ZML - lista1: Regresja logistyczna

Funkcje linkujące, ROC, testowanie hipotez o parametrach, macierz informacji Fishera

Wstęp

Badamy zależność pomiędzy binarną zmienną $Y \in \{0,1\}^n$ a macierzą zmiennych rzeczywistych X rozmiaru $n \times p$. Niech $\mu_i = P(Y_i = 1)$. Model regresji logistycznej opisuje równanie

$$f(\mu_i) = \beta_0 + \beta_1 X_{i,1} + ... + \beta_{p-1} X_{n,p-1},$$

gdzie $f(\mu) : [0,1] \to R$ i f nazywamy **funkcją linkującą**. Standardowo używamy $f(\mu) = logit(\mu) = \log \frac{\mu}{1-\mu}$ i to właśnie od nazwy te (Uwaga: Nazwa "regresja logistyczna" powiązana jest właśnie z użyciem funkcji *logit*. W przypadku innych funkcji linkujących mówimy o uogólnionej regresji liniowej [GLM = generalized linear model] lub używamy nazwy funkcji f. Problem jej wyboru zostanie opisany w dalszej części wstępu). Parametry estymujemy korzystając z algorytmów do optymalizacji wypukłej (nie istnieje wzór na $\hat{\beta}$.) Dla danego zestawu obserwacji jesteśmy w stanie przewidzieć prawdopodobieństwo sukcesu nakładając na wartość $f(\mu)$ odpowiednią funkcję odwrotną (np. $logit(x)^{-1} = sigmoid(x) = \frac{e^x}{e^x+1}$. Jeśli zamiast prawdopodobieństwa chcemy otrzymać wartość zmiennej Y (problem klasyfikacji), ustalamy pewien próg t i przypisujemy Y_i wartość 0, gdy $\mu_i < t$ lub 1 gdy $\mu \geq t$. Różne progi będą skutkować różnymi stosunkami błędów I i II rodzaju w testowaniu hipotez postaci $H_0 : Y = 0$ vs $H_1 : Y = 1$. Do wizualizacji wielu możliwych progów na jednym wykresie stosuje się **krzywą ROC** (Reviever-Operator Curve). Wprowadzamy oznaczenia:

- $TP = \#\{i : \hat{Y}_i = 1, Y_i = 1\}$ (True Positive),
- $FP = \#\{i : \hat{Y}_i = 1, Y_i = 0\}$ (False Positive),
- $TN = \#\{i : \hat{Y}_i = 0, Y_i = 0\}$ (True Negative),
- $FN = \#\{i : \hat{Y}_i = 0, Y_i = 1\}$ (False Negative).

Niech **czułość (True Positive Rate)** $TPR := \frac{TP}{TP+FN}$ oraz **specyficzność (True Negative Rate)** $TNR := TNTN + FP$. Krzywą ROC zwyczajowo rysujemy zaznaczając na osi poziomej wartości $1 - TNR$ (False Negative Rate), a na pionowej TPR dla różnych progów t. W zależności od konkretnego problemu możemy wybrać różne cele optymalizacji t - w ogólności korzystne będzie znalezienie progu odpowiadającego punktowi na krzywej możliwie blisko $(0,1)$. Klasyfikatory można dodatkowo oceniać porównując pole pod krzywą ROC, tzw. AUC (Area Under the Curve) - im większe jest to pole, tym lepszy punkt na wykresie jesteśmy znaleźć.

Możemy testować istotność współczynników β_i. Niech $S(\beta)$ będzie macierzą diagonalną $n \times n$, taką że $S_{i,i} = \mu_i(\hat{\beta}) \cdot (1 - \mu_i(\hat{\beta}))$, gdzie $\mu_i(\hat{\beta})$ to wartość μ_i wyliczona dla konkretnego estymatora $\hat{\beta}$. Zauważmy, że S jest macierzą **teoretycznej kowariancji zmiennej** Y przy założeniu niezależności obserwacji. Ponadto, postać S zależy od wyboru funkcji linkującej. Wektor $\hat{\beta}$ ma asymptotycznie rozkład $N(\beta, J^{-1})$, gdzie

$$J = X^T S(\beta) X.$$

Macierz J nazywamy **macierzą informacji Fishera**. Można w tym punkcie zauważyć analogię do rozkładu estymatora parametrów w regresji liniowej $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$.

Hipotezę postaci $H_{0,i} : \beta_i = 0$ vs $H_{1,i} : \beta_i \neq 0$ testujemy z użyciem statystyki testowej **o rozkładzie zbiegającym do** $N(0,1)$

$$T = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)},$$

gdzie $s(\hat{\beta}_i) = \sqrt{J^{-1}_{i,i}}$ (testowanie innych hipotez opisano w dalszej części). Wyznaczone eksperymentalnie **asymptotyczne przedziały ufności** są postaci

$$\hat{\beta}_i \pm t_c s(\hat{\beta}_i),$$

gdzie $t_c = \Phi^{-1}_{N(0,1)}(1 - \frac{\alpha}{2})$ (uwaga: w regresji liniowej przy wyznaczaniu empirycznych przedziałów używaliśmy rozkładu Studenta.)

Testowanie hipotez w oparciu o statystykę Deviance

Przy modelach liniowych miarą dopasowania modelu do danych była suma kwadratów residuów $RSS = ||\hat{Y} - Y||_2^2$. W modelu regresji logistycznej podobną funkcę pełni statystyka *Deviance* zdefiniowana jako:

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

Można zauważyć, że jest to podwojona suma wyrażeń postaci $\log(\tilde{\mu}_i)^{-1} = -log(\tilde{\mu}_i)$, gdzie $\tilde{\mu}_i$ jest przewidzianym przez model prawdopodobieństwem sukcesu, gdy y_i jest sukcesem lub przewidzianym przez model prawdopodobieństwem porażki, gdy y_i jest porażką. Dla $\tilde{\mu} \in (0, 1)$ funkcja $-\log(\tilde{\mu})$ przyjmuje wartości dodatnie, tym mniejsze im większe jest przewidziane prawdopodobieństwo, zatem gdy model często przewiduje duże prawdopodobieństwa sukcesu dla sukcesów i porażki dla porażek, wartość *Deviance* będzie mała.

Alternatywnie statystykę można wyrazić jako podwojoną sumę funkcji log-wiaogodnosci pomiędzy modelem saturowanym (s) a zredukowanym (r), tzn.

$$D(\hat{\beta}) = 2[l(\hat{\beta}^{(s)}) - l(\hat{\beta}^{(r)})],$$

gdzie przez model zredukowany rozumiemy rozważany przez nas model, dla którego

$$l(\hat{\beta}^{(r)}) = \sum_{i=1}^{n} y_i \log(\hat{\mu}_i(\hat{\beta})) + (1 - y_i)\log(1 - \hat{\mu}_i(\hat{\beta})),$$

a przez model saturowany rozumiemy hipotetyczny model w którym liczba parametrów równa się liczbie obserwacji ($p = n$), a więc jesteśmy w stanie uzyskać idealne dopasowanie do danych a funkcja log-wiarygodności przybiera postać w której wszystkie $\mu_i(\hat{\beta})$ we wzorze powyżej zamieniają się na y_i.

Statystyka *Deviance* jest używana do testowania różnych hipotez statystycznych. Rozważmy niepusty podzbiór $A \subset \{0, ...p\}$ i hipotezę postaci

$$H_0 : \forall(i \in A)\beta_i = 0 \quad \text{vs} \quad \exists(i \in A)\beta_i \neq 0.$$

Niech $\hat{\beta}^{(k)}$ oznacza model stowarzyszony z hipotezą H_k. Wówczas statystyka

$$\chi^2 = D(\hat{\beta}^{(0)}) - D(\hat{\beta}^{(1)})$$

ma przy hipotezie zerowej asymtotycznie rozkład $\chi^2_{|A|}$ z $|A|$ stopniami swobody, a więc odrzucimy H_0 gdy $\chi^2 > t_c$, gdzie t_c jest kwantylem rozkładu $\chi^2_{|A|}$ rzędu α. Zagadnienie to będziemy nazywać **równoczesnym testowaniem istotności wielu parametrów**. Uwaga: wartości zwracane przez `summary` w R to odpowiednio:

- `Null Deviance`: $D(\beta^{(0)})$ gdzie H_0 zakłada, że wszystkie współczynniki **oprócz wyrazu wolnego** β_0 są zerowe.
- `Residual Deviance`: $D(\beta^{(1)})$ gdzie H_1 zakłada, że żaden współczynnik nie jest zerowy.

Wtedy $A = \{1, ...p\}$, a więc statystyka $\chi^2_{|A|}$ ma rozkład z liczbą stopni swobody równej liczbie kolumn macierzy planu. Wiedzę tę wykorzystamy w zadaniu 6.

Różne funkcje linkujące

Logit

Ze względu na wygodne własności matematyczne nazywana **kanoniczną funkcją linkującą** i używana domyślnie. Występujące w niej wyrażenie $\frac{\mu}{1-\mu}$ można interpretować jako stosunek prawdopodobieństwa wygranej do porażki - wielkość ta ma duże znaczenie w hazardzie, gdzie znana jest pod nazwą **odds** ("szanse"). Wartości zmieniają się od $-\inf$ do $\inf$, symetrycznie względem 0.

Probit

Załóżmy, że podejrzewamy że zmienna Y_i jest "obcięciem" regresji liniowej opisanej równaniem $\tilde{Y}_i = -X_i \cdot \beta + \epsilon_i$ z błędem losowym $\epsilon \sim N(0,1)$, tzn. że

$$Y_i = \begin{cases} 1, & \text{gdy } \tilde{Y}_i \leq 0 \\ 0, & \text{gdy } \tilde{Y}_i > 0 \end{cases}$$

$$P(Y_i = 1) = P(\tilde{Y}_i \leq 0) = P(\epsilon_i \leq X_i \cdot \beta) = \Phi(X_i \cdot \beta) \implies X_i \cdot \beta = \Phi^{-1}(\mu_i)$$

Występujący po lewej stronie iloczyn skalarny odpowiada kombinacji liniowej kolumn macierzy planu w regresji liniowej, natomiast wyrażenie po prawej stronie odpowiada nałożeniu na μ_i funkcji kwantylowej rozkładu normalnego standardowego. W celu uzyskania wartości μ_i na kombinację liniową nałożymy funkcję odwrotną, czyli dystrybuantę $\Phi_{N(0,1)}$. Taki model nazywamy regresją probitową.

Cloglog

Analogiczne rozumowanie przeprowadzimy w przypadku, gdy podejrzewamy, że zmienna Y_i powstała jako "obcięcie" zmiennej o rozkładzie Poissona $\tilde{Y}_i \sim Poisson(\lambda_i)$, gdzie $\lambda_i = e^{X_i \cdot \beta}$. Definiujemy następujące obcięcie:

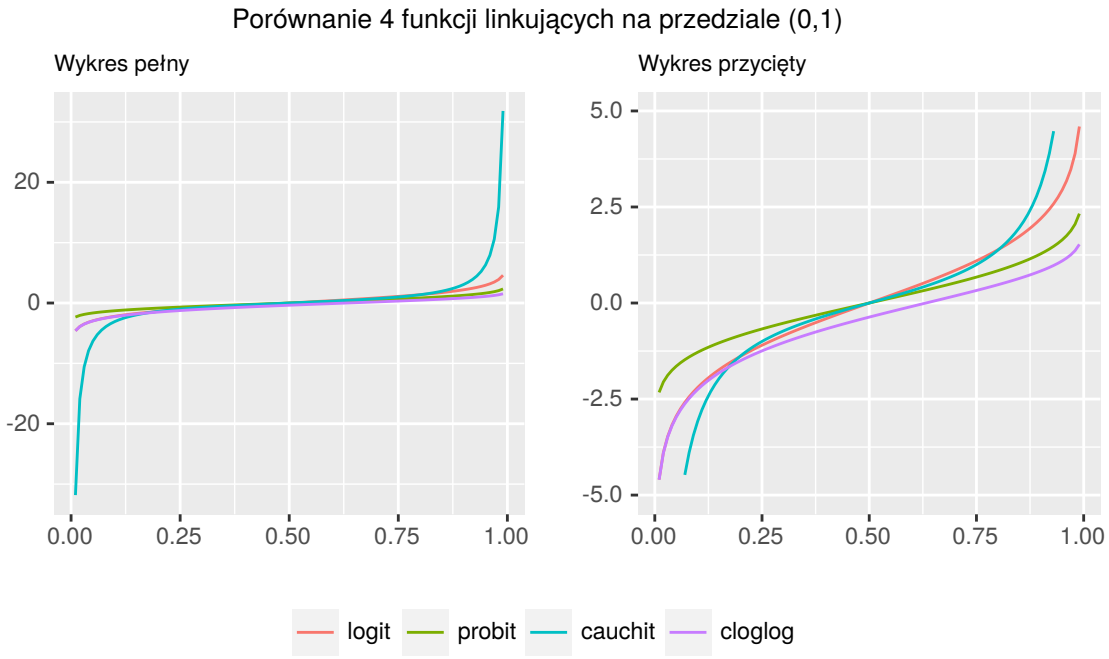
$$Y_i = \begin{cases} 1, & \text{gdy } \tilde{Y}_i \geq 0 \\ 0, & \text{gdy } \tilde{Y}_i = 0 \end{cases}$$

$$P(Y_i = 1) = P(\tilde{Y}_i \geq 0) = 1 - P(\tilde{Y}_i = 0) = 1 - e^{-\lambda_i} = 1 - exp(-exp(X_i \cdot \beta)) \implies X_i \cdot \beta = ln(ln(1 - \mu_i))$$

stąd funkcja linkująca jest postaci $f(x) = ln(-ln(1-x))$ ("complementary log-log function"). Przykładem sytuacji, gdy ma sens zastosowanie tego typu regresji jest badanie w którym obserwowana zmienna kodowałaby odpowiedź na pytanie "czy danego dnia były jakieś wypadki drogowe?" - liczbę wypadków można modelować za pomocą rozkładu Poissona, jednak nas nie obchodzi ich konkretna ilość, tylko czy było ich więcej niż 0. (Analogicznie jako przykład dla regresji probitowej możnabyło podać badanie w którym zmienna koduje np. odpowiedź na pytanie "czy IQ osoby jest powyżej znanej nam średniej?", gdzie zakładamy że IQ ma rozkład normalny.)

Cauchit

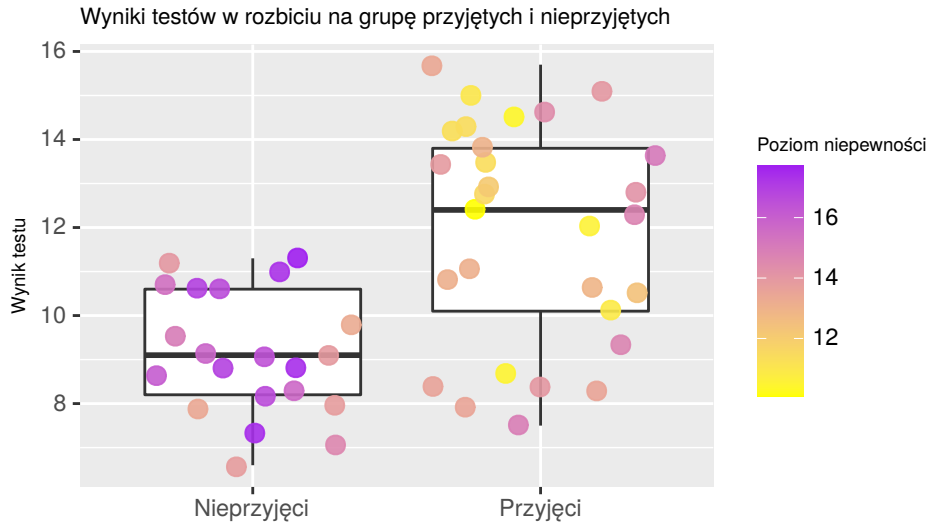
Tak jak w przypadku probitu funkcą linkującą był kwantyl rozkładu normalnego standardowego, tak w przypadku cauchitu jest nią kwantyl rozkładu Cauchy'ego.



Analiza danych

Zadania 1-3

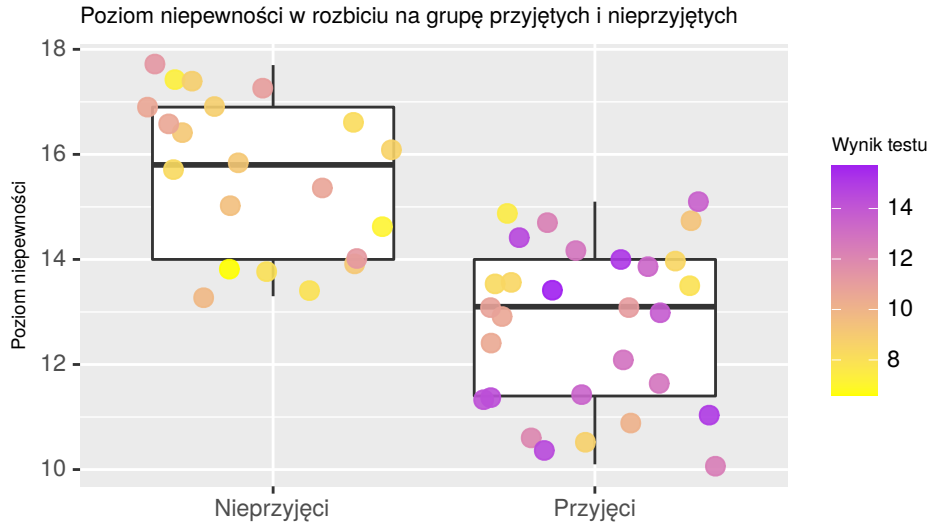
Zbiór danych "Lista_1.csv" zawiera kolumny `X` - numer obserwacji, `numeracy` - wynik testu, `anxiety` - poziom niepewności oraz `success` - czy studenta przyjęto na studia (1/0 - tak/ nie).



Komentarz:

- Zakresy międzykwartylowe dwóch grup są niemal rozłączne, tzn. wartość III kwartyla w grupie nieprzyjętych leży tylko nieznacznie powyżej I kwartyla w grupie przyjętych.
- Średnia wyników przyjętych jest o około 1 punkt większa niż maksymalny wynik drugiej grupy.

- Powyższe obserwacje wskazują na statystycznie znaczącą różnicę pomiędzy grupami. Należy jednak zwrócić uwagę na duży maksymalny rozrzut w pierwszej grupie i występujące na dole wykresu obserwacje (nie są one z definicji odstające, gdyż nie leżą w przedziale $(Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR)$) w pobliżu średniej grupy pierwszej i możliwość błędnej klasyfikacji takich przypadków.
- Po naniesieniu na wykres dodatkowo poszczególnych obserwacji pokolorowanych zależnie od poziomu niepewności można zauważyć, że wyższe jego wartości są charakterystyczne dla osób nieprzyjętych na studia. W ich grupie nie znalazł się nikt z wynikiem poniżej 12 punktów, natomiast wydaje się że w drugiej grupie nie ma nikogo z wynikiem powyżej 16.
- Na wykresach widać występowanie pewnej korelacji pomiędzy zmiennymi objaśniającymi (założenie o niezależności jest często łamane w praktyce).



Komentarz:

- Wygląd wykresu potwierdza wniosek z poprzedniego zadania: grupa osób nieprzyjętych na studia odznacza się statystycznie wyższym poziomem niepewności.
- Można zauważyć, że I kwartyl w pierwszej grupie w przybliżeniu pokrywa się z III kwartylem w grupie drugiej.
- Całkowite rozrzuty wyników są mniej różne niż te w poprzednim zadaniu.
- Po dodatkowym naniesieniu punktów pokolorowanych zgodnie z wartościami zmiennej `numeracy` możemy ponownie zauważyć, że grupa osób przyjętych ma statystycznie wyższe wyniki testu.

Zadanie 4

Z użyciem funkcji `glm` uzyskaliśmy model opisany równaniem $log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 \cdot X_{i,1} + \beta_2 \cdot X_{i,2}$, gdzie: μ_i to prawdopodobieństwo bycia przyjętym na studia przez i-tego studenta, czyli wartość oczekiwana zmiennej Y_i (`success`), $X_{i,1}$ to wartość zmiennej `numeracy` dla i-tego studenta, a $X_{i,2}$ to wartość zmiennej `anxiety` dla i-tego studenta.

Odpowiedzi do zadania:

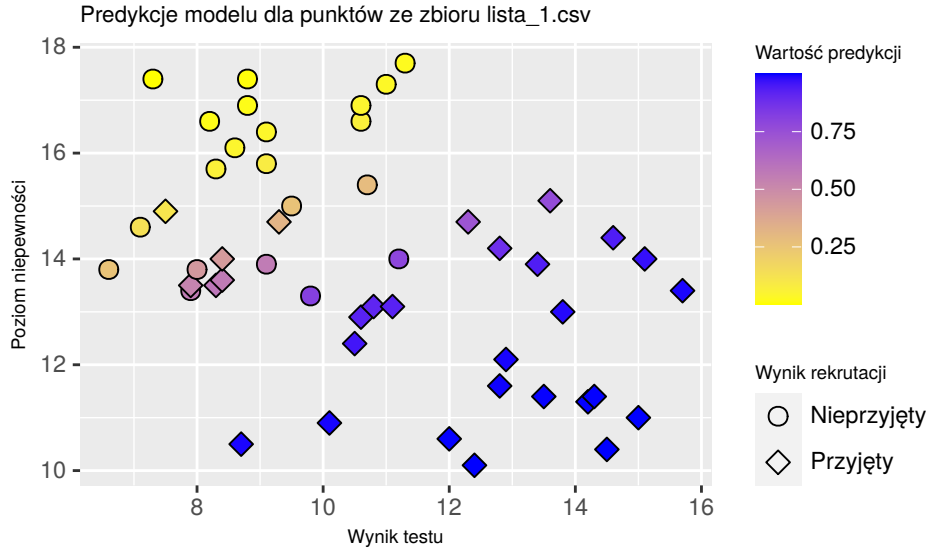
Wartości estymatorów parametrów oraz poziomy istotności możemy odczytać z `summary(model)`

Table 1: Estymatory parametrów i ich p-wartości w zadaniu 4

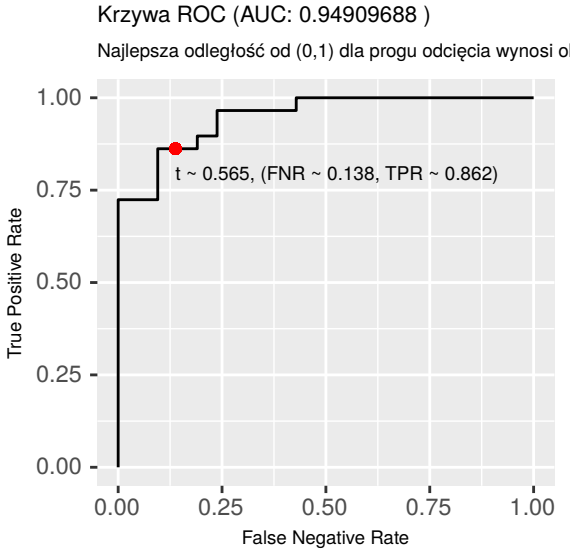
	Estimate	Pr...z..
(Intercept)	14.24	0.04

	Estimate	Pr. . . z..
numeracy	0.58	0.02
anxiety	-1.38	0.00

Prawdopodobieństwo wyznaczone z użyciem `predict.glm` wynosi ok. 88%. Możemy dodatkowo zwizualizować predykowane wartości μ dla wszystkich punktów dostępnych w użytym zbiorze danych i porównać je ze znanym wynikiem testu. W kolejnym podpunkcie ustalimy najlepszy próg odcięcia (wartość t powyżej której mapujemy μ na wartości $Y = 1$.)

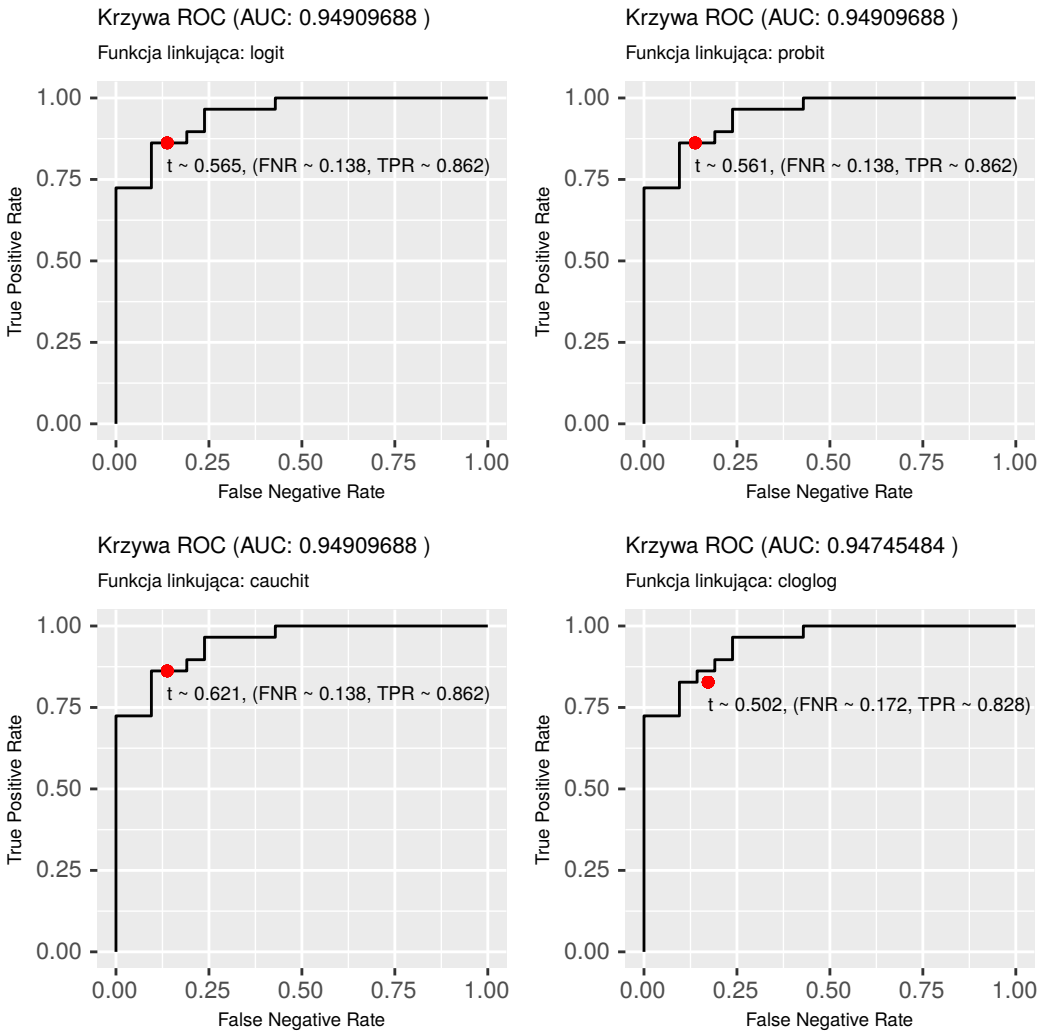


Do narysowania krzywej, obliczenia AUC i znalezienia optymalnego progu używamy funkcji z pakietu `pROC`.



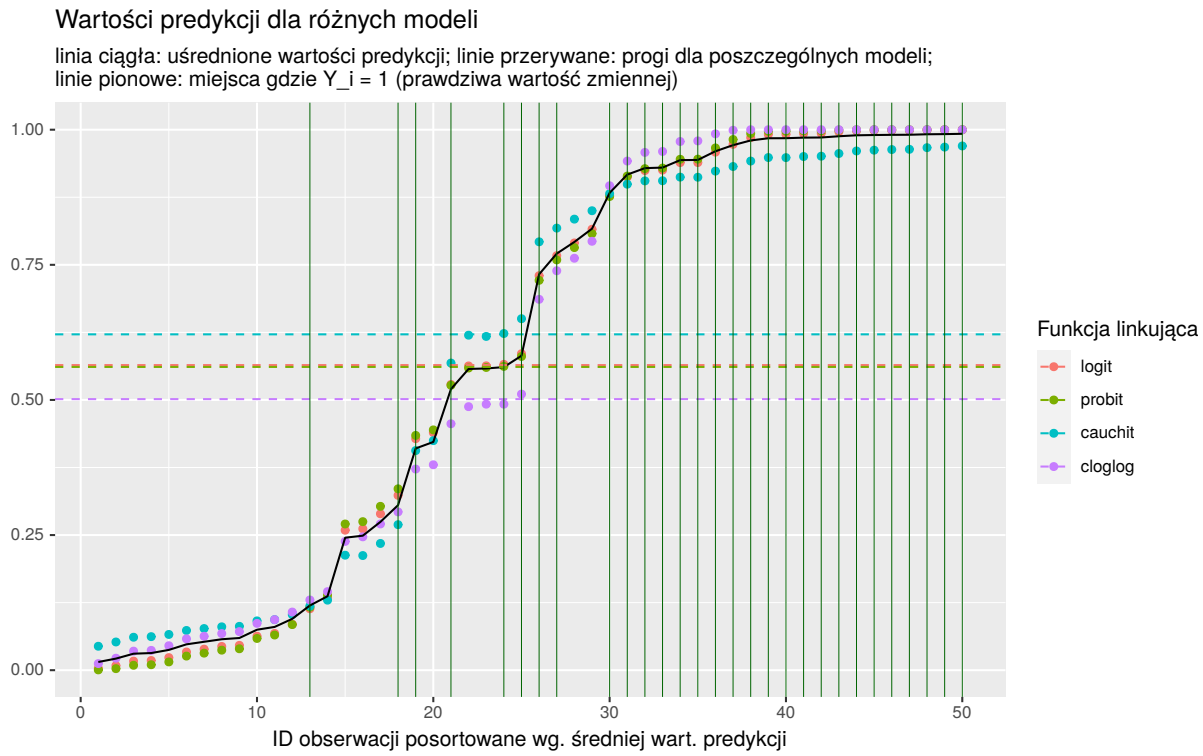
Komentarz: Klasyfikator można ocenić jako całkiem dobry - pole pod wykresem ROC wynosi około 0.95. Punkt optymalny pod względem odległości od $(0, 1)$ odpowiada $FNR \approx 0,14$ i $TPR \approx 0,86$ dla progu odcięcia około $0,56$.

Zadanie 5 - różne funkcje linkujące



Komentarz: Krzywe ROC dla modeli korzystających z różnych funkcji linkujących wyglądają bardzo podobnie. W każdym przypadku AUC wynosi w przybliżeniu 0.95, różnice pod tym względem są naprawdę zaniedbywalne i gdy przybliżamy AUC do 8 miejsc po przecinku trzy modele mają tę samą jego wartość, minimalnie gorszy jest model *cloglog*. Można natomiast zauważyć pewną różnicę w wybranych progach optymalnych pod względem minimalizacji odległości od $(0, 1)$.

Pomocne w zrozumieniu skąd taka różnica się bierze będzie wyrysowanie wartości μ_i zwracanych przez poszczególne modele wraz ze znalezionymi progami. Na wykresie poniżej zaznaczono dodatkowo liniami poziomymi obserwacje, gdzie prawdziwa wartość Y to 1.



Predykcje modeli *logit* i *probit* są bardzo podobne; trzymają się również najbliżej średniej wśród wszystkich modeli. Wartości najbardziej różnią się właśnie w okolicy wyznaczonych progów. Na wykresie widać na przykład, że istnieją 3 obserwacje które leżą poniżej progu odcięcia modelu *cloglog* równocześnie będąc klasyfikowane jako sukces przez pozostałe modele (w tym 2 z nich są rzeczywiście sukcesami). Podobnie istnieją obserwacje, które wszystkie modele oprócz *cauchit* uznają za porażki. Prawdopodobnie przy większej liczbie obserwacji lub innym charakterze danych różnic byłoby więcej.

Zadanie 6

Wracamy do modelu *logit*. Na początku porównamy odchylenia standardowe estymatorów $\hat{\beta}_i$ wyliczone ręcznie (sposobem opisanym we wstępie) z wartościami zwróconymi przez `summary`.

Table 2: Porównanie wartości w zadaniu 6

	(Intercept)	numeracy	anxiety
Obliczone.ręcznie	6.799231	0.2480977	0.4804650
Std..Err..z.summary.model.	6.798519	0.2480840	0.4804027

Widać, że wyniki są z dokładnością do kilku miejsc po przecinku identyczne. W dalszej części zadania testujemy hipotezę postaci $H_0 : \beta_1, \beta_2 = 0$ vs $H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0$. Jak zostało wspomniane we wstępie, można łatwo zrobić wyjmując odpowienie wartości z obiektu modelu. P-wartość statystyki testowej obliczymy jako $1 - F(\chi^2_{|A|})$, za liczbę stopni swobody podstawiając różnicę pomiędzy stopniami swobody modeli s i r.

```
chi2_stat = logitmodel$null.deviance - logitmodel$deviance # statystyka
1 - pchisq(chi2_stat, df = logitmodel$df.null - logitmodel$df.residual) # p-wartość
```

```
## [1] 2.343106e-09
```

Uzyskana p-wartość jest bardzo mała, w szczególności np. $pval < 0.05$ więc odrzucimy hipotezę zerową na poziomie istotnosci 0.05.

W ramach ostatniego podpunktu sprawdzimy, jaki wpływ na model ma parameter `epsilon` podawany do funkcji `glm` podczas jego tworzenia. Parametr ten reguluje rozmiar "kroku" robionego przez wybrany algorytm optymalizacyjny. Wpływa on na dokładność wyznaczania optimum, ale również liczbę iteracji w których algorytm zbiegnie. W dokumentacji można przeczytać, że dopasowywanie modelu regresji logistycznej przebiega z użyciem metody IWLS (Iteratively reweighted least squares), a wywołanie `model$control` pokazuje, że w aktualnym modelu użyliśmy `epsilon = 10^{-8}`. Dodatkowo można z użyciem `model$converged` sprawdzić czy algorytm zbiegł czy zatrzymał się bez znalezienia optimum ze względu na przekroczenie dozwolonej liczby iteracji.

Table 3: Zadanie 6 - wyniki liczbowe podpunktu z parametrem epsilon

epsilon	conv	steps	beta0	beta1	beta2	deviance
0.1	TRUE	3	12.890076428372	0.537584633253287	-1.26395310969864	28.3678360899038
0.01	TRUE	4	14.0924743793738	0.573545076996745	-1.37130612704665	28.2864418201113
0.001	TRUE	5	14.2368342159287	0.577313586875709	-1.38392026253568	28.2856236779218
1e-06	TRUE	6	14.2385811076386	0.577352046950956	-1.38406900945831	28.2856235728691
1e-08	TRUE	6	14.2385811076386	0.577352046950956	-1.38406900945831	28.2856235728691
1e-12	TRUE	7	14.2385813512556	0.577352051413741	-1.3840690296927	28.2856235728691
1e-16	TRUE	7	14.2385813512556	0.577352051413741	-1.3840690296927	28.2856235728691
1e-22	TRUE	7	14.2385813512556	0.577352051413741	-1.3840690296927	28.2856235728691

Komentarz: Dla wszystkich wartości `epsilon` algorytm zbiegł - w każdym przypadku ilość potrzebnych iteracji była dosyć mała, choć dla mniejszych `epsilon` trochę większa niż dla większych (różnica byłaby bardziej widoczna w przypadku "trudniejszych" danych, np. ze znacznie większą macierzą planu lub inną zależnością Y od X.) Jak widać, istnieją subtelne różnice w wartościach znalezionych parametrów i modele o mniejszym kroku są minimalnie lepsze pod względem statystyki *deviance* (`Null Deviance` jest taka sama dla każdego modelu), ale różnica jest naprawdę niewielka i widoczna do 12 miejsca po przecinku tylko poniżej `epsilon = 10^-6`.

Symulacje

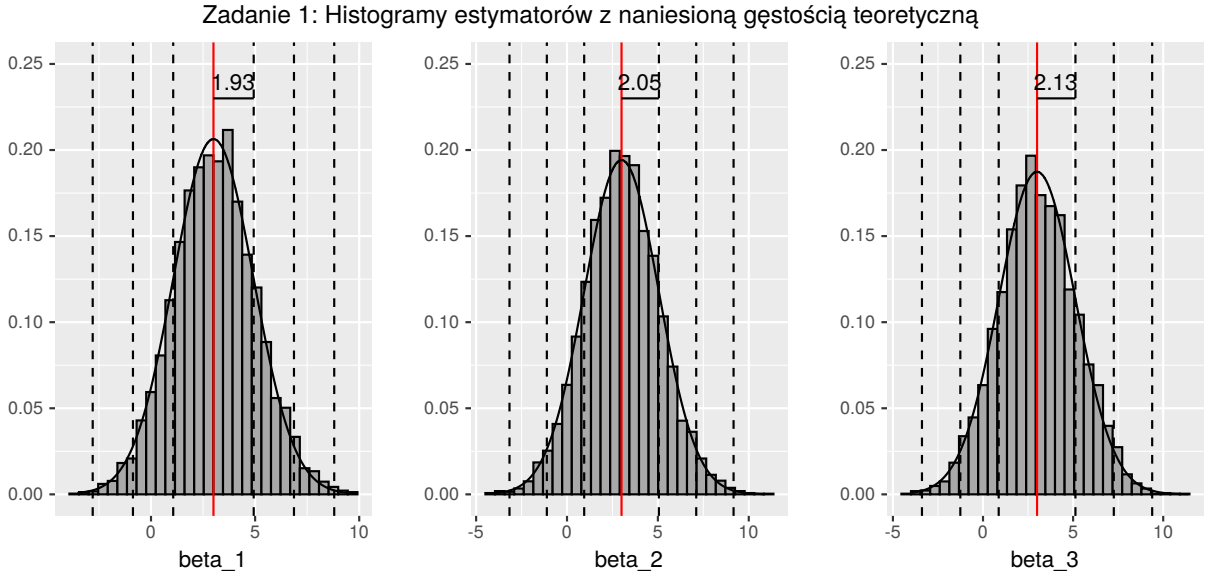
Zadanie 1

Generujemy wektor odpowiedzi długości $n = 4000$ taki, że $log(\frac{\mu_i}{1-\mu_i}) = 3 \cdot X_{i,1} + 3 \cdot X_{i,2} + 3 \cdot X_{i,3}$. Wartość μ_i otrzymujemy nakładając na wyrażenie funkcję *sigmoid*. Zmienne Y_i uzyskujemy losując zmienne z rozkładu zero-jedynkowego z prawdopodobieństwami mu_i.

Obliczamy macierz informacji Fishera z użyciem wygererowanych X i $\beta = (3, 3, 3)$. Asymptotyczna macierz kowariancji powstała przez jej odwrócenie ma postać:

```
##             [,1]        [,2]        [,3]
## [1,]  3.73791830 -0.06729432  0.00648388
## [2,] -0.06729432  4.22192850 -0.48541883
## [3,]  0.00648388 -0.48541883  4.53358219
```

Wariancje estymatorów wyznaczone na podstawie wielu replikacji eksperymentu powinny zbiegać do tych wartości. Na wykresach poniżej widoczne są histogramy uzyskanych w 5000 replikacjach estymatorów. Liniami przerywanymi zaznaczono odległości σ od średniej, a ciągłą linią asymptotyczną gęstość $\hat{\beta}_i$.



Komentarz: Rozkłady estymatorów wyglądają na zgodne z rozkładem asymptotycznym.

Na podstawie wygenerowanych $\hat{\beta}$ estymujemy obciążenie estymatorów

$$b(\hat{\beta}_i) = E(\hat{\beta}_i) - \beta_i$$

gdzie za wartość oczekiwaną $\hat{\beta}_i$, podstawiamy średnią próbkową estymatora. Uzyskamy w ten sposób obciążenia małe, ale niezerowe (nawet przy liczbie powtórzeń zwiększonej do 10000), co sugeruje że estymator MLE jest w tym przypadku obciążony.

```
zad1_results$biases
```

```
## [1] 0.06180907 0.02245828 0.04566158
```

W ramach ostatniego podpunktu porównamy empiryczną macierz kowariancji $\hat{\beta}$ z teoretyczną. Róznice nie są duże:

```
zad1_results$J_inv - zad1_results$estimated_cov
```

```
##              [,1]       [,2]        [,3]
## [1,] -0.12737748 0.07189581  0.05348676
## [2,]  0.07189581 0.02988896  0.06351977
## [3,]  0.05348676 0.06351977 -0.04436557
```

Sprawdzimy jeszcze względny błąd oszacowania wyrazów na przekątnej:

```
mean((diag(zad1_results$J_inv) - diag(zad1_results$estimated_cov))/diag(zad1_results$J_inv))
```

```
## [1] -0.01226122
```

Zadanie 2 - zmniejszona liczba obserwacji

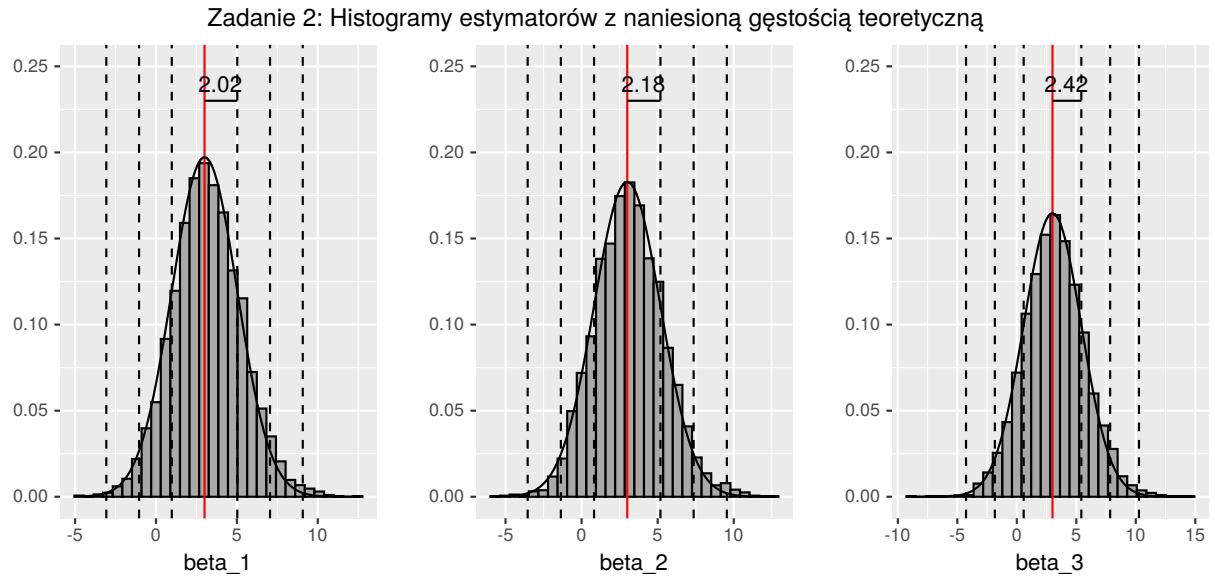
Powtarzamy doświadczenie z macierzą planu przyciętą do pierwszych 100 z 400 wierszy.

Macierz J^{-1} wyznaczona w tym zadaniu jest postaci:

```
##             [,1]       [,2]       [,3]
## [1,]  4.0961026  0.3039696 -0.1547832
```

```
## [2,]  0.3039696  4.7609467 -0.5297449
## [3,] -0.1547832 -0.5297449  5.8707966
```

Można zauważyć wzrost wartości na przekątnej. Intuicyjnie ma to sens, bo dysponując mniejszą liczbą obserwacji oszacujemy parametry mniej dokładnie, a więc ich wariancja wzrośnie.



Można zauważyć, że estymatory są szerzej rozrzucone wokół średnich (wzrost wariancji). Rozkłady ponownie wyglądają na zgodne z asymptotycznymi.

```
zad2_results$biases
```

```
## [1] 0.1905708 0.1449635 0.1876347
```

Obserwujemy wzrost obciążenia, co w przypadku 5000 replikacji eksperymentu nie wydaje się przypadkowe.

```
zad2_results$J_inv - zad1_results$estimated_cov
```

```
##            [,1]       [,2]        [,3]
## [1,]  0.2308068 0.44315969 -0.10778036
## [2,]  0.4431597 0.56890712  0.01919375
## [3,] -0.1077804 0.01919375  1.29284884
```

Różnica pomiędzy teoretyczną asymptotyczną a empirycznie wyznaczoną macierzą kowariancji również wzrosła, choć nie bardzo znacząco.

```
mean((diag(zad2_results$J_inv) - diag(zad2_results$estimated_cov))/diag(zad2_results$J_inv))
```

```
## [1] -0.1092262
```

Zadanie 3 - zależne zmienne objaśniające

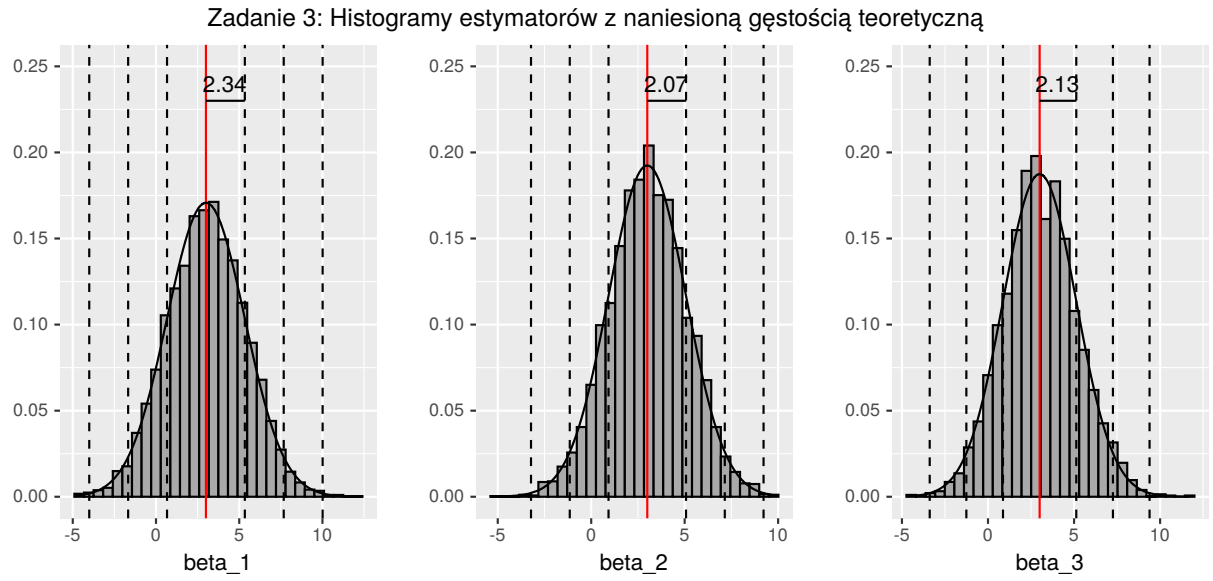
Powtarzamy eksperyment z $n = 400$, ale tym razem zmienne X_i są zależne a kowariancja pomiędzy dwoma różnymi z nich wynosi 0.3.

```
zad3_results$J_inv
```

```
##            [,1]       [,2]       [,3]
## [1,]  5.4603050 -0.8344292 -0.9793438
```

```
## [2,] -0.8344292  4.2997582 -0.8997064
## [3,] -0.9793438 -0.8997064  4.5327503
```

Obserwujemy ponownie wzrost wariancji względem zadania 1 oraz dodatkowo większy niż w zadaniu 2 wzrost wartości wyrazów na przekątnej. Poniżej widać, że rozkłady są zgodne z teoretycznymi:




```
zad3_results$biases
```

```
## [1] 0.03981685 0.04147027 0.03201042
```

Obciążenia są zbliżone do tych z zadania 2.

```
zad3_results$J_inv - zad1_results$estimated_cov
```

```
##            [,1]       [,2]        [,3]
## [1,]  1.5950092 -0.6952391 -0.93234090
## [2,] -0.6952391  0.1077186 -0.35076782
## [3,] -0.9323409 -0.3507678 -0.04519747
```

```
mean((diag(zad3_results$J_inv) - diag(zad3_results$estimated_cov))/diag(zad3_results$J_inv))
```

```
## [1] -0.004560755
```

Największą różnicę względem zadań 1 i 2 widzimy w przypadku różnicy pomiędzy teoretyczną a empiryczną macierzą kowariancji estymatorów, aczkolwiek po sprawdzeniu błędu względnego okazuje się on bardzo mały.

Zadanie 4 - więcej parametrów

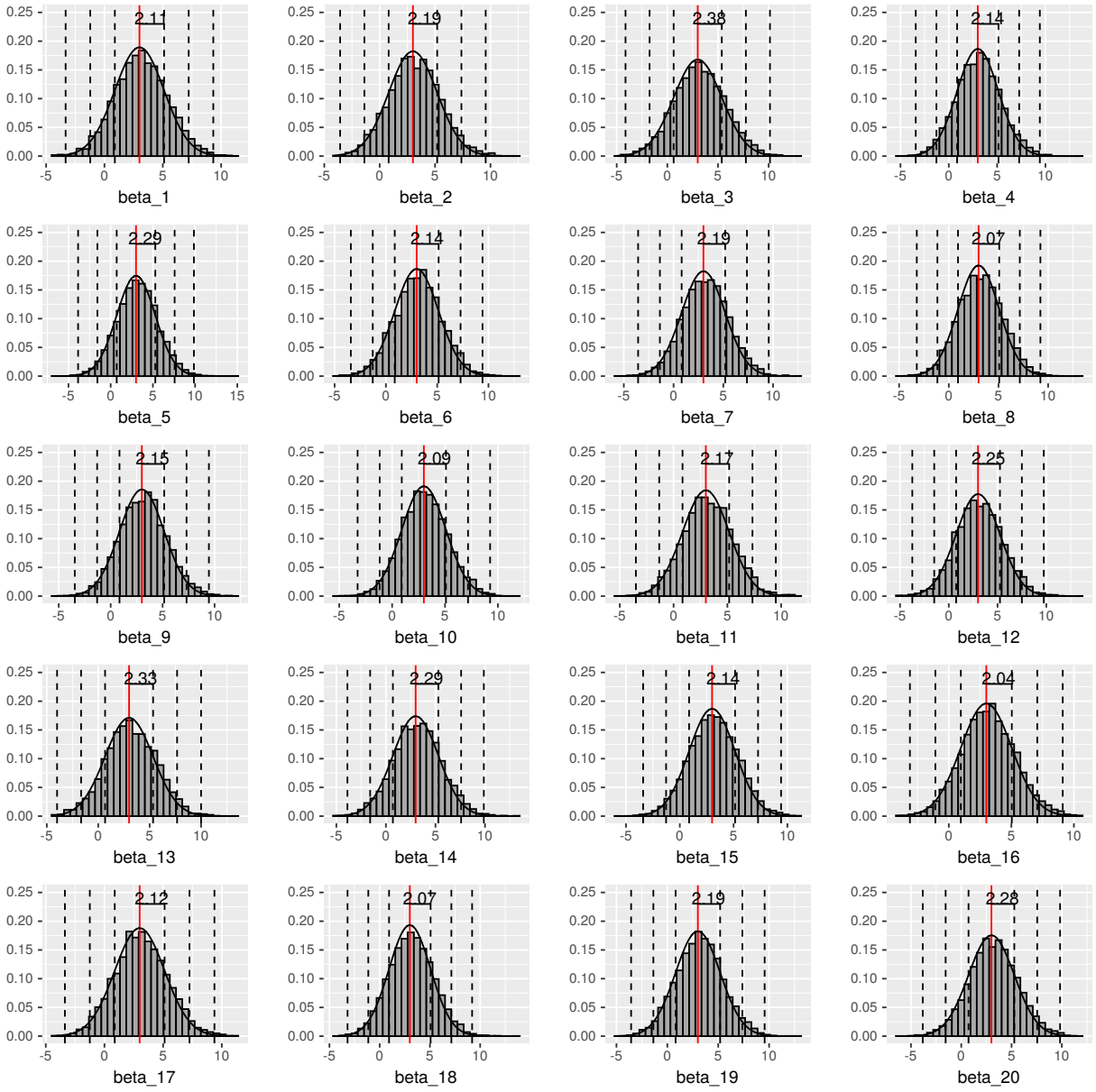
Rozważamy ponownie niezależne kolumny X. Do macierzy planu użytej w zadaniu 1 dołączamy dodatkowo kolumny $X_4, ...X_{20}$ i budujemy model z użyciem wektora $\beta = (3, 3, 3, ...3) \in R^{20}$.

Macierz J^{-1} w tym przypadku jest 20×20, więc wyświetlimy tylko wyrazy na jej przekątnej.

```
diag(zad4_results$J_inv)
```

```
##  [1] 4.447570 4.792903 5.648527 4.579835 5.229717 4.578258 4.784359 4.297474
##  [9] 4.629035 4.359110 4.697004 5.050911 5.445885 5.264790 4.565095 4.146134
## [17] 4.510152 4.268543 4.782535 5.177473
```

Zadanie 4: Histogramy estymatorów z naniesioną gęstością teoretyczną



Obserwowane wariancje estymatorów nie różnią się znacząco od wyników w poprzednich zadaniach, a rozkłady ponownie wyglądają na zgodne z asymptotycznymi. Uzyskane obciążenia przypominają te z zadań 2 i 3:

```
zad4_results$biases
```

```
##  [1] 0.17409634 0.18444216 0.22571553 0.12175686 0.19098142 0.17771147
##  [7] 0.22560957 0.23164490 0.15764731 0.08749677 0.20197133 0.16786487
## [13] 0.20863761 0.20199017 0.15518113 0.19886625 0.13579545 0.18033047
## [19] 0.17147986 0.21420232
```

Sprawdzimy jeszcze błąd względny wyrazów w macierzy kowariancji uśredniony dla całej macierzy:

```
mean((diag(zad4_results$J_inv) - diag(zad4_results$estimated_cov))/diag(zad4_results$J_inv))
```

```
## [1] -0.125655
```

Zadanie 5 - podsumowanie

Wyniki symulacji były we wszystkich przypadkach zgodne z teorią.