Modele Parametryczne: Zespół Policystycznych Jajników (PCOS)

Gomulak Aleksanda Jędrzejczyk Nikola

Podział obowiązków

Każda część projektu została wykonana wspólnie podczas spotkań.

Zbiór danych:

Zbiór danych, na jaki się zdecydowałyśmy, zawiera informacje dotyczące pacjentów przebadanych pod kątem posiadania Zespołu Policystycznych Jajników (PCOS).

W zestawie danych znajdziemy między innymi takie kolumny jak: wiek, waga, regularność miesiączki, czy występują wypryski lub wzmożony porost włosów i tym podobne.

Choroba ta stała się bardzo powszechna na przestrzeni ostatnich lat, w związku z tym kobiety powinny częściej kontrolować wyniki swoich badań i zwracać uwagę na sygnały jakie wysyła im ich ciało.

Co w takim razie może sugerować nam występowanie PCOS? Na co zwrócić uwagę?

Dane pobrałyśmy ze strony Kaggle.

Problem badawczy:

Zespół policystycznych jajników (PCOS) jest jednym z najczęstszych zaburzeń endokrynologicznych u kobiet w wieku rozrodczym. Mimo licznych badań, przyczyny zachorowania na Zespół Policystycznych Jajników pozostają niewyjaśnione, a symptomy znacząco różnią się między poszczególnymi pacjentkami. To prowadzi do trudności w diagnozie i personalizacji leczenia.

Cel badania:

Celem tego badania jest zidentyfikowanie czynników (takich jak nieregularność cyklu, przybranie na wadze, rośnięcie włosów, itp.) i poziomów hormonów (FSH, LH, AMH, itd.), które mogą być związane z występowaniem PCOS oraz ocena ich wpływu na prawdopodobieństwo wystąpienia tej choroby.

Opis zmiennych:

Zmienna:	Opis:
PCOS:	czy pacjent miał zdiagnozowany PCOS? (0 - nie, 1 - tak)
wiek:	wiek pacjenta
BMI:	wartość BMI (Body Mass Index) pacjenta
nieregularnosc_cyklu:	czy cykl miesiączki jest nieregularny? (0 - nie, 1 - tak)
ciaza:	czy pacjent jest w ciąży? (0 - nie, 1 - tak)
poziom_FSH:	poziom hormonu niezbędnego do prawidłowego działania jajników oraz jąder
poziom_LH:	poziom hormonu LH, niski poziom może być oznaką wolnego dojrzewania
stosunek_FSH_do_LH:	stosunek poziomów FSH do LH we krwi pacjenta
poziom_TSH:	poziom hormonu, który kontroluje sposób działania innych hormonów
poziom_AMH:	poziom hormonu, który zapobiega rozwój żeńskich narządów płciowych u płodu męskiego
poziom_PRL:	poziom prolaktyny, hormonu, który odpowiada za laktację czy rozwój piersi
poziom_PRG:	poziom hormonu, który zapobiega rozwój męskich narządów płciowych u płodu żeńskiego
przybranie_na_wadze:	czy pacjent przybrał na wadze? (0 - nie, 1 - tak)
rosniecie_wlosow:	czy pacjentowi rosną włosy? (0 - nie, 1 - tak)
ciemnienie_skory:	czy pacjentowi ciemnieje skóra? (0 - nie, 1 - tak)
wypadanie_wlosow:	czy pacjentowi wypadają włosy? (0 - nie, 1 - tak)
wypryski_na_twarzy:	czy pacjent ma wypryski na twarzy? (0 - nie, 1 - tak)
cisnienie_skurczowe:	skurczowe cisnienie krwi (systoliczne)
cisnienie_rozkurczowe:	rozkurczowe cisnienie krwi (diastoliczne)
pecherzyki_lewy_jajnik:	liczba pęcherzyków w lewym jajniku wykryta podczas badania
pecherzyki_prawy_jajnik:	liczba pęcherzyków w prawym jajniku wykryta podczas badania

Import potrzebnych bibliotek:

```
library("dplyr")
library("GGally")
library("tidyr")
library("ResourceSelection")
library("statmod")
library("car")
library("ggplot2")
library("lmtest")
library("pscl")
library("pROC")
```

Import danych:

```
data <- read.csv("PCOS_data.csv")
```

Usunięcie kolumn, które nas nie interesują, są niepotrzebne lub przekazują podobne informacje (nie uwzględniamy ich w przedstawieniu zmiennych):

```
to_drop <- c("Sl..No", "Patient.File.No.", "Weight..Kg.", "Height.Cm.", "Blood.Group",
"Hip.inch.", "Waist.inch.", "Waist.Hip.Ratio", "Fast.food..Y.N.", "Reg.Exercise.Y.N.",
"Marraige.Status..Yrs.", "II....beta.HCG.mIU.mL.", "RBS.mg.dl.", "Vit.D3..ng.mL.",
"Pulse.rate.bpm.", "RR..breaths.min.", "Hb.g.dl.", "Cycle.length.days.", "No..of.abortions",
"I...beta.HCG.mIU.mL.", "Avg..F.size..L...mm.", "Avg..F.size..R...mm.", "Endometrium..mm.")

data <- data[,!(names(data) %in% to_drop)]
```

Zmiana nazw kolumn dla ułatwienia i większej czytelności:

```
nazwy_kolumn <- c("PCOS", "wiek", "BMI", "nieregularnosc_cyklu", "ciaza", "poziom_FSH",
"poziom_LH", "stosunek_FSH_do_LH", "poziom_TSH", "poziom_AMH", "poziom_PRL",
"poziom_PRG", "przybranie_na_wadze", "rosniecie_wlosow", "ciemnienie_skory",
"wypadanie_wlosow", "wypryski_na_twarzy", "cisnienie_skurczowe", "cisnienie_rozkurczowe",
"pecherzyki_lewy_jajnik", "pecherzyki_prawy_jajnik")

colnames(data)[1:21] <- nazwy_kolumn
```

Sprawdzenie typów zmiennych:

```
str(data)
```

```
## 'data.frame':    541 obs. of  21 variables:
##  $ PCOS                   : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ wiek                   : int  28 36 33 37 25 36 34 33 32 36 ...
##  $ BMI                    : num  19.3 24.9 25.3 29.7 20.1 27.2 26.3 23.1 16 23.1 ...
##  $ nieregularnosc_cyklu   : int  2 2 2 2 2 2 2 2 2 4 ...
##  $ ciaza                  : int  0 1 1 0 1 1 0 1 0 0 ...
##  $ poziom_FSH             : num  7.95 6.73 5.54 8.06 3.98 3.24 2.85 4.86 3.76 2.8 ...
##  $ poziom_LH              : num  3.68 1.09 0.88 2.36 0.9 1.07 0.31 3.07 3.02 1.51 ...
##  $ stosunek_FSH_do_LH     : num  2.16 6.17 6.3 3.42 4.42 3.03 9.19 1.58 1.25 1.85 ...
##  $ poziom_TSH             : num  0.68 3.16 2.54 16.41 3.57 ...
##  $ poziom_AMH             : chr  "2.07" "1.53" "6.63" "1.22" ...
##  $ poziom_PRL             : num  45.2 20.1 10.5 36.9 30.1 ...
##  $ poziom_PRG             : num  0.57 0.97 0.36 0.36 0.38 0.3 0.46 0.26 0.3 0.25 ...
##  $ przybranie_na_wadze    : int  0 0 0 0 0 1 0 1 0 0 ...
##  $ rosniecie_wlosow       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ciemnienie_skory       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ wypadanie_wlosow       : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ wypryski_na_twarzy     : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ cisnienie_skurczowe    : int  110 120 120 120 120 110 120 120 120 110 ...
##  $ cisnienie_rozkurczowe  : int  80 70 80 70 80 70 80 80 80 80 ...
##  $ pecherzyki_lewy_jajnik : int  3 3 13 2 3 9 6 7 5 1 ...
##  $ pecherzyki_prawy_jajnik: int  3 5 15 2 4 6 6 6 7 1 ...
```

Zmiana typu danych, w niektórych kolumnach na zmienne kategorialne, ponieważ są to zmienne dychotomiczne, aby móc w następnych krokach z nich skorzystać:

```
data$PCOS <- as.factor(data$PCOS)
data$nieregularnosc_cyklu <- as.factor(data$nieregularnosc_cyklu)
data$ciaza <- as.factor(data$ciaza)
data$przybranie_na_wadze <- as.factor(data$przybranie_na_wadze)
data$rosniecie_wlosow <- as.factor(data$rosniecie_wlosow)
data$ciemnienie_skory <- as.factor(data$ciemnienie_skory)
data$wypadanie_wlosow <- as.factor(data$wypadanie_wlosow)
data$wypryski_na_twarzy <- as.factor(data$wypryski_na_twarzy)
```

Oraz jedną zmienną na numeryczną:

```
data$poziom_AMH <- as.numeric(data$poziom_AMH)
```

Dla ujednolicenia danych zamieniamy **2 na 0 oraz 4 na 1**.

W zbiorze danych, w zmiennej nieregularnosc_cyklu 2 oznacza, że cykl miesiączkowy jest regularny, a 4, że nie jest regularny. Zatem wartości równe 5 traktujemy jako błąd i nie korzystamy z tych obserwacji.

```
data <- subset(data, nieregularnosc_cyklu != 5)
```

```
data$nieregularnosc_cyklu <- ifelse(data$nieregularnosc_cyklu == 2, 0,
ifelse(data$nieregularnosc_cyklu == 4, 1, data$nieregularnosc_cyklu))
```

Przygotowanie nowej zmiennej kategorialnej:

Zmienną numeryczną BMI zmieniamy na zmienną kategorialną zgodnie z sugestiami WHO, traktując jednak wygłodzenie, wychudzenie oraz niedowagę jako jedną kategorię oraz trzy stopnie otyłości również jako jedną kategorię.

```
data$BMI_kat <- cut(data$BMI,
                    breaks = c(0, 18.5, 25, 30, Inf),
                    labels = c("niedowaga", "poprawna", "nadwaga", "otyłość"),
                    right = FALSE)
table(data$BMI_kat)
```

```
##
## niedowaga  poprawna   nadwaga   otyłość
##        34       277       186        43
```

Pozbycie się możliwie występujących braków:

```
data <- na.omit(data)
```

Sprawdzenie korelacji pomiędzy parami zmiennych:

Prezentujemy pary zmiennych, które cechują się korelacją większą niż 0.7, ponieważ nadmierna korelacja par zmiennych jest niewskazana w przypadku uogólnionych modeli liniowych i zaleca się unikanie takich par zmiennych w modelu.

```
cor_matrix <- cor(data[, sapply(data, is.numeric)])
high_corr <- which(abs(cor_matrix) > 0.7 & abs(cor_matrix) < 1, arr.ind = TRUE)

high_corr_pairs <- data.frame(
  Variable1 = rownames(cor_matrix)[high_corr[, 1]],
  Variable2 = colnames(cor_matrix)[high_corr[, 2]],
  Correlation = cor_matrix[high_corr]
)

high_corr_pairs
```

```
##                 Variable1               Variable2 Correlation
## 1      stosunek_FSH_do_LH              poziom_FSH   0.9719513
## 2              poziom_FSH      stosunek_FSH_do_LH   0.9719513
## 3 pecherzyki_prawy_jajnik  pecherzyki_lewy_jajnik   0.7994650
## 4  pecherzyki_lewy_jajnik pecherzyki_prawy_jajnik   0.7994650
```

Wniosek:

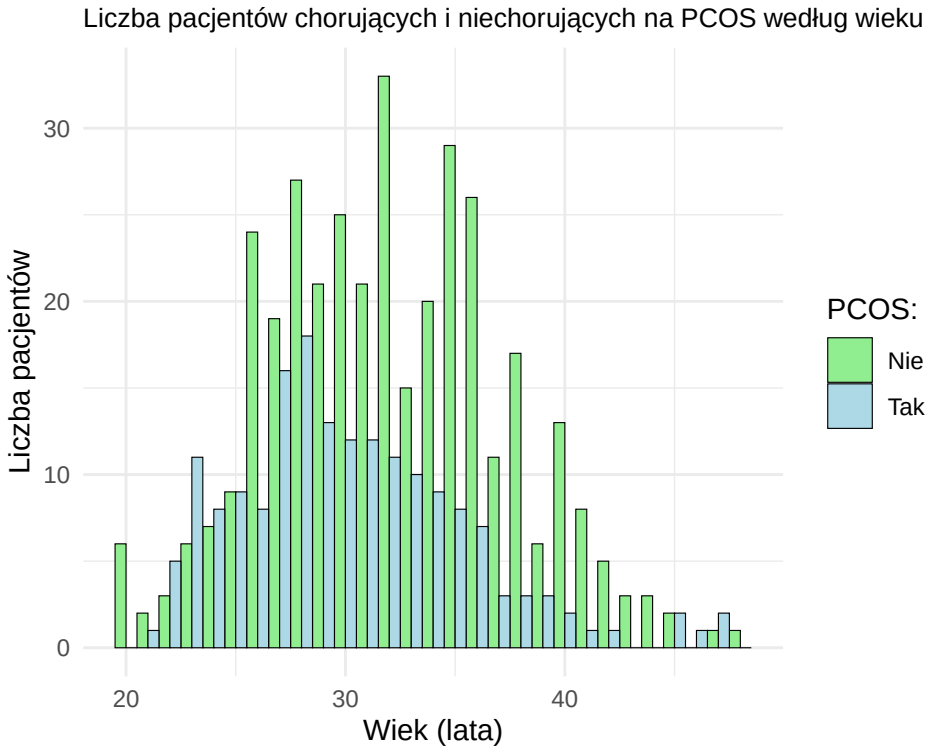
Występują dwie pary zmiennych o nadmiernej korelacji, które nie dziwią.

Decydujemy się na utworzenie nowej zmiennej - średnia ilość pęcherzyków w jajniku oraz usunięcie zmiennej stosunek_FSH_do_LH, która opisuje stosunek poziomów FSH do LH we krwi pacjenta.

```
data <- data[,!(names(data) %in% "stosunek_FSH_do_LH")]
data$Follicle_średnia <- as.integer(rowMeans(data[, c("pecherzyki_lewy_jajnik",
                         "pecherzyki_prawy_jajnik")], na.rm = TRUE))
data <- data[,!(names(data) %in% c("pecherzyki_lewy_jajnik", "pecherzyki_prawy_jajnik"))]
```

Wizualizacja danych:

```
ggplot(data, aes(x = wiek, fill = as.factor(PCOS))) +
  geom_histogram(binwidth = 1, position = "dodge", color = "black", size = 0.2) +
  labs(fill = "PCOS:", x = "Wiek (lata)", y = "Liczba pacjentów",
       title = "Liczba pacjentów chorujących i niechorujących na PCOS według wieku") +
    scale_fill_manual(values = c("0" = "lightgreen", "1" = "lightblue"),
   labels = c("0" = "Nie", "1" = "Tak")) + theme_minimal() +
  theme(plot.title = element_text(size = 10))
```

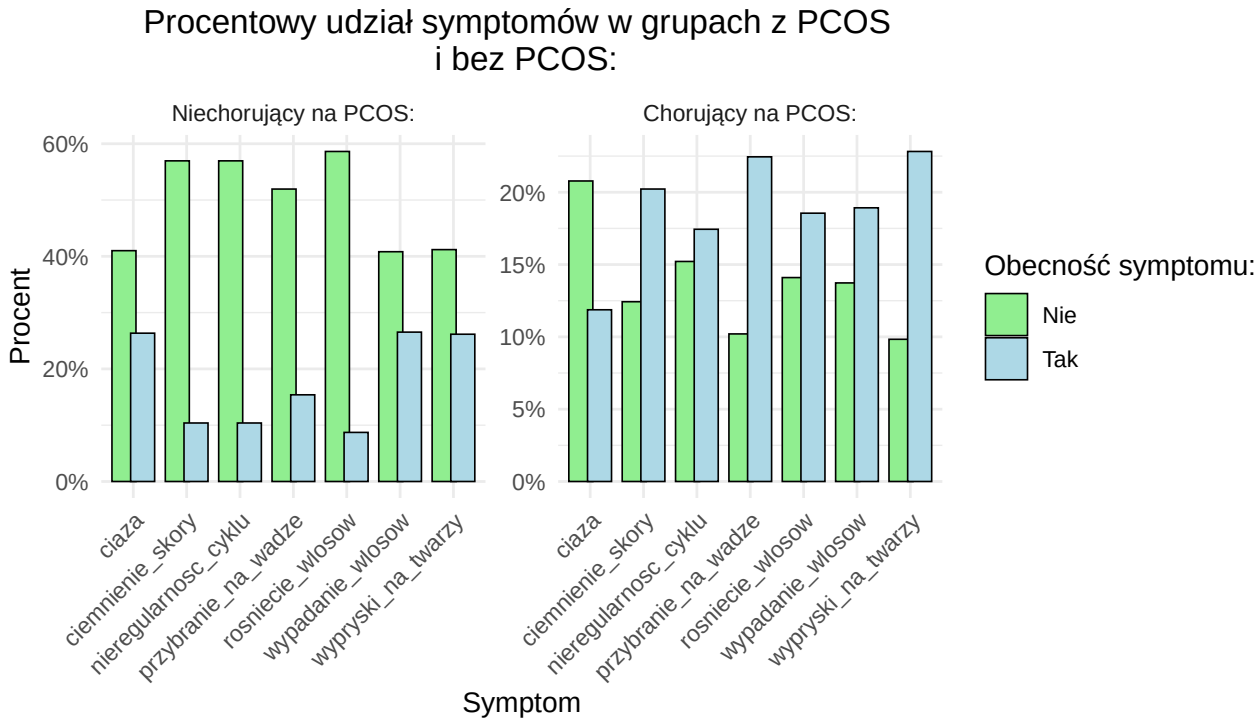


Wniosek:

Można zauważyć wyraźny wzrost zachorowań (lub ich wykrycia) do 30 r.ż. a następnie widoczny spadek. Osoby niechorujące nie wykazują żadnej prawidłowości, co nie budzi zaskoczenia, można jedynie zauważyć, że rozkład osób niechorujących na Zespół Policystycznych Jajników zdaje się być zbliżony do rozkładu normalnego.

```r
data_long <- data %>%
  select(PCOS, nieregularnosc_cyklu, ciaza, przybranie_na_wadze, rosniecie_wlosow,
  ciemnienie_skory, wypadanie_wlosow, wypryski_na_twarzy) %>%
  gather(key = "Symptom", value = "Value", -PCOS) %>%
  group_by(PCOS, Symptom, Value) %>%
  summarise(Count = n(), .groups = 'drop') %>%
  mutate(Frequency = Count / nrow(data))

 ggplot(data_long, aes(x = Symptom, y = Frequency, fill = as.factor(Value))) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.7), color = "black",
           size = 0.3) + facet_wrap(~PCOS, scales = "free_y", labeller =
           as_labeller(c(`0` = "Niechorujący na PCOS:", `1` = "Chorujący na PCOS:"))) +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_fill_manual(values = c("0" = "lightgreen", "1" = "lightblue"),
  labels = c("0" = "Nie", "1" = "Tak")) +
  labs(title = "Procentowy udział symptomów w grupach z PCOS \n i bez PCOS:",
       x = "Symptom",
       y = "Procent",
       fill = "Obecność symptomu:") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5))
```



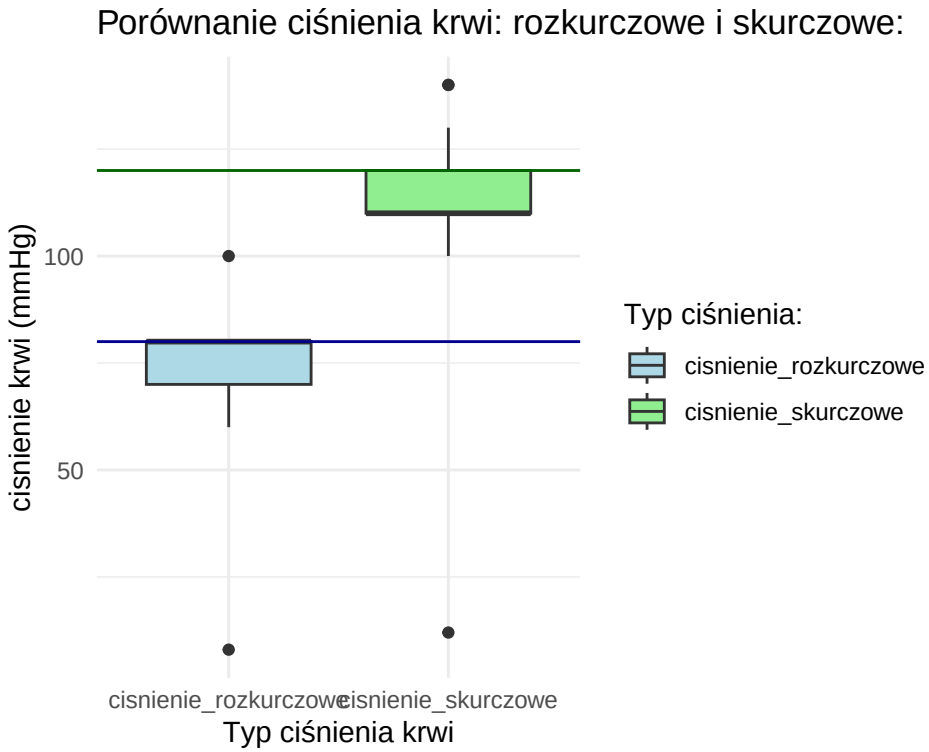
Wniosek:

Zgodnie z oczekiwaniami stosunek osób, które zmagają się z powyższymi symptomami jest zauważalnie większy u osób, u których stwierdzono Zespół Policystycznych Jajników niż u osób, u których tej choroby nie zdiagnozowano.

Warto jednak zwrócić uwagę, że u osób chorujących odpowiedzi "Tak" nie przewyższają tak wyraźnie odpowiedzi "Nie". Można wyciągnąć wniosek, że powyższe symptomy nie są objawami determinującymi Zespół Policystycznych Jajników, jednak ich występowanie powinno dać pacjentom sygnał do spotkania się ze specjalistą.

```
data_long <- data %>%
  gather(key = "Type", value = "BloodPressure", cisnienie_skurczowe, cisnienie_rozkurczowe)

ggplot(data_long, aes(x = Type, y = BloodPressure, fill = Type)) +
  geom_boxplot() +
  geom_hline(yintercept = 120, color = "darkgreen", size = 0.5) +
  geom_hline(yintercept = 80, color = "darkblue", size = 0.5) +
  labs(title = "Porównanie ciśnienia krwi: rozkurczowe i skurczowe:",
       x = "Typ ciśnienia krwi",
       y = "cisnienie krwi (mmHg)",
       fill = "Typ ciśnienia:") +
  theme_minimal() +
  scale_fill_manual(values = c("cisnienie_skurczowe" =
     "lightgreen", "cisnienie_rozkurczowe" = "lightblue"))
```



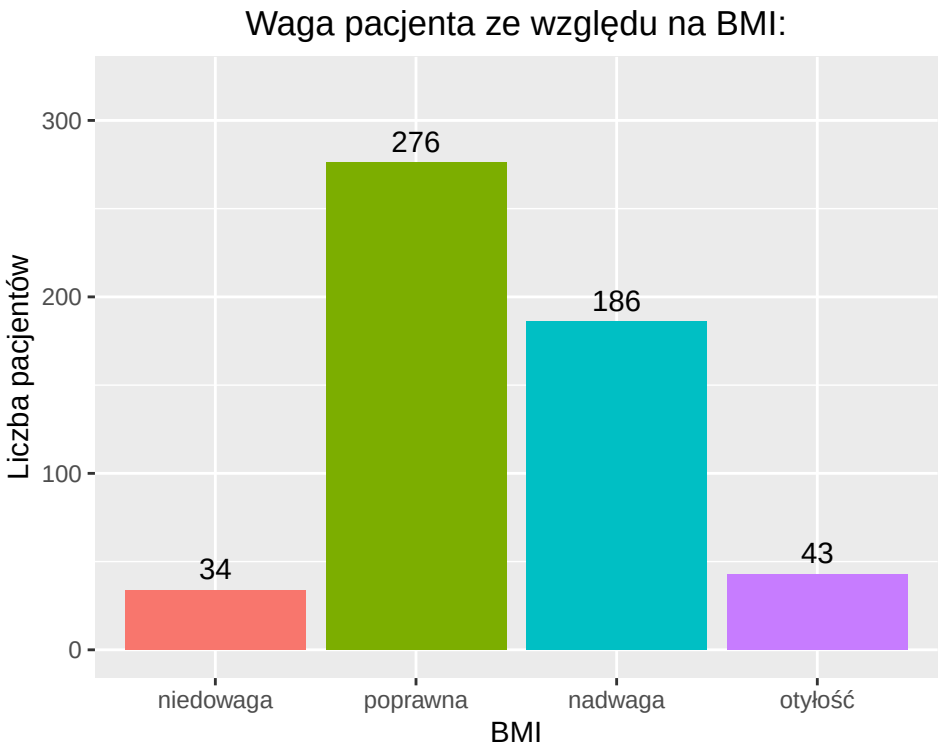
Wniosek:

Linie niebieska i zielona reprezentują górną granicę prawidłowego ciśnienia. Rozstępy ćwiartkowe w obydwu wykresach są do siebie zbliżone lub nawet takie same. Wartości minimalne znajdują się w zauważalnie większej odległości od pierwszego kwartyla niż wartości maksymalne w odległości od trzeciego.

W przypadku ciśnienia rozkurczowego mediana pokrywa się z wartością trzeciego kwartyla, co świadczy o tym, ze większość pacjentów ma ciśnienie niższe niż wartość środkowa. Mówi to o rozkładzie prawoskośnym.

Sytuacja odwrotna zachodzi w przypadku ciśnienia skurczowego, gdzie mediana pokrywa się z pierwszym kwartylem. Oznacza to, że wartości większe od wartości środkowej i mówi to o wykresie lewoskośnym.

```r
ggplot(data, aes(x = factor(BMI_kat), fill = factor(BMI_kat))) +
  geom_bar() +
    geom_text(stat = 'count', aes(label = ..count..), vjust = -0.5, color = "black") +
  theme(legend.position = "none") +
  labs(x = "BMI", y = "Liczba pacjentów", title = "Waga pacjenta ze względu na BMI:") +
  ylim(0, 320) + theme(plot.title = element_text(hjust = 0.5))
```



Wniosek:

Dominującą grupą są pacjenci mieszczący się w poprawnym według WHO indeksie masy ciała. Dużą część pacjentów stanowią osoby z nadwagą. Kategorie skrajne są małoliczne.

Budowa modeli:

Podział zbioru na uczący i testowy w stosunku 70:30. Zbiór uczący służyć nam będzie do budowania modeli, a zbiór testowy do ich oceny.

```
set.seed(1257)
n <- nrow(data)
liczby_losowe <- sample(c(1:n), round(0.7*n), replace = FALSE)
data_uczący <- data[liczby_losowe,]
data_testowy <- data[-liczby_losowe,]
```

Proporcje `PCOS` w podzbiorach danych:

```
summary(data$PCOS) / n
```

```
##         0         1
## 0.6734694 0.3265306
```

```
summary(data_uczący$PCOS) / nrow(data_uczący)
```

```
##         0         1
## 0.6843501 0.3156499
```

```
summary(data_testowy$PCOS) / nrow(data_testowy)
```

```
##         0         1
## 0.6481481 0.3518519
```

Proporcje chorujących i niechorujących na Zespół Policystycznych Jajników są w dużej mierze zachowane w każdym ze zbiorów.

Estymacja modelu dwumianowego logitowego:

Estymacja modelu dla zmiennej dychotomicznej (PCOS) Y `family = binomial` z domyślną funkcją wiążącą probit `link = logit`.

```
model_logit <- glm(PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +  Follicle_średnia +
poziom_AMH + poziom_PRL + poziom_PRG, data = data_uczący, family = binomial)

summary(model_logit)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_TSH + poziom_FSH +
##     poziom_LH + Follicle_średnia + poziom_AMH + poziom_PRL +
##     poziom_PRG, family = binomial, data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.367522   1.025638  -7.183 6.80e-13 ***
## rosniecie_wlosow1     1.519684   0.433427   3.506 0.000455 ***
## wypryski_na_twarzy1   1.110776   0.394394   2.816 0.004856 **
## ciemnienie_skory1     1.693961   0.419015   4.043 5.28e-05 ***
## nieregularnosc_cyklu  1.326693   0.429258   3.091 0.001997 **
## poziom_TSH           -0.004008   0.066351  -0.060 0.951829
## poziom_FSH           -0.029678   0.049041  -0.605 0.545062
## poziom_LH             0.118172   0.081346   1.453 0.146304
## Follicle_średnia      0.558897   0.073550   7.599 2.99e-14 ***
## poziom_AMH            0.035503   0.033230   1.068 0.285337
## poziom_PRL            0.007327   0.014899   0.492 0.622861
## poziom_PRG            0.060394   0.883938   0.068 0.945528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 183.13  on 365  degrees of freedom
## AIC: 207.13
##
## Number of Fisher Scoring iterations: 10
```

Wniosek:

Mamy 5 istotnym parametrów na poziomie istotności $\alpha = 0.05$. Ciekawym faktem jest, że zgodnie z powyższym modelem zmienne, mówiące o poziomach badanych hormonów zdają się nie mieć istotnego statystycznie wpływu na zmienną endogeniczną PCOS.

Testy istotności wszystkich zmiennych niezależnych w modelu:

Test ilorazu wiarygodności i test Walda:

```
lrtest(model_logit)
```

```
## Likelihood ratio test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
## Model 2: PCOS ~ 1
##   #Df   LogLik  Df  Chisq Pr(>Chisq)
## 1  12  -91.564
## 2   1 -235.077 -11 287.03  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(model_logit)
```

```
## Wald test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
## Model 2: PCOS ~ 1
##   Res.Df  Df      F    Pr(>F)
## 1    365
## 2    376 -11 8.1394 9.293e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wniosek:

Zgodnie z obydwoma testami odrzucamy H0, zatem istnieją takie zmienne, które istotnie wpływają na kształtowanie się zmiennej zależnej.

Sprawdzenie czy zmienne objaśniające są współliniowe:

```
vif(model_logit)
```

```
##     rosniecie_wlosow   wypryski_na_twarzy     ciemnienie_skory
##             1.117497             1.039212             1.140977
## nieregularnosc_cyklu          poziom_TSH          poziom_FSH
##             1.149943             1.048832             1.093276
##            poziom_LH     Follicle_średnia          poziom_AMH
##             1.188347             1.204987             1.108076
##           poziom_PRL          poziom_PRG
##             1.034793             1.047829
```

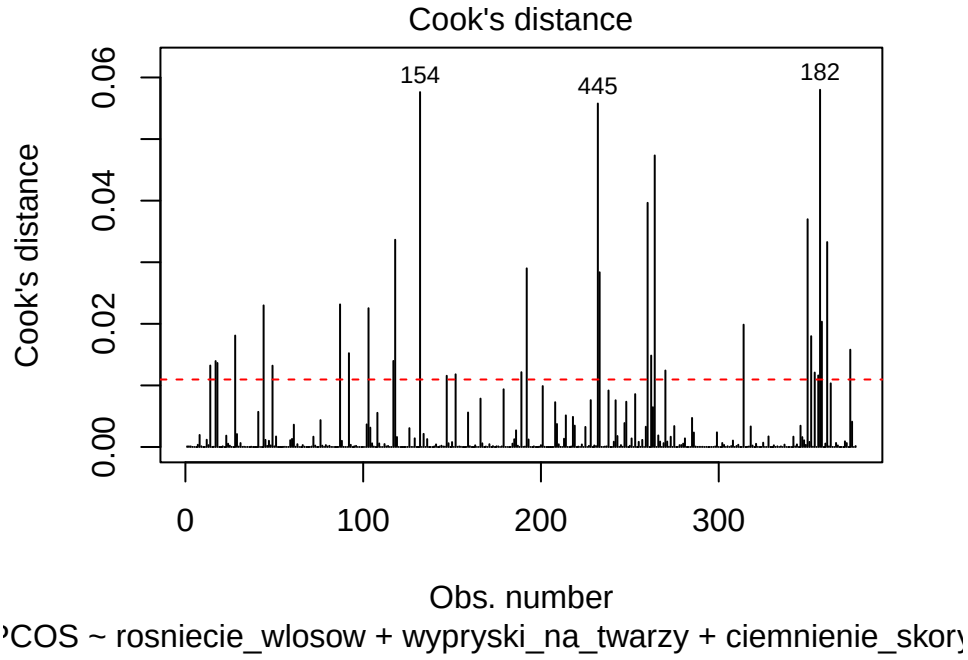
Wniosek:

Zmienne nie są współliniowe, ponieważ oscylują wokół wartości 1, gdzie 1 oznacza sytuację idealną i całkowity brak współliniowości. W literaturze przyjmuje się zakres [1:5] jako umiarkowaną współliniowość, którą można zaakceptować.

Identyfikacja obserwacji wpływowych:

```
c <- 4/(nrow(data_uczący) - (length(model_logit$coefficients) - 1) - 1)

plot(model_logit, which = 4)
abline(h = c, col = "red", lty = 2)
```



Odległość Cook'a mierzy wpływ danej obserwacji na ocenę parametrów β poprzez porównanie ocen wartości teoretycznych y_i wyznaczonych z modelu oszacowanego na podstawie wszystkich obserwacji oraz modelu oszacowanego bez danej obserwacji. Obserwacja jest wpływowa, gdy $D_i > \frac{4}{n-k-1}$.

Zgodnie z przyjętym progiem, model wykazuje wiele obserwacji wpływowych.

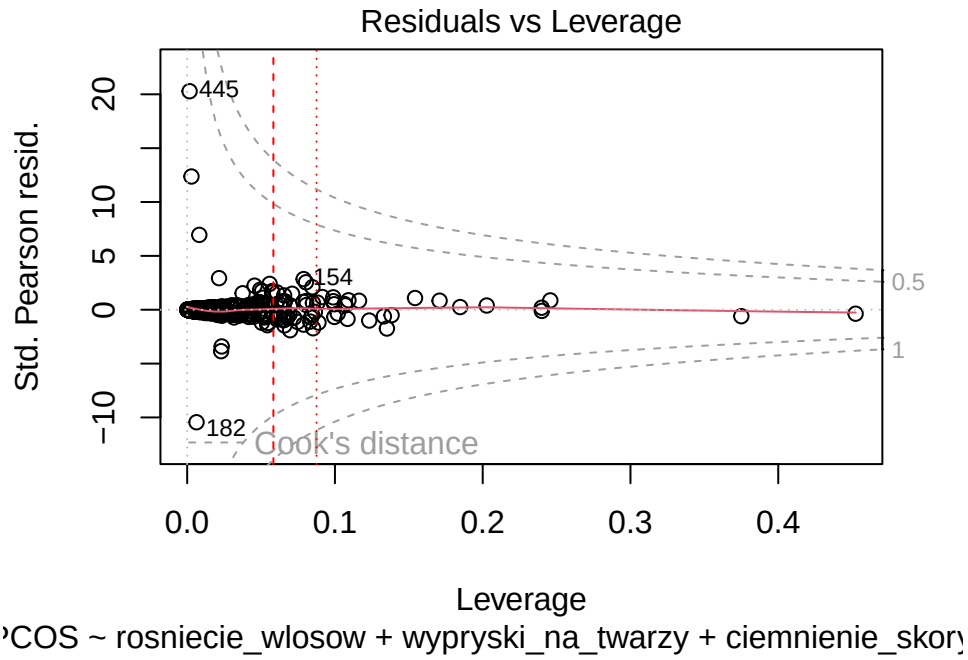
```
sum(cooks.distance(model_logit)>c)
```

```
## [1] 31
```

Zgodnie z wyliczonym progiem, w modelu występuje 31 obserwacji wpływowych, jednak decydujemy się na zostawienie ich. Możliwymi rozwiązaniami na ten problem są usunięcie tych wartości lub zlogarytmowanie zmiennych.

```r
w1 <- 2 *(length(model_logit$coefficients) - 1) / nrow(data_uczący)
w2 <- 3 *(length(model_logit$coefficients) - 1) / nrow(data_uczący)

plot(model_logit, which = 5)
abline(v = c(w1, w2), col = "red", lty = c(2, 3))
```



Wskaźnik dźwigni wyznaczany jest dla każdej zmiennej objaśniającej X z osobna, mierzy dla poszczególnych obserwacji odstępstwo od zmiennej objaśniającej x_i od jej średniego poziomu. Przyjmuje się, że obserwacja jest wpływowa, jeżeli $W_i > \frac{2(k+1)}{n}$ lub $W_i > \frac{3(k+1)}{n}$.

Identyfikacja obserwacji nietypowych:

Bonferroni Outlier Test:

```r
outlierTest(model_logit, n.max = Inf)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 445 3.565096         0.00036372      0.13712
```

Wniosek:

Zgodnie z wynikami procedury Bonferroniego w modelu nie ma istotnie statystycznie odstających wartości.

Estymacja modelu metodą krokową przy minimalizacji kryterium informacyjnego AIC. Wykonujemy ją trzema metodami krokowymi:

```
model_logit_step_both <- step(model_logit)
```

```
## Start:  AIC=207.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_TSH            1   183.13 205.13
## - poziom_PRG            1   183.13 205.13
## - poziom_PRL            1   183.37 205.36
## - poziom_FSH            1   183.65 205.65
## - poziom_AMH            1   184.45 206.45
## <none>                      183.13 207.13
## - poziom_LH             1   189.56 211.56
## - wypryski_na_twarzy    1   191.33 213.33
## - nieregularnosc_cyklu  1   192.91 214.90
## - rosniecie_wlosow      1   195.88 217.88
## - ciemnienie_skory      1   200.35 222.35
## - Follicle_średnia      1   296.37 318.37
##
## Step:  AIC=205.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_PRG            1   183.13 203.13
## - poziom_PRL            1   183.37 203.37
## - poziom_FSH            1   183.65 203.65
## - poziom_AMH            1   184.45 204.45
## <none>                      183.13 205.13
## - poziom_LH             1   189.56 209.57
## - wypryski_na_twarzy    1   191.35 211.35
## - nieregularnosc_cyklu  1   193.10 213.10
## - rosniecie_wlosow      1   195.89 215.89
## - ciemnienie_skory      1   200.54 220.54
## - Follicle_średnia      1   296.48 316.48
##
## Step:  AIC=203.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRL
##
##                        Df Deviance    AIC
## - poziom_PRL            1   183.38 201.38
## - poziom_FSH            1   183.65 201.65
## - poziom_AMH            1   184.45 202.45
## <none>                      183.13 203.13
## - poziom_LH             1   189.56 207.57
## - wypryski_na_twarzy    1   191.39 209.39
```

```
## - nieregularnosc_cyklu  1   193.10 211.10
## - rosniecie_wlosow      1   195.97 213.97
## - ciemnienie_skory      1   200.55 218.55
## - Follicle_średnia      1   300.59 318.59
##
## Step:  AIC=201.38
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH
##
##                         Df Deviance    AIC
## - poziom_FSH             1   183.94 199.94
## - poziom_AMH             1   184.61 200.61
## <none>                       183.38 201.38
## - poziom_LH              1   190.13 206.13
## - wypryski_na_twarzy     1   191.80 207.80
## - nieregularnosc_cyklu   1   193.47 209.47
## - rosniecie_wlosow       1   196.55 212.55
## - ciemnienie_skory       1   200.73 216.73
## - Follicle_średnia       1   300.67 316.67
##
## Step:  AIC=199.94
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia + poziom_AMH
##
##                         Df Deviance    AIC
## - poziom_AMH             1   185.16 199.16
## <none>                       183.94 199.94
## - poziom_LH              1   190.32 204.32
## - wypryski_na_twarzy     1   192.14 206.14
## - nieregularnosc_cyklu   1   194.06 208.06
## - rosniecie_wlosow       1   197.33 211.33
## - ciemnienie_skory       1   202.04 216.04
## - Follicle_średnia       1   301.64 315.64
##
## Step:  AIC=199.16
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia
##
##                         Df Deviance    AIC
## <none>                       185.16 199.16
## - poziom_LH              1   192.44 204.44
## - wypryski_na_twarzy     1   192.83 204.83
## - nieregularnosc_cyklu   1   197.05 209.05
## - rosniecie_wlosow       1   199.69 211.69
## - ciemnienie_skory       1   204.10 216.10
## - Follicle_średnia       1   305.89 317.89
```

```
model_logit_step_for <- step(model_logit, direction = 'forward')
```

```
## Start:  AIC=207.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
```

```r
model_logit_step_back <- step(model_logit, direction = 'backward')
```

```
## Start:  AIC=207.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_TSH            1   183.13 205.13
## - poziom_PRG            1   183.13 205.13
## - poziom_PRL            1   183.37 205.36
## - poziom_FSH            1   183.65 205.65
## - poziom_AMH            1   184.45 206.45
## <none>                      183.13 207.13
## - poziom_LH             1   189.56 211.56
## - wypryski_na_twarzy    1   191.33 213.33
## - nieregularnosc_cyklu  1   192.91 214.90
## - rosniecie_wlosow      1   195.88 217.88
## - ciemnienie_skory      1   200.35 222.35
## - Follicle_średnia      1   296.37 318.37
##
## Step:  AIC=205.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_PRG            1   183.13 203.13
## - poziom_PRL            1   183.37 203.37
## - poziom_FSH            1   183.65 203.65
## - poziom_AMH            1   184.45 204.45
## <none>                      183.13 205.13
## - poziom_LH             1   189.56 209.57
## - wypryski_na_twarzy    1   191.35 211.35
## - nieregularnosc_cyklu  1   193.10 213.10
## - rosniecie_wlosow      1   195.89 215.89
## - ciemnienie_skory      1   200.54 220.54
## - Follicle_średnia      1   296.48 316.48
##
## Step:  AIC=203.13
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRL
##
##                        Df Deviance    AIC
## - poziom_PRL            1   183.38 201.38
## - poziom_FSH            1   183.65 201.65
## - poziom_AMH            1   184.45 202.45
## <none>                      183.13 203.13
## - poziom_LH             1   189.56 207.57
## - wypryski_na_twarzy    1   191.39 209.39
## - nieregularnosc_cyklu  1   193.10 211.10
## - rosniecie_wlosow      1   195.97 213.97
```

```
## - ciemnienie_skory      1   200.55 218.55
## - Follicle_średnia      1   300.59 318.59
##
## Step:  AIC=201.38
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH
##
##                         Df Deviance    AIC
## - poziom_FSH             1   183.94 199.94
## - poziom_AMH             1   184.61 200.61
## <none>                       183.38 201.38
## - poziom_LH              1   190.13 206.13
## - wypryski_na_twarzy     1   191.80 207.80
## - nieregularnosc_cyklu   1   193.47 209.47
## - rosniecie_wlosow       1   196.55 212.55
## - ciemnienie_skory       1   200.73 216.73
## - Follicle_średnia       1   300.67 316.67
##
## Step:  AIC=199.94
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia + poziom_AMH
##
##                         Df Deviance    AIC
## - poziom_AMH             1   185.16 199.16
## <none>                       183.94 199.94
## - poziom_LH              1   190.32 204.32
## - wypryski_na_twarzy     1   192.14 206.14
## - nieregularnosc_cyklu   1   194.06 208.06
## - rosniecie_wlosow       1   197.33 211.33
## - ciemnienie_skory       1   202.04 216.04
## - Follicle_średnia       1   301.64 315.64
##
## Step:  AIC=199.16
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia
##
##                         Df Deviance    AIC
## <none>                       185.16 199.16
## - poziom_LH              1   192.44 204.44
## - wypryski_na_twarzy     1   192.83 204.83
## - nieregularnosc_cyklu   1   197.05 209.05
## - rosniecie_wlosow       1   199.69 211.69
## - ciemnienie_skory       1   204.10 216.10
## - Follicle_średnia       1   305.89 317.89
```

```r
summary(model_logit_step_both)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia,
##     family = binomial, data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.13769    0.74842  -9.537  < 2e-16 ***
## rosniecie_wlosow1     1.58569    0.42623   3.720 0.000199 ***
## wypryski_na_twarzy1   1.05198    0.38561   2.728 0.006370 **
## ciemnienie_skory1     1.74279    0.41118   4.239 2.25e-05 ***
## nieregularnosc_cyklu  1.42400    0.41890   3.399 0.000675 ***
## poziom_LH             0.11768    0.07623   1.544 0.122616
## Follicle_średnia      0.55926    0.07163   7.808 5.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 185.16  on 370  degrees of freedom
## AIC: 199.16
##
## Number of Fisher Scoring iterations: 8
```

```r
summary(model_logit_step_for)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_TSH + poziom_FSH +
##     poziom_LH + Follicle_średnia + poziom_AMH + poziom_PRL +
##     poziom_PRG, family = binomial, data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.367522   1.025638  -7.183 6.80e-13 ***
## rosniecie_wlosow1     1.519684   0.433427   3.506 0.000455 ***
## wypryski_na_twarzy1   1.110776   0.394394   2.816 0.004856 **
## ciemnienie_skory1     1.693961   0.419015   4.043 5.28e-05 ***
## nieregularnosc_cyklu  1.326693   0.429258   3.091 0.001997 **
## poziom_TSH           -0.004008   0.066351  -0.060 0.951829
## poziom_FSH           -0.029678   0.049041  -0.605 0.545062
## poziom_LH             0.118172   0.081346   1.453 0.146304
## Follicle_średnia      0.558897   0.073550   7.599 2.99e-14 ***
## poziom_AMH            0.035503   0.033230   1.068 0.285337
## poziom_PRL            0.007327   0.014899   0.492 0.622861
## poziom_PRG            0.060394   0.883938   0.068 0.945528
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 183.13  on 365  degrees of freedom
## AIC: 207.13
##
## Number of Fisher Scoring iterations: 10
```

```
summary(model_logit_step_back)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia,
##     family = binomial, data = data_uczący)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.13769    0.74842  -9.537  < 2e-16 ***
## rosniecie_wlosow1     1.58569    0.42623   3.720 0.000199 ***
## wypryski_na_twarzy1   1.05198    0.38561   2.728 0.006370 **
## ciemnienie_skory1     1.74279    0.41118   4.239 2.25e-05 ***
## nieregularnosc_cyklu  1.42400    0.41890   3.399 0.000675 ***
## poziom_LH             0.11768    0.07623   1.544 0.122616
## Follicle_średnia      0.55926    0.07163   7.808 5.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 185.16  on 370  degrees of freedom
## AIC: 199.16
##
## Number of Fisher Scoring iterations: 8
```

Wniosek:

Metoda 'backward' dała takie same wyniki, co metoda domyślna 'both'. Mają one mniejsze wyniki dla kryterium informacyjnego AIC niż metoda 'forward'.

Decydujemy się na zbudowanie modelu zgodnie z metodą 'both'/'backward':

```
model_logit2 <- glm(PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
    ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia,
    data = data_uczący, family = binomial)

summary(model_logit2)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia,
##     family = binomial, data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.13769    0.74842  -9.537  < 2e-16 ***
## rosniecie_wlosow1     1.58569    0.42623   3.720 0.000199 ***
## wypryski_na_twarzy1   1.05198    0.38561   2.728 0.006370 **
## ciemnienie_skory1     1.74279    0.41118   4.239 2.25e-05 ***
## nieregularnosc_cyklu  1.42400    0.41890   3.399 0.000675 ***
## poziom_LH             0.11768    0.07623   1.544 0.122616
## Follicle_średnia      0.55926    0.07163   7.808 5.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 185.16  on 370  degrees of freedom
## AIC: 199.16
##
## Number of Fisher Scoring iterations: 8
```

Model ten zawiera zmienne istotne statystycznie co najmniej na poziomie 0.01 oraz jedną zmienną opisującą poziom hormonu LH, która nie jest istotna statystycznie. Należy jednak pamiętać, że metoda krokowa dąży do minimalizacji kryterium AIC a nie do wykorzystania w modelu jedynie zmiennych istotnych statystycznie.

Testy istotności wszystkich zmiennych niezależnych oraz współliniowości w nowym modelu:

Test ilorazu wiarygodności i test Walda:

```
lrtest(model_logit2)
```

```
## Likelihood ratio test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia
## Model 2: PCOS ~ 1
##   #Df   LogLik Df  Chisq Pr(>Chisq)
## 1   7  -92.581
## 2   1 -235.077 -6 284.99  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(model_logit2)
```

```
## Wald test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia
## Model 2: PCOS ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    370
## 2    376 -6 14.969 2.401e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test współliniowości:

```
vif(model_logit2)
```

```
##     rosniecie_wlosow   wypryski_na_twarzy     ciemnienie_skory
##             1.097285             1.006375             1.114442
## nieregularnosc_cyklu            poziom_LH     Follicle_średnia
##             1.113191             1.078438             1.154026
```

Wniosek:

Testy wykazały, że należy H0, zatem istnieją takie zmienne, które istotnie wpływają na kształtowanie się zmiennej zależnej oraz, że nie są one współliniowe.

Estymacja modelu dwumianowego probitowego:

Estymacja modelu dla zmiennej dychotomicznej (PCOS) Y `family = binomial` z domyślną funkcją wiążącą probit `link = probit`.

```
model_probit <- glm(PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +  Follicle_średnia +
poziom_AMH + poziom_PRL + poziom_PRG, data = data_uczący, family = binomial(link = "probit"))

summary(model_probit)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_TSH + poziom_FSH +
##     poziom_LH + Follicle_średnia + poziom_AMH + poziom_PRL +
##     poziom_PRG, family = binomial(link = "probit"), data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.7465023  0.4878294  -7.680 1.59e-14 ***
## rosniecie_wlosow1    0.8271495  0.2288014   3.615  0.00030 ***
## wypryski_na_twarzy1  0.6354548  0.2082747   3.051  0.00228 **
## ciemnienie_skory1    0.8709578  0.2187345   3.982 6.84e-05 ***
## nieregularnosc_cyklu 0.6929611  0.2281116   3.038  0.00238 **
## poziom_TSH          -0.0003833  0.0341528  -0.011  0.99104
## poziom_FSH          -0.0167103  0.0214428  -0.779  0.43580
## poziom_LH            0.0625594  0.0429741   1.456  0.14546
## Follicle_średnia     0.2767503  0.0344308   8.038 9.14e-16 ***
## poziom_AMH           0.0278229  0.0181344   1.534  0.12497
## poziom_PRL           0.0017418  0.0077183   0.226  0.82146
## poziom_PRG          -0.1191791  0.5183469  -0.230  0.81815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 188.63  on 365  degrees of freedom
## AIC: 212.63
##
## Number of Fisher Scoring iterations: 11
```

Wniosek:

Występuje 5 istotnych parametrów na poziomie $\alpha = 0.05$. Podsumowanie modelu jest analogiczne do modelu logitowego, jednak wartość AIC jest większa.

Testy istotności wszystkich zmiennych niezależnych w modelu:

Test ilorazu wiarygodności i test Walda:

```
lrtest(model_probit)
```

```
## Likelihood ratio test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
## Model 2: PCOS ~ 1
##   #Df   LogLik  Df  Chisq Pr(>Chisq)
## 1  12  -94.313
## 2   1 -235.077 -11 281.53  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(model_probit)
```

```
## Wald test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
## Model 2: PCOS ~ 1
##   Res.Df  Df      F    Pr(>F)
## 1    365
## 2    376 -11 10.749 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wniosek:

Zgodnie z obydwoma testami odrzucamy H0, zatem istnieją takie zmienne, które istotnie wpływają na kształtowanie się zmiennej zależnej.

Sprawdzenie czy zmienne objaśniające są współliniowe:

```
vif(model_probit)
```

```
##     rosniecie_wlosow   wypryski_na_twarzy     ciemnienie_skory
##             1.095024             1.042183             1.104765
## nieregularnosc_cyklu          poziom_TSH          poziom_FSH
##             1.135530             1.042715             1.086940
##           poziom_LH     Follicle_średnia          poziom_AMH
##             1.211519             1.071188             1.099938
##          poziom_PRL          poziom_PRG
##             1.018715             1.040608
```

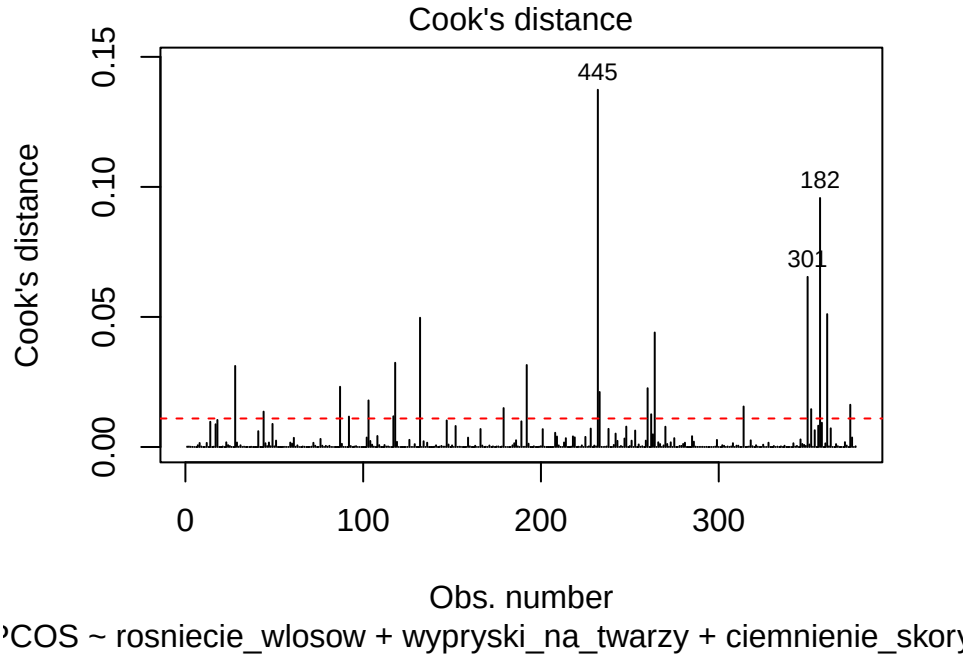
Wniosek:

Zmienne nie są współliniowe.

Identyfikacja obserwacji wpływowych:

```
c <- 4/(nrow(data_uczący) - (length(model_probit$coefficients) - 1) - 1)

plot(model_probit, which = 4)
abline(h = c, col = "red", lty = 2)
```



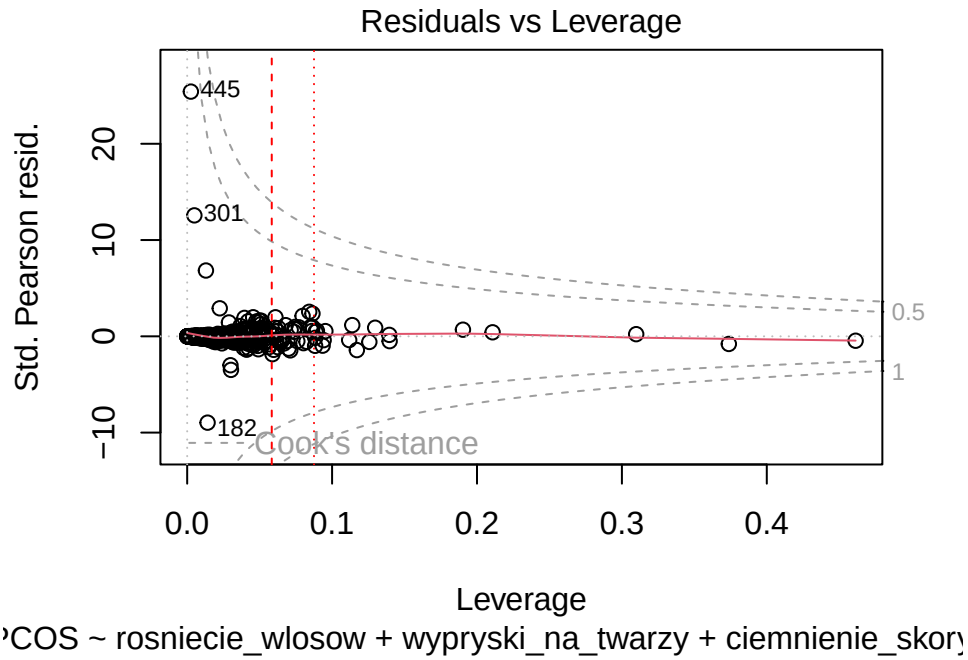

```
sum(cooks.distance(model_probit)>c)
```

```
## [1] 21
```

W modelu probitowym występuje o 10 mniej obserwacji wpływowych niż w modelu logitowym.

```
w1 <- 2 *(length(model_probit$coefficients) - 1) / nrow(data_uczący)
w2 <- 3 *(length(model_probit$coefficients) - 1) / nrow(data_uczący)

plot(model_probit, which = 5)
abline(v = c(w1, w2), col = "red", lty = c(2, 3))
```



W przypadku modelu probitowego również pozostawiamy wartości wpływowe.

Identyfikacja obserwacji nietypowych

Bonferroni Outlier Test:

```
outlierTest(model_probit, n.max = Inf)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 445 3.818842         0.00013408     0.050548
```

Wniosek:

Zgodnie z wynikami procedury Bonferroniego w modelu nie ma istotnie statystycznie odstających wartości.

Estymacja modelu metodą krokową przy minimalizacji kryterium informacyjnego AIC. Wykonujemy ją trzema metodami krokowymi:

```
model_probit_step_both <- step(model_probit)
```

```
## Start:  AIC=212.63
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_TSH            1   188.63 210.63
## - poziom_PRL            1   188.67 210.67
## - poziom_PRG            1   188.68 210.68
## - poziom_FSH            1   189.28 211.28
## <none>                      188.63 212.63
## - poziom_AMH            1   191.26 213.26
## - poziom_LH             1   194.57 216.57
## - nieregularnosc_cyklu  1   197.63 219.63
## - wypryski_na_twarzy    1   198.14 220.14
## - rosniecie_wlosow      1   201.95 223.95
## - ciemnienie_skory      1   204.13 226.14
## - Follicle_średnia      1   298.06 320.06
##
## Step:  AIC=210.63
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_PRL            1   188.67 208.67
## - poziom_PRG            1   188.68 208.68
## - poziom_FSH            1   189.28 209.28
## <none>                      188.63 210.63
## - poziom_AMH            1   191.27 211.27
## - poziom_LH             1   194.58 214.58
## - nieregularnosc_cyklu  1   197.74 217.74
## - wypryski_na_twarzy    1   198.18 218.18
## - rosniecie_wlosow      1   201.96 221.96
## - ciemnienie_skory      1   204.27 224.27
## - Follicle_średnia      1   298.28 318.28
##
## Step:  AIC=208.67
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_PRG            1   188.73 206.73
## - poziom_FSH            1   189.35 207.35
## <none>                      188.67 208.67
## - poziom_AMH            1   191.28 209.28
## - poziom_LH             1   194.84 212.84
## - nieregularnosc_cyklu  1   197.86 215.86
```

```
## - wypryski_na_twarzy    1   198.29 216.29
## - rosniecie_wlosow      1   202.16 220.17
## - ciemnienie_skory      1   204.31 222.31
## - Follicle_średnia      1   298.30 316.30
##
## Step:  AIC=206.73
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH
##
##                         Df Deviance    AIC
## - poziom_FSH             1   189.40 205.40
## <none>                       188.73 206.73
## - poziom_AMH             1   191.36 207.36
## - poziom_LH              1   194.93 210.93
## - nieregularnosc_cyklu   1   197.95 213.95
## - wypryski_na_twarzy     1   198.56 214.56
## - rosniecie_wlosow       1   202.32 218.32
## - ciemnienie_skory       1   204.34 220.34
## - Follicle_średnia       1   302.29 318.28
##
## Step:  AIC=205.4
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia + poziom_AMH
##
##                         Df Deviance    AIC
## <none>                       189.40 205.40
## - poziom_AMH             1   192.00 206.00
## - poziom_LH              1   195.20 209.20
## - nieregularnosc_cyklu   1   198.59 212.59
## - wypryski_na_twarzy     1   198.88 212.88
## - rosniecie_wlosow       1   203.23 217.23
## - ciemnienie_skory       1   205.64 219.64
## - Follicle_średnia       1   302.98 316.98
```

```
model_probit_step_for <- step(model_probit, direction = 'forward')
```

```
## Start:  AIC=212.63
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
```

```
model_probit_step_back <- step(model_probit, direction = 'backward')
```

```
## Start:  AIC=212.63
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_TSH + poziom_FSH + poziom_LH +
##     Follicle_średnia + poziom_AMH + poziom_PRL + poziom_PRG
##
##                         Df Deviance    AIC
## - poziom_TSH             1   188.63 210.63
## - poziom_PRL             1   188.67 210.67
## - poziom_PRG             1   188.68 210.68
```

```
## - poziom_FSH            1   189.28 211.28
## <none>                      188.63 212.63
## - poziom_AMH            1   191.26 213.26
## - poziom_LH             1   194.57 216.57
## - nieregularnosc_cyklu  1   197.63 219.63
## - wypryski_na_twarzy    1   198.14 220.14
## - rosniecie_wlosow      1   201.95 223.95
## - ciemnienie_skory      1   204.13 226.14
## - Follicle_średnia      1   298.06 320.06
##
## Step:  AIC=210.63
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRL + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_PRL            1   188.67 208.67
## - poziom_PRG            1   188.68 208.68
## - poziom_FSH            1   189.28 209.28
## <none>                      188.63 210.63
## - poziom_AMH            1   191.27 211.27
## - poziom_LH             1   194.58 214.58
## - nieregularnosc_cyklu  1   197.74 217.74
## - wypryski_na_twarzy    1   198.18 218.18
## - rosniecie_wlosow      1   201.96 221.96
## - ciemnienie_skory      1   204.27 224.27
## - Follicle_średnia      1   298.28 318.28
##
## Step:  AIC=208.67
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH + poziom_PRG
##
##                        Df Deviance    AIC
## - poziom_PRG            1   188.73 206.73
## - poziom_FSH            1   189.35 207.35
## <none>                      188.67 208.67
## - poziom_AMH            1   191.28 209.28
## - poziom_LH             1   194.84 212.84
## - nieregularnosc_cyklu  1   197.86 215.86
## - wypryski_na_twarzy    1   198.29 216.29
## - rosniecie_wlosow      1   202.16 220.17
## - ciemnienie_skory      1   204.31 222.31
## - Follicle_średnia      1   298.30 316.30
##
## Step:  AIC=206.73
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_FSH + poziom_LH + Follicle_średnia +
##     poziom_AMH
##
##                        Df Deviance    AIC
## - poziom_FSH            1   189.40 205.40
## <none>                      188.73 206.73
## - poziom_AMH            1   191.36 207.36
```

```
## - poziom_LH             1   194.93 210.93
## - nieregularnosc_cyklu  1   197.95 213.95
## - wypryski_na_twarzy    1   198.56 214.56
## - rosniecie_wlosow      1   202.32 218.32
## - ciemnienie_skory      1   204.34 220.34
## - Follicle_średnia      1   302.29 318.28
##
## Step:  AIC=205.4
## PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia + poziom_AMH
##
##                        Df Deviance    AIC
## <none>                      189.40 205.40
## - poziom_AMH            1   192.00 206.00
## - poziom_LH             1   195.20 209.20
## - nieregularnosc_cyklu  1   198.59 212.59
## - wypryski_na_twarzy    1   198.88 212.88
## - rosniecie_wlosow      1   203.23 217.23
## - ciemnienie_skory      1   205.64 219.64
## - Follicle_średnia      1   302.98 316.98
```

```
summary(model_probit_step_both)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia +
##     poziom_AMH, family = binomial(link = "probit"), data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.82251    0.35716 -10.703  < 2e-16 ***
## rosniecie_wlosow1     0.83768    0.22764   3.680 0.000233 ***
## wypryski_na_twarzy1   0.62695    0.20546   3.051 0.002277 **
## ciemnienie_skory1     0.88331    0.21668   4.077 4.57e-05 ***
## nieregularnosc_cyklu  0.69577    0.22692   3.066 0.002169 **
## poziom_LH             0.05352    0.04120   1.299 0.193879
## Follicle_średnia      0.27769    0.03395   8.180 2.83e-16 ***
## poziom_AMH            0.02762    0.01812   1.525 0.127340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 189.40  on 369  degrees of freedom
## AIC: 205.4
##
## Number of Fisher Scoring iterations: 9
```

```r
summary(model_probit_step_for)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_TSH + poziom_FSH +
##     poziom_LH + Follicle_średnia + poziom_AMH + poziom_PRL +
##     poziom_PRG, family = binomial(link = "probit"), data = data_uczący)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.7465023  0.4878294  -7.680 1.59e-14 ***
## rosniecie_wlosow1     0.8271495  0.2288014   3.615  0.00030 ***
## wypryski_na_twarzy1   0.6354548  0.2082747   3.051  0.00228 **
## ciemnienie_skory1     0.8709578  0.2187345   3.982 6.84e-05 ***
## nieregularnosc_cyklu  0.6929611  0.2281116   3.038  0.00238 **
## poziom_TSH           -0.0003833  0.0341528  -0.011  0.99104
## poziom_FSH           -0.0167103  0.0214428  -0.779  0.43580
## poziom_LH             0.0625594  0.0429741   1.456  0.14546
## Follicle_średnia      0.2767503  0.0344308   8.038 9.14e-16 ***
## poziom_AMH            0.0278229  0.0181344   1.534  0.12497
## poziom_PRL            0.0017418  0.0077183   0.226  0.82146
## poziom_PRG           -0.1191791  0.5183469  -0.230  0.81815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 188.63  on 365  degrees of freedom
## AIC: 212.63
##
## Number of Fisher Scoring iterations: 11
```

```r
summary(model_probit_step_back)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia +
##     poziom_AMH, family = binomial(link = "probit"), data = data_uczący)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.82251    0.35716 -10.703  < 2e-16 ***
## rosniecie_wlosow1     0.83768    0.22764   3.680 0.000233 ***
## wypryski_na_twarzy1   0.62695    0.20546   3.051 0.002277 **
## ciemnienie_skory1     0.88331    0.21668   4.077 4.57e-05 ***
## nieregularnosc_cyklu  0.69577    0.22692   3.066 0.002169 **
## poziom_LH             0.05352    0.04120   1.299 0.193879
## Follicle_średnia      0.27769    0.03395   8.180 2.83e-16 ***
## poziom_AMH            0.02762    0.01812   1.525 0.127340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 189.40  on 369  degrees of freedom
## AIC: 205.4
##
## Number of Fisher Scoring iterations: 9
```

Wniosek:

Metoda 'backward' dała takie same wyniki, co metoda domyślna 'both'. Mają one mniejsze wyniki dla kryterium informacyjnego AIC niż metoda 'forward'.

Decydujemy się na zbudowanie modelu zgodnie z metodą 'both'/'backward':

```
model_probit2 <- glm(PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
    ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia +
    poziom_AMH, family = binomial(link = "probit"), data = data_uczący)

summary(model_probit2)
```

```
##
## Call:
## glm(formula = PCOS ~ rosniecie_wlosow + wypryski_na_twarzy +
##     ciemnienie_skory + nieregularnosc_cyklu + poziom_LH + Follicle_średnia +
##     poziom_AMH, family = binomial(link = "probit"), data = data_uczący)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.82251    0.35716 -10.703  < 2e-16 ***
## rosniecie_wlosow1     0.83768    0.22764   3.680 0.000233 ***
## wypryski_na_twarzy1   0.62695    0.20546   3.051 0.002277 **
## ciemnienie_skory1     0.88331    0.21668   4.077 4.57e-05 ***
## nieregularnosc_cyklu  0.69577    0.22692   3.066 0.002169 **
## poziom_LH             0.05352    0.04120   1.299 0.193879
## Follicle_średnia      0.27769    0.03395   8.180 2.83e-16 ***
## poziom_AMH            0.02762    0.01812   1.525 0.127340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 470.15  on 376  degrees of freedom
## Residual deviance: 189.40  on 369  degrees of freedom
## AIC: 205.4
##
## Number of Fisher Scoring iterations: 9
```

Testy istotności wszystkich zmiennych niezależnych oraz współliniowości w nowym modelu:

Test ilorazu wiarygodności i test Walda:

```
lrtest(model_probit2)
```

```
## Likelihood ratio test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     nieregularnosc_cyklu + poziom_LH + Follicle_średnia + poziom_AMH
## Model 2: PCOS ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8  -94.70
## 2   1 -235.08 -7 280.75  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(model_probit2)
```

```
## Wald test
##
## Model 1: PCOS ~ rosniecie_wlosow + wypryski_na_twarzy + ciemnienie_skory +
##     niereguarnosc_cyklu + poziom_LH + Follicle_średnia + poziom_AMH
## Model 2: PCOS ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    369
## 2    376 -7 16.716 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test współliniowości:

```
vif(model_probit2)
```

```
##     rosniecie_wlosow   wypryski_na_twarzy     ciemnienie_skory
##             1.086486             1.016571             1.086253
## niereguarnosc_cyklu           poziom_LH     Follicle_średnia
##             1.129664             1.129952             1.041589
##           poziom_AMH
##             1.098312
```

Wniosek:

Testy wykazały, że należy H0, zatem istnieją takie zmienne, które istotnie wpływają na kształtowanie się zmiennej zależnej oraz, że nie są one współliniowe.

Porównanie dobroci dopasowania modeli logitowych i probitowych:

```
ocena_modelu_dwum <- function(model) {
  kryterium_AIC <- model$aic
  McFadden<- pR2(model)[4]
  Cragg_Uhler<- pR2(model)[6]
  ocena <- data.frame(kryterium_AIC, McFadden, Cragg_Uhler)
  return(ocena)
}
```

Wywołanie powyższej funkcji dla estymowanych modeli:

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

```
##               kryterium_AIC  McFadden Cragg_Uhler
## model_logit         207.1271 0.6104958   0.7478481
## model_logit2        199.1628 0.6061659   0.7442998
## model_probit        212.6256 0.5988007   0.7382199
## model_probit2       205.3994 0.5971548   0.7368536
```

Wniosek:

Według kryterium AIC najlepszy jest model_logit2, który został zbudowany przy pomocy metody krokowej 'backward'/'both'. Według miary PseudoR^2 McFaddena najlepszy jest model_logit zbudowany przez nas. Według miary PseudoR^2 Cragga Uhlera najlepszy jest model model_logit zbudowany przez nas.

```
round(model_logit$coefficients,4)
```

```
##          (Intercept)    rosniecie_wlosow1  wypryski_na_twarzy1
##              -7.3675               1.5197               1.1108
##    ciemnienie_skory1 nieregularnosc_cyklu          poziom_TSH
##               1.6940               1.3267              -0.0040
##           poziom_FSH            poziom_LH    Follicle_średnia
##              -0.0297               0.1182               0.5589
##           poziom_AMH           poziom_PRL          poziom_PRG
##               0.0355               0.0073               0.0604
```

Wzór modelu:

$$logit(p) = -7.3675 + 1.5797 \cdot rosniecie\_wlosow1 + 1.1108 \cdot wypryski\_na\_twarzy + 1.6940 \cdot ciemnienie\_skory1 + 1.3267 \cdot nieregularnosc\_cyklu - 0.0040 \cdot poziom\_TSH - 0.0297 \cdot poziom\_FSH + 0.1182 \cdot poziom\_LH + 0.5589 \cdot Follicle\_średnia + 0.0355 \cdot poziom\_AMH + 0.0073 \cdot poziom\_PRL + 0.0604 \cdot poziom\_PRG$$

Interpretacja modelu:

Ilorazy szans:

```
round(exp(model_logit$coefficients), 4)
```

```
##          (Intercept)    rosniecie_wlosow1  wypryski_na_twarzy1
##               0.0006               4.5708               3.0367
##    ciemnienie_skory1 nieregularnosc_cyklu          poziom_TSH
##               5.4410               3.7686               0.9960
##           poziom_FSH            poziom_LH    Follicle_średnia
##               0.9708               1.1254               1.7487
##           poziom_AMH           poziom_PRL          poziom_PRG
##               1.0361               1.0074               1.0623
```

Intercept Szansa zachorowania na Zespół Policystycznych Jajnikówu pacjentów, którzy nie cierią na nadmierne rośnięcie włosów, wypryski na twarzy, ciemnienie skóry, z regularnymi miesiączkami, poziomami hormonów TSH, FSH, LH, AMH, PRL, PRG na poziomie 0 oraz ze średnią liczbą pęcherczyków w jajnikach równą 0 wynosi 0.0006.

rosniecie_wlosow1(TAK) - Osoby, które skarżą się na nadmierny porost włosów mają ok. 4.5 raza większą szansę na zachorowanie na Zespół Policystycznych Jajników niż osoby, które nie skarżą się na tę dolegliwość przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

wypryski_na_twarzy1(TAK) - Osoby, które skarżą się na występowanie wyprysków na twarzy mają ok. 3 razy większą szansę na zachorowanie na Zespół Policystycznych Jajników niż osoby, które nie skarżą się na tę dolegliwość przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

ciemnienie_skory1(TAK) - Osoby, które skarżą się na ciemnienie skóry mają prawie 5.5 raza większą szansę na zachorowanie na Zespół Policystycznych Jajników niż osoby, które nie skarżą się na tę dolegliwość przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

nieregularnosc_cyklu(TAK) - Osoby z nieregularnym cyklem miesiączkowym mają prawie 4 razy większą szansę na zachorowanie na Zespół Policystycznych Jajników niż osoby, które mają regularny cykl miesiączkowy przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

poziom_TSH - Wraz ze wzrostem poziomu TSH o jedną jednostkę, szansa zachorowania na Zespół Policystycznych Jajników maleje o ok. 0.4% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

poziom_FSH - Wraz ze wzrostem poziomu FSH o jedną jednostkę, szansa zachorowania na Zespół Policystycznych Jajników maleje o ok. 3% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

poziom_LH - Wraz ze wzrostem poziomu LH o jedną jednostkę, szansa zachorowania na Zespół Policystycznych Jajników rośnie o ok. 12.5% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

Follicle_średnia - Wraz ze wzrostem średniej ilości pęcherzyków na jajnikach o jeden, szansa zachorowania na Zespół Policystycznych Jajników rośnie o ok. 75% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

poziom_AMH - Wraz ze wzrostem poziomu AMH o jedną jednostkę, szansa zachorowania na Zespół Policystycznych Jajników wzrośnie o ok. 3.5% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

poziom_PRL - Wraz ze wzrostem poziomu PRL o jedną jednostkę, szansa zachorowania na Zespół Policystycznych Jajników wzrośnie o ok. 0.7% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

poziom_PRG - Wraz ze wzrostem poziomu PRG o jedną jednostkę, szansa zachorowania na Zespół Policystycznych Jajników rośnie o ok. 6% przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

Miary jakości predykcji

Miary oparte na tablicy trafności dla wybranego punktu odcięcia p = 0.5.

Poniższa funkcja `miary_pred` została określona dla argumentów: `model` (model dwumianowy), `dane` (np. zbiór uczący, testowy), `Y` (obserwowany Y 0-1 w analizowanym zbiorze danych).

```
miary_pred <- function(model, dane, Y, p = 0.5) {
 tab <- table(obserwowane = Y, przewidywane = ifelse(predict(model, dane,
                                          type = "response") > p, 1, 0))
 ACC <- (tab[1,1]+tab[2,2])/sum(tab) # dobrze sklasyfikowane
 ER <- (tab[1,2]+tab[2,1])/sum(tab) # źle sklasyfikowane
 SENS <- tab[2,2]/(tab[2,1]+tab[2,2])
 SPEC <- tab[1,1]/(tab[1,1]+tab[1,2])
 PPV <- tab[2,2]/(tab[2,2]+tab[1,2])
 NPV <- tab[1,1]/(tab[1,1]+tab[2,1])
 miary <- data.frame(ACC, ER, SENS, SPEC, PPV, NPV)
 return(miary)
 }
```

Ocena zdolności predykcyjnej na zbiorze uczącym:

```
wyniki_miary_pred <- rbind(
  model_logit = miary_pred(model = model_logit, dane = data_uczący,
                Y = data_uczący$PCOS, 0.5),
  model_logit2 = miary_pred(model = model_logit2, dane = data_uczący,
                Y = data_uczący$PCOS, 0.5),
  model_probit = miary_pred(model = model_probit, dane = data_uczący,
                Y = data_uczący$PCOS,  0.5),
  model_probit2 = miary_pred(model = model_probit2, dane = data_uczący,
                Y = data_uczący$PCOS,  0.5))
wyniki_miary_pred
```

```
##                      ACC         ER      SENS      SPEC       PPV       NPV
## model_logit   0.9071618 0.09283820 0.8235294 0.9457364 0.8750000 0.9207547
## model_logit2  0.9071618 0.09283820 0.8235294 0.9457364 0.8750000 0.9207547
## model_probit  0.9098143 0.09018568 0.8319328 0.9457364 0.8761062 0.9242424
## model_probit2 0.9098143 0.09018568 0.8319328 0.9457364 0.8761062 0.9242424
```

Ocena zdolności predykcyjnej na zbiorze testowym:

```
wyniki_miary_pred2 <- rbind(
  model_logit = miary_pred(model = model_logit, dane = data_testowy,
                           Y = data_testowy$PCOS, 0.5),
  model_logit2 = miary_pred(model = model_logit2, dane = data_testowy,
                            Y = data_testowy$PCOS, 0.5),
  model_probit = miary_pred(model = model_probit, dane = data_testowy,
                            Y = data_testowy$PCOS,  0.5),
    model_probit2 = miary_pred(model = model_probit2, dane = data_testowy,
                               Y = data_testowy$PCOS,  0.5))
wyniki_miary_pred2
```

```
##                      ACC        ER      SENS      SPEC       PPV       NPV
## model_logit   0.8703704 0.1296296 0.8070175 0.9047619 0.8214286 0.8962264
## model_logit2  0.8765432 0.1234568 0.8070175 0.9142857 0.8363636 0.8971963
## model_probit  0.8888889 0.1111111 0.7894737 0.9428571 0.8823529 0.8918919
## model_probit2 0.8827160 0.1172840 0.7894737 0.9333333 0.8653846 0.8909091
```

```
różnice <- wyniki_miary_pred2 - wyniki_miary_pred
różnice
```

```
##                       ACC         ER        SENS         SPEC          PPV
## model_logit   -0.03679143 0.03679143 -0.01651187 -0.040974529 -0.053571429
## model_logit2  -0.03061859 0.03061859 -0.01651187 -0.031450720 -0.038636364
## model_probit  -0.02092543 0.02092543 -0.04245909 -0.002879291  0.006246746
## model_probit2 -0.02709827 0.02709827 -0.04245909 -0.012403101 -0.010721579
##                       NPV
## model_logit   -0.02452830
## model_logit2  -0.02355846
## model_probit  -0.03235053
## model_probit2 -0.03333333
```

Interpretacja:

Zliczeniowy R^{2} - ACC - Udział liczby trafnie sklasyfikowanych jednostek w ogólnej liczbie jednostek.

Wskaźnik błędu - ER - Udział liczby źle sklasyfikowanych jednostek w ogólnej liczbie jednostek.

Czułość - SENS - Udział liczby trafnie oszacowanych 1 w liczbie wszystkich obserwowanych 1.

Swoistość - SPEC - Udział liczby trafnie oszacowanych 0 w liczbie wszystkich obserwowanych 0.

Dodatnia zdolność predykcyjna - PPV - Udział liczby trafnie oszacowanych 1 w liczbie wszystkich prognozowanych 1.

Ujemna zdolność predykcyjna - NPV - Udział liczby trafnie oszacowanych 0 w liczbie wszystkich prognozowanych 0.

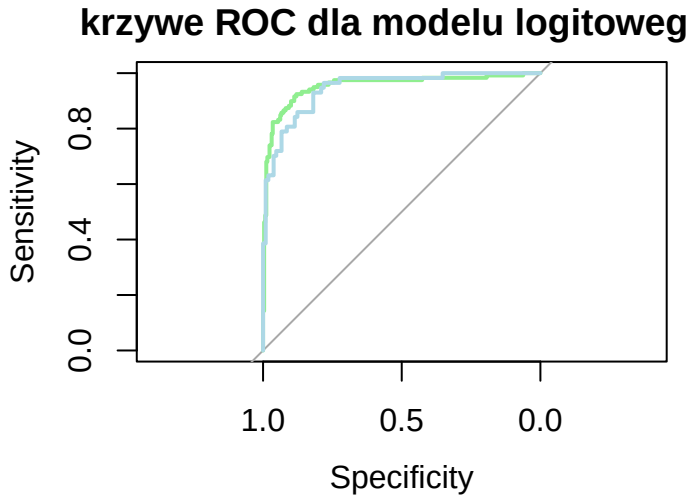
Wniosek: W związku z tym, że otrzymane wyniki niewiele się różnią to nie sugerujemy się nimi, tylko wynikami statystycznymi, otrzymanymi powyżej.

Krzywa ROC:

Krzywa zielona - ROC wyznaczona na zbiorze uczącym Krzywa niebieska - ROC wyznaczona na zbiorze testowym

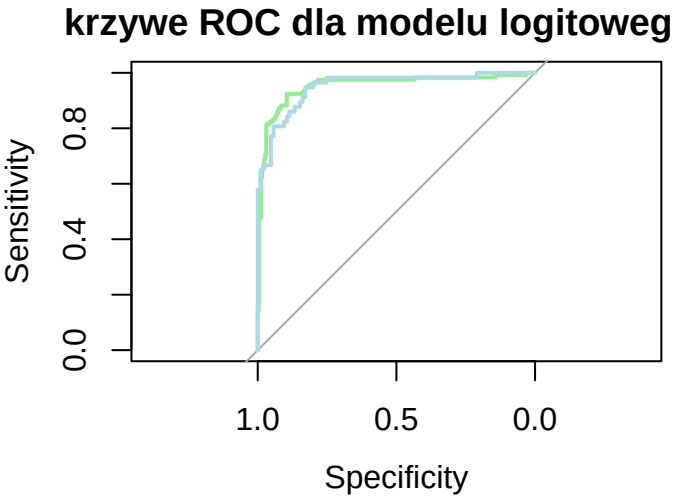
Model_logit:

```
roc_logit <- roc(model_logit$y, model_logit$fitted.values)
roc_logit_pred <- roc(data_testowy$PCOS, predict(model_logit, data_testowy,
                                                 type = "response"))
plot(roc_logit, main = "krzywe ROC dla modelu logitowego:", col = "lightgreen")
lines(roc_logit_pred, col = "lightblue")
```



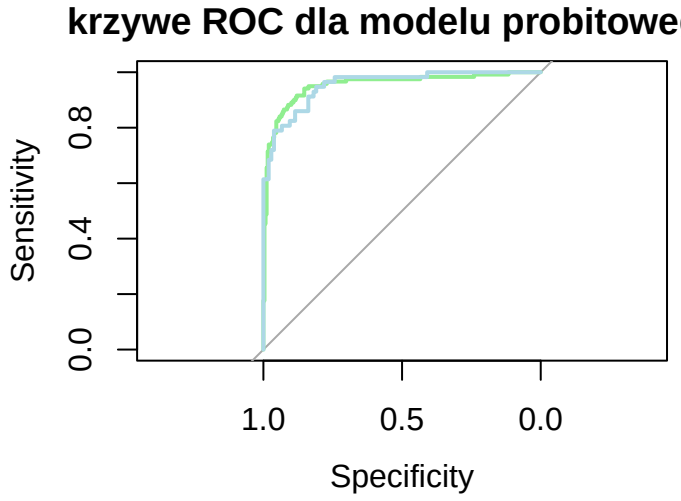
Model_logit_2:

```
roc_logit2 <- roc(model_logit2$y, model_logit2$fitted.values)
roc_logit2_pred <- roc(data_testowy$PCOS, predict(model_logit2, data_testowy,
                                                  type = "response"))
plot(roc_logit2, main = "krzywe ROC dla modelu logitowego2:", col = "lightgreen")
lines(roc_logit2_pred, col = "lightblue")
```



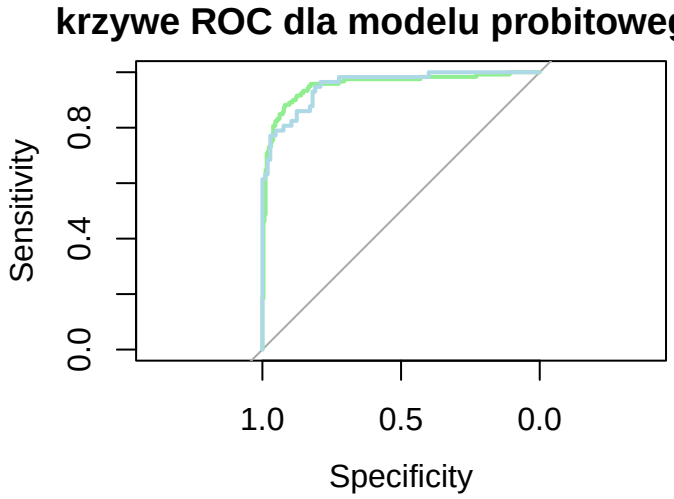
Model_probit:

```
roc_probit <- roc(model_probit$y, model_probit$fitted.values)
roc_probit_pred <- roc(data_testowy$PCOS, predict(model_probit, data_testowy,
                                                  type = "response"))
plot(roc_probit, main = "krzywe ROC dla modelu probitowego:", col = "lightgreen")
lines(roc_probit_pred, col = "lightblue")
```



Model_probit_2:

```
roc_probit2 <- roc(model_probit2$y, model_probit2$fitted.values)
roc_probit2_pred <- roc(data_testowy$PCOS, predict(model_probit2, data_testowy,
                                                   type = "response"))
plot(roc_probit2, main = "krzywe ROC dla modelu probitowego2:", col = "lightgreen")
lines(roc_probit2_pred, col = "lightblue")
```



Wniosek:

Wszystkie wykresy krzywej ROC są podobne. Każdy z nich mówi o bardzo dobrej lub nawet prawie doskonałej jakości predykcji.

Pole powierzchni pod krzywą ROC:

```
cat("AUC dla zbioru uczącego:\n")
```

```
## AUC dla zbioru uczącego:
```

```
auc(roc_logit)
```

```
## Area under the curve: 0.9542
```

```
auc(roc_logit2)
```

```
## Area under the curve: 0.9535
```

```
auc(roc_probit)
```

```
## Area under the curve: 0.9541
```

```
auc(roc_probit2)
```

```
## Area under the curve: 0.9538
```

```
cat("\nAUC dla zbioru testowego:\n")
```

```
##
## AUC dla zbioru testowego:
```

```
auc(roc_logit_pred)
```

```
## Area under the curve: 0.9452
```

```
auc(roc_logit2_pred)
```

```
## Area under the curve: 0.9509
```

```
auc(roc_probit_pred)
```

```
## Area under the curve: 0.9544
```

```
auc(roc_probit2_pred)
```

```
## Area under the curve: 0.953
```

Wniosek:

Na podstawie wielkości pola pod krzywą ROC dla każdego z modeli można stwierdzić, że jakość predykcji jest doskonała. Minimalnie lepszy w tym przypadku jest model_probit.

Interpretacja modelu:

```
model_probit$coefficients
```

```
##          (Intercept)    rosniecie_wlosow1  wypryski_na_twarzy1
##        -3.7465022779         0.8271495443         0.6354547573
##    ciemnienie_skory1 nieregularnosc_cyklu           poziom_TSH
##         0.8709578269         0.6929611286        -0.0003833205
##           poziom_FSH            poziom_LH     Follicle_średnia
##        -0.0167103458         0.0625593643         0.2767502602
##           poziom_AMH           poziom_PRL           poziom_PRG
##         0.0278228648         0.0017417604        -0.1191790783
```

Interpretacja modelu probitowego sprowadza się do określenia czy zmienna objaśniająca jest stymulantą czy destymulantą.

Stymulanta - wzrost wartości zmiennej niezależnej świadczy o wzroście poziomu zmiennej zależnej, a spadek wartości świadczy o spadku poziomu zmiennej zależnej.

Destymulanta - wzrost wartości zmiennej niezależnej świadczy o spadku poziomu zmiennej zależnej, a spadek wartości świadczy o wzroście wartości zmiennej zależnej.

Zatem zmienne: rosniecie_wlosow, wypryski_na_twarzy, ciemnienie_skory, nieregularnosc_cyklu, poziom_LH, Follicle_średnia, poziom_AMH, poziom_PRL są stymulantami, ponieważ wartości współczynników sa większe od 0.

Zmienne: poziom_TSH, poziom_FSH oraz poziom_PRG są destymulantami, ponieważ wartości współczynników są mniejsze od 0.

Estymacja modelu z interakcją:

Zweryfikowanie, które zmienne odnoszące się do zmian wpływających na wygląd ciała istotnie statystycznie wpływają na występowanie Zespołu Policystycznych Jajników.

```
model_inter <- glm(PCOS ~ BMI_kat + rosniecie_wlosow + wypadanie_wlosow +
wypryski_na_twarzy + ciemnienie_skory + przybranie_na_wadze +
wypadanie_wlosow*rosniecie_wlosow, data = data, family = binomial)

summary(model_inter)
```

```
##
## Call:
## glm(formula = PCOS ~ BMI_kat + rosniecie_wlosow + wypadanie_wlosow +
##     wypryski_na_twarzy + ciemnienie_skory + przybranie_na_wadze +
##     wypadanie_wlosow * rosniecie_wlosow, family = binomial, data = data)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -2.2286     0.5076  -4.390 1.13e-05 ***
## BMI_katpoprawna                     -0.5402     0.4999  -1.081  0.27988
## BMI_katnadwaga                      -1.0607     0.5538  -1.915  0.05544 .
## BMI_katotyłość                      -0.5789     0.6633  -0.873  0.38282
## rosniecie_wlosow1                    1.0514     0.3704   2.839  0.00453 **
## wypadanie_wlosow1                   -0.3290     0.3038  -1.083  0.27882
## wypryski_na_twarzy1                  1.0291     0.2510   4.100 4.14e-05 ***
## ciemnienie_skory1                    1.5026     0.2499   6.012 1.83e-09 ***
## przybranie_na_wadze1                 1.5860     0.3097   5.122 3.03e-07 ***
## rosniecie_wlosow1:wypadanie_wlosow1  1.0343     0.5237   1.975  0.04829 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 680.97  on 538  degrees of freedom
## Residual deviance: 438.73  on 529  degrees of freedom
## AIC: 458.73
##
## Number of Fisher Scoring iterations: 5
```

Ilorazy szans:

```
round(exp(model_inter$coefficients), 4)
```

```
##                          (Intercept)                       BMI_katpoprawna
##                               0.1077                               0.5826
##                       BMI_katnadwaga                        BMI_katotyłość
##                               0.3462                               0.5605
##                    rosniecie_wlosow1                     wypadanie_wlosow1
##                               2.8618                               0.7196
##                  wypryski_na_twarzy1                     ciemnienie_skory1
##                               2.7987                               4.4935
##                 przybranie_na_wadze1 rosniecie_wlosow1:wypadanie_wlosow1
##                               4.8840                               2.8130
```

Wzór modelu:

$$logit(p) = -2.7736 - 0.5047 \cdot BMI\_kat\_poprawna + 0.5110 \cdot BMI\_kat\_nadwaga + 0.2197 \cdot BMI\_kat\_otyłość$$
$$+1.0514 \cdot rosniecie\_wlosow1 - 0.3290 \cdot wypadanie\_wlosow1 + 1.0291 \cdot wypryski\_na\_twarzy1$$
$$+1.5026 \cdot ciemnienie\_skory1 + 1.5860 \cdot przybranie\_na\_wadze1 + 1.0343 \cdot rosniecie\_wlosow1 : wypadanie\_wlosow1$$

Intercept – Wśród osób z niedowagą, którym nie rosną nadmiernie włosy, nie wypadają nadmiernie włosy, nie mają wyprysków na twarzy, nie ciemnieje skóra ani nie przybrały na wadze, szansa zachorowania na PCOS wynosi 0.0624.

rosniecie_wlosow1(TAK) – Wśród osób skarżących się na wypadanie włosów, którym rosną nadmiernie włosy, szansa zachorowania na PCOS jest prawie 3 razy większa niż szansa zachorowania na PCOS wśród osób, którym nie rosną nadmiernie włosy, przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

wypryski_na_twarzy1(TAK) - Wśród osób, które mają wypryski na twarzy, szansa zachorowania na PCOS jest prawie 3 razy większa niż szansa zachorowania na PCOS wśród osób, które nie mają wyprysków na twarzy, przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

ciemnienie_skory1(TAK) - Wśród osób, którym ciemnieje skóra, szansa zachorowania na PCOS jest prawie 4.5 razy większa niż szansa zachorowania na PCOS wśród osób, którym nie ciemnieje skóra, przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

przybranie_na_wadze1(TAK) - Wśród osób, które przybrały na wadze, szansa zachorowania na PCOS jest prawie 5 razy większa niż szansa zachorowania na PCOS wśród osób, które nie przybrały na wadze, przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

moderator - wypadanie_wlosów:

rosniecie_wlosów1(TAK):wypadanie_wlosow1(TAK) –

$$\frac{\frac{rosniecie\_wlosow\_TAK}{rosniecie\_wlosow\_NIE} wypadanie\_wlosow\_TAK}{\frac{rosniecie\_wlosow\_TAK}{rosniecie\_wlsow\_NIE} wypadanie\_wlosow\_NIE}$$

Iloraz szans zachorowania na PCOS wśród pacjentów, którym rosną nadmiernie włosy względem osób, którym nie rosną nadmiernie włosy w grupie osób, którym wypadają włosy jest prawie 3 razy większy niż wśród osób, którym nie wypadają włosy przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

moderator - rosniecie_wlosów:

rosniecie_wlosów_1(TAK):wypadanie_wlosow_1(TAK) -

$$\frac{\frac{wypadanie\_wlosow\_TAK}{wypadanie\_wlosow\_NIE} rosniecie\_wlosow\_TAK}{\frac{wypadanie\_wlosow\_TAK}{wypadanie\_wlsow\_NIE} rosniecie\_wlosow\_NIE}$$

Iloraz szans zachorowania na PCOS wśród pacientów, którym wypadają nadmiernie włosy względem osób, którym nie wypadają nadmiernie włosy w grupie osób, którym rosną nadmiernie włosy jest prawie 3 razy większa niż wśród osón, którym nie rosną nadmiernie włosy przy założeniu takich samych wartości pozostałych zmiennych (ceteris paribus).

pozostałe zmienne - W przypadku pozostałych zmiennych interpretacja nie ma sensu, ponieważ nie różni się ona statystycznie istotnie od interpretacji wyrazu wolnego.

Estymacja modelu wielomianowego porządkowego:

Potrzebna biblioteka:

```
library("MASS")
```

W budowie modelu wielomianowego porządkowego zdecydowałyśmy się użyć jako zmiennej objaśnianej, wcześniej utworzonej, nowej zmiennej kategorialnej BMI_kat, która posiada 4 warianty określające wagę ze względu na wartość BMI - "niedowaga", "poprawna", "nadwaga", "otyłość". Jako zmienną objaśniającą wybrałyśmy wypryski_na_twarzy.

```
data$BMI_kat <- as.ordered(data$BMI_kat)
model_wielom_porz <- polr(BMI_kat ~ wypryski_na_twarzy , data = data)
summary(model_wielom_porz)
```

```
## Call:
## polr(formula = BMI_kat ~ wypryski_na_twarzy, data = data)
##
## Coefficients:
##                       Value Std. Error t value
## wypryski_na_twarzy1 0.08264     0.1645  0.5025
##
## Intercepts:
##                     Value    Std. Error t value
## niedowaga|poprawna  -2.6586   0.1936   -13.7317
## poprawna|nadwaga     0.3431   0.1186     2.8938
## nadwaga|otyłość      2.4864   0.1790    13.8890
##
## Residual Deviance: 1170.366
## AIC: 1178.366
```

Wzory:

$$logit(P(Y \leqslant 1)) = -2.6586 + 0.08264 \cdot wypryski\_na\_twarzy1$$
$$logit(P(Y \leqslant 2)) = 0.3431 + 0.08264 \cdot wypryski\_na\_twarzy1$$
$$logit(P(Y \leqslant 3)) = 2.4864 + 0.08264 \cdot wypryski\_na\_twarzy1$$

Interpretacja modelu:

Ilorazy szans:

```
round(exp(model_wielom_porz$coefficients), 4)
```

```
## wypryski_na_twarzy1
##              1.0862
```

```
round(exp(model_wielom_porz$zeta), 4)
```

```
## niedowaga|poprawna   poprawna|nadwaga    nadwaga|otyłość
##             0.0700             1.4094            12.0177
```

wypryski_na_twarzy1(TAK) – Szansa, że BMI będzie na niższych poziomach jest 8,62% większa w grupie osób z wypryskami na twarzy niż u osób bez wyprysków.

Intercept 1 - niedowaga|poprawna – W grupie osób bez wyprysków na twarzy szansa, że wystąpi niedowaga wynosi 0.07.

Intercept 2 - poprawna|nadwaga – W grupie osób bez wyprysków na twarzy szansa, że wystąpi niedowaga lub poprawna waga wynosi 1.4094.

Intercept 3 - nadwaga|otyłość - W grupie osób bez wyprysków na twarzy szansa, że wystąpi niedowaga, poprawna waga lub nadwaga wynosi 12.0177.

Podsumowanie

Merytoryczna ocena modelu dwumianowego logitowego:

Brak istotności statystycznej zmiennych dotyczących poziomu hormonów wydaje się błędny. Zespół Policystycznych Jajników cechuje się zaburzeniami hormonalnymi i wykrycie tych nieprawidłowości pomaga dokonać poprawnej diagnozy.

Według miary pseudo R^2 modelem o najlepszym dopasowaniu jest model_logit (model z większą ilością zmiennych objaśniających). Cechuje się on jednak większą wartością kryterium AIC w porównaniu do modelu z mniejszą ilością zmiennych objaśniających.

Zgodnie z wynikami miar jakości predykcji opartych na tablicy trafności oraz krzywej ROC, zdolność predykcyjna modeli logitowych i probitowych jest niemal doskonała.

Merytoryczna ocena modelu z interakcją:

Zmienną nieistotną statystycznie okazała się zmienna dotycząca poziomu BMI pacjentów, co wydaje się wynikiem błędnym, ponieważ szacuje się że ok. 40-80% osób, u których stwierdzono występowanie Zespołu Policystycznych Jajników zmaga się z nadwagą. Występowanie Zespołu Policystycznych Jajników również idzie w parze z szeregiem zaburzeń metabolicznych.

Merytoryczna ocena modelu wielomianowego porządkowego:

Dodatkowo stworzyłyśmy model wielomianowy porządkowy z utworzoną przez nas zmienną kategorialną BMI. Model nie wnosi informacji na temat zachorowania na Zespół Policystycznych Jajników, ale pokazuje ciekawy wniosek, mówiący o zależności między niższą wagą a problemami ze stanem skóry. Merytorycznie nie wydaje się, aby między tymi zmiennymi występowała jakakolwiek prawidłowość i powodem takiego wyniku jest prawdopodobnie grupa badawcza, u której zachodziły takie tendencje.