Testowanie hipotez badawczych na danych klinicznych

Klaudia Kozakiewicz

Opis zbioru danych

Niniejszy zestaw danych zawiera informacje demograficzne oraz kliniczne pacjentów, dzieląc ich na osoby posiadające i nieposiadające zespół metaboliczny - zbiór zaburzeń zdrowotnych, które w dalszych konsekwencjach mogą prowadzić do poważnych chorób, takich jak miażdżyca, marskość wątroby, a nawet do śmierci.

Zmienne

zmienna	opis
`seqn`	numer identyfikacyjny
`Age`	wiek badanego (w latach)
`Sex`	płeć badanego (kobieta/mężczyzna)
`Marital`	stan cywilny jednostki
`Income`	dochód
`Race`	rasa
`WaistCirc`	pomiar obwodu talii (w cm)
`BMI`	wskaźnik masy ciała
`Albuminuria`	indykator mówiący o ilości albumin (białek) w moczu (wartości: 0, 1, 2)
`UrAlbCr`	stosunek albuminy do kreatyniny w moczu
`UricAcid`	poziom kwasu moczowego we krwi (w mg/dl)
`BloodGlucose`	poziom glukozy we krwi (w mg/dl)
`HDL`	poziom HDL ("dobrego" cholesterolu) (w mg/dl)
`Triglycerides`	poziom trójglicerydów we krwi (w mg/dl)
`MetabolicSyndrome`	indykator mówiący o obecności zespołu metabolicznego (0 = brak, 1 = obecność)

Eksploracja i czyszczenie danych

Wczytuję dane, używając argumentu `stringsAsFactors = TRUE`, dzięki któremu wartości słowne zostaną automatycznie zamienione na zmienne kategoryczne. Dodatkowo zamiany wymagają zmienne liczbowe `Albuminuria` i `MetabolicSyndrome`. Następnie, ponieważ dysponuję dużą ilością rekordów, usuwam te, w których niektóre zmienne przyjmują wartość NA i usuwamy zmienną `seqn` zawierającą numer pacjenta. Na koniec przeanalizuję podsumowanie zbioru danych:

```
# ładowanie zbioru danych i refaktoryzacja
ms <- read.csv("ms.csv", header = TRUE, sep = ";", stringsAsFactors = TRUE)

factors <- c("Albuminuria", "MetabolicSyndrome")
ms[,factors] <- lapply(ms[factors], as.factor)

# usuwanie wartości NA i pierwszej kolumny
ms[ms == ""] <- NA
```

```r
ms <- na.omit(ms)
ms <- ms[,-1]

# wyświetlenie podsumowania
summary(ms)
```

```
##       Age            Sex            Marital         Income
##  Min.   :20.00   Female:1022   Divorced : 219   Min.   : 300
##  1st Qu.:35.00   Male  : 987   Married  :1098   1st Qu.:1600
##  Median :49.00                 Separated:  88   Median :3500
##  Mean   :49.26                 Single   : 460   Mean   :4147
##  3rd Qu.:63.00                 Widowed  : 144   3rd Qu.:6200
##  Max.   :80.00                                  Max.   :9000
##           Race       WaistCirc           BMI        Albuminuria
##  Asian      :295   Min.   : 63.10   Min.   :15.70   0:1761
##  Black      :462   1st Qu.: 86.90   1st Qu.:24.10   1: 200
##  Hispanic   :198   Median : 97.10   Median :27.70   2:  48
##  MexAmerican:198   Mean   : 98.52   Mean   :28.73
##  Other      : 50   3rd Qu.:107.80   3rd Qu.:32.10
##  White      :806   Max.   :170.50   Max.   :68.70
##     UrAlbCr           UricAcid       BloodGlucose      HDL
##  Min.   :   1.40   Min.   : 1.800   Min.   : 39   Min.   : 14.00
##  1st Qu.:   4.46   1st Qu.: 4.500   1st Qu.: 92   1st Qu.: 43.00
##  Median :   6.96   Median : 5.400   Median :100   Median : 51.00
##  Mean   :  42.25   Mean   : 5.491   Mean   :108   Mean   : 53.55
##  3rd Qu.:  13.49   3rd Qu.: 6.400   3rd Qu.:110   3rd Qu.: 62.00
##  Max.   :4462.81   Max.   :11.300   Max.   :382   Max.   :150.00
##  Triglycerides    MetabolicSyndrome
##  Min.   :  26.0   0:1297
##  1st Qu.:  75.0   1: 712
##  Median : 103.0
##  Mean   : 126.9
##  3rd Qu.: 149.0
##  Max.   :1311.0
```

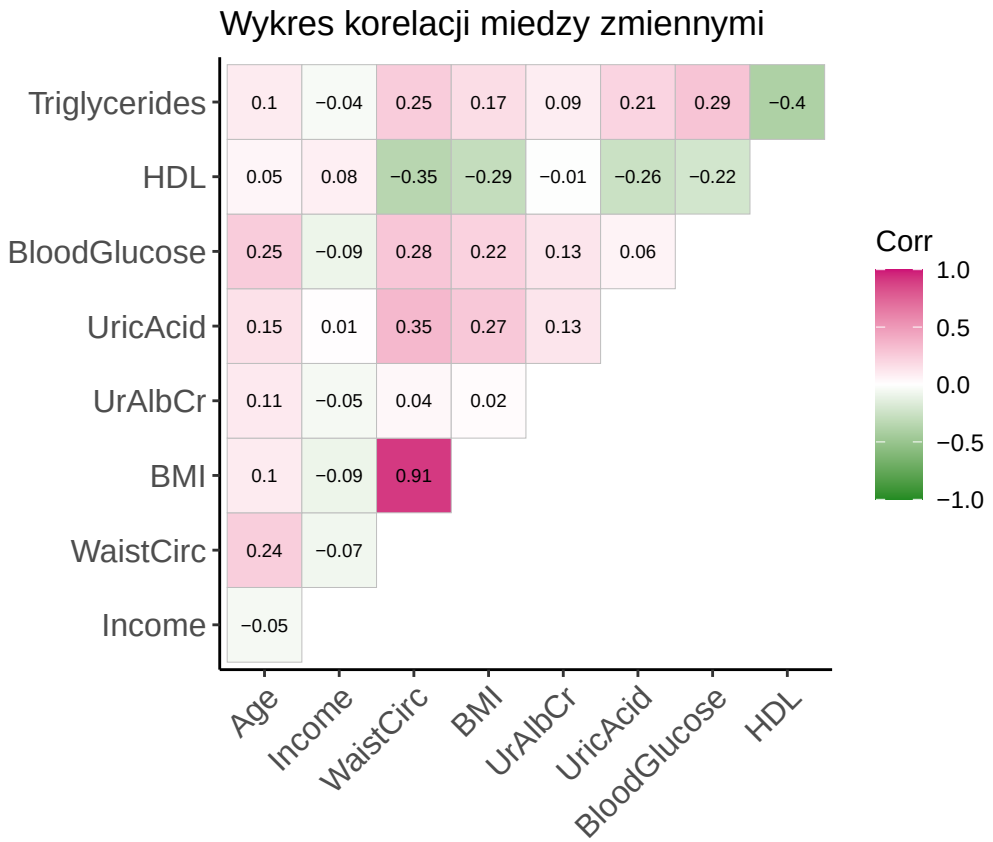
Analizując podsumowanie, dochodzę do poniższych wniosków:

- wiek badanych mieści się w przedziale od 20 do 80 lat, przy czym mediana wynosi 49 lat,
- rozkład płci to w przybliżeniu 51% do 49%, kobiet jest nieznacznie więcej,
- dominującym statusem cywilnym jest "Married", czyli żonaty/mężatka, a rasą - biała,
- w zestawieniu zawarte są dane osób skrajnie wychudzonych (BMI < 16), jak i takich ze skrajną otyłością (BMI > 40),
- albuminuria to zjawisko, gdy organizm wydala z organizmu białka wraz z moczem. Jest to naturalne, jednak wysokie stężenie albumin w moczu może być objawem nieprawidłowości w pracy nerek, spowodowane np. cukrzycą lub nieskutecznie leczonym nadciśnieniem tętniczym. W naszym zbiorze stężenie białek jest indykatorem o wartościach: 0 - normalne stężenie albumin, 1 - podwyższony poziom, 2 - bardzo wysoki poziom albumin. Osoby z podwyższonym poziomem stanowią mniejszość,
- zmienna `UricAcid` informuje o ilości kwasu moczowego we krwi. Jego podwyższone ilości (między 6 a 7 mg/dl w zależności od płci) mogą być skutkiem niewłaściwej diety, otyłości lub cukrzycy. U nas mniej więcej 3/4 osób ma poziom kwasu moczowego w normie,

- poziom glukozy w krwi również przedstawia osoby o skrajnych wynikach - od bardzo niskich, po alarmująco wysokie. Mediana tej wartości wynosi 100 mg/dl, co jest normalną, zdrową wartością,
- HDL to tzw. "zdrowy" cholesterol. W zbiorze danych ok. połowa osób ma za niski lub zdecydowanie za niski poziom HDL (zależnie od płci granica nieprawidłowej ilości HDL wynosi między 40-50 mg/dl),
- wysoki poziom trójglicerydów, czyli inaczej tłuszczów, stanowi główną przyczynę rozwoju miażdżycy, zawału serca lub udaru. Prawidłowe stężenie trójglicerydów nie powinno przekraczać 150 mg/dl, z kolei wartości powyżej 500 są już bardzo niebezpieczne dla zdrowia.

Sprawdźmy, które zmienne liczbowe są ze sobą skorelowane:

```
model.matrix(~0+., data=ms %>%
select(Age, Income, WaistCirc, BMI, UrAlbCr, UricAcid, BloodGlucose, HDL,
       Triglycerides)) %>% cor(use="pairwise.complete.obs") %>%
  ggcorrplot(title = "Wykres korelacji między zmiennymi", type="upper", lab=TRUE,
             colors = c("forestgreen", "white", "deeppink3"), lab_size=2.5,
             ggtheme = ggplot2::theme_classic)
```



Największą korelację wykazują zmienne `BMI` i `WaistCirc` - wskaźnik masy ciała jest ściśle skorelowany z obwodem w pasie. Warto zwrócić też uwagę na korelacje zachodzące między poziomem `HDL` a m.in obwodem w pasie, ilością kwasu moczowego we krwi, poziomem glukozy i trójglicerydami.

Postawienie hipotez badawczych

- **Hipoteza 1:** Model przewidujący poziom HDL oparty na jego 2 najlepszych predyktorach będzie lepiej dopasowany do danych, niż te bazujące tylko na jednym z nich.
- **Hipoteza 2:** Przewidując poziom glukozy we krwi za pomocą modelu z interakcją wskaźnika BMI oraz informacji nt. posiadania *dyslipidemi aterogennej*, otrzymamy lepiej dopasowany model, niż model bez interakcji.
- **Hipoteza 3:** Istnieją istotne różnice w obowdzie w talii, przewidzianym za pomocą BMI między osobami, które mają *syndrom metaboliczy*, a tymi, które nie mają tego syndromu.

Hipoteza 1

Na podstawie macierzy korelacji wybieram 2 najbardziej skorelowane zmienne z HDL: poziom trójglicerydów i obwód w pasie. Należy zauważyć, że w obu przypadkach poziom korelacji jest nieistotny statystycznie (poniżej 0.5), a między zmiennymi `Triglycerides` i `WaistCirc` zachodzi korelacja.

```
# porównanie modeli opierających się na 1 zmiennej wyjaśniającej: trójglicerydach lub obwodzie
model.tri <- lm(HDL~Triglycerides, ms)
model.waist <- lm(HDL~WaistCirc, ms)

summary(model.tri)
```

```
##
## Call:
## lm(formula = HDL ~ Triglycerides, data = ms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.233  -9.621  -2.420   7.107  92.906
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.051452   0.531348   116.8   <2e-16 ***
## Triglycerides -0.066989   0.003418   -19.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.76 on 2007 degrees of freedom
## Multiple R-squared:  0.1606, Adjusted R-squared:  0.1602
## F-statistic: 384.1 on 1 and 2007 DF,  p-value: < 2.2e-16
```

```
summary(model.waist)
```

```
##
## Call:
## lm(formula = HDL ~ WaistCirc, data = ms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.707  -9.681  -2.041   7.452  92.765
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 85.60267    1.91983   44.59   <2e-16 ***
## WaistCirc   -0.32532    0.01922  -16.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 2007 degrees of freedom
## Multiple R-squared:  0.1249, Adjusted R-squared:  0.1244
## F-statistic: 286.4 on 1 and 2007 DF,  p-value: < 2.2e-16
```

Porównując podsumowania tych dwóch modeli regresji liniowej mogę zauważyć, że model bazujący na trójglicerydach wyjaśnia 16% zmienności, a różnica w poziomie HDL między dwoma osobami różniącymi się w wynikach trójglicerydów o 1 mg/dl wynosi 0.067. Model na podstawie obwodu w pasie wyjaśnia ok. 12,5% zmienności.

```
model.combined <- lm(HDL~WaistCirc+Triglycerides, ms)
summary(model.combined)
```

```
##
## Call:
## lm(formula = HDL ~ WaistCirc + Triglycerides, data = ms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.146  -9.106  -1.891   6.547  90.689
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   85.102280   1.802779   47.21   <2e-16 ***
## WaistCirc     -0.248510   0.018644  -13.33   <2e-16 ***
## Triglycerides -0.055693   0.003385  -16.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.19 on 2006 degrees of freedom
## Multiple R-squared:  0.2289, Adjusted R-squared:  0.2282
## F-statistic: 297.8 on 2 and 2006 DF,  p-value: < 2.2e-16
```

Ten model wyjaśnia większy procent zmienności w `HDL`, bo prawie 23%. Mogę zauważyć, że RSE jest niższy niż w przypadku pozostałych dwóch modeli i wynosi 13.19 (model lepiej dopasowuje się do danych). Zarówno `Triglycerides`, jak i `WaistCirc` są negatywnie skorelowane z poziomem cholesterolu. Stwierdzam, że obie zmienne są istotne w wyjaśnianiu zmienności poziomu HDL, a model uwzględniający obie zmienne jest lepszy niż modele zawierające po jednej zmiennej objaśniającej.

Hipoteza 2

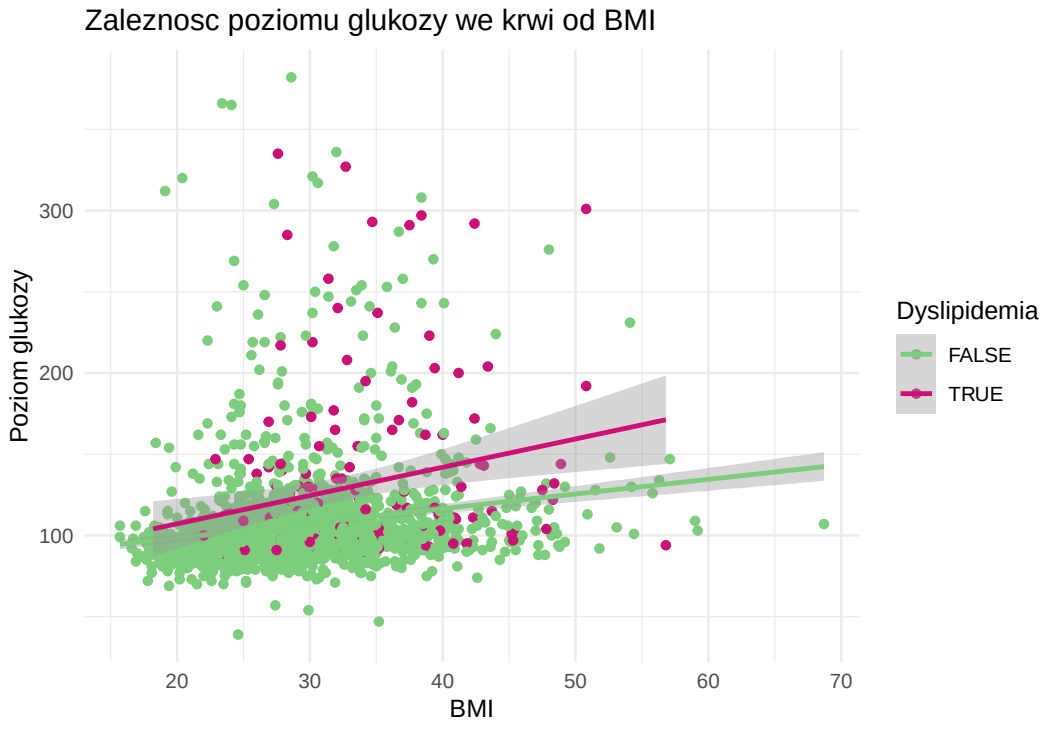
Dyslipidemia aterogenna to zaburzenie lipidowe, które objawia się podwyższonym stężeniem trójglicerydów (powyżej 150 mg/dl) we krwi, przy jednoczesnym niskim stężeniu cholesterolu HDL (poniżej 40 mg/dl). Chcę sprawdzić, czy model z interakcją przewidujący poziom glukozy we krwi oparty na wskaźniku BMI i zmiennej kategorycznej indykującej obecność lub brak dyslipidemi będzie "lepszy" niż model bez interakcji.

Oczekuję, że wpływ BMI na poziom glukozy zmienia się w zależności od tego, czy osoba ma zaburzenie lipidowe.

Zaczynam od stworzenia nowej zmiennej `Dyslipidemia`, która będzie przyjmowała wartości TRUE/FALSE zależnie od tego, czy osoba ma to zaburzenie. Następnie sprawdzę, jak na wykresie prezentują się 3 zmienne:

```
dyslipidemia <- function(HDL, Triglycerides) {
  dyslipidemia <- FALSE

  if (Triglycerides >= 150 & HDL < 40) {
    dyslipidemia <- TRUE
  }

  return(dyslipidemia)
}

# dodanie nowej zmiennej
ms$Dyslipidemia <- mapply(dyslipidemia, ms$HDL, ms$Triglycerides)

# wykres zależności glukozy od BMI, dodając informacje o dyslipidemi
ggplot(ms, aes(x = BMI, y = BloodGlucose, color = Dyslipidemia)) +
geom_point() +
geom_smooth(method = "lm", se = TRUE) +
labs(title="Zależność poziomu glukozy we krwi od BMI", x = "BMI",
y = "Poziom glukozy") +
theme_minimal() +
scale_color_manual(values = c("FALSE" = "palegreen3", "TRUE" = "deeppink3"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Na podstawie wykresu ciężko jednoznacznie stwierdzić, czy zmienna kategoryczna istotnie wpływa na zmienną ciągłą - linie regresji nie przecinają się, jednak widać różnicę w ich nachyleniu. Wydaje się jednak, że istnieje istotny wpływ miedzy tymi zmiennymi. Sprawdźmy, jak prezentują się podsumowania modeli z i bez interakcji:

```
# tworzenie modeli z interakcją i bez
model.wint <- lm(BloodGlucose ~ BMI + Dyslipidemia, ms)
model.int <- lm(BloodGlucose ~ BMI * Dyslipidemia, ms)
summary(model.wint)
```

```
##
## Call:
## lm(formula = BloodGlucose ~ BMI + Dyslipidemia, data = ms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.742 -14.494  -6.634   3.097 275.755
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       78.0900     3.2563  23.982  < 2e-16 ***
## BMI                0.9844     0.1116   8.824  < 2e-16 ***
## DyslipidemiaTRUE  18.7355     2.6026   7.199 8.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.42 on 2006 degrees of freedom
## Multiple R-squared:  0.07204,    Adjusted R-squared:  0.07111
## F-statistic: 77.86 on 2 and 2006 DF,  p-value: < 2.2e-16
```

Wartość DyslipidemiaTRUE (18.7355) oznacza różnicę w średnim poziomie glukozy między grupą osób z dyslipidemią (DyslipidemiaTRUE) a grupą bez tego zaburzenia, przy założeniu stałego BMI. Wartość BMI (0.9844) oznacza, że przy założeniu braku dyslipidemii, każda jednostka wzrostu BMI jest związana ze wzrostem średniego poziomu glukozy o 0.9844 jednostki.

```
summary(model.int)
```

```
##
## Call:
## lm(formula = BloodGlucose ~ BMI * Dyslipidemia, data = ms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.236 -14.429  -6.694   2.988 275.772
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           80.4336     3.4168  23.541  < 2e-16 ***
## BMI                    0.9019     0.1174   7.684  2.4e-14 ***
## DyslipidemiaTRUE      -8.0718    12.2361  -0.660   0.5095
## BMI:DyslipidemiaTRUE   0.8389     0.3741   2.242   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 32.39 on 2005 degrees of freedom
## Multiple R-squared:  0.07436,    Adjusted R-squared:  0.07297
## F-statistic: 53.69 on 3 and 2005 DF,  p-value: < 2.2e-16
```

Intercept (80.43) to wartość średniego poziomu glukozy dla osób bez dyslipidemii (DyslipidemiaFALSE) i BMI równego zeru. BMI (0.9019) oznacza wpływ jednostkowego wzrostu BMI na poziom glukozy dla osób bez dyslipidemii. DyslipidemiaTRUE (-8.0718) oznacza różnicę w średnim poziomie glukozy między grupą osób z dyslipidemią a grupą bez tego zaburzenia, ale przy BMI równym zeru. BMI:DyslipidemiaTRUE (0.8389) to efekt interakcji między BMI a dyslipidemią. Oznacza to, że wpływ wzrostu BMI na poziom glukozy jest różny w zależności od obecności dyslipidemii.

Oba modele wyjaśniają mały ułamek zmienności `BloodGlucose`, przy czym model z interakcją wypada odrobinę lepiej i wyjaśnia ok. 7,4% zmienności. Analizując oba modele, możemy zauważyć, że dodanie interakcji nieznacznie poprawia dopasowanie modelu do danych. Jednakże, istotność statystyczna współczynnika interakcji (BMI:DyslipidemiaTRUE) jest również ważna (p-value = 0.0251), co sugeruje, że wpływ BMI na BloodGlucose różni się w zależności od obecności dyslipidemii.

Hipoteza 3

Chcę opracować procedurę pozwalającą oszacować rozmiar obwodu w talii na podstawie BMI danej osoby. Użyję tej procedury, żeby przewidzieć jaki rozmiar talii będzie miała losowo wybrana część osób z używanego zbioru danych. Następnie, aby zbadać czy istnieją istotne różnice w wielkości tej zmiennej, na jednym histogramie przedstawię osoby, które nie mają syndromu metabolicznego i te, które go mają.

```r
# Podział danych na dwa zbiory - treningowy i testowy - zawierające tylko
# potrzebne komórki

# Tworzenie zbioru treningowego z losowo wybranymi 1009 wierszami
set.seed(123)  # Ustawienie ziarna dla powtarzalności wyników
random_index <- sample(1:nrow(ms), 1009, replace = FALSE)
train_set <- ms[random_index, c("WaistCirc", "BMI", "MetabolicSyndrome")]

# Tworzenie zbioru testowego z pozostałymi 1000 wierszami
test_set <- ms[-random_index, c("BMI", "MetabolicSyndrome")]

# Dzielę ten zbiór pod względem tego, czy osoby mają syndrom metaboliczy
set_MS0 <- test_set[test_set$MetabolicSyndrome == 0, ]
set_MS1 <- test_set[test_set$MetabolicSyndrome == 1, ]

# Tworzę model na zbiorze treningowym
model <- lm(WaistCirc ~ BMI, train_set)

# Używam funkcji predict dla zbiorów testowych
predicted_waist_MS0 <- predict(model, set_MS0)
predicted_waist_MS1 <- predict(model, set_MS1)

# Zocaczmy podsumowanie dla predykcji
summary(predicted_waist_MS0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   71.41   85.32   92.27   94.13  100.18  155.76
```

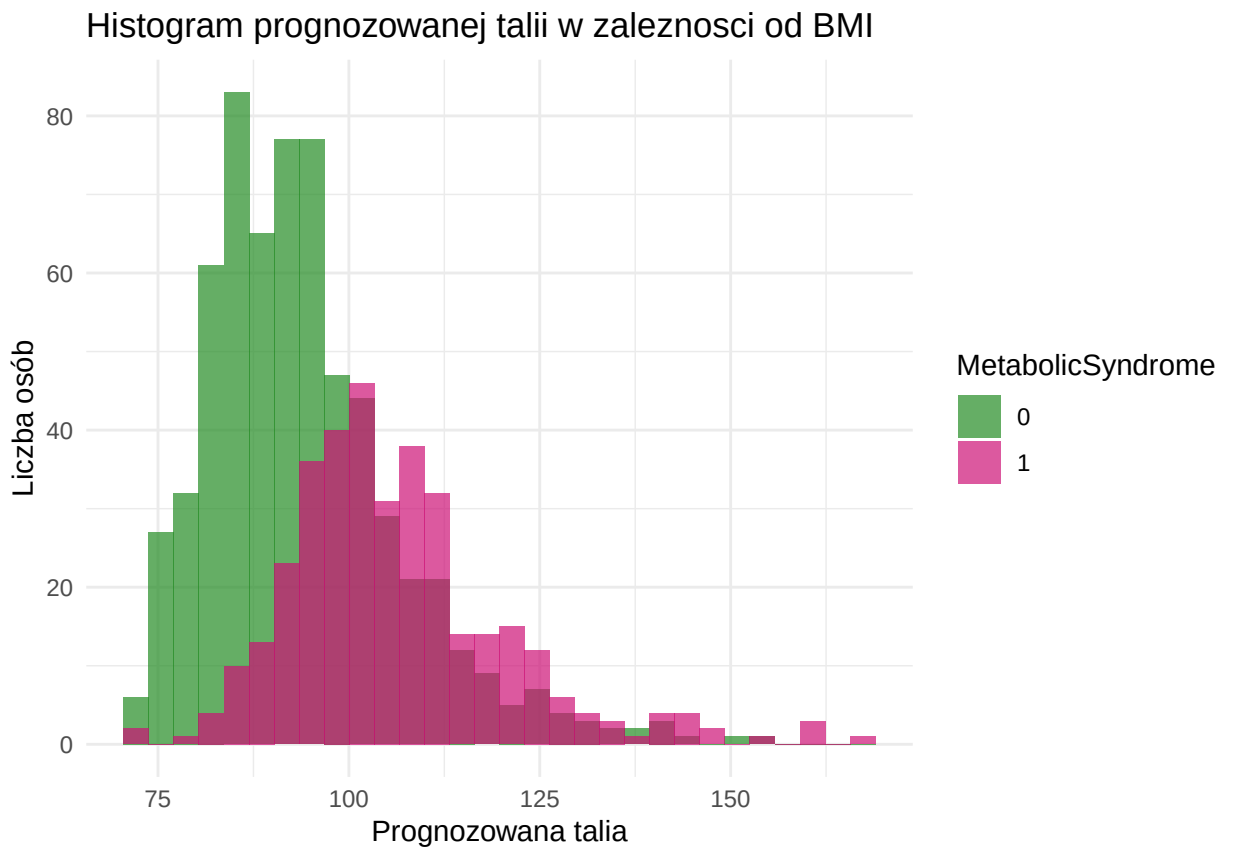
```r
summary(predicted_waist_MS1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   72.08   97.21  103.94  106.45  112.46  166.53
```

Dla osób zdrowych: Mediana wynosi 92.27. Średnia wartość prognozowanej talii wynosi 94.13.

Dla osób mających *syndrom metaboliczny*: Mediana wynosi 103.94. Średnia wartość prognozowanej talii wynosi 106.45.

```r
# Tworzenie ramki danych dla histogramu
df <- data.frame(
  Predicted_Waist = c(predicted_waist_MS0, predicted_waist_MS1),
  MetabolicSyndrome = factor(rep(c(0, 1),  times = c(length(set_MS0[,1]),
                                                     length(set_MS1[,1]))))
)

# Tworzenie histogramu
ggplot(df, aes(x = Predicted_Waist, fill = MetabolicSyndrome)) +
  geom_histogram(position = "identity", alpha = 0.7, bins = 30) +
  scale_fill_manual(values = c("forestgreen", "deeppink3"), name = "MetabolicSyndrome") +
  labs(title = "Histogram prognozowanej talii w zależności od BMI",
       x = "Prognozowana talia",
       y = "Liczba osób") +
  theme_minimal()
```



Na wykresie także widać, że znaczna większość zdrowych osób ma talię w przedziale 75cm - 100cm. Natomiast osoby mające syndrom metaboliczny mają większy obwód talii.

Przyjżyjmy się jeszcze dokładniej poniższemu modelowi.

```
model <- lm(WaistCirc ~ BMI + MetabolicSyndrome, ms)
summary(model)
```

```
##
## Call:
## lm(formula = WaistCirc ~ BMI + MetabolicSyndrome, data = ms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.3425  -4.1241  -0.3237   4.3308  22.5887
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        36.02621    0.68011   52.97   <2e-16 ***
## BMI                 2.12424    0.02461   86.31   <2e-16 ***
## MetabolicSyndrome1  4.12227    0.33850   12.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.546 on 2006 degrees of freedom
## Multiple R-squared:  0.839,  Adjusted R-squared:  0.8389
## F-statistic:  5228 on 2 and 2006 DF,  p-value: < 2.2e-16
```

Widzimy, że wielkość talii osób mających syndrom metaboliczny jest o ponad 4cm większa niż w przypadku osób zdrowych o tym samym BMI. Co ważniejsze, wszystkie zmienne są istotne statystycznie, a model wyjaśnia aż 83.9% zmienności. Zatem istnieją istotne różnice w obowdzie w talii między grupą osób zdrowych, grupą mającą *syndrom metaboliczy*.