Pakiety statystyczne
Sprawozdanie 3

Michał Cerazy, Kamil Zawistowski
229969, 229924

24 czerwca 2018

Spis treści

1 Cel

Celem sprawozdania jest odszukanie występowania zależności między statystykami z amerykańskiej koszykówki przy użyciu modeli regresji liniowej.

2 Zbiór danych

Przedmiotem badań są statystyki poszczególnych zawodników występujących na parkietach NBA w sezonie 2014/2015. Podczas tego sezonu w lidze wystąpiło 490 zawodników (każda drużyna może zatrudniać jednocześnie 15 graczy), co przekłada się na taką samą długość próbki. Polecenie wymaga co najmniej 500 pomiarów, jednak w związku z atrakcyjnością omawianych danych, zdecydowaliśmy się na ich użycie. Dane pochodzą ze strony *kaggle.com*.

2.1 Opis zmiennych

Zmienne kategoryczne wyróżnione są poprzez nawiasy zawierające wartości, które zmienna może przyjąć. W sprawozdaniu zostały użyte następujące zmienne:

- „Name" – imię oraz nazwisko,
- „Age" – wiek,
- „Birth Place" – miejsce urodzenia (US, NONUS),
- „Height"– wzrost w centymetrach,
- „Pos" – pozycja, na której gra dany zawodnik (PG, SG, C, PF, SF),
- „Team" – ostatni zespół, w którym grał zawodnik,
- „Weight" – waga zawodnika w kilogramach,
- „BMI" – wskaźnik Body Mass Index,
- „Games Played" - ilość rozegranych meczy w sezonie,
- „MIN" – ilość minut spędzonych na boisku w sezonie,
- „PTS" – ilość punktów zdobytych w sezonie,
- „FGM" – ilość trafionych rzutów z gry w sezonie,
- „FGA" – ilość rzutów z gry w sezonie,
- „FGp" – procentowa skuteczność rzutów z gry w sezonie,

- „ThreePM” – ilość trafionych rzutów za 3 punkty w sezonie,
- „ThreePA” – ilość rzutów za 3 punkty w sezonie,
- „ThreePp” – procentowa skuteczność rzutów za trzy punkty w sezonie,
- „FTM” – ilość trafionych rzutów osobistych w sezonie,
- „FTA” – ilość rzutów osobistych w sezonie,
- „FTp” – procentowa skuteczność rzutów osobistych w sezonie,
- „OREB” – ilość zbiórek ofensywnych w sezonie,
- „DREB” – ilość zbiórek defensywnych w sezonie,
- „REB” – ilość zbiórek w sezonie,
- „AST” – ilość asyst w sezonie,
- „STL” – ilość przechwytów w sezonie,
- „BLK” – ilość bloków w sezonie,
- „TOV” – ilość strat w sezonie,
- „PF” – ilość fauli osobistych w sezonie,
- „EFF” – wydajność zawodnika,
- „AST/TOV” – stosunek ilości asyst do ilości strat,
- „STL/TOV” – stosunek ilości przechwytów do ilości strat.

Dodatkowo, na potrzeby modeli wprowadziliśmy dodatkowe kolumny takie jak:

- „PPG” – średnia ilość punktów zdobywanych na mecz,
- „MPG” – średnia ilość minut rozegranych na mecz,
- „FTAPG” – średnia ilość rzutów osobistych na mecz,
- „FGAPG” – średnia ilość rzutów z pola na mecz,
- „ThreeAPG” – średnia ilość rzutów za 3 punkty na mecz.

3 Dopasowanie modelu

Jako zmienną zależną będziemy rozważać średnią ilość punktów zdobywanych na mecz. Zasugerowane przez nas modele będą się różnić doborem zmiennych niezależnych, jednakże w każdym z nich używać będziemy średniej ilości rozegranych minut na mecz, gdyż jest to niezbędna informacja do analizy modelu. Dodatkowo, w naszych rozważaniach korzystać będziemy z cech fizycznych zawodników, statystyk rzutowych, pochodzenia i pozycji, dlatego usuniemy dane dotyczące asyst, bloków, zbiórek, przechwytów, strat oraz drużyn, w których skończyli sezon.

3.1 Poprawność modeli

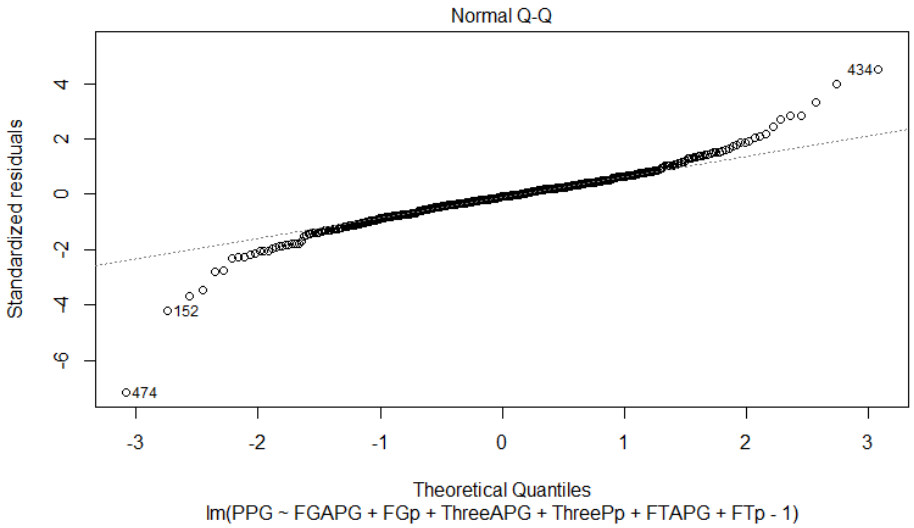
Na wstępie należy zaznaczyć, że żaden z naszych modeli nie spełnia warunku normalności residuów. Pomimo wielu podjętych prób nie udało nam się przekształcić danych tak, by residua modelu pochodziły z rozkładu normalnego. Normalizacji residuów próbowaliśmy dokonać poprzez:

- nałożenie logarytmu na zmienną zależną,
- nałożenie pierwiastków na zmienną zależną,
- nałożenie potęgi na zmienną zależną,
- usuwanie wartości odstających,
- usuwanie wartości o dużej dźwigni.

Dodatkowo, w każdy zaproponowany przez nas model nie posiada stałej, gdyż po sprawdzeniu kilku możliwości zauważyliśmy, że modele bez niej charakteryzują się lepszym dopasowaniem.

3.2 Model I: umiejętności rzutowe zawodników

Przy pierwszym ze sprawdzanych przez nas modeli skupimy się na zdolnościach rzutowych, a więc zmiennymi objaśniającymi będą średnia ilość rzutów z pola, skuteczność z pola, średnia ilość rzutów za 3 punkty, skuteczność rzutów za 3 punkty, średnia ilość rzutów osobistych, skuteczność rzutów osobistych oraz średnia ilość minut przegranych w meczu. W tabeli 1 przedstawione zostały wartości współczynników oraz pozostałe statystyki dla omawianego modelu.

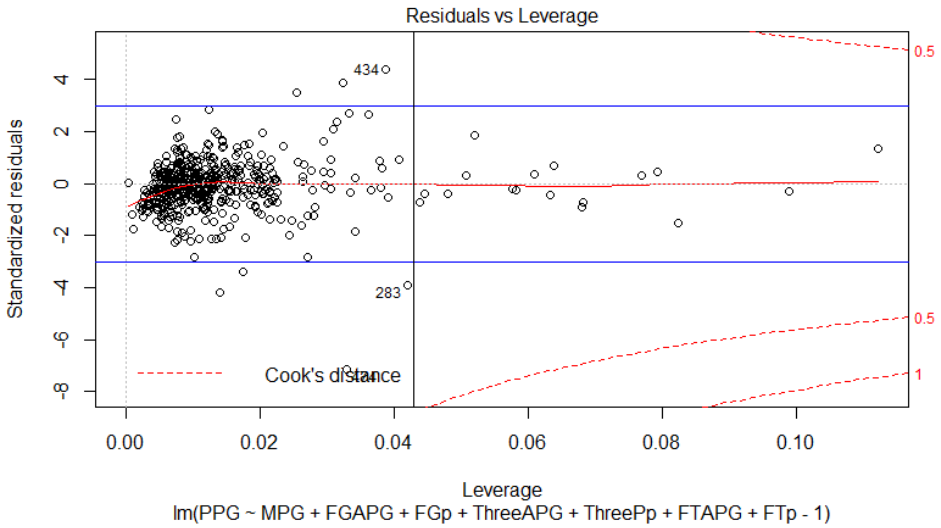


Rysunek 1: QQplot dla modelu rzutowego

Zmienna	Estymacja	Błąd standardowy	T-wartość	Pr($<\|t\|$)
FGAPG	0.961931	0.025304	38.015	<2e-16
FGp	0.012008	0.003441	3.490	0.000528
ThreeAPG	0.130223	0.032714	3.981	7.93e-05
ThreePp	0.001808	0.002931	0.617	0.537764
FTAPG	0.842080	0.044256	19.027	<2e-16
FTp	-0.007247	0.002059	-3.519	0.000474

Tabela 1: Model rzutowy - współczynniki modelu

Jak można zauważyć, wszystkie parametry poza skutecznością rzutów za 3 punkty są istotne(co może dziwić przy obecnych realiach gry w koszykówkę, gdzie coraz większą wagę przykłada się do szybkich rzutów trzypunktowych). Jak się okazało, w przypadku tego modelu założenia nie są spełnione: test Shapiro-Wilka i spojrzenie na Rysunek 1 odrzuciły hipotezę o normalności residuów, a testy badające średnią i wariancję(polegające na wielokrotnym losowaniu próbki 50 elementów i przeprowadzaniu t.testu i var.testu) wykazały, że średnia jest różna od zera oraz wariancja nie jest stała na poziomie istotności $\alpha = 0.05$. Wyniki zawarte zostały w Tabeli 2.



Rysunek 2: Wykres residuów i dźwigni dla modelu rzutowego

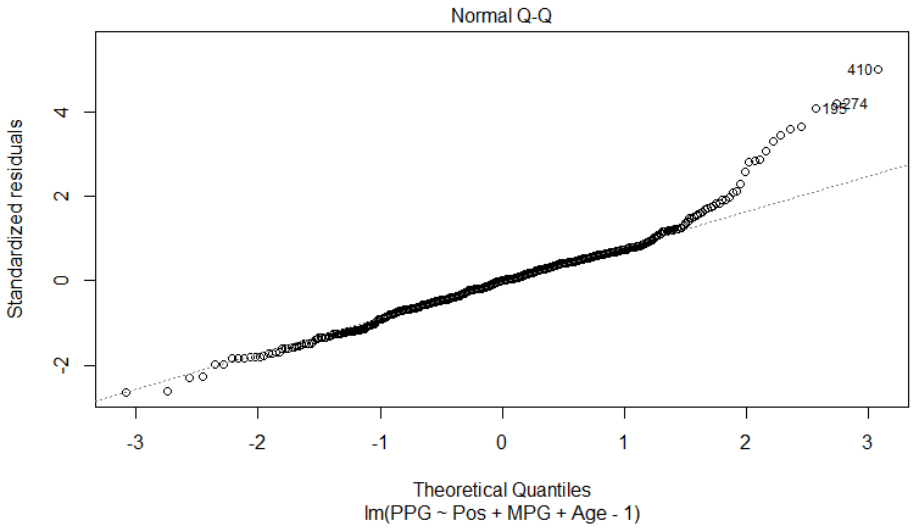
W	p-value	ConstMeanMC	ConstVarMC
0.92953	2.028e-14	0.1019	0.2892

Tabela 2: Model rzutowy - testy założeń

Na Rysunku 2 niebieskimi liniami zaznaczono próg dla wartości odstających, czarną natomiast tych o wysokiej dźwigni. Niestety, po ich usunięciu model nadal nie spełniał założeń i pojawiły się kolejne problematyczne obserwacje(próbowaliśmy eliminować je do skutku, jednak również to nie przyniosło oczekiwanych rezultatów).

3.3 Model II: pozycje zawodników

Podczas tworzenia tego modelu staraliśmy się określić, czy pozycja zawodnika wpływa na ilość zdobywanych przez niego punktów, dlatego jako zmienne objaśniające przyjęliśmy pozycję danego gracza, średnią ilość czasu spędzanego na boisku oraz jego wiek. W tabeli 3 przedstawione są współczynniki dopasowanego modelu. Jak widać, wszystkie zmienne poza wiekiem okazały się być istotne.



Rysunek 3: QQplot dla modelu uwzględniającego pozycje

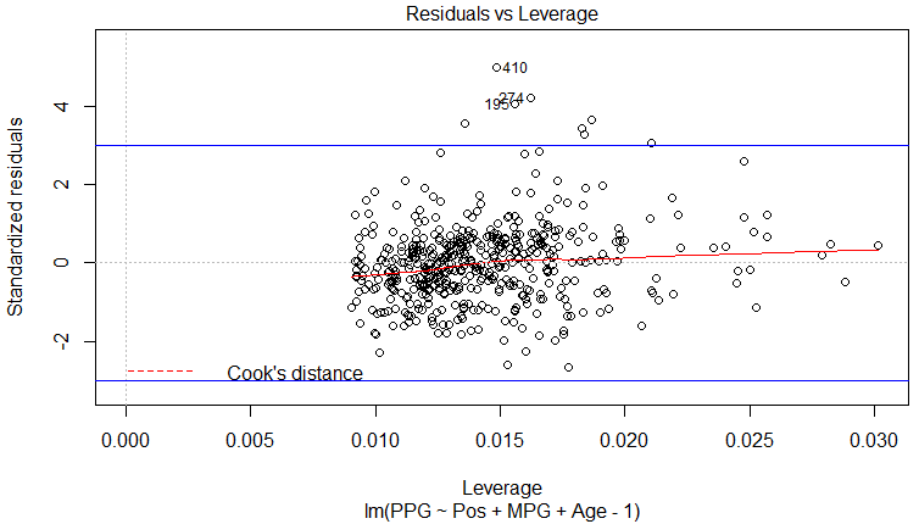
Zmienna	Estymacja	Błąd standardowy	T-wartość	Pr($<\|t\|$)
PosC	-2.654071	0.807253	-3.288	0.001084
PosPF	-2.535002	0.793808	-3.193	0.001498
PosPG	-2.506314	0.799289	-3.136	0.001819
PosSF	-3.007021	0.810325	-3.711	0.000231
PosSG	-2.355508	0.782667	-3.010	0.002753
MPG	0.542043	0.012523	43.282	$< 2e-16$
Age	-0.008325	0.026757	-0.311	0.755835

Tabela 3: Model pozycyjny - współczynniki modelu

W tabeli 4 przedstawione są wyniki testu Shapiro-Wilka oraz testów średniej i wariancji residuów. Jak wyżej wspomnieliśmy, model nie spełnia założenia normalności badanych wartości resztowych(jak widać również na Rysunku 3). Na Rysunku 4 zaznaczone zostały wartości odstające, których usunięcie również nie pomogło w normalizacji residuów.

W	p-value	ConstMeanMC	ConstVarMC
0.95198	1.644e-11	0.0424	0.1784

Tabela 4: Model pozycyjny - testy założeń



Rysunek 4: Wykres residuów i dźwigni dla modelu uwzględniającego pozycje

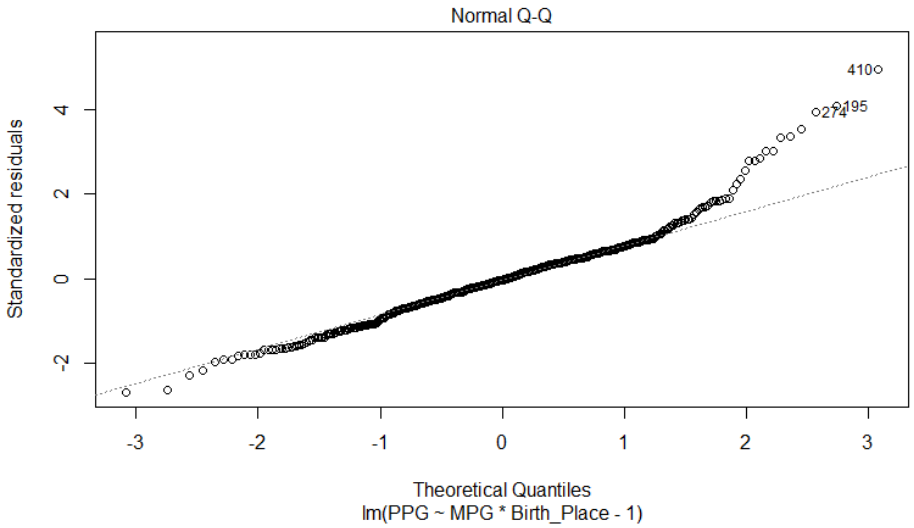
3.4 Model III: pochodzenie zawodników

Przy ostatnim z badanych modeli chcieliśmy zbadać, czy zawodnicy ze Stanów Zjednoczonych, stanowiący większość ligi, są lepszymi punktującymi niż obcokrajowcy. Jako zmienne objaśniające przyjęliśmy średnią ilość minut na mecz oraz pochodzenie zawodnika. W Tabeli 5 przedstawione są współczynniki dobranego modelu, gdzie jak można zauważyć, wszystkie parametry poza iloczynem średniej ilości minut na mecz i indykatora urodzenia w USA okazały się być istotnymi.

Zmienna	Estymacja	Błąd standardowy	T-wartość	$\Pr(<\|t\|)$
MPG	0.49746	0.02803	17.746	$< 2e-16$
Birth Placenon us	-2.05640	0.60703	-3.388	0.000762
Birth Placeus	-3.02104	0.30683	-9.846	$< 2e-16$
MPG:Birth Placeus	0.05583	0.03123	1.788	0.074431

Tabela 5: Model uwzględniający pochodzenie - współczynniki modelu

W tabeli 6 zaprezentowane są wyniki testów przeprowadzonych dla powyższego modelu. Na podstawie tych wyników można stwierdzić, że residua nie pochodzą z rozkładu normalnego. Podjęte przez nas próby normalizacji residuów nie przyniosły efektów również w tym przypadku.



Rysunek 5: QQplot dla modelu uwzględniającego pochodzenie

W	p-value	ConstMeanMC	ConstVarMC
0.95802	1.347e-10	0.0443	0.1693

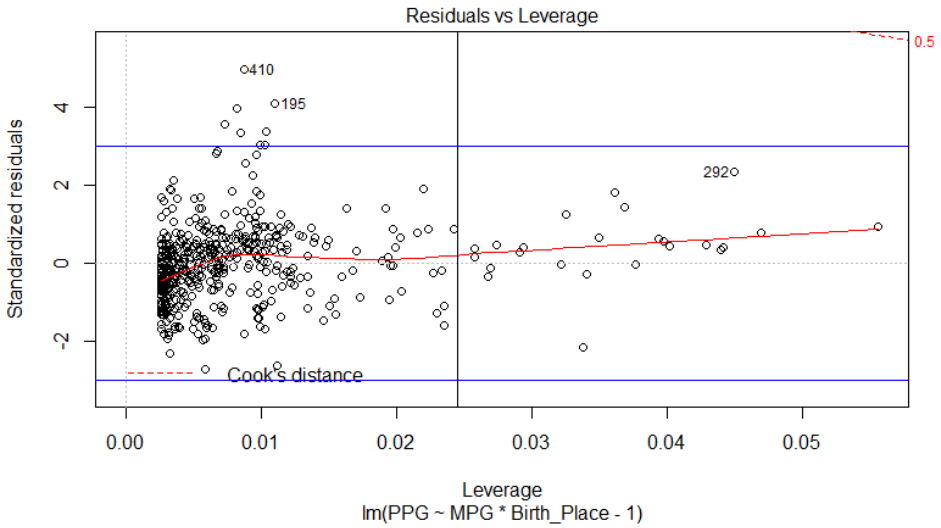
Tabela 6: Model uwzględniający pochodzenie - testy założeń

Na Rysunku 5 przedstawiony jest wykres kwantyl–kwantyl, którego ocena utwierdza nas w przekonaniu, że residua nie pochodzą z rozkładu normalnego. Rysunek 6 przedstawia wartości odstające oraz o dużej dźwigni, których usunięcie również nie wpłynęło pozytywnie na poprawność modelu.

3.5 Porównanie modeli

Znając współczynniki dobranych modeli wyznaczyliśmy:

- współczynnik determinacji, oznaczany jako R^2.
- skorygowany współczynnik determinacji, oznaczany jako aR^2,
- kryterium Akaikego, oznaczane jako AIC,
- logarytm wskaźnika wiarygodności, oznaczany jako $LogLik$.



Rysunek 6: Wykres residuów i dźwigni dla modelu uwzględniającego pochodzenie

Model	R^2	aR^2	AIC	LogLik
I	0.99317	0.993	1204.576	-594.288
II	0.9374	0.9365	2283.041	-1133.52
III	0.9375	0.937	2281.154	-1135.577

Tabela 7: Porównanie modeli

Podsumowując wyżej wymienione modele można stwierdzić, że najlepszym z nich jest model uwzględniający umiejętności rzutowe zawodników. Taki wniosek nasuwa się po analizie Tabeli 7. Współczynniki determinacji oraz logarytm wskaźnika wiarygodności są największe dla tego właśnie modelu, co oznacza, że jest on najlepiej dopasowany do naszych danych. Taką konkluzję potwierdza również kryterium informacyjne Akaikego, które najmniejszą wartość przyjmuje dla modelu nr I.

4 Inne testowane modele

4.1 Model uwzględniający cechy motoryczne zawodników

Podjęliśmy trzy próby dobrania najlepszego modelu, który uzależniony jest od predyspozycji fizycznych zawodników. Oczywiście, każdy z tych modeli musi być rozszerzony o ilość rozegranych minut na mecz, gdyż jest to niezbędna informacja do analizy modelu.

4.1.1 Model WWW

Pierwszy z dopasowywanych modeli uwzględnia wzrost, wiek oraz wagę zawodników. Jedyną wpływową zmienną okazała się być ilość minut na mecz (jak zakładaliśmy wyżej, jest to najważniejszy czynnik). W tabeli 8 przedstawione zostały wartości współczynników oraz pozostałe statystyki dla omawianego modelu.

Zmienna	Estymacja	Błąd standardowy	T-wartość	$Pr(<\|t\|)$
Minuty na mecz	0.542353	0.012281	44.161	<2e-16
Wzrost	-0.023874	0.007604	-3.140	0.00179
Wiek	-0.013982	0.025582	-0.547	0.58495
Waga	0.022851	0.012485	1.830	0.06783

Tabela 8: Model WWW - współczynniki modelu

4.1.2 Model WW

Drugi z dopasowywanych modeli uwzględnia średni czas gry, wzrost oraz wagę zawodników. Odjęcie wieku ma na celu sprawdzenie jak zachowa się model po odjęciu jednej zmiennej nie będącej zmienną istotną. W tabeli 9 przedstawione zostały współczynniki modelu WW. Jak można zauważyć zmienne istotne dla nie zmieniły się, jednakże waga jest na granicy zostania zmienną istotną.

Zmienna	Estymacja	Błąd standardowy	T-wartość	$Pr(<\|t\|)$
Minuty na mecz	0.541663	0.012207	44.372	<2e-16
Wzrost	-0.026028	0.006499	-4.005	7.17e-05
Waga	0.023444	0.012429	1.886	0.0599

Tabela 9: Model WW - współczynniki modelu

4.1.3 Model WB

Trzeci z dopasowywanych modeli uwzględnia wiek oraz wskaźnik BMI. Omawiany wskaźnik jest popularną statystyką mówiącą o stanie zdrowia fizycznego i wyliczana jest na podstawie poniższego wzoru:

$$BMI = \frac{waga}{wzrost^2}.$$

W związku z tym niejawnie do modelu włączamy wagę i wzrost zawodników ligi. Wyniki tego zabiegu przedstawione są w tabeli 10. Ponownie, wiek okazał się być nieistotny, natomiast BMI oraz ilość minut na mecz są istotne.

Model	R^2	AIC	LogLik
WWW	0.9369	2281.809	-1135.905
WW	0.9358	2280.111	-1136.055
WB	0.937	2288.041	-1140.021

Tabela 11: Porównanie modeli

Zmienna	Estymacja	Błąd standardowy	T-wartość	Pr($<\|t\|$)
Minuty na mecz	0.53943	0.01245	43.329	<2e-16
Wiek	-0.03886	0.02571	-1.512	0.1313
BMI	-0.06666	0.02896	-2.302	0.0218

Tabela 10: Model WB - współczynniki modelu

4.2 Porównanie modeli

W tabeli 11 widać, że model WB cechuje się najlepszym dopasowaniem do danych według współczynnika R^2. Kryterium informacyjne Akaikego przemawia za wybraniem drugiego modelu jako najlepiej opisującego badany zbiór wartości. Logarytm wskaźnik wiarygodności wskazuje z kolei na model WWW. Na podstawie zaprezentowanych wyników nie możemy stwierdzić, który z nich jest najlepszy, gdyż każde sprawdzane kryterium wskazuje na inny model. Dodatkowo, żaden nie spełnia założeń, gdyż ich residua nie pochodzą z rozkładu normalnego.

5 Podsumowanie

Model opisujący zdolności rzutowe okazał się być istotnie najlepszym spośród wszystkich testowanych przez nas.