Analiza liczby wygranych gier w koszykówce, przy użyciu technik Bayesowskich.

Wstęp

Celem projektu jest konstrukcja modelu opisującego czynniki wpływające na liczbę wygranych gier w Amerykańskiej lidze piłki koszykowej drużyn należących do NCAA Divison przy użyciu technik Bayesowskich. W pierwszym rozdziale opisany został zbiór danych oraz wybrane zmienne. Drugi rozdział został poświęcony estymacji Bayesowskiej. Podzielony został na 3 podrozdziały. Pierwszy podrozdział skupia się na opisie konstrukcji rozkładu *a priori*. W następnym zostaje przeprowadzona konstrukcja rozkładu *a posteriori*. Ostatni podrozdział drugiego rozdziału skupia się na graficznym przestawieniu wyników estymacji Bayesowskiej wraz z analizą zamieszczonych tam wykresów. Ostatni rozdział to podsumowanie wyników projektu.

1.Zbiór danych

Do przeprowadzenia analizy skorzystano ze zbioru „College Basketball Dataset”[1]. Zbiór ten zawiera informacje o statystykach drużyny należących do NCAA Division I[2] w podziale na sezony. Do analizy wybrano dwa sezony: 2019 będący docelowym sezonem oraz sezon do niego poprzedni – 2018. Dane z sezonu 2018 posłużyły do zbudowania modelu, który zapewnił informacje *a priori* dla modelu docelowego. Zbiory danych zawierały 23 zmienne oraz odpowiednio 353 i 351 obserwacji dla sezonu 2019 i 2018. W celu doboru zmiennych oraz redukcji wymiarowości zdecydowano się na użycie regresji *stepwise* typ *backward*, gdzie jako kryterium doboru zmiennych wybrano Bayesowskie Kryterium Informacyjne. Następnie usunięto zmienne, które odpowiadały za współliniowość modelu przy użyciu współczynnika VIF. Ostatecznie do analizy pozostały zmienne:

- *W* – Zmienna dyskretna. Zmienna objaśniana. Liczba wygranych gier.
- *G* – Zmienna dyskretna. Liczba gier zagranych.
- *EFG_O* - Zmienna ciągła. Effective Field Goal Percentage Shot - mierzy skuteczność zawodnika z uwzględnieniem różnej punktacji w zależności od rodzaju rzutu.
- *TOR* - Zmienna ciągła. Turnover Percentage Allowed (Turnover Rate) – statystyka mierząca ile razy drużyna straciła piłkę zanim gracz wykonał rzut do kosza w stosunku do rzutów do kosza.
- *TORD* - Zmienna ciągła. Turnover Percentage Committed (Steal Rate) – statystyka odwrotna do turnover, mówiąca ile razy drużyna przejęła piłkę bez faulu przeciwnikowi, który wykonywał rzut do kosza.
- *ORB* - Zmienna ciągła. Offensive Rebound Percentage – procent piłek, które „uderzyły w tablice i powróciły do gracza” po „spudłowaniu” rzutu wolnego.
- *DRB* - Zmienna ciągła. Defensive Rebound Percentage – procent piłek, które drużyna przejmie po „spudłowaniu” rzutu wolnego przez drużynę przeciwną.

[1] https://www.kaggle.com/andrewsundberg/college-basketball-dataset (10.01.2020)
[2] https://en.wikipedia.org/wiki/NCAA_Division_I

Dla tak dobranych zmiennych został skonstruowany model regresji liniowej. Zakładamy, że składniki losowe mają takie same rozkłady normalne z wartością oczekiwaną 0, które są od siebie niezależne.

2. Estymacja Bayesowska

Rozkład *a priori*

W pracy przyjęto, że parametry β_i mają rozkład *a priori* wielowymiarowy normalny-gamma, który można zapisać jako:

$$\beta \sim N\left(\underline{\beta}, h^{-1}\underline{U}\right)$$

$$h = \frac{1}{\sigma^2} \qquad h \sim \Gamma\left(\beta = \frac{\underline{v}\underline{s}^2}{2}, \alpha = \frac{\underline{v}}{2}\right)$$

W celu otrzymania pozostałych parametrów *a priori* niezbędnych do skonstruowania rozkładów *a posteriori*, estymowano model KMNK dla sezonu 2018 oraz uzyskano w wyniku następujące oszacowane wartości parametrów β_i oraz wyrazu wolnego:

Tabela 1. Oszacowania KMNK parametrów dla sezonu 2018

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-56.43262	5.07263	-11.125	< 2e-16	***
G	0.82945	0.07496	11.065	< 2e-16	***
EFG_O	0.93457	0.06455	14.477	< 2e-16	***
TOR	-0.59364	0.09630	-6.165	1.98e-09	***
TORD	0.76326	0.08305	9.190	< 2e-16	***
ORB	0.35379	0.04503	7.857	5.09e-14	***
DRB	-0.46653	0.05906	-7.899	3.82e-14	***

Pozostałe parametry *a priori*:

$\underline{h^{-1}U}$ - macierz kowariancji oszacowanych parametrów β_i z modelu dla sezonu 2018 przemnożona przez odwrotność wariancji reszt z tego modelu

$\underline{v}$ - liczba stopni swobody modelu dla sezonu 2018

$\underline{s^{-2}}$ - precyzja, odwrotność wariancji reszt modelu dla sezonu 2018

Rozkład *a posteriori*

W pracy w celu oszacowania rozkładu *a posteriori* użyto opisane w poprzednim podrozdziale informacje a priori jak i informacje pochodzące z danych czyli oszacowanie modelu liniowego dla danych z sezonu 2019. Przyjęto założenia o niezależności obserwacji oraz normalności rozkładu składnika losowego w modelu. Zdefiniowano również, że rozkład *a posteriori* jest rozkładem wielowymiarowym normalnym-gamma, gdzie jego parametry są opisane następująco:

$$\overline{\beta} = \left(\underline{U}^{-1} + X^T X\right)^{-1} \left(\underline{U}^{-1}\underline{\beta} + X^T X\hat{\beta}\right)$$
$$\overline{U} = \left(\underline{U}^{-1} + X^T X\right)^{-1}$$
$$\overline{v} = \underline{v} + N$$
$$\overline{vs}^2 = \hat{v}\hat{s}^2 + \underline{vs}^2 + \left(\hat{\beta} - \underline{\beta}\right)^T \left[\underline{U} + \left(X^T X\right)^{-1}\right]^{-1} \left(\hat{\beta} - \underline{\beta}\right)$$

Gdzie:

X – macierz obserwacji zmiennych objaśniających

$\hat{\beta}$ - wektor oszacowań KMNK parametrów modelu dla sezonu 2019

N – liczba obserwacji

$\hat{s}^2$ – suma kwadratów reszt modelu dla sezonu 2019

Dodatkowo postanowiono przeprowadzić test normalności Shapiro-Wilka dla reszt:

Tabela 2. Test normalności Shapiro-Wilka dla reszt

Reszty z modelu KMNK	**P-value**
2018	0.9583
2019	0.4423

W obu przypadkach nie ma podstaw do odrzucenia H_0: rozkład badanej cechy ma rozkład normalny.

Następnie przystąpiono do oszacowania parametrów rozkładu *a posteriori*, gdzie wyniki oszacowań wartości oczekiwanej β_i pokazano w poniższej tabeli:

Tabela 3. Wartości oczekiwane parametrów a posteriori

β	
G	0.7988597
EFG_0	0.9150662
TOR	-0.6917571
TORD	0.7902432
ORB	0.4099115
DRB	-0.4496453

Znaki przy wartościach oczekiwanych parametrów *a posteriori* są takie same jak w oszacowaniu OLS z danych.

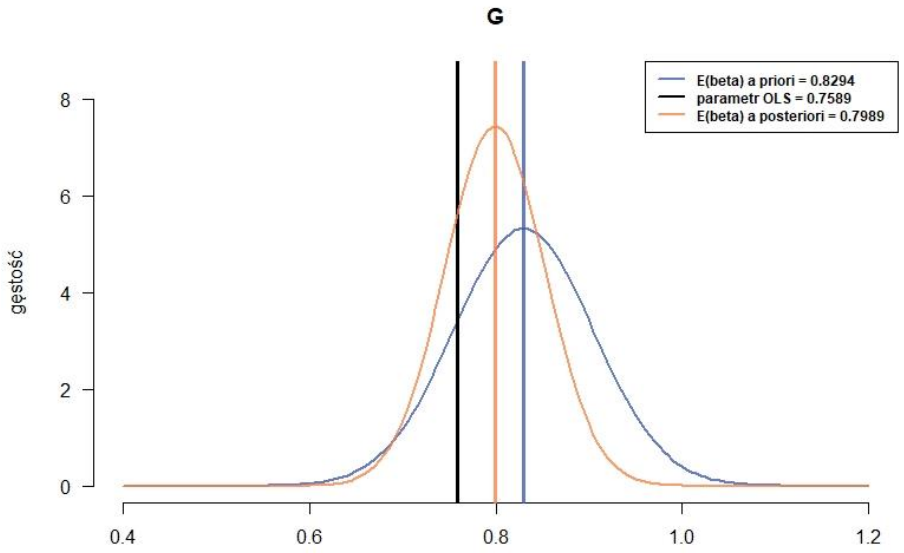
Rozkłady brzegowe

Aby zilustrować rozkłady *a posteriori* dla poszczególnych zmiennych należało skonstruować rozkłady brzegowe dla wielowymiarowego rozkładu gamma-normalnego *a posteriori*. Rozkład brzegowy dla parametru β_i jest jednowymiarowym rozkładem t:

$$\beta_i|\boldsymbol{y} \sim t(\overline{\beta}_i, \overline{s}^2\overline{U}_{i,i}, \overline{v})$$

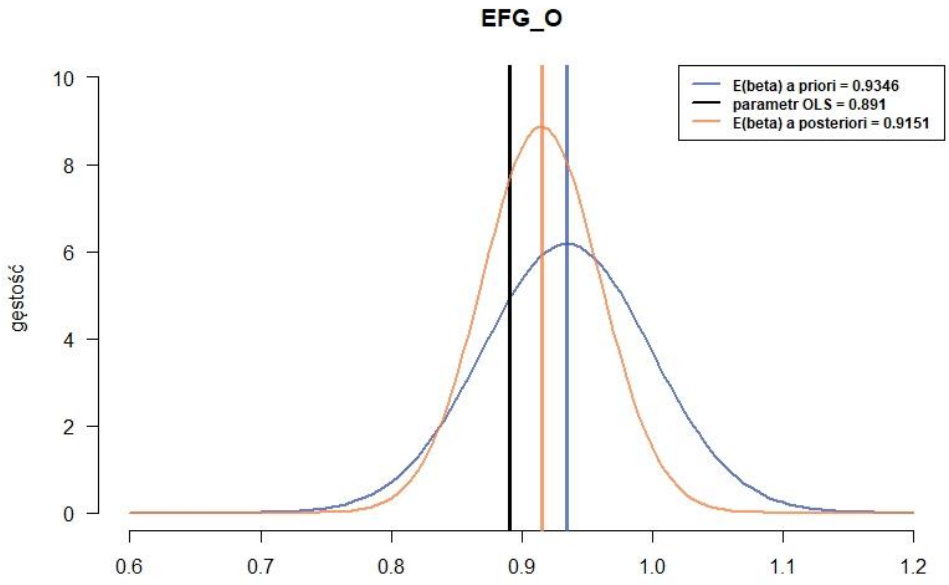
Rozkłady brzegowe dla poszczególnych parametrów prezentowały się w następujący sposób:

Rysunek 1. Gęstości brzegowe parametru G



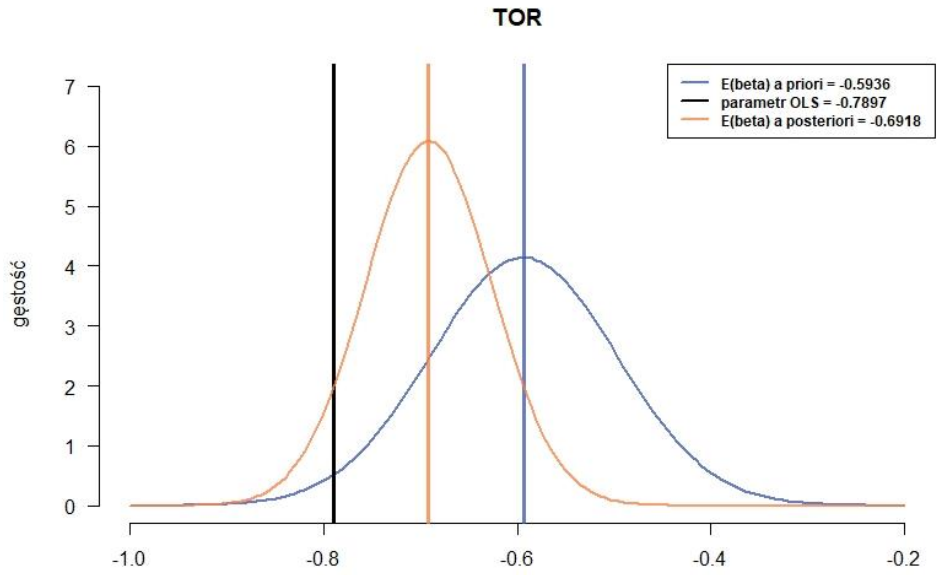
Jak widać na rysunku pierwszym dla zmiennej G informacje *a priori* w zauważalnym stopniu wpłynęły na rozkład *a posteriori*. Jest on węższy od rozkładu a priori co znaczy, że jego większa część masy prawdopodobieństwa jest blisko wartości oczekiwanej z tego rozkładu. Kierunek przesunięcia rozkładu *a posteriori* również się zgadza. Co ciekawe znacząca wielkość masy prawdopodobieństwa rozkładu *a posteriori* znajduje się na prawo od oszacowania OLS, co świadczy o dużej wartości jaką niesie informacja *a priori* z naszego modelu z sezonu 2018. Patrząc na rozkład *a posteriori* i wartość oczekiwaną moglibyśmy stwierdzić, że na podstawie informacji z zeszłego sezonu wpływ tej zmiennej jest większy na liczbę wygranych gier, niż w przypadku oszacowania OLS.

Rysunek 2. Gęstości brzegowe parametru EFG_O



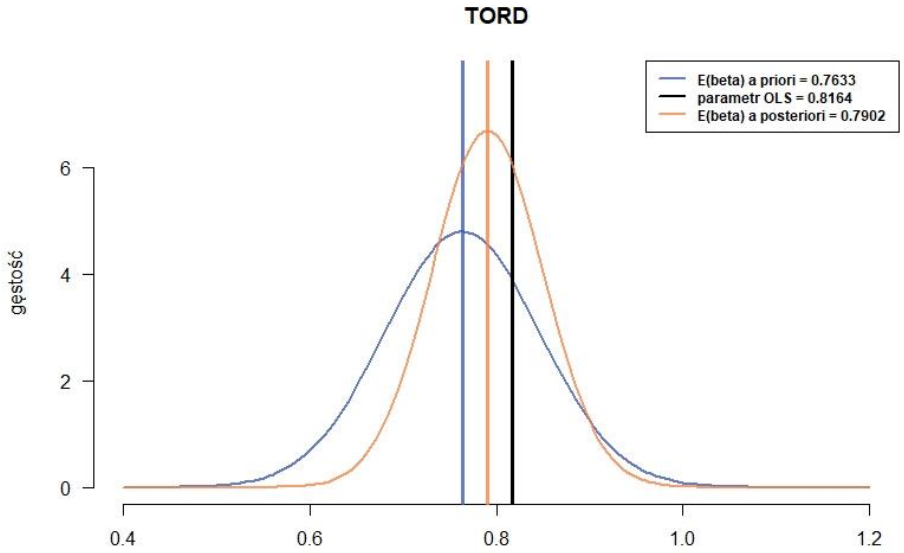
Rysunek drugi przedstawia rozkład brzegowy zmiennej EFG_O. Podobnie jak w zmiennej G kierunek przesunięcia rozkładu *a posteriori* jest poprawny, wskazuje na istotny wpływ informacji *a priori.* Rozkład *a posteriori* jest węższy od rozkładu *a priori*. Wartość oczekiwana rozkładu *a posteriori* zawiera się pomiędzy oszacowaniem OLS jak i wartością oczekiwaną rozkładu *a priori* co oznacza, że wpływ informacji z danych jak i z poprzedniego sezonu był równie istotny. Wynika to z estymowanych modeli liniowych, które dla obu sezonów miały bardzo dobrą jakość oszacowania parametrów. W przypadku tej zmiennej wartość oszacowania OLS jest bardzo bliska wartości oczekiwanej rozkładu *a posteriori* więc nasze przekonania co do wielkości wpływu tej zmiennej na liczbę wygranych gier nie zmienią się w znaczny sposób.

Rysunek 3. Gęstości brzegowe parametru TOR



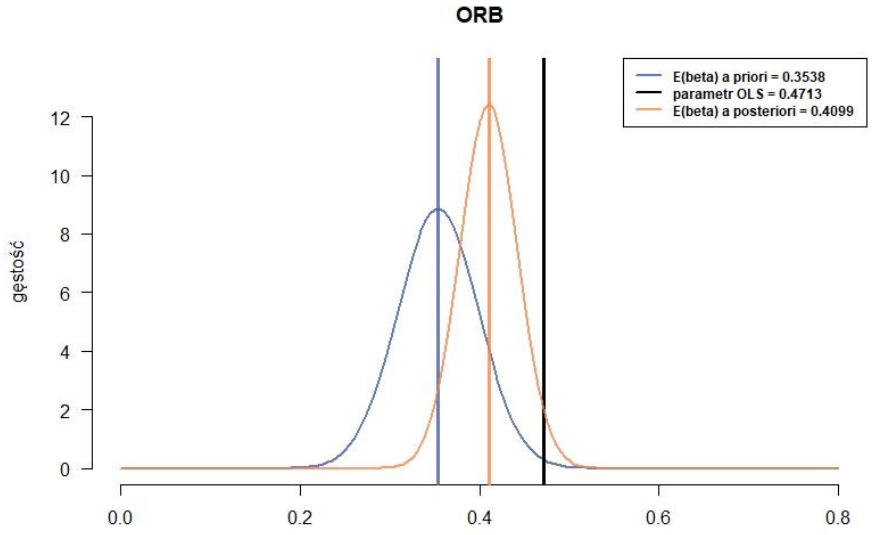
Na rysunku trzecim widzimy gęstości brzegowe parametru TOR. Rozkład *a posteriori* zachowuje się w bardzo zbliżony sposób jak dla poprzednich zmiennych. Jego wartość oczekiwana znajduje się prawie idealnie pomiędzy oszacowaniem OLS jak i wartością oczekiwaną rozkładu *a priori*. Jednakże w przypadku tej zmiennej, nasze przekonania co do wielkości wpływu na liczbę wygranych gier po uwzględnieniu informacji *a priori* zeszłego sezonu może ulec osłabieniu. Wynika to z faktu, że większość masy prawdopodobieństwa znajduje się na prawo od oszacowania OLS parametru.

Rysunek 4.Gęstości brzegowe parametru TORD



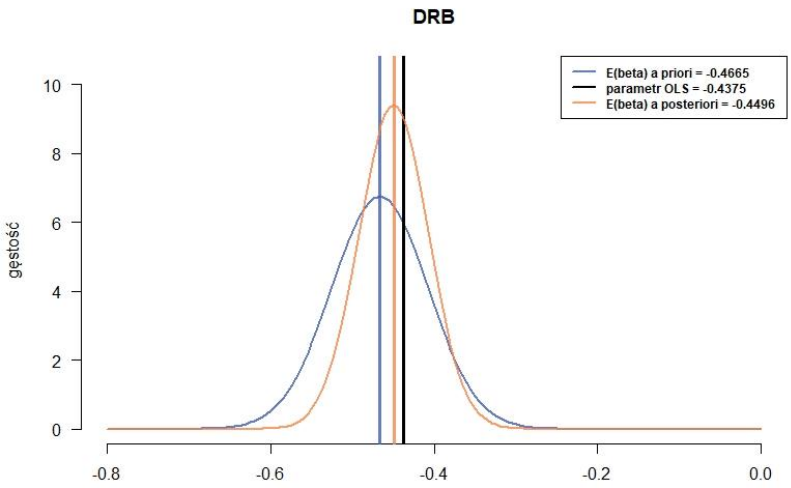
Na rysunku czwartym widzimy rozkłady brzegowe dla parametru TORD. Podobnie jak przy poprzednich zmiennych wartość oczekiwana rozkładu *a posteriori* znajduję się pomiędzy wartością oczekiwaną rozkładu *a priori* oraz wartością oszacowania OLS. Widać, że wartość oszacowania OLS parametru jest bardzo blisko wartości oczekiwanej *a posteriori* więc nasze przekonanie co do wpływy tej zmiennej nie ulegnie dużej zmianie.

Rysunek 5. Gęstości brzegowe parametru ORB



W przypadku zmiennej ORB również przekonanie co do wpływu zmiennej może ulec osłabieniu patrząc na rozkład *a posteriori*.

Rysunek 6. Gęstości brzegowe parametru DRB



Na rysunku 6 widzimy gęstości brzegowe parametru DRB. Jest to o tyle ciekawszy przypadek spośród zmiennych gdyż wartość oczekiwana *a posteriori* jest bardziej zbliżona do oszacowania parametru OLS, może wynikać to z faktu większej niepewności informacji *a priori* co do tego parametru.

HPDI

W celu dalszej analizy rozkładów *a posteriori* parametrów skonstruowano 95% przedziały ufności o najwyższej gęstości, wyniki prezentują się w sposób następujący:

Rysunek 7. HPDI dla zmiennych

Zmienna	HPDI
G	(0.77 ; 0.82)
EFG_0	(0.89 ; 0.94)
TOR	(-0.73 ; -0.66)
TORD	(0.76 ; 0.82)
ORB	(0.39 ; 0.42)
DRB	(-0.47 ; -0.43)

Dla wszystkich zmiennych przedziały wskazują, że rozkłady *a posteriori* przyjmowały dosyć wąski kształt. Żadna ze zmiennych na zadanym przedziale ufności nie zmienia znaku co nie daje żadnych podstaw do obaw o jakość informacji z tych rozkładów.

3. Podsumowanie

Po przeprowadzonej analizie można wyciągnąć następujące wnioski. Przy uwzględnieniu informacji na temat statystyk z zeszłego sezonu możemy stwierdzić, że nasze przekonania co do wpływu zmiennych: G, TOR i ORB może ulec osłabieniu. W pozostałych przypadkach przekonania mogły się umocnić.

Wyniki estymacji Bayesowskiej dla poszczególnych zmiennych były bardzo zbliżone do siebie pod względem charakterystyk. Prawdopodobnie wynika to z oszacowań modeli liniowych dla sezonu 2019 oraz 2018, gdzie modele te przedstawiały bardzo dobre oszacowania parametrów z niskimi błędami standardowymi. Ponadto w procesie wybierania zmiennych odrzucono te mało istotne oraz wyeliminowano te powodujące występowanie współliniowość w modelu.