Raport - Lista 4 ZML

Erwin Jasic

5 czerwca 2021

Na początku ustalimy pewne oznaczenia, które ułatwią zapis:

- n - liczba obiektów,
- k - liczba pomiarów na każdym obiekcie,
- p - liczba kolumn w macierzy planu,
- $N = n \cdot k$ - liczba zmiennych objaśnianych y_{ij}.

Cel raportu:

Celem raportu jest przeprowadzenie analizy jak zmieniają się poszczególne estymatory (β, γ, ρ) podczas, gdy manimulujemy wartościami n, k i p.

Zadanie 1:

W tym zadaniu przyjmujemy $n = 20$, $k = 3$, $p = 4$.

W podpunkcie (a) generujemy macierz zgodnie z treścią oraz dzielimy ją na $n = N/k$ podmacierzy. Ustalamy $\beta = (3, 3, 0)' \in \mathbb{R}^{p-1}$. Tworzymy macierz postaci jak w treści z parametrami $\rho = 0.3$ oraz $\gamma = 2$.

Następnie w podpunkcie (b) generujemy n niezależnych wektorów losowych następującej postaci:

$$y_i = (y_{i1}, ..., y_{ik})' \sim N(X_i\beta, \Sigma) \in \mathbb{R}^k,$$

gdzie $i = 1, 2, ...n$. Zapisujemy dane w postaci jednowymiarowej, a następnie skorzystamy z funkcji *gls()* (z metodą "REML") z pakietu *nlme*. Tworzymy ten model zgodnie z treścią.

W podpunkcie (c) będziemy porównywać rzeczywiste wartości parametrów β, ρ i γ z ich estymatorami. Z wykładu wiemy, że wzór na estymator $\hat{\beta}$ jest równy

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}^{-1} y_i \right).$$

Z kolei wzór na jego kowariancje jest postaci

$$cov(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}^{-1} X_i \right)^{-1}.$$

Porównananie wyników uzyskanych poprzez funkcję *vcov()* oraz estymacje wzorami.

Norma supremum różnicy:

	Norma
$\hat{\beta}$	0.054
cov	0.090

Porównanie odpowiednich parametrów z ich estymatorem:

	Parametr	Estymator
ρ	0.3	0.397
γ	2.0	2.198

Wnioski:

Funkcja *gls()* daje wartości bliskie teoretycznym oszacowaniom. Widzimy to, ponieważ wartościw tabeli są małe.

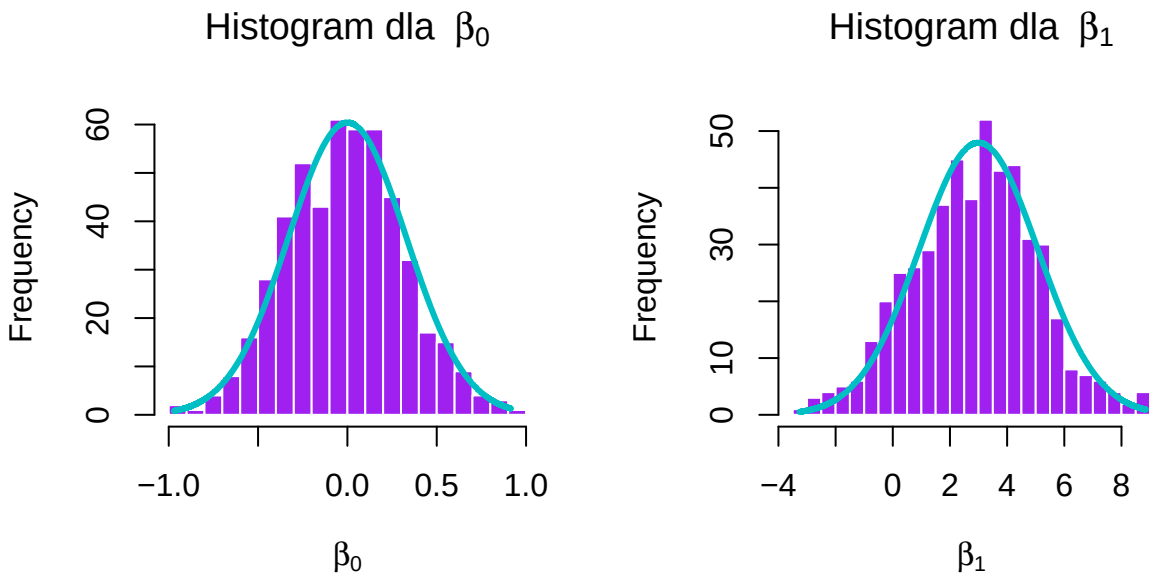
Parametry ρ i γ zostały dobrze oszacowane. Wartości wyestymowane i prawdziwe nie różnią się znacząco. Może się wydawać, że jest to spowodoane tym, że zadaliśmy odpowiednią strukture w funkcji *gls()*.

Zadanie 2

W tym zadaniu mamy powtórzyć podpunkt (a) z zadania 1 i następnie powtórzyć 500 razy podpunkt (b) z zadania 1. Uzyskamy 500 modeli i na ich podstawie wyznaczymy ciąg 500-ciuset estymatorów β, ρ i γ. Następnie na ich podstawie przetestujemy asymtotyczne własności estymatora β oraz asymtotyczne własności estymatora Σ.

Na początek wykonamy polecenia dla β, a następnie dla Σ.

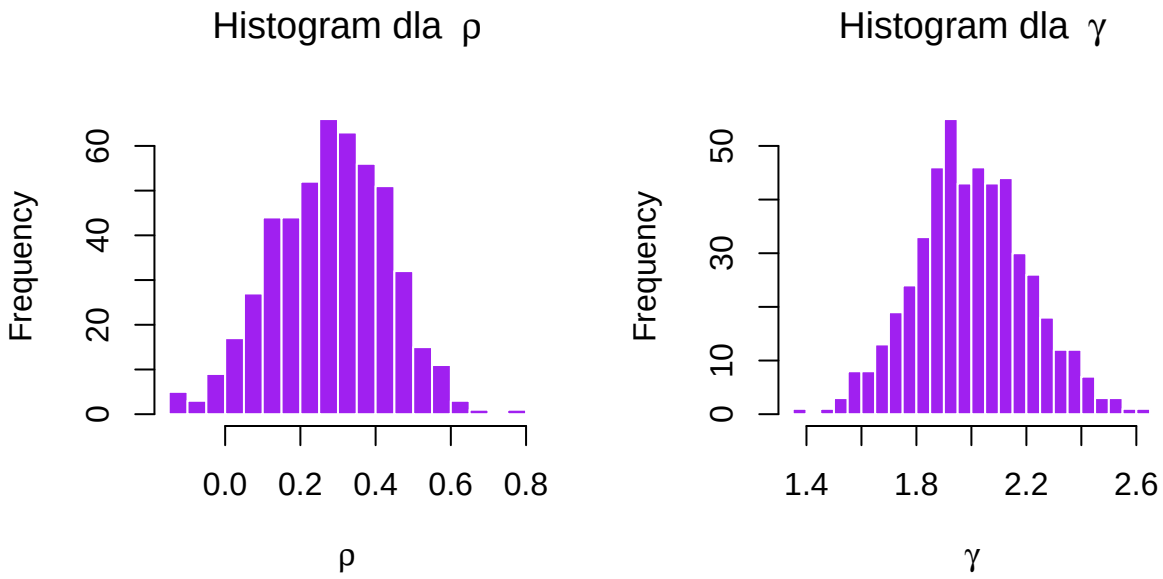
Histogramy dla β_0 oraz β_1:



Obciążenia dla β_0 oraz β_1:

	Dane.wyjściowe
Supremum	0.283
Średnia	0.108

Histogramy dla ρ oraz γ:



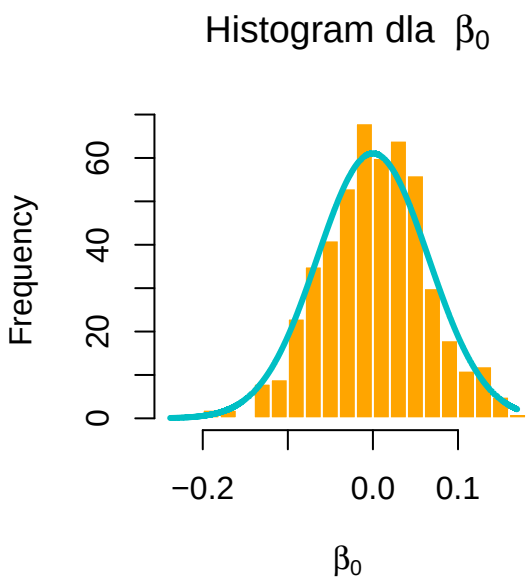
Obciążenia dla ρ i γ:

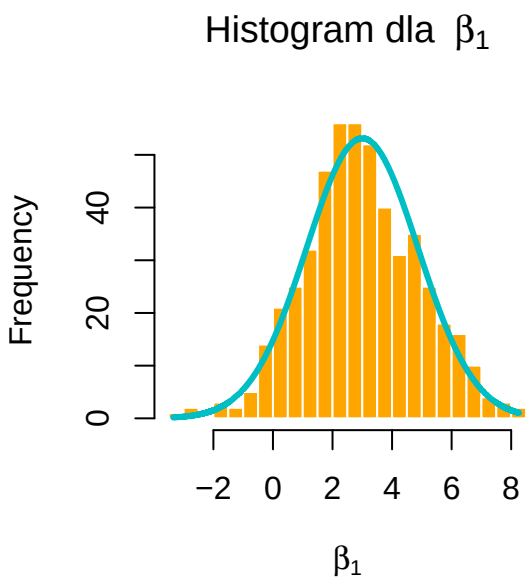
	Dane.wyjściowe
γ	0.00184
ρ	-0.01992

Zadanie 3

Teraz powtórzymy zadanie 2 dla $n = 500$.

Histogramy dla β_0 oraz β_1:





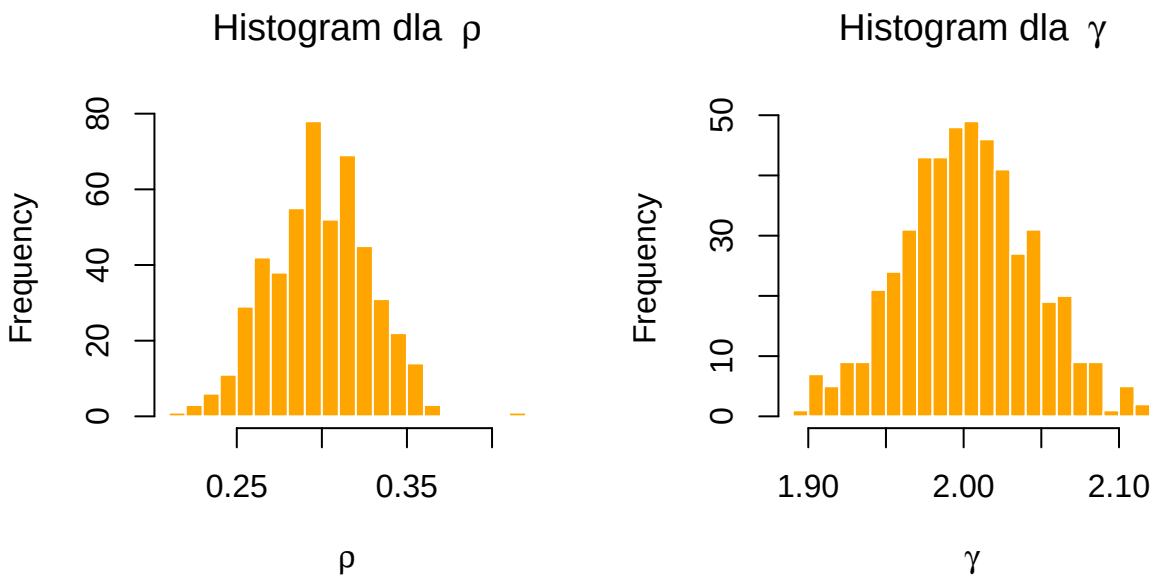
Możemy zaobserwować lepsze dopasowanie do swojego asymptotycznego rozkładu w przypadku modelu z ilością obiektów równą 500.

Obciążenia dla β_0 oraz β_1:

	n. . . 500
Supremum	0.144
Średnia	0.057

Wyraźnie widać spadek w wartościach obciążeń. Można wysnuć wniosek, że duża ilość obserwacji zmniejsza obciążenie estymatora.

Histogramy dla ρ oraz γ:



W przypadku estymatora dla ρ widzimy, że większa liczba obiektów powoduje zmniejszenie rozrzutu w estymatorze. Histogram dla dużego n jest bardziej skupiony w okół prawdziwej wartości niż dla mniejszego n.

Obciążenia dla ρ i γ:

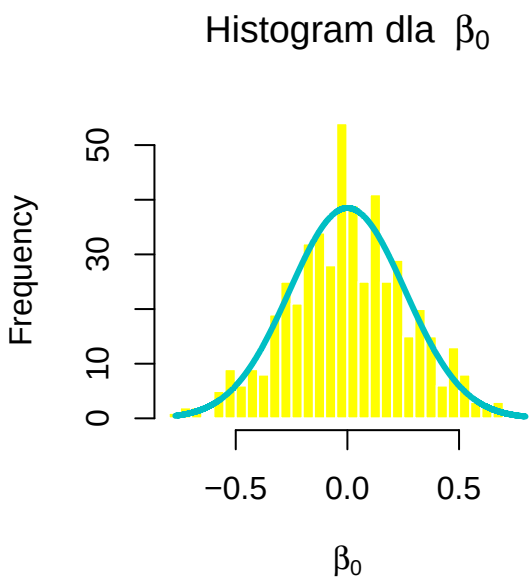
	n. . . 500
γ	0.00249
ρ	-0.00144

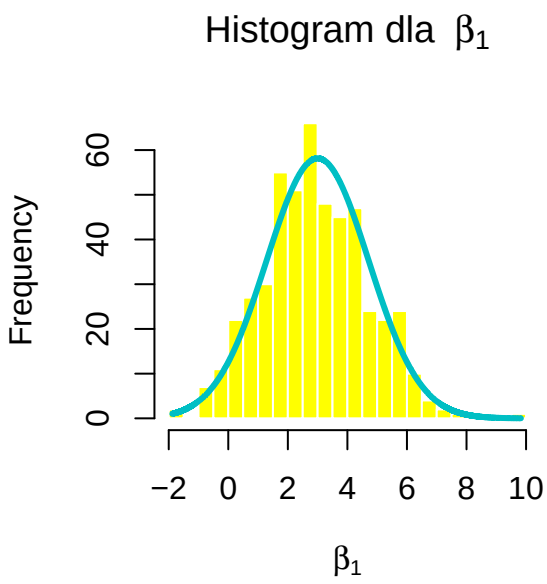
Nie widzimy specjlanych różnic w porównaniu z wyjściowymi danymi.

Zadanie 4

Teraz powtórzymy zadanie 2 dla $k = 30$.

Histogramy dla β_0 oraz β_1:





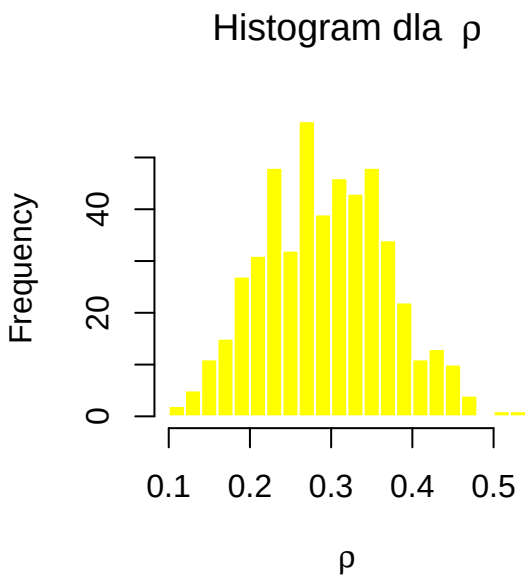
Widzimy, że histogramy są bardziej skoncentrowane.

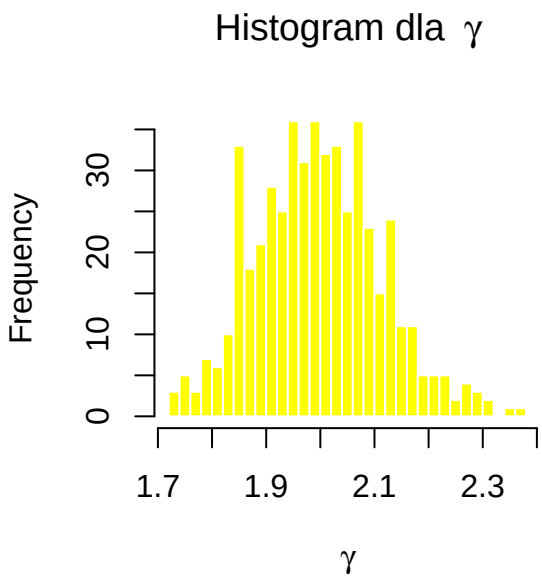
Obciążenia dla β_0 oraz β_1:

	k. . . 30
Supremum	0.168
Średnia	-0.057

Obiążenia dla większego k są trochę mniejsze.

Histogramy dla ρ oraz γ:





Tak samo jak dla β widzimy, że histogramy są bardziej skoncentrowane dla większego k.

Obciążenia dla ρ i γ:

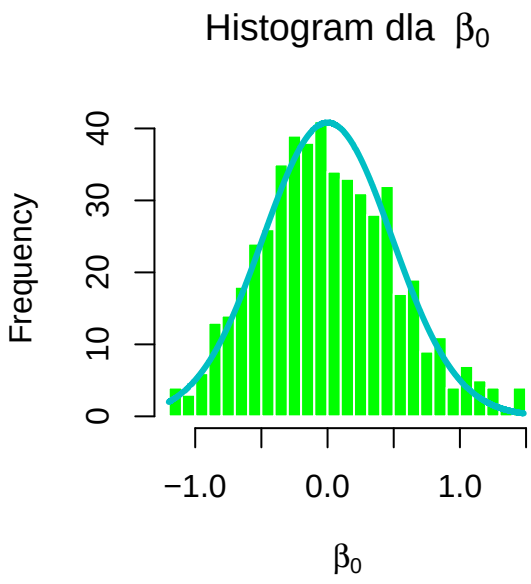
	k...30
γ	-0.00297
ρ	-0.00738

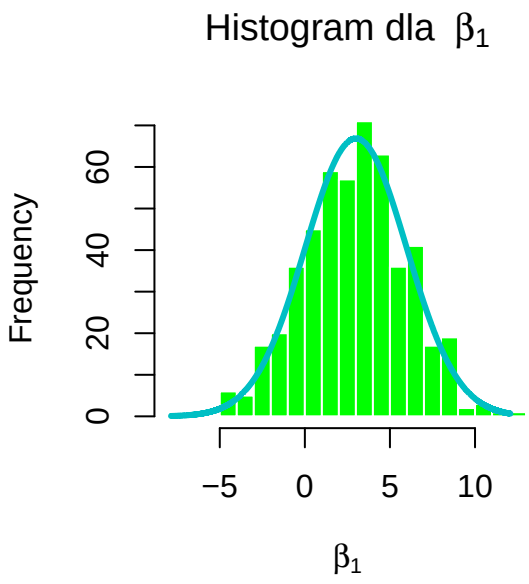
Obciążenie dla γ jest podobne, ale dla ρ wyraźne mniejsze.

Zadanie 5

Powtórzymy zadanie 2 dla przypadku, gdzie $p = 40$.

Histogramy dla β_0 oraz β_1:





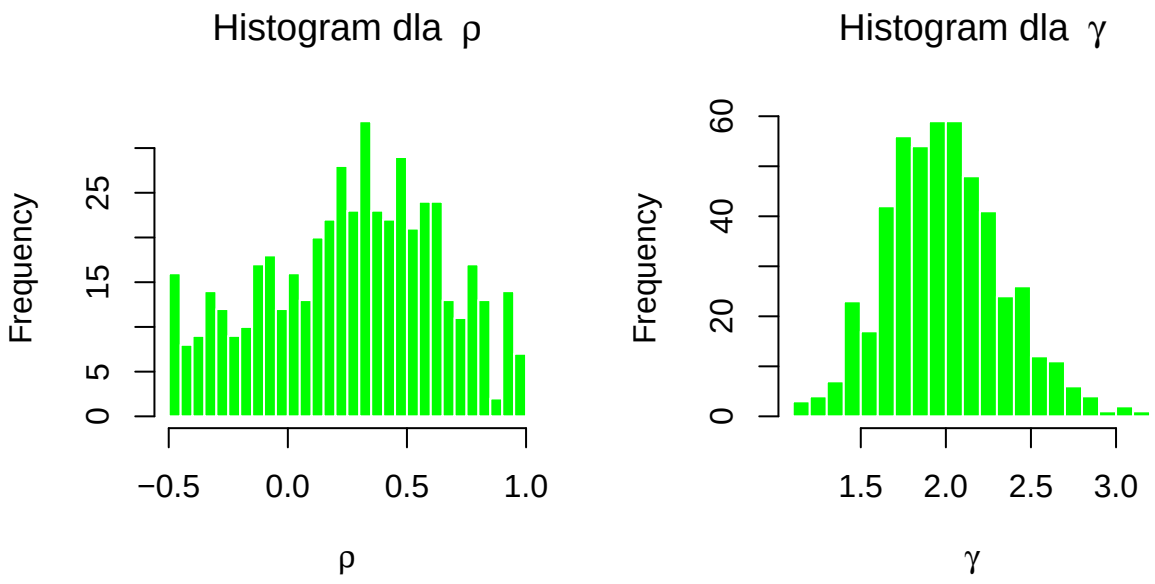
Widzimy, że histogramy są dużo szersze (antagonistycznie do przypadku z zadania 4).

Obciążenia dla β_0 oraz β_1:

	p...40
Supremum	0.249
Średnia	0.013

Norma supremum jest wyższa niż dla wyjściowych danych, natomiast średnie obciążenie jest lepsze.

Histogramy dla ρ oraz γ:



Widzimy wyraźne pogorszenie wykresów. Histogram dla ρ nie przypomina rozkładu normalnego.

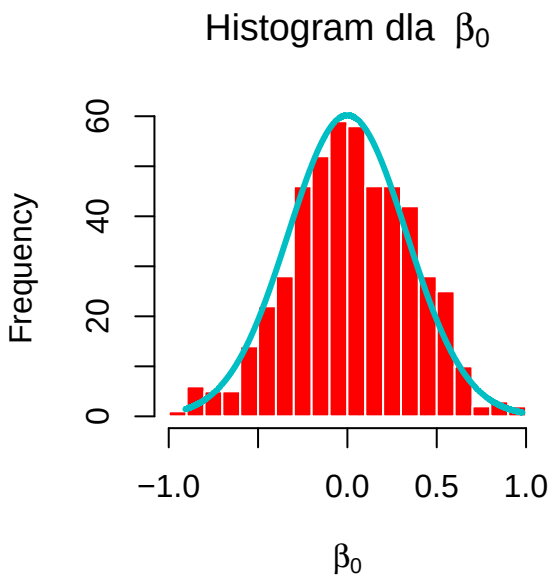
Obciążenia dla ρ i γ:

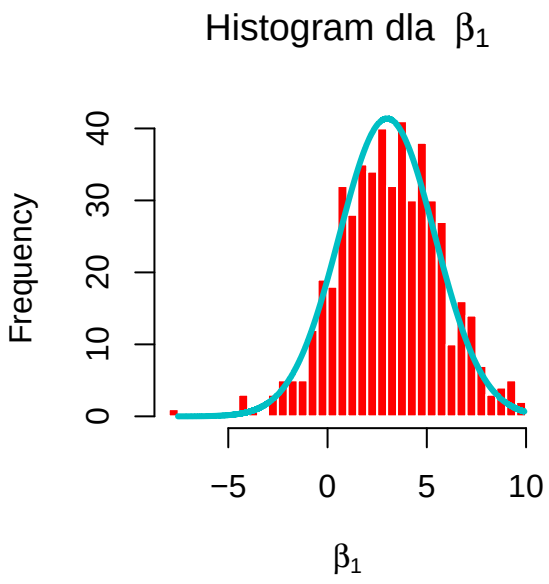
	p...40
γ	-0.00727
ρ	-0.03112

Obciążenie dla ρ wyraźne wzrosło, ale dla γ jest lepsze.

Zadanie 6

Powtórzymy teraz zadanie 2, ale dla modelu gdzie użyjemy metody “ML” zamiast “REML”.





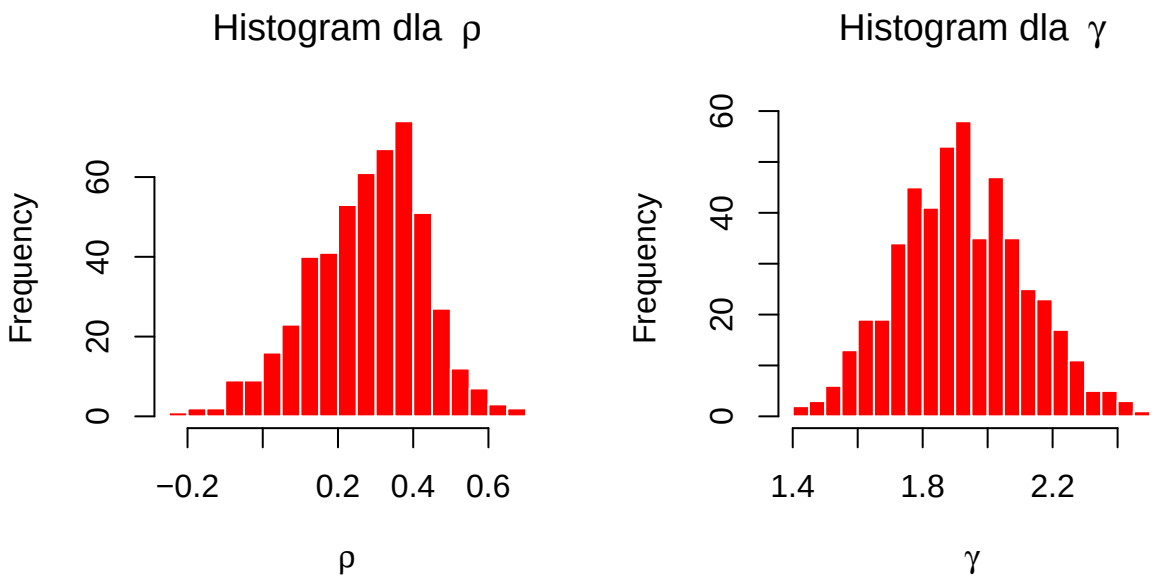
Histogramy dla parametrów β wyglądają podobnie do metody z zadania 2 ("REML").

Obciążenia dla β_0 oraz β_1:

	ML
Supremum	0.141
Średnia	-0.063

Widzimy, że obciążenie jest trochę lepsze dla metody "ML".

Histogramy dla ρ oraz γ:



Histogramy wyglądają podobnie jak dla tych z metody "ML".

Obciążenia dla ρ i γ:

	ML
γ	-0.08171
ρ	-0.02244

Parametr γ jest znacznie większy w przypadku metody "ML", natomiast parametr ρ jest podobny do tego z metody "REML".

Podsumowanie:

Przeprowadzona analiza z zadań 2-5 pokazuje jak poszczególne parametry n, k i p oddziałowują na estymatory β, γ i ρ. W zależności od tego, który parametr zwiększamy, zmieniają nam się histogramy dla estymatorów, podobnie to wygląda z ich obciążeniem. W zadaniu 6 zobaczyliśmy, że jeśli użyjemy metody "ML" zamiast "REML" w funkcji *gls()*, to widzimy, że zwiększa nam się obciążenie estymatora wariancji.