Zaawansowane Modele Liniowe - Lista 4

Pomiary wielokrotne; ogólny model liniowy

Wstęp

W uogólnionych modelach liniowych zakładaliśmy, że zmienne wynikowe $y_1, ...y_n$ nie są ze sobą w żaden sposób skorelowane. Takie założenie traci sens, gdy analizujemy dane w których różne obserwacje pochodzą od tych samych obiektów badanych w kilku punktach czasu. Do modelowania tego typu danych posłużymy się *ogólnym modelem liniowym*. Zakładamy, że zmienne wynikowe możemy zapisać w macierzy $n \times k$, gdzie $y_{i,j}$ oznacza wynik obserwacji i- tego obiektu w j-tym momencie. Z każdą obserwacją $y_{i,j}$ związany jest wektor zmiennych objaśniających $X^{(i,j)} \in R^{p-1}$ tak, że

$$y_{i,j} = \beta_0 + X_1^{(i,j)} \cdot \beta_1 + ... + X_{p-1}^{(i,j)} \cdot \beta_{p-1} + \epsilon_{i,j},$$

gdzie $\epsilon_{i,j} \sim N(0, \sigma_{i,j}^2)$. Ponadto zakładamy, że nie ma korelacji pomiędzy błędami losowymi stowarzyszonymi z różnymi obiektami ($i \neq k \implies cor(\epsilon_{i,j}, \epsilon_{k,l}) = 0$), ale istnieje pewna struktura korelacji pomiędzy kolejnymi pomiarami dla tego samego obiektu. Jeśli przez ϵ_i oznaczymy wektor błędów losowych związanych z wierszem y_i, to można zapisać że $\forall_{i \in \{1,...n\}} cov(\epsilon_i) = \Sigma$ (Σ jest macierzą symetryczną $k \times k$). Oznacza to, że macierze kowariancji pomiędzy pomiarami w różnych momentach są takie same dla wszystkich obiektów (uogólnienie założenia o stałości wariancji). Możemy estymować macierz Σ i wektor współczynników β oraz testować hipotezy dotyczące estymatorów. Estymator wektora parametrów β jest wektorem losowym o **asymptotycznym rozkładzie** normalnym o wartości oczekiwanej β i macierzy kowariancji

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

Estymator dany jest jawnym wzorem

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

W przypadku, gdy nie znamy prawdziwych parametrów możemy szacować macierz kowariancji $cov(\hat{\beta})$ podstawiając w miejsce macierzy Σ jej estymator, którego obliczenie może odbyć się za pomocą różnych narzędzi. Będziemy rozważać estymator ML (metoda Maximum Likelihood) oraz REML (Restricted Maximum Likelihood) wprowadzający pewną poprawkę do metody ML i wygrywający z nią w przypadku małej liczby obserwacji. Nie podajemy tutaj postaci estymatorów $\hat{\Sigma}$, skorzystamy z zaimplementowanych w R metod.

W ramach raportu dopasujemy modele do wygenerowanych symulacyjnie danych i zbadamy własności uzyskanych estymatorów porównując je z teorią. Do modelowania zastosujemy funkcję `gls` z biblioteki `nmle`.

Zadanie 1

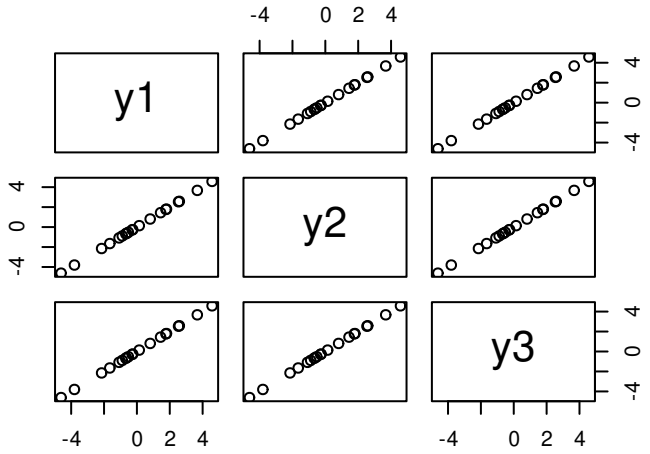
Generujemy dane zgodnie z założeniami ogólnego modelu liniowego z $n = 20, k = 3, p = 4, \beta = (3, 3, 0)^T$ i macierzą Σ taką, że wyrazy na przekątnej są równe i wynoszą $\gamma^2 = 4$ (są to wariancje obserwacji w poszczególnych tygodniach), a poza nią $\gamma^2 \rho = 4 \cdot 0.3 = 1.2$ (korelacje pomiędzy dowolnymi dwoma

momentami pomiaru są takie same). Dopasowując później model zadamy mu właśnie taką strukturę macierzy kowariancji podając odpowiednie argumenty do funkcji `gls`, tzn. `correlation = corCompSymm(form = ~1|id)` i `weights = varIdent(form = ~1)`).

Table 1: Pierwsze 6 wierszy danych (macierz 60 x 6)

y	id	t	X1	X2	X3
0.142	1	1	0.163	0.046	-0.015
0.967	1	2	-0.042	-0.001	0.026
-0.153	1	3	0.172	-0.121	-0.138
-0.258	2	1	0.164	-0.015	-0.104
-1.845	2	2	0.054	-0.105	-0.144
-2.406	2	3	-0.199	0.031	0.204

Wykres poniżej (`pairs`) przedstawia korelacje pomiędzy czasami 1, 2 i 3 (widać, że są one takie same dla wszystkich czasów):



Dopasowujemy model z Interceptem. W ramach zadania 1 porównujemy estymatory zwrócone przez model z wartościami obliczonymi na podstawie wzorów z wykładu ($\hat{\beta}, cov(\hat{\beta})$) oraz z prawdziwymi wartościami ($\hat{\gamma}, \hat{\rho}$).

Table 2: Zadanie 1, wyniki do 5 miejsca po przecinku

max(|beta_model - beta_wzory|)	0.00000
max(|cov_beta_model - cov_beta_wzory|)	0.00000
|gamma_true - gamma_model|	0.02200
|rho_true - rho_model|	0.18784

Komentarz:

- estymatory $\hat{\beta}$ i $cov(\hat{\beta})$ wyznaczone przez funkcję `gls` są zgodne z teorią z wykładu (drobne różnice mogą wynikać z ograniczeń dokładności komputerów).
- estymatory γ i ρ wyznaczone przez model są zbliżone do prawdziwych wartości. Więcej na ten temat można powiedzieć w kolejnych zadaniach, przy większej ilości powtórzeń eksperymentu.

Zadanie 2

Powtórzymy 500-krotnie eksperyment z zadania 1 w celu zbadania asymptotycznych własności estymatorów. Macierz kowariancji współczynników **wyliczona w oparciu o wartości** β, Σ **i macierz planu użyte do generowania danych** jest postaci

$$cov(\hat{\beta}) = \begin{bmatrix} 0.11 & -0.01 & -0.01 & -0.02 \\ -0.01 & 3.2 & -1.42 & 0.26 \\ -0.01 & -1.42 & 6.07 & -0.96 \\ -0.02 & 0.26 & -0.96 & 3.21 \end{bmatrix}$$

a odchylenia standardowe współczynników to odpowiednio: $0.33, 1.79, 2.46, 1.79$. Można zauważyć, że $\hat{\beta}_2$ ma znacząco większą wariancję niż pozostałe estymatory, jednak pewnie jest to kwestia tego jaka postać macierzy X została wylosowana do generowania danych.

Estymatory $\hat{\beta}$

Wykresy poniżej przedstawiają histogramy gęstości estymatorów $\hat{\beta}_0$ i $\hat{\beta}_1$ uzyskanych w 500 replikacjach eksperymentu wraz z dorysowanymi **teoretycznymi** gęstościami obliczonymi na podstawie wzorów z wykładu z użyciem prawdziwej macierzy Σ. (Wykresy z zadań 2-6 zostały dodatkowo zebrane w jednym miejscu na końcu dokumentu.)

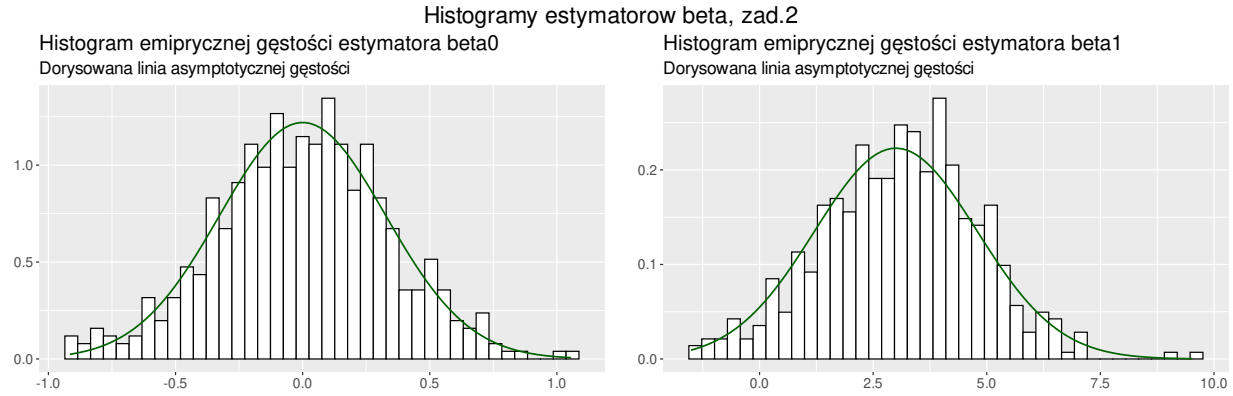


Tabela poniżej zawiera wyestymowane obciążenia i normy supremum błędu oszacowania współczynników, gdzie przez błąd oszacowania w i-tej iteracji rozumiemy różnicę $\hat{\beta}_i - \beta_i$, a obciążenie obliczamy jako wartość oczekiwaną powyższej różnicy, uśredniając wyniki.

Table 3: Obciążenia i norma supremum błędu estymatorów, zad. 2

	beta0	beta1	beta2	beta3
obciazenie	0.000	0.096	0.075	0.096
norma_supremum_bledu	1.056	6.521	7.473	5.960

Komentarz:

- Średnie wartości $\hat{\beta}_0, \hat{\beta}_1$ wynoszą odpowiednio około 0 i 3, a odchylenia standardowe $0.34, 1.72$. Wartości te są blisko teoretycznych, a wygląd wykresów potwierdza zgodność asymptotycznych rozkładów z teoretycznymi.
- Obciążenia dla estymatorów są bardzo małe, przy czym można zauważyć że dla 500 replikacji różnica pomiędzy prawdziwą wartością a estymatorem jest wciąż widoczna dla $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, podczas gdy dla $\hat{\beta}_0$ jest ona już bliska 0. Ma to intuicyjnie sens, gdyż odchylenie standardowe $\hat{\beta}_0$ jest znacznie mniejsze

niż dla pozostałych współczynników i można spodziewać się, że średnia szybciej stabilizuje się wokół prawdziwej wartości. W celu potwierdzenia tego wykonałam dodatkowo eksperyment z liczbą powtórzeń równą 3000.

Table 4: Obciążenia i norma supremum błędu estymatorów, zad. 2
Dodatkowe wyniki dla 3000 powtórzeń.

	beta0	beta1	beta2	beta3
obciazenie	-0.006	-0.033	-0.016	-0.008
norma_supremum_bledu	1.122	6.989	6.510	8.200

Widać, że obciążenia innych estymatorów zmalały. Można spodziewać się, że wraz ze wzrostem liczby powtórzeń będą one coraz bliższe 0. Dodatkowo można zaobserwować, że zwiększyły się normy supremum błędu, co wynika z faktu że przy większej ilości prób zwiększamy szanse na wylosowanie bardzo skrajnej wartości. Podobnie, współczynnik o największym błędzie i obciążeniu to $\hat{\beta}_2$, czyli ten o największej wariancji.

Estymatory $\hat{\gamma}, \hat{\rho}$

W tej części zbadamy parametry wpływające na postać macierzy Σ.

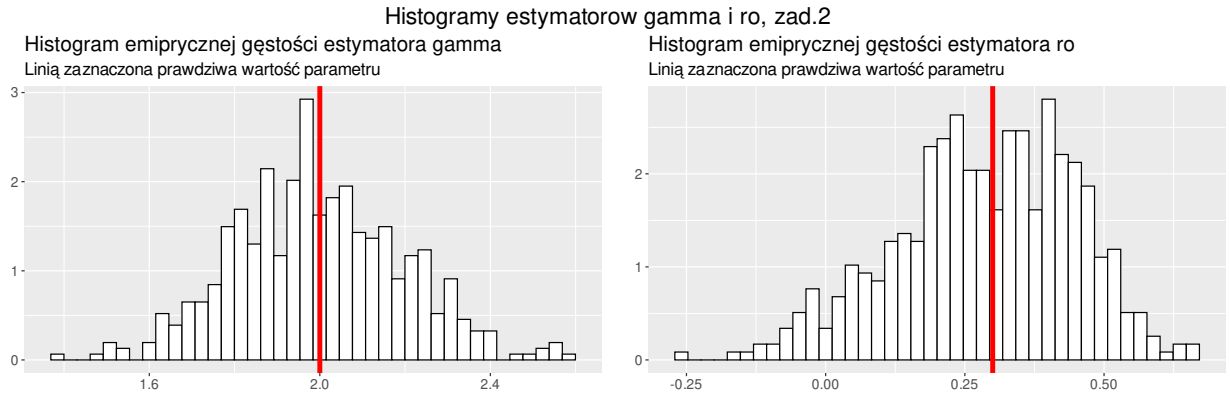
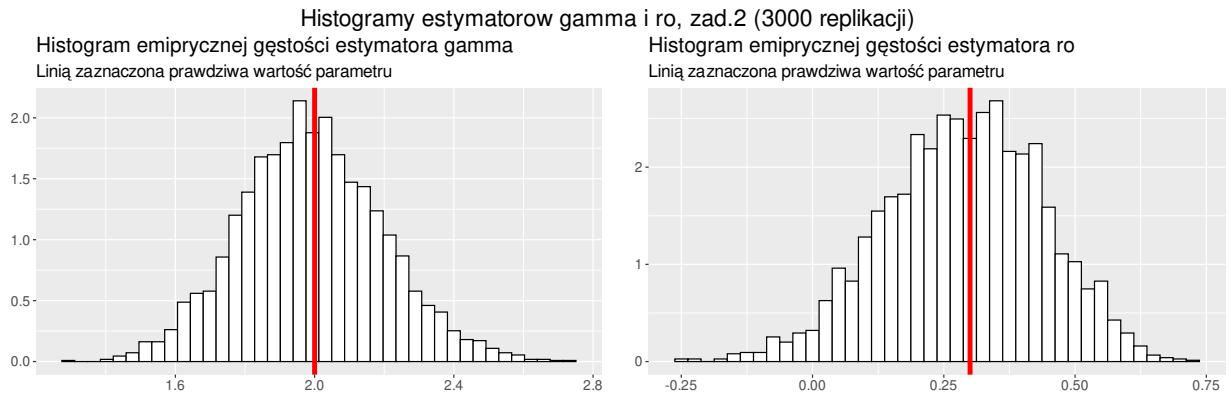


Table 5: Obciążenia i norma supremum estymatorów, zad.2

	gamma	ro
obciazenie	-0.003	-0.015
norma_supremum_bledu	0.615	0.549

Komenatrz:

- Uzyskane obciążenia dla obu estymatorów są bardzo małe, tak samo jak normy supremum błędu. Oznacza to, że model dobrze szacuje te parametry.
- Wykresy w przybliżeniu przypominają rozkład normalny (lekko skośny dla estymatora $\hat{\rho}$). Możemy wykorzystać wygenerowane wyniki dla 3000 replikacji w celu dokładniejszego ocenienia rozkładów.



Widać, że wykresy bardziej przypominają rozkład normalny skupiony wokół prawdziwych wartości parametrów. Pozwala to przypuszczać, że ich rozkłady są asymptotycznie normalne. Sprawdziłam również, jak zmienia się **odchylenie standardowe** estymatorów. W przypadku 500 replikacji dla $\hat{\gamma}, \hat{\rho}$ jest to odpowiednio: $0.2, 0.16$, a w przypadku 3000 replikacji: $0.21, 0.15$, czyli można spodziewać się że wartości te stabilizują się wokół około 0.21 i 0.15.

Zadanie 3 (wzrasta liczba obserwacji, n = 500)

Macierz kowariancji współczynników wyliczona w oparciu o wartości β, Σ i macierz planu użyte do generowania danych w tym zadaniu jest postaci

$$cov(\hat{\beta}) = \begin{bmatrix} 0 & 0.01 & 0 & 0 \\ 0.01 & 3.53 & 0.01 & -0.04 \\ 0 & 0.01 & 3.37 & 0.03 \\ 0 & -0.04 & 0.03 & 3.34 \end{bmatrix}$$

a odchylenia standardowe współczynników to odpowiednio: $0.07, 1.88, 1.84, 1.83$. Są one bardziej zbliżone niż w zadaniu 2 - większa ilość danych zmniejszyła wpływ losowości w generowaniu macierzy planu.

Estymatory $\hat{\beta}$

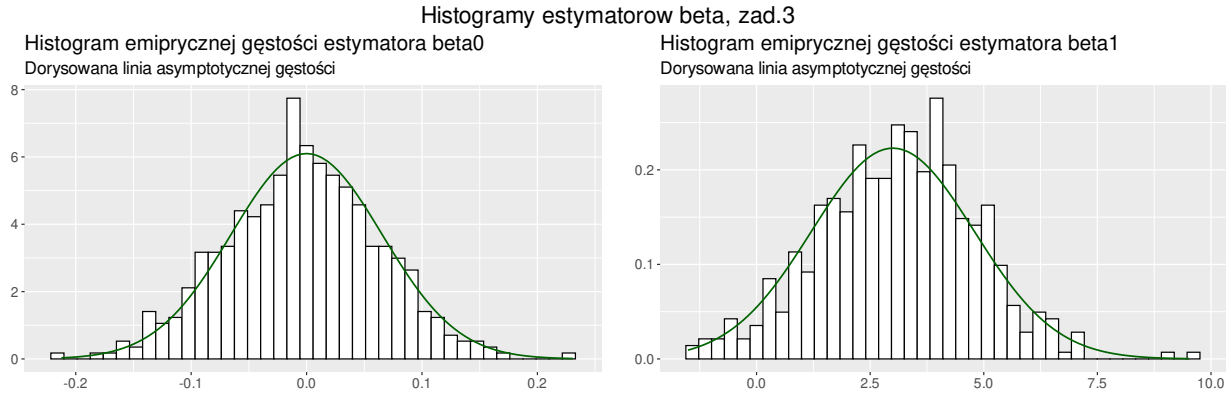


Table 6: Obciążenia i norma supremum błędu estymatorów, zad. 3

	beta0	beta1	beta2	beta3
obciazenie	-0.004	-0.048	-0.153	-0.016

	beta0	beta1	beta2	beta3
norma_supremum_bledu	0.231	5.845	6.946	6.513

Komenatarz:

- Wygląd wykresów jest na pierwszy rzut oka podobny, jednak zmieniła się skala na osi OX (można łatwo zauważyć to na porównaniu wykresów na końcu raportu.) Wzrost liczby obserwacji zmniejsza odchylenie standardowe estymatorów.
- Obciążenia ponownie są małe, w tym prawie zerowe dla $\hat{\beta}_0$.
- W porównaniu do zadania 2 obciążenia estymatorów $\hat{\beta}_1, \hat{\beta}_3$ zmalały, ale obciążenie estymatora $\hat{\beta}_2$ wzrosło. Można podejrzewać, że wzrost obciążenia wynika z losowości. W celu sprawdzenia tego oraz porównania odchyleń standardowych generujemy wyniki dla większej (3000) liczby powtórzeń.

Table 7: Obciążenia i norma supremum błędu estymatorów, zad. 3
Dodatkowe wyniki dla 3000 powtórzeń.

	beta0	beta1	beta2	beta3
obciazenie	0.000	-0.021	0.030	0.017
norma_supremum_bledu	0.243	7.263	7.284	7.409

Okazuje się, że nie widać znaczących różnic w obciążeniach estymatorów uzyskanych dla małej i dużej próby.

Estymatory $\hat{\gamma}, \hat{\rho}$

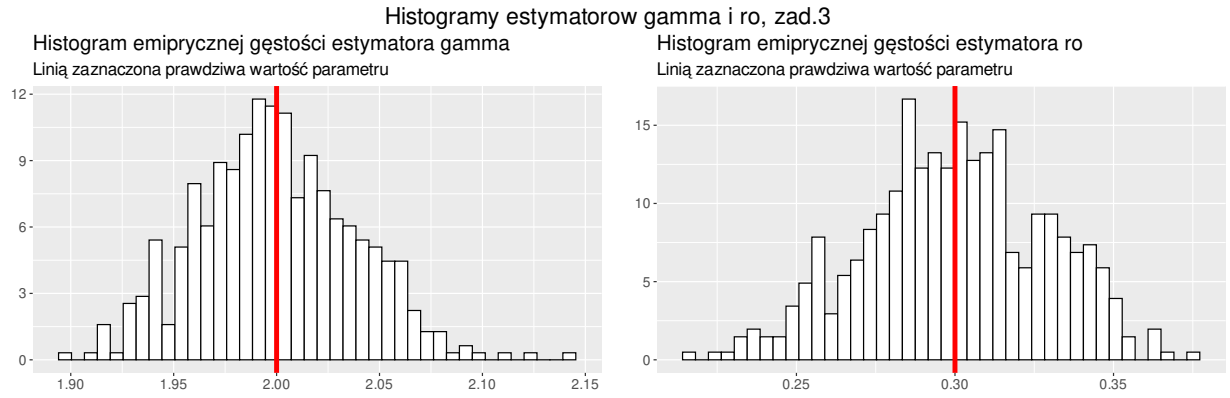
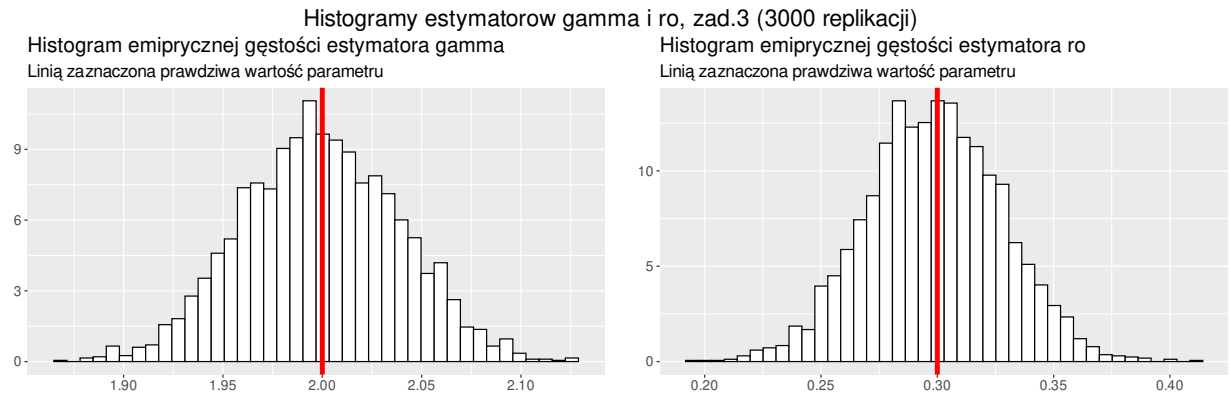


Table 8: Obciążenia i norma supremum estymatorów, zad.3

	gamma	ro
obciazenie	0.000	0.000
norma_supremum_bledu	0.141	0.086

Komentarz: Liczba obserwacji wydaje się mieć wpływ na estymację Σ - histogramy już dla 500 replikacji wyglądają bardziej normalnie dla $n = 500$ niż dla $n = 20$. Widać też znaczący spadek odchyleń standardowych estymatorów. Poniżej dodatkowe wykresy dla 3000 replikacji:



Wykresy dla $n = 500$ zdają się być bardziej skupione wokół średnich - obserwacje tę potwierdza zbadanie odchyleń standardowych: w przypadku 500 replikacji dla $\hat{\gamma}, \hat{\rho}$ jest to odpowiednio: $0.2, 0.16$, a w przypadku 3000 replikacji: $0.04, 0.03$. Odchylenia dla $n = 500$ stabilizują się na niższym poziomie niż dla małej próby.

Zadanie 4 (wzrasta liczba pomiarów, k = 30)

Macierz kowariancji współczynników wyliczona w oparciu o wartości β, Σ i macierz planu użyte do generowania danych w tym zadaniu jest postaci

$$cov(\hat{\beta}) = \begin{bmatrix} 0.06 & 0.01 & 0 & -0.01 \\ 0.01 & 3.04 & -0.07 & 0.19 \\ 0 & -0.07 & 2.89 & 0.16 \\ -0.01 & 0.19 & 0.16 & 2.85 \end{bmatrix}$$

a odchylenia standardowe współczynników to odpowiednio: $0.25, 1.74, 1.7, 1.69$.

Estymatory $\hat{\beta}$

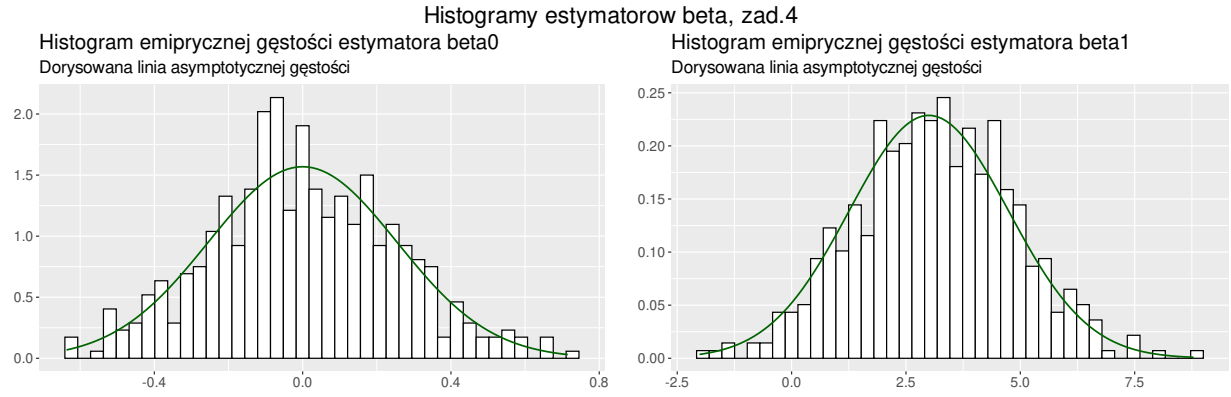


Table 9: Obciążenia i norma supremum błędu estymatorów, zad. 4

	beta0	beta1	beta2	beta3
obciazenie	0.000	0.120	-0.125	0.052
norma_supremum_bledu	0.715	5.787	4.867	4.793

Wykresy ponownie wyglądają jak rozkład normalny, a obciążenia są małe. Porównałam ze sobą uzyskane w zadaniach 2-4 wykresy, tabele i wartości odchyleń standardowych z następującymi wnioskami:

- choć wykresy wydają się bardziej chaotyczne niż w zadaniu 2, odchylenie standardowe estymatorów zmalało, podobnie jak obciążenie.
- zwiększenie liczby pomiarów do 30 również wpłynęło na zmniejszenie odchylenia standardowego estymatorów β w mniejszym stopniu niż zwiększenie liczby obserwacji do 500. Może to wynikać z faktu, że zmiana była mniejsza.

Estymatory $\hat{\gamma}, \hat{\rho}$

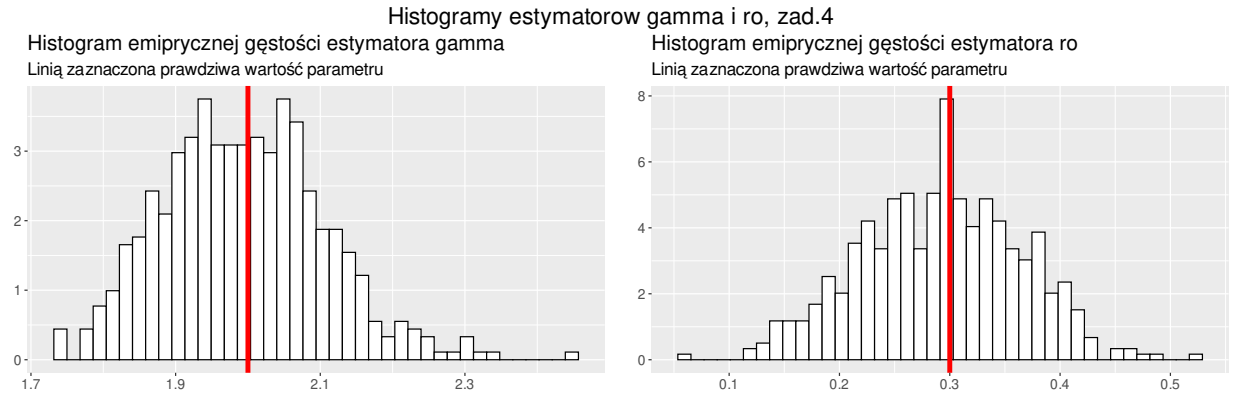


Table 10: Obciążenia i norma supremum estymatorów, zad.4

	gamma	ro
obciazenie	-0.006	-0.009
norma_supremum_bledu	0.442	0.238

Można zauważyć spadek wariancji estymatorów.

Zadanie 5 (wzrasta liczba kolumn macierzy planu)

W tym zadaniu zwiększamy liczbę kolumn macierzy planu (liczba prawdziwych istotnych zmiennych nie zmienia się).

Macierz kowariancji $cov(\hat{\beta})$ jest w tym przypadku bardzo duża, więc nie została zawarta w raporcie. Odchylenia standardowe pierwszych 4 współczynników to odpowiednio: $0.61, 2.9, 4.25, 3.49$.

Estymatory $\hat{\beta}$

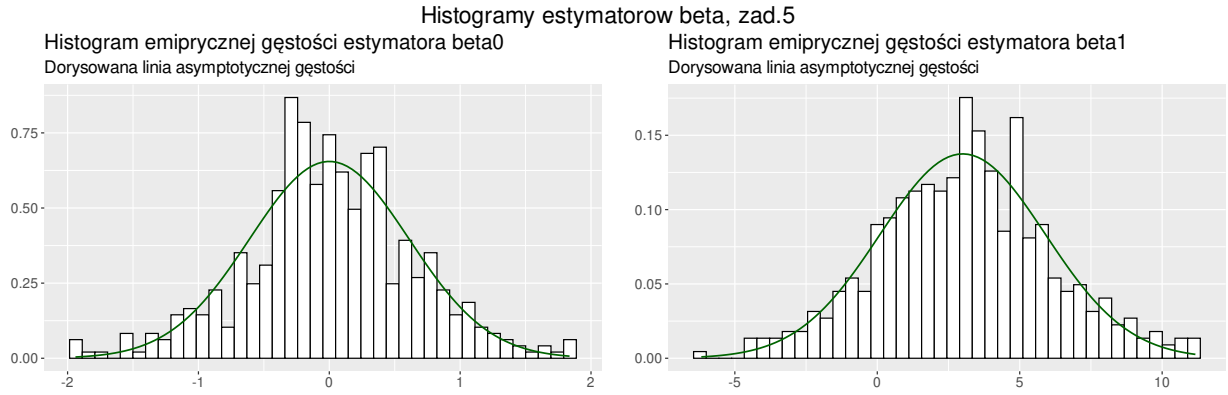


Table 11: Obciążenia i norma supremum błędu estymatorów, zad. 5 Pierwsze 4 współczynniki

	beta0	beta1	beta2	beta3
obciazenie	0.006	0.066	-0.074	0.138
norma_supremum_bledu	1.940	9.177	13.508	16.209

Komentarz: Norma supremum błędu jest wyższa niż w poprzednich zadaniach. Widać również wzrost wariancji estymatorów $\hat{\beta}$ (na wykresach na końcu raportu widać, że takie zjawisko występuje tylko w tym zadaniu).

Estymatory $\hat{\gamma}, \hat{\rho}$

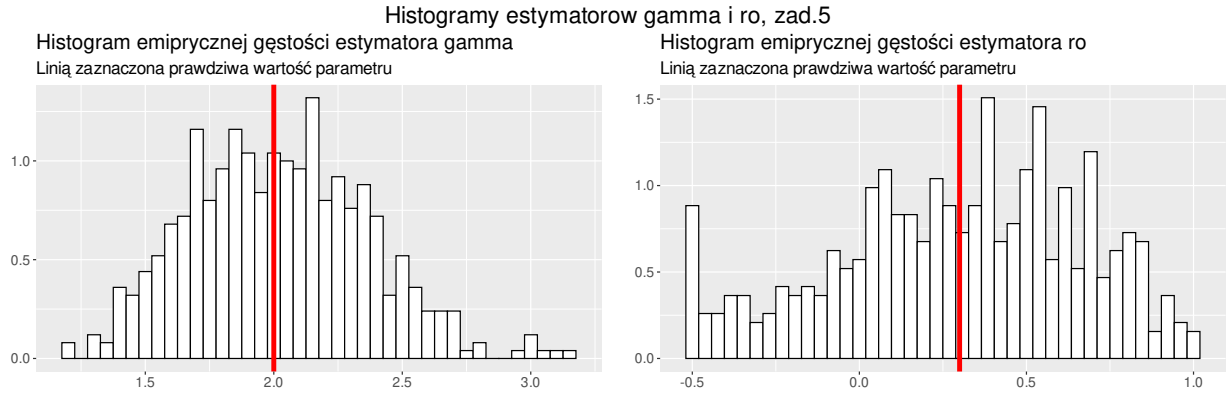


Table 12: Obciążenia i norma supremum estymatorów, zad.5

	gamma	ro
obciazenie	0.020	-0.01
norma_supremum_bledu	1.129	0.80

Komentarz: Tutaj również dodanie kolumn w macierzy planu wpływa negatywnie na estymację - zwiększyła się względem poprzednich zadań norma supremum błędu i wariancja estymatorów.

Zadanie 6 (zamiana metody REML na ML)

Macierz kowariancji współczynników wyliczona w oparciu o wartości β, Σ i macierz planu użyte do generowania danych w tym zadaniu jest postaci

$$cov(\hat{\beta}) = \begin{bmatrix} 0.11 & 0.02 & 0.06 & -0.04 \\ 0.02 & 3.32 & -0.31 & 0.53 \\ 0.06 & -0.31 & 4.17 & -0.43 \\ -0.04 & 0.53 & -0.43 & 3.18 \end{bmatrix}$$

a odchylenia standardowe współczynników to odpowiednio: $0.33, 1.82, 2.04, 1.78$.

Estymatory $\hat{\beta}$

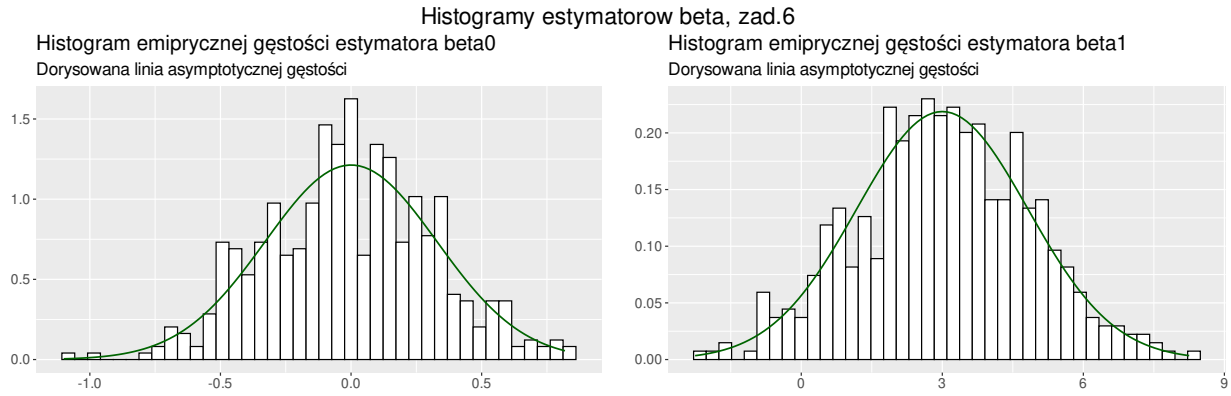


Table 13: Obciążenia i norma supremum błędu estymatorów, zad. 6

	beta0	beta1	beta2	beta3
obciazenie	-0.008	0.025	0.045	-0.063
norma_supremum_bledu	1.100	5.258	6.622	6.092

Nie widać znaczących różnic względem zadania 2.

Estymatory $\hat{\gamma}, \hat{\rho}$

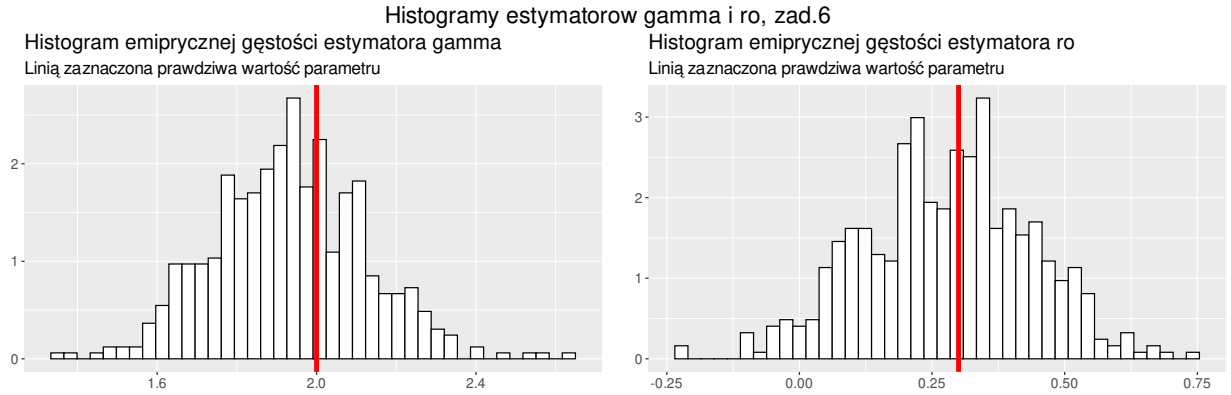


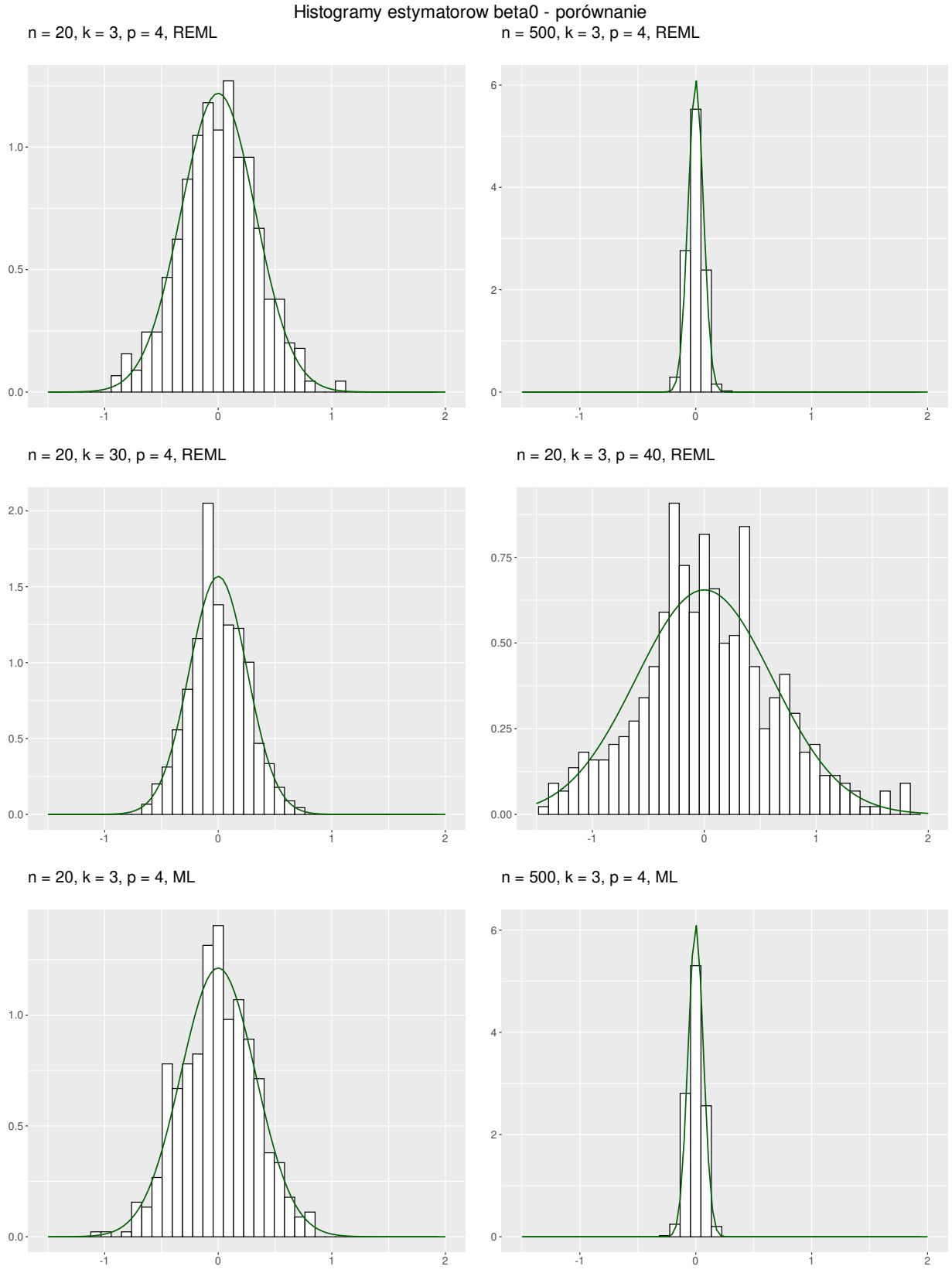
Table 14: Obciążenia i norma supremum estymatorów, zad.6

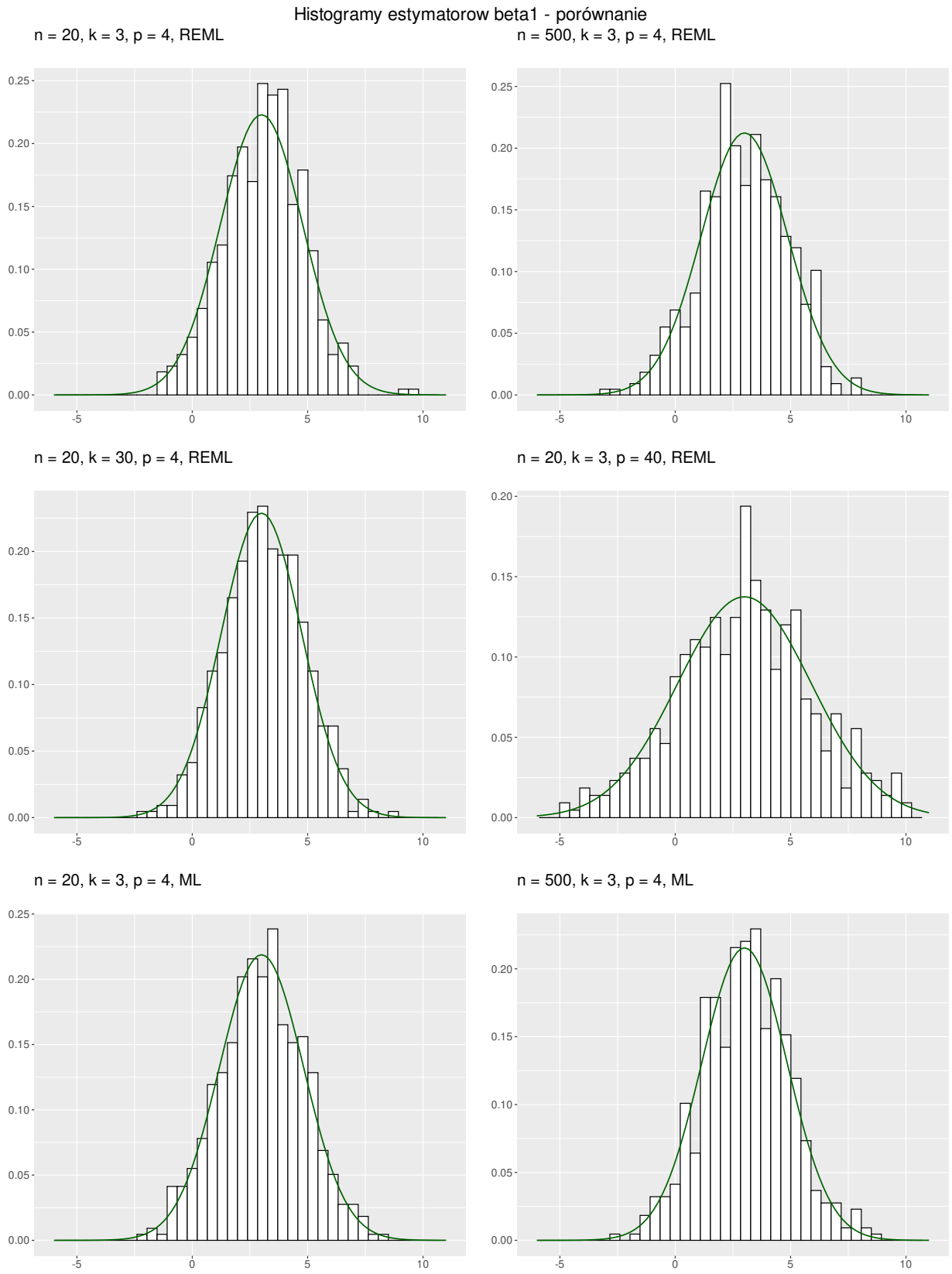
	gamma	ro
obciazenie	-0.069	-0.022
norma_supremum_bledu	0.666	0.520

Histogram γ jest lekko przesunięty w lewo - potwierdza to obecny w notatkach z wykładu fakt, że dla małej próby estymator ML jest obciążony w stronę 0. Wykonałam dodatkowe symulacje z metodą ML i $n = 500$, wyniki widoczne są na wykresach poniżej (dla $n = 500$ różnica pomiędzy metodami REML i ML na oko zaniknęła).

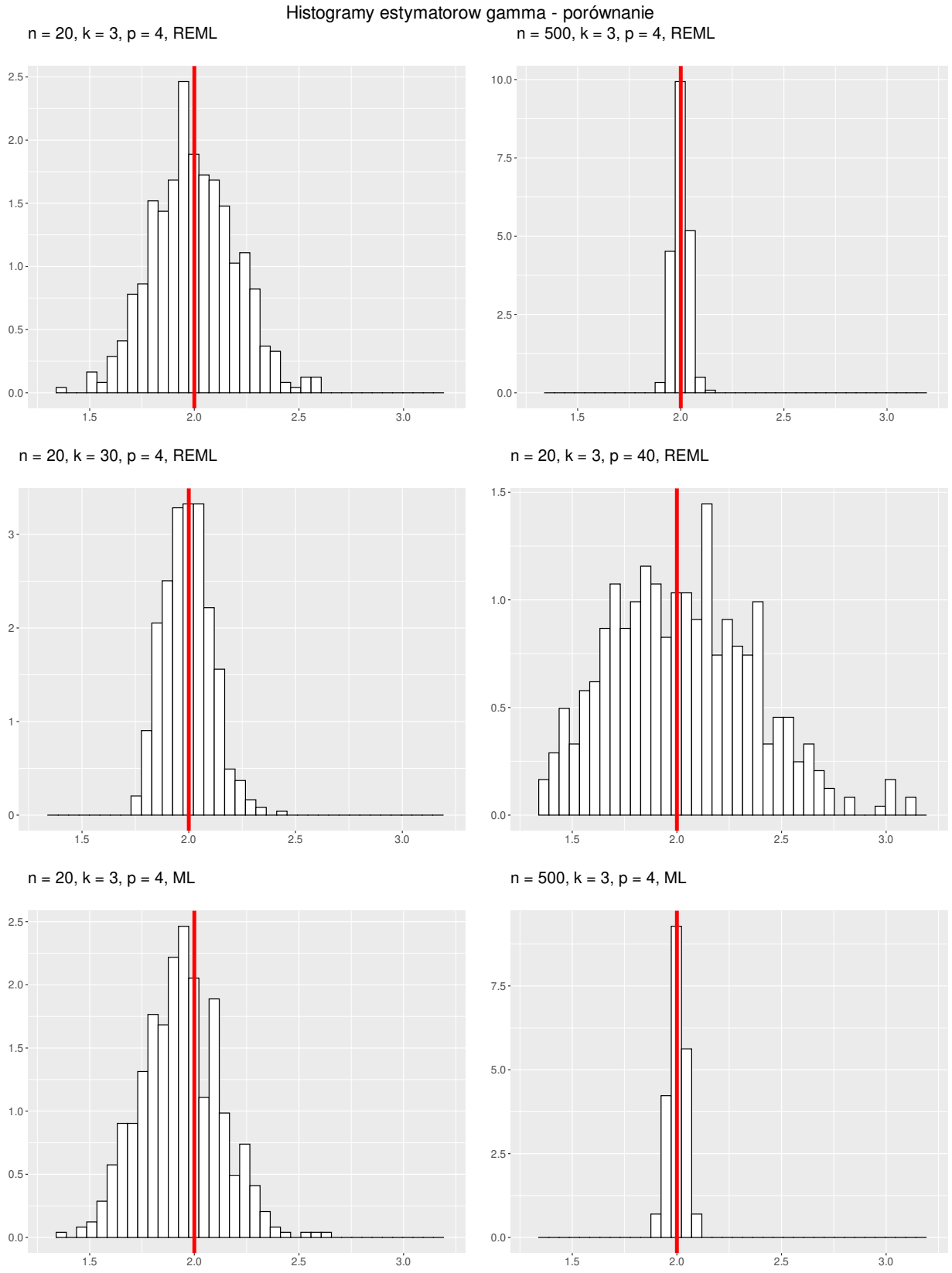
Porównanie wyników zadań 2-6 dla 500 replikacji

Estymatory $\hat{\beta}$



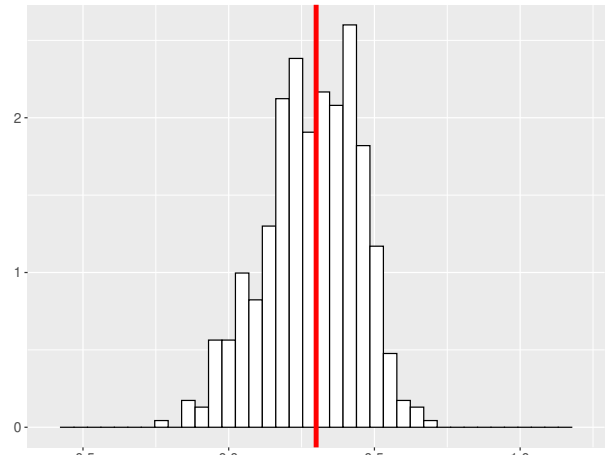

Histogramy estymatorow beta1 - porównanie
n = 20, k = 3, p = 4, REML
n = 500, k = 3, p = 4, REML
n = 20, k = 30, p = 4, REML
n = 20, k = 3, p = 40, REML
n = 20, k = 3, p = 4, ML
n = 500, k = 3, p = 4, ML

Histogramy γ, ρ

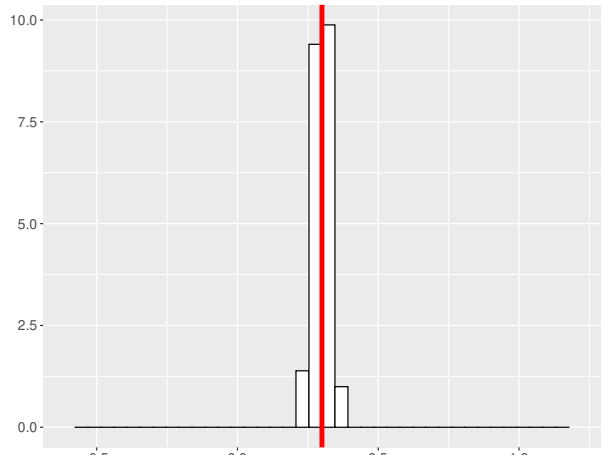


Histogramy estymatorow ro - porównanie

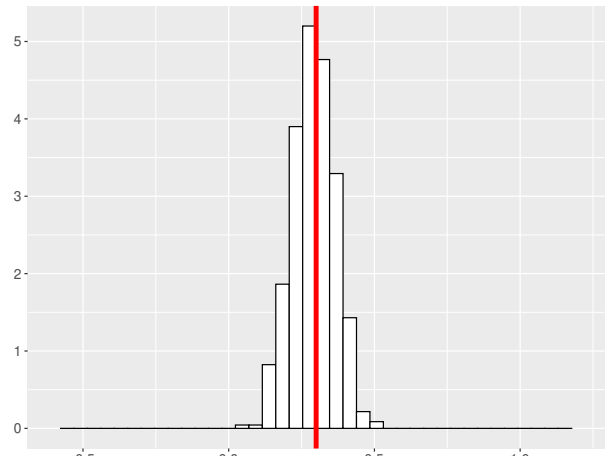
n = 20, k = 3, p = 4, REML



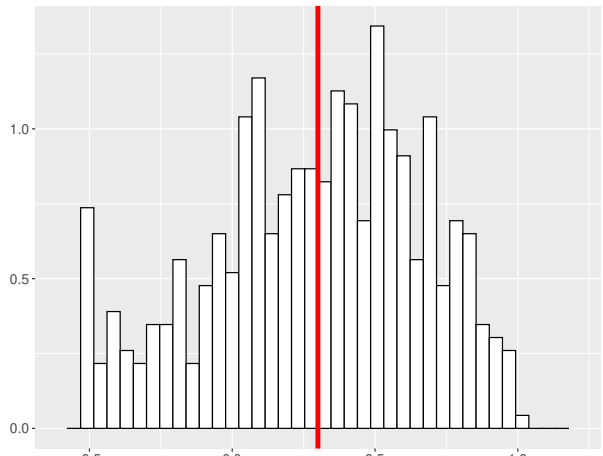
n = 500, k = 3, p = 4, REML



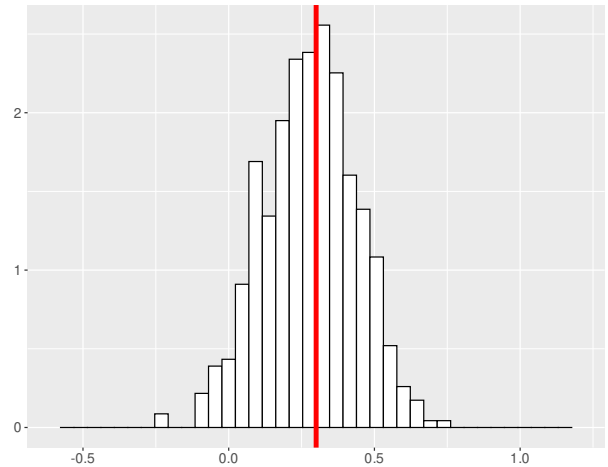
n = 20, k = 30, p = 4, REML



n = 20, k = 3, p = 40, REML



n = 20, k = 3, p = 4, ML



n = 500, k = 3, p = 4, ML