Zaawansowane modele liniowe raport nr 4

Dominik Mika

13 października 2021

Zadanie 1

Na początku wprowadźmy następujące oznaczenia:

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

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

Najpierw wygenerujemy macierzy $X \in \mathbb{M}_{N \times p-1}$ tak, że jej elementy są niezależnymi realizacjami z rozkładu $N(0, \sigma = 1/\sqrt{N})$. Podzielimy macierz X na $n = N/k$ podmacierzy $X_1, \ldots, X_n \in \mathbb{M}_{k \times p-1}$ w taki sposób że X_1 zawiera pierwszych k wierszy z X; X_2 zawiera kolejnych k wierszy z X; ... ; X_n zawiera ostatnich k wierszy z X.

Niech $\beta = (3, 3, 0)' \in \mathbb{R}^{p-1}$ oraz

$$\Sigma = \gamma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \in \mathbb{M}_{k \times k},$$

gdzie $\rho = 0.3$ i $\gamma = 2$.

Następnie wygenerujemy n niezależnych wektorów losowych:

$$y_i = (y_{i1}, \cdots, y_{ik})' \sim N(X_i\beta, \Sigma) \in \mathbb{R}^k, \quad i = 1, 2, \cdots, n.$$

Na koniec zapiszemy nasze wygenerowane dane w ramce danych w jednowymiarowej reprezentacji w celu zbudowania modelu. Model stworzymy w założeniu symetryczności macierzy kowariancji oraz stałej wariancji(czyli tak jak generowaliśmy nasze dane).

Z tak stworzonym modelem będziemy chcieli porównać wyniki teoretyczne z ich estymatorami oraz wyniki, które R zwraca nam automatycznie oraz obliczone wzorami.

Najpierw porównamy własności wektora beta. Na podstawie skonstruowanego modelu obliczymy estymator $\hat{\Sigma}$ i przy jego pomocy obliczymy estymator $\hat{\beta}$, który zadaje się wzorem:

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

oraz jego macierz kowariancji zgodnie z następującą formułą:

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

W celu porównania estymatorów obliczymy normę supremum ich różnicy.

Na koniec porównamy parametry ρ oraz γ z ich estymatorami, czyli wyznaczymy normę supremum różnicy.

Wyniki przedstawiają się w tabeli poniżej:

Parametr	β	$\mathrm{Cov}(\beta)$	ρ	γ
norma	$2.66 \cdot 10^{-15}$	$1.78 \cdot 10^{-15}$	0.218	0.22

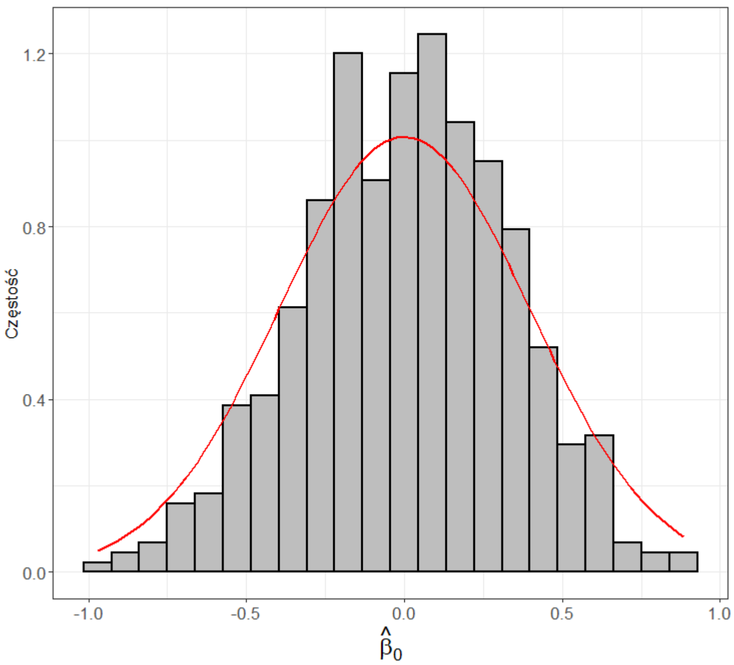
Widzimy, że norma dla β oraz $\mathrm{Cov}(\beta)$ jest bardzo mała, czyli estymacja uzyskana z modelu obliczonego prze R zgadza się ze wzorami podanymi wyżej. Natomiast patrząc na normę różnicy między parametrami macierzy Σ, a ich estymatorami widzimy, że nie są one duże jak na pojedyńcze powtórzenie. Możemy powiedzieć, że estymator jest w miarę blisko faktycznego parametru.

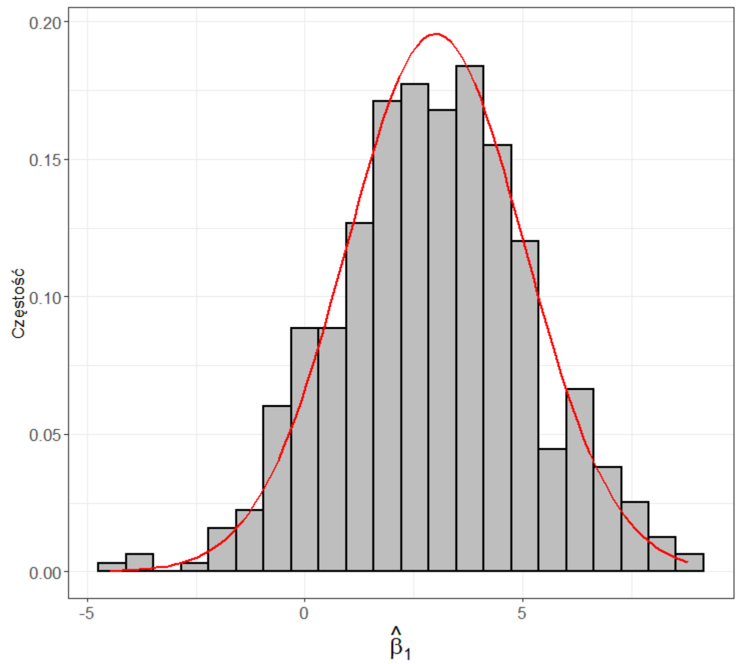
Zadanie 2

W tym zadaniu wykonamy jednokrotnie podpunkt a) z poprzedniego zadania. Następnie dla tak uzyskanych danych wykonamy 500 razy podpunkt b) z poprzedniego zadania. Na podstawie wszystkich obliczonych modeli, wyznaczymy ciąg 500-ciuset estymatorów β, ρ oraz γ.

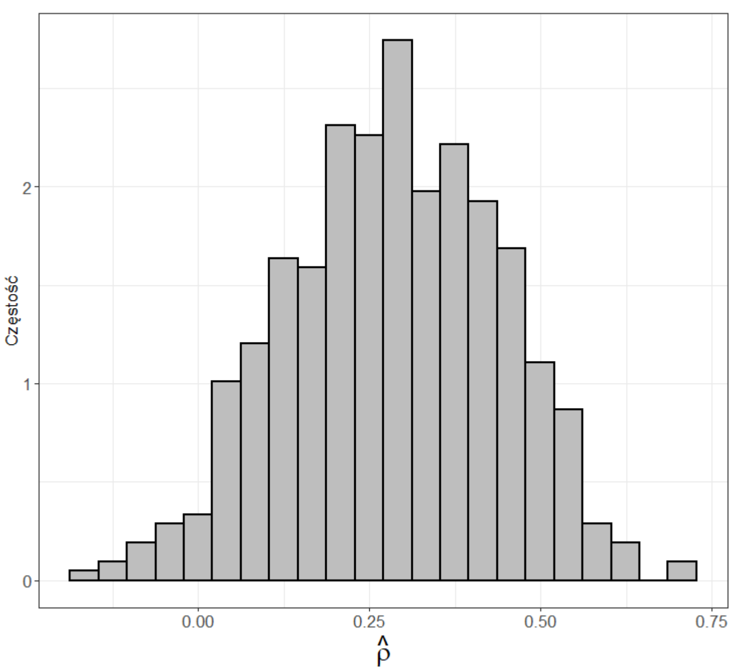
Następnie na ich podstawie narysujemy histogramy dla $\hat{\beta}_0, \hat{\beta}_1, \hat{\rho}$ oraz $\hat{\gamma}$. Do histogramów bet dorysujemy krzywą gęstości oraz obliczymy normę supremum średniego obciążenia tego estymatora. Natomiast dla estymatorów ρ oraz γ wyznaczymy średnie obciążenie.

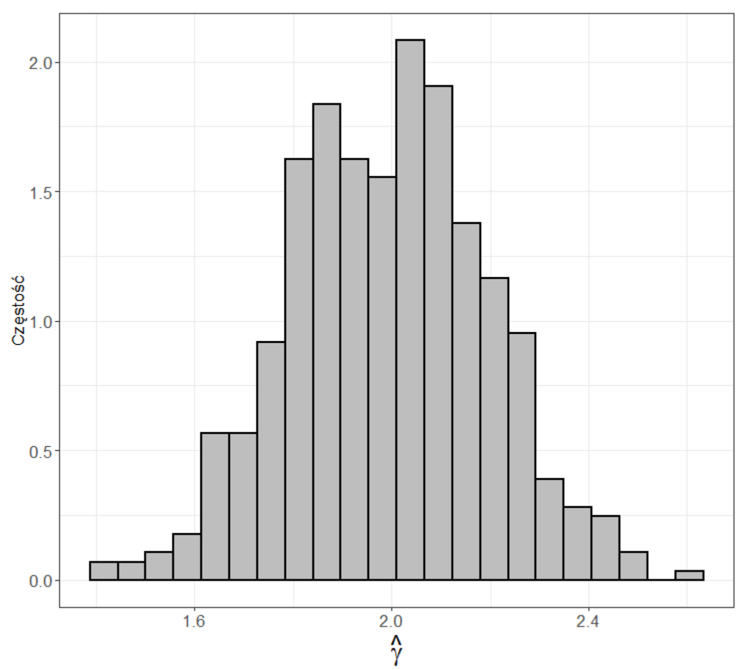
Poniżej znajdują się histogramy odpowiednich estymatorów:





Pierwsze co możemy zauważyć to, to że wartości estymatory faktycznie często są bliskie prawdziwej wartości. W przypadku $\hat{\beta}_0$- 0, a $\hat{\beta}_1$- 3. Natomiast histogram w przypadku $\hat{\beta}_0$ jest dość słabo dopasowany do krzywej gęstości rozkładu normalnego. Z drugiej strony histogram dla $\hat{\beta}_1$ jest dobrze dopasowany do tej krzywej.





Na histogramach dla estymatorów ρ i γ możemy zauważyć, że estymatory najczęściej przyjmują wartość bliską tej faktycznej.

Poniżej znajduje się tabelka z wyznaczonymi średnimi obciążeniami estymatorów ρ oraz γ, a dla β jest to norma supremum ze średniego obciążęnia:

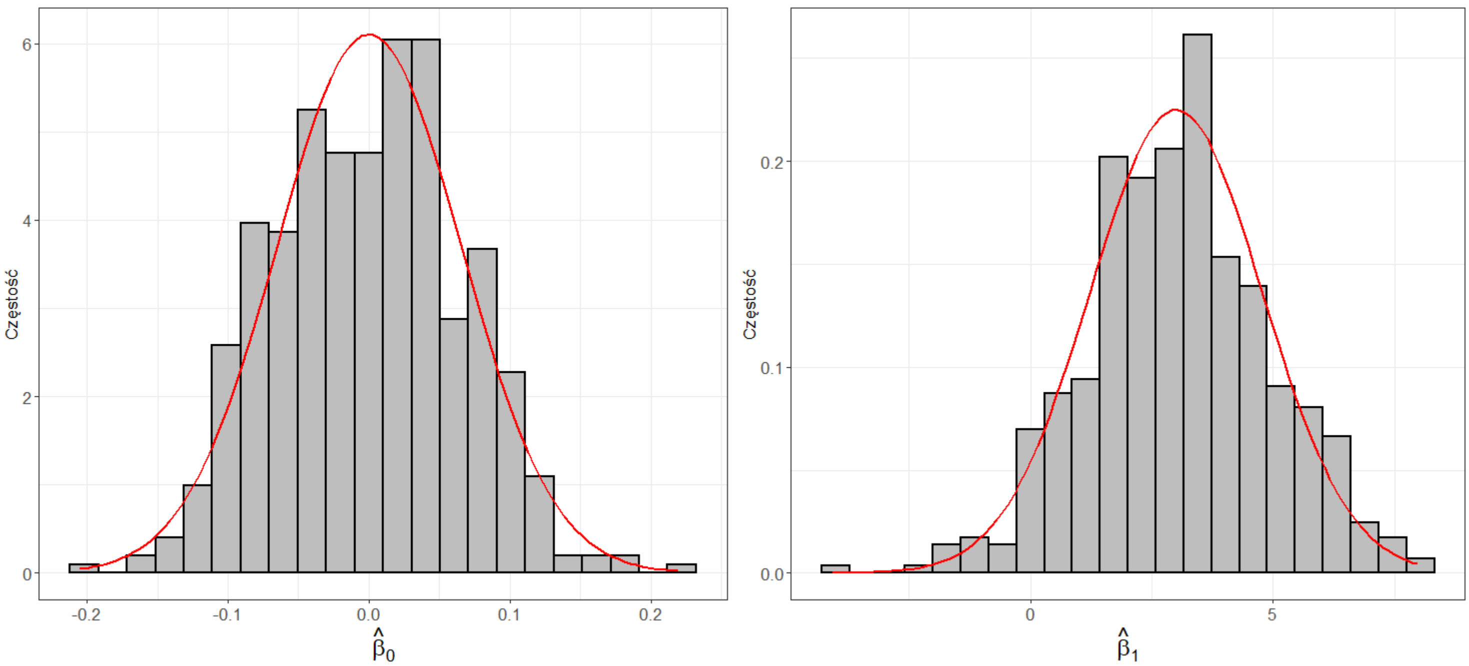
Parametr	β	ρ	γ
Wartość	0.048	-0.015	-0.006

Widzimy, że norma dla wektora β jest dość mała. Natomiast w porównaniu z poprzednim zadaniem estymatory ρ i γ wydają się być dużo lepsze. Średnie obciążenie w ich przypadku jest bardzo małe.

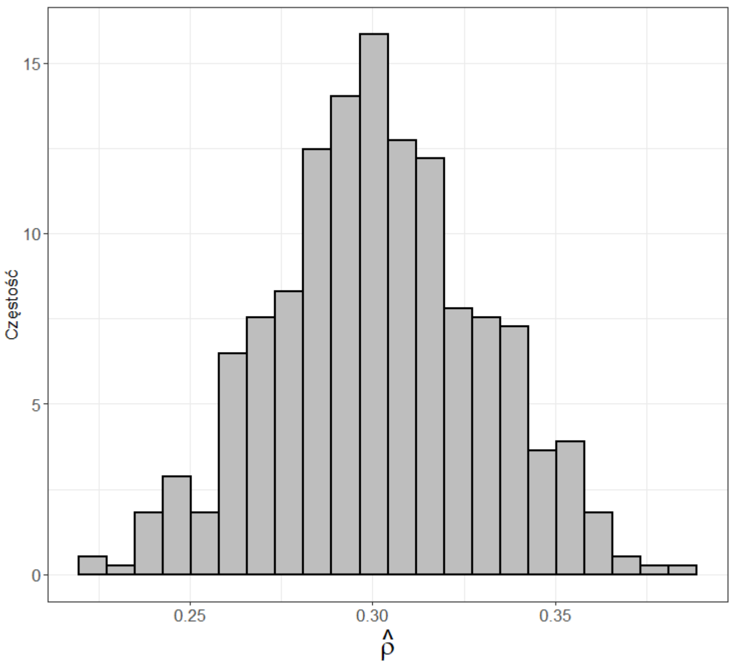
Zadanie 3

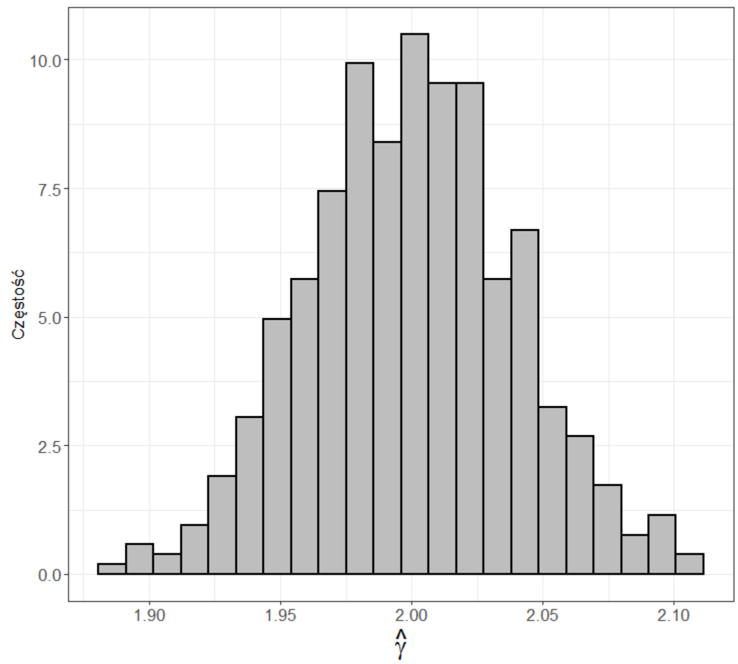
W tym zadaniu zbadamy wpływ liczby obserwacji na własności naszych estymatorów. W tym celu powtórzymy zadanie nr 2 dla $n = 500$.

Histogramy znajdują się poniżej:



Jako, że zwiększyliśmy liczbę obserwacji spodziewalibyśmy się, że nasze wyniki będą dużo lepsze. W przypadku estymatora β_0 widzimy poprawę, ale dla β_1 można powiedzieć, że się nawet pogorszył. Jeden słupek zdecydowanie wystaje poza krzywą gęstości.





Wykresy dla estymatorów parametrów Σ wydają się być podobne do tych z poprzedniego zadania. Znów wartości w obu przypadkach są bardzo blisko tej prawdziwej. Oczywiście wpływ liczby obserwacji poprawił naszą estymacje.

Poniżej znajduje się tabelka ze średnimi obciążeniami:

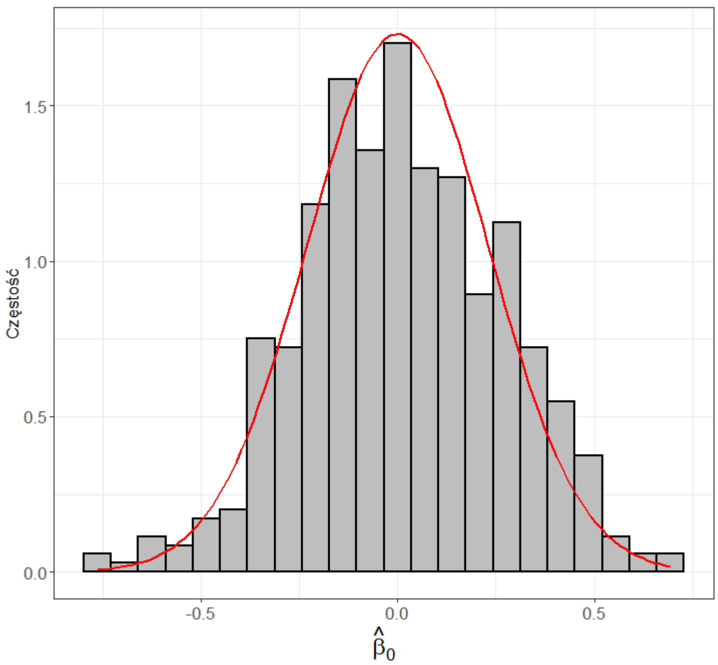
Parametr	β	ρ	γ
Wartość	0.048	0.001	0

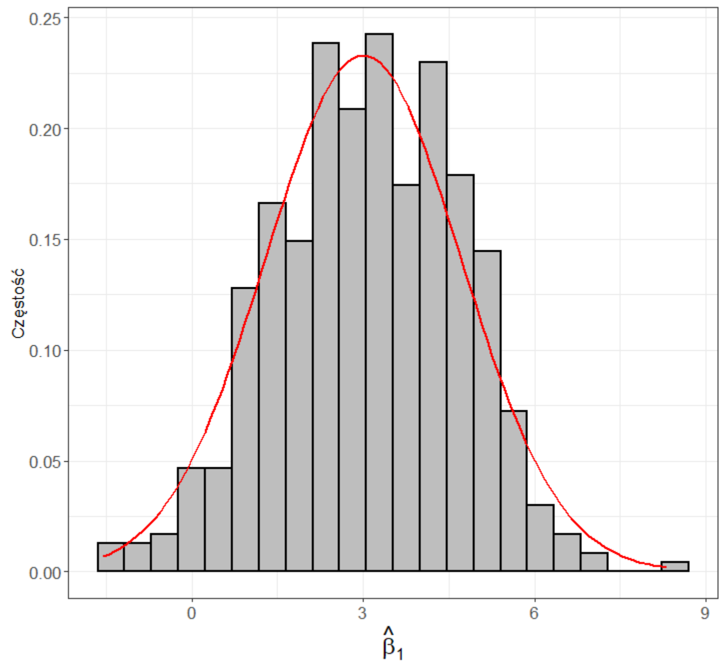
Dla estymatora β nasza wartość nie zmieniła się od tej z poprzedniego zadania, lecz dla pozostałych dwóch estymatorów średnie obciążenia zmniejszył się i są bardzo małe. Możemy, więc powiedzieć, że liczba obserwacji na pewno poprawia estymacje naszych parametrów w szczególności macierzy Σ.

Zadanie 4

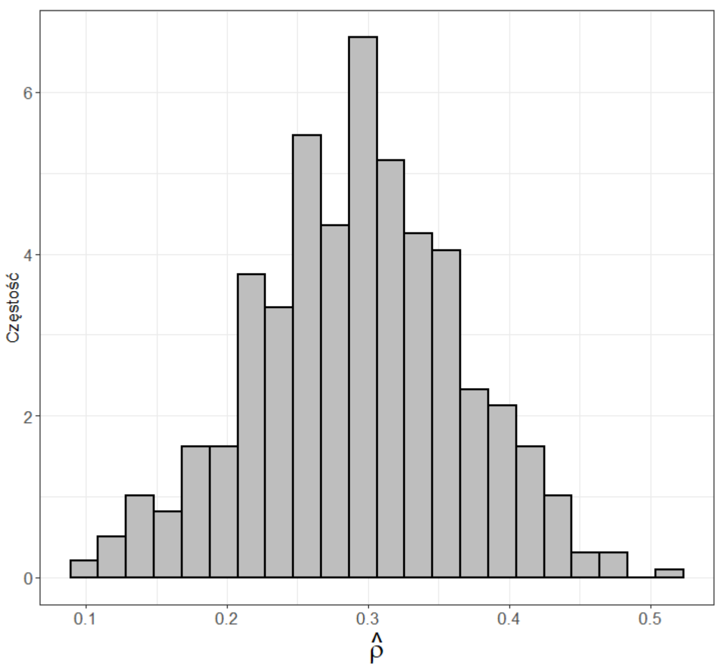
Natomiast w tym zadaniu badamy wpływ liczby pomiarów, czyli znów powtórzymy zadanie 2 tylko dla $k = 30$.

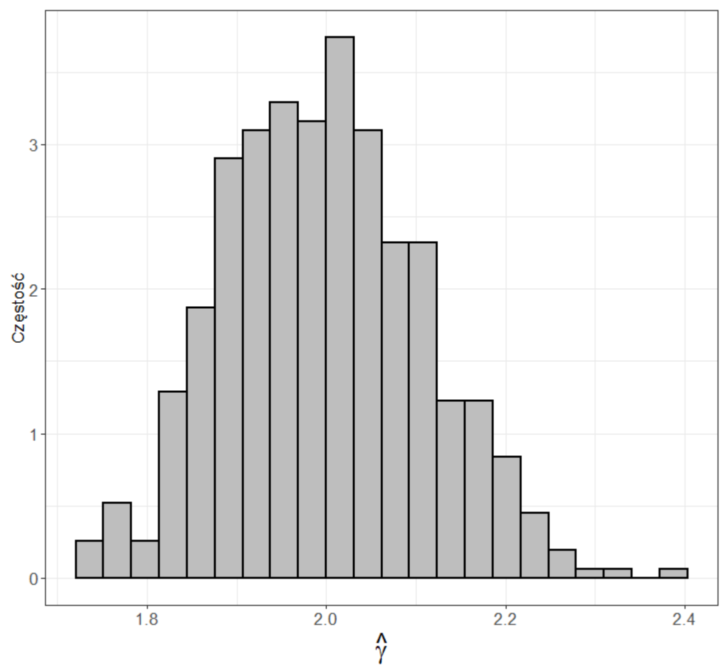
Histogramy znajdują się poniżej:





Widzimy, że liczba pomiarów również nieco poprawiła nasze histogramy. Nie są one idealnie dopasowane do krzywej, ale na pewno w przypadku β_0 widzimy znaczną poprawę, gdzie w drugim zadaniu histogram wychodził ponad krzywą gęstości.





Estymatory ρ i γ również wypadają lepiej w tym przypadku. Nie są one lepsze od tych z poprzedniego zadania, ale możemy zauważyć poprawę.

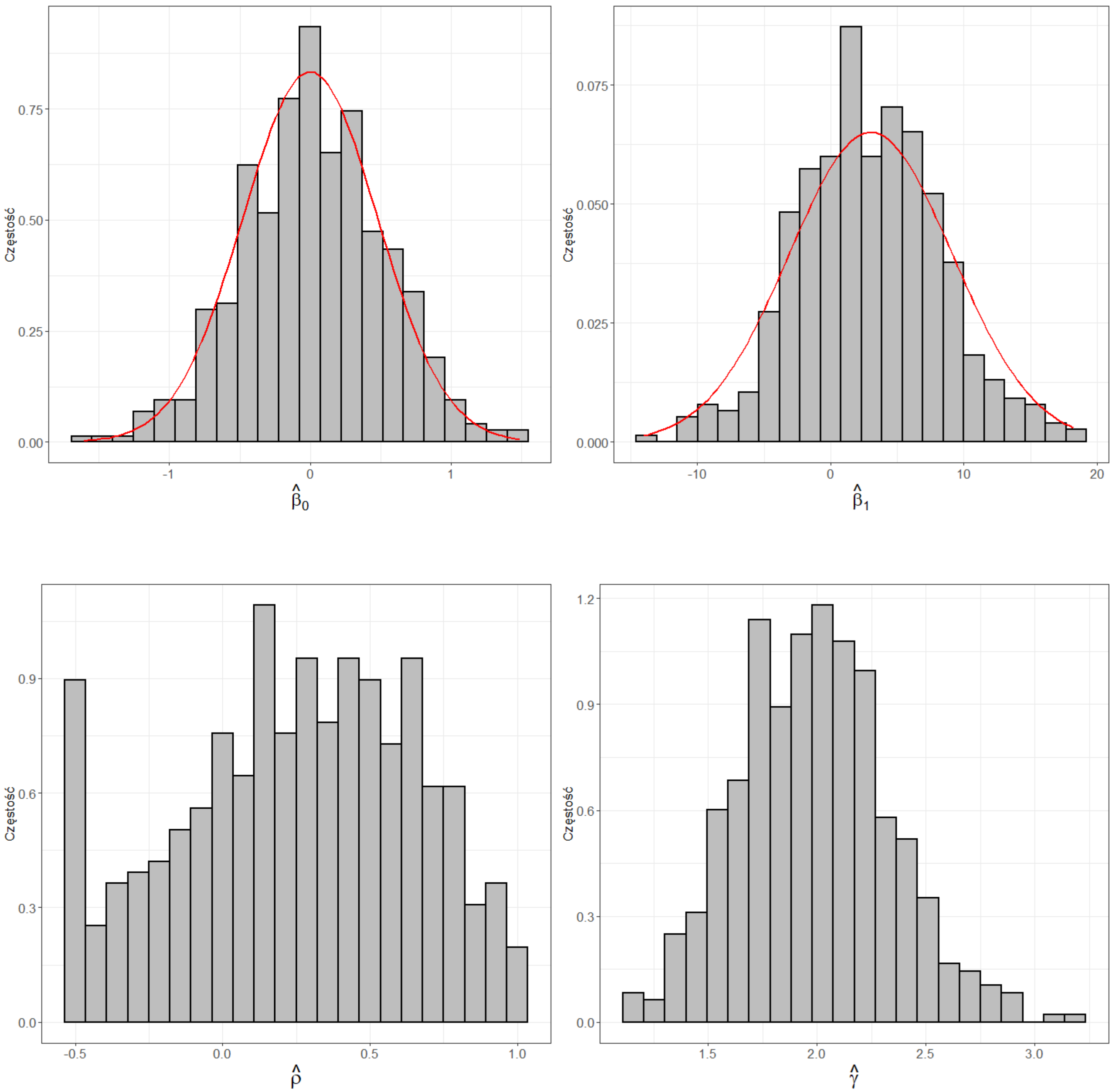
Parametr	β	ρ	γ
Wartość	0.093	-0.007	-0.002

Z powyższej tabeli widzimy, że obciążenie dla β pogorszyło się nieco w porównaniu z dwoma poprzednimi zadaniami. Natomiast dla ρ i γ obciążenia nie są mniejsze od tych z poprzedniego zadania, ale są mniejsze od tych z zadania nr 2.

Zadanie 5

To zadanie będzie analogiczne do pozostałych, tylko tym razem wykonamy je dla $p = 40$.

Histogramy znajdują się poniżej:



W tym zadaniu spodziewamy się, że nasze wyniki znacznie się pogorszą. Wynika to z tego, że wprowadzamy dodatkowy szum do naszego modelu. Na histogramach dla bet nie jest widoczna ogromna różnica, ale widzimy, że dopasowują się one do krzywej gęstości znacznie gorzej niż te z poprzednich zadań.

W przypadku estymatora parametru γ widzimy, że wartości często są bliskie faktycznej. Natomiast dla ρ histogram wygląda zupełnie inaczej niż w poprzednich zadaniach. Wartości tego estymatora układają się bardziej jednostajnie niż normalnie. Zastanawiające może być to, że bardzo często otrzymywaliśmy wartość bliską 0.5, co widać na histogramie.

Spójrzmy jeszcze na obciążenia:

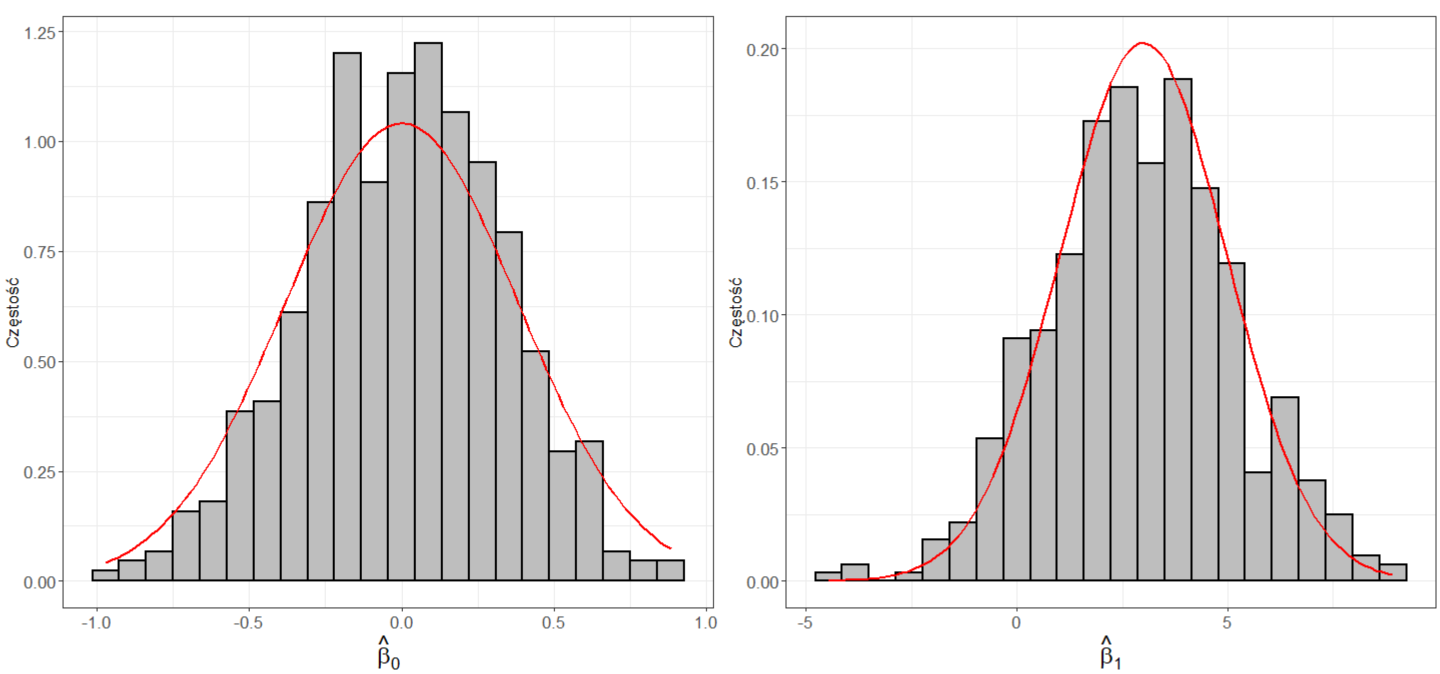
Parametr	β	ρ	γ
Wartość	0.454	-0.048	-0.012

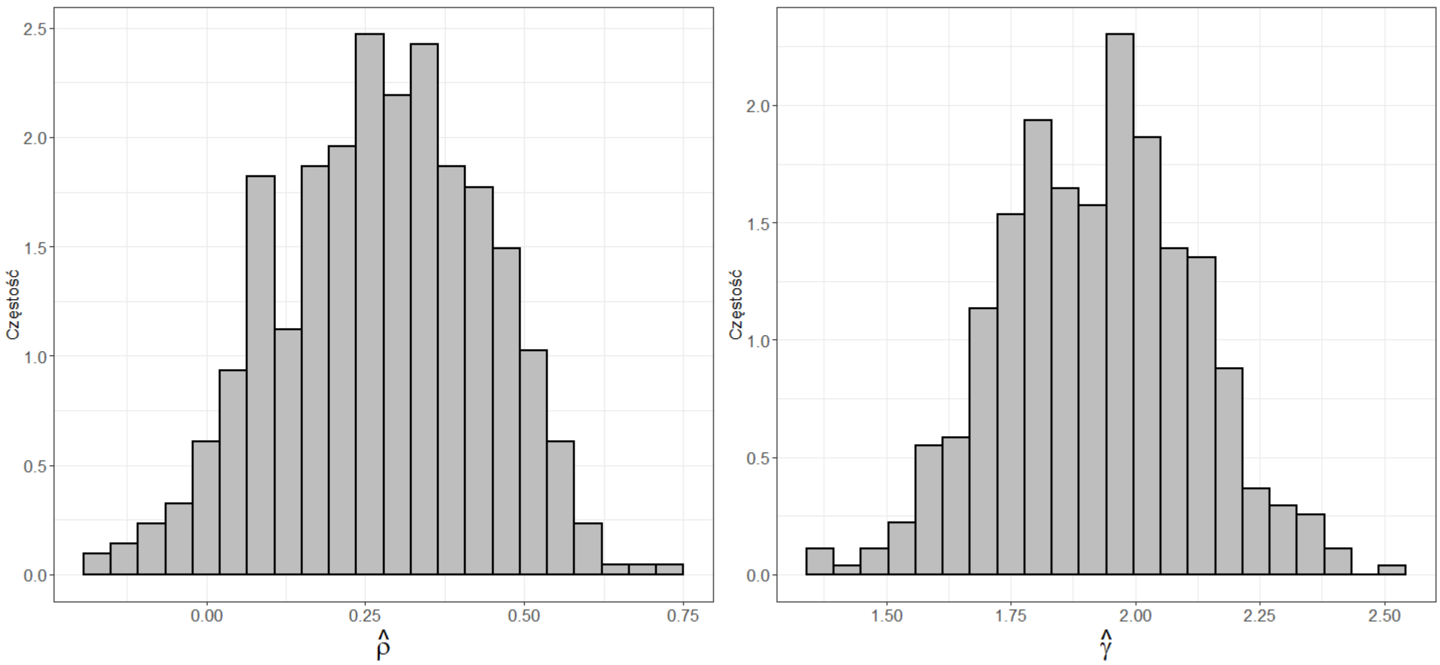
Z tabelki możemy wywnioskować, że obciążenia znacznie wzrosły. W porównaniu do zadania 2 nawet dziesięciokrotnie.

Podsumowując poprzednie zadania możemy powiedzieć, że modyfikacje parametrów n, k i p zmieniają własności estymacji parametrów regresji. Poprawę własności estymatorów mogliśmy obserwować, gdy zwiększyliśmy liczby n i k w porównaniu do drugiego zadania. Poprawiały się różne wartości w zależności od tego parametru, ale nie obserwowaliśmy żadnego pogorszenia. Natomiast pogorszenie tych własności występował dla zwiększenia liczby p. Największą różnicę widzieliśmy na histogramie estymatora ρ.

Zadanie 6

Na koniec powtórzymy zadanie 2 z tymi samymi parametrami tylko do estymacji użyjemy metody "ML". Wyniki porównamy w takim sam sposób jak poprzednio.





Porównując histogramy dla bet z tymi z zadania drugiego nie obserwujemy znaczącej różnicy. Wydaje się, że histogram dla $\hat{\beta}_0$ nieznacznie się poprawił, a dla $\hat{\beta}_1$ nieco pogorszył.

Natomiast histogramy dla parametrów macierzy Σ nieco wyglądają gorzej. Histogram dla $\hat{\rho}$ jest w miarę podobny do tego z zadania drugiego, ale histogram $\hat{\gamma}$ jest nieco gorszy. Przede wszystkim na pewno nie jest on symetryczny względem 2, czyli faktycznej wartości. Jego środek wydaje się być umiejscowiony w nieco mniejszej wartości.

Obliczymy jeszcze obciążenia:

Parametr	β	ρ	γ
Wartość	0.047	-0.025	-0.076

W tabelce widzimy, że nasze obserwacje z histogramów mają w niej pokrycie. Obciążenie dla $\hat{\beta}$ jest niemal identyczne jak to z zadania drugiego. Obciążenie dla $\hat{\rho}$ zwiększyło się około dwukrotnie, a dla $\hat{\gamma}$ ponad dziesięciokrotnie.

Podsumowując możemy, więc stwierdzić, że metoda estymacji "ML" nie zmienia za bardzo własności estymatora β, natomiast zdecydowanie gorzej radzi sobie z estymowaniem macierzy kowariancji.