Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

Mateusz Gomulski

Nr albumu: 293646

And the Oscar goes to...
Analiza czynników ekonomicznych i pozaekonomicznych wpływających na prawdopodobieństwo otrzymania Oscara

Praca na kierunku: Informatyka i Ekonometria

Praca wykonana pod kierunkiem
mgr. Rafała Woźniaka
z Katedry Statystyki i Ekonometrii
WNE UW

Warszawa, czerwiec 2013

Spis treści

Streszczenie

Niniejsza praca podejmuje tematykę czynników wpływających na prawdopodobieństwo zdobycia Oscara. Pierwsze rozdziały poświęcone są przeglądowi literatury dotyczącej problemu, postawieniu hipotez badawczych oraz zdefiniowaniu zmiennych i zbioru danych. W kolejnych częściach pracy zaprezentowane są trzy modele ekonometryczne mogące posłużyć do oszacowania badanego zjawiska, są nimi: Liniowy Model Prawdopodobieństwa, model logitowy i model probitowy. Dla modeli najlepiej reprezentujących dane wejściowe przeprowadzana jest interpretacja i weryfikacja hipotez badawczych. Ostatni rozdział niniejszej pracy skupia się na wyciągnięciu odpowiednich wniosków z przeprowadzonych badań.

Wstęp

Każdego roku od ponad 80 lat Amerykańska Akademia Sztuki i Wiedzy Filmowej (w dalszej części pracy zwana Akademią) przyznaje najbardziej prestiżowe w świecie filmu nagrody - Oscary. Początkowo statuetki przyznawane były w 13 kategoriach, obecnie liczba kategorii wzrosła do 24. Są nimi między innymi: najlepszy film, najlepszy reżyser, najlepszy scenariusz adaptowany, najlepszy scenariusz oryginalny, najlepszy aktor, najlepsza aktorka, . Na początku każdego roku blisko 6 tysięcy członków Akademii wybiera w tajnym głosowaniu po pięć filmów w każdej kategorii, aby nominować je do nagrody. W kolejnej turze głosowania spośród nominowanych wyłaniani są zwycięzcy poszczególnych kategorii, a ich ogłoszenie ma miejsce na oficjalnej gali. Gala rozdania Oscarów jest jednym z najważniejszych wydarzeń kulturalnych nie tylko w Stanach Zjednoczonych, ale i na całym świecie. Z zapartym tchem co roku śledzą ją setki milionów ludzi, informują o niej praktycznie wszystkie serwisy informacyjne świata.

Kwestia tego kto otrzyma Oscara od wielu lat budzi duże zainteresowanie zarówno mediów, jak i ogromnej rzeszy ludzi. Jest to temat wielu dyskusji w czasie tygodni poprzedzających galę, pojawiają się liczne teorie na ten temat, a nawet przyjmowane są zakłady bukmacherskie. Mimo to rozstrzygnięcia w poszczególnych kategoriach są często dla opinii publicznej ogromną niespodzianką. Stąd też wynika potrzeba oszacowania czynników, które mogą wpływać na prawdopodobieństwo otrzymania Oscara. Celem niniejszej pracy będzie więc skonstruowanie i przeanalizowanie odpowiedniego modelu ekonometrycznego oraz wyznaczenie zmiennych, które w sposób istotny mogą wpływać na szanse oskarowe filmu. Jest to zagadnienie rzadko pojawiające się w literaturze, nie powstało zbyt wiele publikacji na ten temat, z tego też powodu wydaje się bardzo interesujące pod kątem badawczym.

Na prawdopodobieństwo zdobycia Oscara wpływa niewątpliwie mnóstwo niemierzalnych czynników takich jak gra aktorska, klimat filmu czy ciekawość scenariusza. Istnieje jednak szereg czynników mierzalnych, które mogą zostać użyte w badaniu. Można je podzielić na trzy główne kategorie: zmienne ekonomiczne (np. budżet czy przychody filmu), zmienne charakteryzujące film (np. czas trwania, gatunek) i zmienne weryfikujące jakość filmu (liczba nagród/nominacji w innych konkursach). Wpływ tych trzech typów zmiennych na prawdopodobieństwo otrzymania Oscara będzie podlegał badaniu ekonometrycznemu przedstawionemu w niniejszej pracy. Naturalnym w tej sytuacji wydaje się posłużenie się jednym z modeli wykorzystującym binarną zmienną zależną. Jako sukces (oznaczymy przez 1) przyjmiemy zdobycie choćby jednej statuetki i porażkę (oznaczymy przez 0) zdefiniujemy jako nie zdobycie żadnej. W poniższej pracy przeanalizuję i porównam trzy modele binarnej zmiennej zależnej: Liniowy Model Prawdopodobieństwa, model Logit i model Probit. Z modeli tych wybiorę ten, który najlepiej będzie reprezentował dane.

Rozdział 1

Podstawy teoretyczne modelu

W rozdziale tym opisane zostaną wcześniejsze najbardziej znane badania ekonometryczne poruszające zagadnienie modelowania prawdopodobieństwa zdobycia nagrody Amerykańskiej Akademii Sztuki i Wiedzy Filmowej. Szczególny nacisk zostanie położony na wskazanie celów poszczególnych badań, hipotez badawczych, metod estymacji, sposobu doboru danych oraz ostatecznych wniosków. W dalszej części rozdziału zaprezentowane zostaną hipotezy badawcze niniejszej pracy będące konsekwencją analizy wcześniej przytoczonej literatury oraz wyrosłe z wiedzy teoretycznej. Każda hipoteza zostanie szczegółowo opisana i przeanalizowana, tak by móc ją poddać stosownej weryfikacji w finalnej części pracy.

1.1 Literatura dotycząca problemu

Iain Pardoe i Dean K. Simonton w artykule *Applying discrete choice models to predict Academy Award winners* [9] opisali swoje badania dotyczące prawdopodobieństwa otrzymania Oscara w czterech kategoriach: najlepszy film, najlepszy reżyser, najlepszy aktor pierwszoplanowy i najlepsza aktorka pierwszoplanowa. W swoich badaniach wykorzystali oni dane portalu *us.imdb.com* dotyczące filmów nominowanych do Oscara w latach 1938-2006. Celem ich badań było przewidzenie zwycięzców Oscarów w poszczególnych kategoriach w latach 1938-2006 wykorzystując informacje dotyczące nominowanych filmów/osób dostępne przed ogłoszeniem werdyktu Amerykańskiej Akademii Filmowej. W tym celu posłużono się modelem **mixed logit**, który jest rozszerzeniem klasycznego logitu wielomianowego. Głównymi hipotezami badawczymi w artykule I. Pardoe'a i D.K. Simonton'a były: na prawdopodobieństwo zdobycia Oscara w czterech głównych kategoriach wpływa liczba nominacji i liczba wygranych Złotych Globów w tożsamych kategoriach, na szansę zdobycia Oscara dla najlepszego reżysera ma wpływ to czy był on już nim nagradzany i czy dostał nagrodę *Directors Guild of America Award* [1], większe prawdopodobieństwo otrzymania Oscara ma aktor/aktorka, która była już wcześniej nominowana/wygrała tę nagrodę. Autorzy w swym badaniu wykorzystali następujące zmienne niezależne: całkowita liczba nominacji, nominacja dla reżysera, nominacja dla najlepszego filmu, wygrany Złoty Glob w kategorii dramat, wygrany Złoty Glob w kategorii komedia lub musical, wygrana przez reżysera nagroda Gildii, poprzednie oscarowe wygrane i nominacje dla ludzi zaangażowanych w film. Główne wnioski wypływające z omawianego w tym paragrafie artykułu wskazują na fakt, iż czynniki wpływające na zdobycie Oscara w poszczególnych kategoriach

[1] Directors Guild of America Award to nagroda przyznawana przez Amerykańską Gildię Reżyserów dla najlepszego reżysera danego roku, jej wręczenie ma miejsce około miesiąc przed galą rozdania Oscarów

zmieniają się w czasie. Co nie zmieniło się od 1938 roku to pozytywny wpływ liczby nominacji (kształtował się on na poziomie około 0,5%) oraz zdobycia Złotych Globów w tożsamych kategoriach (szacowany wpływ - między 2, a 4%). Badanie przeprowadzone w przytoczonym w tym rozdziale artykule pozwoliło z zadowalającym wynikiem zakwalifikować zwycięzców w kategoriach: najlepszy aktor(77% dobrych trafień), najlepszy film (70%), najlepsza aktorka (77%) oraz najlepszy film (93%).

Andrew B. Bernard w swoim artykule *An Index of Oscar-Worthiness: Predicting the Academy Award for Best Picture* [1] starał się zbadać czynniki, które mogą wpłynąć na prawdopodobieństwo zdobycia Oscara w najbardziej prestiżowej kategorii, czyli najlepszy film. Do tego celu zebrał on informacje o filmach nominowanych do Oscara z lat 1984-2003. Zmienne w badaniu podzielił na 2 kategorie: charakteryzujące film i mierzące liczbę wyróżnień filmu. Oprócz zmiennych przytoczonych już w artykule omawianym w poprzednim paragrafie autor wykorzystał takie zmienne jak: pochodzenie głównego bohatera, niepełnosprawność głównego bohatera, nieszczęśliwa miłość, ekranizacja, czy wojenny charakter filmu. Po wstępnym oszacowaniu modelu **probit** okazało się, że zmiennymi istotnymi w badaniu są: całkowita liczba nominacji do Oscara, całkowita liczba wygranych Złotych Globów, wygranie Złotego Globa w kategorii najlepszy film, pochodzenie głównego bohatera, zero-jedynkowa zmienna wskazująca czy główny bohater jest geniuszem, oraz zero-jedynkowa zmienna wskazująca czy film jest komedią czy nie. Najlepiej przewidującym zdobycie Oscara (18 na 20 poprawnie zakwalifikowanych zwycięzców) okazał się model z trzema zmiennymi niezależnymi: całkowitą liczbą nominacji do Oscara (każda kolejna nominacja zwiększa prawdopodobieństwo otrzymania Oscara o 4,5%),całkowitą liczbą wygranych Złotych Globów (każdy kolejny Złoty Glob zwiększa prawdopodobieństwo otrzymania Oscara o 10,2%) oraz zero-jedynkową zmienną wskazującą czy film jest komedią czy nie (fakt bycia komedią w 100% przewidywał porażkę).

Ostatnim artykułem, który zostanie omówiony w tej części rozdziału jest artykuł Davida Kaplana *And the Oscar Goes to…A Logistic Regression Model for Predicting Academy Award Results* [4], który podobnie jak praca Andrew B. Bernard'a skupia się na modelowaniu prawdopodobieństwa otrzymania Oscara w kategorii najlepszy film. Kaplan do swoich badań wykorzystał dwie bazy danych dotyczące filmów z lat 1966-2006: *Videohound's Golden Movie Retriever* i *Internet Movie Database (IMDB)*. Jako model badawczy przyjął model **logit** analizując informacje dotyczące 200 filmów, dla których utworzył on 22 zmienne niezależne. Tylko 9 zmiennych z pierwotnych 22 okazało się istotne w badaniu i mogło służyć do właściwego wnioskowania, były nimi (w nawiasach ilorazy szans poszczególnych zmiennych): zero-jedynkowa zmienna wskazująca czy film jest biografią (0,0167), zero-jedynkowa zmienna wskazująca czy film miał najwięcej nominacji w stawce (15,7824), zero-jedynkowa zmienna wskazująca czy film wygrał Złotego Globa w kategorii komedia/musical (7,1135), zero-jedynkowa zmienna wskazująca czy film wygrał Złotego Globa w kategorii dramat (36,8664), zero-jedynkowa zmienna wskazująca czy reżyser dostał Złotego Globa (0,1816), zero-jedynkowa zmienna wskazująca czy film był wybitny i był biografią (645,6128), liczba poprzednich nominacji reżysera do Oscara (0,3599), zero-jedynkowa zmienna wskazująca czy reżyser filmu dostał za ten film nagrodę Amerykańskiej Gildii Reżyserów (307,3847) oraz liczba Oscarów dla najlepszego aktora/aktorki, które obsada zdobyła przed nakręceniem danego filmu (0,3046). Pseudo R^2 modelu, którym posługiwał się David Kaplan wyniosło około 0,5. Wyniki badania przeprowadzonego przez autora artykułu omawianego w tym podrozdziale potwierdzają główne wnioski płynące z pozostałych artykułów - najbardziej istotnym czynnikiem wpływającym na prawdopodobieństwo otrzymania Oscara są poprzednie nagrody, które uzyskał film, lub osoby tworzące film.

Mankamentem większość badań poruszających zagadnienie czynników wpływających na szanse oskarowe filmu jest fakt, iż koncentrują się one na konkretnej kategorii nagrody nie

starając się oszacować modelu ogólnego, czyli prawdopodobieństwa otrzymania Oscara w jakiejkolwiek kategorii. W badaniach pomijane są również czynniki ekonomiczne takie jak budżet filmu, przychody z kas kinowych,czy zyskowność filmu. Najczęściej próbką testową dla badaczy są filmy nominowane do nagrody Akademii, a nie losowa próbka filmów, co zdecydowanie zawęża pole wnioskowania i zuboża analizy. W niniejszej pracy postaram się wypełnić tą niewątpliwą lukę badawczą tworząc model dla losowej próbki filmów, będący uogólnieniem innych badań - skupiając się na prawdopodobieństwie otrzymania Oscara bez podziału na kategorię oraz umieszczając w modelu zmienne niezależne o charakterze ekonomicznym.

1.2 Hipotezy badawcze

Analizując literaturę przytoczoną w poprzedniej części tego rozdziału można dojść do wniosku, iż istnieje co najmniej kilka czynników istotnie wpływających na prawdopodobieństwo otrzymania Oscara w poszczególnych kategoriach nagrody. Większość z nich wpływa dodatnio na badane zjawisko. Celem niniejszej pracy będzie weryfikacja czy wnioski wynikające z literatury można uogólnić na wszystkie kategorie oskarowe jako całość oraz znalezienie innych istotnych czynników wpływających na modelowane prawdopodobieństwo. Wiąże się to z odpowiedzią na następujące pytania badawcze:

- Czy gatunek wiodący filmu ma wpływ na jego szanse oskarowe?

  Hipoteza: Literatura i analiza danych historycznych wskazuje, iż niektóre gatunki filmowe są preferowane przez członków Akademii (np. dramat), inne zaś prawie nigdy nie zdobywają statuetki (komedia). Co skłania do postawienia hipotezy, iż taki wpływ istnieje.

- Czy wielkość budżetu może wpłynąć na prawdopodobieństwo zdobycia najważniejszej statuetki w świecie filmowym?

  Hipoteza: Ze względu na charakter niektórych kategorii (np. najlepszy krótkometrażowy film animowany) budżet nie powinien istotnie wpływać na modelowane prawdopodobieństwo. Na fakt ten wskazuje również Simonton D. K. w swoim artykule *Cinematic creativity and production budgets: does money make the movie?*[11].

- Czy film będący ekranizacją ma większe szanse na Oscara?

  Hipoteza: Filmy uhonorowane największą liczbą statuetek to zazwyczaj filmy będące adaptacją jakiejś znanej powieści, czy sztuki. Stąd też hipoteza, iż czynnik ten powinien wpływać na szacowane zjawisko.

- Czy przychody z kas kinowych lub inne mierniki sukcesu ekonomicznego filmu zwiększają prawdopodobieństwo zdobycia Oscara?

  Hipoteza: Przychody z kas kinowych są pewnym rodzajem głosowania ludzi "przez portfele". Im wyższe przychody z kas do momentu głosowania przez członków Akademii tym większe szanse na zwycięstwo w choć jednej kategorii powinien mieć film.

- Czy amerykańskie filmy mają większe szanse oskarowe niż filmy spoza ameryki?

  Hipoteza: Ze względu na miejsce odbywania się gali i narodowość osób głosujących filmy ze Stanów Zjednoczonych powinny mieć większe szanse w walce o Oscara niż filmy z poza USA.

- Jak na prawdopodobieństwo zdobycia Oscara wpływa liczba zdobytych nominacji?

  Hipoteza: Zarówno logika, jak i literatura wskazują wyraźnie, iż im więcej film ma nominacji tym większą ma szansę na zdobycie choć jednej statuetki.

- Czy liczba zdobytych statuetek w innych konkursach filmowych takich jak Złote Globy czy BAFTA wpływa na prawdopodobieństwo zdobycia najważniejszej z nagród?

  Hipoteza: Większość przeprowadzonych w tym obszarze badań wskazuje , iż taka zależność istnieje i jest silna, szczególnie w przypadku Złotych Globów, które są przyznawane około miesiąc przed Oscarami.

- Czy wątek miłosny w filmie może zwiększyć jego szanse oskarowe?

  Hipoteza: Wydaje się, iż nie powinien to być czynnik istotny w przypadku tak prestiżowych i profesjonalnych nagród filmowych jakimi są Oscary.

- Czy czas trwania filmu może wpłynąć na prawdopodobieństwo zdobycia przez niego statuetki?

  Hipoteza: Czas trwania filmu jest typowo technicznym parametrem, który nie powinien wpływać na jego ocenę.

Rozdział 2

Opis zbioru danych i definicje zmiennych

W pierwszej części rozdziału przedstawiony zostanie sposób zgromadzenia bazy danych, najważniejsze informacje o niej oraz najważniejsze źródła pochodzenia danych. W dalszej części rozdziału szczegółowo opisana zostanie każda zmienna użyta w badaniu, łącznie z jej najważniejszymi statystykami opisowymi.

2.1 Baza danych

Baza danych zawiera 1663 obserwacje 20 zmiennych, z czego tylko 12 najważniejszych zostało użytych w modelowaniu ekonometrycznym. Użyte zostały zmienne, które wydawały się najbardziej istotne w obliczu teorii i mogły dać najciekawsze wnioski badawcze. Sposób doboru filmów do bazy danych był pseudolosowy, o wyborze filmu decydował algorytm napisany przez autora niniejszej pracy. Niestety z powodu braków danych i sposobu ich gromadzenia nie udało się uzyskać czystej losowości. Informacje o filmach zostały sczytane z następujących stron internetowy:

- http://www.imdb.com/,
- http://www.filmweb.pl/,
- http://www.boxofficemojo.com/,
- http://www.oscars.org/awardsdatabase,
- http://www.boxoffice.com/,
- http://www.cinematoria.com/.

W związku z tym, iż żadne z powyższych źródeł nie udostępnia swojej bazy danych, informacje o każdym filmie musiały zostać zebrane oddzielnie poprzez sczytanie ich przez program komputerowy z odpowiednich miejsc na stronach internetowych. Co oczywiste otrzymany zbiór danych nie były nigdy wykorzystywany w innych badaniach – został specjalnie zgromadzony na potrzeby modelu. W modelu wykorzystano 1638 obserwacji ze znajdujących się w bazie 1663 ze względu na braki danych niektórych zmiennych.

2.2 Charakterystyka zmiennych

2.2.1 Zmienna zależna

➤ **oscar**

Zmienna binarna, przyjmująca następujące wartości:

1 - gdy film lub osoba związana z filmem zdobyła Oscara

0 - w przeciwnym wypadku

Tabela 1. Charakterystyki zmiennej zależnej oscar.

oscar	Freq.	Percent	Cum.
0	1,441	86.65	86.65
1	222	13.35	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Jak widać w bazie przeważają filmy, które nie zdobyły żadnego Oscara, jest ich 1441 czyli blisko 87% wszystkich obserwacji. Tylko 13% filmów (222 z 1663) zdobyło choć jedną statuetkę.

2.2.2 Zmienne niezależne

Zgodnie z tym co zostało napisane we wstępie do niniejszej pracy zmienne niezależne zostały podzielone na 3 główny grupy: zmienne ekonomiczne, zmienne charakteryzujące film i zmienne weryfikujące jakość filmu.

- **Zmienne ekonomiczne:**

  Do tej grupy należą zmienne o charakterze ekonomicznym, mierzące koszty wyprodukowania filmu oraz przychody/zyski które dany film wygenerował. Mają one weryfikować wpływ czynników finansowych na prawdopodobieństwo otrzymania Oscara.

  ➤ **budzet2000**

  Ciągła zmienna ilościowa, powstała poprzez przeliczenie budżetów poszczególnych filmów na dolary w cenach stałych z 2000 roku.

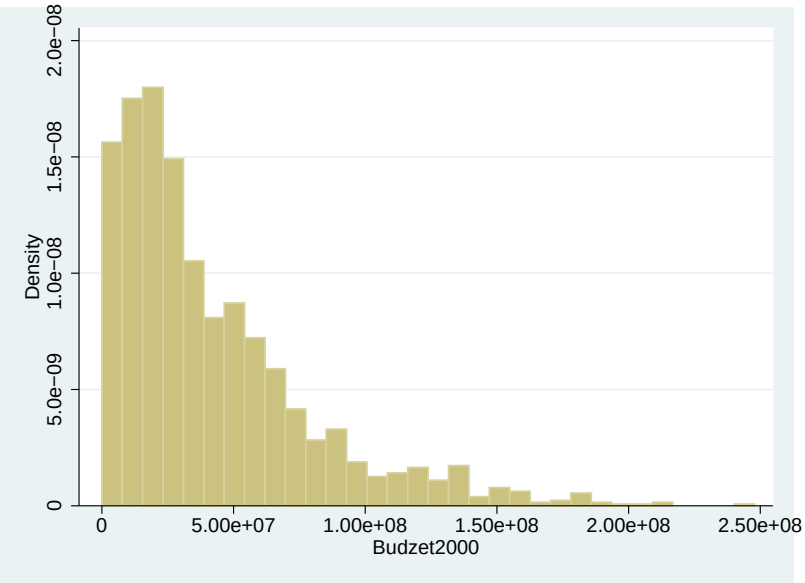
Tabela 2. Charakterystyki zmiennej budzet2000.

Variable	Obs	Mean	Std. Dev.	Min	Max
budzet2000	1644	4.19e+07	3.72e+07	12389.04	2.48e+08

Źródło: Opracowanie własne.

Zmienna budzet2000 przyjmuje wartości w przedziale od około 12 tysięcy do 248 milionów dolarów, ze średnią na poziomie 42 milionów.

Rysunek 1: Histogram zmiennej budzet2000.



Źródło: Opracowanie własne.

➤ **przychody2000**

Ciągła zmienna ilościowa, powstała poprzez przeliczenie przychodów poszczególnych filmów z kas biletowych na dolary w cenach stałych z 2000 roku. Można by przypuszczać, iż na wartość tej zmiennej może wpływać wartość zmiennej zależnej, czyli fakt zdobycia Oscara może implikować wyższe przychody filmu. Jednak ze względu na termin gali oskarowej - koniec lutego, najmłodszy film który może wygrać statuetkę może mieć dwa miesiące, czyli jest już u schyłku swojej kinowej popularności. Fakt ten sprawia, iż wpływ zdobycia Oscara na przychody z kas jest marginalny, co potwierdzają badania E. Helmera przedstawione w artykule *The Value of an Oscar* [3].

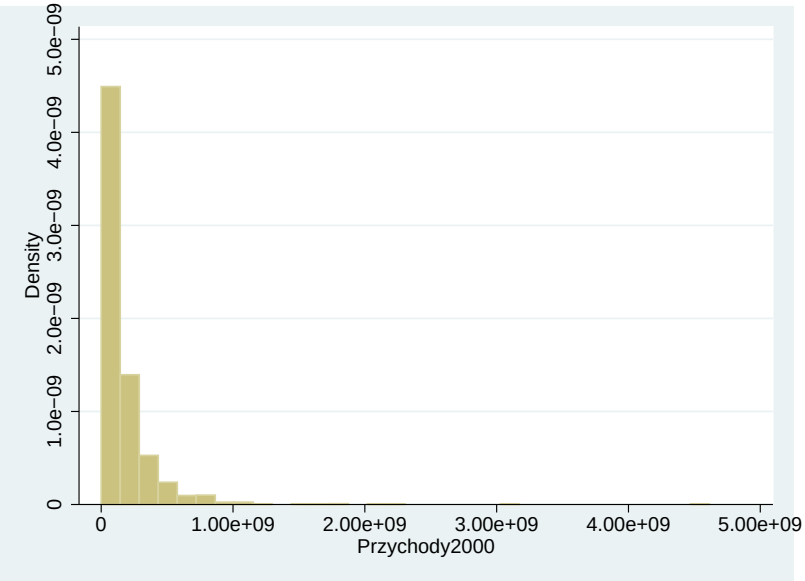
Tabela 3. Charakterystyki zmiennej przychody2000.

Variable	Obs	Mean	Std. Dev.	Min	Max
przycho~2000	1657	1.61e+08	2.41e+08	23714.63	4.61e+09

Źródło: Opracowanie własne.

Średnia wartość zmiennej przychody2000 wynosi około 161 mln dolarów, minimum tej zmiennej wyniosło około 24 tysiące, a maksimum 4,6 miliarda dolarów.

Rysunek 2: Histogram zmiennej przychody2000.



Źródło: Opracowanie własne.

➤ **roi**

Ciągła zmienna ilościowa, jest to wskaźnik zwrotu z inwestycji w film (w %), wyliczony według wzoru:

$$roi = \frac{przychody2000 - budzet2000}{budzet2000} * 100 \tag{1}$$

Tabela 4. Charakterystyki zmiennej roi.

Variable	Obs	Mean	Std. Dev.	Min	Max
roi	1638	1615.718	33486.27	-99.54134	1288939

Źródło: Opracowanie własne.

Zmienna roi przyjmuje wartości od -99.54% do blisko 1288939%, przy średniej 1616%.

- **Zmienne charakteryzujące film:**

Jest to grupa zmiennych, która przedstawia najważniejsze cechy filmu takie jak gatunek, czas trwania czy miejsce wyprodukowania. Są to czynniki niewątpliwie bezpośrednio lub pośrednio brane pod uwagę przez członków Amerykańskiej Akademii Sztuki i Wiedzy Filmowej, gdy decydują się oni na głosowanie na konkretny film.

➤ **gatunek**

Dyskretna zmienna jakościowa przedstawiająca wiodący gatunek filmu. Przyjmuje ona następujące poziomy:

0 - Dramat
1 - Komedia
2 - Film animowany
3 - Film akcji
4 - Fantasy
5 - Sci-Fi
6 - Sensacyjny
7 - Thriller
8 - Horror
9 - Inny gatunek

Tabela 5. Charakterystyki zmiennej gatunek.

gatunek10	Freq.	Percent	Cum.
Dramat	624	37.52	37.52
Komedia	383	23.03	60.55
Animowany	85	5.11	65.66
Akcja	63	3.79	69.45
Fantasy	74	4.45	73.90
Sci-Fi	118	7.10	81.00
Sensacyjny	82	4.93	85.93
Thriller	98	5.89	91.82
Horror	95	5.71	97.53
Inny	41	2.47	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Rozkład poszczególnych gatunków w próbce wydaje się dobrze reprezentować strukturę współczesnej kinematografii. Pierwszy poziom zmiennej (dla dramatu) został przyjęty jako poziom bazowy i nie wprowadzony do modelu, pozostałe zmienne zostały rozkodowane do binarnych postaci: _Igatunek_1,_Igatunek_2, _Igatunek_3, _Igatunek_4, _Igatunek_5, _Igatunek_6, _Igatunek_7, _Igatunek_8, _Igatunek_9, gdzie 1 oznacza, że film należy do danego gatunku, a 0 że nie należy.

➤ **ekranizacja**

Binarna zmienna jakościowa wskazująca czy film jest ekranizacją powieści, artykułu, biografii lub sztuki teatralnej. Zmienna przyjmuje następujące poziomy:

1 - gdy film jest ekranizacją

0 - w przeciwnym wypadku

Tabela 6. Charakterystyki zmiennej ekranizacja.

Ekraniacja	Freq.	Percent	Cum.
0	1,216	73.12	73.12
1	447	26.88	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Ponad jedna czwarta filmów w próbce (26,88%) jest ekranizacją.

➤ **kraj_prod**

Zmienna kraj_prod jest zmienną binarną, jakościową, reprezentującą pochodzenie filmu na zasadzie:

1 - gdy film jest produkcją wyłącznie amerykańską

0 - gdy film jest produkcją spoza Stanów Zjednoczonych, lub z udziałem Stanów jednoczonych i innych krajów

Taki podział pochodzenia filmów został przyjęty ze względu fakt, iż w Stanach Zjednoczonych od zarania kinematografii tworzyło się najwięcej filmów.

Tabela 7. Charakterystyki zmiennej kraj_prod.

kraj_produkcji	Freq.	Percent	Cum.
0	735	44.20	44.20
1	928	55.80	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Tabel 7 przedstawia, iż podział filmów w próbce na wyłącznie amerykańskie i inne jest prawie równomierny, nieznacznie przeważają te pierwsze, jest ich 928, czyli prawie 56%.

➤ **milosc**

Binarna zmienna jakościowa wskazująca czy w filmie pojawia się wątek miłosny, czy też nie. Przyjmuje ona następujące poziomy:

1 - gdy osią filmu jest wątek miłosny

0 - w przeciwnym razie

Tabela 8. Charakterystyki zmiennej milosc.

Milosc	Freq.	Percent	Cum.
0	1,531	92.06	92.06
1	132	7.94	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Jak widać tylko w 132 filmach spośród 1663 osią filmu jest wątek miłosny, stanowi to zaledwie 8% wszystkich filmów w próbce.

➤ **czas_trwania**

Quasi-ciągła zmienna ilościowa reprezentująca czas trwania filmu w minutach.

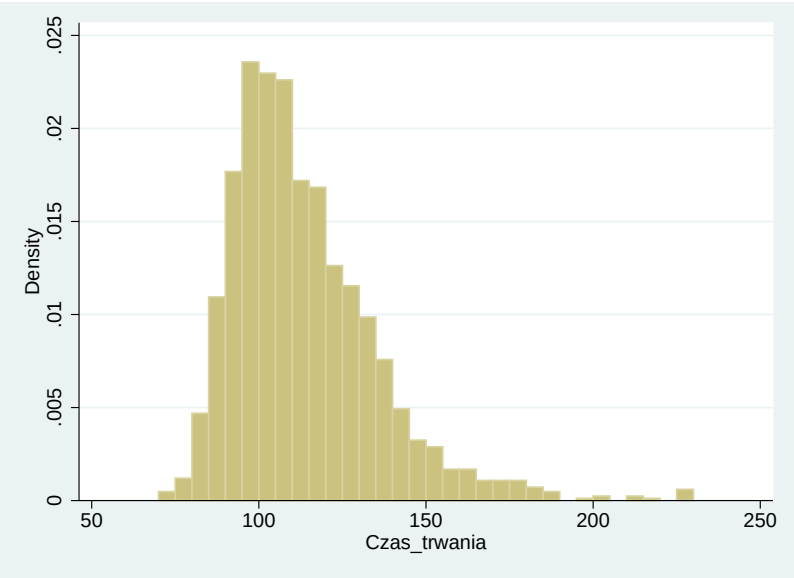
Tabela 9. Charakterystyki zmiennej czas_trwania.

Variable	Obs	Mean	Std. Dev.	Min	Max
czas_trwania	1663	112.9357	21.73786	70	230

Źródło: Opracowanie własne.

Średni czas trwania filmu wynosi blisko 113 minut, najkrótszy film trwa 70 minut, a najdłuższy 230, czyli blisko 4 godziny.

Rysunek 3: Histogram zmiennej czas_trwania.



Źródło: Opracowanie własne.

- **Zmienne weryfikujące jakość filmu:**

Grupa zmiennych reprezentująca uznanie jakie dany film uzyskał w środowisku filmowym, mierzone w liczbie nagród na festiwalach i liczbie nominacji do Oscara. Literatura wskazuje, iż zmienne z tej grupy najbardziej znacząco wpływają na prawdopodobieństwo zdobycia Oscara.

➤ **nominacje**

Zmienna ilościowa przyjmująca wartości ze zbioru liczb naturalnych, reprezentująca liczbę nominacji do Oscara jaką dany film uzyskał.

Tabela 10. Charakterystyki zmiennej nominacje.

Nominacje	Freq.	Percent	Cum.
0	1,126	67.71	67.71
1	206	12.39	80.10
2	77	4.63	84.73
3	51	3.07	87.79
4	47	2.83	90.62
5	29	1.74	92.36
6	20	1.20	93.57
7	29	1.74	95.31
8	26	1.56	96.87
9	15	0.90	97.78
10	14	0.84	98.62
11	11	0.66	99.28
12	6	0.36	99.64
13	5	0.30	99.94
15	1	0.06	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Jak można odczytać z tabli 10 tylko około 32% filmów w bazie uzyskało nominacje do statuetki, z czego większość, bo 12 punktów procentowych uzyskało tylko jedną nominacje. Pod tym względem próbka użyta w badaniu nie jest reprezentatywna, gdyż nominacje do Oscara może uzyskać rocznie maksymalnie 120 filmów/osób związanych z filmami (24 kategorie, 5 nominacji w kategorii), co stanowi promil rocznej światowej produkcji, szczególnie w ostatnich latach gdy przemysł filmowych zaczął się bujnie rozwijać w Indiach i Nigerii. Maksymalna liczba nominacji uzyskanych przez jeden film wynosi 15.

➤ **zlote_globy**

Zmienna ilościowa przyjmująca wartości ze zbioru liczb naturalnych, reprezentująca liczbę Złotych Globów jaką dany film uzyskał.

Tabela 11. Charakterystyki zmiennej zlote_globy.

Zlote_globy	Freq.	Percent	Cum.
0	1,489	89.54	89.54
1	101	6.07	95.61
2	32	1.92	97.53
3	22	1.32	98.86
4	12	0.72	99.58
5	4	0.24	99.82
6	3	0.18	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Tylko 174 spośród 1663 filmów znajdujących się w bazie danych uzyskało co najmniej jednego Złotego Globa, z czego większość (101) uzyskała tylko jedną tego typu nagrodę. Maksymalna liczba statuetek zdobytych przez jeden film wynosi 6.

➤ **bafta**

Zmienna ilościowa przyjmująca wartości ze zbioru liczb naturalnych, reprezentująca liczbę nagród Brytyjskiej Akademii Sztuk Filmowych i Telewizyjnych (BAFTA) jaką dany film uzyskał.

Tabela 12. Charakterystyki zmiennej bafta.

BAFTA	Freq.	Percent	Cum.
0	1,461	87.85	87.85
1	100	6.01	93.87
2	50	3.01	96.87
3	24	1.44	98.32
4	11	0.66	98.98
5	7	0.42	99.40
6	7	0.42	99.82
7	3	0.18	100.00
Total	1,663	100.00	

Źródło: Opracowanie własne.

Podobnie jak w przypadku Złotych Globów, większość filmów w próbce nie zdobyło nagrody BAFTA (87,85%), z tych które zdobyły ponad połowa otrzymała tylko jedną statuetkę. Maksymalna liczba wyróżnień jakie dany film uzyskał wyniosła 7.

Rozdział 3

Oszacowania modeli

W rozdziale tym przedstawione zostaną trzy modele ekonometryczne mogące służyć oszacowaniu czynników wpływających na prawdopodobieństwo zdobycia Oscara. Będą to: Liniowy Model Prawdopodobieństwa (LMP), Model Logit i Model Probit. Z modeli tych zostanie wybrany najlepszy, który potem posłuży do dalszych szczegółowych analiz.

3.1 Liniowy Model Prawdopodobieństwa

W tym podrozdziale oszacowany zostanie najprostszy do wyestymowania model z omawianych, czyli Liniowy Model Prawdopodobieństwa (LMP).

Tabela 13. Ogólny model LMP

```
. xi: reg oscar  budzet2000 i.gatunek ekranizacja roi przychody2000 kraj_prod nominacje zlote_globy bafta
 milosc czas_trwania
i.gatunek         _Igatunek_0-9       (naturally coded; _Igatunek_0 omitted)
```

Source	SS	df	MS		
Model	113.326224	19	5.96453809	Number of obs	= 1638
Residual	76.3935565	1618	.047214806	F(19, 1618)	= 126.33
				Prob > F	= 0.0000
Total	189.71978	1637	.115894795	R-squared	= 0.5973
				Adj R-squared	= 0.5926
				Root MSE	= .21729

oscar	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
budzet2000	-7.00e-11	1.83e-10	-0.38	0.703	-4.30e-10	2.90e-10
_Igatunek_1	-.0109619	.0164266	-0.67	0.505	-.0431815	.0212578
_Igatunek_2	.0653794	.0302081	2.16	0.031	.0061283	.1246306
_Igatunek_3	.0277427	.0301426	0.92	0.358	-.0313799	.0868654
_Igatunek_4	-.0364617	.0286802	-1.27	0.204	-.0927159	.0197925
_Igatunek_5	.0179434	.0240767	0.75	0.456	-.0292814	.0651683
_Igatunek_6	.0440391	.0271553	1.62	0.105	-.0092242	.0973023
_Igatunek_7	.0028486	.0241708	0.12	0.906	-.0445608	.0502579
_Igatunek_8	-.0056704	.0255728	-0.22	0.825	-.0558297	.0444888
_Igatunek_9	-.0388769	.0355595	-1.09	0.274	-.1086245	.0308706
ekranizacja	-.0099141	.0128952	-0.77	0.442	-.0352071	.015379
roi	-1.84e-08	1.62e-07	-0.11	0.910	-3.37e-07	3.00e-07
przychody2000	4.56e-11	2.77e-11	1.65	0.100	-8.70e-12	9.99e-11
kraj_produkcji	-.0277771	.0113187	-2.45	0.014	-.0499779	-.0055763
nominacje	.0812346	.003667	22.15	0.000	.0740421	.0884271
zlote_globy	.0745881	.0108014	6.91	0.000	.053402	.0957743
bafta	.0487341	.0086398	5.64	0.000	.0317878	.0656804
milosc	.0559628	.0201809	2.77	0.006	.0163793	.0955463
czas_trwania	-.0007682	.0003289	-2.34	0.020	-.0014133	-.0001231
_cons	.1058333	.0387433	2.73	0.006	.029841	.1818256

Źródło: Opracowanie własne.

W modelu ogólnym występuje kilka zmiennych nieistotnych na 5%-owym poziomie istotności[2], takich jak: budzet2000, _Igatunek_1, _Igatunek_3, _Igatunek_4, _Igatunek_5, _Igatunek_6, _Igatunek_7, _Igatunek_8, _Igatunek_9, ekranizacja, roi, przychody2000. W związku z tym postanowiono posłużyć się procedurą od ogólnego do szczegółowego, aby kolejno usuwając zmienne z modelu i badając ich łączną nieistotność (przy użyciu testu Walda) uzyskać model zagnieżdżony ze wszystkimi zmiennymi istotnymi na zadanym poziomie istotności. W tym miejscu zostanie pokazany tylko ostatni krok procedury, czyli sprawdzenie łącznej nieistotności wszystkich usuniętych w procedurze zmiennych i model finalny (tabela 14 i 15).

Tabela 14. Test Walda dla wszystkich zmiennych usuwanych

```
. test _Igatunek_7 roi _Igatunek_8 budzet2000 _Igatunek_9 ekranizacja _Igatunek_1  _Igatunek_5 _Igatunek_3
 _Igatunek_4 przychody2000
 ( 1)  _Igatunek_7 = 0
 ( 2)  roi = 0
 ( 3)  _Igatunek_8 = 0
 ( 4)  budzet2000 = 0
 ( 5)  _Igatunek_9 = 0
 ( 6)  ekranizacja = 0
 ( 7)  _Igatunek_1 = 0
 ( 8)  _Igatunek_5 = 0
 ( 9)  _Igatunek_3 = 0
 (10)  _Igatunek_4 = 0
 (11)  przychody2000 = 0
       Constraint 11 dropped
       F( 10,  1618) =    0.67
            Prob > F =    0.7525
```

Źródło: Opracowanie własne.

Statystyka testu Walda na poziomie 0,67 i p-value tego testu równe 0,7525, czyli wskazują, iż nie ma podstaw do odrzucenia hipotezy zerowej o łącznej nieistotności zmiennych, czyli zmienne mogą zostać usunięte z modelu bez obaw o występowanie obciążenia Lovella.

Tabela 15. Zagnieżdżony model LMP

```
. reg oscar _Igatunek_2 _Igatunek_6 kraj_prod nominacje zlote_globy bafta milosc czas_trwania
      Source |       SS       df       MS              Number of obs =    1663
-------------+------------------------------           F(  8,  1654) =  300.73
       Model |  113.994174     8  14.2492717            Prob > F      =  0.0000
    Residual |  78.3702278  1654  .047382242            R-squared     =  0.5926
-------------+------------------------------           Adj R-squared =  0.5906
       Total |  192.364402  1662  .115742721            Root MSE      =  .21767

------------------------------------------------------------------------------
       oscar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igatunek_2 |   .0755959   .0252044     3.00   0.003       .02616    .1250318
  _Igatunek_6 |   .0504197   .0249408     2.02   0.043     .0015008    .0993386
kraj_produkcji |  -.0262216   .0110038    -2.38   0.017    -.0478044   -.0046388
    nominacje |   .0825513   .0034519    23.91   0.000     .0757806    .0893219
  zlote_globy |   .0748217   .0106572     7.02   0.000     .0539186    .0957248
        bafta |   .0490981    .008559     5.74   0.000     .0323105    .0658856
       milosc |   .0495576   .0198606     2.50   0.013     .0106031    .0885122
  czas_trwania |  -.0007373   .0002819    -2.62   0.009    -.0012903   -.0001844
        _cons |   .0990725    .032692     3.03   0.002     .0349504    .1631946
------------------------------------------------------------------------------
```

Źródło: Opracowanie własne.

[2] Autor pracy założył poziom istotności dla wszystkich testów równy 5%

Statyka testu F równa 300,73 i p-value na poziomie 0.0000, czyli mniejsze od zakładanego 5%-owego poziomu istotności sugerują, iż należy odrzucić hipotezę zerową o łącznej nieistotności zmienny objaśniających. Każda ze zmiennych objaśniających ma p-value mniejsze od 5%, co oznacza, że zmienne są również osobno istotne (dla każdej z nich odrzucamy h0 o nieistotności). Model pod tym względem jest więc poprawnie zbudowany. Zmienności zmiennej zależnej jest wyjaśniane w 59,26% przez zmienne niezależne.

Liniowy model prawdopodobieństwa cechują jednak dwie poważne wady: nie ma gwarancji, że wartości dopasowane będą należeć do przedziału [0,1] oraz występuje w nim heteroskedastyczność, która przyczynia się niewłaściwego oszacowania błędów standardowych poszczególnych parametrów, co może wpłynąć na wnioski dotyczące istotności lub nieistotności zmiennych. Tabela 16 przedstawia test Breuscha-Pagana na heroskedastyczność reszt w modelu.

Tabela 16. Test Breuscha-Pagana na heroskedastyczność reszt

```
. hettest,rhs
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: _Igatunek_2 _Igatunek_6 kraj_produkcji nominacje
                    zlote_globy bafta milosc czas_trwania
         chi2(8)      =   806.27
         Prob > chi2  =   0.0000
```

Źródło: Opracowanie własne.

Statystyka testowa równa 806,27 i p-value równe 0,0000 - mniejsze od zakładanego poziomu istotności, skłaniają do odrzucenia hipotezy zerowej o homoskedastyczności w modelu. Rozwiązaniem tego problemu może być oszacowanie modelu z użyciem błędów standardowych odpornych na heteroskedastyczność, czyli zastosowanie macierzy wariancji-kowariancji White'a (polecenie robust w programie Stata).

Tabela 17. Estymacja LMP z zastosowaniem macierzy odpornej White'a

```
. reg oscar _Igatunek_2 _Igatunek_6 kraj_prod nominacje zlote_globy bafta milosc czas_trwania, robust
Linear regression                               Number of obs =    1663
                                                F(  8,  1654) =  150.19
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5926
                                                Root MSE      =  .21767
```

oscar	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Igatunek_2	.0755959	.0329737	2.29	0.022	.0109212	.1402705
_Igatunek_6	.0504197	.0244259	2.06	0.039	.0025107	.0983287
kraj_produkcji	-.0262216	.0111622	-2.35	0.019	-.0481151	-.0043282
nominacje	.0825513	.0057522	14.35	0.000	.0712689	.0938336
zlote_globy	.0748217	.0174394	4.29	0.000	.040616	.1090274
bafta	.0490981	.0149869	3.28	0.001	.0197028	.0784934
milosc	.0495576	.0239683	2.07	0.039	.0025462	.0965691
czas_trwania	-.0007373	.0002821	-2.61	0.009	-.0012906	-.0001841
_cons	.0990725	.0324942	3.05	0.002	.0353384	.1628066

Źródło: Opracowanie własne.

Ze względu na fakt, iż program Stata/IC 11 nie pozwala na zastosowanie komendy *hettest,rhs* dla modelu z macierzą odporną White'a zastosowano inny równoważny test na heteroskeda-

styczność - test White'a (tabela 18).

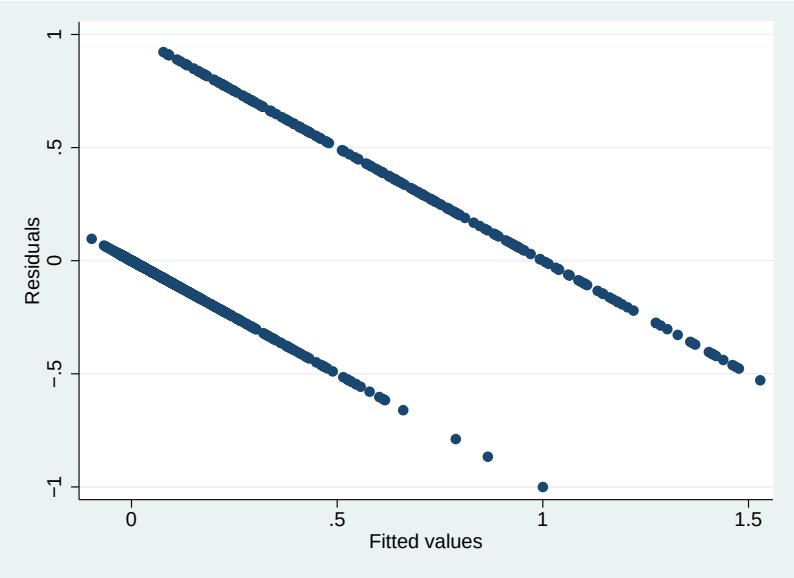
Tabela 18. Test White'a na heteroskedastyczność reszt

```
. whitetst
White s general test statistic :  449.6622  Chi-sq(38)  P-value =  8.3e-72
```

Źródło: Opracowanie własne.

Statystyka testu White'a równa 449,6622 i p-value równe 8.3e-72, czyli zdecydowanie mniejsze niż 5% wskazują, iż problem heteroskedastyczność reszt dalej występuje, potwierdza to również przedstawiony poniżej rysunek 4.

Rysunek 4: Reszty vs. wartości przewidywane



Źródło: Opracowanie własne.

Dla pełności analizy należy sprawdzić jeszcze czy wszystkie wartości dopasowane należą do przedziału [0,1], jeśli nie to jaki ich procent wykracza poza przedział.

Tabela 19. Statystyki dotyczące wartości dopasowanych

```
. predict xb1,xb
. sum xb1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         xb1 |       1663    .1334937    .2618942   -.096738   1.528635

. count if xb1<0 | xb1>1
  425
```

Źródło: Opracowanie własne.

Jak wynika z analizy tabeli 19 wartości dopasowane należą do przedziału od około -0.097 do 1,53, czyli znacząco wykraczają poza dopuszczalny przedział prawdopodobieństwa. Obserwacji wykraczających poza ten przedział jest dokładnie 425, czyli ponad jedna czwarta wszystkich.

Otrzymane wyniki testów na heteroskedastyczność reszt i zakres, w którym znajdują się wartości dopasowane, każą jednoznacznie wykluczyć Liniowy Model Prawdopodobieństwa jako narzędzie badania czynników wpływających na prawdopodobieństwo zdobycia Oscara.

3.2 Model Logit

Kolejnym modelem omawianym w tym rozdziale pracy jest model Logit, różni się on od LMP dystrybuantą rozkładu prawdopodobieństwa. W LMP dystrybuanta miała postać funkcji liniowej, natomiast w modelu logitowym dystrybuantą jest dystrybunta rozkładu logistycznego o postaci:

$$\Lambda(x_i\beta) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \tag{2}$$

Podobnie jak w LMP analizę modelu logitowego należy zacząć od oszacowania modelu ogólnego, który przedstawia poniższa tabela:

Tabela 20. Ogólny model logitowy

```
. xi: logit oscar  budzet2000 i.gatunek ekranizacja roi przychody2000 kraj_prodnominacje zlote_globy
bafta milosc czas_trwania
i.gatunek         _Igatunek_0-9       (naturally coded; _Igatunek_0 omitted)

Iteration 0:   log likelihood = -644.32291
Iteration 1:   log likelihood = -281.35559
Iteration 2:   log likelihood =  -239.5106
Iteration 3:   log likelihood = -236.66683
Iteration 4:   log likelihood =  -236.1946
Iteration 5:   log likelihood = -235.78428
Iteration 6:   log likelihood = -235.78225
Iteration 7:   log likelihood = -235.78225

Logistic regression                               Number of obs   =       1638
                                                  LR chi2(19)     =     817.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -235.78225                       Pseudo R2       =     0.6341

------------------------------------------------------------------------------
       oscar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   budzet2000 |  -2.59e-09   4.68e-09    -0.55   0.580    -1.18e-08    6.57e-09
  _Igatunek_1 |  -.2373909   .4828453    -0.49   0.623     -1.18375    .7089686
  _Igatunek_2 |   .9471718   .6072435     1.56   0.119    -.2430034    2.137347
  _Igatunek_3 |   .8159699    .639625     1.28   0.202     -.437672    2.069612
  _Igatunek_4 |  -.5872744   .7910166    -0.74   0.458    -2.137639    .9630898
  _Igatunek_5 |   .4131266   .6023232     0.69   0.493    -.7674051    1.593658
  _Igatunek_6 |   .9000649   .5851035     1.54   0.124    -.2467169    2.046847
  _Igatunek_7 |   .0142123   .6206547     0.02   0.982    -1.202249    1.230673
  _Igatunek_8 |   .0969241   .8132038     0.12   0.905    -1.496926    1.690774
  _Igatunek_9 |  -.4927431   1.191666    -0.41   0.679    -2.828366     1.84288
  ekranizacja |  -.1438909   .2883711    -0.50   0.618    -.7090879    .4213061
          roi |  -.0001334   .0000993    -1.34   0.179     -.000328    .0000611
 przychody2000 |   2.11e-09   8.44e-10     2.50   0.012     4.58e-10    3.77e-09
kraj_produkcji |  -.7313622   .2766945    -2.64   0.008    -1.273674    -.189051
     nominacje |   .7179235   .0710326    10.11   0.000     .5787021    .8571449
   zlote_globy |   1.709133   .2826415     6.05   0.000     1.155166      2.2631
         bafta |   .7203465   .1709316     4.21   0.000     .3853267    1.055366
        milosc |   1.216106   .3967675     3.07   0.002     .4384561    1.993756
  czas_trwania |  -.0066497   .0080065    -0.83   0.406    -.0223422    .0090428
         _cons |  -3.372513    .958756    -3.52   0.000     -5.25164   -1.493386
------------------------------------------------------------------------------
Note: 2 failures and 0 successes completely determined.
```

Źródło: Opracowanie własne.

Jak można wywnioskować z tabeli 20, w modelu znajduje się kilka zmiennych nieistotnych na 5%-owym poziomie istotności, są nimi: budzet2000, _Igatunek_1, _Igatunek_2, _-

Igatunek_3, _Igatunek_4, _Igatunek_5, _Igatunek_6, _Igatunek_7, _Igatunek_8, _Igatunek_9, ekranizacja, roi i czas trwania. W związku z tym zdecydowano się przeprowadzić podobnie jak w przypadku LMP procedurę od ogólnego do szczegółowego. Model zagnieżdżony otrzymany po przeprowadzeniu procedury został przedstawiony poniżej.

Tabela 21. Zagnieżdżony model Logit i test Walda dla zmiennych usuniętych z modelu

```
. logit oscar _Igatunek_2 kraj_prod przychody2000 nominacje zlote_globy bafta milosc
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood =  -288.8549
Iteration 2:   log likelihood = -247.94559
Iteration 3:   log likelihood = -245.42423
Iteration 4:   log likelihood = -245.41989
Iteration 5:   log likelihood = -245.41989
Logistic regression                             Number of obs   =       1657
                                                LR chi2(7)      =     814.47
                                                Prob > chi2     =     0.0000
Log likelihood = -245.41989                     Pseudo R2       =     0.6240
------------------------------------------------------------------------------
       oscar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Igatunek_2 |    1.09894   .4186544     2.62   0.009     .2783924    1.919488
kraj_produkcji |  -.7350574   .2651866    -2.77   0.006    -1.254814   -.2153012
 przychody2000 |   1.31e-09   4.83e-10     2.71   0.007     3.61e-10    2.26e-09
     nominacje |    .677297   .0601785    11.25   0.000     .5593492    .7952448
   zlote_globy |   1.642281   .2649008     6.20   0.000     1.123085    2.161477
         bafta |   .7662303    .168816     4.54   0.000      .435357    1.097104
        milosc |   1.040389   .3741386     2.78   0.005      .307091    1.773687
         _cons |  -4.07565   .2378775   -17.13   0.000    -4.541881   -3.609418
------------------------------------------------------------------------------

. test _Igatunek_7 _Igatunek_8 _Igatunek_9 _Igatunek_1 ekranizacja budzet2000 Igatunek_5 czas_trwania
roi _Igatunek_5 _Igatunek_3 _Igatunek_6
 ( 1)  [oscar]_Igatunek_7 = 0
 ( 2)  [oscar]_Igatunek_8 = 0
 ( 3)  [oscar]_Igatunek_9 = 0
 ( 4)  [oscar]_Igatunek_1 = 0
 ( 5)  [oscar]ekranizacja = 0
 ( 6)  [oscar]budzet2000 = 0
 ( 7)  [oscar]_Igatunek_5 = 0
 ( 8)  [oscar]czas_trwania = 0
 ( 9)  [oscar]roi = 0
 (10)  [oscar]_Igatunek_5 = 0
 (11)  [oscar]_Igatunek_3 = 0
 (12)  [oscar]_Igatunek_6 = 0
       Constraint 10 dropped
           chi2( 11) =    8.74
         Prob > chi2 =    0.6455
```

Źródło: Opracowanie własne.

Statystyka testu ilorazu wiarygodności o wartości 814,47 i p-value równe 0,0000 każą odrzucić hipotezę zerową testu o łącznej nieistotności zmiennych. Wszystkie zmienne osobno również mają p-value mniejsze od 5% co oznacza, że należy dla każdej z nich odrzucić hipotezę zerową o nieistotności. Statystyka testu Walda równa 8,74 i p-value wyraźnie większe od 5% (0,6455 > 0,05) wskazują, iż nie ma podstaw do odrzucenia hipotezy zerowej o łącznej nieistotności zmiennych usuniętych z modelu.

Porównanie modelu ogólne i zagnieżdżonego na podstawie kryteriów informacyjnych BIC i AIC nie jest możliwe ze względu na różną liczbę obserwacji - model zagnieżdżony mający ich więcej niż ogólny byłby preferowany za względu na sposób wyliczania obu kryteriów (dzielenie przez N). Różnica w liczbie obserwacji bierze się z faktu, iż w modelu bez ograniczeń występuje zmienna budzet2000 dla której brakuje 19 obserwacji, a zmiennej tej nie ma w modelu z ograniczeniami.

3.2.1 Diagnostyka modelu Logit i testy jakości dopasowania

Przed przystąpieniem do interpretacji oszacowań modelu należy sprawdzić czy uzyskany model ma poprawną formę funkcyjną, czy jest dobrze dopasowany do danych. Do tego celu posłużą 3 testy diagnostyczne: linktest - test typu związku, test jakości dopasowania Perasona i test Hosmera-Lemeshow'a.

Tabela 22. Test linktest dla modelu zagnieżdżonego

```
. linktest
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -296.83394
Iteration 2:   log likelihood = -279.56721
Iteration 3:   log likelihood = -237.62155
Iteration 4:   log likelihood = -235.86757
Iteration 5:   log likelihood = -235.84334
Iteration 6:   log likelihood = -235.84323
Iteration 7:   log likelihood = -235.84323
Logistic regression                               Number of obs   =       1657
                                                  LR chi2(2)      =     833.62
                                                  Prob > chi2     =     0.0000
Log likelihood = -235.84323                       Pseudo R2       =     0.6386
```

oscar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	.9408812	.055074	17.08	0.000	.8329381	1.048824
_hatsq	-.0504623	.0066712	-7.56	0.000	-.0635376	-.037387
_cons	.2421468	.1626919	1.49	0.137	-.0767234	.561017

Źródło: Opracowanie własne.

Zarówno dla _hat, jak i _hatsq odrzucamy hipotezę zerową o nieistotności co oznacza, że dla wartości dopasowanych ważne są ich pierwsze i kolejne potęgi - forma funkcyjna jest niepoprawna.

Tabela 23. Test Hosmera-Lemeshowa dla modelu zagnieżdżonego

```
. lfit, group(10) table
Logistic model for oscar, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0086	0	1.4	167	165.6	167
2	0.0094	0	1.5	165	163.5	165
3	0.0115	0	1.7	166	164.3	166
4	0.0171	0	2.6	165	162.4	165
5	0.0184	1	2.9	165	163.1	166
6	0.0234	0	3.4	166	162.6	166
7	0.0374	6	4.9	159	160.1	165
8	0.0854	13	9.5	153	156.5	166
9	0.5610	55	41.7	111	124.3	166
10	1.0000	147	152.4	18	12.6	165

```
     number of observations =      1657
           number of groups =        10
     Hosmer-Lemeshow chi2(8) =     21.72
                Prob > chi2 =      0.0055
```

Źródło: Opracowanie własne.

Statystyka testu Pearsona wynosi 1981,2, a p-value 0.0000 < 0,05 co oznacza, że należy odrzucić hipotezę zerową wskazującą na poprawność formy funkcyjnej. Również test Hosmera-Lemeshowa wskazuje na niepoprawną formę funkcyjną - statystyka testowa=21,72 i p-value=0,0055 (mniejsze od 5%), skłaniają do odrzucenia hipotezy zerowej o poprawnej postaci funkcyjnej modelu.

Tabela 24. Test Pearsona dla modelu zagnieżdżonego

```
. estat gof
Logistic model for oscar, goodness-of-fit test
       number of observations =      1657
 number of covariate patterns =      1654
           Pearson chi2(1646) =      1981.20
                 Prob > chi2 =         0.0000
```

Źródło: Opracowanie własne.

W tej sytuacji gdy wszystkie testy wskazują na niepoprawność formy funkcyjnej modelu należy zastanowić się nad poszukaniem dodatkowych istotnych zmiennych objaśniających, które można by dodać do modelu, lub nad wprowadzeniem interakcji pomiędzy zmiennymi niezależnymi. W przypadku tego modelu oba sposoby zawiodły - nie udało się odnaleźć dodatkowych istotnych zmiennych ani interakcji. W takich sytuacjach jak sugerują autorzy książki *Logistic Regression with Stata* [2] można się posłużyć modelem Boxa-Tidwella (komenda *boxtid* w programie Stata)[3]. Model ten transformuje zmienną niezależną używając transformacji potęgowej i znajduje najlepszą jej potęgę na podstawie oszacowania największej wiarygodności (odpowiednik transformacji Boxa-Coxa). Transformacja zmiennej objaśniającej x jest dokonywana według wzoru:

$$\beta_0 + \beta_1 x^p \tag{3}$$

Oszacowana potęga p (na wydruku ze Staty pod nazwą $p1$) przyjmuje wartości ze zbioru liczb rzeczywistych, aby ułatwić późniejszą interpretację przetransformowanych zmiennych przyjęło się przybliżać p do jednej z liczb ze zbioru $P = \{-2, -1, -0.5, 0, 0.5, 1, 2\}$. W przypadku, gdy oszacowane p jest bliskie zera to zmienną niezależną x transformuje się używając logarytmu naturalnego do postaci $ln(x)$, w pozostałych przypadkach zmienną niezależną podnosi się do najbliższej potęgi ze zbioru P.[4] W tabeli 25 zaprezentowane jest wydruk testu transformacji zmiennych niezależnych przy użyciu komendy *boxtid*. Opcja zero() w tym teście została użyta ze względu na przeważającą liczbę obserwacji zerowych dla zmiennych nominacje, zlote_globy i bafta. Opcja ta zapewnia nie uwzględnianie obserwacji zerowych przy dopasowywaniu modelu z potęgami - pod uwagę brane są tylko obserwacje mogące wpływać na nieliniowość zmiennej objaśniającej.

[3] Przykład zastosowania modelu Boxa-Tidwella do transformacji zmiennej niezależnej w modelu Logit można odnaleźć w trzecim rozdziale (*Logistic Regression Diagnostics*) wyżej wspomnianej książki

[4] Szczegółowy opis działania komendy *boxtid* i samego modelu Boxa-Tidewella został zawarty w artykule P. Roystona i G. Ambler *Nonlinear regression models involving power or exponential functions of covariates* opublikowanym w 1999 roku w *Stata Technical Bulletin 49 [10]*.

Tabela 25. Test transformacji zmiennych dla modelu zagnieżdżonego Logit

```
. boxtid logit oscar _Igatunek_2 kraj_prod przychody2000 nominacje zlote_globy bafta milosc,
zero(nominacje zlote_globy bafta)
Iteration 0:  Deviance =  427.4704
(unprofitable step attempted, step length divided by 10)
Iteration 1:  Deviance =  410.3915 (change = -17.07888)
Iteration 2:  Deviance =  410.3759 (change = -.0156017)
Iteration 3:  Deviance =  410.3746 (change = -.0012795)
Iteration 4:  Deviance =  410.3745 (change =  -.000127)
-> gen double Iprzy__1 = X^0.5491-.3674405338 if e(sample)
-> gen double Iprzy__2 = X^0.5491*ln(X)+.6700151336 if e(sample)
   (where: X = przychody2000/1000000000)
-> gen double Inomi__1 = X^0.1671-.8424693853 if e(sample)
-> gen double Inomi__2 = X^0.1671*ln(X)+.8642909775 if e(sample)
   (where: X = nominacje/10)
-> gen double Izlot__1 = zlote_globy^0.2492-1.161232965 if e(sample)
-> gen double Izlot__2 = zlote_globy^0.2492*ln(zlote_globy)-.6965615002 if e(sa
> mple)
-> gen double Ibaft__1 = bafta^0.2163-1.167894033 if e(sample)
-> gen double Ibaft__2 = bafta^0.2163*ln(bafta)-.8380787452 if e(sample)
[Total iterations: 16]
Box-Tidwell regression model

Logistic regression                               Number of obs   =       1657
                                                  LR chi2(11)     =     894.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -205.18721                       Pseudo R2       =     0.6856

------------------------------------------------------------------------------
       oscar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Iprzy__1 |   1.029977   .6067958     1.70   0.090    -.1593207    2.219275
    Iprzy_p1 |   .0053793   .5716119     0.01   0.992    -1.114959    1.125718
    Inomi__1 |   9.716826   2.955026     3.29   0.001     3.925082    15.50857
    Inomi_p1 |  -.0017302   .6638651    -0.00   0.998    -1.302882    1.299421
    Izlot__1 |   1.809972   .3147959     5.75   0.000     1.192983    2.426961
    Izlot_p1 |   .0003923   .5272507     0.00   0.999       -1.033    1.033785
    Ibaft__1 |   1.265426   .3160349     4.00   0.000     .6460094    1.884843
    Ibaft_p1 |    .000038   .3781499     0.00   1.000    -.7411223    .7411983
 _Igatunek_2 |   .6675307   .4445605     1.50   0.133    -.2037919    1.538853
kraj_produ~i |  -.6652295   .2711986    -2.45   0.014    -1.196769     -.13369
      milosc |   1.119758    .413176     2.71   0.007     .3099482    1.929568
       _cons |   2.689804    .428151     6.28   0.000     1.850643    3.528964
-------------+----------------------------------------------------------------
przychody2000| 7.37e-10   4.24e-10      1.739 Nonlin. dev. 0.922   (P = 0.337)
      p1 |   .5490613   .5305962      1.035
-------------+----------------------------------------------------------------
nominacje|      .6618   .0606394     10.914   Nonlin. dev. 56.291  (P = 0.000)
      p1 |   .1670898   .0681398      2.452
-------------+----------------------------------------------------------------
zlote_globy| 1.284998   .2322075      5.534   Nonlin. dev. 5.699   (P = 0.017)
      p1 |   .2492006   .2913101      0.855
-------------+----------------------------------------------------------------
bafta    |   .6230049   .1535889      4.056   Nonlin. dev. 6.261   (P = 0.012)
      p1 |   .2162805   .2988773      0.724
------------------------------------------------------------------------------
Deviance:  410.375.
```

Źródło: Opracowanie własne.

Dla trzech zmiennych p-value testu na nieliniowość ma wartość mniejszą od 5%, co oznacza konieczność odrzucenia hipotezy zerowej o liniowości tych zmiennych, są nimi: nominacje (P = 0,000), zlote_globy (P = 0,017) i bafta (P = 0,012). Dla zmiennej przychody2000 p-value równe 0,337 wskazuje na brak podstaw do odrzucenia h0 o liniowej formie zmiennej. Dla wszystkich trzech zmiennych $p1$ jest bliższe 0 niż 0,5, dla zmiennej nominacje jest to około 0,167, dla zmiennej zlote_globy 0,2492, a dla zmiennej bafta 0,2163. W tej sytuacji należy przetransformować wszystkie trzy zmienne przy użyciu logarytmu naturalnego. Taka postać transformacji jest prawdopodobnie spowodowana spadającym krańcowym wpływem danej zmiennej objaśniającej na objaśnianą. Czyli z każdą kolejną nominacją/nagrodą wpływ na prawdopodobieństwo zdobycia Oscara jest mniejszy. W związku z wynikami testu sprawdzającego nieliniowość zmiennych

niezależnych oszacowano nowy model logitowy ze zmiennymi ln_nom, ln_zg i ln_baf będącymi logarytmami naturalnymi zmiennych: nominacje, zlote_globy i bafta.

Tabela 26. Model Logit po transformacji

```
. logit oscar _Igatunek_2 kraj_prod przychody2000 ln_nom ln_zg ln_baf milosc
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -343.52428
Iteration 2:   log likelihood = -293.05786
Iteration 3:   log likelihood = -286.11262
Iteration 4:   log likelihood = -286.07875
Iteration 5:   log likelihood = -286.07874

Logistic regression                               Number of obs   =       1657
                                                  LR chi2(7)      =     733.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -286.07874                       Pseudo R2       =     0.5617
```

oscar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Igatunek_2	1.298812	.4099903	3.17	0.002	.495246	2.102379
kraj_produkcji	-.6795015	.2417108	-2.81	0.005	-1.153246	-.205757
przychody2000	9.58e-10	4.65e-10	2.06	0.039	4.73e-11	1.87e-09
ln_nom	2.43787	.1658775	14.70	0.000	2.112757	2.762984
ln_zg	1.91387	.7098043	2.70	0.007	.5226786	3.30506
ln_baf	1.620144	.4197213	3.86	0.000	.7975055	2.442783
milosc	1.076732	.3544051	3.04	0.002	.3821111	1.771354
_cons	-3.707948	.2136529	-17.36	0.000	-4.1267	-3.289196

Źródło: Opracowanie własne.

Wszystkie zmienne w modelu są łącznie istotne (LR chi2(7)=733,15, p-value=0,0000 < 0,05) i każda zmienna z osobna ma p-value mniejsze od 5% co wskazuje na konieczność odrzucenia hipotezy zerowej o nieistotności poszczególnych zmiennych. Psudo R^2 wskazuje, iż zmienność została wyjaśniona w około 56,15%. Przeprowadzenie ponownych testów poprawności funkcyjnej modelu pozwoli odpowiedzieć na pytanie czy transformacja zmiennych przyniosła pożądany skutek.

Tabela 27. Linktest dla modelu z transformacjami

```
. linktest
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -340.18544
Iteration 2:   log likelihood = -300.42787
Iteration 3:   log likelihood = -290.19594
Iteration 4:   log likelihood = -289.39801
Iteration 5:   log likelihood = -289.36772
Iteration 6:   log likelihood = -289.36772

Logistic regression                               Number of obs   =       1657
                                                  LR chi2(2)      =     726.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -289.36772                       Pseudo R2       =     0.5566
```

oscar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	.9450873	.071231	13.27	0.000	.8054771	1.084698
_hatsq	-.0251742	.0231046	-1.09	0.276	-.0704584	.02011
_cons	.0769665	.1584652	0.49	0.627	-.2336197	.3875527

Źródło: Opracowanie własne.

Tym razem zmienna *_hat* okazała się istotna (p-value = 0 - należy odrzucić h0 o nieistotności zmiennej na każdym poziomie istotności), a w przypadku zmiennej *_hatsq* niema podstaw do

odrzucenia h0 o nieistotności (p.value = 0,276 > 0,05). Model z trzema zmiennymi w logarytmach ma poprawną formę funkcyjną.

Tabela 28. Test Pearsona i test Hosmera-Lemeshowa dla modelu z transformacjami

```
. estat gof
Logistic model for oscar, goodness-of-fit test

       number of observations =      1657
 number of covariate patterns =      1654
          Pearson chi2(1646) =      1440.50
                Prob > chi2 =         0.9999

. lfit, group(10) table

Logistic model for oscar, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0142	1	2.3	165	163.7	166
2	0.0151	1	2.4	165	163.6	166
3	0.0170	2	2.6	164	163.4	166
4	0.0235	4	3.3	161	161.7	165
5	0.0244	4	4.0	162	162.0	166
6	0.0261	5	4.2	161	161.8	166
7	0.0408	3	4.9	162	160.1	165
8	0.0799	10	9.8	156	156.2	166
9	0.5908	49	45.4	117	120.6	166
10	0.9999	143	143.1	22	21.9	165

```
      number of observations =      1657
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =         3.23
                Prob > chi2 =         0.9194
```

Źródło: Opracowanie własne.

Tabela 29. Porównanie modeli przed i po transformacji trzech zmiennych niezależnych

```
. fitstat, using(m2)
Measures of Fit for logit of oscar
                               Current             Saved        Difference
Model:                           logit             logit
N:                                1657              1657                 0
Log-Lik Intercept Only         -652.655          -652.655             0.000
Log-Lik Full Model             -286.079          -245.420           -40.659
D                               572.157(1649)     490.840(1649)      81.318(0)
LR                              733.153(7)        814.471(7)         81.318(0)
Prob > LR                         0.000             0.000                .
McFadden s R2                     0.562             0.624             -0.062
McFadden s Adj R2                 0.549             0.612             -0.062
ML (Cox-Snell) R2                 0.358             0.388             -0.031
Cragg-Uhler(Nagelkerke) R2        0.656             0.712             -0.056
McKelvey & Zavoina s R2           0.610             0.751             -0.142
Efron s R2                        0.590             0.635             -0.044
Tjur s Discrimination Coef        0.596             0.650             -0.053
Variance of y*                    8.425            13.221             -4.796
Variance of error                 3.290             3.290              0.000
Count R2                          0.937             0.941             -0.004
Adj Count R2                      0.532             0.559             -0.027
AIC                             588.157           506.840            81.318
AIC/N                             0.355             0.306              0.049
BIC                             631.460           550.142            81.318
Difference of   81.318 in BIC provides very strong support for saved model.
Note: p-value for difference in LR is only valid if models are nested.
```

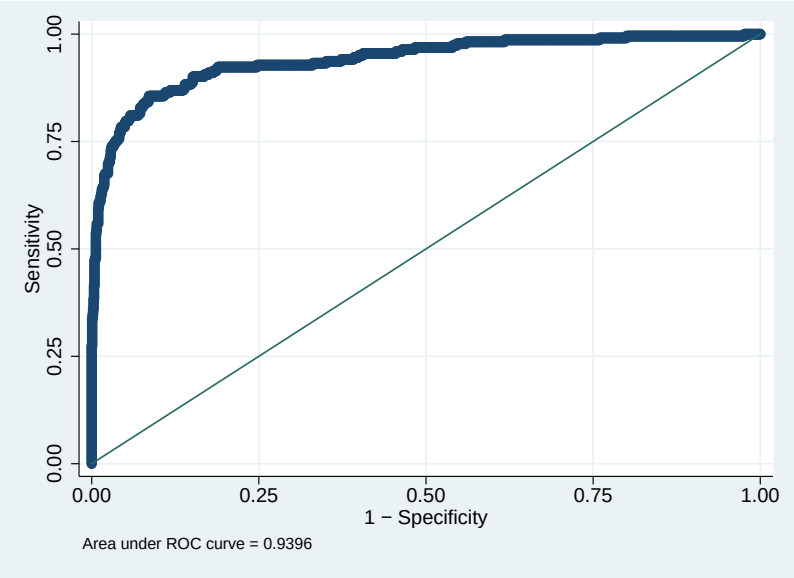
Źródło: Opracowanie własne.

Podobnie jak linktest tak i dwa kolejne testy sprawdzające dopasowanie modelu do danych wskazały, iż nowy model ma prawidłową formę funkcyjną. Statystyka testu Pearsona wyniosła

1440,50, a p-value 0,9999, co wskazuje na przyjęcie hipotezy zerowej mówiącej o poprawności zastosowanej formy funkcyjnej. Podobnie statystyka testu Hosmera-Lemeshowa równa 3,23 i p-value równe 0,9194 (>0,05) wskazują na brak podstaw do odrzucenia h0 mówiącej o prawidłowości formy funkcyjnej. W tym miejscu należałoby porównać model przed transformacją i po transformacji zmiennych, aby zobaczyć jak zlogarytmowanie wpłynęło na najważniejsze miary dopasowania modelu.

Tabel 29 przedstawia porównanie modeli przed i po transformacji trzech zmiennych niezależnych. *Current logit* oznacza model po transformacjach, a *Saved logit* oznacza model przed transformacjami. Niemal wszystkie współczynniki determinacji modelu wskazują, iż model ze zmiennymi niezlogarytmowanymi lepiej reprezentował dane, lepiej przewidywał sukcesy i porażki, miał niższe kryteria informacyjne. Mimo to należy posłużyć się modelem gorzej dopasowanym , ale z poprawną formą funkcyjną. Tak więc model po transformacjach dla zmiennej ukrytej y_i^* byłby wyjaśniany w 61%, gdyby zmienna ta była bezpośrednio obserwowalna (McKelvey-Zavoina R2). Przewiduje on trafnie sukcesy i porażki w 93,7% na co wskazuje liczebnościowe R^2. Skorygowane liczebnościowe R^2 (Adj Count R2) pokazuje natomiast, że dzięki zmiennym objaśniającym model trafnie przewiduje sukcesy i porażki w 53,2%.

Rysunek 5: Krzywa ROC



Źródło: Opracowanie własne.

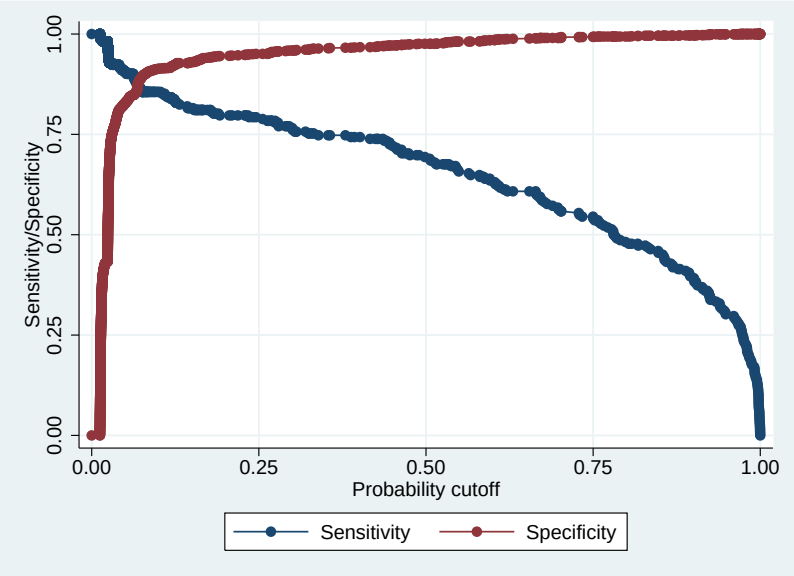
Tabela 30. Pole pod krzywą ROC

```
. lroc
Logistic model for oscar
number of observations =     1657
area under ROC curve   =   0.9396
```

Źródło: Opracowanie własne.

Pole pod krzywą ROC jest to procent dobrze zakwalifikowanych pozytywnych i negatywnych odpowiedzi przy różnych wyborach punktu rozdzielającego decyzję. Im bliższa jedynce wartość pola tym model jest lepiej dopasowany do danych wejściowych. Tabela 30 wskazuje na bardzo dobre dopasowanie modelu - pole pod krzywą ROC wynosi 0,9396.

Rysunek 6: Wrażliwość i Specyficzność



Źródło: Opracowanie własne.

Wykres wrażliwości i specyficzności wskazuje prawdopodobieństwo ucięcia, czyli poziom rozgraniczenia maksymalizującego trafność klasyfikacji wartości dopasowanych. Z rysunku 6 wynika, iż najlepsze ucięcie powinno mieć miejsce w punkcie 0,05. Dla tego punktu wyznaczona została tablica klasyfikacji.

Tabela 31. Tablica klasyfikacji dla punktu odcięcia równego 0,05

```
. lstat,cutoff(0.05)
Logistic model for oscar
```

Classified	True D	True ~D	Total
+	201	248	449
-	21	1187	1208
Total	222	1435	1657

```
Classified + if predicted Pr(D) >= .05
True D defined as oscar != 0
```

Sensitivity	Pr(+| D)	90.54%
Specificity	Pr(-|~D)	82.72%
Positive predictive value	Pr(D| +)	44.77%
Negative predictive value	Pr(~D| -)	98.26%
False + rate for true ~D	Pr(+|~D)	17.28%
False - rate for true D	Pr(-| D)	9.46%
False + rate for classified +	Pr(~D| +)	55.23%
False - rate for classified -	Pr(D| -)	1.74%
Correctly classified		83.77%

Źródło: Opracowanie własne.

Tabela 31 wskazuje iż na 222 filmy, które zdobyły Oscara, model prawidłowo przewidział aż 201 z nich, czyli blisko 91% (wrażliwość), a jedynie 21 zostało błędnie zakwalifikowanych (9%). Natomiast spośród 1435 filmów, które nie otrzymały żadnej statuetki, model prawidłowo

przewidział 1187, czyli około 83% (specyficzność), błędnie zakwalifikowanych obserwacji było w tej kategorii około 17%. Łącznie 83,77% wszystkich obserwacji zostało trafnie zakwalifikowanych, co wydaje się bardzo dobrym wynikiem.

Tabela 32. Test Współliniowości zmiennych

```
. collin _Igatunek_2 kraj_prod przychody2000 ln_nom ln_zg ln_baf milosc
(obs=1657)
  Collinearity Diagnostics
                       SQRT                  R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
_Igatunek_2      1.05    1.02    0.9531      0.0469
kraj_produkcji      1.03    1.02    0.9678      0.0322
przychody2000      1.17    1.08    0.8528      0.1472
    ln_nom      1.83    1.35    0.5470      0.4530
     ln_zg      1.52    1.23    0.6579      0.3421
    ln_baf      1.56    1.25    0.6404      0.3596
    milosc      1.01    1.00    0.9923      0.0077
----------------------------------------------------
  Mean VIF      1.31

                         Cond
        Eigenval         Index
---------------------------------
    1     3.1127          1.0000
    2     1.5024          1.4394
    3     0.9743          1.7874
    4     0.8126          1.9572
    5     0.5384          2.4044
    6     0.4980          2.5001
    7     0.3291          3.0754
    8     0.2324          3.6594
---------------------------------
 Condition Number         3.6594
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.3797
```

Źródło: Opracowanie własne.

Ostatnim koniecznym testem jest test na występowanie współliniowości zmiennych objaśniających w modelu. Współliniowość zmiennych ma miejsce gdy jedną zmienną można uzyskać jako kombinację liniową pozostałych. Występowanie współliniowości może stanowić obciążenie dla oszacowanego ilorazu szans, gdyż powoduje ono znaczący wzrost błędów standardowych w modelu. Aby sprawdzić czy zjawisko współliniowości zmiennych objaśniających występuje w oszacowanym modelu logitowym posłużono się komendą *collin* w programie Stata. Tabela 32 przedstawia wyniki tego testu. Niska wartość statystyk *VIF*, *Condition Number* i *Condition Index* (poniżej 10) wskazuje na brak współliniowości pomiędzy zmiennymi niezależnymi i ich stabilność.

3.3 Model Probit

Model probitowy podobnie jak model logitowy jest modelem dedykowanym binarnej zmiennej zależnej. Główna różnica pomiędzy Logitem, a Probitem polega na postaci dystrybuanty rozkładu prawdopodobieństwa - dla modelu logitowego jest to dystrybuanta rozkładu logitowego $\Lambda(x_i\beta)$, a dla modelu probitowego jest nią dystrybuanta standaryzowanego rozkładu normalnego $\Phi(x_i\beta)$. Podobnie jak w przypadku modeli Logit i LMP na początku wyestymowano ogólny model Probit (tabela 33).

Tabela 33. Model Probit ogólny

```
. xi: probit oscar  budzet2000 i.gatunek ekranizacja roi przychody2000 kraj_prod nominacje zlote_globy bafta
 milosc czas_trwania
i.gatunek         _Igatunek_0-9       (naturally coded; _Igatunek_0 omitted)

Iteration 0:   log likelihood = -644.32291
Iteration 1:   log likelihood = -252.37591
Iteration 2:   log likelihood = -233.42969
Iteration 3:   log likelihood = -233.07162
Iteration 4:   log likelihood = -232.57344
Iteration 5:   log likelihood = -232.57049
Iteration 6:   log likelihood = -232.57049

Probit regression                                 Number of obs   =       1638
                                                  LR chi2(18)     =     823.50
                                                  Prob > chi2     =     0.0000
Log likelihood = -232.57049                       Pseudo R2       =     0.6390

------------------------------------------------------------------------------
       oscar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   budzet2000 |  -2.05e-09   2.36e-09    -0.87   0.386    -6.67e-09    2.58e-09
  _Igatunek_1 |  -.1380634   .2283552    -0.60   0.545    -.5856314    .3095046
  _Igatunek_2 |   .4729934   .3124234     1.51   0.130    -.1393452    1.085332
  _Igatunek_3 |   .3290652    .334237     0.98   0.325    -.3260273    .9841578
  _Igatunek_4 |  -.2879733   .3982407    -0.72   0.470    -1.068511    .4925641
  _Igatunek_5 |   .1873941   .2964899     0.63   0.527    -.3937154    .7685035
  _Igatunek_6 |   .3683275   .2977027     1.24   0.216     -.215159     .951814
  _Igatunek_7 |  -.0186585   .3088218    -0.06   0.952    -.6239381    .5866212
  _Igatunek_8 |  -.0237881   .3833369    -0.06   0.951    -.7751147    .7275385
  _Igatunek_9 |  -.2951531    .613072    -0.48   0.630    -1.496752    .9064459
  ekranizacja |  -.0461109   .1434598    -0.32   0.748     -.327287    .2350652
          roi |  -.0000491   .0000469    -1.05   0.295     -.000141    .0000428
 przychody2000 |   1.03e-09   4.26e-10     2.43   0.015     1.98e-10    1.87e-09
kraj_produkcji |  -.3755098   .1370391    -2.74   0.006    -.6441015    -.106918
     nominacje |   .3835013   .0353914    10.84   0.000     .3141353    .4528672
   zlote_globy |   .8039639   .1324897     6.07   0.000     .5442889    1.063639
         bafta |    .375868   .0860155     4.37   0.000     .2072806    .5444554
        milosc |   .5883756   .2029476     2.90   0.004     .1906055    .9861456
  czas_trwania |  -.0036616   .0039782    -0.92   0.357    -.0114588    .0041355
         _cons |  -1.769618   .4736227    -3.74   0.000    -2.697901   -.8413347
------------------------------------------------------------------------------
Note: 2 failures and 15 successes completely determined.
```

Źródło: Opracowanie własne.

Na 5%-owym poziomie istotności nie ma podstaw do odrzucenia hipotezy zerowej o nieistotności dla następujących zmiennych: budzet2000, _Igatunek_1, _Igatunek_2, _Igatunek_3, _Igatunek_4, _Igatunek_5, _Igatunek_6, _Igatunek_7, _Igatunek_8, _Igatunek_9, ekranizacja, roi, czas_trwania. W celu pozbycia się zmiennych nieistotnych z modelu zastosowano procedurę od ogólnego do szczegółowego, po przeprowadzeniu której uzyskano model ze wszystkimi zmiennymi istotnymi. W tym miejscu zostanie zaprezentowany tylko ostatni etap procedury.

Tabela 34. Test Walda dla zmiennych usuniętych z modelu

```
. test _Igatunek_7 _Igatunek_8 ekranizacja _Igatunek_9 _Igatunek_1 _Igatunek_4  czas_trwania roi
budzet2000 _Igatunek_3 _Igatunek_5 _Igatunek_6
 ( 1)  [oscar]_Igatunek_7 = 0
 ( 2)  [oscar]_Igatunek_8 = 0
 ( 3)  [oscar]ekranizacja = 0
 ( 4)  [oscar]_Igatunek_9 = 0
 ( 5)  [oscar]_Igatunek_1 = 0
 ( 6)  [oscar]_Igatunek_4 = 0
 ( 7)  [oscar]czas_trwania = 0
 ( 8)  [oscar]roi = 0
 ( 9)  [oscar]budzet2000 = 0
 (10)  [oscar]_Igatunek_3 = 0
 (11)  [oscar]_Igatunek_5 = 0
 (12)  [oscar]_Igatunek_6 = 0
           chi2( 12) =    8.11
         Prob > chi2 =    0.7761
```

Źródło: Opracowanie własne.

Statystyka testu Walda (8,11) i p-value(0,7761 < 0,05) wskazują, iż nie ma podstaw do odrzucenia h0 o łącznej nieistotności zmiennych usuniętych z modelu.

Tabela 35. Model Probit zagnieżdżony

```
. probit oscar _Igatunek_2 kraj_prod przychody2000 nominacje zlote_globy bafta milosc
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -260.18232
Iteration 2:   log likelihood = -241.49363
Iteration 3:   log likelihood = -241.25808
Iteration 4:   log likelihood = -241.25791
Iteration 5:   log likelihood = -241.25791
Probit regression                                 Number of obs   =       1657
                                                  LR chi2(6)      =     822.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -241.25791                       Pseudo R2       =     0.6303
------------------------------------------------------------------------------
         oscar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
    _Igatunek_2 |    .5389003   .2270996     2.37   0.018     .0937931    .9840074
kraj_produkcji |   -.3809342   .1314571    -2.90   0.004    -.6385853   -.1232831
 przychody2000 |    6.53e-10   2.34e-10     2.79   0.005     1.94e-10    1.11e-09
      nominacje |    .3678263   .0306592    12.00   0.000     .3077353    .4279173
    zlote_globy |    .7954267   .1267459     6.28   0.000     .5470092    1.043844
          bafta |    .3983077   .0842226     4.73   0.000     .2332344    .5633811
         milosc |    .5150889    .194828     2.64   0.008     .1332331    .8969446
          _cons |   -2.212964   .1096436   -20.18   0.000    -2.427861   -1.998066
------------------------------------------------------------------------------
Note: 0 failures and 15 successes completely determined.
```

Źródło: Opracowanie własne.

Statystyka LR chi2(6) = 822,79 i p-value=0,0000 < 0,05 wskazuje na konieczność odrzucenia hipotezy zerowej o łącznej nieistotności wszystkich zmiennych. Również p-value każdej zmiennej z osobna jest mniejsze od 5% co każe odrzucić h0 dla każdej zmiennej o nieistotności - wszystkie zmienne są osobno istotne. Model wyjaśnia około 63,03% zmienności (Pseudo R^2).

Podobnie jak w przypadku modelu Logit nie możemy porównywać modeli ogólnego i zagnieżdżonego na podstawie kryteriów informacyjnych BIC i AIC ze względu na różnicę w liczbie obserwacji w obu modelach - wynika ona z braku 19 obserwacji zmiennej budzet2000.

3.3.1 Diagnostyka modelu Probit i testy jakości dopasowania

Aby móc interpretować oszacowania modelu należy najpierw sprawdzić czy model zagnieżdżony ma poprawną formę funkcyjną, czy jest dobrze dopasowany. W tym celu wykorzystane zostaną trzy testy diagnostyczne: test typu związku (linktest), test jakości dopasowania (goodness of fit) i test Hosmera-Lemeshow'a.

Tabela 36. Test linktest dla zagnieżdżonego modelu Probit

```
. linktest
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -244.58488
Iteration 2:   log likelihood = -230.27775
Iteration 3:   log likelihood =  -230.1471
Iteration 4:   log likelihood = -230.14708
Iteration 5:   log likelihood = -230.14708
Probit regression                                 Number of obs   =       1657
                                                  LR chi2(2)      =     845.02
                                                  Prob > chi2     =     0.0000
Log likelihood = -230.14708                       Pseudo R2       =     0.6474
```

oscar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	.9539365	.049118	19.42	0.000	.857667	1.050206
_hatsq	-.0997172	.013172	-7.57	0.000	-.125534	-.0739005
_cons	.1917916	.090433	2.12	0.034	.0145461	.369037

Źródło: Opracowanie własne.

Test typu związku wskazuje, iż dla zmiennej *_hat* należy odrzucić hipotezę zerową o nieistotności (p-value=0,000 < 0,05), podobnie dla *_hatsq* odrzucamy hipotezę o nieistotności (p-value=0,0000 < 0,05), co oznacza, że dla wartości dopasowanych istotne są również dalsze potęgi.

Tabela 37. Test jakości dopasowania dla zagnieżdżonego modelu Probit

```
. estat gof
Probit model for oscar, goodness-of-fit test
       number of observations =      1657
 number of covariate patterns =      1654
          Pearson chi2(1646) =      6344.33
                Prob > chi2 =         0.0000
```

Źródło: Opracowanie własne.

Również test jakości dopasowania Pearsona wskazuje na złe dopasowanie modelu. Statystyka testowa tego test uje równa 6344,33, a p-value 0,0000 co oznacza, iż należy odrzucić hipotezę zerową o prawidłowej formie funkcyjnej modelu.

Tabela 38. Test Hosmera-Lemeshow'a dla zagnieżdżonego modelu Probit

```
. lfit, group(10) table
Probit model for oscar, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0052	0	0.8	167	166.2	167
2	0.0059	0	0.9	165	164.1	165
3	0.0078	0	1.1	166	164.9	166
4	0.0139	0	2.1	165	162.9	165
5	0.0153	0	2.4	166	163.6	166
6	0.0201	1	2.9	165	163.1	166
7	0.0379	6	4.7	159	160.3	165
8	0.0903	12	9.9	154	156.1	166
9	0.5247	56	41.7	110	124.3	166
10	1.0000	147	149.5	18	15.5	165

```
      number of observations =      1657
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =        16.46
                Prob > chi2 =         0.0363
```

Źródło: Opracowanie własne.

Statystyka testowa testu Hosmera-Lemeshowa wynosi 16,46, p-value tego testu jest równe 0,0363, co oznacza, iż należy odrzucić hipotezę zerową o poprawności funkcyjnej modelu.

Podobnie jak w przypadku zagnieżdżonego modelu Logit wszystkie testu diagnostyczne wskazały, iż zagnieżdżony model Probit ma niepoprawną formę funkcyjną na 5%-owym poziomie istotności. Oznacza to, iż należy poszukać istotnych interakcji pomiędzy zmiennymi lub dodać dodatkowe istotne zmienne objaśniające do modelu. Oba sposoby nie przyniosły w przypadku tego modelu efektu, co skłania do sprawdzenia za pomocą modelu Boxa-Tidwella

czy wszystkie zmienne niezależne mają liniową postać (tabela 39).

Tabela 39. Test transformacji zmiennych dla zagnieżdżonego modelu Probit

```
. boxtid probit oscar _Igatunek_2 kraj_prod przychody2000 nominacje zlote_globy bafta milosc, zero (nominacje
 zlote_globy bafta)
Iteration 0:  Deviance =  419.4762
(unprofitable step attempted, step length divided by 10)
Iteration 1:  Deviance =  409.3456 (change = -10.13062)
Iteration 2:  Deviance =   409.208 (change = -.1375801)
Iteration 3:  Deviance =  409.2024 (change = -.0056166)
Iteration 4:  Deviance =  409.2012 (change = -.0012525)
Iteration 5:  Deviance =  409.2006 (change = -.0005835)
-> gen double Iprzy__1 = X^0.6074-.330388106 if e(sample)
-> gen double Iprzy__2 = X^0.6074*ln(X)+.6024513102 if e(sample)
   (where: X = przychody2000/1000000000)
-> gen double Inomi__1 = X^0.2478-.7755143751 if e(sample)
-> gen double Inomi__2 = X^0.2478*ln(X)+.7956017025 if e(sample)
   (where: X = nominacje/10)
-> gen double Izlot__1 = zlote_globy^0.1801-1.114093141 if e(sample)
-> gen double Izlot__2 = zlote_globy^0.1801*ln(zlote_globy)-.6682848434 if e(sa
> mple)
-> gen double Ibaft__1 = bafta^0.2500-1.196456662 if e(sample)
-> gen double Ibaft__2 = bafta^0.2500*ln(bafta)-.8585752382 if e(sample)
[Total iterations: 20]
Box-Tidwell regression model
Probit regression                                 Number of obs   =       1657
                                                  LR chi2(11)     =     896.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -204.59967                       Pseudo R2       =     0.6865

------------------------------------------------------------------------------
       oscar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Iprzy__1 |   .6022548   .3187901     1.89   0.059    -.0225624    1.227072
    Iprzy_p1 |  -.0110813    .314466    -0.04   0.972    -.6274232    .6052607
    Inomi__1 |   4.245622   .8706235     4.88   0.000     2.539232    5.952013
    Inomi_p1 |  -.0002092   .3302994    -0.00   0.999    -.6475841    .6471656
    Izlot__1 |   1.050233   .1796496     5.85   0.000     .6981265     1.40234
    Izlot_p1 |  -.0003976   .2929421    -0.00   0.999    -.5745537    .5737585
    Ibaft__1 |   .7007768   .1780864     3.94   0.000      .351734     1.04982
    Ibaft_p1 |  -.0000487   .2047614    -0.00   1.000    -.4013738    .4012763
 _Igatunek_2 |   .3711311   .2601759     1.43   0.154    -.1388043    .8810665
kraj_produ~i |  -.4123544   .1529418    -2.70   0.007    -.7121149   -.1125939
      milosc |   .6354977   .2378797     2.67   0.008     .1692622    1.101733
       _cons |    1.50005   .2259557     6.64   0.000     1.057185    1.942915
-------------+----------------------------------------------------------------
przychody2000| 4.52e-10   2.37e-10      1.902 Nonlin. dev. 0.832   (P = 0.362)
      p1 |   .6073531    .617754      0.983
-------------+----------------------------------------------------------------
nominacje|   .3536865   .0312209     11.329   Nonlin. dev. 49.843  (P = 0.000)
      p1 |   .2478102   .0777897      3.186
-------------+----------------------------------------------------------------
zlote_globy| .6752742   .1194133      5.655   Nonlin. dev. 7.509   (P = 0.006)
      p1 |   .1801139   .2790091      0.646
-------------+----------------------------------------------------------------
bafta    |   .3426729    .082032      4.177   Nonlin. dev. 5.934   (P = 0.015)
      p1 |   .2499509   .2920985      0.856
------------------------------------------------------------------------------
Deviance:  409.201.
```

Źródło: Opracowanie własne.

Test na nieliniowość zmiennych niezależnych każe odrzucić hipotezę zerową mówiącą o liniowości dla trzech zmiennych: nominacje (P = 0,000 < 0,05), zlote_globy (P = 0.006 < 0,05) i bafta (P = 0.015 < 0,05). Sugerowana potęgi $p1$ dla tych trzech zmiennych (kolejno 0,2478, 0,1801 i 0,24995) są bliższe zeru niż 0,5, więc każdą z tych zmiennych należy zlogarytmować, choć w przypadku zmiennej bafta wartość potęgi równa 0,24995 jest niejako na granicy transformacji logartymicznej i pierwiastkowej ($p = 0,5$). Wnioski uzyskane z tego testu są tożsame z wnioskami z wnioskami uzyskanymi dla Logitu. Nie trzeba wiec dokonywać dodatkowych transformacji zmiennych, gdyż zostały one już wygenerowane dla modelu logitowego. Model

Probit po transformacjach ma następującą postać:

Tabela 40. Model Probit po transformacji trzech zmiennych niezależnych

```
. probit oscar _Igatunek_2 kraj_prod przychody2000 ln_nom ln_zg ln_baf milosc
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -298.19042
Iteration 2:   log likelihood = -284.49931
Iteration 3:   log likelihood = -284.05301
Iteration 4:   log likelihood = -284.05187
Iteration 5:   log likelihood = -284.05187
Probit regression                                 Number of obs   =       1657
                                                  LR chi2(6)      =     737.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -284.05187                       Pseudo R2       =     0.5648
```

oscar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Igatunek_2	.657707	.2129003	3.09	0.002	.2404301	1.074984
kraj_produkcji	-.3684023	.1202409	-3.06	0.002	-.6040703	-.1327344
przychody2000	4.70e-10	2.25e-10	2.09	0.036	2.99e-11	9.11e-10
ln_nom	1.330217	.0860817	15.45	0.000	1.1615	1.498934
ln_zg	1.042435	.3593798	2.90	0.004	.3380632	1.746806
ln_baf	.8936561	.2194941	4.07	0.000	.4634555	1.323857
milosc	.5213194	.1806683	2.89	0.004	.1672159	.8754228
_cons	-1.99513	.0969444	-20.58	0.000	-2.185137	-1.805122

Źródło: Opracowanie własne.

Zmienne w modelu Probit po transformacji trzech zmiennych zależnych są łącznie istotne, wskazują na to wartości statystyki testowej (LR chi2(6) = 737.21) i p-value (Prob > chi2 = 0.0000 < 0,05). Również wszystkie zmienne są osobno istotne - p-value każdej z nich jest mniejsze niż 5%. Na podstawie Pseudo R^2 można stwierdzić, iż model wyjaśnia 56,48% zmienności zmiennej objaśnianej. Przeprowadzenie ponownych testów poprawności funkcyjnej modelu pozwoli odpowiedzieć na pytanie czy transformacja zmiennych przyniosła pożądany skutek.

Tabela 41. Test linktest dla modelu Probit po transformacji

```
. linktest
Iteration 0:   log likelihood = -652.65521
Iteration 1:   log likelihood = -295.01947
Iteration 2:   log likelihood = -284.04447
Iteration 3:   log likelihood = -283.70692
Iteration 4:   log likelihood = -283.70401
Iteration 5:   log likelihood = -283.70401
Probit regression                                 Number of obs   =       1657
                                                  LR chi2(2)      =     737.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -283.70401                       Pseudo R2       =     0.5653
```

oscar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	.9586166	.0668595	14.34	0.000	.8275743	1.089659
_hatsq	-.035424	.0407709	-0.87	0.385	-.1153334	.0444854
_cons	.0404994	.09177	0.44	0.659	-.1393666	.2203653

Źródło: Opracowanie własne.

Test linktest wskazuje, iż należy odrzucić hipotezę zerową o nieistotności zmiennej *_hat* (p-value = 0,0000 < 0,05) przy 5%-owym poziomie istotności. P-value zmiennej *_hatsq* równe 0,385 czyli zdecydowanie większe od 0,05 sugeruje, iż nie ma podstaw do odrzucenia hipotezy zerowej mówiącej o nieistotności tej zmiennej. Oznacza to, iż dla wartości dopasowanych istotne są tylko pierwsze potęgi zmiennych - forma funkcyjna modelu jest poprawna.

Tabela 42. Test jakości dopasowania dla modelu Probit po transformacji

```
. estat gof
Probit model for oscar, goodness-of-fit test
       number of observations =      1657
 number of covariate patterns =      1654
          Pearson chi2(1646) =      1462.64
                Prob > chi2 =         0.9995
```

Źródło: Opracowanie własne.

Statystyka testu Pearsona równa 1462,64 i p-value wynoszące 0,9995 (zdecydowanie większe od 5%) wskazują na brak podstaw do odrzucenia h0 o poprawnej formie funkcyjnej modelu.

Tabela 43. Test Hosmera-Lemeshowa dla modelu Probit po transformacji

```
. lfit, group(10) table
Probit model for oscar, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0096	1	1.5	165	164.5	166
2	0.0102	1	1.6	165	164.4	166
3	0.0116	1	1.8	165	164.2	166
4	0.0231	3	2.6	162	162.4	165
5	0.0240	4	3.9	162	162.1	166
6	0.0260	6	4.1	160	161.9	166
7	0.0367	1	4.9	164	160.1	165
8	0.0911	15	10.4	151	155.6	166
9	0.5817	49	48.3	117	117.7	166
10	1.0000	141	142.1	24	22.9	165

```
      number of observations =      1657
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =         7.27
                Prob > chi2 =         0.5078
```

Źródło: Opracowanie własne.

Tabela 44. Porównanie modeli Probit przed i po transformacji

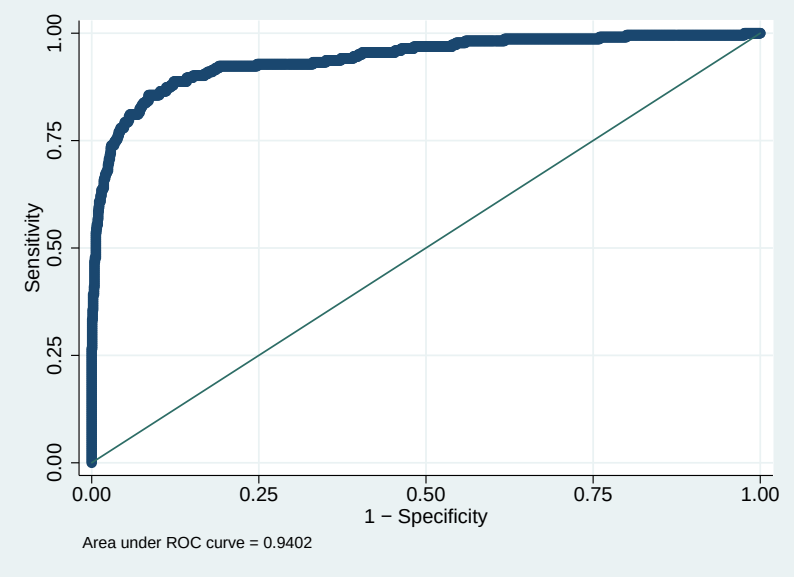
```
. fitstat, using(m1) force
Measures of Fit for probit of oscar
                               Current             Saved        Difference
Model:                          probit            probit
N:                                1657              1657                 0
Log-Lik Intercept Only         -652.655          -652.655             0.000
Log-Lik Full Model             -284.052          -241.258           -42.794
D                               568.104(1649)     482.516(1649)      85.588(0)
LR                              737.207(6)        822.795(6)         85.588(0)
Prob > LR                         0.000             0.000                .
McFadden s R2                     0.565             0.630            -0.066
McFadden s Adj R2                 0.553             0.618            -0.066
ML (Cox-Snell) R2                 0.359             0.391            -0.032
Cragg-Uhler(Nagelkerke) R2        0.659             0.718            -0.059
McKelvey & Zavoina s R2           0.602             0.728            -0.127
Efron s R2                        0.591             0.636            -0.045
Tjur s Discrimination Coef        0.594             0.641            -0.047
Variance of y*                    2.511             3.681            -1.170
Variance of error                 1.000             1.000             0.000
Count R2                          0.937             0.941            -0.004
Adj Count R2                      0.527             0.559            -0.032
AIC                             584.104           498.516            85.588
AIC/N                             0.353             0.301             0.052
BIC                             619.993           534.405            85.588
Difference of   85.588 in BIC provides very strong support for saved model.
Note: p-value for difference in LR is only valid if models are nested.
```

Źródło: Opracowanie własne.

Podobnie statystyka testowa Hosmera-Lemeshowa równa 7,27 i p-value większe od 5% (Prob > chi2 = 0.5078) sugerują, iż nie ma podstaw do odrzucenia hipotezy zerowej o prawidłowości formy funkcyjnej modelu na 5% poziomie istotności.

Wszystkie testy diagnostyczne wykazały, iż model po transformacjach ma prawidłową formą funkcyjną, aby zobaczyć jak zlogarytmowanie trzech zmiennych objaśniających wpłynęło na współczynniki determinacji modelu należy posłużyć się komendą *fitstat* w programie Stata. Porównanie modeli przed (*Saved model*) i po transformacji (*Current model*) prezentuje tabela 44. Niestety podobnie jak w przypadku modelu Logit tak i w tym przypadku zlogarytmowanie trzech zmiennych niezależnych spowodowało pogorszenie niemal wszystkich statystyk. Kreteria informacyjne BIC i AIC są wyraźnie niższe dla modelu przed transformacjami, lepiej przewiduje on też sukcesy i porażki w modelu. Mimo to w dalszych analizach zostanie zastosowany model ze zlogarytmowanymi zmiennymi: nominacje, zlote_globy i bafta, gdyż ma on prawidłową formę funkcyjną. Model ten dla zmiennej ukrytej y_i^* byłby wyjaśniany w 60,2%, gdyby zmienna ta była bezpośrednio obserwowalna (McKelvey-Zavoina R2). Przewiduje on trafnie sukcesy i porażki w 93,7% na co wskazuje liczebnościowe R^2. Skorygowane liczebnościowe R^2 (Adj Count R2) pokazuje natomiast, że dzięki zmiennym objaśniającym model trafnie przewiduje sukcesy i porażki w 52,7%.

Rysunek 7: Krzywa ROC



Źródło: Opracowanie własne.

Tabela 45. Pole pod krzywą ROC

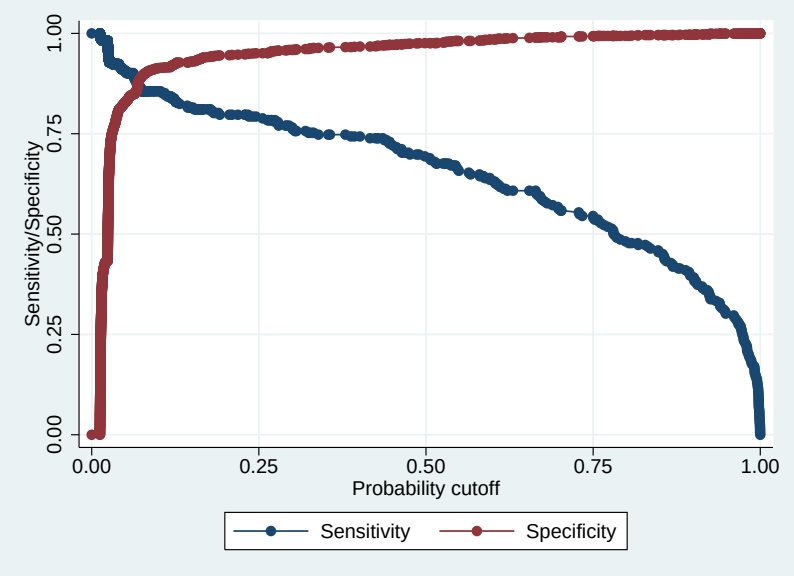
```
. lroc
Probit model for oscar
number of observations =      1657
area under ROC curve   =    0.9402
```

Źródło: Opracowanie własne.

Pole pod krzywą ROC jest to procent dobrze zakwalifikowanych pozytywnych i negatywnych odpowiedzi przy różnych wyborach punktu rozdzielającego decyzję. Im bliższa jedynce

wartość pola tym model jest lepiej dopasowany do danych wejściowych. Tabela 45 wskazuje na bardzo dobre dopasowanie modelu - pole pod krzywą ROC wynosi 0,9402.

Rysunek 8: Wrażliwość i Specyficzność



Źródło: Opracowanie własne.

Wykres wrażliwości i specyficzności wskazuje prawdopodobieństwo ucięcia, czyli poziom rozgraniczenia maksymalizującego trafność klasyfikacji wartości dopasowanych. Z rysunku 8 wynika, iż najlepsze ucięcie powinno mieć miejsce w punkcie 0,05. Dla tego punktu wyznaczona została tablica klasyfikacji.

Tabela 46. Tablica klasyfikacji dla punktu odcięcia 0,05

```
. lstat,cutoff(0.05)
Probit model for oscar
                 ———— True ————
Classified |         D          ~D |     Total
-----------+----------------------+----------
         + |       202         249 |       451
         - |        20        1186 |      1206
-----------+----------------------+----------
     Total |       222        1435 |      1657
Classified + if predicted Pr(D) >= .05
True D defined as oscar != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   90.99%
Specificity                     Pr( -|~D)   82.65%
Positive predictive value       Pr( D| +)   44.79%
Negative predictive value       Pr(~D| -)   98.34%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   17.35%
False - rate for true D         Pr( -| D)    9.01%
False + rate for classified +   Pr(~D| +)   55.21%
False - rate for classified -   Pr( D| -)    1.66%
-------------------------------------------------
Correctly classified                        83.77%
-------------------------------------------------
```

Źródło: Opracowanie własne.

Tablica klasyfikacji dla punktu odcięcia równego 0,05 wskazuje, iż model poprawnie przewidział aż 202 spośród 222 filmów, które zdobyły Oscara, co oznacza, że wrażliwość modelu wyniosła 90,99%. Specyficzność z kolei równa 82,65 wskazuje, iż model poprawnie zakwalifikował 1189 obserwacji spośród 1435 jako te, które statuetki nie uzyskały. Ogólny wskaźnik klasyfikacji modelu równy 83,77%, wydaje się bardzo dobrym wynikiem. W związku z tym, iż współliniowość zmiennych objaśniających: _Igatunek_2, kraj_prod, przychody2000, ln_nom, ln_zg, ln_baf, milosc była już badana przy okazji modelu Logit nie ma potrzeby przeprowadzać testu na współliniowość ponownie.

3.4 Wybór najlepszego modelu

W związku z tym, iż Liniowy Model Prawdopodobieństwa cechował się heteroskedastycznością reszt i wartościami dopasowanymi wykraczającymi poza przedział [0,1], wyboru modelu najlepiej reprezentującego dane należy dokonać pomiędzy modelem logitowym i probitowym. Dla obu tych modeli ich zagnieżdżone wersje nie miały prawidłowej formy funkcyjnej stąd też wybór będzie dokonywany pomiędzy dwoma modelami zawierającymi zlogarytmowane zmienne niezależne: ln_nom, ln_zg i ln_baf.

Tabela 47. Porównanie modeli Logit i Probit po transformacji

```
. fitstat, dif force
Measures of Fit for probit of oscar
Warning: Current model estimated by probit, but saved model estimated by logit
                              Current             Saved        Difference
Model:                         probit             logit
N:                               1657              1657                 0
Log-Lik Intercept Only        -652.655          -652.655             0.000
Log-Lik Full Model            -284.052          -286.079             2.027
D                              568.104(1649)     572.157(1649)        4.054(0)
LR                             737.207(6)        733.153(7)           4.054(1)
Prob > LR                        0.000             0.000              0.044
McFadden s R2                    0.565             0.562              0.003
McFadden s Adj R2                0.553             0.549              0.003
ML (Cox-Snell) R2                0.359             0.358              0.002
Cragg-Uhler(Nagelkerke) R2       0.659             0.656              0.003
McKelvey & Zavoina s R2          0.602             0.610             -0.008
Efron s R2                       0.591             0.590              0.000
Tjur s Discrimination Coef       0.594             0.596             -0.002
Variance of y*                   2.511             8.425             -5.914
Variance of error                1.000             3.290             -2.290
Count R2                         0.937             0.937             -0.001
Adj Count R2                     0.527             0.532             -0.005
AIC                            584.104           588.157             -4.054
AIC/N                            0.353             0.355             -0.002
BIC                            619.993           631.460            -11.467
Difference of   11.467 in BIC provides very strong support for current model.
Note: p-value for difference in LR is only valid if models are nested.
```

Źródło: Opracowanie własne.

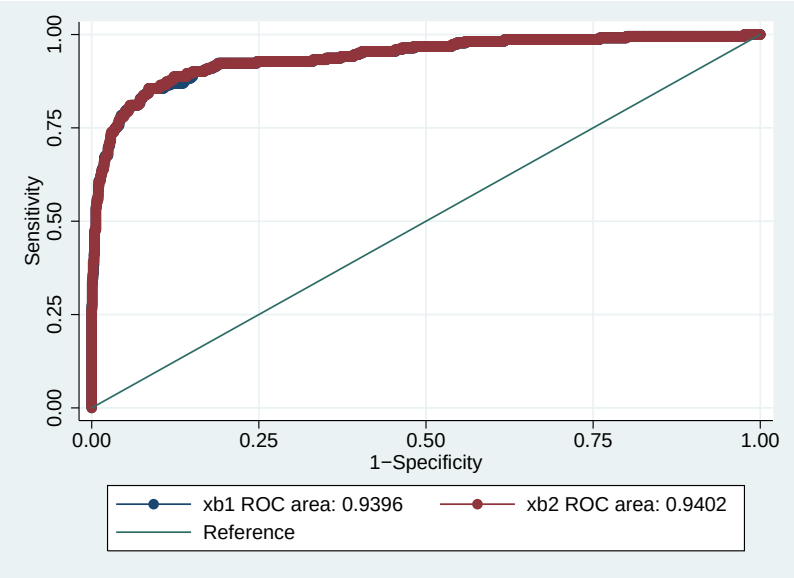
Różnica pomiędzy modelami dla większości wskaźników determinacji jest niewielka, modele te są ze sobą silnie skorelowane (tabela 48). R^2 McFaddena, skorygowane R^2 McFaddena, R^2 Efrona, kryteria informacyjne AIC i BIC oraz pole pod krzywą ROC (rysunek 7, tabela 49) wskazują, iż lepiej dopasowanym modelem jest model probitowy. Z kolei R^2 McKelvey'a & Zavoina, skorygowane liczebnościowe R^2 oraz analiza literatury wskazują, iż do interpretacji oszacowań modelu należałoby użyć modelu logitowego.

Tabela 48. Współczynnik korelacji pomiędzy modelami Logit i Probit

```
. pwcorr prlogit prprobit
             | prlogit prprobit
-------------+------------------
     prlogit |  1.0000
    prprobit |  0.9994   1.0000
```

Źródło: Opracowanie własne.

Rysunek 9: Krzywe ROC dla Logitu (niebieska linia) i Probitu (czerwona linia)



Źródło: Opracowanie własne.

Tabela 49. Porównanie krzywych ROC dla Logitu i Probitu

```
. roccomp oscar logit_fin probit_fin, graph summary
                                ROC                    -Asymptotic Normal--
                    Obs       Area     Std. Err.      [95% Conf. Interval]
---------------------------------------------------------------------------
logit_fin          1657     0.9396       0.0100        0.91998     0.95917
probit_fin         1657     0.9402       0.0100        0.92067     0.95975
---------------------------------------------------------------------------
Ho: area(logit_fin) = area(probit_fin)
    chi2(1) =     4.60       Prob>chi2 =   0.0319
```

Źródło: Opracowanie własne.

W związku z wyraźnym wskazaniem bayesowskiego kryterium informacyjnego na model Probit (BIC mniejsze o 11,467) oraz z większością wskaźników determinacji bardziej korzystnych dla niego to właśnie ten model powinien być uznawany za najlepiej opisujący prawdopodobieństwo otrzymania nagrody Amerykańskiej Akademii Sztuki i Wiedzy Filmowej.

W załączniku nr 1 do niniejszej pracy przedstawiona jest tabela zawierająca oszacowania parametrów i najważniejsze statystyki wszystkich modeli przedstawionych w tym rozdziale.

Rozdział 4

Interpretacja wyników

W rozdziale tym przedstawiona zostanie interpretacja wyników regresji dla modeli Logit i Probit z trzema zmiennymi zlogarytmowanymi, gdyż tylko te modele posiadają prawidłową formę funkcyjną. Interpretacja oszacowań modelu LMP ze względu na heteroskedastyczność reszt mogłaby prowadzić do nieprawidłowych wniosków wynikających z możliwych błędnych wskazań modelu co do istotności niektórych zmiennych. Interpretacji w modelach logitowym i probitowym podlegać będą znaki przy oszacowanych parametrach oraz efekty cząstkowe. Dodatkowo dla modelu Logit wyznaczone zostaną ilorazy szans.

4.1. Interpretacja znaków przy parametrach modeli

Interpretację wyników należy rozpocząć od analizy znaków przy parametrach zmiennych niezależnych w modelach Logit i Probit. Samych współczynników w żadnym z tych modeli nie należy interpretować wprost - w przypadku modelu probit nie mają one żadnej bezpośredniej interpretacji, a w przypadku modelu logit oznaczają procentowy wpływ jednostkowej zmiany wartości zmiennej objaśniającej na iloraz szans.

Tabela 50. Parametry zmiennych niezależnych w modelach Logit i Probit

```
. est table log_final prob_final
```

Variable	log_final	prob_final
_Igatunek_2	1.2988123	.65770695
kraj_produ~i	-.67950148	-.36840234
przycho~2000	9.583e-10	4.704e-10
ln_nom	2.4378704	1.3302167
ln_zg	1.9138695	1.0424347
ln_baf	1.6201441	.89365612
milosc	1.0767324	.52131938
_cons	-3.7079477	-1.9951296

Źródło: Opracowanie własne.

Interpretacja znaków przy parametrach jest częściowo zgodna z kierunkami zawartymi w hipotezach badawczych niniejszej pracy:

- Film animowany ma większe prawdopodobieństwo zdobycia Oscara niż dramat (nie zakładano tego zjawiska w hipotezach),
- Film wyprodukowany wyłącznie przez Stany Zjednoczone ma mniejsze prawdopodobieństwo zdobycia statuetki niż film wyprodukowany przez inne kraje lub przez Stany Zjed-

noczone we współpracy z innymi krajami (hipoteza badawcza zakładała, iż filmy amerykańskie powinny mieć większe prawdopodobieństwo zdobycia nagrody),

- Wraz ze wzrostem przychodów z kas biletowych prawdopodobieństwo zdobycia nagrody Akademii rośnie (zgodne z hipotezą),
- Im więcej nominacji otrzymał dany film tym większe prawdopodobieństwo, że zdobędzie Oscara (zgodne z hipotezą) [5],
- Im więcej Złotych Globów zdobył film tym większe prawdopodobieństwo, że zdobędzie Oscara (zgodne z hipotezą)[6],
- Im więcej nagród BAFTA otrzymał film tym większe jest prawdopodobieństwo, że otrzyma on również Oscara (zgodne z hipotezą) [7],
- Film z wyraźnie zarysowanym wątkiem miłosnym ma większe prawdopodobieństwo zdobycia statuetki niż film pozbawiony tego wątku (hipoteza zakładała, iż wątek miłosny nie powinien wpływać na prawdopodobieństwo zdobycia Oscara).

4.2. Efekty cząstkowe w modelach Logit i Probit

Efekty cząstkowe mierzą jednostkową reakcję zmiennej zależnej na jednostkową zmianę zmiennej niezależnej, oblicza się je według następującego wzoru:

$$\frac{\partial E(y|x)}{\partial x_i} = f(x\beta)\beta_i = \Phi(x\beta)\beta_i \tag{4}$$

Efekty cząstkowe są niezwykle użyteczne, gdyż można je bezpośrednio porównywać pomiędzy modelami.

4.2.1. Model Logit

Tabela 51. Efekty cząstkowe w modelu Logit

```
. mfx compute
Marginal effects after logit
      y  = Pr(oscar) (predict)
         =  .05364139
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
_Igatu_2*	.1123638	.05289	2.12	0.034	.008703	.216025	.051298
kraj_p_i*	-.0362143	.01344	-2.70	0.007	-.062551	-.009878	.558841
prz_2000	4.86e-11	.00000	2.03	0.042	1.7e-12	9.6e-11	1.6e+08
ln_nom	.123756	.01334	9.28	0.000	.097607	.149905	.29546
ln_zg	.0971556	.03981	2.44	0.015	.019137	.175175	.045141
ln_baf	.082245	.02409	3.41	0.001	.035021	.129469	.063922
milosc*	.0830142	.03823	2.17	0.030	.008088	.157941	.079662

```
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Źródło: Opracowanie własne.

[5] Fakt zlogarytmowania zmiennej nie wpływa na znak przy jej oszacowanym parametrze.
[6] Patrz odnośnik 5.
[7] Patrz odnośnik 5.

Gdy zmienne objaśniające przyjmują wartości na poziomie swoich średnich prawdopodobieństwo zdobycia Oscara wynosi około 5,36% ($y = Pr(oscar)(predict) = .05364139$).

Efekty cząstkowe (dy/dx) w tym modelu powinny być interpretowane w następujący sposób:

- Film animowany ma wyższe prawdopodobieństwo dostania Oscara o około 11,24 punktu procentowego od filmu, którego gatunkiem jest dramat, przy charakterystykach na poziomie średnich w próbie,
- Film wyprodukowany wyłącznie przez Stany Zjednoczone ma prawdopodobieństwo zdobycia Oscara nisze o około 3,62 punkty procentowe niż film wyprodukowany przez inne kraje lub przez Stany Zjednoczone we współpracy z innymi krajami, przy charakterystykach na poziomie średnich w próbie,
- Wzrost przychodów z kas kinowy filmu o jeden milion dolarów powoduje wzrost prawdopodobieństwa zdobycia Oscara o 0,00486 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,
- Filmy z wyraźnym wątkiem miłosnym mają wyższe prawdopodobieństwo dostania statuetki o około 8,3 punktu procentowego niż filmy bez tego wątku, przy charakterystykach na poziomie średnich w próbie,

W przypadku zlogarytmowanych zmiennych niezależnych efekty cząstkowe interpretowane są nieco inaczej. Na przykład dla zmiennej nominacje prawidłowo policzony efekt cząstkowy powinien wyglądać następująco:

$$\frac{\partial F(\beta_{ln\_nom} ln\_nom)}{\partial nominacje} = f(x\beta)\frac{\beta_{ln\_nom}}{nominacje} \tag{5}$$

Aby móc zinterpretować efekty cząstkowe należy posłużyć się średnimi wartościami zmiennych nominacje, złote_globy i bafta, które przedstawia tabela 52.

Tabela 52. Charakterystyki zmiennych nominacje, zlote_globy i bafta

```
. sum nominacje zlote_globy bafta
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   nominacje |       1663    1.157547    2.441459         0         15
zlote_globy |       1663    .1906194    .6791433         0          6
      bafta |       1663    .2489477     .836493         0          7
```

Źródło: Opracowanie własne.

Faktyczne efekty cząstkowe dla zmiennych nominacje, zlote_globy i bafta wynoszą:

$$\frac{\partial F(\beta_{ln\_nom} ln\_nom)}{\partial nominacje} = f(x\beta)\frac{\beta_{ln\_nom}}{\overline{nominacje}} = \frac{0,123756}{1,157547} \approx 0,1071 \tag{6}$$

$$\frac{\partial F(\beta_{ln\_zg} ln\_zg)}{\partial zlote\_globy} = f(x\beta)\frac{\beta_{ln\_zg}}{\overline{zlote\_globy}} = \frac{0,0971556}{0,1906194} \approx 0,5097 \tag{7}$$

$$\frac{\partial F(\beta_{ln\_baf} ln\_baf)}{\partial bafta} f(x\beta)\frac{\beta_{ln\_baf}}{\overline{bafta}} = \frac{0,082245}{0,2489477} \approx 0,3304 \tag{8}$$

Interpretacja dla prawidłowych efektów cząstkowych będzie następująca:

- Wzrost liczby nominacji o jednostkę powoduje wzrost prawdopodobieństwa zdobycia Oscara o około 10,71 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,
- Wzrost liczby Złotych Globów o jednostkę powoduje wzrost prawdopodobieństwa zdobycia Oscara o około 50,97 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,
- Wzrost liczby nagród BAFTA o jednostkę powoduje wzrost prawdopodobieństwa zdobycia Oscara o około 33,04 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,

4.2.1. Model Probit

Tabela 53. Efekty cząstkowe w modelu Probit

```
. mfx compute
Marginal effects after probit
      y  = Pr(oscar) (predict)
         =  .06026116
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
_Igatu_2*	.1202109	.0528	2.28	0.023	.016717	.223705	.051298
kraj_p_i*	-.0458747	.01531	-3.00	0.003	-.075875	-.015874	.558841
prz_2000	5.62e-11	.00000	2.07	0.039	2.9e-12	1.1e-10	1.6e+08
ln_nom	.1589991	.01512	10.52	0.000	.129364	.188634	.29546
ln_zg	.1246009	.04652	2.68	0.007	.033425	.215776	.045141
ln_baf	.1068176	.0288	3.71	0.000	.05038	.163256	.063922
milosc*	.0862263	.03858	2.24	0.025	.010611	.161842	.079662

```
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Źródło: Opracowanie własne.

Gdy zmienne objaśniające przyjmują wartości na poziomie swoich średnich prawdopodobieństwo zdobycia Oscara wynosi około 6,03% ($y = Pr(oscar)(predict) = .06026116$).

Efekty cząstkowe (dy/dx) w tym modelu powinny być interpretowane w następujący sposób:

- Film animowany ma wyższe prawdopodobieństwo dostania Oscara o około 12,02 punktu procentowego od filmu, którego gatunkiem jest dramat, przy charakterystykach na poziomie średnich w próbie,
- Film wyprodukowany wyłącznie przez Stany Zjednoczone ma prawdopodobieństwo zdobycia Oscara nisze o około 4,59 punkty procentowe niż film wyprodukowany przez inne kraje lub przez Stany Zjednoczone we współpracy z innymi krajami, przy charakterystykach na poziomie średnich w próbie,
- Wzrost przychodów z kas kinowy filmu o jeden milion dolarów powoduje wzrost prawdopodobieństwa zdobycia Oscara o 0,00562 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,
- Filmy z wyraźnym wątkiem miłosnym mają wyższe prawdopodobieństwo dostania statuetki o około 8,62 punktu procentowego niż filmy bez tego wątku, przy charakterystykach na poziomie średnich w próbie,

Podobnie jak w przypadku efektów cząstkowych dla modelu logitowego, aby móc interpretować efekty cząstkowe zmiennych nominacje, zlote_globy i bafta należy podzielić oszacowane efekty (dla zmiennych zlogarytmowanych) przez średnią wartość tych zmiennych. Faktyczne efekty cząstkowe dla zmiennych nominacje, zlote_globy i bafta wynoszą:

$$\frac{\partial F(\beta_{ln\_nom} ln\_nom)}{\partial nominacje} = f(x\beta)\frac{\beta_{ln\_nom}}{\overline{nominacje}} = \frac{0,1589991}{1,157547} \approx 0,1373 \tag{9}$$

$$\frac{\partial F(\beta_{ln\_zg} ln\_zg)}{\partial zlote\_globy} = f(x\beta)\frac{\beta_{ln\_zg}}{\overline{zlote\_globy}} = \frac{0,1246009}{0,1906194} \approx 0,6567 \tag{10}$$

$$\frac{\partial F(\beta_{ln\_baf} ln\_baf)}{\partial bafta} = f(x\beta)\frac{\beta_{ln\_baf}}{\overline{bafta}} = \frac{0,1068176}{0,2489477} \approx 0,4291 \tag{11}$$

Interpretacja dla prawidłowych efektów cząstkowych będzie następująca:

- Wzrost liczby nominacji o jednostkę powoduje wzrost prawdopodobieństwa zdobycia Oscara o około 13,73 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,
- Wzrost liczby Złotych Globów o jednostkę powoduje wzrost prawdopodobieństwa zdobycia Oscara o około 65,67 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,
- Wzrost liczby nagród BAFTA o jednostkę powoduje wzrost prawdopodobieństwa zdobycia Oscara o około 42,91 punktu procentowego, przy charakterystykach na poziomie średnich w próbie,

Efekty cząstkowe dla niektórych zmiennych takich jak zlote_globy, cz bafta mogą wydawać się zadziwiająco wysokie. Wynika to z faktu, iż są one obliczane dla wartości średnich zmiennych niezależnych, co sprawia, iż dla faktycznie zaobserwowanych wartości mogą się one znacząco różnić.

Tabela 54. Efekty cząstkowe Logit i Probit

	Logit	Probit
_Igatunek_2	0.112**	0.120**
kraj_produkcji	-0.0362***	-0.0459***
przychody2000	4.86e-11**	5.62e-11**
ln_nom	0.1071***	0.1373***
ln_zg	0.5097**	0.6567***
ln_baf	0.3304***	0.4291***
milosc	0.0830**	0.0862**

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Źródło: Opracowanie własne.

Efekty cząstkowe w modelu Probit dla wszystkich zmiennych niezależnych okazały się wyższe niż efekty cząstkowe w modelu Logit. W związku z tym, iż jako lepiej dopasowany model uznany został Probit to jego oszacowania powinny być przede wszystkim brane pod uwagę

przy weryfikacji hipotez badawczych. Ciekawym przypadkiem do interpretacji jest zmienna przychody2000, pomimo, iż jest ona istotna w obu modelach na 5%-owym poziomie istotności to jednak jej oszacowany wpływ na prawdopodobieństwo zdobycia Oscara jest minimalny - przy przychodach z kas biletowych na poziomie średnim z próby (1,61 mln dolarów) zmienna ta wpływa na prawdopodobieństwo badanego zjawiska z siłą słabszą niż 1 punkt procentowy (około 0,8 pp).

W załączniku nr 2 do niniejszej pracy znajduje się tabela z efektami cząstkowymi najważniejszych modeli LMP, Logit i Probit oszacowanych w tej pracy. Dzięki efektom cząstkowym można te modele bezpośrednio porównywać.

Rozdział 5

Wnioski końcowe

Przeprowadzona estymacja miała na celu oszacowanie czynników wpływających na prawdopodobieństwo zdobycia Oscara przez film choć w jednej kategorii. Badanie zawarte w niniejszej pracy miało stanowić uogólnienie przeprowadzonych dotychczas badań dotyczących prawdopodobieństw zdobycia nagrody Amerykańskiej Akademii Sztuki i Wiedzy Filmowej w konkretnych kategoriach.

Wyniki przeprowadzonej w niniejszej pracy analizy wskazują, iż na prawdopodobieństwo zdobycia Oscara wpływa ograniczona liczba czynników, z których najważniejszymi są te dotyczące innych nagród i nominacji dla danego filmu. Wnioski te pokrywają się z wnioskami zawartymi w literaturze dotyczącej badanego zagadnienia. Głównymi czynnikami wpływającymi na prawdopodobieństwo zdobycia najważniejszej w świeci filmowym statuetki są: gatunku, kraju produkcji, przychodów z kas kinowych, liczby nominacji do Oscara, liczby zdobytych Złotych Globów, liczby zdobytych nagród Brytyjskiej Akademii Sztuk Filmowych i Telewizyjnych i faktu czy w filmie występuje wiodący wątek miłosny.

Spośród zmiennych nieistotnych w modelu najbardziej dziwi nieistotność zmiennej zerojedynkowej ekranizacja, która reprezentowała filmy będące adaptacją powieści, biografii, sztuki lub artykułu. Większość najwybitniejszych dzieł filmowych powstało właśnie w oparciu o scenariusz jakiegoś wybitnego dzieła literackiego, stąd na wstępie pracy została zawarta hipoteza, iż zmienna ta będzie istotnie wpływać na badane zjawisko. Dziwi również, iż z dziewięciu głównych gatunków filmowych tylko filmy animowane okazały się istotnie wpływać na prawdopodobieństwo zdobycia Oscara. Hipotezy o tym, iż gatunki dramat i komedia powinny mieć wpływ na szacowane zjawisko należy odrzucić jako nieistotne. Są to wnioski stojące w opozycji do badań Dawida Kaplana i Andrew B. Bernarda omówionych na wstępie pracy.

Zarówno badanie zawarte w niniejszej pracy jak i badania omówione w literaturze wskazują wyraźnie, iż czynnikami najsilniej wpływającymi na prawdopodobieństwo zdobycia Oscara są: liczba nominacji, liczba zdobytych Złotych Globów oraz liczba zdobytych nagród BAFTA. Zmienne te były istotne we wszystkich oszacowanych modelach na niemal każdym poziomie istotności. Również ich krańcowy wpływ na modelowane prawdopodobieństwo znacznie przewyższał wpływ pozostałych zmiennych.

Nie potwierdziły się hipotezy badawcze dotyczące kraju produkcji filmu oraz wiodącego wątku miłosnego. Zakładały one, iż filmy powstałe wyłącznie w Stanach Zjednoczonych będą lepiej oceniane przez członków Akademii niż pozostałe filmy oraz występowanie wiodącego wątku miłosnego miało nie wpływać na prawdopodobieństwo zdobycia Oscara. Okazało się, że filmy tworzone poza USA lub przy współpracy USA mają większą szansę na statuetkę niż filmy czysto amerykańskie. Również wiodący wątek miłosny był zmienną istotną w modelu

wpływając znacznie na modelowane zjawisko.

Jedyną zmienną ekonomiczną, która okazała się być istotną w estymowanych modelach ekonometrycznych była zmienna przychody z kas kinowych. Jej wpływ okazał się jednak marginalny - efekty cząstkowe pokazały, iż dopiero wzrost przychodów liczony w setkach milionów dolarów wpływa znacząco na szacowane prawdopodobieństwo. W związku z takimi wynikami regresji należy stwierdzić, iż czynniki ekonomiczne nie wpływają, lub wpływają w zupełnie minimalnym stopniu na prawdopodobieństwo zdobycia Oscara. Umieszczenie czynników ekonomicznych w modelu pośrednio na przekór literaturze nie opłaciło się, model zweryfikował tę decyzję na korzyść literatury.

Wydaje się, iż zaprezentowane w niniejszej pracy badanie przyczynia się do lepszego zrozumienia czynników wpływających na prawdopodobieństwo zdobycia Oscara, wskazuje które czynniki są najbardziej istotne i mają największy wpływ na decyzje podejmowane przez członków Amerykańskiej Akademii Sztuki i Wiedzy Filmowej.

Bibliografia

[1] Bernard A.B. [2005], *An Index of Oscar-Worthiness: Predicting the Academy Award for Best Picture*, Hanover.

[2] Chen X. i in. [2010], *Logistic Regression with Stata*, `http://www.ats.ucla.edu/stat/stata/webbooks/logistic/`, dostęp: czerwiec 2013, rozdz. 1-3.

[3] Helmer E. [2013], *The Value of an Oscar*, `http://boxofficequant.com/the-value-of-an-oscar/`, Reuters, dostęp: czerwiec 2013.

[4] Kaplan D. [2006], *And the Oscar Goes to...A Logistic Regression Model for Predicting Academy Award Results*, Journal Of Applied Economics and Policy, XXV, s.23-41.

[5] Krauss J. i in. [2008], *Predicting movie success and acadamy awards through sentiment and social network analysis*, Koeln.

[6] Mycielski J. [2010], *Ekonometria*, Warszawa.

[7] Nelson R. A. i in. [2001], *What's an Oscar Worth?*, Economic Inquiry, 39(1).

[8] Pardoe I., Simonton D.K. [2008], *Applying discrete choice models to predict Academy Award winners*, The Journal of the Royal Statistical Society, Series A, s.375-394.

[9] Pardoe I. [2012], *Applied Regression Modeling*, Oregon, rozdz.7.2.

[10] Royston P., Ambler G. [1999], *Nonlinear regression models involving power or exponential functions of covariates*, Stata Technical Bulletin 49, s.25-30.

[11] Simonton D. K. [2005], *Cinematic creativity and production budgets: does money make the movie?*, Journal of Creative Behavior, nr 39, s.1-15.

Załączniki

Załącznik 1. Oszacowania parametrów modeli LMP, Logit i Probit

	Logit			Probit			Liniowy Model Prawdopodobieństwa (LMP)	
	Logit Ogólny	Logit Zagnieżdżony[1]	Logit po transformacji[2]	Probit Ogólny	Probit Zagnieżdżony	Probit po transformacji	LMP Ogólny	LMP Zagnieżdżony
budzet2000	-2.59e-09 (4.67e-09)			-2.05e-09 (2.36e-09)			-7.00e-11 (1.83e-10)	
_Igatunek_1	-0.237 (0.483)			-0.138 (0.228)			-0.0118 (0.0164)	
_Igatunek_2	0.947 (0.607)	**1.099*****[4] (0.419)	**1.299***** (0.410)	0.473 (0.312)	**0.539**** (0.227)	**0.658***** (0.213)	**0.0668**** (0.0302)	**0.0756***** (0.0252)
_Igatunek_3	0.816 (0.639)			0.343 (0.335)			0.0291 (0.0302)	
_Igatunek_4	-0.587 (0.791)			-0.288 (0.398)			-0.0369 (0.0287)	
_Igatunek_5	0.413 (0.602)			0.193 (0.296)			0.0179 (0.0241)	
_Igatunek_6	0.900 (0.585)			0.383 (0.298)			**0.0453*** (0.0272)	**0.0504**** (0.0249)
_Igatunek_7	0.014 (0.620)			-0.0401 (0.313)			0.00251 (0.0242)	
_Igatunek_8	0.09692 (0.813)			-0.0275 (0.385)			-0.00506 (0.0256)	
_Igatunek_9	-0.493 (1.192)			-0.300 (0.611)			-0.0411 (0.0356)	
ekranizacja	-0.144 (0.288)			-0.0508 (0.144)			-0.00985 (0.0129)	
roi	-0.000133 (0.0000994)			-0.0000484 (0.0000469)			-1.70e-08 (0.000000162)	
przychody2000	**2.11e-09**** (8.44e-10)	**1.31e-09***** (4.83e-10)	**9.58e-10**** (4.65e-10)	**1.03e-09**** (4.26e-10)	**6.53e-10***** (2.34e-10)	**4.70e-10**** (2.25e-10)	4.52e-11 (2.77e-11)	
kraj_prod	**-0.731***** (0.277)	**-0.735***** (0.265)	**-0.680***** (0.242)	**-0.374***** (0.137)	**-0.381***** (0.131)	**-0.368***** (0.120)	**-0.0278**** (0.0113)	**-0.0262**** (0.0110)
nominacje	**0.718***** (0.0710)	**0.677***** (0.0602)		**0.384***** (0.0355)	**0.368***** (0.0307)		**0.0812***** (0.00367)	**0.0826***** (0.00345)

zlote_globy	**1.709*****	**1.642*****		**0.806*****	**0.795*****		**0.0748*****	**0.0748*****
	(0.283)	(0.265)		(0.132)	(0.127)		(0.0108)	(0.0107)
bafta	**0.720*****	**0.766*****		**0.373*****	**0.398*****		**0.0485*****	**0.0491*****
	(0.171)	(0.169)		(0.0859)	(0.0842)		(0.00864)	(0.00856)
milosc	**1.216*****	**1.040*****	**1.077*****	**0.578*****	**0.515*****	**0.521*****	**0.0558*****	**0.0496****
	(0.397)	(0.374)	(0.354)	(0.204)	(0.195)	(0.181)	(0.0202)	(0.0199)
czas_trwania	-0.00665			-0.00361			**-0.00076****	**-0.000737*****
	(0.008001)			(0.00399)			(0.000329)	(0.000282)
ln_nom			**2.438*****			**1.330*****		
			(0.166)			(0.0861)		
ln_zg			**1.914*****			**1.042*****		
			(0.710)			(0.359)		
ln_baf			**1.620*****			**0.894*****		
			(0.420)			(0.219)		
_cons	**-3.414*****	**-4.076*****	**-3.708*****	**-1.793*****	**-2.213*****	**-1.995*****	**0.103*****	**0.0991*****
	(0.964)	(0.238)	(0.214)	(0.476)	(0.110)	(0.0969)	(0.0388)	(0.0327)
Liczba Obserwacji	1638	1657	1657	1638	1657	1657	1638	1663
Skorygowane R^2							0.593	0.591
Pseudo R^2	0.0.634	0.624	0.562	0.639	0.630	0.565		
AIC	511.6	506.8	588.2	505.1	496.5	582.1	-332.6	-343.0
BIC	619.6	550.1	631.5	605.8	534.4	620.0	-233.0	-294.2
Log Wiarygodności	-235.8	-245.4	-286.1	-232.6	-241.3	-284.1	186.2	180.5
Chi-kwadrat	817.8	814.5	733.2	823.5	822.8	737.2		
Statystyka F							120.1	300.7

[1] Modele zagnieżdżone oznaczają ostatnie modele otrzymane po zastosowaniu procedury od ogólnego do szczegółowego.
[2] Modele po transformacji oznaczają modele powstałe po transformacji zmiennych nominacje, zlote_globy, bafta na zmienne zlogarytmowane ln_nom, ln_zg, ln_baf.
[3] W nawiasie znajdują się wartości odchyleń standardowych uzyskanych parametrów.
[4] Liczba gwiazdek przy zmiennej oznacza, że dana zmienna jest istotna na odpowiednim poziomie istotności: 10% (*), 5% (**), 1% (***).

Źródło: Opracowanie własne.

Załącznik 2. Efekty cząstkowe modeli LMP, Logit i Probit

	Logit			Probit			Liniowy Model Prawdopodobieństwa (LMP)	
	Logit Ogólny	Logit Zagnieżdżony	Logit po transformacji	Probit Ogólny	Probit Zagnieżdżony	Probit po transformacji	LMP Ogólny	LMP Zagnieżdżony
budzet2000	-1.12e-10			-2.10e-10			-7.00e-11	
_Igatunek_1	-0.00973			-0.0134			-0.0110	
_Igatunek_2	0.0613	**0.0896*****	**0.112*****	0.0680	**0.0895****	**0.120*****	**0.0654****	**0.0756*****
_Igatunek_3	0.0504			0.0432			0.0277	
_Igatunek_4	-0.0202			-0.0238			-0.0365	
_Igatunek_5	0.0211)			0.0219			0.0179	
_Igatunek_6	0.0572			0.0494			**0.0440***	**0.0504****
_Igatunek_7	0.000620			-0.00189			0.00285	
_Igatunek_8	0.00437			-0.00240			-0.00567	
_Igatunek_9)	-0.0174			-0.0240			-0.0389	
ekranizacja	-0.00606			-0.00466			-0.00991	
roi	-0.00000579			-0.00000504			-1.84e-08	
przychody2000	**9.16e-11****	**6.80e-11*****	**4.86e-11****	**1.06e-10****	**7.56e-11*****	**5.62e-11****	4.56e-11	
kraj_produkcji	**-0.0336*****	**-0.0403*****	**-0.0362*****	**-0.0405*****	**-0.0461*****	**-0.0459*****	**-0.0278****	**-0.0262****
nominacje	**0.0311*****	**0.0352*****		**0.0394*****	**0.0426*****		**0.0812*****	**0.0826*****
zlote_globy	**0.0741*****	**0.0854*****		**0.0826*****	**0.0921*****		**0.0746*****	**0.0748*****
bafta	**0.0312*****	**0.0398*****		**0.0386*****	**0.0461*****		**0.0487*****	**0.0491*****
milosc	**0.0858*****	**0.0808*****	**0.0830*****	**0.0893*****	**0.0827*****	**0.0862*****	**0.0560*****	**0.0496****
Czas_trwania	-0.000288			-0.000376			**-0.000768****	**-0.000737*****
ln_nom[1]			**0.124*****			**0.159*****		
ln_zg[2]			**0.0972*****			**0.125*****		
ln_baf[3]			**0.0822*****			**0.107*****		

Liczba gwiazdek przy zmiennej oznacza, że dana zmienna jest istotna na odpowiednim poziomie istotności: 10% (*), 5% (**), 1% (***).
[1][2][3] Aby uzyskać efekty cząstkowe dla zmiennych niezlogarytmowanych należy podzielić oszacowania w tabeli przez wartości średnie zmiennych nominacje, zlote_globy i bafta.
Źródło: Opracowanie własne.