Raport 1

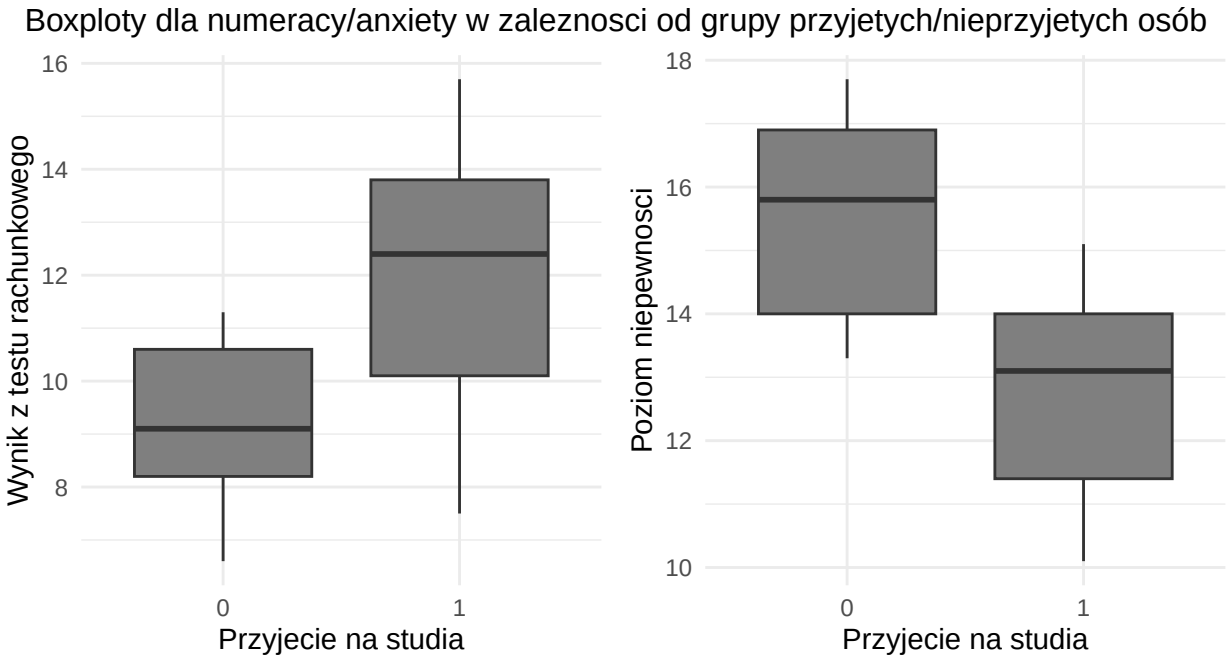
Magdalena Potok

2024-04-17

Celem raportu jest badanie relacji między p-stwami przyjęcia na studia, wynikami z testów rachunkowych oraz poziomem niepewności, z wykorzystaniem analizy danych oraz symulacji, w celu zrozumienia wpływu różnych czynników na modele regresji logistycznej oraz oceny ich efektywności.

Zadanie 1,2 i 3

Zbiór danych **"Lista_1.csv"** opisuje relacje między p-stwami przyjęcia na studia (success), a wynikami z testów rachunkowych (numeracy) i poziomu niepewności (anxiety). Poniżej zanjdują się boxploty dla zmiennej "numeracy" oraz "anxiety" w rozbiciu na grupę przyjętych/nieprzyjętych osób.



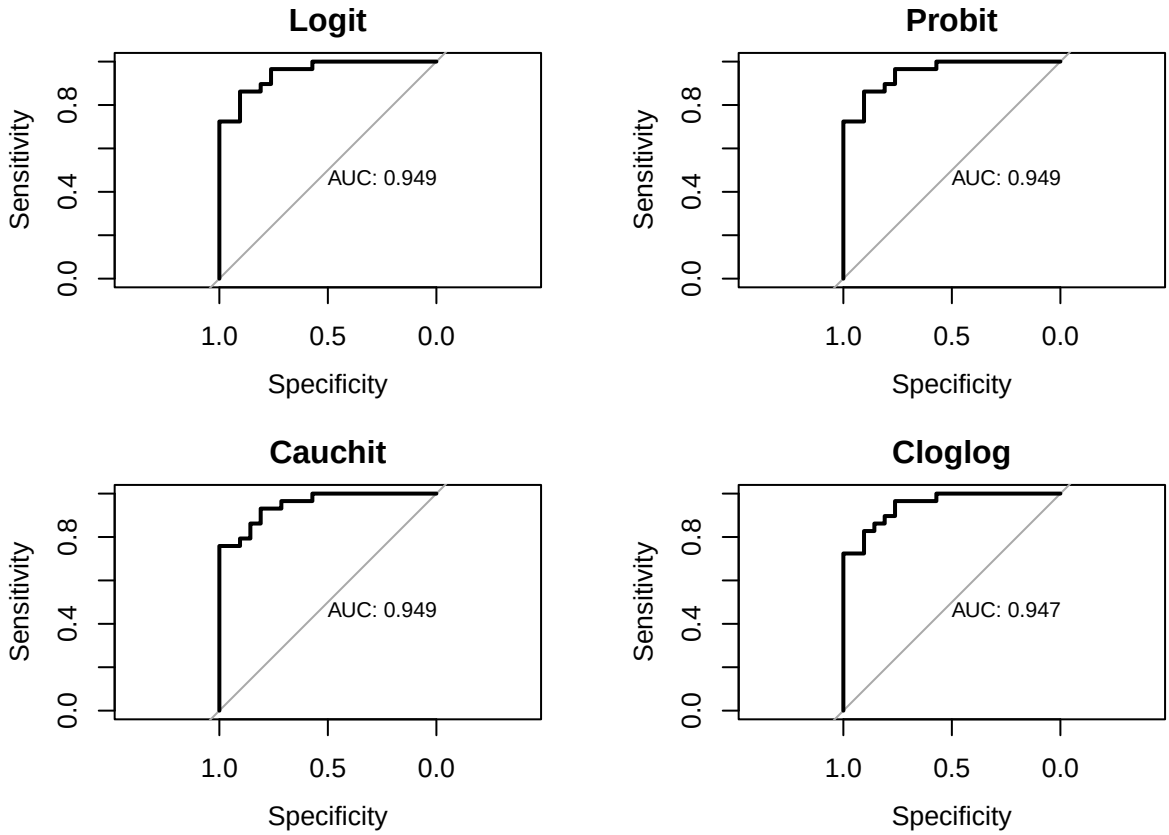
Wykresy przedstawiają rozkład zmiennych numeracy oraz poziomu niepewności (anxiety) w zależności od grupy przyjętych/nieprzyjętych osób. Z analizy boxplotów wynika, że obie zmienne - numeracy i anxiety - wydają się mieć istotny wpływ na sukces kandydatów.
Pierwszy wykres pokazuje, że osoby przyjęte na studia tendencjonalnie uzyskiwały wyższe wyniki z testów rachunkowych (numeracy) w porównaniu do tych, które nie zostały przyjęte. Możemy wnioskować, że wyniki z testów rachunkowych mogą odgrywać istotną rolę w procesie rekrutacji, przy czym osoby z wyższymi wynikami mają większe szanse na przyjęcie na studia.
Drugi wykres ukazuje, że osoby, które nie zostały przyjęte na studia, zazwyczaj wykazywały wyższy poziom niepewności (anxiety) w porównaniu do osób przyjętych. To sugeruje, że poziom niepewności może negatywnie wpływać na szanse przyjęcia na studia, przy czym osoby o niższym poziomie niepewności mogą mieć większe szanse na sukces.

Zadanie 4 i 5

Skonstruuję model regresji logistycznej dla powyższych danych używając różnych funkcji linkujących (logit, probit, cauchit, cloglog). Podam dla każdej z funkcji estymatory parametrów i wyniki testów istotności, wyznaczę przewidywane p-stwo sukcesu u studenta, którego anxiety = 13, numeracy = 10. Narysuję krzywą ROC dla dopasowanych modelów statystycznych.

Model	E_Intercept	E_Numeracy	E_Anxiety	p_Intercept	p_Numeracy	p_Anxiety	prawdop	AIC
Logit	14.239	0.577	-1.384	0.036	0.020	0.004	0.883	34.286
Probit	8.257	0.337	-0.804	0.025	0.014	0.001	0.881	33.854
Cauchit	18.383	0.732	-1.774	0.135	0.122	0.074	0.885	37.115
Cloglog	9.001	0.402	-0.939	0.053	0.008	0.005	0.896	34.000

W tabelce kolumny zaczynające się "E_" oznaczają estymatory współczynników konkretnych parametrów, natomiast "p_" to p-wartości, które mówią o ich istotności. W przypadku funkcji logit oraz probit wszystkie zmienne są istotne, co potwierdzają (przy założonym poziomie istotności $\alpha = 0.05$) wartości p-value dla współczynników w modelach. Natomiast w przypadku funkcji cloglog tylko dla interceptu nie odrzucamy hipotezy zerowej $H_0 : \beta_0 = 0$. Te wyniki są zgodne z obserwacjami płynącymi z analizy boxplotów. Funkcja cauchit nie wskazuje na istotność żadnego z parametrów, co sugeruje, że jej skuteczność może być ograniczona w tym konkretnym przypadku. Prawdopodobieństwo sukcesu dla studenta z numeracy=10 i anxiety=13 jest równe około 0.88, co jest stosunkowo wysoką wartością, sugerującą dość wysokie szanse na sukces w kontekście tych wartości predykcyjnych. Według kryterium AIC, model z funkcją cauchit jest najlepiej dopasowany. Jednakże, należy podkreślić, że wartość AIC jest tylko jednym z kryteriów oceny modeli, a istotność parametrów również stanowi istotny czynnik w ocenie jakości modeli. Wartość AIC w modelach logitowych jest zbliżona, z czego wynika, że model logitowy jest drugim najlepszym modelem pod względem dopasowania.



Przy ocenie jakości klasyfikatora pomocny jest wykres ROC (Receiver Operating Characteristic), który ilustruje zależność między czułością (sensitivity) a specyficznością (specificity) predyktora w zależności od wartości parametru s. Czym większa powierzchnia pod krzywą ROC, tym lepsza jakość klasyfikacji. Dla analizowanych funkcji linkujących osiągamy podobne dopasowanie modelu do danych. Najgorsze wyniki uzyskujemy dla funkcji cloglog.

Zadanie 6

Dla modelu z funkcją linkującą "logit":
> Wyznaczę estymator macierzy kowariancji wektora estymatorów parametrów w modelu regresji logistycznej, następnie porównam wartości na przekątnej z estymatorami odchyleń standardowych zwracanych przez R. Macierz kowariancji wygląda następująco:

```
##             (Intercept)    numeracy     anxiety
## (Intercept)   46.229542 -0.24821804 -3.06203011
## numeracy      -0.248218  0.06155249 -0.02382592
## anxiety       -3.062030 -0.02382592  0.23084660
```

Natomiast estymatory parametrów wynoszą:

	Estymator_odchyl_stand	Estymator_wariacji
(Intercept)	6.7985	46.2199
numeracy	0.2481	0.0615
anxiety	0.4804	0.2308

Estymator wariancji ma takie same co do drugiej liczby po przecinku wartości, co wyrazy na przekątnej asymptotycznej macierzy kowariancji.

> Przetestuję jedną hipotezę, że obie zmienne objaśniające nie mają wpływu na zmienną odpowiedzi. Zatem nasza hipoteza:

$$H_0 : \forall i \in \{1,2\}\ \beta_i = 0 \quad vs. \quad H_1 : \exists i \in \{1,2\}\ \beta_i \neq 0$$

Statystyka χ^2 wynosi:

```
round(summary(modellog)$null.deviance - summary(modellog)$deviance,3)
```

```
## [1] 39.744
```

natomiast kwantyl rozkładu wynosi:

```
qchisq(1-0.05, 2)
```

```
## [1] 5.991465
```

Statystyka jest większa, więc odrzucamy H_0 na poziomie istotności $\alpha = 0.05$. Oznacza to, że przynajmniej jeden regresor jest istotny.

> Wykonam ponownie obliczenia stosując wartości epsilon ze zbioru $10^{-1}, 10^{-2}, 10^{-3}, 10^{-6}$. Porównam liczbę iteracji i wartości estymatorów poszczególnych parametrów.

Epsilon traktujemy jako wartość tolerancji, która określa granice błędów numerycznych akceptowanych w obliczeniach. Domyślna wartość epsilonu wynosi 10^{-8}. Jest to wartość, poniżej której różnice między wartościami numerycznymi są uznawane za pomijalne, a dalsze zmniejszanie epsilonu może być zbędne lub prowadzić do błędów związanych z ograniczeniami maszynowymi.

Epsilon	10^{-1}	10^{-2}	10^{-3}	10^{-6}
Liczba iteracji	3	4	5	6

Zatem im mniejsza jest wartość epsilon, tym więcej iteracji wykonujemy, co może prowadzić do dokładniejszego modelu.

Symulacje

Zadanie 1 i 2

Wygeneruję macierz X wymiaru $n = 400, p = 3$ oraz $n = 100, p = 3$, której elementy są zmiennymi losowymi z rozkładu $N(0, \sigma^2 = 1/400)$. Założę, że binarny wektor odpowiedzi jest wygenerowany zgodnie z modelem regresji logistycznej z wektorem $\beta = (3, 3, 3)$. Wyznaczę macierz informacji Fishera w punkcie β i asymptotyczną macierz kowariancji estymatorów największej wiarogosności. Wygeneruję 1000 replikacji wektora odpowiedzi zgodnie z powyższym modelem i na podstawie każdej replikacji wyznaczę estymator wektora β.

```
X = matrix(rnorm(1200,0,1/20),400,3)
B = c(3,3,3)
eta = X%*%B
mu = c(plogis(eta)) #nakladam funkcje linkujaca
Y = rbinom(400,1,mu)
model = glm(Y~X-1,family = "binomial")

S = diag(400)
S_diag = mu * (1-mu)
S = S * S_diag
fisher_matrix = t(X) %*% S %*% X
fisher_matrix
```

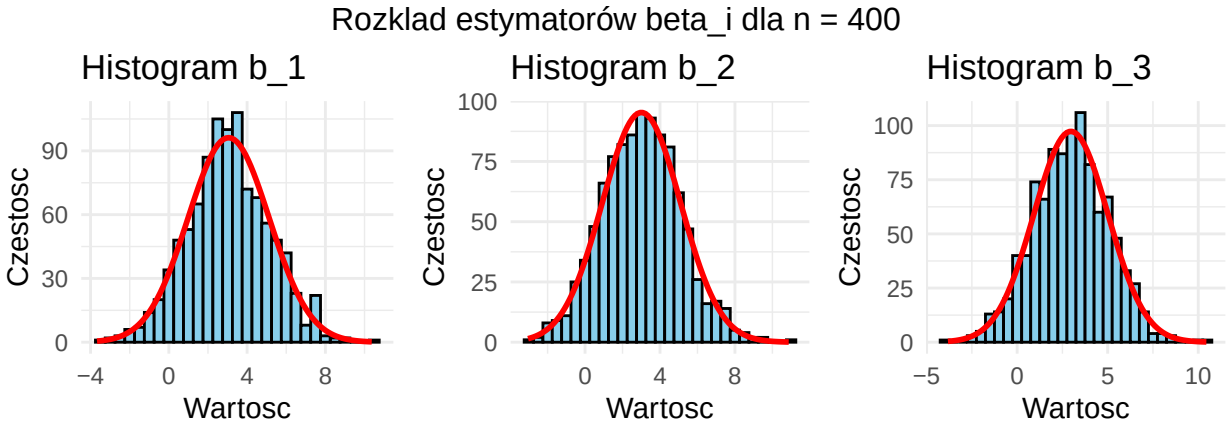
```
##              [,1]          [,2]          [,3]
## [1,]  0.228013379  0.0029682800 -0.0024343741
## [2,]  0.002968280  0.2823964487 -0.0008068508
## [3,] -0.002434374 -0.0008068508  0.2619129611
```

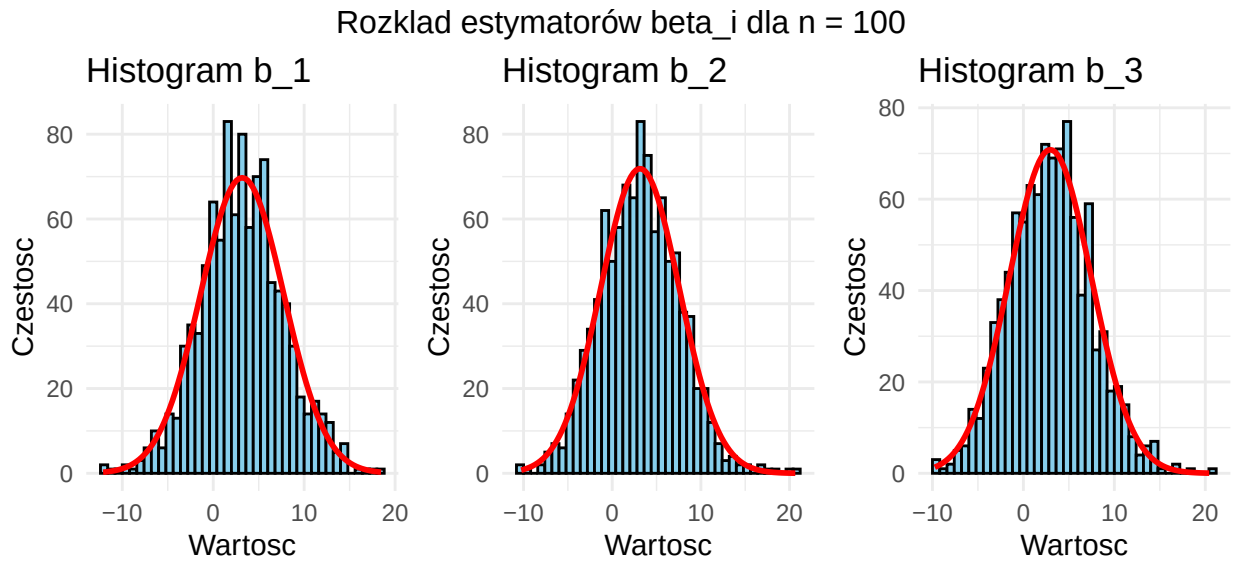
Macierz asymptotycznej kowariancji jest postaci:

```
solve(fisher_matrix)
```

```
##             [,1]        [,2]       [,3]
## [1,]  4.38674010 -0.04599311 0.04063127
## [2,] -0.04599311  3.54163440 0.01048290
## [3,]  0.04063127  0.01048290 3.81847224
```

> Narysuję histogramy estymatorów $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ i porównam ich rozkładami asymptotycznymi.





Dla $n = 400$ wszystkie trzy histogramy wyglądają podobnie, tak samo dla $n = 100$. Czerwoną linią na wykresie zostały zaznaczone rozkłady asymptotyczne, widać, że wszystkie są histogramy są dość zbliżone do tego rozkładu. Można zauważyć, że histogramy dla $n = 100$ są szersze, wynika to z większej wariancji estymatora.
$>$ Wyestymuję obciążenie estymatorów $\hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
n=100	0.181	0.134	-0.038
n=400	0.055	0.008	-0.031

Dla $n = 100$ mamy większe obciążenia.
$>$ Wyestymuję macierz kowariancji wektora estymatorów $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ i porównam z asymptotyczną macierzą kowariancji.
Macierz kowariancji wektora estymatorów dla $n = 400$ jest postaci:

```
##              [,1]         [,2]       [,3]
## [1,] 4.051885621  0.003245324  0.1254295
## [2,] 0.003245324  4.043459470 -0.1684307
## [3,] 0.125429524 -0.168430747  4.1550707
```

Wyniki są zbliżone do macierzy asymptotycznej kowariancji. Minimalnie większe są różnice dla $n = 100$, ponieważ tam mamy większą wariancję estymowanych parametrów.

Zadanie 3

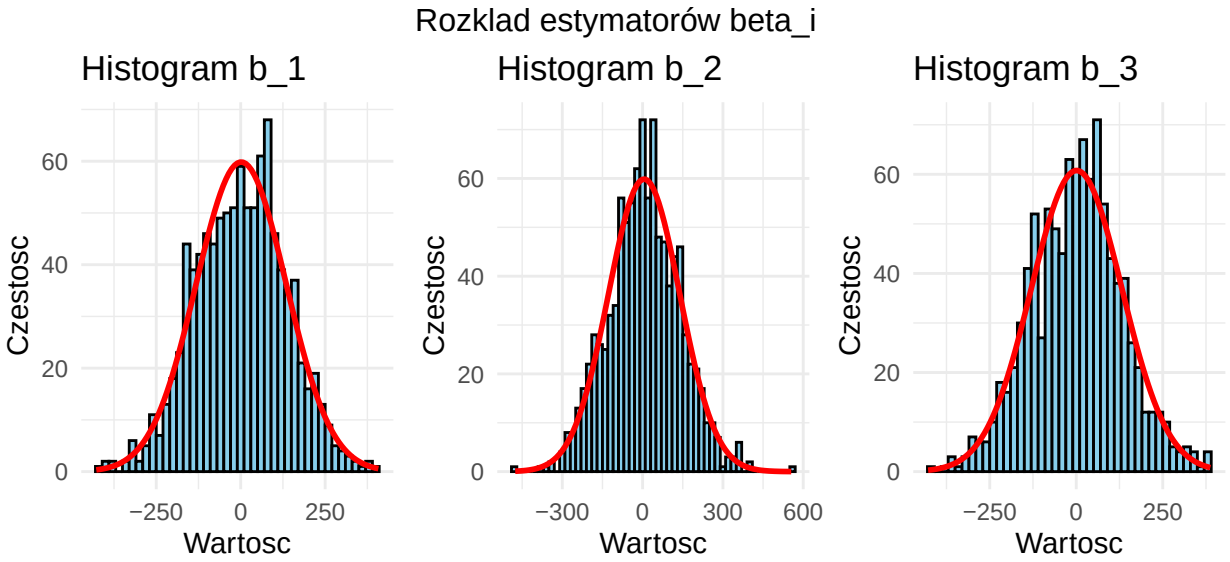
W zadaniu 3 powtórzę punkt 1 w przypadku, gdy wiersze macierzy X są niezależnymi wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$ z macierzą kowariancji $\Sigma = \frac{1}{n}S$, gdzie $S_{ii} = 1$, a dla $i \neq j,\ S_{ij} = 0.3$.

Macierz Fishera jest postaci:

```
##              [,1]          [,2]          [,3]
## [1,]  0.228013379  0.0029682800 -0.0024343741
## [2,]  0.002968280  0.2823964487 -0.0008068508
## [3,] -0.002434374 -0.0008068508  0.2619129611
```

Macierz asymptotycznej kowariancji jest postaci:

```
##             [,1]        [,2]       [,3]
## [1,]  4.38674010 -0.04599311 0.04063127
## [2,] -0.04599311  3.54163440 0.01048290
## [3,]  0.04063127  0.01048290 3.81847224
```



Obciążenie estymatorów:

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
-1.615	2.631	-1.982

Macierz kowariancji wektora estymatorów jest postaci:

```
##            B1          B2          B3
## B1 17777.4301   520.7170   303.0526
## B2   520.7170 17755.3069   820.2622
## B3   303.0526   820.2622 17232.8447
```

Korelacja między regresorami prowadzi do wzrostu wariancji estymowanych parametrów, co można zaobserwować przy wzroście obciążenia. Można również zauważyć różnice w macierzy asymptotycznej kowariancji oraz estymowanej macierzy kowariancji. Histogram wygląda podobnie, jak w przypadku zadania 1 i 2, gdzie nyły niezależne regresory, jednak można zauważyć, że jest dużo bardziej rozciągniety, co wynika ze wzrostu wariancji. W tym zadaniu dobór korelacji był praktycznie zerowy. Zwiększenie korelacji mogłoby prowadzić do problemów z współliniowością danych.

Zadanie 4

W tym zadaniu trzeba powtórzyć punkt 1 w przypadku, gdy wiersze macierzy X są niezależne, a $p = 20$. Analiza modelu wskazuje, że mamy około 10 istotnych współczynników, co stanowi połowę oczekiwanej liczby. Dla pierwszych 10 estymatorów wyniki są podobne do wyników z zadania 1, dla pozostałych wzrasta wariancja estymatorów oraz ich obciążenie.

Wnioski

- Ze wszystkich histogramów mogliśmy odczytać, że estymatory oscylują wokół prawdziwej wartości - 3. W zadaniu 3. rozkład bet wyszedł o wiele szerszy (ich wariancja jest większa). Pokrywały się również wszystkie wyestymowane rozkłady β_i wraz z ich asymptotycznymi rozkładami.
- W zadaniu 3 i 4 wzrosło obciążenie estymatorów, co oznacza, że istnienie korelacjii między regresorami oraz liczba regresorów wpływają na ten parametr.
- Wartości na przekątnych macierzy kowariancji estymatorów oraz asymptotycznej macierzy kowariancji były zbliżone, w zadaniu 3 różnica wartości była nieco większa, ale również same wartości były znacząco większe, niż w poprzednich zadaniach.