Międzynarodowe zróżnicowanie popularności whiskey irlandzkiej - analiza z wykorzystaniem metod bayesowskich

Jakub Cierocki
Ekonometria bayesowska, MIESI, SGH
2021, 13 czerwca

Spis treści

Wprowadzenie

W niniejszej pracy zostanie przeanalizowana sprzedaż whiskey irlandzkiej per capita w latach 1990-2016 w krajach EU14 (sprzed Brexitu), USA oraz Kanadzie. Jej konsumpcja ustawicznie rośnie na przestrzeni ostatnich lat, jest jednak bardzo zróżnicowana między krajami i ze względu na klasy jakości. Celem tej pracy będzie w tej sytuacji pogłębienie oczywistych wniosków wynikających z opisowej analizy danych oraz ich sformalizowanie w postaci modelu matematycznego. Podejście bayesowskie jest w tym przypadku szczególnie obiecujące ze względu na liczne braki danych ładność elicytacji a priori.

Zbiór danych

W dalszej analizie zostaną wykorzystane dane zebrane przez ***The ISWR***, a opublikowane przez ***Irish Food Board***, z okazji obchodzów Dnia Św. Patryka 2018, na potrzeby prostego konkursu wizualizacyjnego. Obecnie dostęp do danych można uzyskać za pomocą portalu ***data.world***.

Surowe dane zawierają 5 kolumn po 4131 wierszy i mają charakter makro-panelu w formacie wzdłużnym: każdy wiersz zawiera informacje o skumulowanej sprzedaży whiskey irlandzkiej dla konkretnego kraju, roku oraz klasy jakości (standardowa, premium lub "super premium").

W celu ograniczenia liczby anomalii, braków danych oraz redukcji wymiaru (po konwersji zmiennych kategoryzowanych na binarne) zbiór danych ograniczono do 14 krajów EU14 (na rok 2018) oraz państw anglosaskich Ameryki Płn.: USA i Kanady. W dalszej kolejności ze zbioru usunięto dane o sprzedaży whiskey "Super Premium" z powodu małej liczby (2) obserwacji. Równocześnie okazało się koniecznym usunąć ze zbioru danych Grecję ze względu na brak jakichkolwiek danych dla klas "Premium" i "Super Premium". Dane o oryginalnym wymiarze również zostały przetestowane, ale ich użycie wiązało się z poważnymi trudności w zakresie obliczeń numerycznych, które w optymistycznym przypadku sprowadzały się do kilkudziesięciominutowych czasowych pracy algorytmu NUTS, a uzyskane wyniki, w szczególności wykresy, były bardzo trudne do analizy.

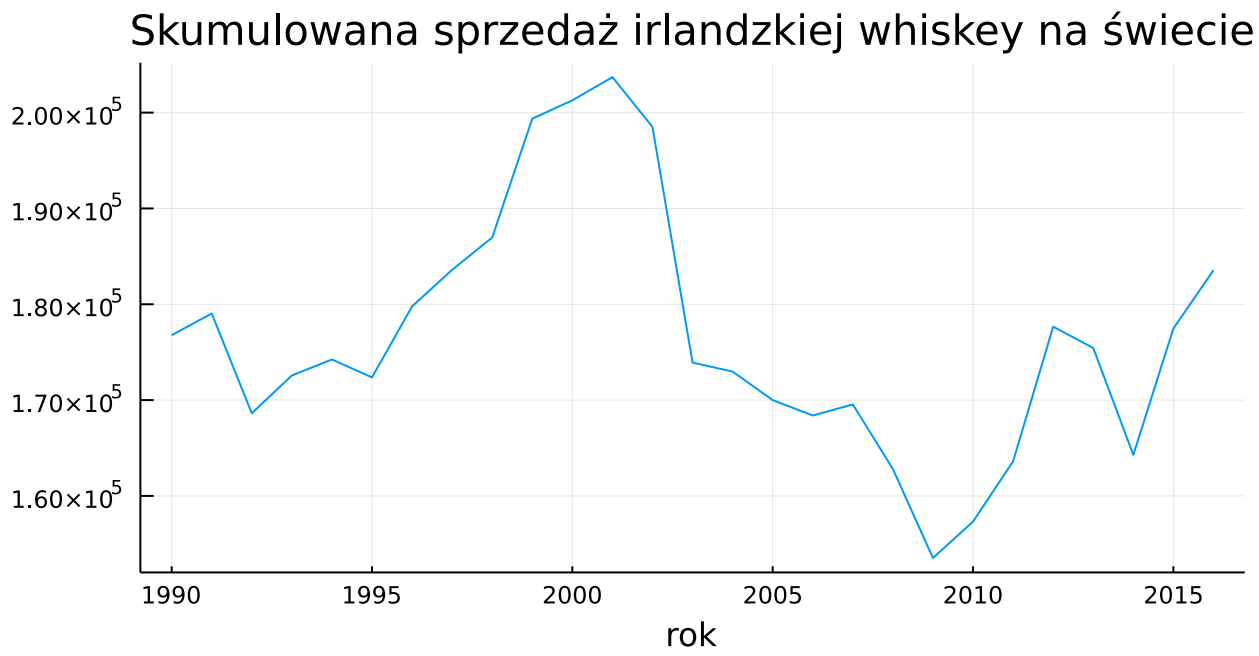
Dane oryginalne były wyrażone w wartościach bezwzględnych i w celu ich przeskalowania do postaci per capita (skumulowana sprzedaż roczna na 1 mln mieszkańców: `popularity`) zostały wykorzystane dane demograficzne publikowane przez *Bank Światowy*. Dodano ponadto jej opóźnienie 1 rzędu: `popularity_lag`, które będzie dalej wykorzystywane jako zmienna objaśniająca.

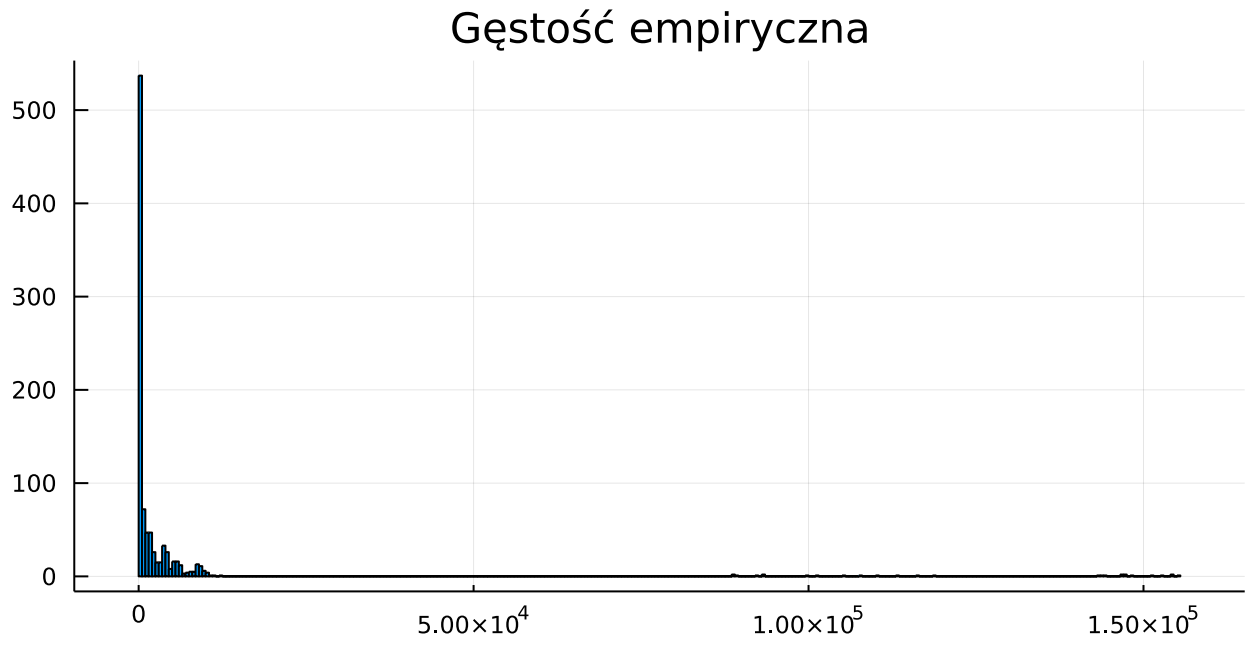
Proces wstępnej obróbki danych omówionych powyżej został zaimplementowany w pliku `src/preproc.jl`. W wyniku jego zastosowania otrzymaliśmy tabelę w postaci:

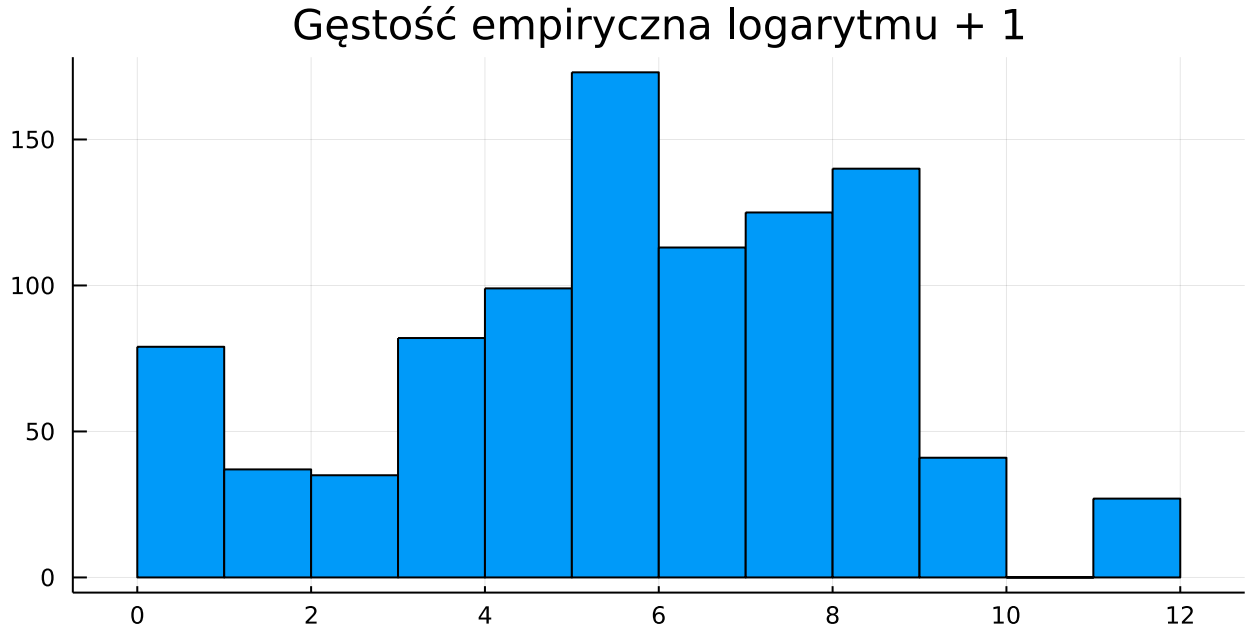
951 rows × 5 columns

	country String	quality String	year Int64	popularity Float64	popularity_lag Float64?
1	Austria	Premium	1998	31.3409	*missing*
2	Austria	Premium	1999	50.048	31.3409
3	Austria	Premium	2000	43.6868	50.048
4	Austria	Premium	2001	52.2239	43.6868
5	Austria	Premium	2002	77.9514	52.2239
6	Austria	Premium	2003	118.206	77.9514
7	Austria	Premium	2004	117.475	118.206
8	Austria	Premium	2005	173.8	117.475
9	Austria	Premium	2006	226.156	173.8
10	Austria	Premium	2007	231.451	226.156
11	Austria	Premium	2008	372.529	231.451
12	Austria	Premium	2009	367.959	372.529
13	Austria	Premium	2010	394.576	367.959
14	Austria	Premium	2011	399.207	394.576
15	Austria	Premium	2012	421.116	399.207
16	Austria	Premium	2013	430.434	421.116
17	Austria	Premium	2014	362.728	430.434
18	Austria	Premium	2015	439.677	362.728
⋮	⋮	⋮	⋮	⋮	⋮

Zmienna objaśniana dla zredukowanego zbioru danych ma postać:




Gęstość empiryczna
500
400
300
200
100
0
0
5.00×10^4
1.00×10^5
1.50×10^5


Gęstość empiryczna logarytmu + 1
150
100
50
0
0
2
4
6
8
10
12

Ze względu na bardzo silną prawostronną skośność zmiennej objaśnianej w dalszej analizie zostanie wykorzystany jej logarytm (analogicznie dla zmiennej opóźnionej), który w dalszym ciągu jednak jest daleki od rozkładu normalnego.

Zmienna `quality`, wyrażająca klasę jakości trunku, po początkowych próbach z konwersją na binarne, została potraktowana jako zmienna liczbowa z wartościami odpowiednio:

- Standard: -1,
- Premium: 0,
- Super Premium: 1,

co było możliwe z racji jej uporządkowania i generowania monotonicznej zależności ze zmienną objaśnianą.

Zmienna `country` została z kolei, zgodnie z wcześniejszymi zapowiedziami, przekonwertowana na zmienne binarne odpowiadające poszczególnym krajom, z wyłączeniem Włoch, które zostały pominięte w celu uniknięcia współliniwości. Wybór Włoch wynika z najniższego dla tego kraju średniego wskaźnika spożycia, który czyni je dobrym punktem odniesienia.

	country	quality	year	y	y_lag	Austria	Belgium and Luxembourg	Canad
1	"Austria"	0.0	1998	3.47633	missing	1.0	0.0	0.0
2	"Austria"	0.0	1999	3.93277	3.47633	1.0	0.0	0.0
3	"Austria"	0.0	2000	3.79968	3.93277	1.0	0.0	0.0
4	"Austria"	0.0	2001	3.97451	3.79968	1.0	0.0	0.0
5	"Austria"	0.0	2002	4.36883	3.97451	1.0	0.0	0.0
6	"Austria"	0.0	2003	4.78085	4.36883	1.0	0.0	0.0
7	"Austria"	0.0	2004	4.7747	4.78085	1.0	0.0	0.0
8	"Austria"	0.0	2005	5.16364	4.7747	1.0	0.0	0.0
9	"Austria"	0.0	2006	5.42564	5.16364	1.0	0.0	0.0
10	"Austria"	0.0	2007	5.44868	5.42564	1.0	0.0	0.0

Klasyczny model ekonometryczny

Wyestymujemy teraz klasyczny model OLS, który docelowo będzie stanowić punkt odniesienia dla dalszej analizy bayesowskiej.

StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Vector{Float64}}, GLM.DensePrec

y ~ 1 + quality + y_lag + Austria + Belgium and Luxembourg + Canada + Denmark + Finlanc

Coefficients:

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	0.347639	0.0603428	5.76	<1e-07	0.229209	0.46607
quality	-0.232949	0.037598	-6.20	<1e-09	-0.30674	-0.159158
y_lag	0.909103	0.0111767	81.34	<1e-99	0.887167	0.931039
Austria	0.188324	0.0728922	2.58	0.0099	0.0452637	0.331385
Belgium and Luxembourg	0.140243	0.076131	1.84	0.0658	-0.00917413	0.28966
Canada	0.184901	0.0752013	2.46	0.0141	0.0373087	0.332494
Denmark	0.166732	0.0738344	2.26	0.0242	0.0218217	0.311642
Finland	0.208736	0.0778194	2.68	0.0074	0.0560049	0.361467
France	0.250303	0.0806362	3.10	0.0020	0.092044	0.408563
Germany	0.174014	0.0701957	2.48	0.0134	0.0362456	0.311783
Ireland	0.458696	0.0953705	4.81	<1e-05	0.271518	0.645873
Netherlands	0.162529	0.0766244	2.12	0.0342	0.0121431	0.312914
Portugal	0.246123	0.0768114	3.20	0.0014	0.09537	0.396875
Spain	0.0500235	0.0773717	0.65	0.5181	-0.101829	0.201876
Sweden	0.266012	0.0810692	3.28	0.0011	0.106903	0.425122
United Kingdom	0.198154	0.0797128	2.49	0.0131	0.0417068	0.354601
United States	0.233144	0.0734455	3.17	0.0016	0.0889976	0.377291

```
begin
    using GLM

    df_lm = @pipe df_model |>
        select(_, Not([:country, :year])) |>
        dropmissing

    lm(Term(:y) ~ sum(Term.(Symbol.(names(df_lm[:, Not(:y)])))), df_lm)
end
```

Model bayesowski

Specyfikacja

$$
\begin{array}{lcl}
X & - & \text{macierz zmiennych objaśniających} \\
K & - & \text{liczba obserwacji zawierających braki danych} \\
\sigma^2 & \sim & InvGamma(\underline{\alpha}_{\sigma^2}, \underline{\beta}_{\sigma^2}) \\
\alpha & \sim & \mathcal{N}(\underline{\mu}_{\alpha}, \underline{\sigma}_{\alpha}) \\
\beta & \sim & \mathcal{N}(\underline{\mu}_{\beta}, \underline{\sigma}_{\beta}) \\
y_i & \overset{\forall_{k=1,..K}}{\sim} & \mathcal{N}(\hat{\mu}_{y_i}, \hat{\sigma}_{y_i}) \\
\overline{\mathbf{y}} & \sim & \mathcal{N}(\alpha + \mathbf{X} \times \beta, \sigma^2)
\end{array}
$$

Opisany model posida łącznie 36 parametrów *a priori*, po 2 dla stałej i odchylenia, oraz po 2 dla każdego z 16 parametrów modelu.

Elicytacja parametrów *a priori*

Przyjmijmy $\underline{\alpha}_{\sigma^2}, \underline{\beta}_{\sigma^2} = 1$, daje to nam rozkład o długim ogonie i duże wariancji dobrze obrazujący naszą ograniczoną wiedzę na temat wariancji składnika losowego.

Podobne podejście zostosujmy do stałej: $\underline{\mu}_{\alpha} = 3, \underline{\sigma}_{\alpha} = 25$, duża wariancja odpowiada ograniczonej wiedzy o rozkładzie parametru.

Zmienne objaśniające zawarte w $\mathbf{X}$ można podzielić na 3 grupy:

- opóźnioną zmienną objaśnianą y_{t-1} (zlogarytmowaną liczbową)
- zmienną jakościową uporządkowaną $quality$
- zmienne binarne krajów

W celu elicytacji parametrów rozkładów współczynników z wektora β korzystamy z własności modelu log-liniowego, pozwalającej określić przybliżone interpretacje parametrów:

- zależność log-log: zmiana x o $1\ \%$ powoduje zmianę y o około $\beta\ \%$ ceteris paribus
- zależność log-raw: zmiana x o 1 powoduje zmianę y o około $100 * \beta\ \%$

W naszym modelu mamy do czynienia ze (w sposób oczywisty) niestacjonarnym procesem autoregresyjnym, tj. sprzedaż w danym roku jest silnie powiązana z ubiegłoroczną i reprezentuje podobny rząd wielkości. Przyjmijmy w tej sytuacji $\mu = 0.8$ co oznacza, zgodnie z wcześniej przedstawionymi regułami, że y wzrośnie średnio o $0.8\ \%$ przy wzroście x o $1\ \%$ oraz $\sigma = 0.2$ co w świetle reguły 3 sigm oznacza, że $75\ \%$ masy rozkładu parametru znajdzie się w przedziale 0.61.

Zdroworoządnowo, a w dodatku mając w papięci wcześniejszej analizy klasycznej, zakładamy, że "wzrost" jakoś trudnku powinien wpływać na spadek jego popularności. Przyjmijmy w tej sytuacji $\mu = -0.4$ co odpowiada spadkowi sprzedaży o $40\ \%$ przy przejściu do wyższej klasy jakości. Z racji, że wartości tej jesteśmy już mniej pewni, dobieramy do niej relatywnie większe $\sigma = 0.2$.

W przypadku zmiennych binarnych, interpretacja w modelu ze zlogarytmowaną zmienną objaśnianą jest dość prosta, a mianowicie zmiana wartości z "NIE" na "TAK" skutkuje wzrostem wartości zmiennej objaśnianej o $100 * \beta \%$. Nie mają dostępu do szczegółowych badań na poziomach krajowych wyróżnijmy 4 różne grupy krajówi odpowiadające im wartości oczekiwane rozkładów:

- macierzystą Irlandię: 0.5
- kraje anglosaskie: 0.3
- pozostałe kraje Europy Środokowo-Zachodniej i Północnej: 0.2
- kraje basenu Morza Śródziemnego (Hiszpania): : 0.1

Nie będąc w stanie wyznacznyć miarodajnie wartości dla każego kraju z osobna, ani nawet nie do końca dla grup, skupiliśmy się na wyodrębnieniu 4 grup możliwie różnych pod względami kultury, klimatu i zamożności, w sposób powiązany z konsumpcją droższych alkoholi wysokoprocentowych oraz przedstawieniu dysproporcji międzygrupowych. Zakładamy, że te rzeczywiste będą zbliżone do różnic między wartościami oczekiwanymi rozkładów *a priori*. W celu uwzględnienia przede wszystkim zmienności wewnątrzgrupowej przyjmujemy relatywnie duże $\sigma = 0.2$.

Zaporponowana specyfikacja uwzględnia również imputację bayesowską brakujących wartości zmiennej zależnej, ale w tym celu zostaną wykorzystane rozkłady empiryczne momenty dla poszczególnych podgrup (kraj-klasa jakości).

```
0.2
```

Implementacja modelu

Opisany wyżej model został zaimplementowany z użyciem pakietu *Turing.jl*, napisanego od zera w Julii subjęzyka probabilistycznego pozwalającego budować modele w tym samym języku co resztę analizy zachowując ponadto wydajność zbliżoną do *Stan*'a oraz liczne analogie w zakresie logicznej struktury kodu. Wśród funkcjonalności tego narzędzia, które zostaną mocniej wykorzystane w niniejszej analizie jest automatyczna detekcja brakujących wartości bez konieczności definiowania dodatkowej flagi i wyrażenia warunkowego jak w *Stan*'ie.

Kod samego modelu prezentuje się następująco:

```
using Turing, LazyArrays
```

```
mvar_reg1 (generic function with 1 method)
```

```
begin
    @model function mvar_reg1(y, y_lag, X)
        σ² ~ InverseGamma(1, 1)

        α ~ Normal(intercept_mu_prior, intercept_sigma_prior)

        β₁ ~ Normal(lag_mu_prior, lag_sigma_prior)
        β ~ arraydist(LazyArray(@~ Normal.(X_mu_prior, X_sigma_prior)))

        for i in eachindex(y_lag)
            y_lag[i] ~ Normal(default_mu[i], default_sigma[i])
        end

        μ = α .+ y_lag .* β₁ .+ X * β
        y ~ MvNormal(μ, sqrt(σ²))
    end
end
```

Model będziemy próbkować z użyciem *No-U-Turn-Sampler* (`NUTS()`) z limited kroków optymalizatora ustawionym na 1000 i tolerancją 0.65. Wartość początkową parametrów funkcji wiarygodności pozostawimy niezdefiniowaną co poskutkuje jej dobraniej za pomocą domyślnej procedury heurystycznej.

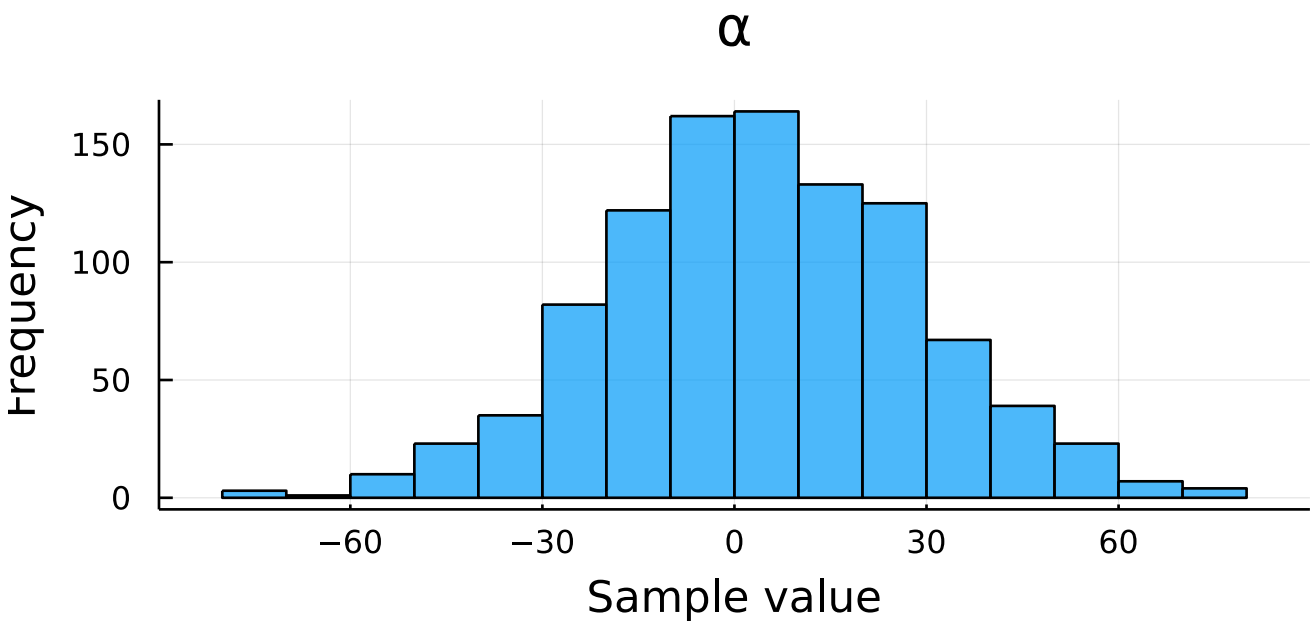
```
NUTS(1000, 0.65, 10, 1000.0, 0.0)
```

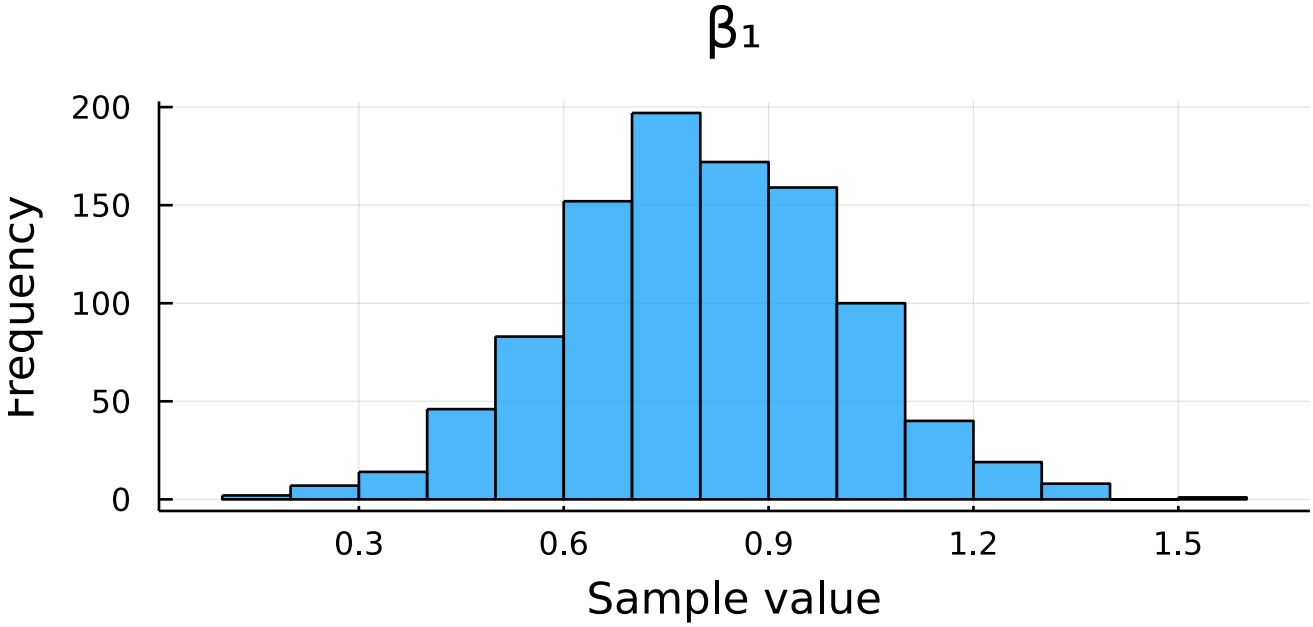
```
begin
    y = df_model.y
    y_lag = df_model.y_lag
    X = @pipe df_model |>
        select(_, Not([:country, :year, :y, :y_lag])) |>
        Matrix

    model = mvar_reg1(y, y_lag, X)
    alg = NUTS(1000, 0.65)
end
```

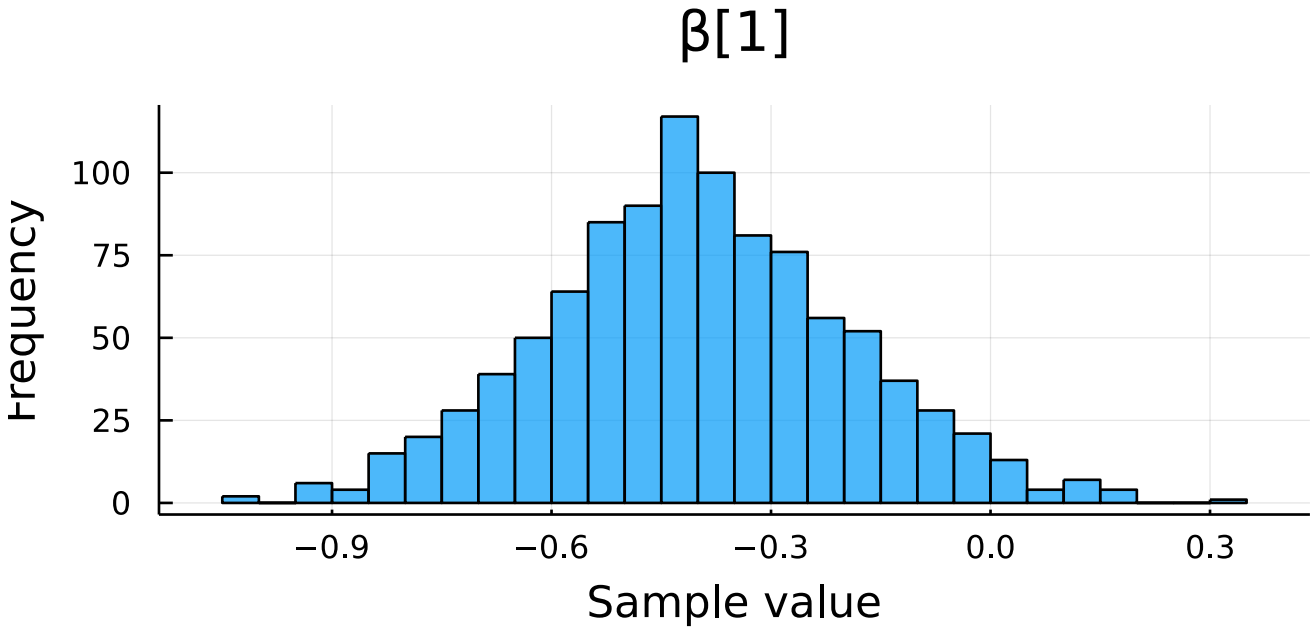
Przeprowadzimy najpierw próbkowanie rozkładu *a priori*

	parameters	mean	std	naive_se	mcse	ess	rhat
1	:α	4.24689	24.5219	0.77545	0.841375	927.438	0.999007
2	:β₁	0.805257	0.204933	0.00648056	0.0105223	858.454	0.999194
3	Symbol("β[1]")	-0.402669	0.206663	0.00653525	0.00723795	1101.13	0.999017
4	Symbol("β[2]")	0.198947	0.199659	0.00631379	0.00841627	911.433	1.00081
5	Symbol("β[3]")	0.201758	0.195142	0.00617094	0.00541096	983.167	0.999577
6	Symbol("β[4]")	0.304628	0.202495	0.00640346	0.00669461	970.946	1.00091
7	Symbol("β[5]")	0.208464	0.201308	0.00636592	0.00650908	903.401	0.999978
8	Symbol("β[6]")	0.191756	0.197316	0.00623968	0.00374637	873.348	0.999009
9	Symbol("β[7]")	0.206457	0.201695	0.00637817	0.00823314	809.521	1.00079
10	Symbol("β[8]")	0.495908	0.195474	0.00618145	0.00664091	920.779	0.999209
11	Symbol("β[9]")	0.505265	0.204752	0.00647482	0.00589205	893.285	0.999144

```
begin
    using MCMCChains

    model_params = vcat("α", "β₁", ["β[$idx]" for idx in 1:15])

    prior_chain = sample(model, Prior(), 1000)[model_params]

    summarize(prior_chain)
end
```


α
Frequency
150
100
50
0
−60
−30
0
30
60
Sample value


β₁
Frequency
200
150
100
50
0
0.3
0.6
0.9
1.2
1.5
Sample value


β[1]
Frequency
100
75
50
25
0
−0.9
−0.6
−0.3
0.0
0.3
Sample value

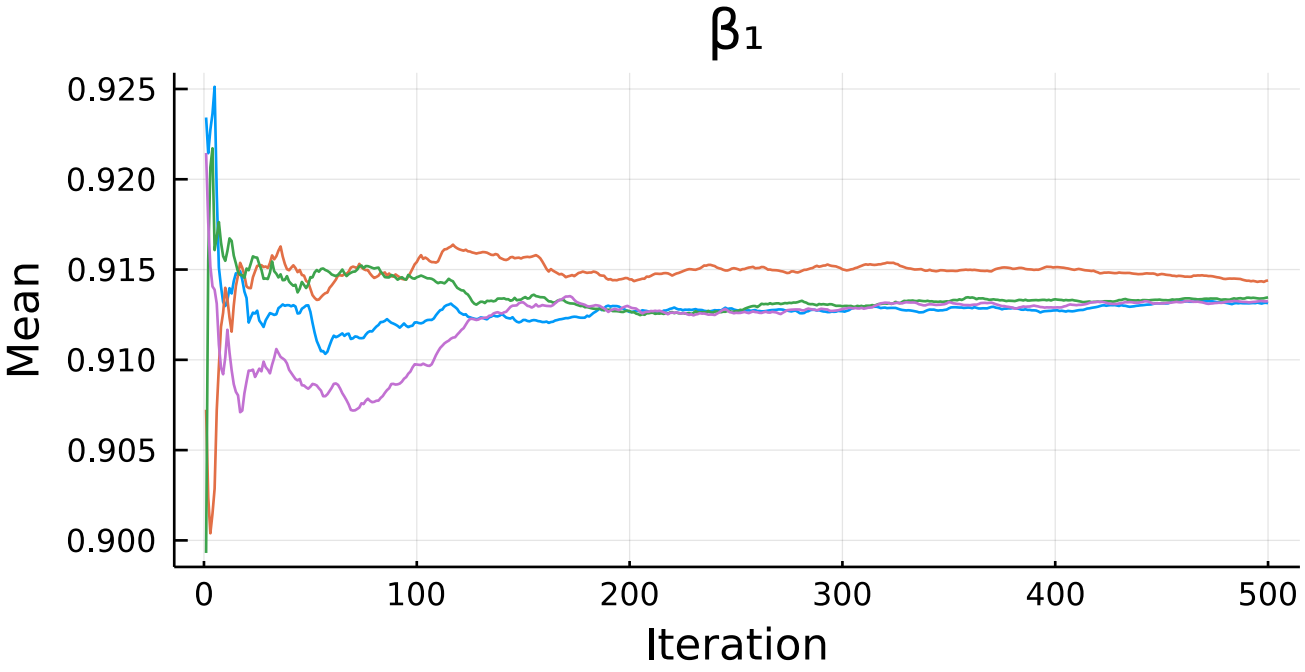
Ewaluacja modelu *a posteriori*

Na tym etapie przechodzimy do właściwego modelu bayesowskiego: modelu regresji Normalnego-Gamma z imputacją bayesowską brakujących wartości.

	parameters	mean	std	naive_se	mcse	ess	rhat
1	:α	0.292126	0.0553263	0.00123713	0.0014314	942.047	1.00168
2	:β₁	0.913563	0.0101811	0.000227657	0.00038341	951.66	1.00153
3	Symbol("β[1]")	-0.224615	0.0349273	0.000780999	0.0010927	1060.53	0.99998
4	Symbol("β[2]")	0.211895	0.06104	0.0013649	0.00154505	986.337	1.00056
5	Symbol("β[3]")	0.168112	0.066586	0.00148891	0.0013348	1177.85	0.99966
6	Symbol("β[4]")	0.211842	0.0615954	0.00137732	0.00149364	1083.63	0.99911
7	Symbol("β[5]")	0.197604	0.0621046	0.0013887	0.00162177	1026.06	1.00138
8	Symbol("β[6]")	0.22936	0.0622586	0.00139214	0.00163565	985.051	0.99951
9	Symbol("β[7]")	0.24697	0.0640528	0.00143226	0.00191058	864.765	1.00049
10	Symbol("β[8]")	0.216998	0.0569539	0.00127353	0.00130607	941.003	0.99927
11	Symbol("β[9]")	0.477082	0.0739556	0.0016537	0.00270507	790.991	1.00164

```
begin
    chain = sample(model, alg, MCMCThreads(), 500, 4)[model_params]

    summarize(chain)
end
```

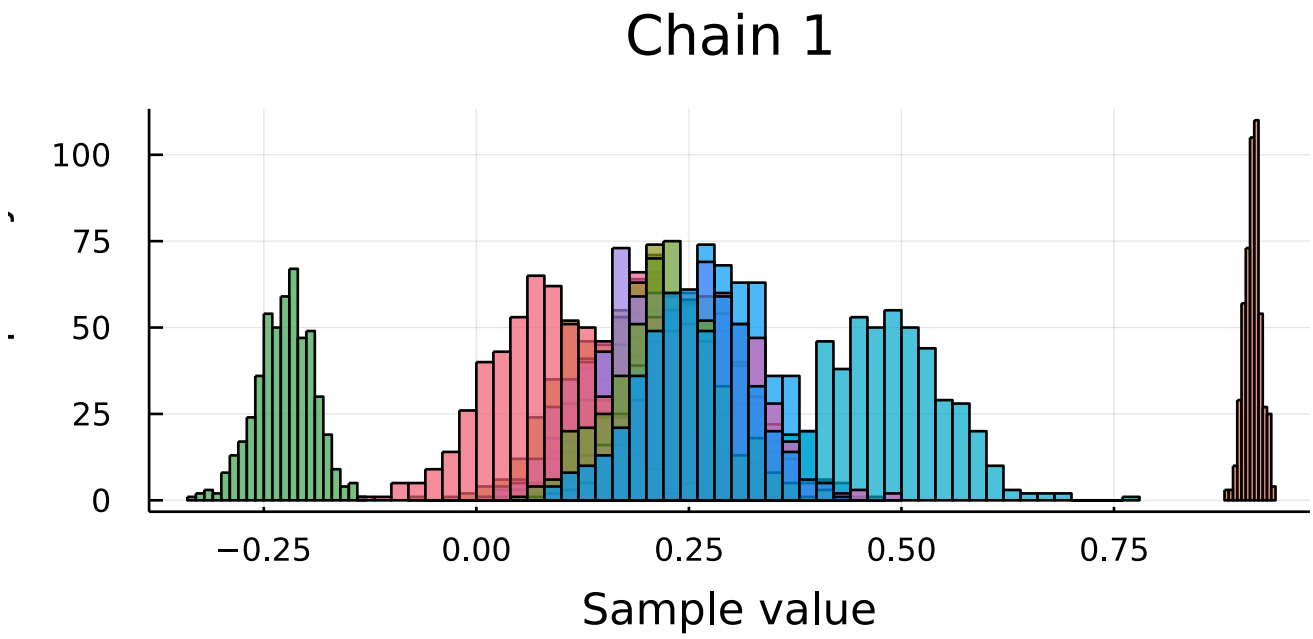
W oparciu o wartości statystyki `rhat`, która w przypadku każdego ze współczynników modelu znajduje się w okolicach 1-ki, możemy stwierdzić zbieżność próbkowanych łańcuchów. Zweryfikujmy to dodatkowo przy użyciu adekwatnego wykresu, kontrolnie tylko dla parametru β_1:

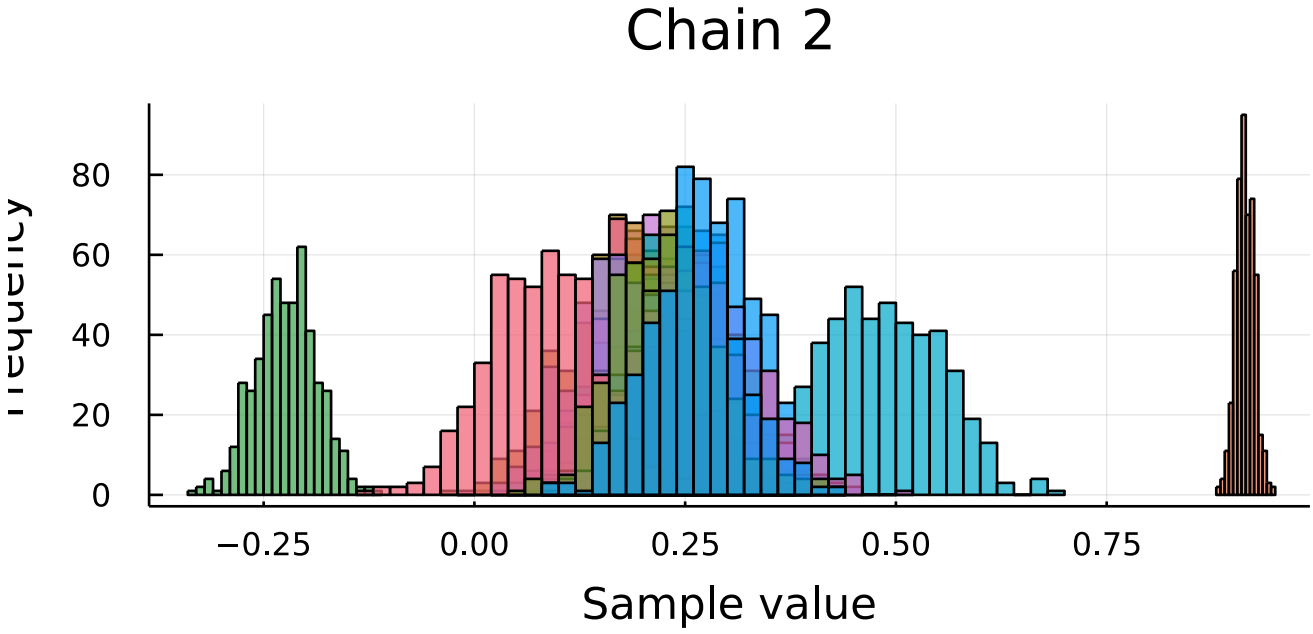


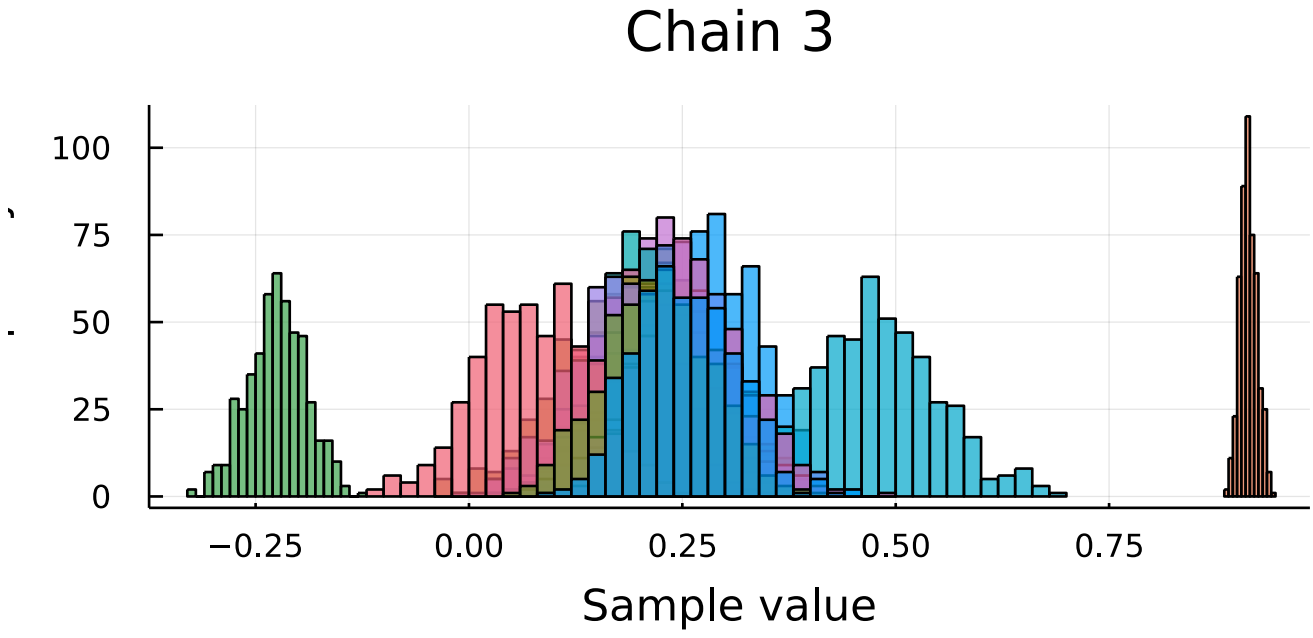
O ustabilizowaniu wartości średniej z łańcucha na zbliżonych poziomach dla wszystkich 4 łańcuchów można już było mówić de facto w okolicach 200 iteracji.

Same wartości oczekiwane współczynników zmieniły się jedynie nieznacznie w stosunku do modelu nieuwzględniającego informacji z poza próby, co jest dość naturalne z racji na relatywnie dużą liczbę obserwacji w zbiorze uczącym (951).

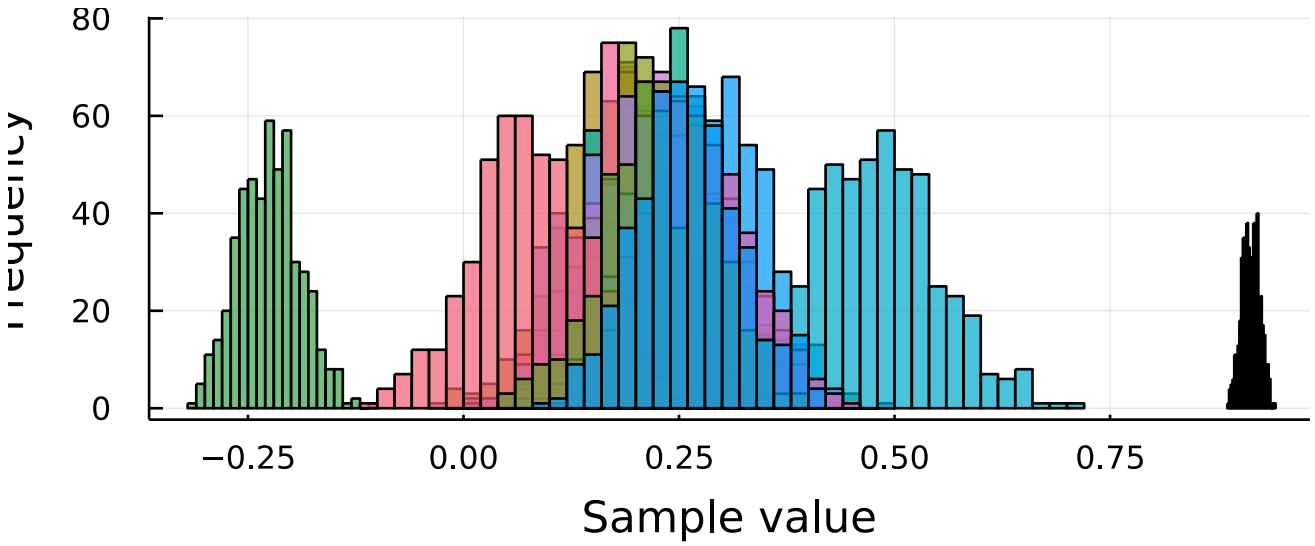
Zweryfikujmy teraz uzyskane rozkłady:


Chain 1
100
75
50
25
0
−0.25
0.00
0.25
0.50
0.75
Sample value


Chain 2
Frequency
80
60
40
20
0
−0.25
0.00
0.25
0.50
0.75
Sample value


Chain 3
100
75
50
25
0
−0.25
0.00
0.25
0.50
0.75
Sample value


Chain 4



Dla wszystkich łańcuchów poza 4-tym uzyskaliśmy bardzo analogiczne rozkłady współczynników *a posteriori*:

- β_1 (opóźniona zmienna objaśniana) - czarny/brązowy histogram po prawej, charakteryzujacy się bardzo małą wariancją;
- β_2 (klasa jakości) - zielony histogram po lewej;
- $\beta_3 : \beta_K$ (zmienne binarne określające kraj) - zgrupowanie histogramów w środku wykresu.

Użyta platforma sprzętowa

```
Julia Version 1.6.1
Commit 6aaedecc44 (2021-04-23 05:59 UTC)
Platform Info:
  OS: Linux (x86_64-pc-linux-gnu)
  CPU: AMD Ryzen 9 5900X 12-Core Processor
  WORD_SIZE: 64
  LIBM: libopenlibm
  LLVM: libLLVM-11.0.1 (ORCJIT, generic)
Environment:
  JULIA_REVISE_WORKER_ONLY = 1
```

Bibliografia

- Ge, H., Xu, K. & Ghahramani, Z.. (2018). Turing: A Language for Flexible Probabilistic Inference. Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 84:1682-1690 Available from http://proceedings.mlr.press/v84/ge18b.html .
- Hoffman, M. D., & Gelman, A. (2011, November 18). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. arXiv.org. https://arxiv.org/abs/1111.4246.
- 2018/W11: Growth in Irish Whiskey Sales - dataset by makeovermonday. data.world. (2018, March 11). https://data.world/makeovermonday/2018w11-growth-in-irish-whiskey-sales.