wooldrigde, 2013] W przeprowadzonym teście próby zostały uzyskane przez podzielenie wszystkich obserwacji na dwie grupy podobnej wielkości.

Hipotezy testu:

H_0: $parametry\ są\ stabilne$

H_1: $parametry\ nie\ są\ stabilne$

Wnioski: Wartość p dla testu Chowa wynosi 0, więc należy odrzucić hipotezę zerową i uznać, że parametry modelu nie są stabilne.

2.3.11 Współliniowość

Współliniowość jest zjawiskiem występującym, gdy dwie lub więcej zmiennych w modelu są ze sobą silnie skorelowane.

Zmienne zostały dobrane przy użyciu meotdy Hellwiga, która wyklucza współliniowość między zmiennymi objaśniającymi. W modelu nie ma współliniowości.

2.3.12 Koincydencja

Model jest koincydencjalny, gdy znak współczynnika korelacji między zmienną x_i oraz zmienną objaśnianą będzie taki sam jak znak przy wyestymowanym parametrze. Brak koincydencji może świadczyć o o współliniowości zmiennych objaśniających. [Gładysz,..]

$$sign\big(r(x_i, y)\big) = sign(a_i)$$

$gdzie$: $r(x_i, y) - korelacja\ między\ zmienną\ x_i\ objaśniającą\ a\ zmienną\ zależną$

$a_i - wartość\ współczynnika\ w\ modelu\ ekonometrycznym\ przy\ zmiennej\ x_i$

W badanym modelu znaki korelacji pokrywają się ze znakami przy współczynnikach. Świadczy to, że oba z tych modeli są koincydencjalne.

$$R_0 = \begin{bmatrix} 0{,}47617 \\ 0{,}25140 \\ -0{,}39056 \end{bmatrix}$$

Równanie 1. Macierz korelacji zmiennych objaśniających ze zmienną zależną

2.3.13 Interpretacja parametrów modelu

Przy estymacji modelu zostały użyte logarytmy, co ułatwia interpretację parametrów, ponieważ wpływ czynników na jakość wina możemy opisywać procentowo.

Gdy ilość siarczynów (*sulphates*) wzrasta o 1% to ocena jakości wina wzrasta średnio o 0.14% pod warunkiem, że pozostałe czynniki zostały niezmienione.

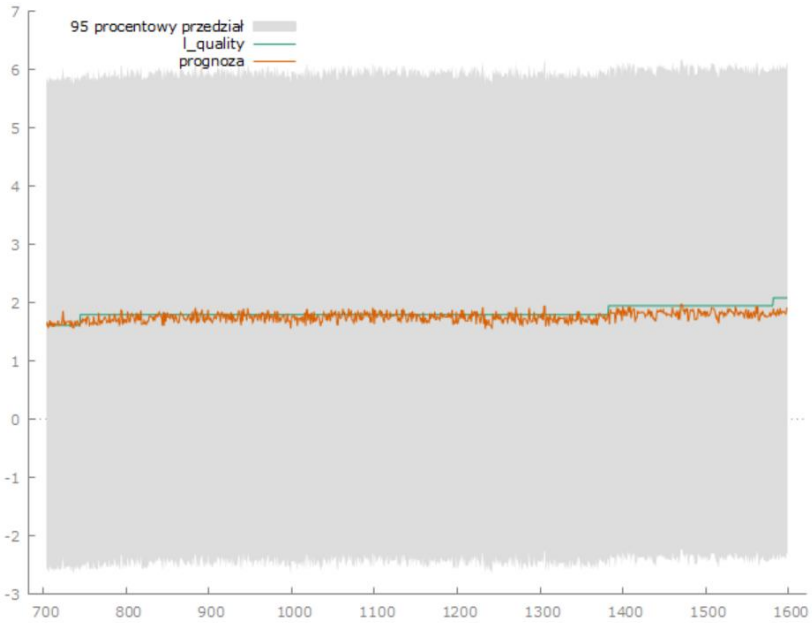
Gdy kwasowość lotna (*volatile acidity*) wzrasta o 1% to ocena jakości wina obniża się średnio o 0.11% pod warunkiem, że pozostałe czynniki zostały niezmienione.

Gdy poziom alkoholu (*alcohol*) wzrasta o 1% to ocena jakosci wina wzrasta srednio o 0.49% pod warunkiem, że pozostałe czynniki zostały niezmienione.

Można zauważyć, że największy wpływ na ocenę jakości wina ma poziom alkoholu.

2.3.14 Predykcja wraz z 95% przedziałem ufności

Przygotowano prognozę dla modelu po korekcie heteroskedastyczności dla 95% przedziału ufności. Na początku obliczono, że prognoza punktowa *log(quality)* powinna wynosić 1,71. Natomiast średni błąd wynosi 2,14 i przewyższa wartość prognozy, co świadczy o niedokładności przewidywań. Po kalkulacji dla *quality* wraz z uwzględnieniem wartości wariancji obliczono, że prognoza dla zmiennej powinna wynosić 55,1. Otrzymany wynik wydaje się być nieuzasadniony i prawdopodobnie obarczony błędem biorąc po uwagę, że w analizowanym zbiorze danych ocena jakości może przyjmować wartości od 1 do 10.



Rysunek 6. Prognoza oparta na uogólnionym modelu MNK

Przyjmując za współczynnik ufności 95% obliczono prognozę przedziałową dla *log(quality).* Wygenerowany wykres dla prognozy oraz dokonane obliczenia wskazują, że przy przyjętym przedziale ufności prognozą przedziałową dla zmiennej *log(quality)* jest przedział od -2,48 do

5,92, co oznaczałoby, że prognozą dla zmiennej *quality* jest przedział od 0,083 do 371,2. Zakres prognozy został przedstawiony na powyższym wykresie. (*Rysunek 6*) Podobnie jak w przypadku prognozy przedziałowej wyniki są wątpliwe i prawdopodobnie obarczone błędem. Miary dokładności prognoz zostały zaprezentowane w *Tabeli 11*.

Miary dokładności prognoz ex post	Oznaczenie	Wartość
Średni błąd predykcji	ME	0,0712
Średni błąd procentowy	MPE	8,99%
Średni błąd absolutny	MAE	3,7946
Średni absolutny błąd procentowy	MAPE	4,8590

Tabela 11. Miary dokładności prognoz ex post dla modelu

Błędy obliczone przy prognozowaniu nie wskazują uzasadnienie otrzymanych prognoz. Średni błąd procentowy wynosi ok. 9%, co można uznać za umiarkowaną wartość. Badając jakość podobny błąd mogły spowodować zaniżenie lub zawyżenie oceny o 1 jednostkę.

Dokonano również prognozy dla modelu logitowego. Dla modelu dokonano prognozy ex post przy użyciu 896 obserwacji. Miary dokładności prognoz zostały zaprezentowane w *Tabeli 12*.

Miary dokładności prognoz ex post	Oznaczenie	Wartość
Średni błąd predykcji	ME	0,4063
Średni błąd procentowy	MPE	6,2160
Średni błąd absolutny	MAE	0,4777
Średni absolutny błąd procentowy	MAPE	7,4213
Pierwiastek błędu średniokwadr.	RMSE	0,7212

Tabela 7. Miary dokładności prognoz ex post dla modelu

Model na podstawie parametrów przewidywał ocenę jakości wina, a następnie porównał przewidywania do wartości realnych. (*Tabela 13*)

ocena predykcji	liczba obserwacji	udział w całości
poprawna, różnica między prognozą quality a quality wynosi 0	487	54,4%
różnica między prognozą quality a quality wynosi 1	390	43,5%
różnica między prognozą quality a quality wynosi 2	19	2,1%
Total	896	100,0%

Tabela 13. Ocena predykcji jakości wina w modelu logitowym

Przy 896 testowanych obserwacjach model poprawnie dopasował ocenę jakości do 54,4% z nich. Błędne prognozy wyniosły 45,6% wszystkich przewidywań.

Z otrzymanych obliczeń wynika, że model logitowy jest modelem lepiej dopasowanym do analizowanych danych, a prognozy oparte na tym modelu są bardziej wiarygodne i obarczone mniejszym ryzykiem błędu.

3. Podsumowanie

Zaproponowane w pracy modele wraz z właściwościami opisują zależność między oceną jakości wina a jego cechami charakterystycznymi. Zarówno modele wyestymowane metodą MNK (oraz UMNK) jak i modele logitowe szacują prognozy obarczone błędami. Jednak każdy z modeli opisuje czynniki takie jak kwasowość, zawartość siarki i siarczanów oraz poziom alkoholu jako istotne przy ocenie jakości wina. Analizowane obserwacje wykazują się dużą rozbieżnością danych – co z jednej strony mogłoby być pomocne w dokładnej ocenie zjawiska, ale z drugiej jest problem może leżeć w sposobie oceny zmiennej jakości.

Analizowane modele pomogły zweryfikować przyjęte hipotezy badawcze. Zauważono, że lepszą oceną jakości cechowały się wina o niższej kwasowości lotnej. Zaskakującym może być fakt, że zawartość siarczanów nie wpłynęła negatywnie na ocenę jakości wina. Jednak nie zawyżyła jej znacząco. Najwyższe znaczenie na poprawę oceny miał poziom alkoholu zawartego w winie. Przytoczone wnioski wynikają głównie ze wstępnej oceny jakości przez badaczy w grupie obserwacji, które posłużyły do wytrenowania sieci klasyfikującej jakość. Badacze uznali, że wina o niższej kwasowości, a wyższej zawartości siarczanów i alkoholu cechowały się najwyższą jakością.

4. Bibliografia

1. „Modeling wine preferences by data mining from physicochemical properties” Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, Jose Reis, 2009
2. „Modelowanie ekonometryczne – Studium przypadku” Barbara Gładysz, Jacek Mercik; Wrocław, 2007
3. „Principles of Econometrics” R. Carter Hill, William E. Griffiths, Guay C. Lim; 2011
4. „Ekonometria” Marek Gruszczyński, Maria Podgórska; Warszawa, 2004
5. „Introductory Econometrics – A modern approach” Joeffrey M. Wooldrigde; Mason, 2013
6. „Using gretl for Principles of Econometrics” Lee C. Adkins, Oklahoma State University, 2012