Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

|  |  |
| --- | --- |
| Mateusz Majak  Numer albumu: 384912 | Wojciech Nowakowski  Numer albumu:388298 |

**Determinanty oceny filmu wystawionej przez użytkowników portalu TMDb**

Praca przygotowana w ramach zajęć

z Ekonometrii

pod kierunkiem mgr Kateryny Zabariny

Warszawa, styczeń 2019

**Wprowadzenie**

Celem poniższej pracy jest analiza czynników, które mają wpływ na średnią ocenę filmów przez widzów. Za źródło danych posłużyła internetowa baza filmów i seriali TMDb.

Jako zmienną objaśnianą wybraliśmy średnią ocenę filmu (*vote\_average*) wystawioną przez użytkowników TMDb. Znaczenie ocen użytkowników jest bardzo duże, ponieważ ludzie często uważają średnie oceny z serwisów agregujących opinie za godne zaufania i opierają decyzję właśnie przy ich pomocy[[1]](#footnote-1). W przypadku filmów, konsument decyduje często czy ponieść koszt biletu lub wypożyczenia w serwisach VOD, a także swojego wolnego czasu. Wpływ decyzji i opinii konsumentów na zyski i sukcesy firmy są przedmiotem licznych badań i prac. Nasza praca koncentruje się na badaniu wpływu zmiennych, które uznaliśmy za jednoznacznie obiektywne (pominęliśmy na przykład gatunki filmów, ze względu na coraz szerzej klasyfikowane filmy, znajdujące się często w kilku kategoriach gatunków filmowych).

Ocena filmu, podobnie jak innych dzieł kultury i sztuki, jest kwestią bardzo indywidualną. Wyjaśnienie więc dokładnego wpływu na końcową ocenę może być trudne, jednak ze względu na rozmiar branży filmowej można przypuszczać, że wpływ na ocenę widza mają również techniczne cechy filmów.

Na zmienne objaśniające wybraliśmy budżet i dochód filmu, wiek, długość trwania i liczba ocen.

**Przegląd literatury**

Kwestie związków, które wpływają na ocenę filmów poruszają w swej pracy zatytułowanej ,,Correlations Between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database’’ Max Wasserman, Satyam Mukherjee, Konner Scott, Xiao Han T. Zeng. Badali oni wpływ ocen użytkowników portali oceniających filmy na zysk. Przeprowadzili oni badanie korelacji na wybranym zbiorze filmów, by złagodzić skutki błędów związane z językiem i krajem w którym film powstawał. Zauważyli oni silną korelację między liczbą głosów użytkowników, a zyskami osiągniętymi z filmów. Autorzy artykułu zebrali dane z portali IMD i przeprowadzili regresje liniowe, w celu kwantyfikacji korelacji pomiędzy danymi głosowania użytkowników a informacjami dostarczanymi przez użytkowników. Wzięli bazę filmów krótkometrażowych jak i długometrażowych. Krótkie filmy były konieczne dla ich badania, ponieważ bez nich pominięte byłyby filmy wyprodukowane przed 1920 rokiem. Filmy zostały podzielone pod względem języka, w którym zostały nakręcone oraz kraju pochodzenia. Pierwszą grupę tworzą filmy anglojęzyczne wyprodukowane w USA. Druga grupa to filmy anglojęzyczne wyprodukowane poza USA, a trzecią grupę stanowią filmy nieanglojęzyczne i wyreżyserowane poza USA. Pierwsza grupa ma największą liczebność. Zauważono, że od dwóch dekad co roku powstaje coraz więcej filmów, recesja zaczęła się w 2009 roku i trwała do 2011, jednak autorzy tłumaczą, że ten fakt jest spowodowany poprzez kryzys 2007/2008.

Autorzy badania zdecydowali się na zlogarytmowanie budżetu i całkowitej liczby głosów oddanych przez użytkowników portalu. Logarytm został zastosowany, ponieważ wartości rozciągały się o kilka rzędów wartości. Zostały znormalizowane statystyki monetarne w ujęciu rocznym, uwzględniając inflację, wzrost cen biletów do kina oraz wzrost liczby ludności. Wyniki badań wskazały że liczba głosów użytkowników jest wskaźnikiem znaczenia filmu. Jednak czołowe miejsca niekoniecznie są związane z sukcesem, ponieważ wiele filmów osiąga sukces z innych powodów takich jak nominacje do dużych nagród filmowych. Ustalono również, że wpływ na zyski z filmu mają środki przeznaczone na reklamę i promocje filmu. Autorzy doszli do wniosku że jakość filmu - w postaci przeciętnej oceny użytkownika - nie ma znacznego znaczenia w kontekście zastosowania w połączeniu z budżetem w modelu liniowym.

Drugi artykuł autorstwa doktora Scott Wallsten ,,Rotten Tomatoes and IMDB Reviews Strongly Correlated With Movie Revenues’’. Na potrzeby swojego badania stworzył bazę danych zawierającą wyniki Rotten Tomato dla prawie 2000 filmów oraz wyniki IMDB dla ponad 4000 filmów wydanych między 1925 a 2016 r. Jednak więcej filmów z zestawu danych została wyprodukowana w latach 2004-2016. W badaniu analizowano ogólną korelację między wynikami a przychodami, ale nie kontrolowaliśmy innych czynników wpływających na przychody, takich jak rodzaj filmu, ocena i inne czynniki wpływające na przychody. Autor korzystał z trzech źródeł danych Budżet i dochody są dostarczane w dolarach nominalnych, więc zamieniłem je na stałe w 2016 r. Za pomocą deflatora PKB. Wyniki pracy sugerują, że wzrost tomatometer wyniku o jeden punkt procentowy w skali związany jest ze wzrostem światowych przychodów o 1,12 miliona USD. Film oceniany na 60 procent zarobiłby o 11,2 miliona więcej niż film oceniony na 50 procent. Podobnie, mają się wyniki z bazy IMD, film oceniany jako 6.0 na zarabiałby o 38 milionów więcej niż film oceniony 5.0. Biorąc pod uwagę, że średnie światowe przychody wynoszą około 116 milionów USD. Autor zwraca uwagę na to, że są to dosyć duże liczby. Scott Wallsten nie uważa, że oceny na Rotten Tomatoes i IMDB nie szkodzą sprzedaży biletów do kina. Ponieważ są recenzje pozytywne jaki i negatywne, te negatywne zmniejszają dochód z filmów, ale za to te pozytywne poprawiają jego zyski.

**Opis bazy danych**

Źródłem bazy danych do modelu była strona internetowa: [www.themoviedb.org](https://www.themoviedb.org/).

Jest to strona internetowa poświęcona filmom i serialom. Użytkownicy (zarejestrowani) mogą oddawać głos na wybrany film lub serial (w skali od 1 do 10). Strona udostępnia interesujące informacje na temat produkcji, takie jak: obsada, ekipa filmowa, czas trwania, budżet, dochód, miejsce produkcji i inne. W internecie można znaleźć inne, również większe bazy (IMDb czy Filmweb). Za badaną bazę wybraliśmy TMDb ze względu na informację o budżecie i dochodach, które to uznaliśmy za prawdopodobnie ważne zmienne, które posłużą nam w estymacji modelu.

Pobrana baza zawierała 4807 rekordów. Z podanych kolumn wybraliśmy te, które uznaliśmy za istotne dla naszego modelu. Dodatkowo zamieniliśmy daty wydania filmów na wiek w latach, liczony od 10. stycznia 2019 roku.

Po oczyszczeniu bazy z rekordów zawierających puste lub błędne, niemożliwe do uzupełnienia komórki (brak pola vote\_count czy revenue był dla nas podstawą do usunięcia rekordu) otrzymaliśmy bazę zawierającą 3112 filmów.

Poniżej znajduje się lista zmiennych objaśniających, które ostatecznie zostały użyte w modelu, wraz z opisami.

**Opis zmiennych**

W poniższej tabeli znajdują się wszystkie wykorzystane w modelu zmienne objaśniające wraz z krótkim opisem. Szczegółowe opisy znajdują się poniżej tabeli.

|  |  |  |
| --- | --- | --- |
| **Zmienna objaśniająca** | **Opis zmiennej** | **Nazwa zmiennej** |
| Liczba ocen filmu | Liczba ocen wystawionych przez użytkowników strony TMDb. Użytkownik musi zarejestrować się na stronie i zalogować, aby móc ocenić produkcję. | vote\_count |
| Długość trwania filmu | Długość trwania filmu wyrażona w pełnych minutach | runtime |
| Dochód | Dochód z dystrybucji filmu wyrażony w dolarach amerykańskich (USD) | revenue |
| Budżet | Budżet przeznaczony na film wyrażony w dolarach amerykańskich (USD) | budget |
| Popularność | Miara popularności stworzona przez TMDb. Składają się na nią: Liczba ocen filmu danego dnia, liczba wyświetleń strony z filmem danego dnia, liczba użytkowników którzy dodali film do ulubionych danego dnia, liczba użytkowników którzy dodali film do listy życzeń (“watchlist”), data wydania filmu, Całkowita liczba ocen filmu, wartości miernika *popularity* z poprzednich dni. | popularity |
| Wiek | Wiek filmu wyrażony w latach (mierzony od 10. stycznia 2019 roku). | years\_old |

**Hipotezy badawcze**

W naszej pracy postawiliśmy następujące hipotezy:

* Im większa liczba ocen tym wyższa średnia ocena filmu (duża liczba ocen świadczy o popularności, a tym samym prawdopodobnie o atrakcyjności dla konsumenta)
* Im większy budżet filmu tym wyższa średnia ocena filmu (wysoki budżet przekłada się często na lepsze efekty specjalne, obsadę oraz ekipę filmową)
* Im większy dochód tym wyższa średnia ocena filmu
* Im większa popularność filmu tym wyższa średnia ocena filmu
* Im mniejsza długość trwania filmu tym tym wyższa średnia ocena filmu (wydaje się, że krótsze filmy są bardziej preferowane przez konsumentów)
* Im starszy film tym wyższa ocena filmu

**Statystyki opisowe zmiennych**

Tabela poniżej przedstawia podstawowe statystyki opisowe (średnia, odchylenie standardowe, wartość minimalna oraz wartość maksymalna) zmiennych ciągłych wykorzystanych w modelu.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| VARIABLES | N | mean | sd | min | max |
| vote\_average | 3112 | 6.334 | 0.837 | 2.300 | 8.500 |
| runtime | 3112 | 111.0 | 20.96 | 63 | 338 |
| popularity | 3112 | 30.03 | 36.46 | 0.179 | 875.6 |
| years\_old | 3112 | 16.99 | 12.74 | 3 | 103 |
| revenue | 3112 | 1.256e+08 | 1.884e+08 | 6399 | 2.788e+09 |
| budget | 3112 | 4.187e+07 | 4.471e+07 | 7000 | 3.800e+08 |
| vote\_count | 3112 | 1013 | 1428 | 20 | 13752 |

Tabela 1. Statystyki opisowe

Uwagę zwracają wartości minimalne i maksymalne zmiennych *popularity, years\_old, revenue, budget i vote\_count* ze względu na duże różnice między nimi. Dla lepszego oszacowania modelu spróbujemy wykluczyć mniejsze lub większe wartości części tych zmiennych w dalszej części pracy.

**Macierz korelacji**

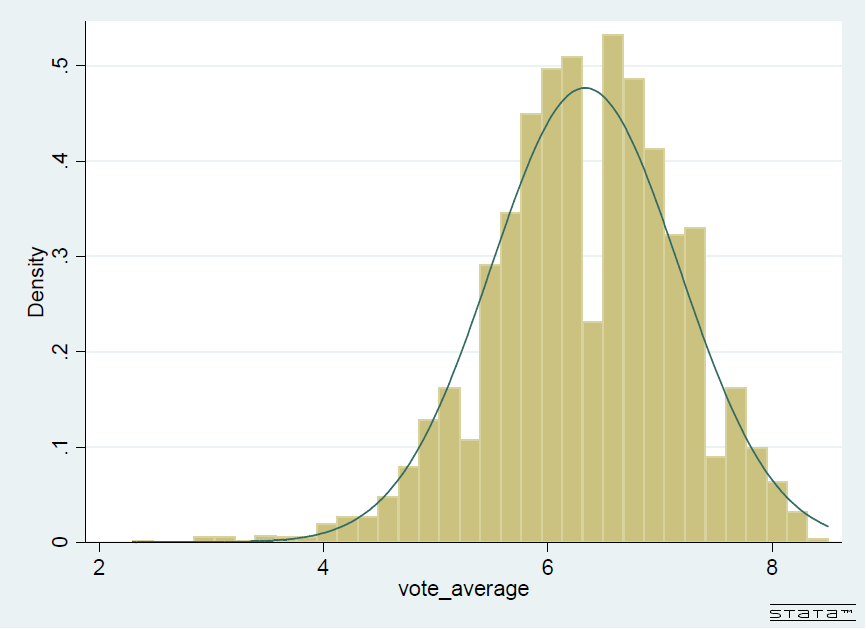
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **vote\_average** | **vote\_count** | **runtime** | **years\_old** | **popularity** | **revenue** | **budget** |
| **vote\_average** | **1** |  |  |  |  |  |  |
| **vote\_count** | **0.384\*\*\*** | **1** |  |  |  |  |  |
| **runtime** | **0.398\*\*\*** | **0.255\*\*\*** | **1** |  |  |  |  |
| **years\_old** | **0.225\*\*\*** | **-0.168\*\*\*** | **0.167\*\*\*** | **1** |  |  |  |
| **popularity** | **0.286\*\*\*** | **0.744\*\*\*** | **0.177\*\*\*** | **-0.154\*\*\*** | **1** |  |  |
| **revenue** | **0.183\*\*\*** | **0.752\*\*\*** | **0.230\*\*\*** | **-0.141\*\*\*** | **0.596\*\*\*** | **1** |  |
| **budget** | **-0.0524\*\*** | **0.532\*\*\*** | **0.225\*\*\*** | **-0.265\*\*\*** | **0.421\*\*\*** | **0.701\*\*\*** | **1** |

**\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001**

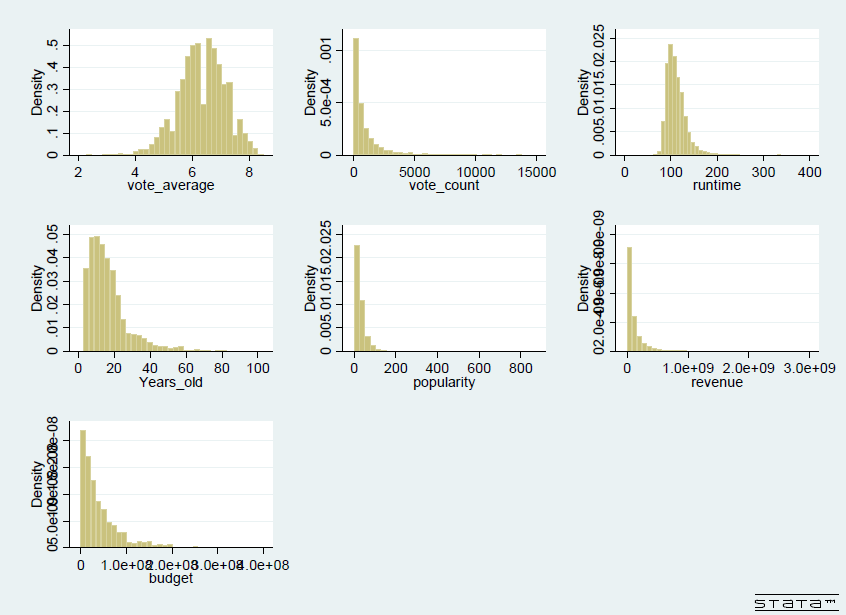
Tabela 2. Macierz korelacji

Z macierzy korelacji możemy odczytać duże zależności między *popularity* i *vote\_average*, *revenue* i *vote\_average*, *budget* i *vote\_average*, *revenue* i *popularity*, *budget* i *popularity* oraz *budget* i *revenue*. Może to wynikać z faktu, że budżet silnie wpływa na dochody (intuicyjnie większy budżet może przekładać się na większe wydatki na reklamę i marketing, a także zwiększać zasięg dystrybucji, a co za tym idzie niezależnie od jakości filmu generować większe dochody), co z kolei silnie wpływa na popularność filmu. Trudno jasno określić wpływ zmiennych na vote\_average. Aby sprawdzić ten wpływ, przeprowadzimy estymację w dalszej części naszej pracy.

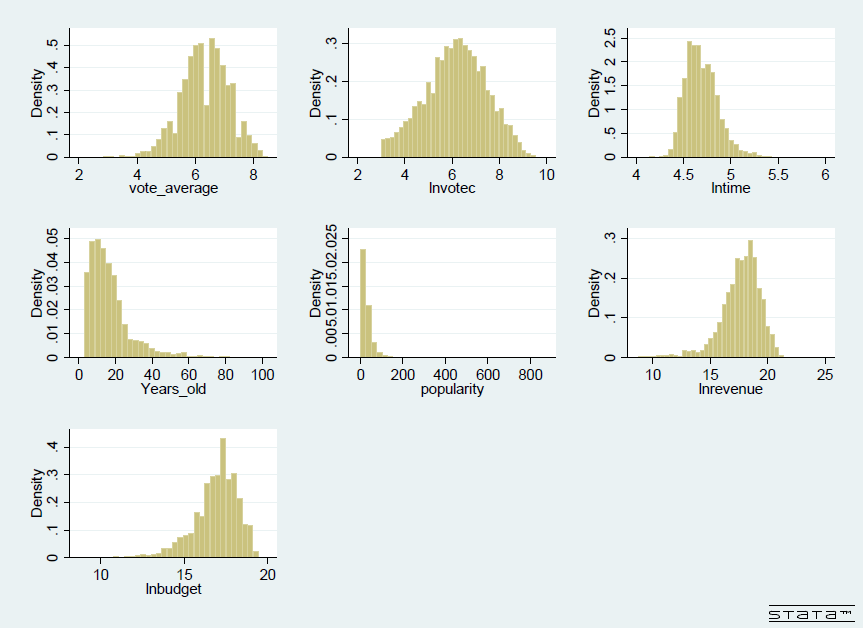
**Histogramy zmiennych**

****

Rysunek 1. Histogram zmiennej objaśnianej *vote\_average*

Histogram z nałożoną gęstością rozkładu normalnego wskazuje, że zmienna vote\_average nie posiada rozkładu normalnego (występuje asymetria). Zlogarytmowanie zmiennej nie poprawia jej histogramu oraz nie wynika z literatury.

Rysunek 2. Histogramy zmiennych objaśniających

Histogramy użytych w modelu zmiennych ciągłych pozwalają stwierdzić, iż prawidłową formą funkcyjną modelu jest forma liniowa. Większość rozkładów jest silnie asymetryczna (obserwacje nietypowe).

Rysunek 3. Histogramy zmiennych ( w tym zlogarytmowanych)

Histogramy zlogarytmowanych wartości, w szczególności *Revenue* i *Budget* (zgodnie z literaturą - logaarytmowanie dochodów), przypominają bardziej rozkład normalny. Ta informacja może przydać nam się przy budowaniu modelu.

**Forma funkcyjna oraz estymacja modelu**

W części empirycznej pracy oszacowana została forma funkcyjna modelu, zawierająca analizowane zmienne. Następnie przeprowadziliśmy regresję, mającą na celu weryfikację postawionych hipotez. Przyjęliśmy następujące równanie:

**vote\_average = β0 + β1lnbudget + β2years\_old + β3lnrevenue + β4runtime + β5vote\_count + β6popularity + εi**

Następnie przeprowadzona została regresja z użyciem wszystkich zmiennych, której wyniki prezentuje poniższa tabela.

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | vote\_average |
|  |  |
| lnbudget | -0.242\*\*\* |
|  | (0.0119) |
| years\_old | 0.00692\*\*\* |
|  | (0.00105) |
| lnrevenue | 0.0483\*\*\* |
|  | (0.00963) |
| runtime | 0.0137\*\*\* |
|  | (0.000609) |
| vote\_count | 0.000224\*\*\* |
|  | (1.32e-05) |
| popularity | 0.000687 |
|  | (0.000484) |
| Constant | 7.693\*\*\* |
|  | (0.177) |
|  |  |
| Observations | 3,112 |
| R-squared | 0.390 |

Standard errors in parentheses

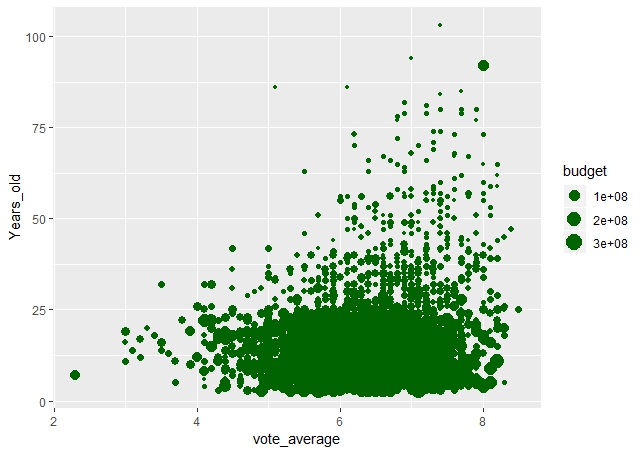
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

W wyniku pierwszej regresji wszystkie zmienne poza *popularity* okazały się istotne. R2 wyniosła 0.39. Test RESET odrzucił jednak hipotezę zerową o tym, że liniowa specyfikacja modelu jest właściwa, dlatego postanowiliśmy wprowadzić zmiany w modelu.

Zmienną runtime zamieniliśmy na logarytmowaną zmienną *lntime,*a zmienną *vote\_count* na logarytmowaną zmienną *lnvotec*. Wyniki drugiej regresji w tabeli poniżej:

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | vote\_average |
|  |  |
| lnbudget | -0.225\*\*\* |
|  | (0.0114) |
| years\_old | 0.0102\*\*\* |
|  | (0.00101) |
| lnrevenue | -0.0428\*\*\* |
|  | (0.0105) |
| lntime | 1.822\*\*\* |
|  | (0.0694) |
| lnvotec | 0.334\*\*\* |
|  | (0.0140) |
| popularity | 0.00134\*\*\* |
|  | (0.000400) |
| Constant | 0.0652 |
|  | (0.321) |
|  |  |
| Observations | 3,112 |
| R-squared | 0.446 |

Jak pokazuje wynik regresji, wszystkie zmienne okazał się istotne, a R2 wzrosła, jednak Test RESET wciąż odrzucał hipotezę zerową o tym, że liniowa specyfikacja modelu jest właściwa. Zbadaliśmy więc zmienne o wprowadziliśmy kolejne zmiany. Obrazek poniżej pokazuje zależność między wiekiem filmu, a wysokością ocen.

****

Rysunek 4. Wykres vote\_average i Years\_old

Jak widać, zdecydowanie więcej filmów w bazie to filmy młodsze niż 25 lat. Co więcej, wraz z przyrostem lat oceny produkcji są coraz wyższe, co może wskazywać na obecność w bazie tylko tych filmów, które zyskały miano “kultowych”. Ze względu na tę obserwację, a także na fakt wielu osobliwych wartości ocen w przedziale (0;85) zmiennej *vote\_count,* przeprowadziliśmy regresję z pominięciem filmów staraszych niż 25 lat i ocenionych przez mniej niż 85 użytkowników. Wyeliminowaliśmy w ten sposób 775 filmów. Dodaliśmy także zmienną *yearsquared,* oznaczającą wiek podniesiony do kwadratu.

Wyniki regresji poniżej:

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | vote\_average |
|  |  |
| lnbudget | -0.249\*\*\* |
|  | (0.0140) |
| years\_old | 0.0547\*\*\* |
|  | (0.0107) |
| yearsquare | -0.00121\*\*\* |
|  | (0.000398) |
| lnrevenue | -0.0515\*\*\* |
|  | (0.0133) |
| lntime | 1.889\*\*\* |
|  | (0.0839) |
| lnvotec | 0.406\*\*\* |
|  | (0.0186) |
| popularity | 0.00113\*\*\* |
|  | (0.000415) |
| Constant | -0.474 |
|  | (0.392) |
|  |  |
| Observations | 2,337 |
| R-squared | 0.429 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Wszystkie zmienne okazały się istotne, R2 nieznacznie spadła. Następnie przeprowadziliśmy test RESET:

**Ramsey RESET test using powers of the fitted values of vote\_average**

**Ho: model has no omitted variables**

**F(3, 2326) = 1.99**

**Prob > F = 0.1130**

Wynik testu odrzuca hipotezę alternatywną o tym, że liniowa specyfikacja modelu nie jest właściwa. Będziemy więc interpretować wyniki regresji powyżej.

Przeciwne znaki *years\_old* i *yearsquare,* które będziemy potem interpretować, zainspirowały nas jednak do kolejnej modyfikacji zmiennych, w celu porównania efektu wieku filmu.

Utworzyliśmy trzy zmienne binarne według grup wiekowych:

*year10 =*

*year25 =*

*year100 =*

Poniżej wyniki regresji przeprowadzonej na zmiennych binarnych na zbiorze początkowym (bez eliminacji filmów starszych niż 25 lat) (zmienne *years\_old* i *yearsquare* zastąpione zmiennymi binarnymi):

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | vote\_average |
|  |  |
| lnbudget | -0.244\*\*\* |
|  | (0.0113) |
| lnrevenue | -0.0438\*\*\* |
|  | (0.0104) |
| lntime | 1.890\*\*\* |
|  | (0.0684) |
| lnvotec | 0.340\*\*\* |
|  | (0.0141) |
| popularity | 0.00153\*\*\* |
|  | (0.000400) |
| o.year10 | - |
|  |  |
| year25 | 0.226\*\*\* |
|  | (0.0258) |
| year100 | 0.362\*\*\* |
|  | (0.0378) |
| Constant | 0.0494 |
|  | (0.320) |
|  |  |
| Observations | 3,112 |
| R-squared | 0.449 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Zmienne *year25* i *year100* są istotne statystycznie, a także mają duże wartości współczynników β, co sugeruje dodatni wpływ wieku na ocenę filmu. Z wcześniejszej analizy wartości wiemy jednak, że filmy mające ponad 25 lat cechowały się wyższymi ocenami całej grupy niż młodsze filmy, ze względu na brak odczytów z niskimi ocenami. Sugerowanie dodatniego wpływu z przedziałów wiekowych zawartych w bazie, jest więc błędne, a wyniki regresji ze zmienną *yearsquare* wydają się być bardziej intuicyjne.Potwierdza to wynik testu RESET, który odrzuca hipotezę zerową o poprawności funkcyjnej modelu.

**Testy diagnostyczne**

W tej części pracy zostały przedstawione wyniki przeprowadzonych testów diagnostycznych.

1. **Test na współliniowość**

Po przeprowadzeniu estymacji został wykonany test na współliniowość.

|  |  |  |
| --- | --- | --- |
| **Variable** | **VIF** | **1/VIF** |
| **years\_old** | **23.58** | **0.042413** |
| **yearsquare** | **23.08** | **0.043322** |
| **lnrevenue** | **2.66** | **0.375827** |
| **lnvotec** | **2.53** | **0.395525** |
| **lnbudget** | **1.88** | **0.532851** |
| **popularity** | **1.71** | **0.586321** |
| **lntime** | **1.16** | **0.858637** |
| **Mean VIF** | **8.09** |  |

Na podstawie wyników testu VIF (testu na współliniowość) możemy uznać, że zmienne nie są ze sobą skorelowane. Jedynie wartości *years\_old* i *yearsquare* są powyżej 10, ale nie budzi to żadnych podejrzeń w związku z celową korelacją tych zmiennych. W modelu nie występuje więc problem współliniowości.

1. **Test RESET**

**Ramsey RESET test using powers of the fitted values of vote\_average**

**Ho: model has no omitted variables**

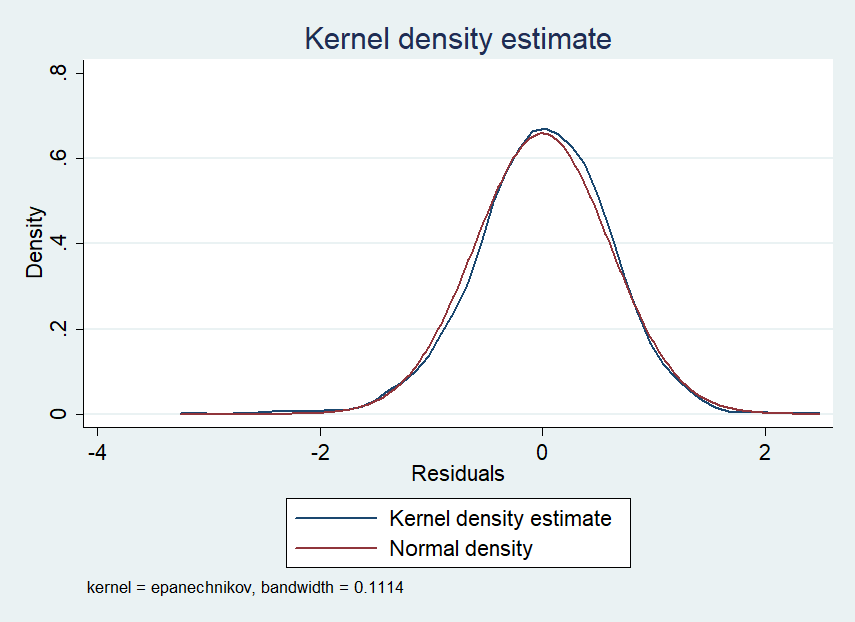
**F(3, 2326) = 1.99**

**Prob > F = 0.1130**

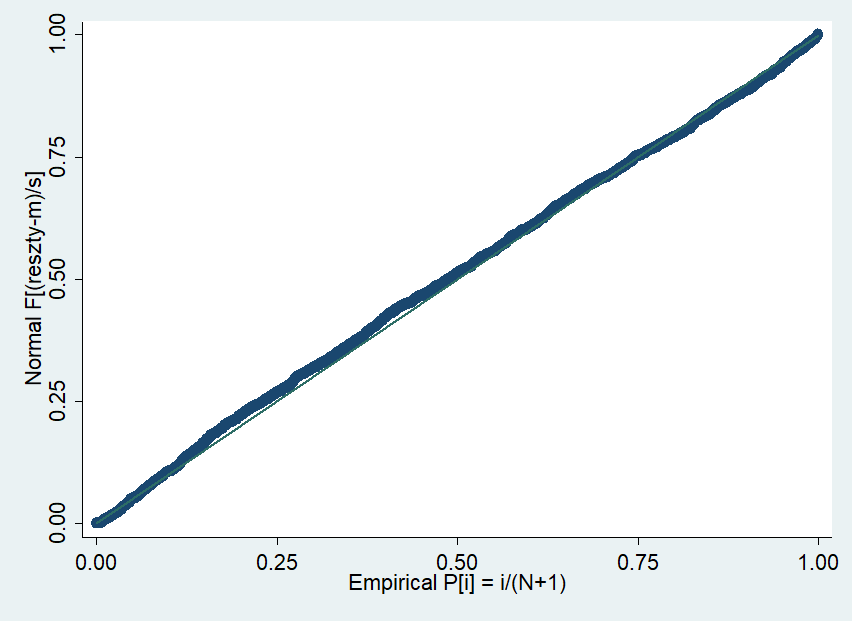
1. **Test na normalność składnika losowego (analiza graficzna)**

H0: Składnik losowy ma rozkład normalny.

H1: Składnik losowy nie ma rozkładu normalnego.

****

Rysunek 5. Test na normalność składnika losowego

****

Z analizy graficznej wynika, że nie ma podstaw do odrzucenia hipotezy zerowej o normalności składnika losowego.

**Weryfikacja hipotez**

Hipotezy o wpływie liczby ocen, popularności i wieku filmu potwierdziły się. Wbrew naszej pierwotnej intuicji, wyniki regresji odrzuciły hipotezy o dodatnim wpływie dochodu i budżetu, a także o ujemnym wpływie długości filmu na ocenę filmu.

**Interpretacja wyników estymacji**

Wyestymowane parametry interpretujemy jako:

**β1 -** wzrost budżetu o 1% przekłada się na spadek oceny o 0.00249 jednostki

**β2 -** wzrost wieku filmu o rok przekłada się na wzrost oceny o 0.05468 jednostki

***yearsquare*** - vote\_average = -0.0012127yearsquare + 0.0546784years\_old

Jest to parabola z ramionami skierowanymi do dołu, czyli wraz ze wzrostem wieku ocena rośnie, ale coraz wolniej. Maksimum funkcji jest osiągane dla filmu w wieku powyżej 22 lat.

Dla filmu powyżej 22 lat zależność między wiekiem a średnią oceną filmu jest ujemna – ocena maleje coraz szybciej wraz ze wzrostem wieku.

**β3 -** wzrost dochodu o 1% przekłada się na spadek oceny o 0.00051 jednostki

**β4 -** wzrost długości czasu trwania filmu o 1% przekłada się na wzrost oceny o 0.01889 jednostki

**β5 -** wzrost liczby ocen filmu o 1% przekłada się na wzrost oceny o 0.00406 jednostki

**β6 -** wzrost popularności o 1 przekłada się na wzrost oceny o 0.00113 jednostki

**Podsumowanie i wnioski**

Zaprezentowany w pracy model wskazał czynniki determinujące średnią ocenę filmów na podstawie ocen z TMDb. Wyniki badania pokazały, że wszystkie zmienne okazały się istotne. Potwierdzono, że liniowa specyfikacja modelu jest właściwa i wyjaśnia on zmienność zależnej w 42.94%. Wartość dopasowanego R2  jest równa 42.77%.

W drodze przeprowadzonych dyskusji i analizy wyników modelu można wyciągnąć wnioski, że w analizie średniej oceny filmów warto uwzględnić nowe czynniki jak liczba lub sam fakt zdobycia prestiżowych nagród, a także rozważyć analizy na różnych szeregach czasowych (dodając na przykład występujące osoby z obsady i ekipy filmowej, co nie byłoby istotne przy analizie w tak dużym, analizowanym w badaniu, przedziale czasowym.

**Bibliografia**

* + - 1. <https://www.themoviedb.org/>
      2. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/asi.23213>
      3. Prag, J., & Casavant, J. (1994). An empirical study of the determinants of

revenues and marketing expenditures in the motion picture industry.

Journal of Cultural Economics, 18(3), 217–235

1. <https://techpolicyinstitute.org/2017/09/15/rotten-tomatoes-and-imdb-reviews-strongly-correlated-with-movie-revenues/?fbclid=IwAR0KdbKEMUZe0XVYbbAMXf-SgoI5oA21SpaQ8UsUgpGGllU-7LuN9HO33Ic>
2. Petrykowski, K. (2016) Micro and macro film rating effects as marketing strategies. Journal of film and video, 17(2) , 32-46
3. <https://www.researchgate.net/publication/305800376_Do_MPAA_Ratings_Affect_Box_Office_Revenues>
4. <https://businessinsider.com.pl/lifestyle/produkcja-filmow-biznes-i-zyskiwywiad-z-krzysztofem-terejem/ymzcm30?fbclid=IwAR0x6Ke8nFGW-mbZRk4tOP0YO4SwFLa0W8TkluWRJif16bVfeaFdqBA5uz4>

1. Według badania przeprowadzonego przez BrightLocal, 88 procent konsumentów ufa opiniom w internecie w takim samym stopniu, jak rekomendacjom przekazywanym na żywo. [↑](#footnote-ref-1)