

Pakiety statystyczne

Sprawozdanie 3

Michał Ceraży, Kamil Zawistowski
229969, 229924

24 czerwca 2018

Spis treści

1	Cel	3
2	Zbiór danych	3
2.1	Opis zmiennych	3
3	Dopasowanie modelu	5
3.1	Model I: umiejętności rzutowe zawodników	5
3.2	Model II: pozycje zawodników	6
3.3	Model III: pochodzenie zawodników	6
3.4	Porównanie modeli	7
4	Inne testowane modele	8
4.1	Model uwzględniający cechy motoryczne zawodników	8
4.1.1	Model WWW	8
4.1.2	Model WW	9
4.1.3	Model WB	10
4.2	Porównanie modeli	10
5	Podsumowanie	12

1 Cel

Celem prowadzonych badań jest odszukanie występowania zależności między statystykami, w koszykówce, przy użyciu modeli regresji.

2 Zbiór danych

Przedmiotem badań są statystyki poszczególnych zawodników występujących na parkietach NBA w sezonie 2014/2015. W omawianym okresie, 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 katagoryczne 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,
- „Pos” – pozycja, na której gra dany zawodnik (PG, SG, C, PF, SF),
- „Team” – ostatni zespół, w którym grał zawodnik,
- „Weight” – waga zawodnika,
- „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 doбором zmiennych niezależnych, jednakże w każdym z nich używać będziemy średniej ilości rozegranych minut na mecz, gdyż jest niezbędna informacja do analizy modelu. Dodatkowo, w naszych modelach będziemy korzystać z cech fizycznych zawodników i statystyk rzutowych, dlatego usuniemy dane dotyczące asyst, bloków, zbiórek, przechwyty i strat.

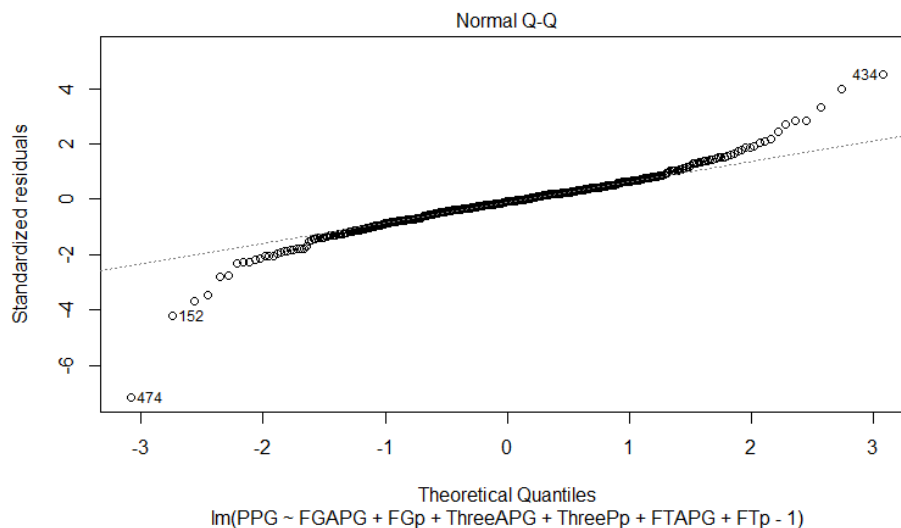
3.1 Model I: umiejętności rzutowe zawodników

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.

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 trzy-punktowych). P-value dla testu F wynosi $p-value : < 2.2e - 16$, a więc nasz model jest lepszy niż pewna stała. 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 stałą 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 1: QQplot dla modelu rzutowego

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

Tabela 2: Model rzutowy - testy założeń

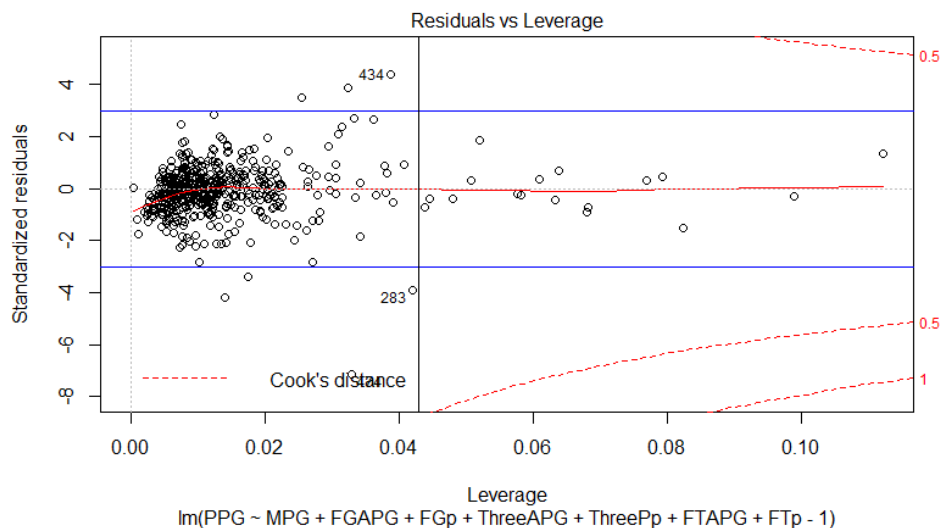
Na Rysunku 2 niebieskimi liniami zaznaczono próg dla wartości odstających, czarną natomiast dla tych o wysokiej dźwigni. Niestety, po ich usunięciu model nadal nie spełniał założeń i pojawiły się kolejne problematyczne obserwacje(próbowaliśmy eliminować je do skutku, jednak nawet to nie przyniosło oczekiwanych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

3.2 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.

3.3 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.



Rysunek 2: Wykres residuów i dźwigni dla modelu rzutowego

3.4 Porównanie modeli

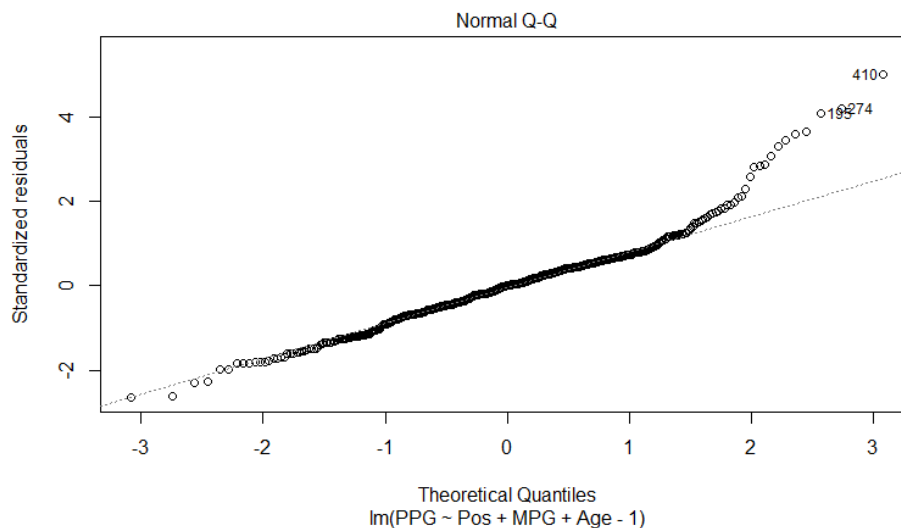
Znając współczynniki dobranych modeli wyznaczyliśmy:

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

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 3: Porównanie modeli

Niestety, żaden z dopasowanych modeli nie jest modelem poprawnym, gdyż ich residua nie mają rozkładu normalnego. Sprawdzenia tej hipotezy dokonaliśmy przy użyciu testu Shapiro-Wilka (długość naszej próbki nie przekracza 5000). Statystyka W jest większa od liczby 0.947 (wartość zaczerpnięta z tablic), co implikuje odrzucenie hipotezy zerowej o normalności residuów. W tabeli 3 widać, że model I cechuje się najlepszym dopasowaniem



Rysunek 3: QQplot dla modelu uwzględniającego pozycje

do danych według wszystkich przedstawionych kryteriów, tak więc możemy założyć, że statystyki rzutowe najlepiej opisują zdobycze punktowe zawodników.

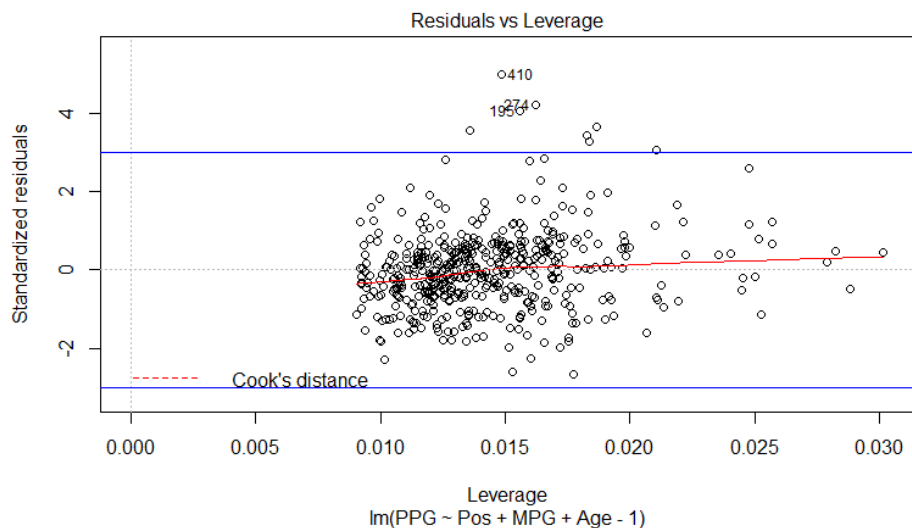
4 Inne testowane modele

4.1 Model uwzględniający cechy motoryczne zawodników

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

4.1.1 Model WWW

Pierwszy z dopasowywanych modeli uwzględnia wzrost, wiek oraz wagę zawodników. Zmienne istotne dla dobranego modelu to minuty na mecz (jak zakładaliśmy wyżej, jest to najważniejszy czynnik) oraz waga. W tabeli 4 przedstawione zostały wartości współczynników oraz pozostałe statystyki dla omawianego modelu.



Rysunek 4: Wykres residuów i dźwigni dla modelu uwzględniającego pozycję

Zmienna	Estymacja	Błąd standardowy	T-wartość	$\Pr(< t)$
Minuty na mecz	0.54219	0.01242	43.638	$<2e-16$
Wzrost	-0.03394	0.02154	-1.576	0.1156
Wiek	-0.01660	0.02667	-0.622	0.5340
Waga	0.02721	0.01462	1.861	0.0634
Stała	1.63120	3.35589	0.486	0.6271

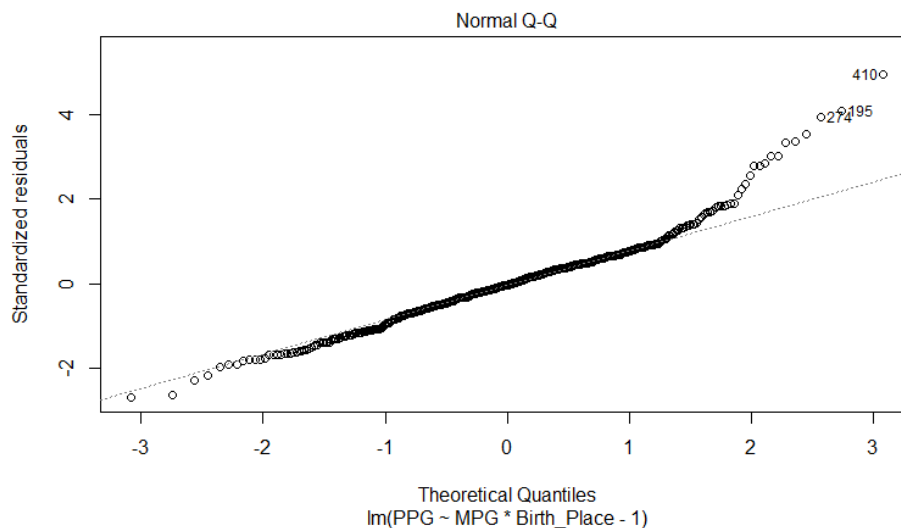
Tabela 4: Model WWW - współczynniki modelu

4.1.2 Model WW

Drugi z dopasowywanych modeli uwzględnia wiek oraz wagę zawodników. Odjęcie wzrostu ma na celu sprawdzenie jak zachowa się model po odjęciu jednej zmiennej nie będącej zmienną istotną. W tabeli 5 przedstawione zostały współczynniki modelu WW. Jak można zauważyć zmienne istotne dla modelu nie są już takie same, do minut na mecz dołączyła wartość stała. Ponadto waga zawodnika przestała należeć do istotnych zmiennych.

Zmienna	Estymacja	Błąd standardowy	T-wartość	$\Pr(< t)$
Minuty na mecz	0.54356	0.01241	43.78	$<2e-16$
Wiek	-0.01304	0.02661	-0.490	0.62428
Waga	0.00852	0.00857	0.994	0.32059
Stała	-3.35012	1.13018	-2.964	0.00318

Tabela 5: Model WW - współczynniki modelu



Rysunek 5: QQplot dla modelu uwzględniającego pochodzenie

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 6. Podobnie jak w modelu WW istotne zmienne to liczba minut na boisku oraz stała.

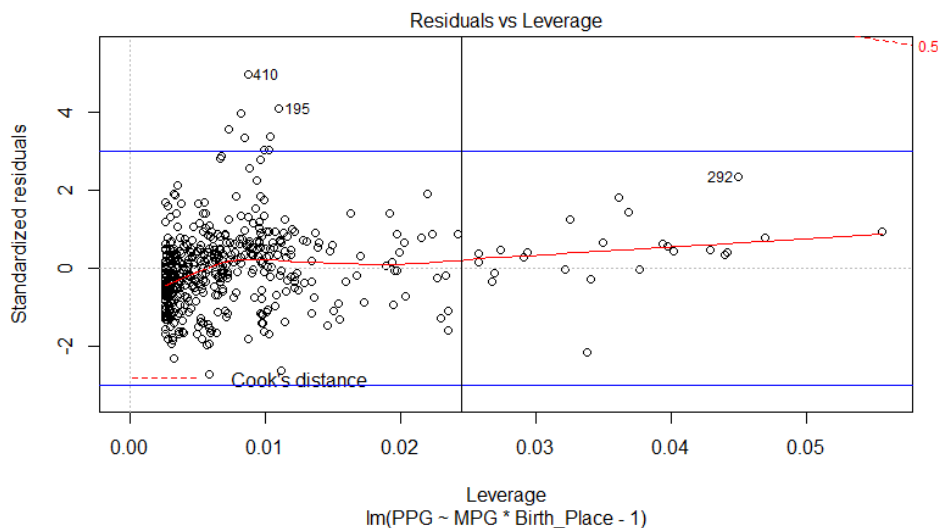
Zmienna	Estymacja	Błąd standardowy	T-wartość	Pr(< t)
Minuty na mecz	0.54259	0.01238	43.838	<2e-16
Wiek	-0.01608	0.02667	-0.603	0.54676
BMI	0.09639	0.05878	1.640	0.10166
Stała	-4.84181	1.60703	-3.013	0.00272

Tabela 6: Model WB - współczynniki modelu

4.2 Porównanie modeli

Znając współczynniki dobranych modeli wyznaczyliśmy:

- skorygowany współczynnik determinacji, oznaczany jako R^2 ,
- kryterium Akaikego, oznaczane jako AIC ,



Rysunek 6: Wykres residuów i dźwigni dla modelu uwzględniającego pochodzenie

- logarytm wskaźnika wiarygodności, oznaczany jako LogLik ,
- wartość W z testu Shapiro-Wilka, oznaczana jako W .

Model	R^2	AIC	LogLik	W
WWW	0.7977	2282.972	-1135.486	0.95495
WW	0.7971	2283.475	-1136.737	0.95403
WB	0.7978	2281.766	-1135.883	0.95493

Tabela 7: Porównanie modeli

Każdy z dopasowanych modeli jest „poprawnym modelem”, gdyż ich residua mają rozkład normalny. Sprawdzenia tej hipotezy dokonaliśmy przy użyciu testu Shapiro-Wilka (długość naszej próbki nie przekracza 5000). Wartość W jest mniejsza od liczby 0.947 (wartość zaczerpnięta z tablic), co oznacza, że nie ma podstaw by odrzucić hipotezę o normalności residuów. W tabeli 7 widać, że model WB cechuje się najlepszym dopasowaniem do danych według współczynnika R^2 . Kryterium informacyjne Akaikego również przemawia za wybraniem trzeciego modelu jako najlepiej opisującego badany zbiór wartości. Jediną statystyką przemawiającą za innym modelem niż WB jest logarytm wskaźnik wiarygodności, który wskazuje, że to pierwszy model jest najlepszy. Na podstawie zaprezentowanych wyników należy wybrać model uwzględniający wiek oraz BMI zawodników.

5 Podsumowanie