Zaawansowane modele liniowe raport nr 1

Dominik Mika

22 marca 2021

Cele raportu

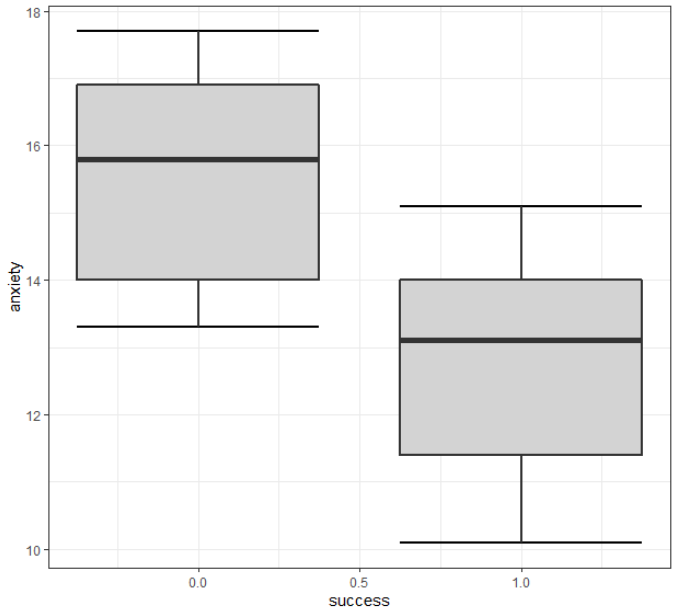
1. Analiza zbioru, który opisuje relacje między p-stwami przyjęcia na studia (*success*), a wynikami testów rachunkowych (*numeracy*) i poziomu niepewności (*anxiety*), przy pomocy modelu regresji logistycznej, gdzie zmienna celu to success.
2. Przeprowadzenie symulacji i zbadanie zachowania się estymatorów wektora β.

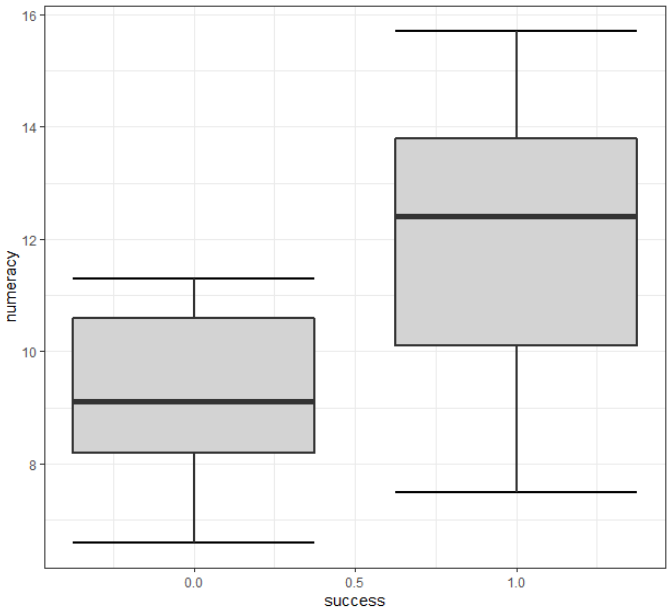
Analiza danych

Zadanie 2, 3

W tych zadaniach narysujemy wykresy pudełkowe dla zmiennych *„numeracy"* oraz *„anxiety"* w rozbiciu na grupę przyjętych/nieprzyjętych osób.

Wykresy znajdują się poniżej:





Widzimy, że w przypadku obu wykresów prawdopodobieństwo dostania się na studia, różni się w zależności od wartości zmiennych *numeracy* i *success*, możemy więc podejrzewać, że zmienne te wpływają w jakiś sposób na przyjęcie studentów.

Zadanie 4

W tym zadaniu dla powyższych danych skontruujemy model regresji logistycznej. Następnie wykonamy następujące polecenia:

- Podamy estymatory parametrów i wyniki testów istotności.
- Wyznaczymy przewidywane p-stwo sukcesu u studenta, którego *anxiety*=13, a *numeracy*=10.
- Narysujemy krzywą ROC dla dopasowanego modelu statystycznego.

W poniższej tabeli przedstawiam estymatory paramtrów oraz p-wartości testów ich istotności.

zmienna	w. estymatora	p-wartość
numeracy	0.5774	0.01995
anxiety	-1.3841	0.00396

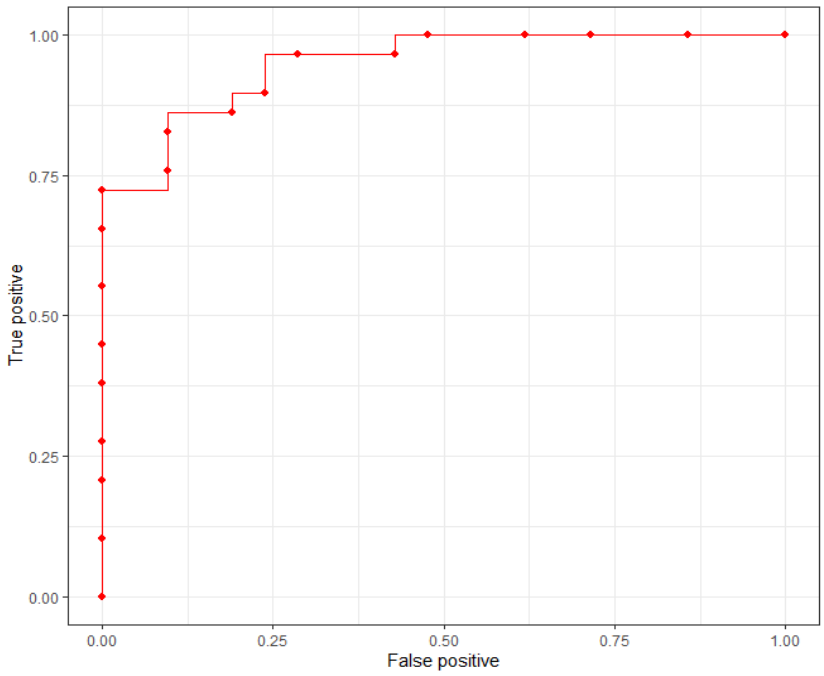
Widzimy, że p-wartości są dość małe. Na poziomie istotności $\alpha = 0.05$ odrzucilibyśmy hipotezę zerową dla obu parametrów.

Następnie chcielibyśmy wyznaczyć p-stwo sukcesu dla konkrentego studenta. Zadaje się ono wzorem:

$$\eta_{10,13} = \frac{1}{1 + exp(-(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2))}$$

i wynosi 0.883.

Następnie mieliśmy narysować krzywą ROC dla tego modelu. Znajduje się ona poniżej.



Jak można zauważyć krzywa ROC dla zera szybko zbliża się do 1 i już dla wartości False positive równej około 0.45 osiąga 1.

Podsumowując wszystkie wyniki, możemy stwierdzić, że nasze zmienne odpowiedzi dość dobrze wyjaśniają nasz model.

Zadanie 5

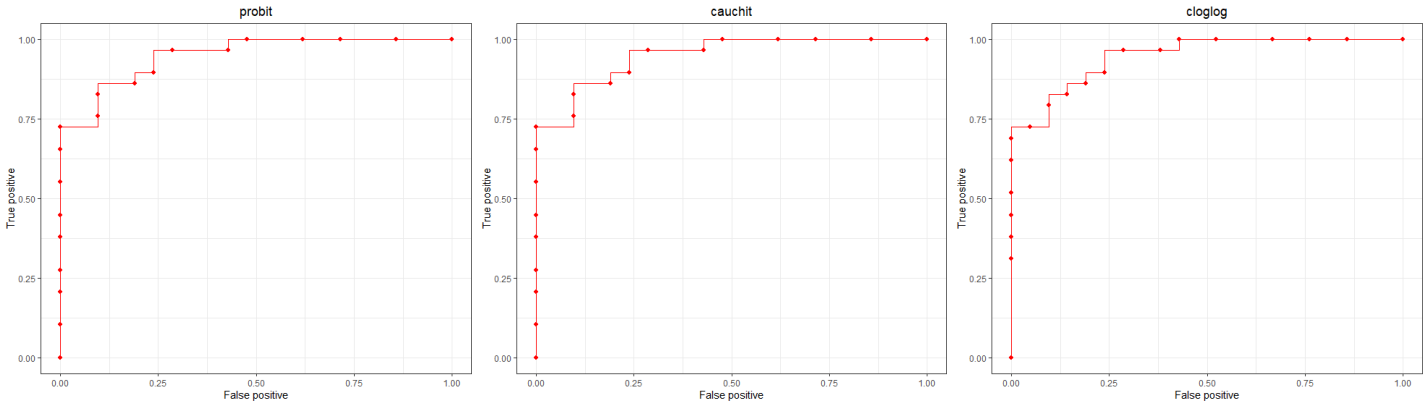
Tym razem powtórzymy poprzednie zadanie dla różnych funkcji linkujących („probit", „cauchit", „cloglog") i ocenimy, który model zapewnia najlepsze dopasowanie do naszych danych.

Zadanie wykonujemy w analogiczny sposób do poprzedniego, tylko w przypadku tworzenia modelu zmieniamy funkcje linkujące.

Wyniki przedstawiamy w tabelach poniżej.

zmienna	w. estymatora	p-wartość	p-stwo
probit			
numeracy	0.3371	0.01374	0.88
anxiety	-0.8039	0.00142	
cauchit			
numeracy	0.7323	0.1224	0.885
anxiety	-1.7741	0.0735	
cloglog			
numeracy	0.4024	0.00819	0.896
anxiety	-0.939	0.0047	

Widzimy, że wartości parametrów nieco się różnią od siebie, lecz największą różnicę widzimy dla modelu z funkcją linkującą „cauchit". Testy istotności w prawie wszystkich przypadkach odrzucają hipotezę zerową. Wyjątkiem są testy dla modelu z funkcją „cauchit". Prawdopodobieństwa sukcesu studenta z poprzedniego zadania są bardzo podobne do siebie. Najmniejsze jest dla funkcji „cauchit".



Wszystkie krzywe ROC, wraz z tą z poprzedniego zadania są bardzo do siebie podobne. Za wyjątkiem krzywej „cloglog" są nawet identyczne.

Do podsumowania spójrzmy jeszcze na miarę dopasowania danych do modelu przy pomocy statystyki deviance.

	logit	*probit*	*cauchit*	*cloglog*
Deviance	28.286	27.854	31.115	28

Podsumowując zadanie, możemy stwierdzić, że nie ma jasno najlepszego modelu z pośród czterech rozważanych. Jedynie nieco wyróżniał się model z funkcją „cauchit". Możemy powiedzieć, że dawał nam najgorsze dopasowanie z tej czwórki, a najlepsze „probit".

Zadanie 6

W tym zadaniu rozważać będziemy model z funkcją linkującą „logit". Dla tego modelu wykonamy następujące eksperymenty:

- Wyznaczymy estymator macierzy kowariancji wektora estymatorów parametrów modelu i porównamy jej wyrazy na przekątnej z estymatorami odchyleń standardowych zwracanych przez R.
- Przetestujemy hipotezę, że obie zmienne nie mają wpływu na zmienną celu.
- Podamy definicję parametru „epsilon" i jego wartość domyślną przyjmowaną w funkcji *glm()*. Następnie stworzymy ponownie modele regresji logistycznej ze zmienionym parametrem epsilon ze zbioru: $10^{-1}, 10^{-2}, 10^{-3}, 10^{-6}$. Dla takich modeli porównamy liczbę iteracji i wartości estymatorów parametrów.

Wyznaczymy teraz estymator macierzy kowariancji wektora estymatorów. Z asymptotycznego rozkładu $\hat{\beta}$ wiemy, że jest to macierz J^{-1}, gdzie $J = X'S(\beta)X$, a $S(\beta)$ to macierz diagonalna, gdzie $S_{ii} = \mu_i(b)(1-\mu_i(b))$. Przedstawia się ona następująco:

$$\begin{pmatrix} 46.2295 & -0.2482 & -3.062 \\ -0.2482 & 0.0616 & -0.0238 \\ -3.062 & -0.0238 & 0.2308 \end{pmatrix}.$$

Natomiast wartości kwadratów odchyleń standardowych jakie zwrócił nam R, to: 46.2199, 0.0615, 0.2308. Porównując je z wyrazami na przekątnej naszej macierzy, widzimy, że są bardzo bliskie sobie.

Teraz chcielibyśmy przetestować, czy obie zmienne objaśniające mają wpływ na zmienną odpowiedzi. W tym celu wykorzystamy test dla statystki Deviance, która wygląda następująco:

$$D(M(\hat{\beta})) = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\mu_i(\hat{\beta})}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\mu_i(\hat{\beta})}\right)\right].$$

Testujemy, więc hipotezę:

$$H_0 : \forall_i \beta_i = 0 \quad vs \quad H_1 : \exists_i \beta_i \neq 0.$$

Oznaczmy M_1 jako model pełny, a M_0 jako model zredukowany, wtedy statystyka testowa ma postać:

$$\chi^2 = D(M_0) - D(M_1)$$

przy H_0 ma ona asymptotycznie rozkład χ^2 z 2 stopniami swobody.

W naszym przypadku statystyka wynosi:

$$\chi^2 = 68.029 - 28.286 = 39.743$$

Natomiast p-wartość wynosi $2.34 \cdot 10^{-9}$. Widzimy, więc że p-wartość jest bardzo mała i test odrzuciłby hipotezę zerową dla standardowych poziomów istotności. Stąd możemy wnioskować, że zmienne wnoszą informacje do modelu.

Na koniec tego zadania chcieliśmy się zastanowić nad parametrem „epsiolon". Jest to parametr w funkcji *glm()*, który ma związek z przybliżaniem naszych parametrów regresji. Mianowicie „epsilon" to taki poziom, dla którego nasza medota estymacji zatrzymuje się w momencie, gdy spełniona jest nierówność:

$$\frac{|dev - dev_{old}|}{|dev| + 0.1} < \epsilon,$$

gdzie dev i dev_{old}, to odpowiednio statystyki deviance, dla modelu w tym i poprzednim kroku.

Jego domyślna wartość w tej funkcji to 10^{-8}. Teraz rozważymy modele stworzone z ustalonym parametrem „epsilon".

Wyniki dla różnych parametrów znajdują się w tabeli poniżej:

epsilon	zmienna	parametr	iteracje
10^{-1}	numeracy anxiety	0.5376 -1.264	3
10^{-2}	numeracy anxiety	0.5735 -1.3713	4
10^{-3}	numeracy anxiety	0.5773 -1.3839	5
10^{-6}	numeracy anxiety	0.5774 -1.3841	6

Widzimy, że liczba iteracji zwiększa się wraz ze spadkiem wartości epsilon, co jest oczywiste. Natomiast widzimy, że dla $\epsilon = 10^{-6}$ metoda otrzymała takie same wyniki jak dla domyślnej wartości, czyli jest to najbardziej optymalna wartość epsilona z tych.

Symulacje

Zadanie 1

W tym zadaniu chcemy wylosować macierz planu $X_{n\times p}$, gdzie $n = 400$ i $p = 3$. Jej elemantami są zmienne losowe z rozkładu $N(0, \sigma^2 = 1/400)$. Następnie wylosujemy wektor binarny Y zgodnie z modelem regresji logistycznej z wektorem $\beta = (3, 3, 3)$. Wyznaczymy macierz informacji Fishera w punkcie β i asymptotyczną macierz kowariancji estymatorów największej wiarogodności.

Następnie wygenerujemy 1000 replikacji wektora odpowiedzi zgodnie z powyższym modelem i na podstwie ich wykonamy następujące polecenia:

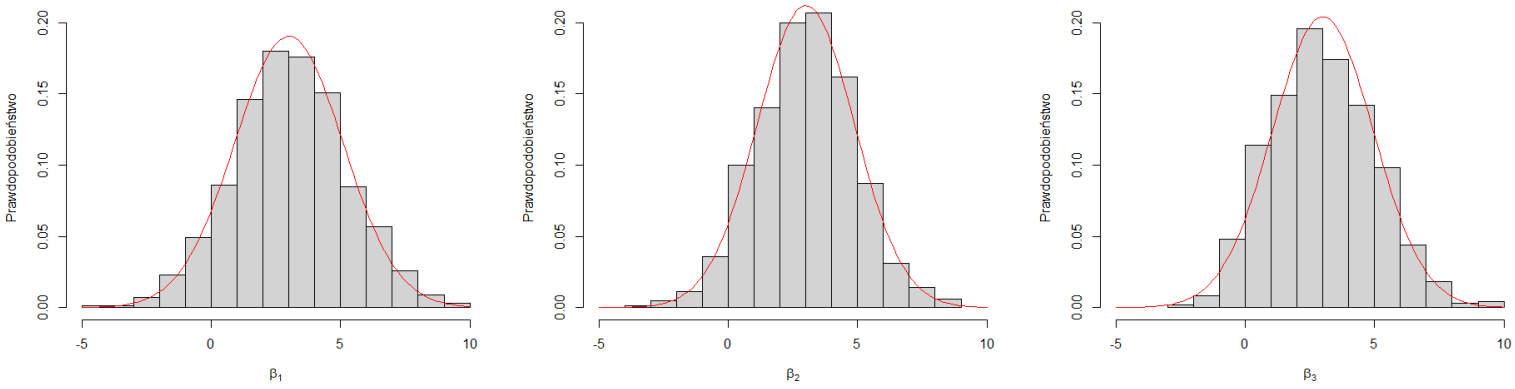
- Narysujemy histogramy estymatorów $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ i porównamy je z ich rozkładami asymptotycznymi.
- Wyestymujemy obciążenie estymatorów $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$.
- Wyestymujemy macierz kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ i porównamy ją z asymptotyczną macierzą kowariancji.

Do wyznaczenia wektora odpowiedzi Y skorzystamy z tego, że $\eta = \mathbb{E}(Y) = X\beta$. Mając wartość oczekiwaną Y możemy wyznaczyć wektor prawdopodobieństw:

$$\mu = \frac{1}{1 + exp(-\eta)}.$$

Teraz możemy wylosować wektor odpowiedzi Y z rozkładu dwupunktowego, z prawdopodobieństwem μ określonym dla każdego powtórzenia.

Po wykonaniu odpowiednich obliczeń otrzymaliśmy estymatory wektora β. Ich histogramy z krzywą gęstości rozkładu normalnego prezentujemy poniżej:



Widzimy, że wszystkie trzy histogramy dość dobrze dopasowują się do krzywej z ich rozkładu asymptotycznego. Oznacza to, że nasze estymatory zgadzają się z tymi teoretycznymi.

Natomiast obciążenia estymatorów wyniosły odpowiednio 0.051, -0.011, 0.005. Są one dość małe.

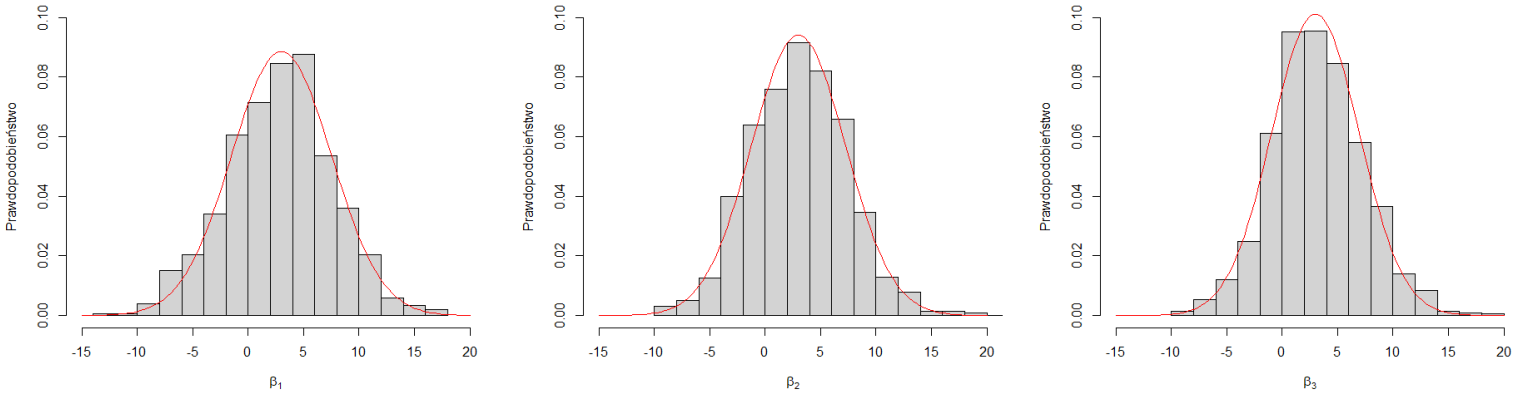
Wyestymowana macierz kowariancji oraz asymptotyczna macierz kowariancji wyglądają następująco.

$$J^{-1} = \begin{pmatrix} 4.387 & -0.046 & 10.041 \\ -0.046 & 3.542 & 0.01 \\ 0.041 & 0.01 & 3.818 \end{pmatrix}, \quad \text{Cov}(\hat{\beta}) = \begin{pmatrix} 4.699 & -0.07 & -0.187 \\ -0.07 & 3.573 & 0.08 \\ -0.187 & 0.08 & 3.967 \end{pmatrix}$$

W celu zbadania różnic pomiędzy wyestymowaną macierzą kowariancji wektora $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ i asymptotyczną macierzą kowariancji, skorzystamy z normy L_2. Policzymy normę L_2 wekora różnic tych macierzy. W tym przypadku wynosi ona 0.4842735. Widzimy, że jest ona mała i nasze macierze są bardzo podobne.

Zadanie 2

W tym zadaniu zbadamy wpływ liczby obserwacji na nasz poprzedni eksperyment, więc tym razem stworzymy macierz planu, gdy $n = 100$.

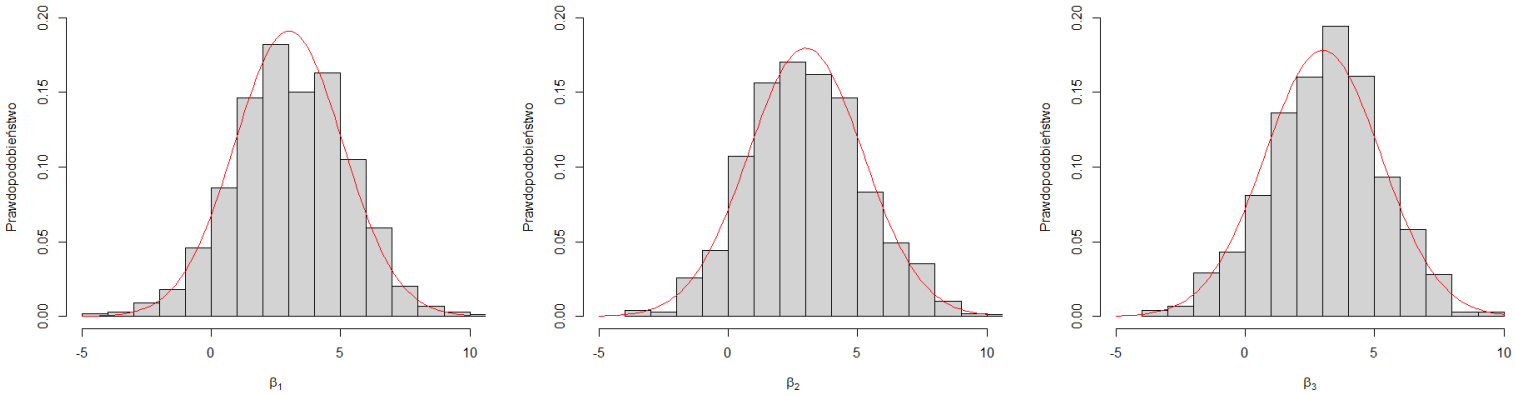


Obciążenia estymatorów wynoszą -0.129, 0.104, 0.206. Widzimy, że zwiększyły się one nieco w porównaniu z poprzednim modelem, ale nadal nie są duże. Norma różnicy estymowanej macierzy kowariancji i tej asymptotycznej wynosi 3.202201. Zwiększyła się w porównaniu z poprzednim podpunktem. Różnice w tym przypadku są już zauważalne.

Zadanie 3

Tym razem chcemy zbadać wpływ korelacji między regresorami na naszą estymację. Powtórzymy zadanie 1, gdy wiersze macierzy X są niezależnymi wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$ z miacierzą kowariancji $\Sigma = \frac{1}{n}S$, gdzie $S_{ii} = 1$, a dla $i \neq j$, $S_{ij} = 0.3$.

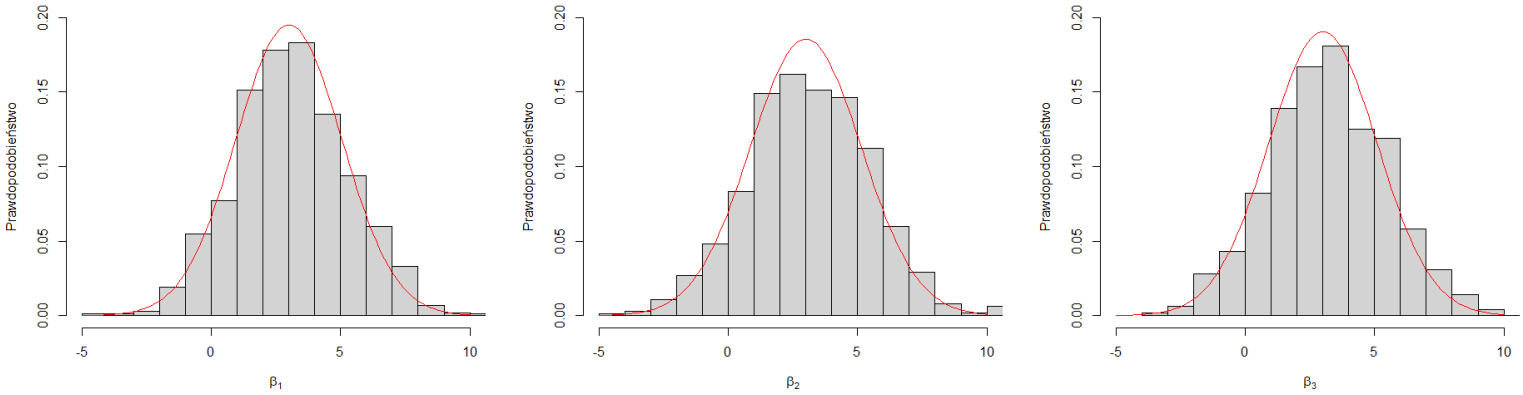
W podanym przypadku histogramy wyglądają następująco:



Obciążenia estymatorów wynoszą 0.096, 0.03, 0.12. Są mniejsze niż w poprzednim podpunkcie, ale nieco większe od tych z zadania 1. Norma różnicy naszych macierzy wyniosła 3.202201, czyli również, te macierze są bliskie.

Zadanie 4

Teraz zbadamy wpływ liczby regresorów na nasz model. Naszą macierz X stworzymy analogicznie jak w zadaniu 1, tylko dla $p = 20$.



Obciążenia estymatorów wynoszą 0.145, 0.16, 0.215. Są one widocznie największe z wszystkich naszych rozważanych przypadków. Podobnie jest z normą macierzową. Wynosi ona 4.565544.

Zadanie 5

Podsumowując, widzimy, że niektóre zmiany nie zmieniły rozkładu i własności estymatora β. Zmiana liczby obserwacji wpłynęła na to, co nie powinno nas dziwić, ponieważ nie mamy tyle informacji, aby dobrze wyestymować te parametry. Podobnie, było w przypadku, gdy zwiększyliśmy liczbę regresorów. Wprowadziły one pewien „szum" do naszego modelu. Natomiast, jak mogliśmy zaobserwować korelacja pomiędzy zmiennymi nie wpłynęła w znacznym stopniu na nasz model.