Raport 1 ZML

Erwin Jasic

23 marca 2021

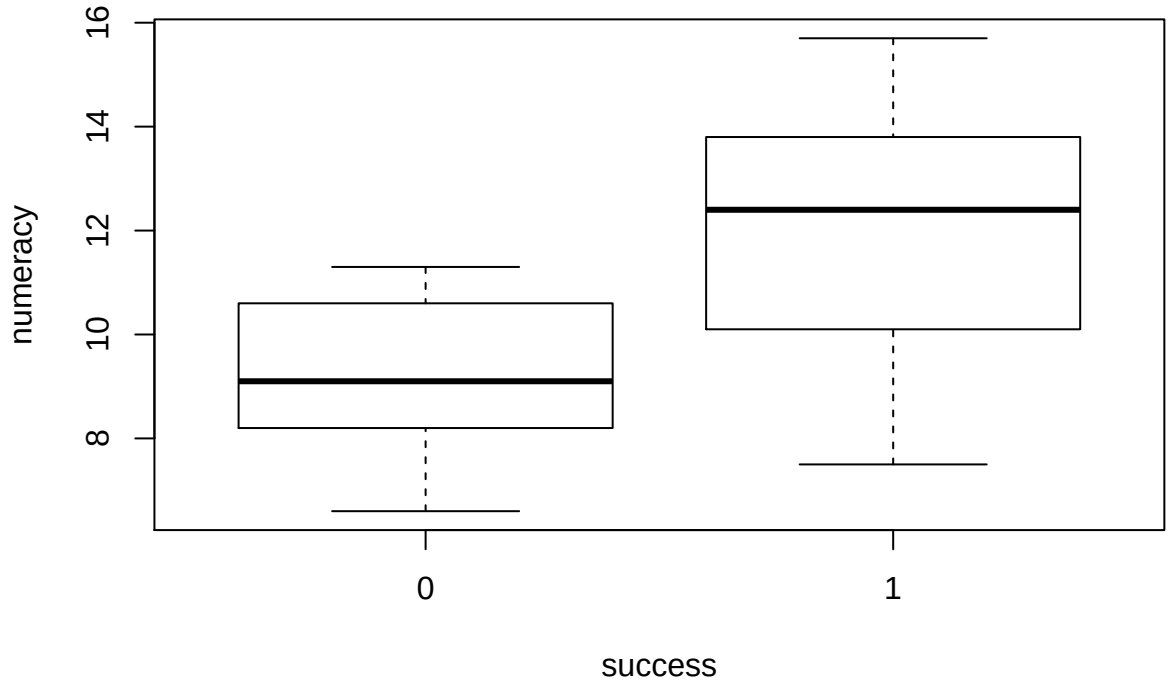
Cel raportu:

W raporcie zajmiemy się zastosowaniem teorii z wykładu dotyczącego regresji logistycznej w praktyce. Porównamy ze sobą różne metody powszechne przy tym zagadnieniu oraz postaramy się wyciągnąć wnioski z otrzymanych wyników.

Analiza danych

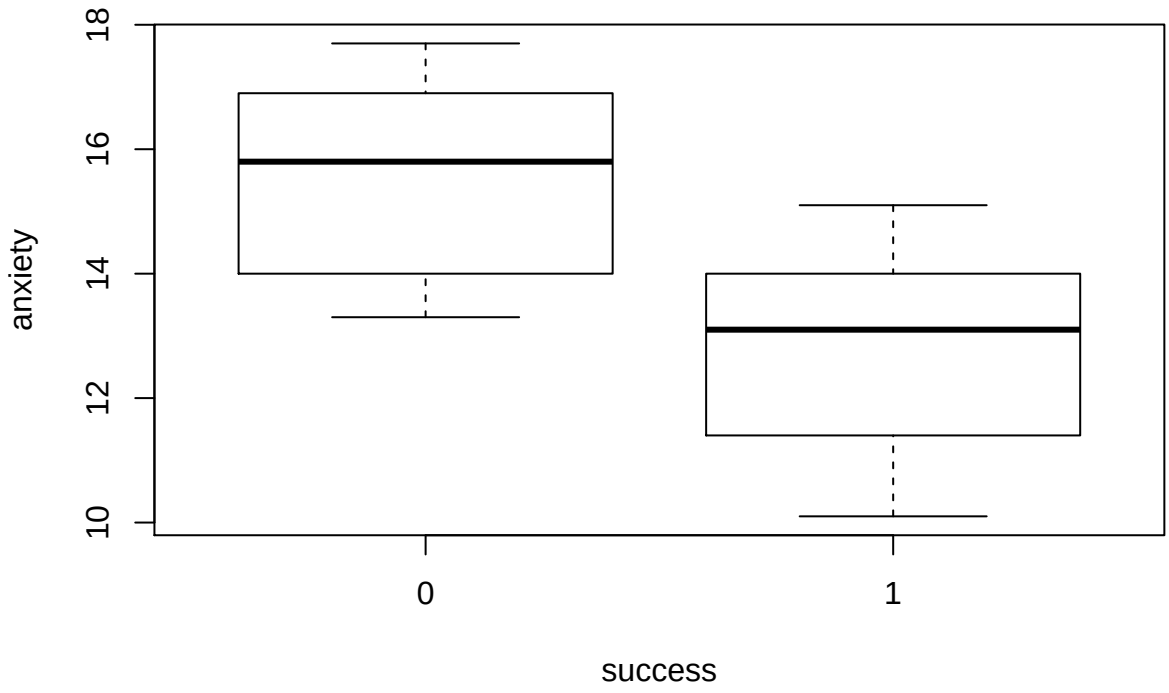
W tym zadaniu skupimy się analizie danych dotyczących studentów przyjętych i nieprzyjętych na studia. Mianowicie, w pliku "Lista_1.csv" znajduje się zbiór danych, który opisuje relacje między prawdopodobieństwami przyjęcia na studia (success), a wynikami z testów rachunkowych (numeracy) oraz poziomu niepewności (anxiety).

Wkonajmy podstawowe czynności, żeby zobaczyć jak wyglądają nasze dane.



Rysujemy boxplot, żeby zobaczyć czy istnieje zależnośc między success, a zmienną numeracy. Widzimy, że

takowa istnieje, ponieważ jeden box jest wyraźnie niżej na wykresie niż ten drugi. Zatem z tego wykresu możemy wyciągnąć wnioski, że im lepszy wynik z testu rachunkowego (numeracy), tym większe prawdopodobieństwo dostania się na uczelnie.



Drugi boxplot opisuje zależność między success, a zmienną anxiety. Tutaj również jest zauważalna wyraźna zależność. Im mniejsze anxiety, tym większe prawdopodobieństwo przyjęcia na studia.

W następnych krokach przeprowadzimy dokładniejszą analizę tych danych. Z wykresów wynika, że anxiety wpływa na success oraz numeracy wpływa na success, ale chcemy się dowiedzieć między innymi, czy obie te zmienne są istotne (możliwe, że jedna zmienna jest wyjaśniana przez drugą).

Przeprowadźmy dokładniejszą analizę zależności między success, a anxiety. W tym celu stworzymy model regresji logistycznej i zobaczymy, czy intercept oraz zmienna anxiety są rzeczywiście istotne w takim modelu.

	estymatory	p-wartości	czy HA zachodzi?
Intercept	19.5819361839021	0.000559916641581152	Tak
anxiety	-1.35557953189789	0.000645606048086719	Tak

Tabela pokazuje, że hipoteza alternatywna zachodzi zarówno dla interceptu jak i dla anxiety (beta 1). P-wartości są bardzo małe, więc te zmienne mają istotny wpływ na success w tym modelu.

Teraz powtórzymy tę analizę zależności, ale dla zmiennej numeracy. W tym celu stworzymy model regresji logistycznej i ponownie zobaczymy, czy intercept oraz zmienna anxiety są rzeczywiście istotne w takim modelu.

	estymatory	p-wartości	czy HA zachodzi?
Intercept	-6.1414134799398	0.00113778832952452	Tak
numeracy	0.624292192238598	0.000762514407240585	Tak

Tutaj również widzimy, że hipoteza alternatywna zachodzi zarówno dla interceptu jak i dla numeracy (beta 1.

P-wartości są bardzo małe, więc te zmienne mają istotny wpływ na success w tym modelu.

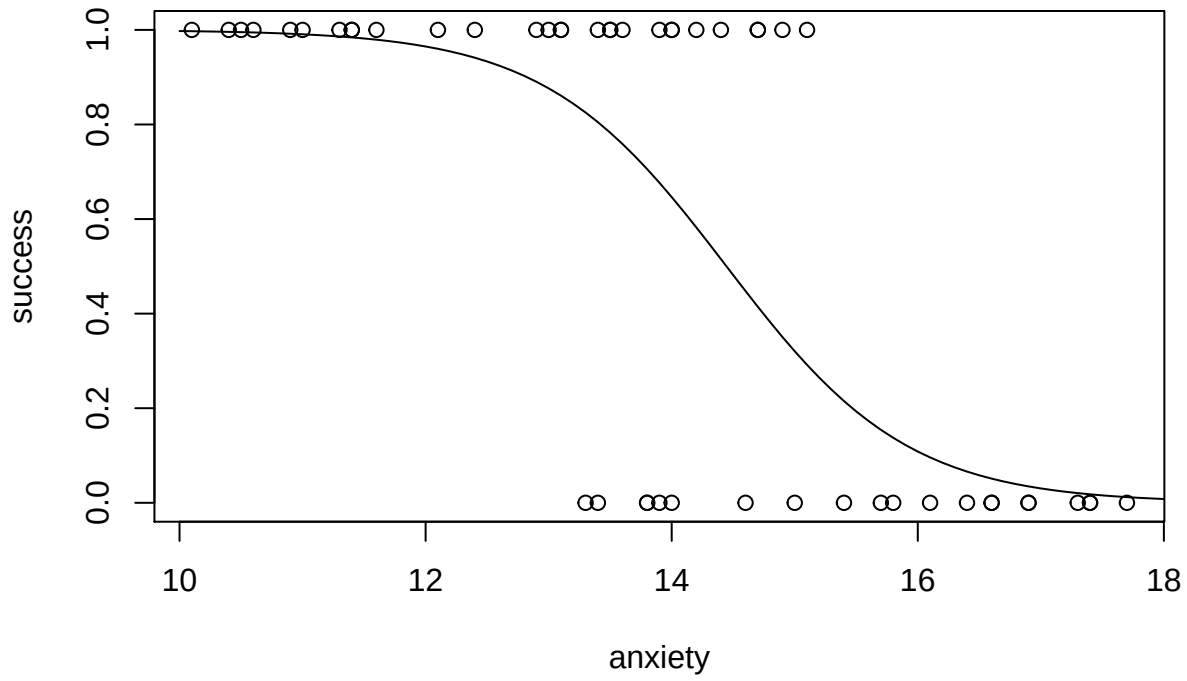
Sprawdźmy teraz jak to będzie wyglądało w pełnym modelu ze wszystkimi zmiennymi.

	estymatory	p-wartości	czy HA zachodzi?
Intercept	14.2385811076387	0.0362274723477621	Tak
numeracy	0.577352046950955	0.0199522732347329	Tak
anxiety	-1.38406900945831	0.00396339819417636	Tak

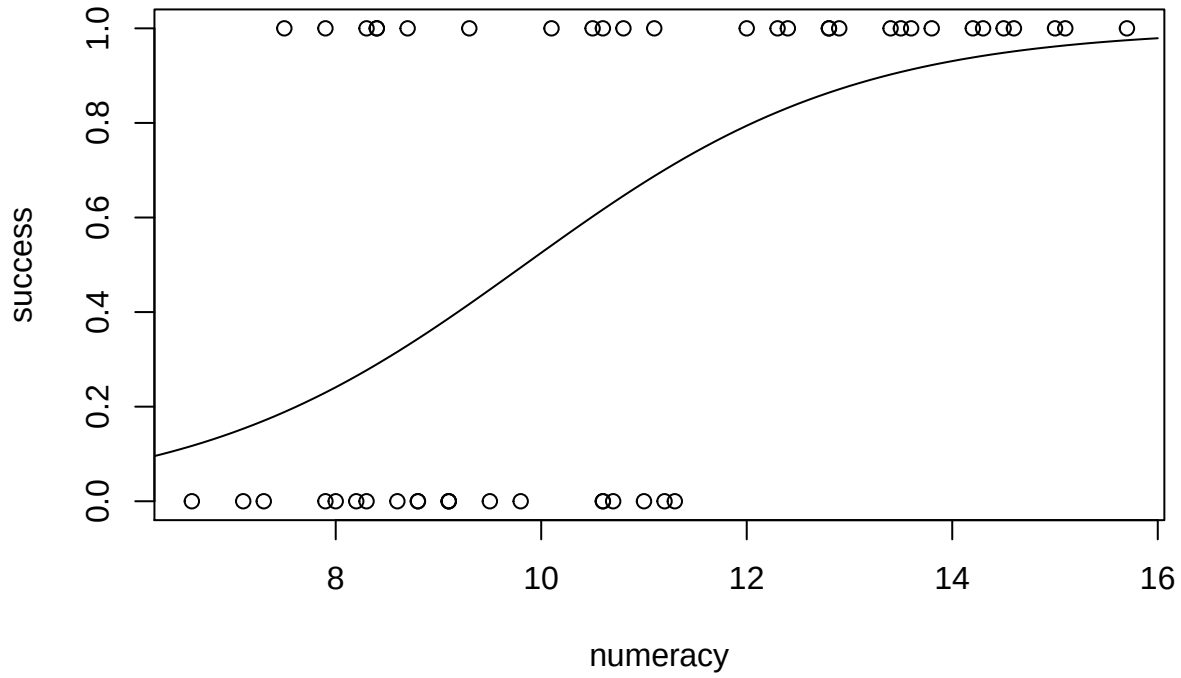
W pełnym modelu również każda zmienna jest istotna na poziomie 95%. P-wartości są trochę większe niż w przypadku modelów z jedna zmienną objaśniającą, ale mimo wszystko na tyle małe, że na poziomie istotności 95%, odrzucają hipotezy zerowe.

Teraz zobaczmy jak to się prezentuje na wykresach.

Ten wykres przedstawia krzywą prawdopodobieństwa, gdy mamy do czynienia z modelem z jedną zmienną anxiety.



Z kolei ten wykres ukazuje krzywą prawdopodobieństwa, kiedy model posiada jedną zmienną numeracy.



Porównując te dwa wykresy możemy dojść do wniosku, że bardziej na przyjecie na studia (success) wpływa zmienna anxiety. Zauważamy to poprzez nachylenie krzywej prawdopodobieństwa. Na górnym wykresie jest ona ostra (stroma), a na dolnym wydaje się na łagodniejszą.

Policzmy dodatkowo jakie prawdopodobieństwo na dostanie się na studia ma osoba, która z testu rachunkowego otrzymała 10 punktów (numeracy = 10) oraz poziom niepewności wynosi u niej 13 (anxiety = 13).

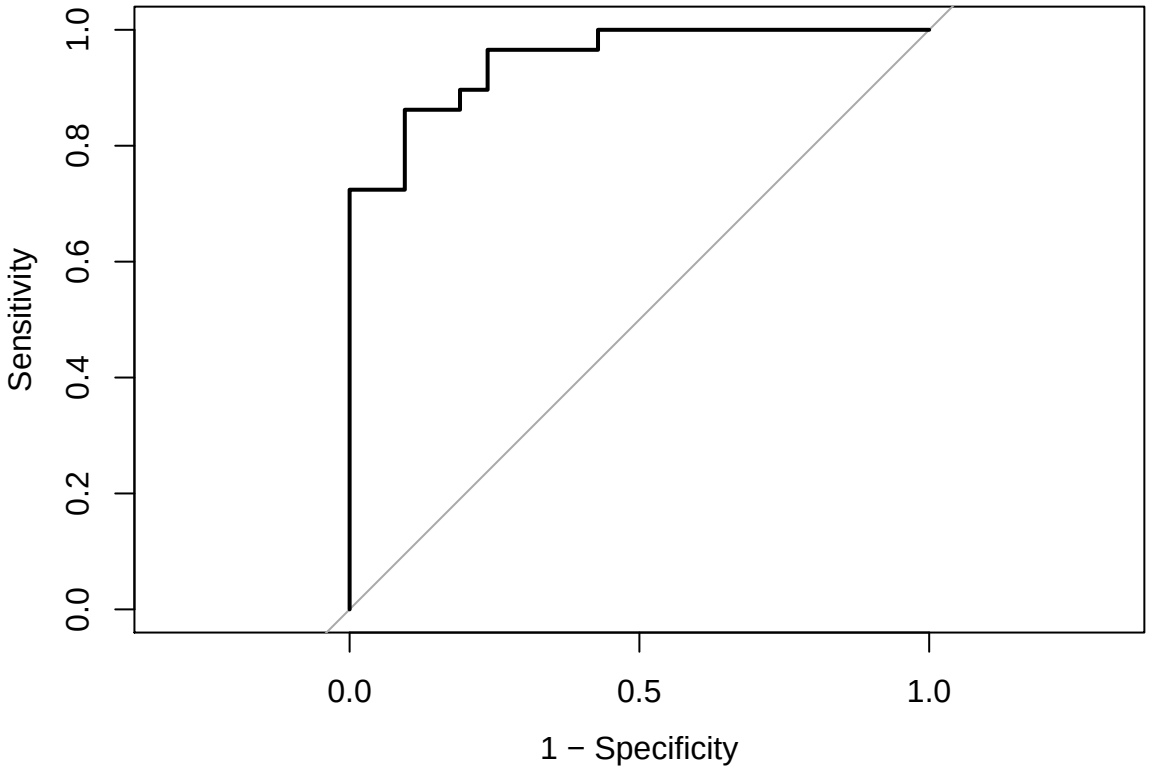
```
##         1
## 0.8827987
```

Prawdopodobieństwo przyjęcia na studia z takimi wynikami jest wysokie, bo wynosi ono ponad 88%.

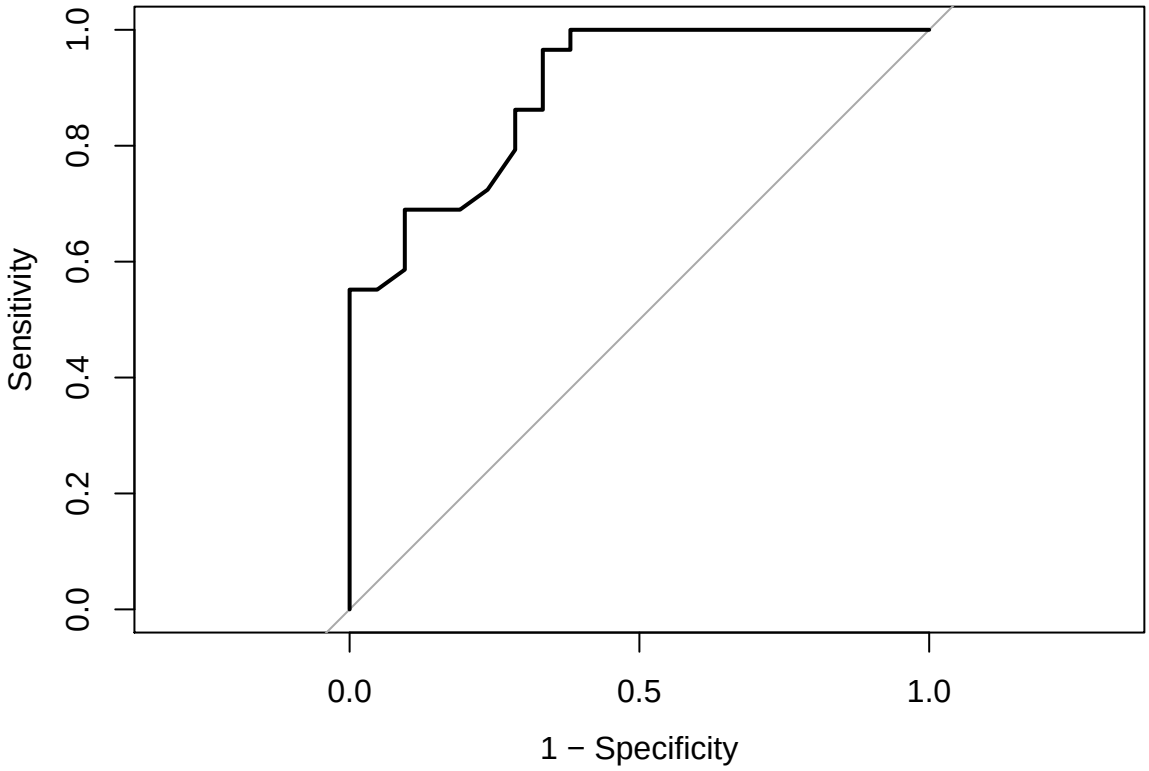
Teraz zaprezentujemy jak wygląda krzywa ROC dla modeli z jedną zmienną oraz w pełnym modelu. Z wykładu wiemy, że im większe pole pod krzywą ROC, tym większa zależność między zmiennymi a zmienną objaśnianą.

```
## Setting levels: control = 0, case = 1
```

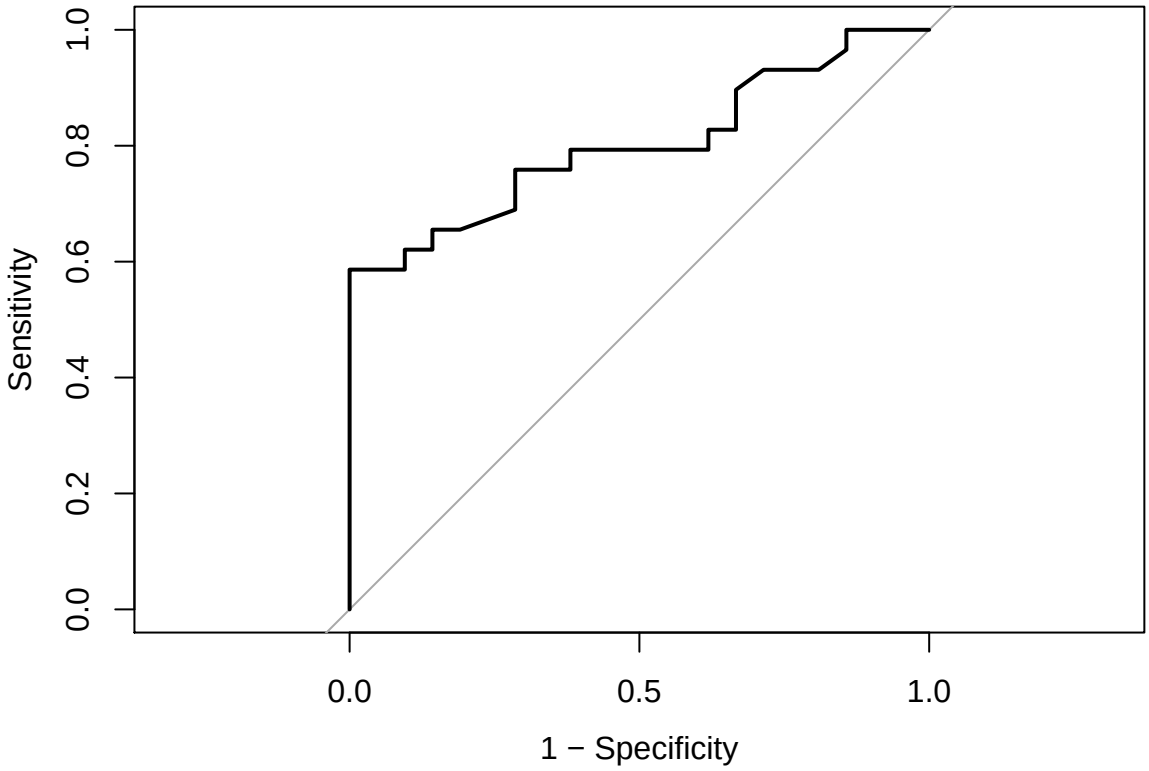
```
## Setting direction: controls < cases
```




```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Tak jak mogliśmy się spodziewać, największe pole po krzywą ROC jest dla modelu pełnego, na drugim miejscu jest model ze zmienną anxiety (podobny wniosek wysnuliśmy na podstawie wykresów z krzywą prawdopodobieństwa) i na ostatnim miejscu model ze zmienną numeracy.

Teraz powtórzymy tę analizę dla róznych funkcji linkujących (probit, cauchit, cloglog).

Tabela (probit, model ze zmienną anxiety)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	11.7188656999143	0.000157850756303377	Tak
beta 1	-0.812038273332444	0.00018184180550088	Tak

Tabela (cauchit, model ze zmienną anxiety)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	19.3599053253719	0.0300205238609546	Tak
beta 1	-1.3247489816864	0.0317496494283829	Tak

Tabela (cloglog, model ze zmienną anxiety)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	12.5743216430078	0.00103889239564819	Tak
beta 1	-0.905762806276495	0.000974967759386304	Tak

Tabela (probit, model ze zmienną numeracy)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	-3.76788423964788	0.000394474591795683	Tak
beta 1	0.384165400709649	0.000203019237983459	Tak

Tabela (cauchit, model ze zmienną numeracy)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	-5.54423668540842	0.0243347380525979	Tak
beta 1	0.551090282135641	0.0237323259050512	Tak

Tabela (cloglog, model ze zmienną numeracy)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	-4.87837628968975	0.00020220328998475	Tak
beta 1	0.448332044783392	0.000175653563422144	Tak

Tabela (probit, model pełny)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	8.25725513325112	0.0246616624148335	Tak
beta 1	0.337104041956133	0.0137377954571997	Tak
beta 2	-0.8038678239762	0.00141776350868082	Tak

Tabela (cauchit, model pełny)

	estymatory	p-wartości	czy HA zachodzi?
Intercept	18.3829650819089	0.135011663633679	Nie
beta 1	0.732273887978906	0.122428220067293	Nie
beta 2	-1.77407876596497	0.0735350484509499	Nie

Tabela (cloglog, model pełny)

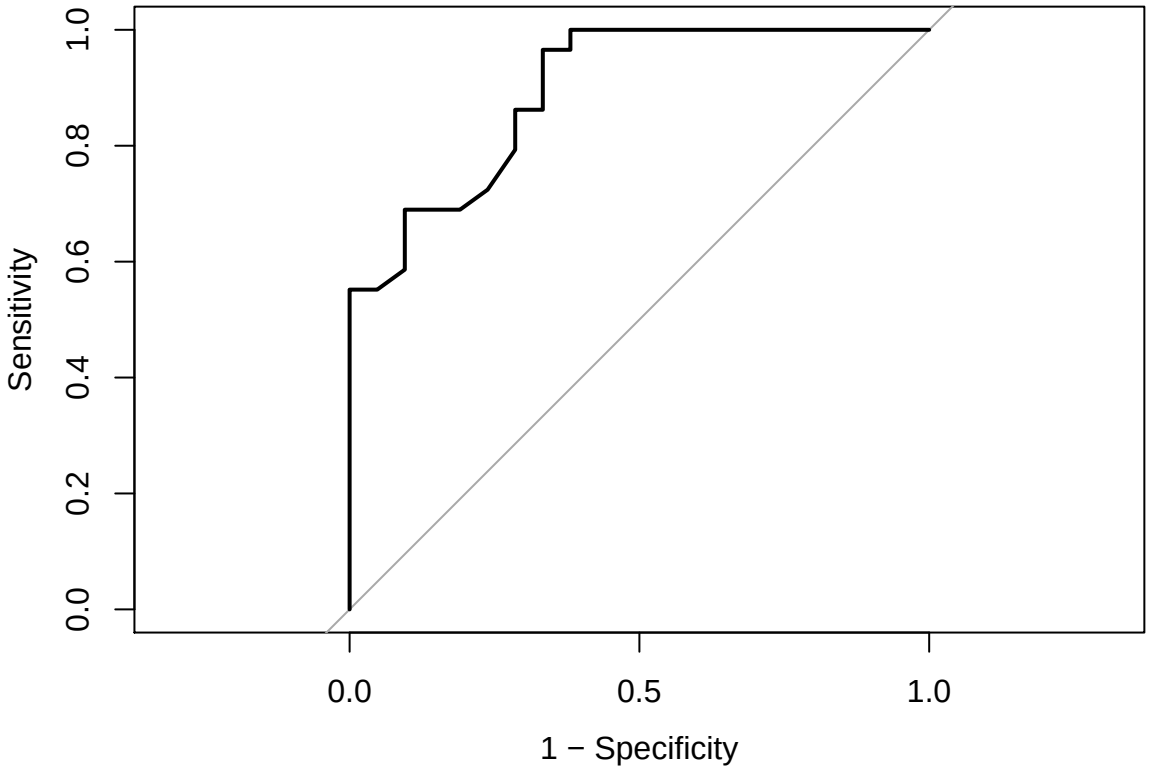
	estymatory	p-wartości	czy HA zachodzi?
Intercept	9.0005810101381	0.0530398559082637	Nie
beta 1	0.402424995015592	0.00819235500531036	Tak
beta 2	-0.938974805302969	0.00470235690067275	Tak

Porównując ze sobą te tabele, możemy zauważyć, że jeżeli mamy jedną zmienną to zawsze zachodzą hipotezy alternatywne, ale w modelu pełnym już jest inaczej. Widzimy, że tylko 'probit' dla każdej zmiennej odrzuca hipotezę zerową.

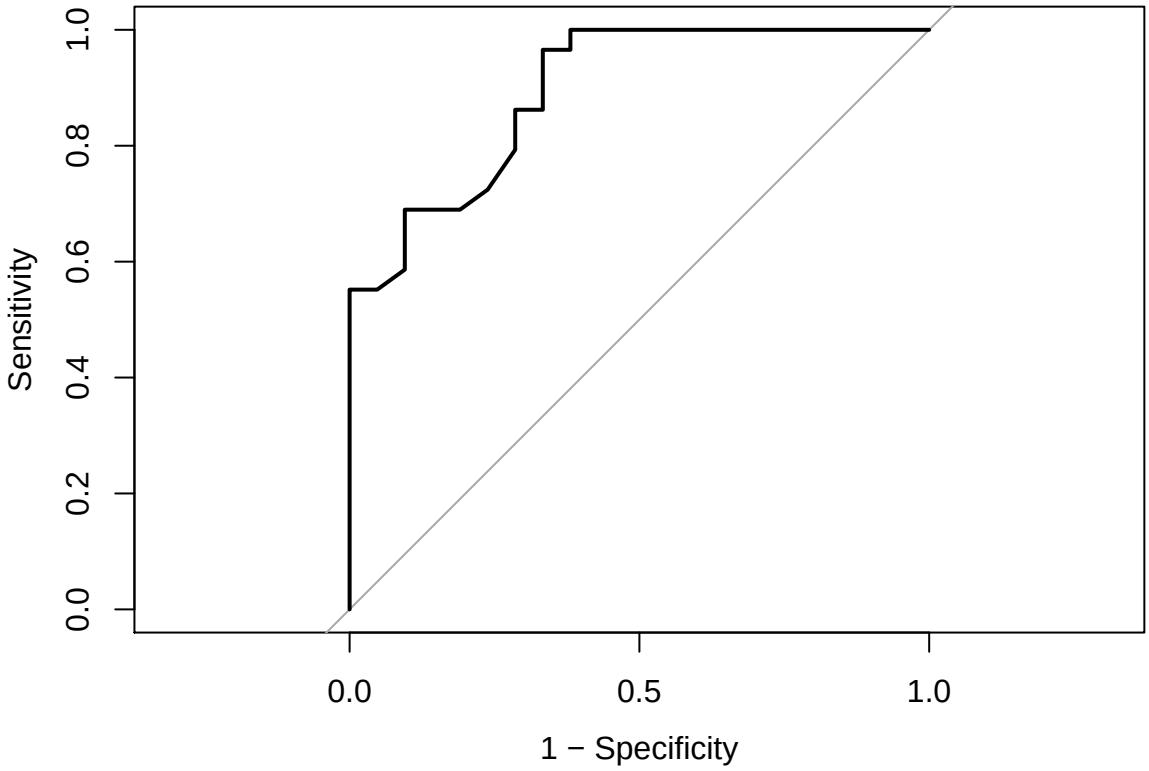
Zobaczmy jak zachowują się krzywe ROC dla róznych modeli dla zmiennej anxiety.

```
## Setting levels: control = 0, case = 1
```

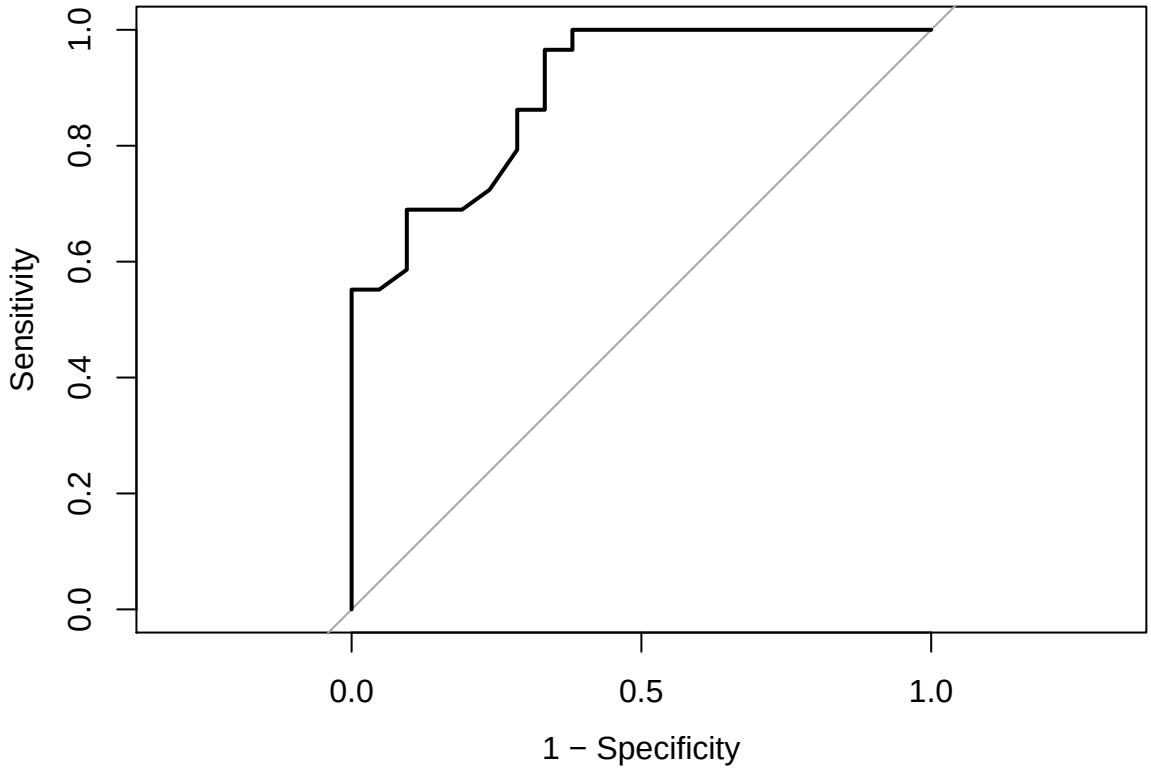
```
## Setting direction: controls < cases
```




```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Wszystkie wychodzą identycznie.

Teraz skupimy się na analizie modelu z funkcją linkującą logit.

Wyznaczmy estymatory odchyleń standardowych:

```
## (Intercept)    numeracy     anxiety
##   6.7985192   0.2480840   0.4804027
```

Jeszcze zobaczmy czym jest parametr epsilon w funkcji glm.

```
##
## Call:  glm(formula = success ~ numeracy + anxiety, family = "binomial",
##     data = data_l1, epsilon = 10^(-1))
##
## Coefficients:
## (Intercept)     numeracy      anxiety
##     12.8901       0.5376      -1.2640
##
## Degrees of Freedom: 49 Total (i.e. Null);  47 Residual
## Null Deviance:       68.03
## Residual Deviance: 28.37     AIC: 34.37
```

```
##
## Call:  glm(formula = success ~ numeracy + anxiety, family = "binomial",
##     data = data_l1, epsilon = 10^(-2))
##
## Coefficients:
## (Intercept)     numeracy      anxiety
```

```
##     14.0925       0.5735      -1.3713
##
## Degrees of Freedom: 49 Total (i.e. Null);  47 Residual
## Null Deviance:       68.03
## Residual Deviance: 28.29     AIC: 34.29
```

```
##
## Call:  glm(formula = success ~ numeracy + anxiety, family = "binomial",
##     data = data_l1, epsilon = 10^(-3))
##
## Coefficients:
## (Intercept)     numeracy      anxiety
##     14.2368       0.5773      -1.3839
##
## Degrees of Freedom: 49 Total (i.e. Null);  47 Residual
## Null Deviance:       68.03
## Residual Deviance: 28.29     AIC: 34.29
```

```
##
## Call:  glm(formula = success ~ numeracy + anxiety, family = "binomial",
##     data = data_l1, epsilon = 10^(-6))
##
## Coefficients:
## (Intercept)     numeracy      anxiety
##     14.2386       0.5774      -1.3841
##
## Degrees of Freedom: 49 Total (i.e. Null);  47 Residual
## Null Deviance:       68.03
## Residual Deviance: 28.29     AIC: 34.29
```

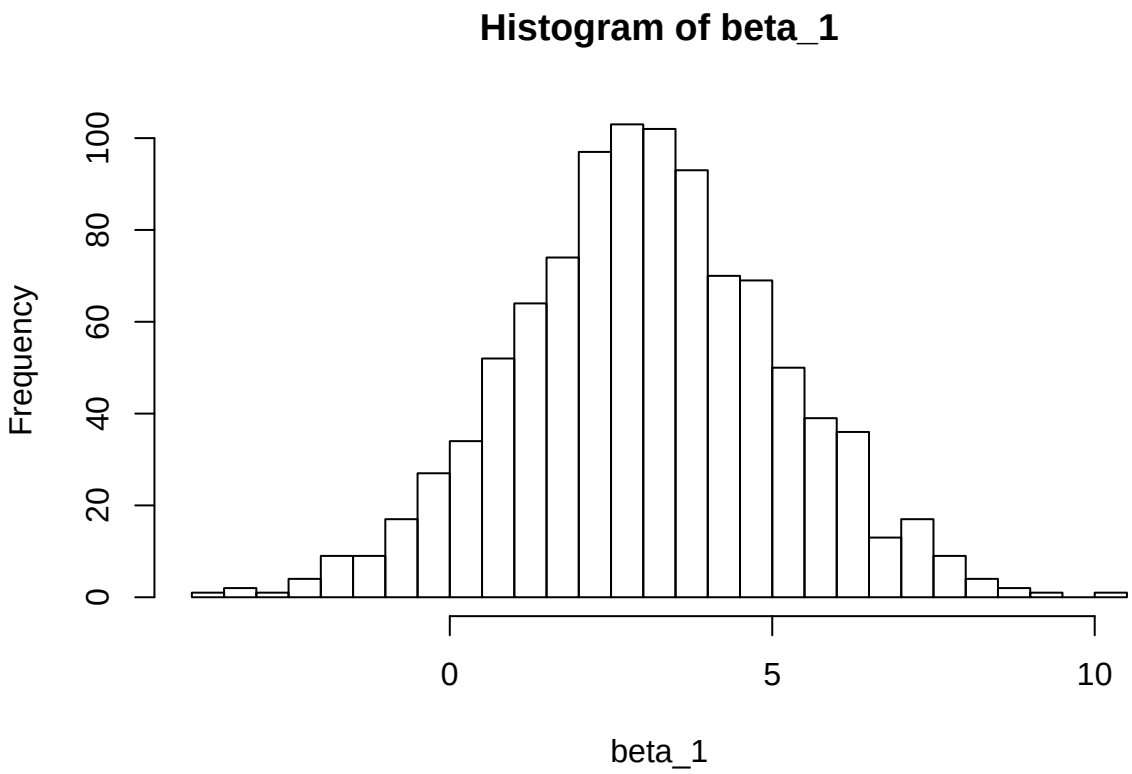
Widzmy, że jeśli zmieniamy epsilon to estymatory też się zmieniają. Im mniejszy epsilon tym dokładniejsze estymatory. Przypuszczamy, że domyślna wartość epsilon w funckcji glm to 10^{-6}, ponieważ z takim parametrem jak i bez niego (wartość domyślna) otrzymujemy takie same estymatory.

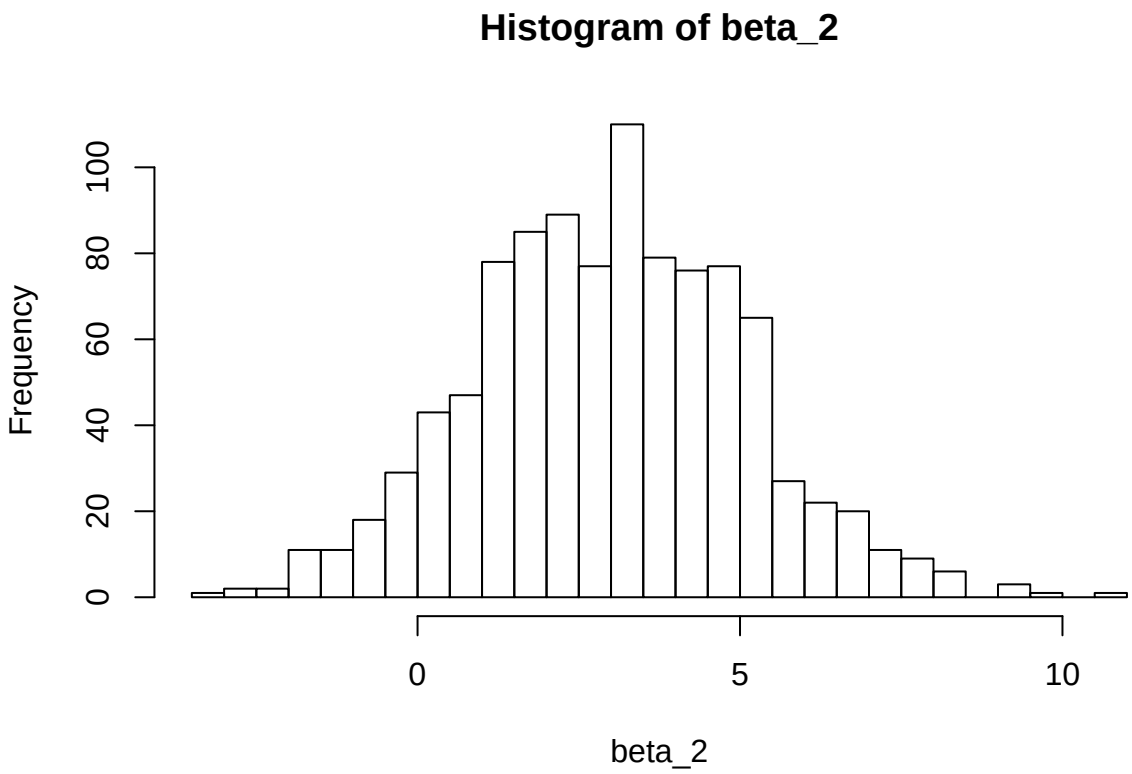
Teraz powiemy sobie jak się to ma do liczby iteracji. Mianowicie im mniejszy parametr epsilon tym więcej jest iteracji. W naszym zadaniu dla epsilon = 10^{-1} mamy 3 iteracje, a już dla epsilon = 10^{-6} otrzymujemy 6 iteracji.

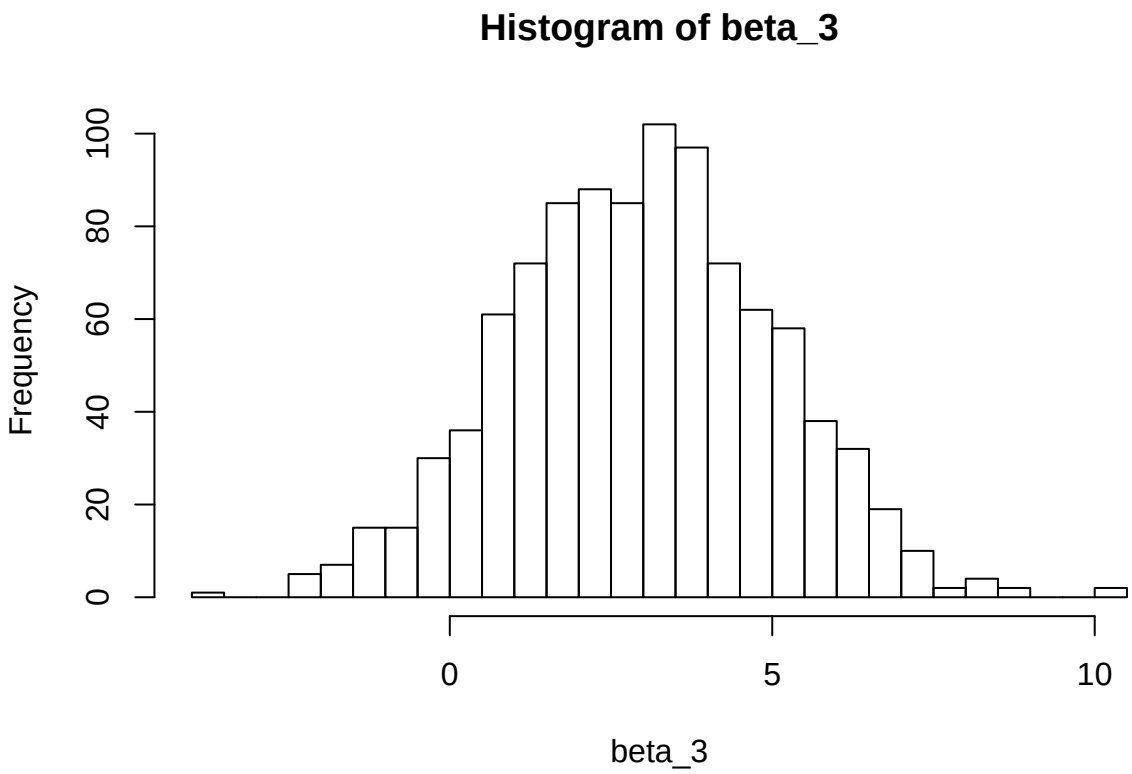
Symulacje

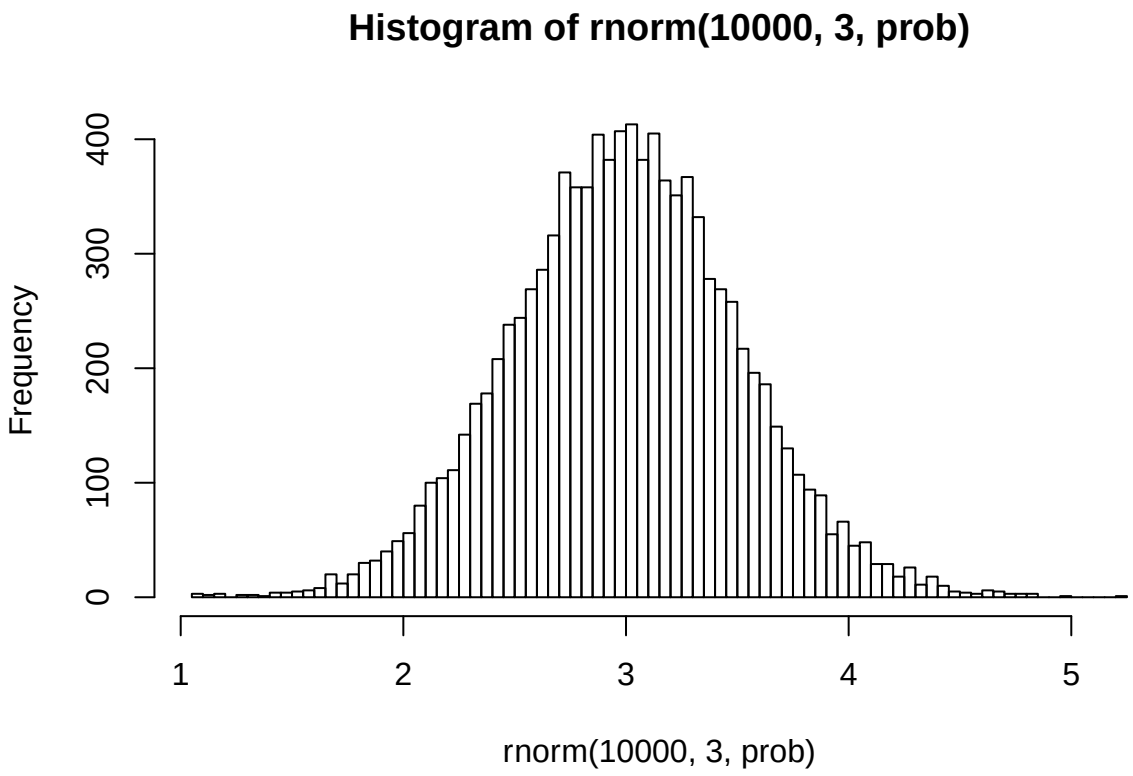
W tym zadaniu mamy wygenerować macierz planu o wymierach n na p i na podstawie uzyskanego zbioru estymatorów przeprowadzić analizę. Zbadamy wpływ liczby obserwacji n, wpływ korelacji między regresorami oraz wpływ liczby regresorów.

Doświadczenie przeprowadzamy dla n=400, p=3 oraz $\sigma = 0.05$.


Histogram of beta_1
Frequency
0
20
40
60
80
100
0
5
10
beta_1


Histogram of beta_2
Frequency
0
20
40
60
80
100
0
5
10
beta_2


Histogram of beta_3
Frequency
0
20
40
60
80
100
0
5
10
beta_3

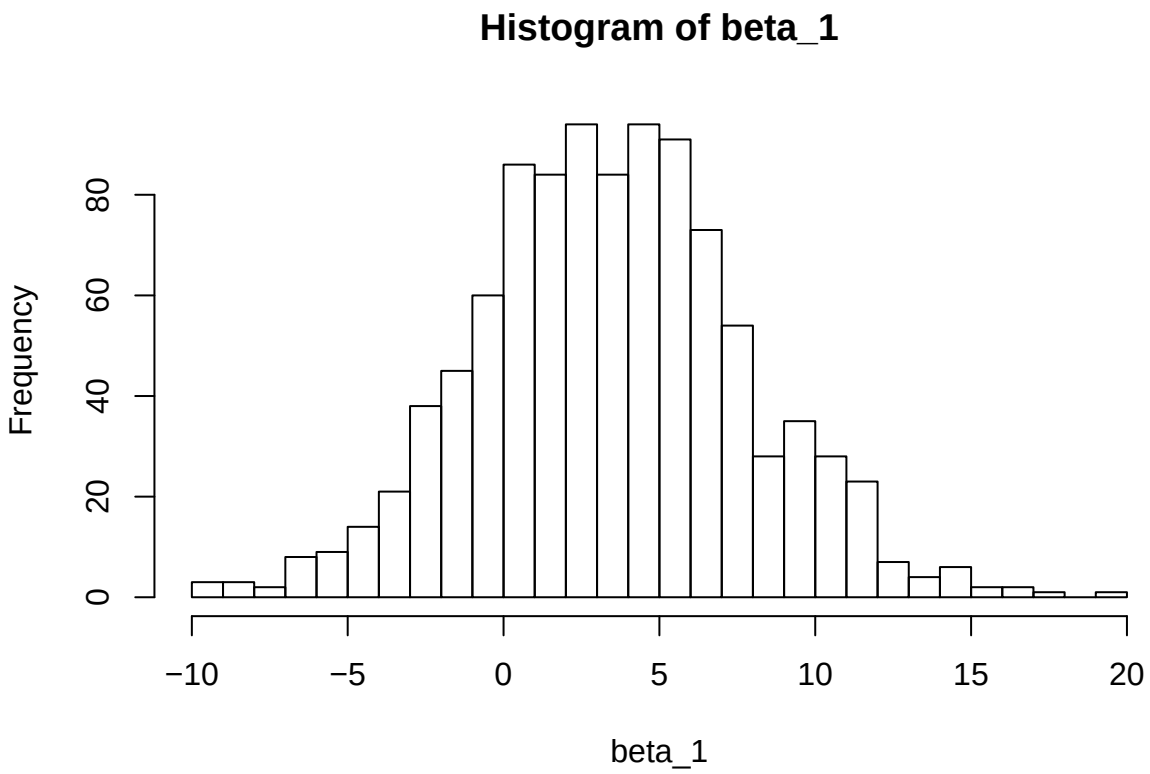


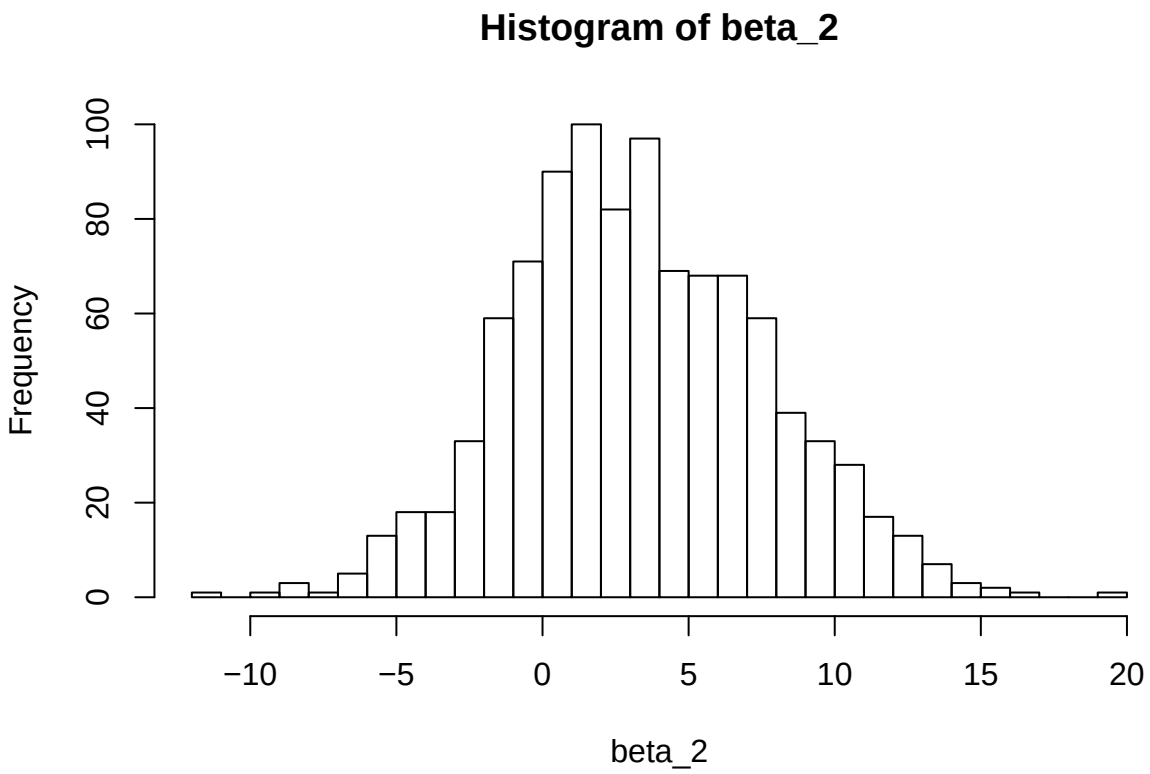

```
## [1] -55.97373
```

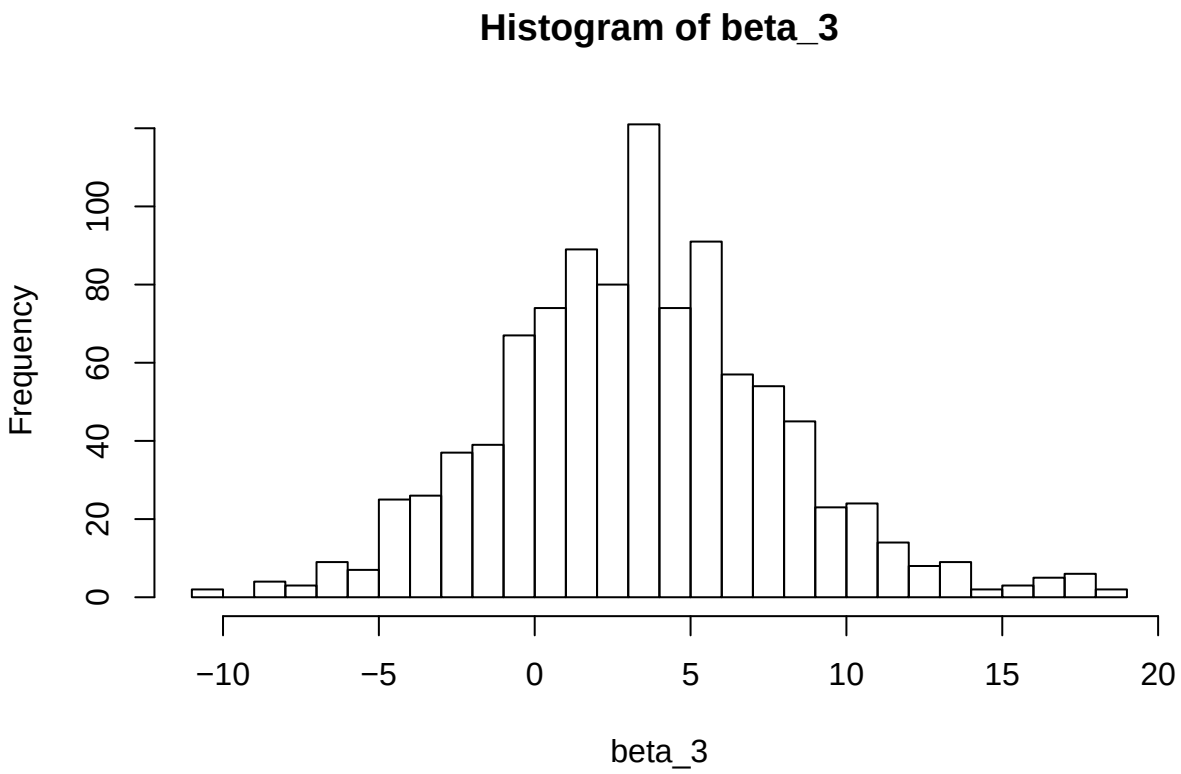
```
## [1] -6.750867
```

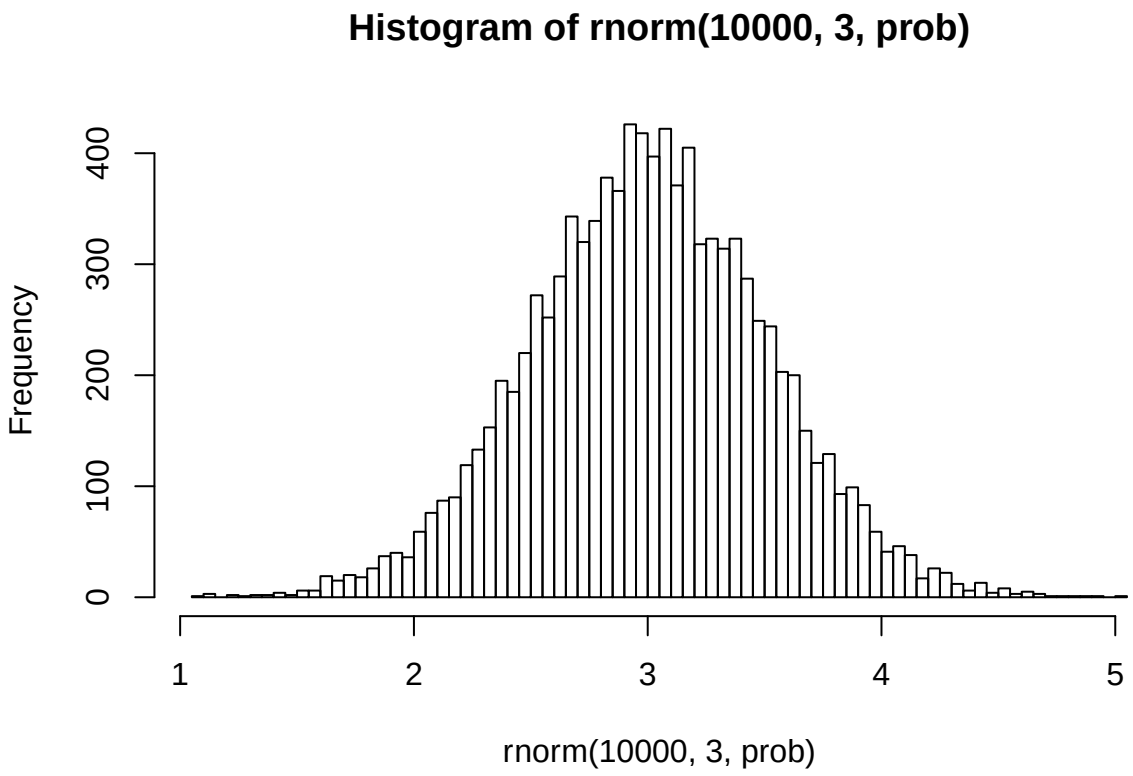
```
## [1] 33.09231
```

Powtarzamy doświadczenie dla n=100, pozostałe parametry bez zmian.


Histogram of beta_1
Frequency
0
20
40
60
80
−10
−5
0
5
10
15
20
beta_1


Histogram of beta_2
Frequency
0
20
40
60
80
100
−10
−5
0
5
10
15
20
beta_2


Histogram of beta_3
Frequency
0
20
40
60
80
100
−10
−5
0
5
10
15
20
beta_3



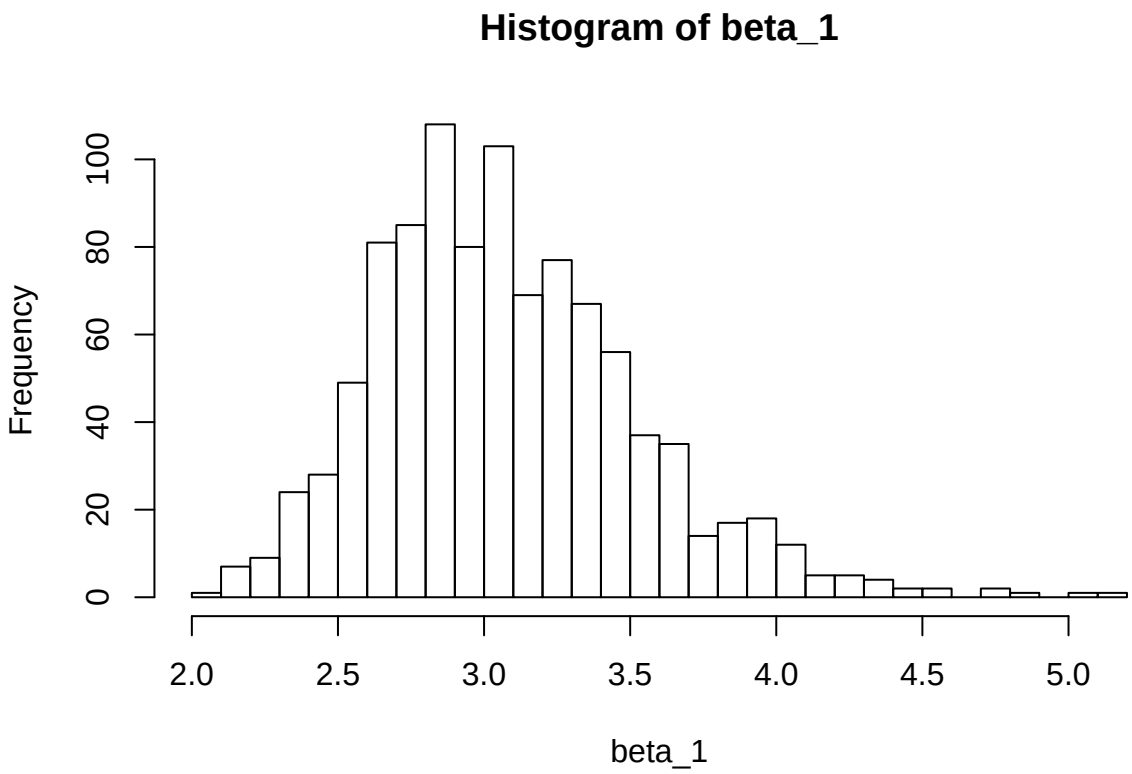
Obciążenie estymatorów beta:

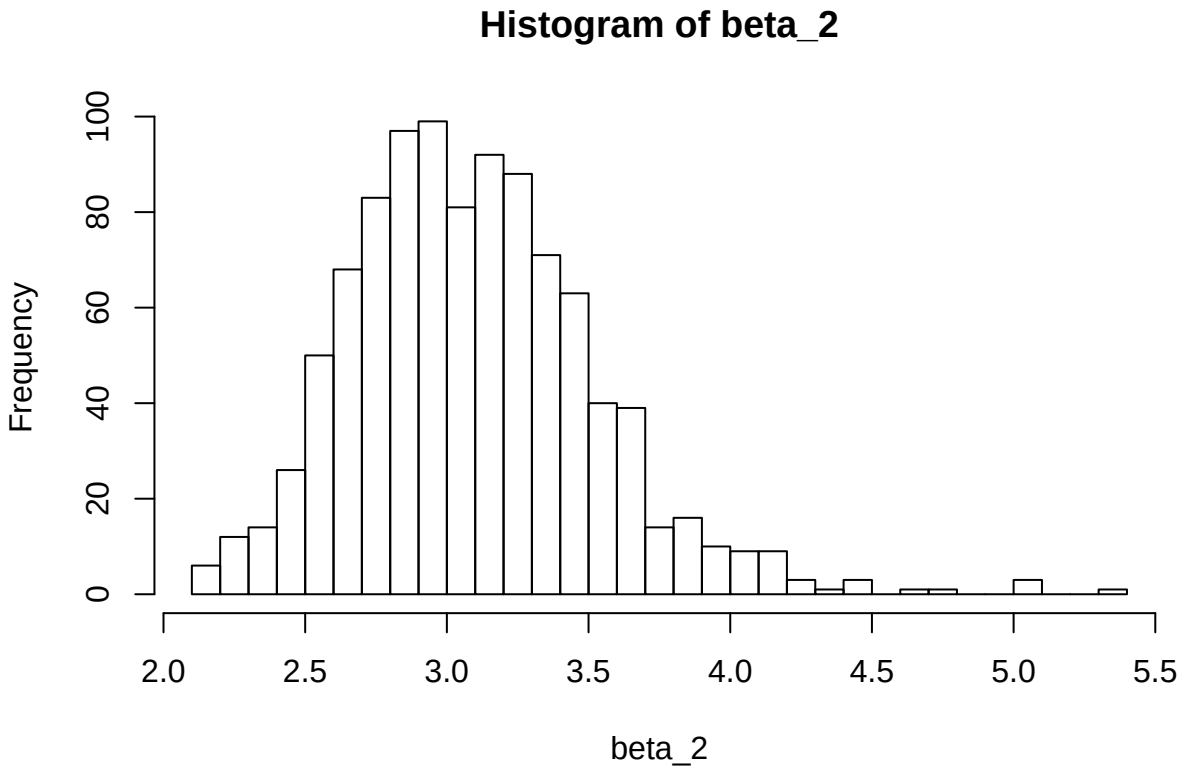
```
## [1] -464.1608
```

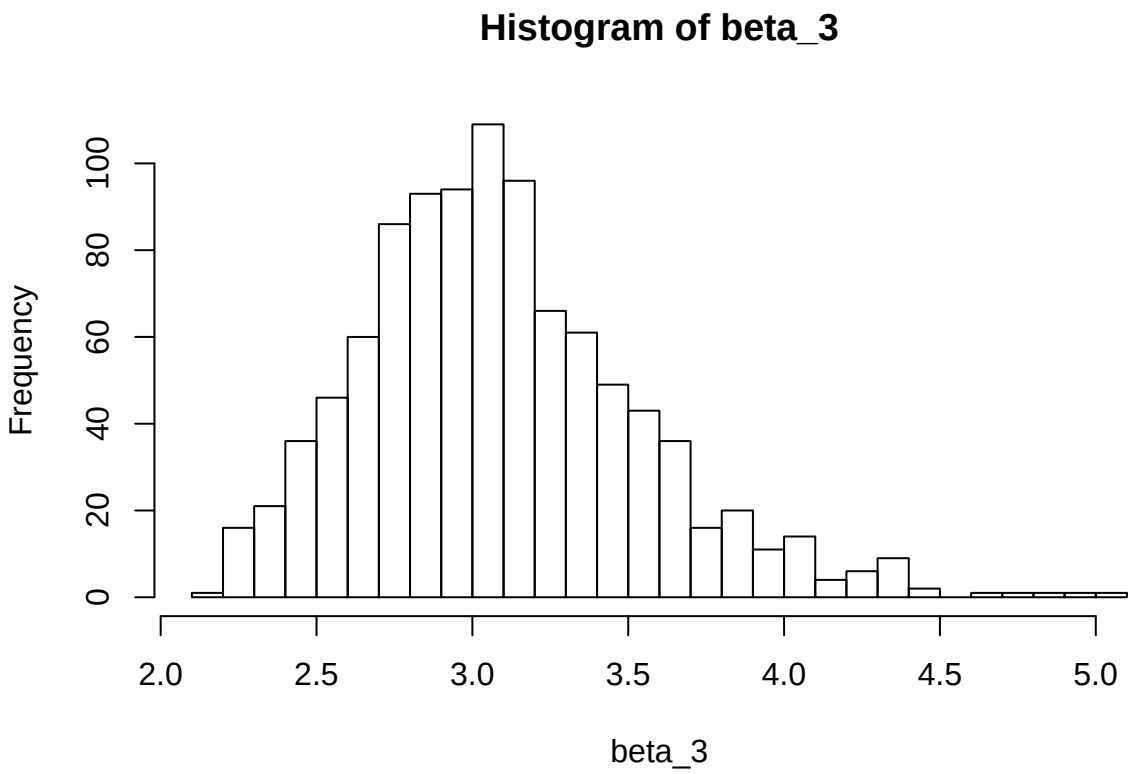
```
## [1] -308.4713
```

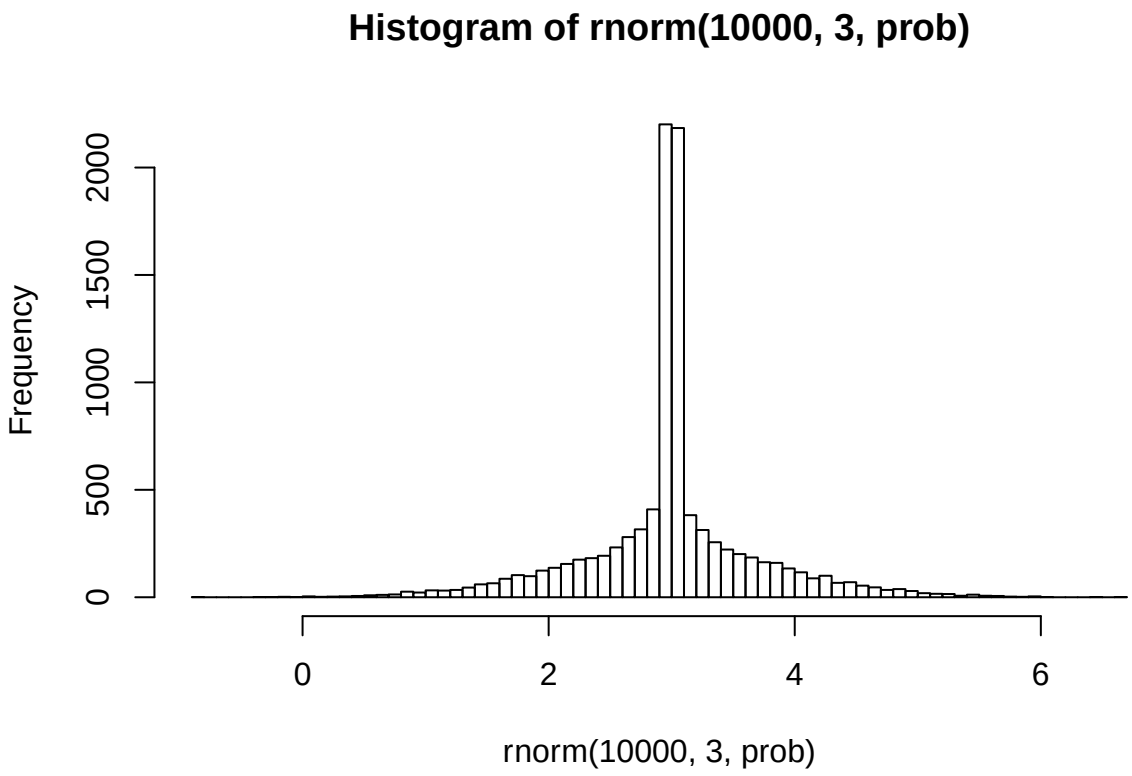
```
## [1] -381.14
```

Teraz wykonujemy to doświadczenie dla parametrów takie jak w punkcie 1, ale wiersze macierzy X będą pochodzić z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$ z macierzą kowariancji $\Sigma = \frac{1}{n}S$, gdzie $S_{ii} = 1$, a dla $i \neq j$, $S_{ij} = 0.3$.


Histogram of beta_1
Frequency
0
20
40
60
80
100
2.0
2.5
3.0
3.5
4.0
4.5
5.0
beta_1


Histogram of beta_2
Frequency
0
20
40
60
80
100
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
beta_2


Histogram of beta_3
Frequency
0
20
40
60
80
100
2.0
2.5
3.0
3.5
4.0
4.5
5.0
beta_3



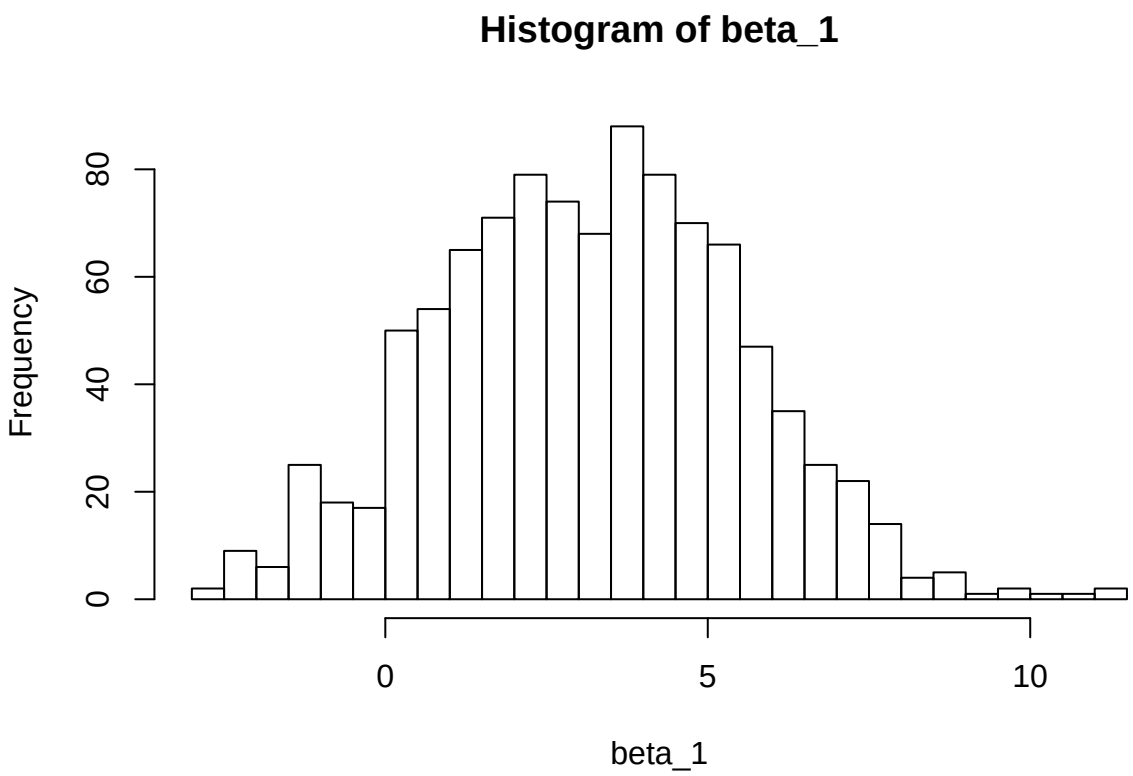
Obciążenie estymatorów beta:

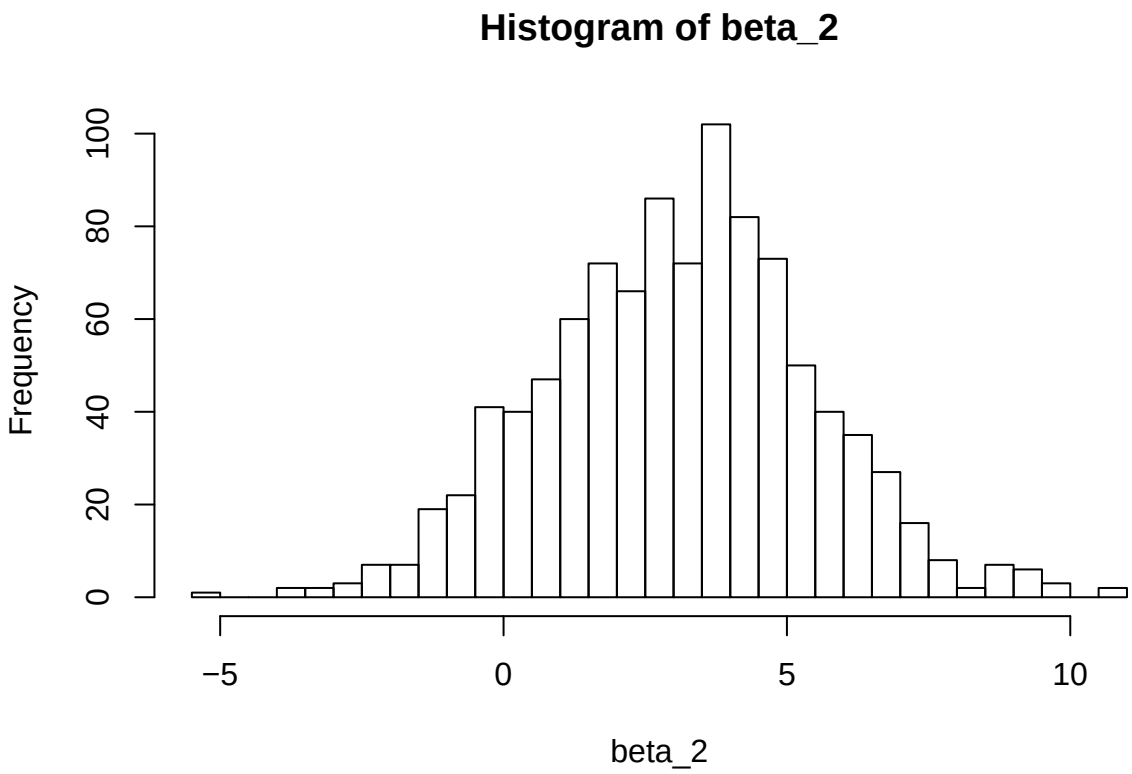
```
## [1] -77.26527
```

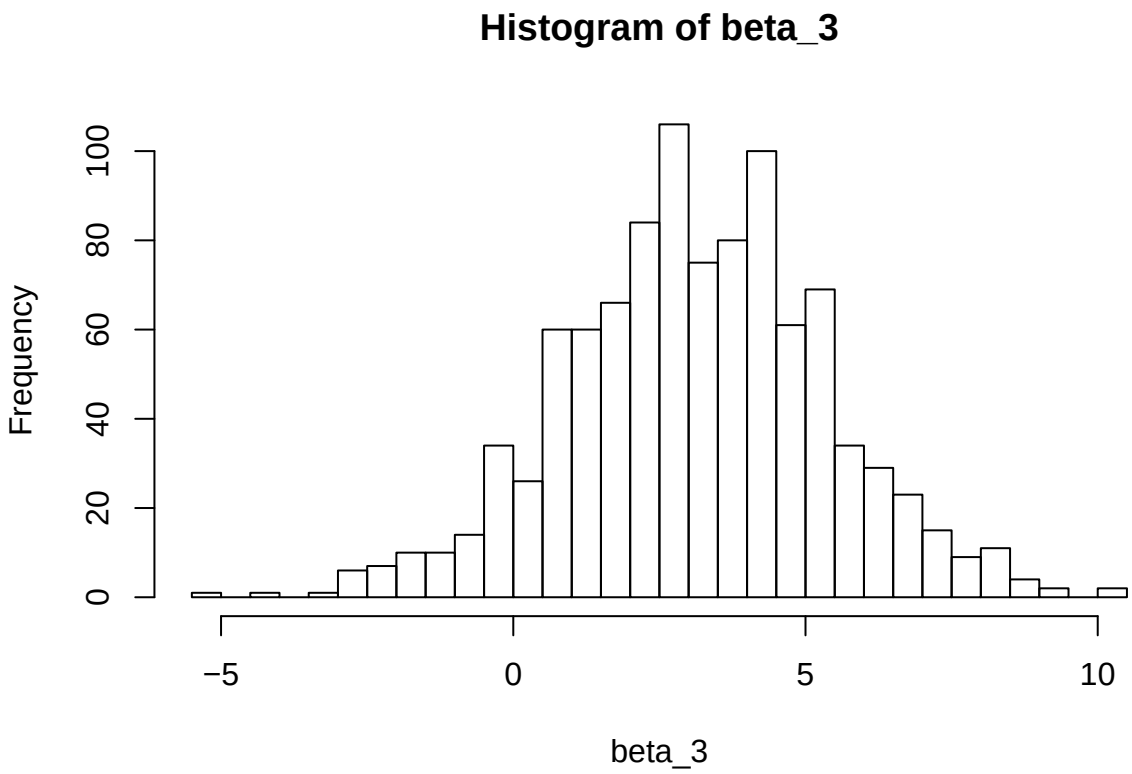
```
## [1] -90.32703
```

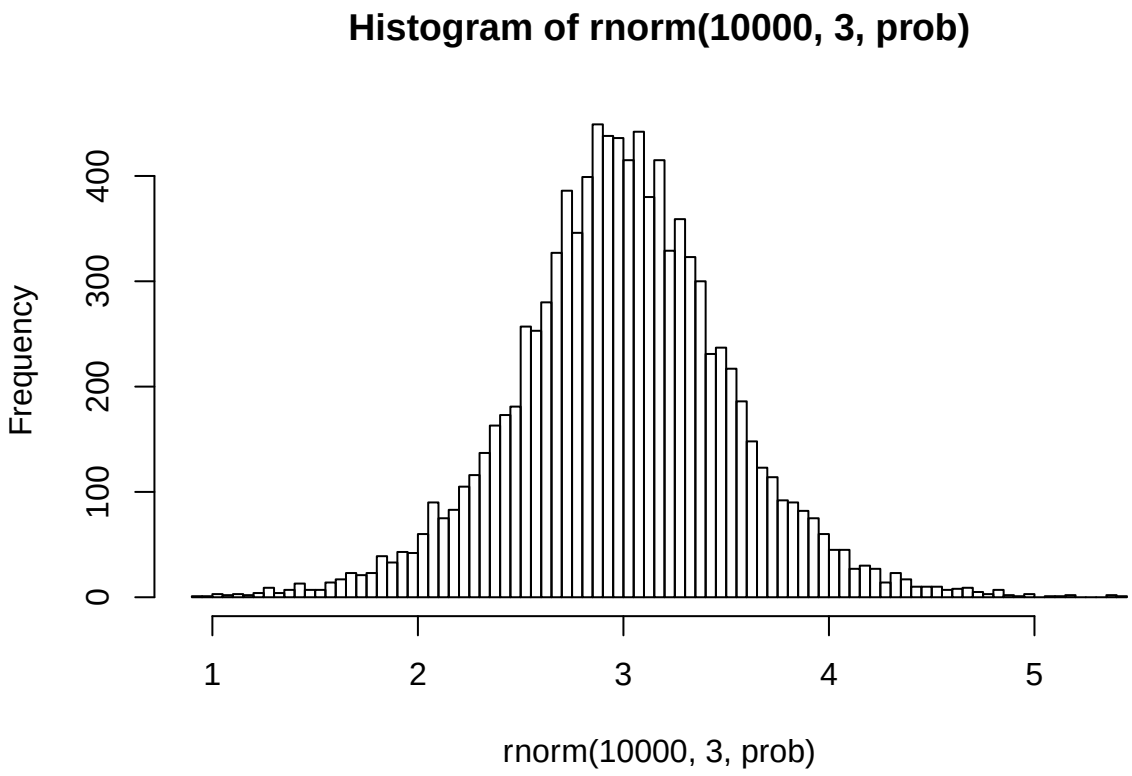
```
## [1] -89.63998
```

Na koniec powtarzamy doświadczenie z parametrami z podpunktu pierwszego, ale p=20.


Histogram of beta_1
Frequency
0
20
40
60
80
0
5
10
beta_1


Histogram of beta_2
Frequency
0
20
40
60
80
100
−5
0
5
10
beta_2


Histogram of beta_3
Frequency
0
20
40
60
80
100
−5
0
5
10
beta_3



Obciążenie estymatorów beta:

```
## [1] -190.1751
```

```
## [1] -95.48275
```

```
## [1] -123.9528
```

Z histogramów ciężko wysnuć jakieś konkretne wnioski na piewszy rzut oka, bo wszystkie wyglądają podobnie, ale różnią się rozpiętością na osi x. Widzimy, że jeśli mamy n=100 to rozpiętość histogramu jest większa niż w przypadku n=400. Dla podponktu z macierzą kowariancji Σ praktycznie wszystkie obserwacje estymatorów beta mieszczą się na przedziale od 1.5 do 4.5 (mała rozpiętość). Histogramy dla p=20 i p=3 wyglądaja podobnie.

Popatrzmy też na obciążenie estymatorów beta. Dla n=100 dużo większe obciązenie niż dla n=400. Dla doświadczenia z macierzą kowariancji Σ mamy nieco większe obciążenia niż w podpunkcie 1, a dla p=20 mamy większe obciążenie dla bet niż w przypadku p=3.

Podsumowanie:

Dzięki wykonaniu poleceń dotyczących regresji logitycznej widzimy, że mamy dużo możliwości analizowania danych pochodzących z takiego modelu oraz dzięki krzywej ROC możemy łatwo porównywać ze sobą modele pod względem dopasowania do danych. Poprzez wykonanie symulacji zobaczyliśmy jak niektóre parametry mają wpływ na rozkład estymatorów beta oraz jak to się ma do ich obciążenia.