Projekt zaliczeniowy - Uogólnione Modele Liniowe

Gomulak Aleksanda, Jędrzejczyk Nikola

Importujemy potrzebne biblioteki:

```
library("dplyr")
library("GGally")
library("tidyr")
library("ResourceSelection")
library("statmod")
```

Zbiór danych:

Zbiór danych na jaki się zdecydowałyśmy zawiera wszelkiego rodzaju 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?

Na podstawie wybranej bazy danych spróbujemy zbadać hipotezy o symptomach choroby, które naszym zdaniem są prawdziwe.

Dane pobrałyśmy ze strony Kaggle.

Importujemy dane:

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

Opis zmiennych:

zmienna:	opis:
PCOS:	czy pacjent miał zdiagnozowany PCOS? (0 - nie, 1 - tak)
Age..yrs.:	wiek pacjenta
BMI:	wartość BMI (Body Mass Index) pacjenta
Pulse.rate.bpm.:	liczbę uderzeń serca (na min)
RR..breaths.min.:	liczba oddechów (na min)
Hb.g.dl.:	poziom hemoglobiny
Cycle.R.I.:	czy cykl miesiączki jest nieregularny? (0 - nie, 1 - tak)
Cycle.length.days.:	ilość dni trwania cyklu miesiączki
Pregnant.Y.N.:	czy pacjent jest w ciąży? (0 - nie, 1 - tak)
No..of.abortions:	liczba wykonanych aborcji w całym życiu
I. . . beta.HCG.mIU.mL.:	poziom kormonu hCG, który jest produkowany w trakcie ciąży
FSH:	poziom hormonu niezbędnego do prawidłowego działania jajników oraz jąder
LH.mIU.mL.:	poziom hormonu LH, niski poziom może być oznaką wolnego dojrzewania
FSH.LH :	stosunek poziomów FSH do LH we krwi pacjenta
TSH:	ilość hormonu, który kontroluje sposób działania innych hormonów
AMH.ng.mL.:	poziom hormonu, który zapobiega rozwój żeńskich narządów płciowych u płodu męskiego
PRL.ng.mL.:	poziom prolaktyny, hormonu, który odpowiada za laktację czy rozwój piersi
PRG.ng.mL:	poziom hormonu, który zapobiega rozwój męskich narządów płciowych u płodu żeńskiego
Weight.gain.Y.N.:	czy pacjent przybrał na wadze? (0 - nie, 1 - tak)
hair.growth.Y.N.:	czy pacjentowi rosną włosy? (0 - nie, 1 - tak)
Skin.darkening..Y.N.:	czy pacjentowi ciemnieje skóra? (0 - nie, 1 - tak)
Hair.loss.Y.N.:	czy pacjentowi wypadają włosy? (0 - nie, 1 - tak)
Pimples.Y.N.:	czy pacjent ma wypryski na twarzy? (0 - nie, 1 - tak)
BP._Systolic..mmHg.:	skurczowe ciśnienie krwi (systoliczne)
BP._Diastolic..mmHg.:	rozkurczowe ciśnienie krwi (diastoliczne)
Follicle.No. . . L.:	liczba pęcherzyków w lewym jajniku wykryta podczas badania
Follicle.No. . . R.:	liczba pęcherzyków w prawym jajniku wykryta podczas badania
Avg..F.size..L. . . mm.:	średnia wielkość pęcherzyków w lewym jajniku (w mm)
Avg..F.size..R. . . mm.:	średnia wielkość pęcherzyków w prawym jajniku (w mm)
Endometrium..mm.:	grubość błony śluzowej macicy (w mm)

Hipotezy badawcze:

Zdecydowałyśmy się na zbadanie hipotez, które naszym zdaniem mogą być prawdziwe.

Hipoteza pierwsza:

Osoby z zespołem policystycznych jajników mają podwyższone ciśnienie, wyższe BMI i FSH poza normą.

Hipoteza druga:

Ciemnienie skóry, pojawianie się wyprysków i nieregularnych miesiączek oznacza chorowanie na zespół policystycznych jajników.

Hipoteza trzecia:

Hormon FSH u kobiet pobudza wzrost pęcherzyków w jajniku.

Wstępne czyszczenie i przygotowanie danych do pracy:

Aby doprowadzić do stanu zaprezentowanego powyżej, musimy wykonać pare potrzebnych króków.

Zaczynamy od usunięcia kolumn, które nas nie interesują lub są niepotrzebne:

```
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.")

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

Na wszelki wypadek usuwamy braki:

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

Zauważyłyśmy błędne dane w trakcie przygotowywania projektu, w związku z tym czyścimy dane tak, aby ich nie zawierały:

```
data <- subset(data, Cycle.R.I. != 5)
```

```
data$Cycle.R.I. <- ifelse(data$Cycle.R.I. == 2, 0, ifelse(data$Cycle.R.I. == 4, 1, data$Cycle.R.I.))
```

Następnie zmianiemy typ danych, w niektórych kolumnach na factor, aby móc w następnych krokach z nich skorzystać:

```
data$PCOS..Y.N.<-as.factor(data$PCOS..Y.N.)
data$Cycle.R.I.<-as.factor(data$Cycle.R.I.)
data$Weight.gain.Y.N.<-as.factor(data$Weight.gain.Y.N.)
data$hair.growth.Y.N.<-as.factor(data$hair.growth.Y.N.)
data$Skin.darkening..Y.N.<-as.factor(data$Skin.darkening..Y.N.)
data$Hair.loss.Y.N.<-as.factor(data$Hair.loss.Y.N.)
data$Pimples.Y.N.<-as.factor(data$Pimples.Y.N.)
data$AMH.ng.mL. <- as.double(data$AMH.ng.mL.)
```

Sprawdzamy jak wyglądają ilości odpowiedzi, w każdej z kolumn i podstawowe wartości:

```
lapply(data, summary)
```

```
## $PCOS..Y.N.
##   0   1
## 364 176
##
## $Age..yrs.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   28.00   31.00   31.45   35.00   48.00
##
## $BMI
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.40   21.60   24.20   24.30   26.62   38.90
##
```

```
## $Pulse.rate.bpm.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   72.00   72.00   73.25   74.00   82.00
##
## $RR..breaths.min.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   18.00   18.00   19.25   20.00   28.00
##
## $Hb.g.dl.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   10.50   11.00   11.16   11.72   14.80
##
## $Cycle.R.I.
##   0   1
## 390 150
##
## $Cycle.length.days.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.000   5.000   4.937   5.000  12.000
##
## $Pregnant.Y.N.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3815  1.0000  1.0000
##
## $No..of.abortions
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2889  0.0000  5.0000
##
## $I...beta.HCG.mIU.mL.
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.30     1.99    19.38   665.56   298.04 32460.97
##
## $FSH.mIU.mL.
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.210    3.322    4.855   14.625    6.412 5052.000
##
## $LH.mIU.mL.
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.020    1.020    2.300    6.481    3.680 2018.000
##
## $FSH.LH
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    1.417    2.165    6.911    3.962 1372.830
##
## $TSH..mIU.L.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.040   1.480   2.260   2.985   3.570  65.000
##
## $AMH.ng.mL.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.100   2.010   3.700   5.624   6.950  66.000       1
##
## $PRL.ng.mL.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    0.40   14.50   21.92   24.34   29.91  128.24
##
## $PRG.ng.mL.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0470  0.2500  0.3200  0.6116  0.4525 85.0000
##
## $Weight.gain.Y.N.
##   0   1
## 336 204
##
## $hair.growth.Y.N.
##   0   1
## 393 147
##
## $Skin.darkening..Y.N.
##   0   1
## 375 165
##
## $Hair.loss.Y.N.
##   0   1
## 295 245
##
## $Pimples.Y.N.
##   0   1
## 275 265
##
## $BP._Systolic..mmHg.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0   110.0   110.0   114.7   120.0   140.0
##
## $BP._Diastolic..mmHg.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   70.00   80.00   76.94   80.00  100.00
##
## $Follicle.No...L.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   5.000   6.126   9.000  22.000
##
## $Follicle.No...R.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   6.000   6.635  10.000  20.000
##
## $Avg..F.size..L...mm.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.00   15.00   15.02   18.00   24.00
##
## $Avg..F.size..R...mm.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.00   16.00   15.45   18.00   24.00
##
## $Endometrium..mm.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.000   8.500   8.474   9.800  18.000
```

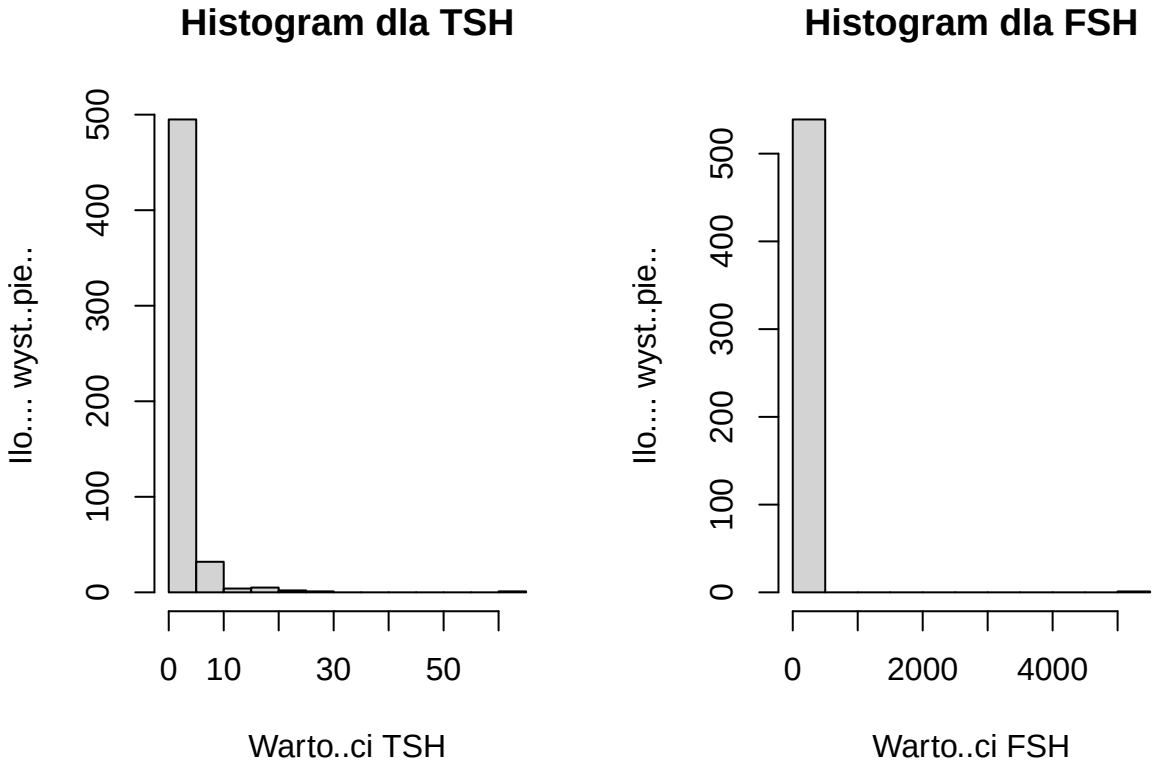
Dla wygody zmieniamy niektóre nazwy kolumn:

```
names(data)[names(data) == "PCOS..Y.N."] <- "PCOS"
names(data)[names(data) == "TSH..mIU.L."] <- "TSH"
names(data)[names(data) == "FSH.mIU.mL."] <- "FSH"
```

Eksploracyjna analiza danych:

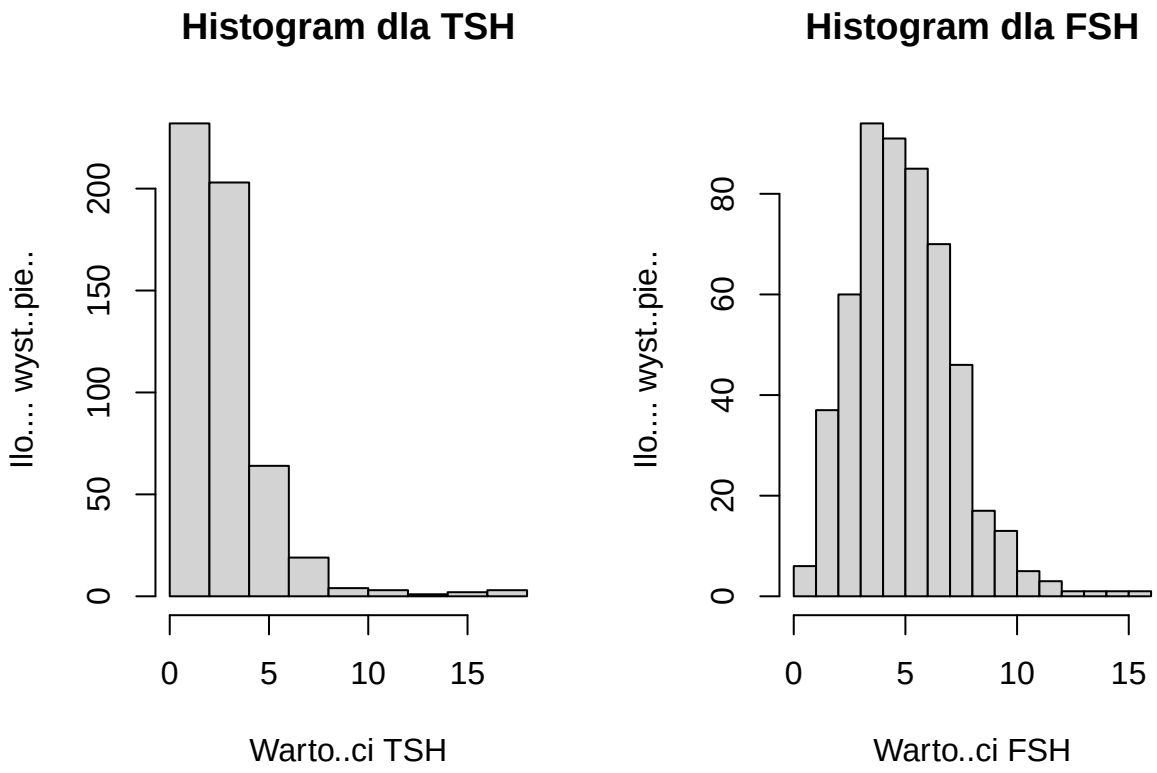
Na początek sprawdźmy jak prezentują się histogramy dla najbardziej istotnych hormonów: TSH i FSH:

```
par(mfrow = c(1, 2))
hist(data$TSH, main = "Histogram dla TSH", xlab = "Wartości TSH", ylab = "Ilość wystąpień")
hist(data$FSH, main = "Histogram dla FSH", xlab = "Wartości FSH", ylab = "Ilość wystąpień")
```



Musimy usunąć pojedyńcze, duże wartości, aby histogramy były czytelniejsze.

```r
data <- subset(data, TSH < 20)
data <- subset(data, FSH < 20)
par(mfrow = c(1, 2))
hist(data$TSH, main = "Histogram dla TSH", xlab = "Wartości TSH", ylab = "Ilość wystąpień")
hist(data$FSH, main = "Histogram dla FSH", xlab = "Wartości FSH", ylab = "Ilość wystąpień")
```



Wniosek:

W przypadku TSH widzimy, że większość pacjentów mieści się w normie. Sporo z nich jednak ma ten poziom skrajnie wysoki. Pojawia się też grupa osób, która kilkukrotnie przekracza normę.

Ciężko jednak wywnioskować coś dla poziomu FSH, ponieważ mogą to być jednocześnie osoby, w konkretnym cyklu miesiączki i osoby, które mają wartość poza normę.

Spójrzmy teraz jak rozkładają się różne wartości względem występowania PCOS. To na PCOS skupiamy większość swojej uwagi, ponieważ jest to kluczowa kolumna.

```r
tab1 <- xtabs(~ PCOS + Cycle.R.I., data %>% select(PCOS, Cycle.R.I.))

p <- c(tab1[1,2], tab1[2,2])
n <- c(tab1[1,2] + tab1[1,1], tab1[2,1] + tab1[2,2])
tab1
```

```
##     Cycle.R.I.
## PCOS   0   1
##    0 305  55
##    1  81  90
```

Wniosek:

Najliczniejszą grupą są badani, którzy nie chorują na PCOS i nie występują u nich nieregularne miesiączki.

```
tab2 <- xtabs(~ PCOS + Skin.darkening..Y.N., data %>% select(PCOS, Skin.darkening..Y.N.))

p <- c(tab2[1,2], tab2[2,2])
n <- c(tab2[1,2] + tab2[1,1], tab2[2,1] + tab2[2,2])
tab2
```

```
##     Skin.darkening..Y.N.
## PCOS   0   1
##    0 304  56
##    1  66 105
```

Wniosek:

Sytuacja osób, kóre nie mają PCOS się nie zmieniła, wyniki są takie same. Jednak dla osób, które chorują na PCOS ciemnienie skóry jest częstsze niż nieregularne miesiączki.

```
tab3 <- xtabs(~ PCOS + Pimples.Y.N., data %>% select(PCOS, Pimples.Y.N.))

p <- c(tab3[1,2], tab3[2,2])
n <- c(tab3[1,2] + tab3[1,1], tab3[2,1] + tab3[2,2])
tab3
```

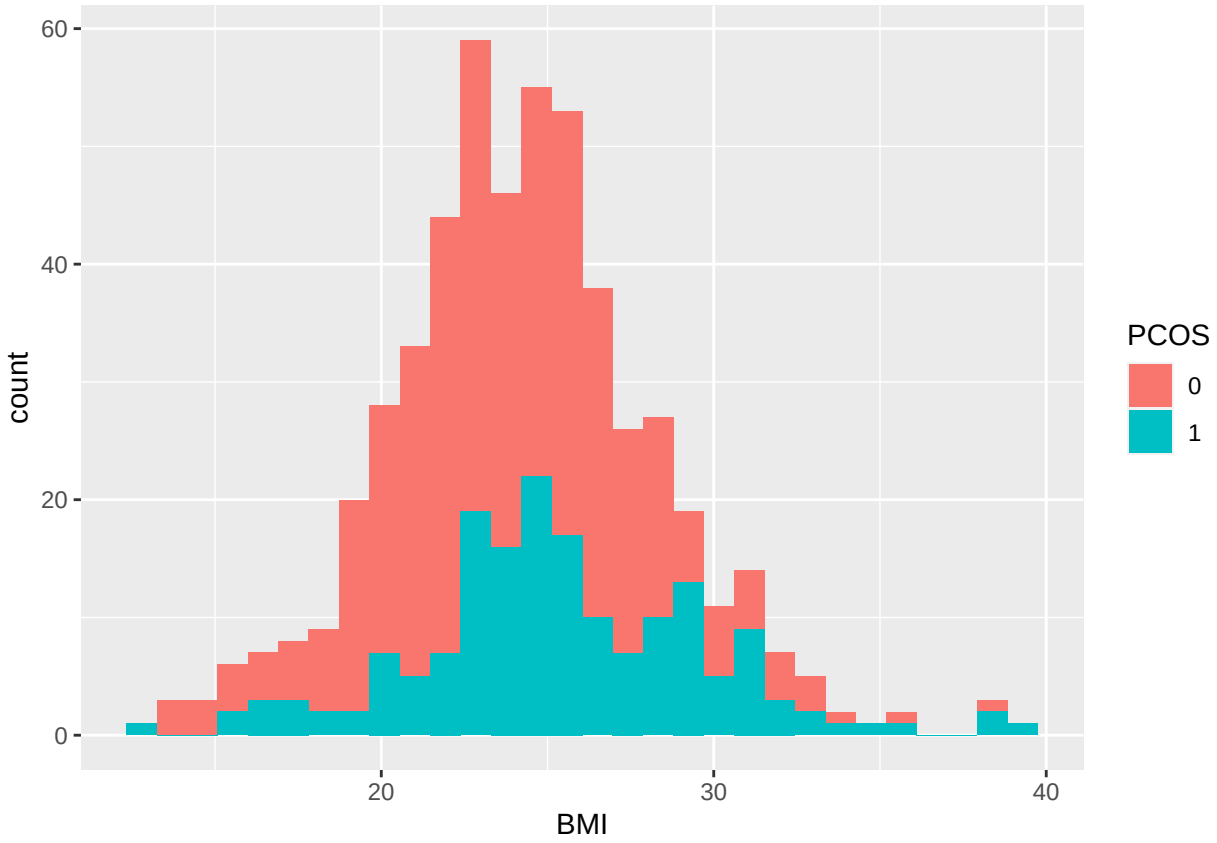
```
##     Pimples.Y.N.
## PCOS   0   1
##    0 221 139
##    1  53 118
```

Wniosek:

Najmniej liczną grupą są osoby chorujące na PCOS i nie posiadające wyprysków. Było to do przewidzenia. Jesteśmy jednak w szoku, że około połowa badanych nie posiada wyprysków, ponieważ jest to coś powszechnego, zależnego nie tylko od PCOS.

Porównajmy jeszcze na wykresie BMI i PCOS:

```
ggplot(data, aes(x = BMI, fill = PCOS)) + geom_histogram()
```



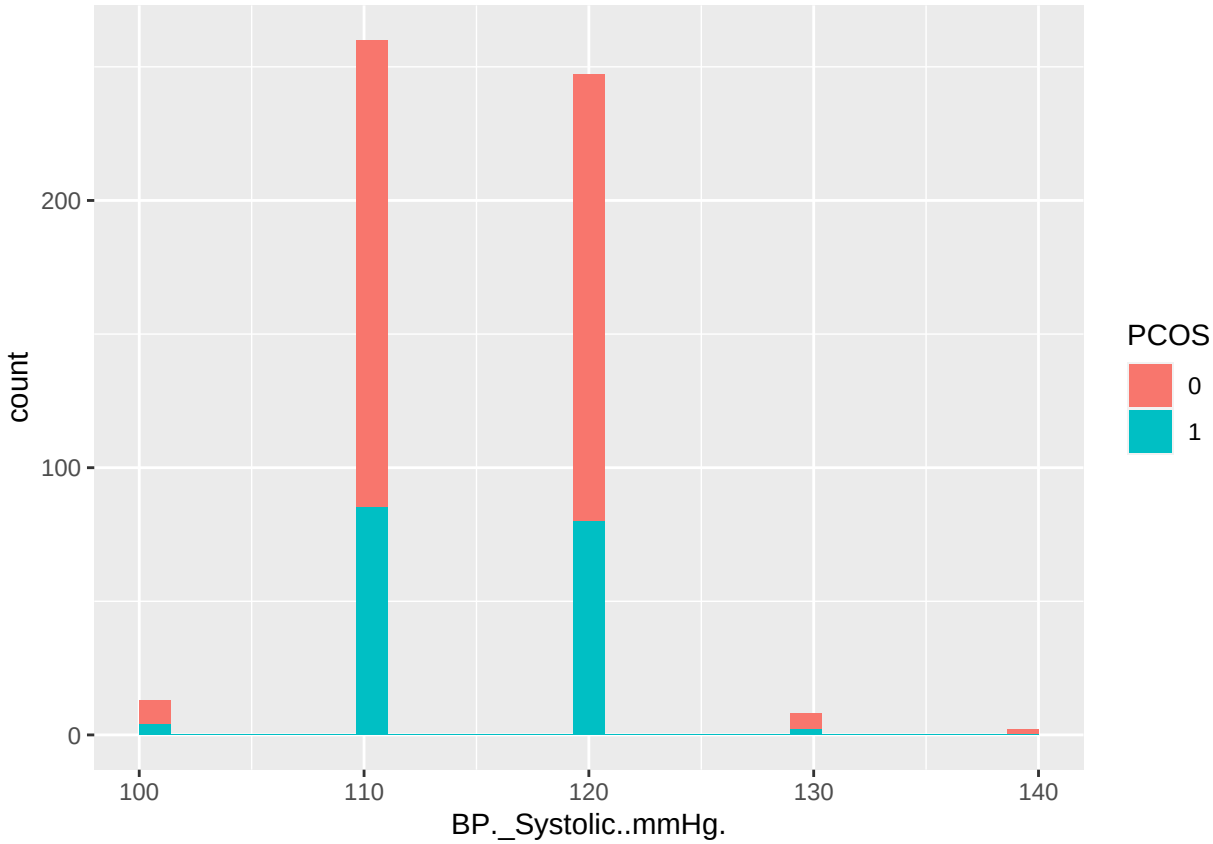
Wniosek:

Rozkład wygląda podobnie, a różnica wynika z tego, że w bazie danych jest dużo mniej osób chorych.

Porównamy jeszcze oba rodzaje ciśnienia krwi:

Ciśnienie skórczowe:

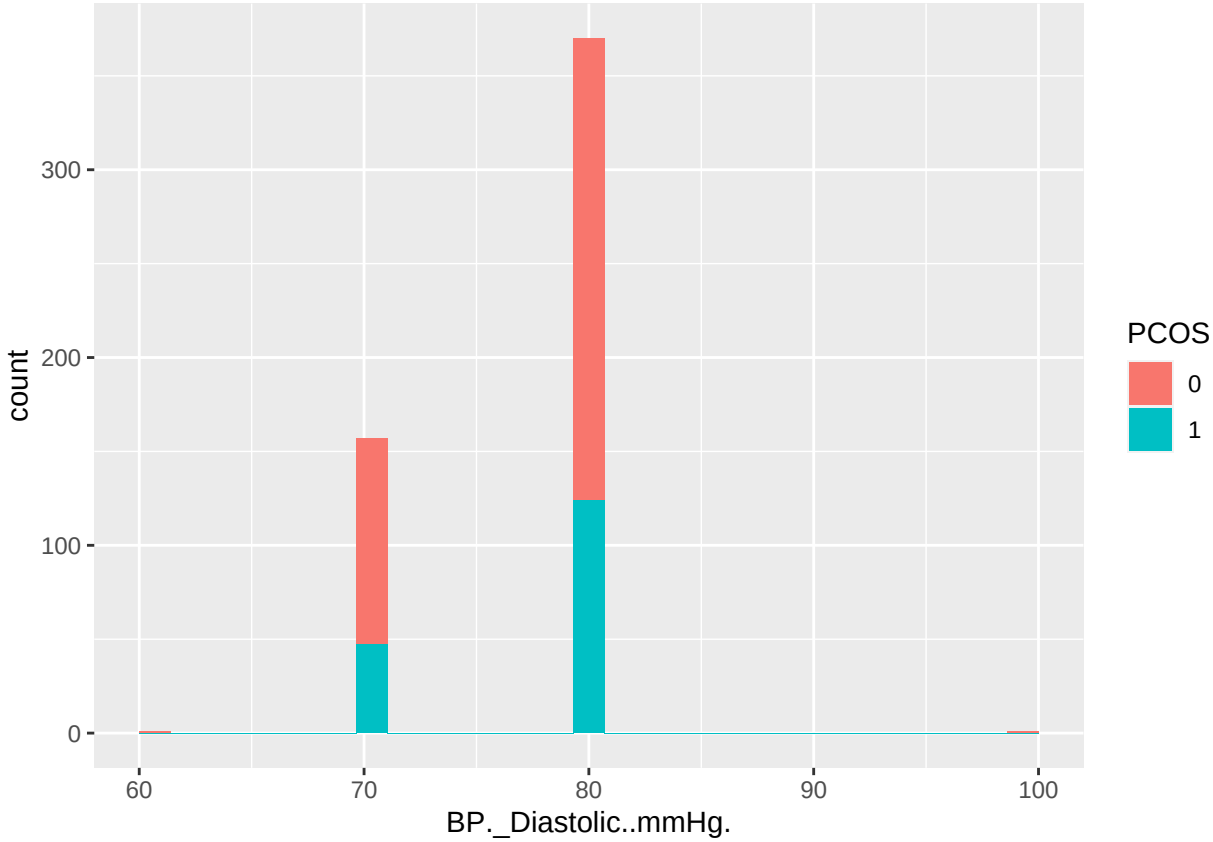
```
data <- subset(data, BP._Systolic..mmHg. > 75)
ggplot(data, aes(x = BP._Systolic..mmHg., fill = PCOS)) + geom_histogram()
```



Wniosek: Większość badanych ma trochę za niskie ciśnienie skurczowe. Norma jest pomiędzy 120 a 129.

Ciśnienie rozkórczowe:

```
data <- subset(data, BP._Diastolic..mmHg. > 55)
ggplot(data, aes(x = BP._Diastolic..mmHg., fill = PCOS)) + geom_histogram()
```

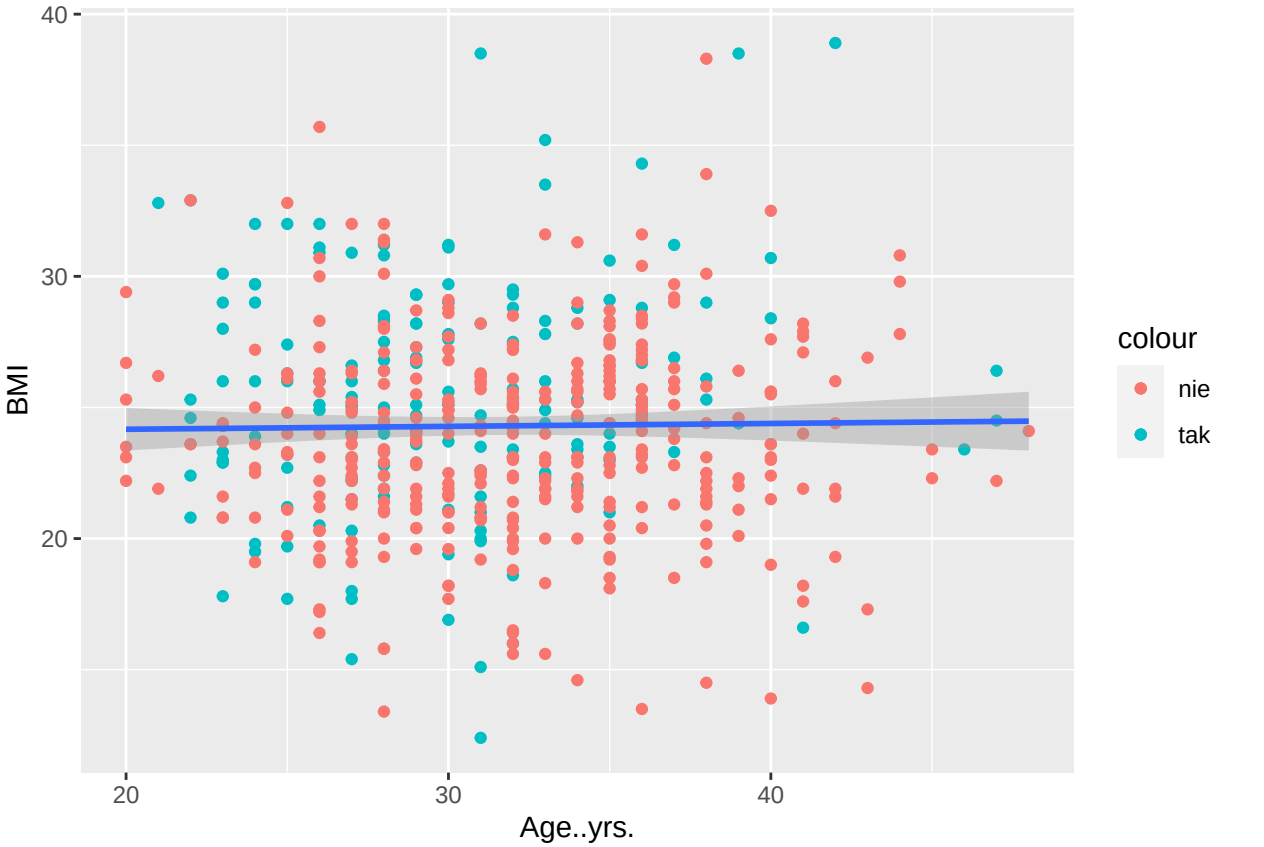


Wniosek: W tym przypadku większość osób ma poprawne ciśnienie rozkurczowe, jednak jest spora grupa osób, która ma to ciśnienie za niskie.

Dodatkowo spójrzmy jeszcze na występowanie PCOS w zależności od wieku i BMI:

```
data_nie <- data[data$PCOS == 0,]
data_tak <- data[data$PCOS == 1,]

ggplot(data, aes(x = Age..yrs., y = BMI)) +
geom_point(
data = data_tak,
aes(colour = "tak")) +
geom_point(
data = data_nie,
aes(colour = "nie")) + geom_smooth(method = "lm")
```



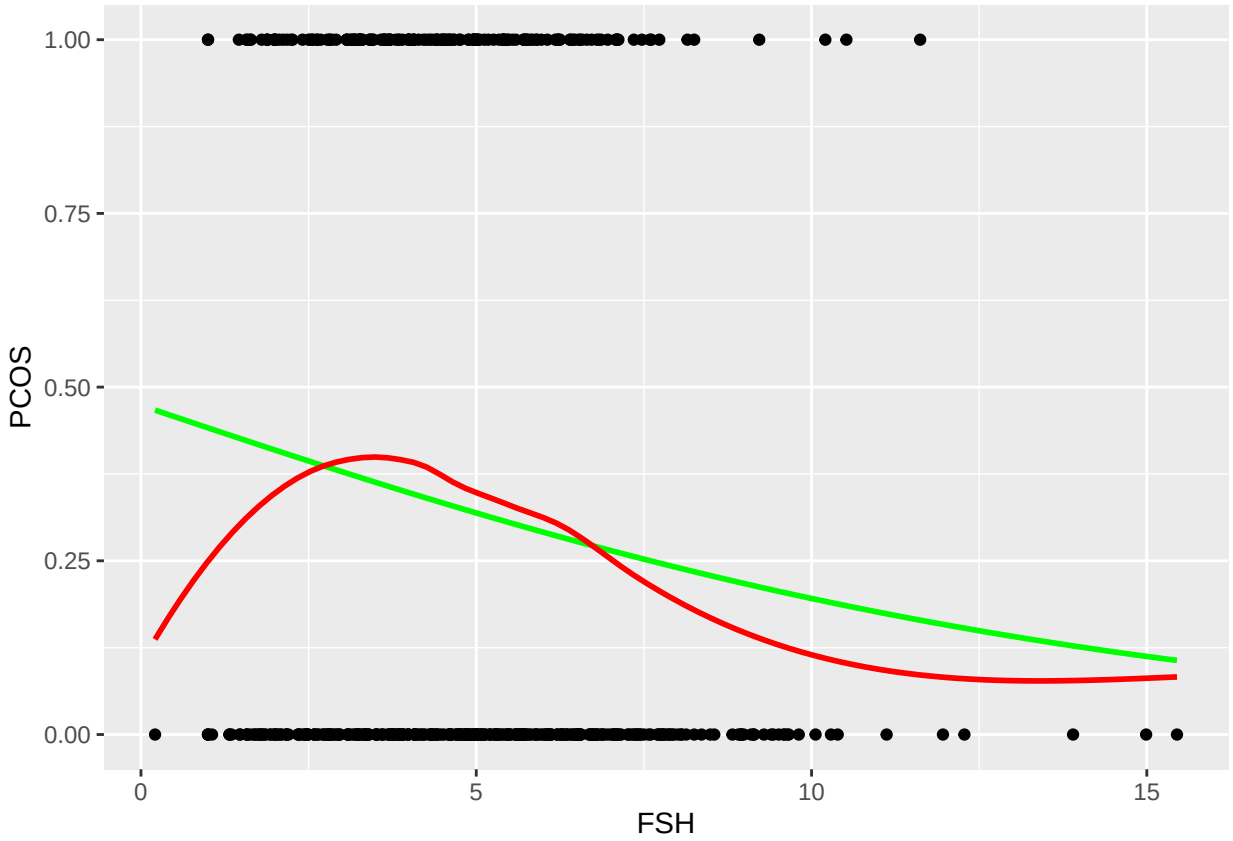
Wniosek:

Zauważamy, że więcej osób przed 35 choruje na PCOS niż po 35 roku życia. Wartości BMI rozkładaja się równomiernie. Widoczny jest jedynie bardzo delikatny wzrost BMI wraz z wiekiem.

W następnej kolejności wykonujemy wykres punktowy wykreślając zmienną PCOS względem nieprzekształconej zmiennej FSH:

```
data$PCOS <- as.numeric(as.character(data$PCOS))

ggplot(data, aes(x = FSH, y = PCOS)) + geom_point() +
  geom_smooth(method = "glm", color = "green", se = FALSE,
              method.args = list(family = binomial)) +
  geom_smooth(method = "loess", se  = FALSE, color = 'red')
```

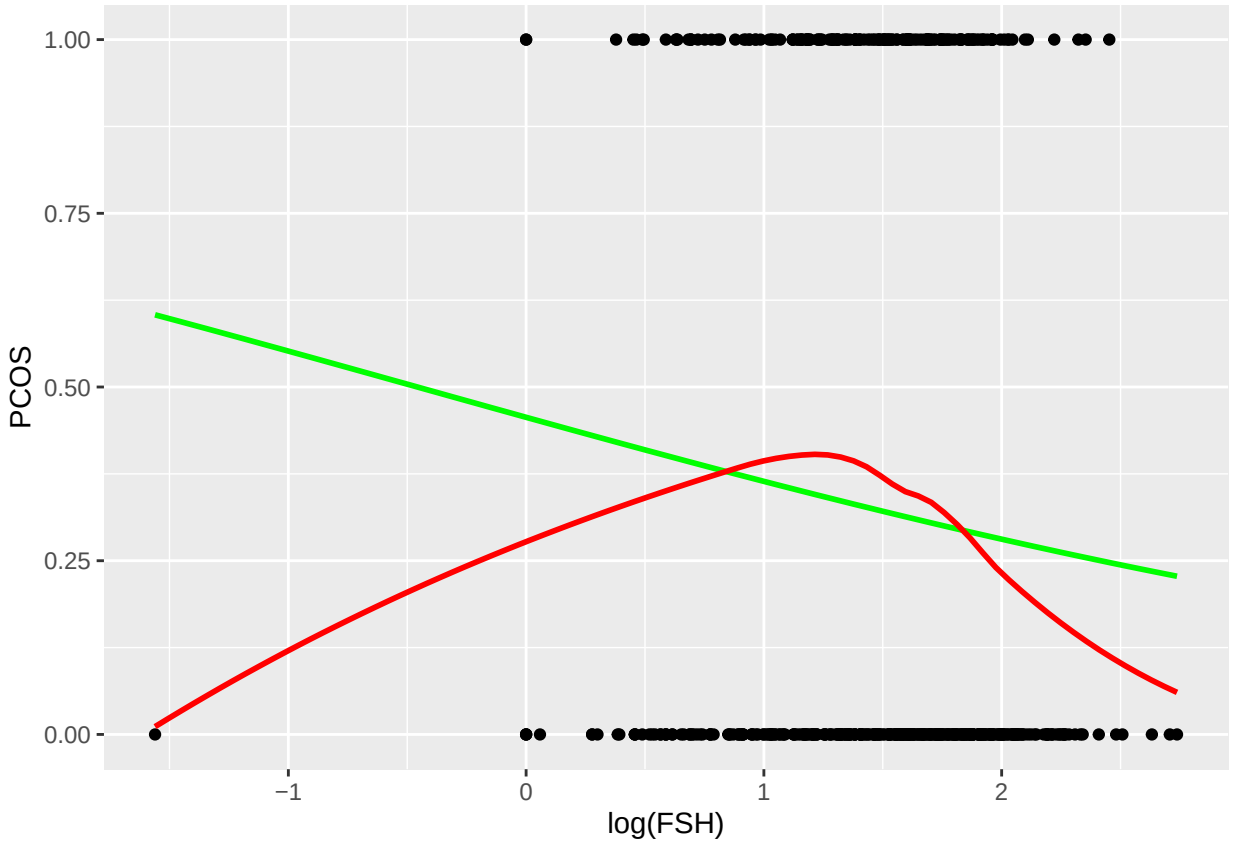


Wniosek:

Metoda "loess" pokazuję nam linię trendu, która na początku rośnie, żeby następnie "gładko" zmaleć. Uważamy, że metoda "loess" lepiej oddaje stosunek stężenia FSH do występowania zespołu policystycznych jajników.

Porównajmy teraz ten sam przypadek dla zmiennej przekształconej:

```
ggplot(data, aes(x = log(FSH), y = PCOS)) + geom_point() +
  geom_smooth(method = "glm", color = "green", se = FALSE,
              method.args = list(family = binomial)) +
  stat_smooth(method = "loess", se  = FALSE, color = 'red')
```



Wniosek:

Ponownie widzimy lepsze zachowanie dla metody “loess”. Ciężko stwierdzić czy różnica przy zmiennej przekształconej jest dużo lepsza.

Zbudujemy teraz modele dla obu rodzajów zmiennych:

Zmienna nieprzekształcona:

```
data$PCOS <- as.factor(data$PCOS)
model.fsh <- glm(PCOS ~ FSH, family = 'binomial', data)
summary(model.fsh)
```

```
##
## Call:
## glm(formula = PCOS ~ FSH, family = "binomial", data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.10504    0.23170  -0.453  0.65030
## FSH         -0.13077    0.04482  -2.918  0.00353 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.80  on 528  degrees of freedom
## Residual deviance: 656.79  on 527  degrees of freedom
## AIC: 660.79
##
## Number of Fisher Scoring iterations: 4
```

Wniosek:

W powyższym przypadku wartość Intercept nie jest statystycznie istotna, natomiast współczynnik FSH jest istotny statystycznie ($\Pr(>|z|) = 0.00353$)

Null deviance mówi o odchyleniu modelu zerowego. Służy on do porównania dopasowania dla innych modeli.

Residual deviance mówi o odchyleniu dla dopasowanego modelu. Różnica pomiędzy null deviance a residual deviance wskazuje jak dobrze dopasowany jest model. Im niższa wartość residual deviance, tym model jest lepiej dopasowany.

AIC mierzy jakość modelu statystycznego. Zawiera w sobie zarówno dopasowanie modelu jak i jego złożoność. Im niższe AIC, tym model jest lepiej dopasowany.

Podsumowując, powyższy model mówi, że FSH jest istotnym statystycznie predyktorem występowania zespołu policystycznych jajników. Wraz ze wzrostem FSH o jedną jednostkę, prawdopodobieństwo występowania PCOS maleje o 0.13077, co jest jak najbardziej prawdopodobne, ponieważ osoby z PCOS mają tendencję do posiadania poziomu FSH poniżej normy.

Zmienna przekształcona:

```
model.fsh.log <- glm(PCOS ~ log(FSH), family = 'binomial', data)
summary(model.fsh.log)
```

```
##
## Call:
## glm(formula = PCOS ~ log(FSH), family = "binomial", data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1748     0.2840  -0.616   0.5382
## log(FSH)     -0.3823     0.1836  -2.083   0.0373 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.80  on 528  degrees of freedom
## Residual deviance: 661.45  on 527  degrees of freedom
## AIC: 665.45
##
## Number of Fisher Scoring iterations: 4
```

Wniosek:

Nie zauważamy większych zmian przy zlogarytmowaniu zmiennej FSH. Dodanie logarytmu nie poprawiło modelu.

Wykonujemy dla nich test Hosmera-Lemeshowa:

```
data$PCOS <- as.numeric(as.character(data$PCOS))

hoslem.test((data %>% select(PCOS, FSH) %>% drop_na())$PCOS, fitted(model.fsh))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, FSH) %>% drop_na())$PCOS, fitted(model.fsh)
## X-squared = 8.2997, df = 8, p-value = 0.4048
```

```
hoslem.test((data %>% select(PCOS, FSH) %>% drop_na())$PCOS, fitted(model.fsh.log))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, FSH) %>% drop_na())$PCOS, fitted(model.fsh.log)
## X-squared = 12.129, df = 8, p-value = 0.1455
```

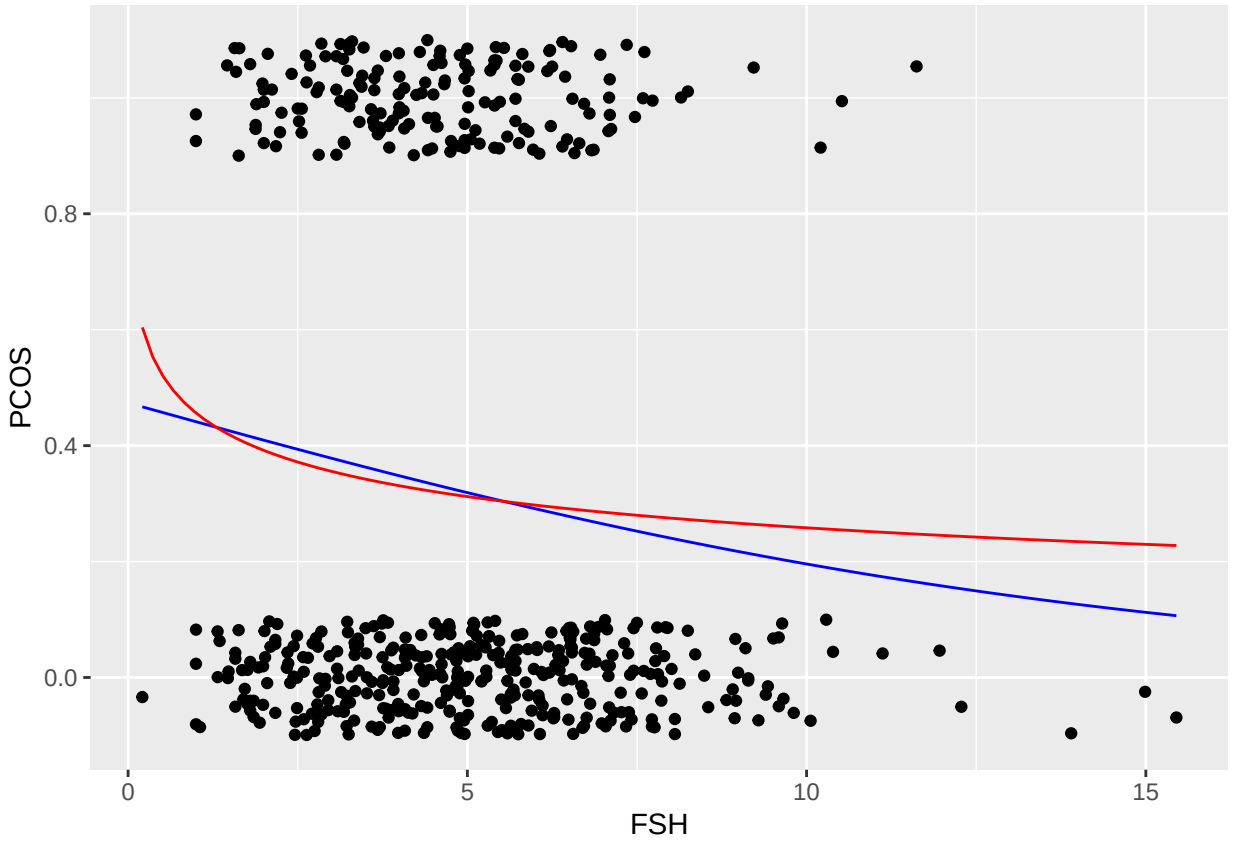
Wniosek:

W obu przypadkach widzimy brak istotnych różnic między obserwowanymi danymi a danymi przewidywanymi przez model.

Wartość p-value wynosi 0.4048 i 0.1455, sugeruje to, że nie ma powodów do odrzucenia hipotezy zerowej, a to oznacza, że oba modele są odpowiednio dopasowane.

Graficzne porównanie modelu bez przekształconej zmiennej objaśniającej oraz modelu ze zlogarytmowaną zmienną objaśniającą:

```
ggplot(data, aes(x = FSH, y = PCOS)) + geom_point(position = position_jitter(height = 0.1)) +
      stat_function(fun = function(x) predict(object = model.fsh, newdata = data.frame(FSH = x),
      type = 'response'), color = 'blue') + stat_function(fun = function(x)
      predict(object = model.fsh.log, newdata = data.frame(FSH = x), type = 'response'),
      color = 'red')
```



W następnej kolejności postanowiłyśmy przetestować model z dwoma predyktorami dla zmiennych PCOS i FSH:

Budowa modelu i przeprowadzenie testu:

```
data$PCOS <- as.factor(data$PCOS)
model.two <- glm(PCOS ~ FSH + log2(FSH), binomial, data)
```

```
data$PCOS <- as.numeric(as.character(data$PCOS))

hoslem.test((data %>% select(PCOS, FSH) %>% drop_na())$PCOS, fitted(model.two))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, FSH) %>% drop_na())$PCOS, fitted(model.two)
## X-squared = 2.3091, df = 8, p-value = 0.97
```

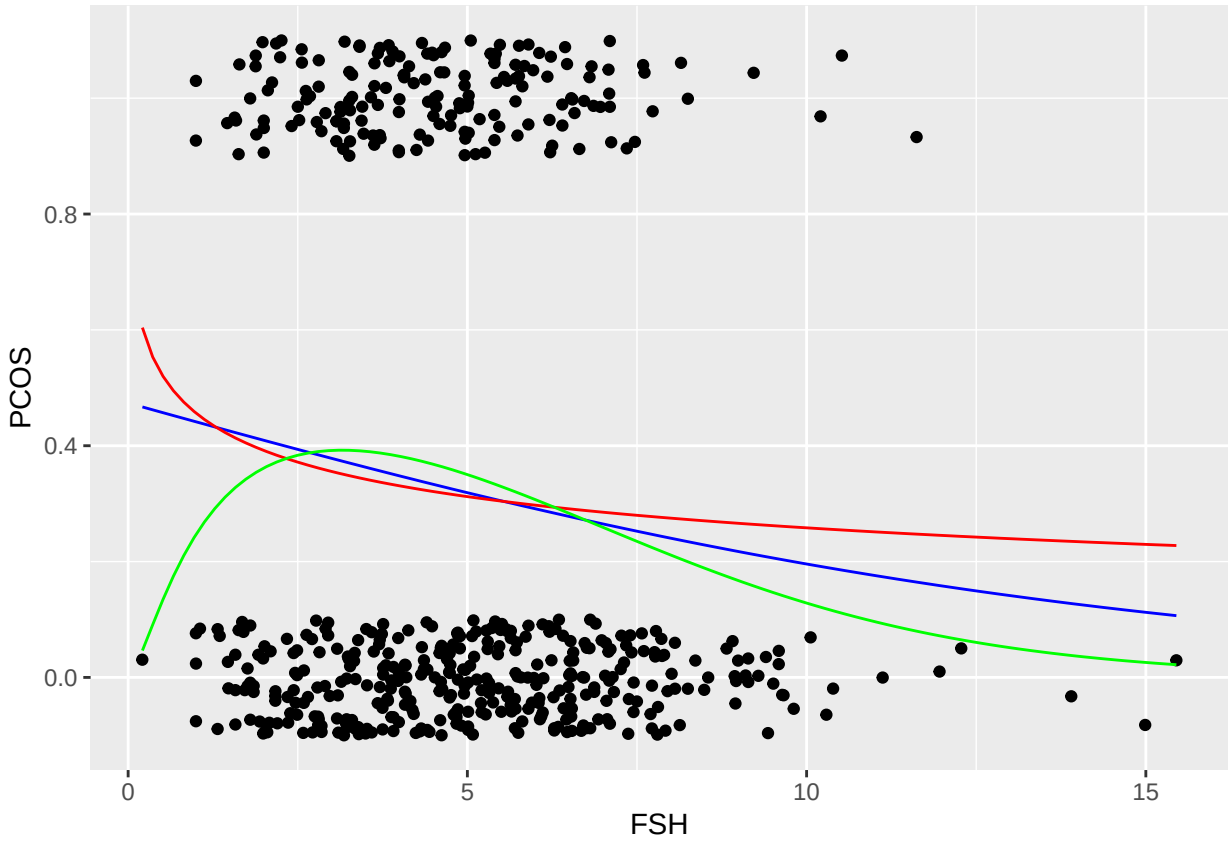
Wniosek:

Statystyka testowa X-squared wynosi 2.3091, a stopnie swobody df wynoszą 8. Oznacza to, że testowane dane miały niewielkie różnice od przewidywanych wartości modelu. Ponadto p-value wynosi 0.97.

Model zatem jest bardzo dobrze dopasowany.

Graficzne porównanie wszystkich trzech modeli z PCOS jako zmienną objaśnianą i FSH jako zmienną objaśniającą:

```
ggplot(data, aes(x = FSH, y = PCOS)) +
  geom_point(position = position_jitter(height = 0.1)) +
  stat_function(fun = function(x) predict(object = model.fsh, newdata = data.frame(FSH = x),
  type = 'response'), color = 'blue') + stat_function(fun = function(x) predict(object = model.fsh.log,
  newdata = data.frame(FSH = x), type = 'response'), color = 'red') +
  stat_function(fun = function(x)
  predict(object = model.two, newdata = data.frame(FSH = x), type = 'response'),
  color = 'green')
```



Wniosek:

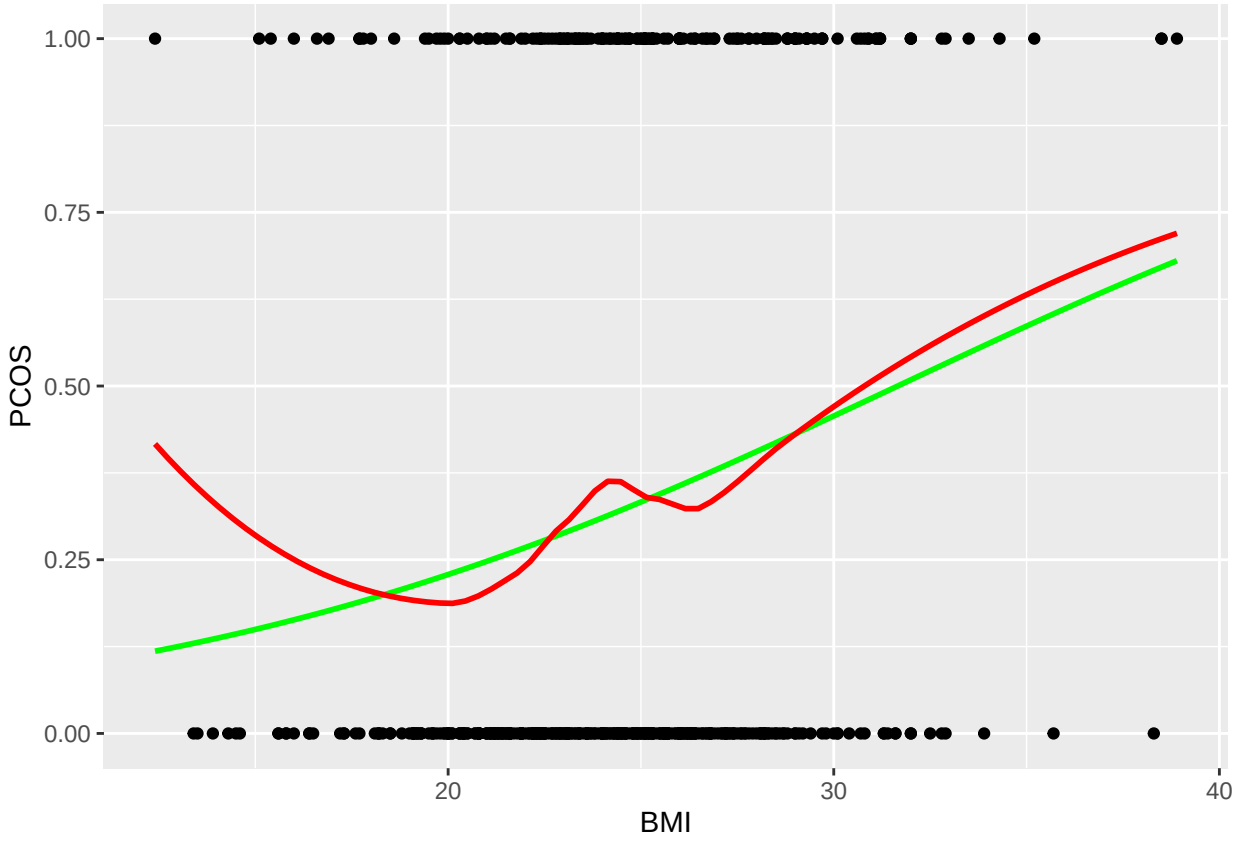
Tak jak potwierdziły to testy Hosmera-Lemeshowa, model z dwoma predyktorami wykazuje najlepsze dopasowanie.

W następnym kroku przeprowadzimy taką samą analizę, ale tym razem dla zmiennej BMI:

Dla zmiennej nieprzekształconej:

```
data$PCOS <- as.numeric(as.character(data$PCOS))

ggplot(data, aes(x = BMI, y = PCOS)) + geom_point() +
  geom_smooth(method="glm", color = "green", se = FALSE,
              method.args = list(family = binomial)) +
  geom_smooth(method = "loess", se  = FALSE, color = 'red')
```

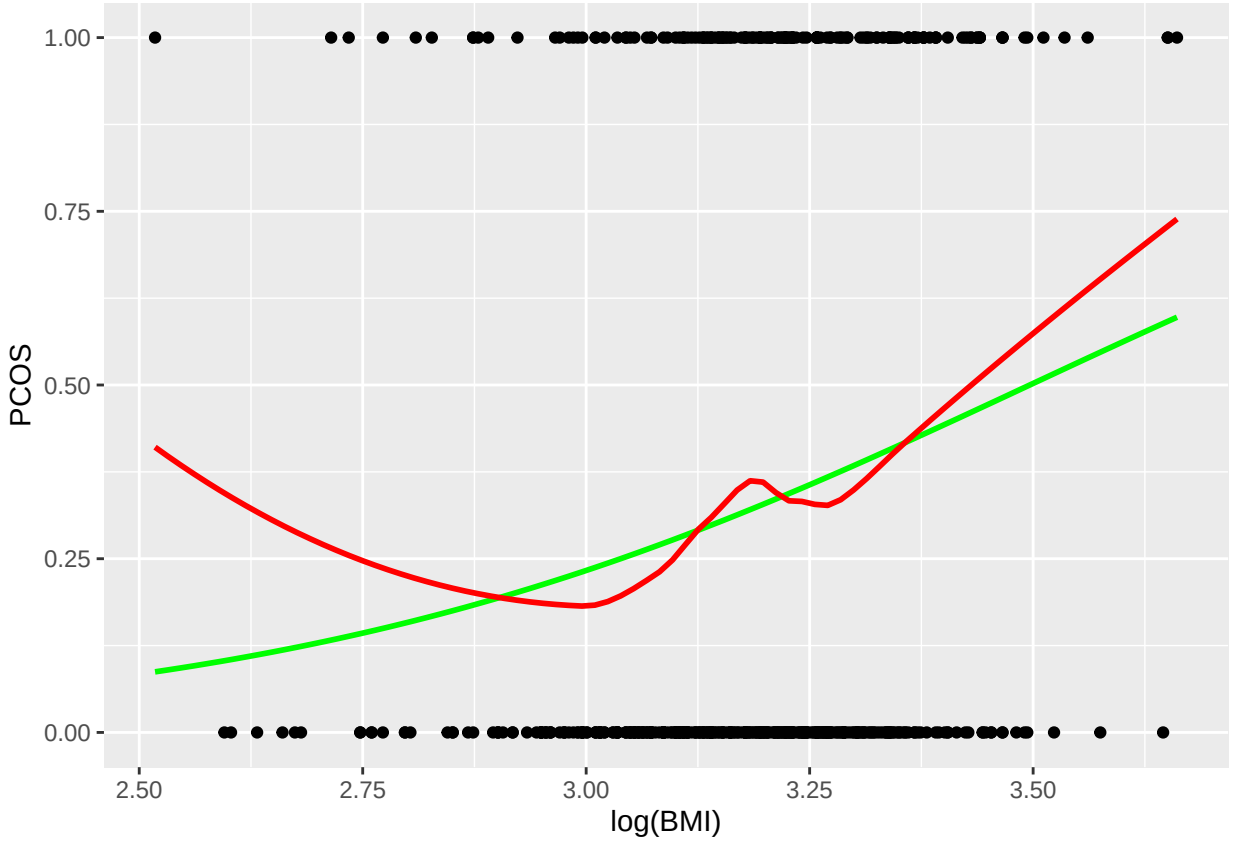


Wniosek:

Wraz ze wzrostem BMI zauważamy wzrost prawdopodobieństwa wystąpienia PCOS.

Dla zmiennej przekształconej:

```
ggplot(data, aes(x = log(BMI), y = PCOS)) + geom_point() +
  geom_smooth(method="glm", color="green", se = FALSE,
              method.args = list(family=binomial)) +
  stat_smooth(method = "loess", se  = FALSE, color = 'red')
```



Wniosek: Zlogarytmowanie zmiennej objaśniającej nie poprawiło ani nie pogorszyło dopasowania modelu.

Zbudujemy teraz modele dla obu rodzajów zmiennych:

Zmienna nieprzekształcona:

```
data$PCOS <- as.factor(data$PCOS)
model.bmi <- glm(PCOS ~ BMI, family = 'binomial', data)
summary(model.bmi)
```

```
##
## Call:
## glm(formula = PCOS ~ BMI, family = "binomial", data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.30055    0.61150  -5.397 6.76e-08 ***
## BMI          0.10426    0.02435   4.282 1.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.80  on 528  degrees of freedom
## Residual deviance: 646.33  on 527  degrees of freedom
## AIC: 650.33
##
## Number of Fisher Scoring iterations: 4
```

Wniosek:

Z modelu wynika, że na wzrost BMI o jedną jednostkę, szansę na cierpienie na zespół policystycznych jajników wzrasta o 0.10426. Ponadto można zauważyć, że zarówno Intercept jak i współczynnik BMI są bardzo istotne statystycznie. Sugeruje to mocny związek pomiędzy BMI a występowaniem PCOS.

Parametr dyspersji mierzy dopasowanie modelu, gdy wybrana rodzina to "binomial". Parametr ten, przybierając wartość 1, wskazuje na dobre dopasowanie modelu.

Residual deviance (646.33) jest mniejsza niż null deviance (665.80), co wskazuje na to, że model z BMI jako predyktorem jest lepiej dopasowany niż model zerowy.

Podsumowując, wyniki sugerują, że wraz ze wzrostem poziomu BMI, rośnie szansa na występowanie zespołu policystycznych jajników.

Zmienna przekształcona:

```
data$PCOS <- as.factor(data$PCOS)
model.bmi.log <- glm(PCOS ~ log(BMI), family = 'binomial', data)
summary(model.bmi.log)
```

```
##
## Call:
## glm(formula = PCOS ~ log(BMI), family = "binomial", data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.3910     1.9171  -4.377 1.20e-05 ***
## log(BMI)      2.4000     0.5984   4.011 6.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.8  on 528  degrees of freedom
## Residual deviance: 648.4  on 527  degrees of freedom
## AIC: 652.4
##
## Number of Fisher Scoring iterations: 4
```

Wniosek:

Ponownie zlogarytmowanie zmiennej objaśniającej nie zmieniło nic, biorąc pod uwagę dopasowanie modelu.

Teraz również wykonamy test Hosmera-Lemeshowa:

```
data$PCOS <- as.numeric(as.character(data$PCOS))

hoslem.test((data %>% select(PCOS, BMI) %>% drop_na())$PCOS, fitted(model.bmi))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, BMI) %>% drop_na())$PCOS, fitted(model.bmi)
## X-squared = 15.054, df = 8, p-value = 0.05811
```

```
hoslem.test((data %>% select(PCOS, BMI) %>% drop_na())$PCOS, fitted(model.bmi.log))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, BMI) %>% drop_na())$PCOS, fitted(model.bmi.log)
## X-squared = 16.545, df = 8, p-value = 0.03521
```

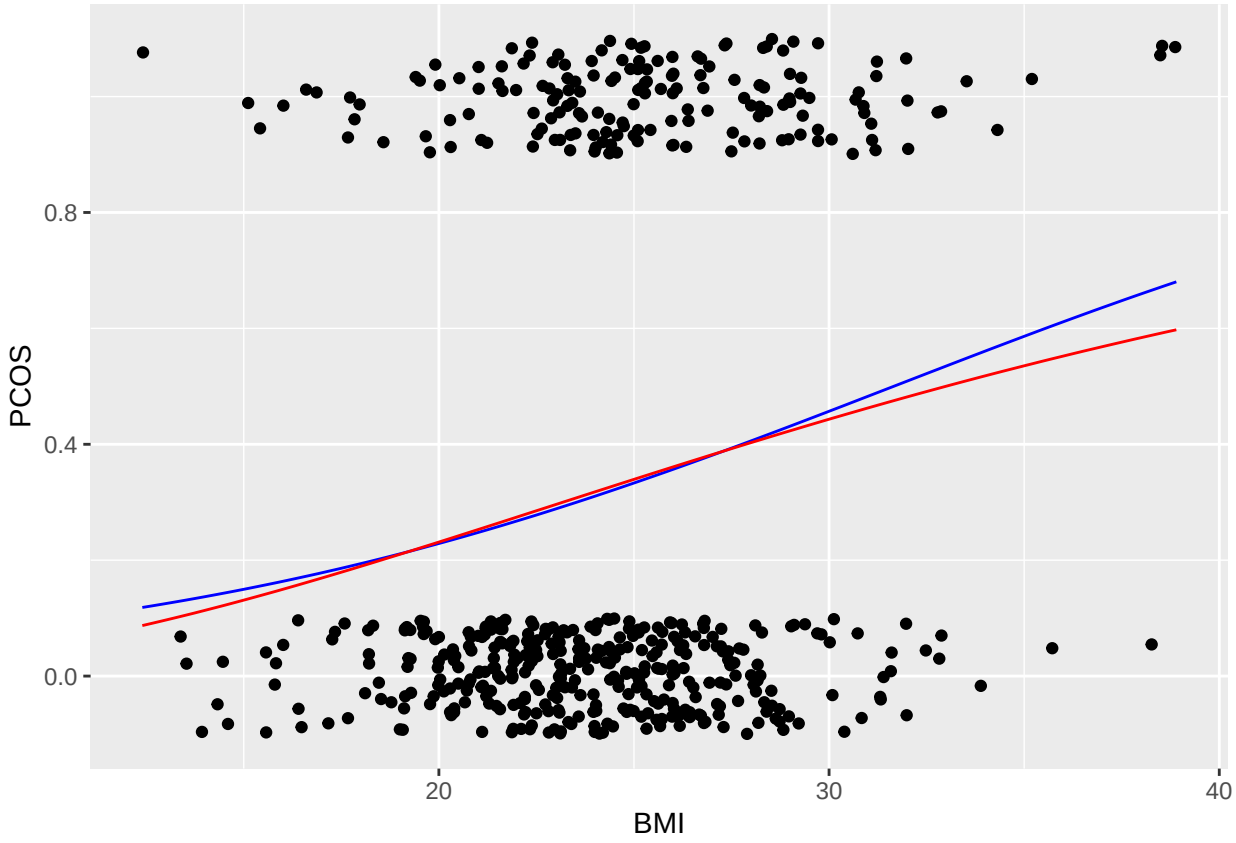
wniosek:

P-value dla pierwszego testu wynosi 0.05811, a to sugeruje brak istotnych różnic między danymi obserwowanymi a danymi przewidywanymi przez model.

W przypadku drugiego testu wartość p-value wynosi 0.03521, co sugeruje, że istnieją istotne różnice między danymi obserwowanymi a danymi przewidywanymi przez model.

Graficzne porównanie obydwu modeli

```
ggplot(data, aes(x = BMI, y = PCOS)) + geom_point(position = position_jitter(height = 0.1)) +
  stat_function(fun = function(x) predict(object = model.bmi, newdata = data.frame(BMI = x),
  type = 'response'), color = 'blue') + stat_function(fun = function(x) predict(object = model.bmi.log,
  newdata = data.frame(BMI = x), type = 'response'), color = 'red')
```



Tak jak poprzednio budujemy model z dwoma predyktorami:

```
data$PCOS <- as.factor(data$PCOS)
model.twoo <- glm(PCOS ~ BMI + log2(BMI), binomial, data)
```

```
data$PCOS <- as.numeric(as.character(data$PCOS))

hoslem.test((data %>% select(PCOS, BMI) %>% drop_na())$PCOS, fitted(model.twoo))
```

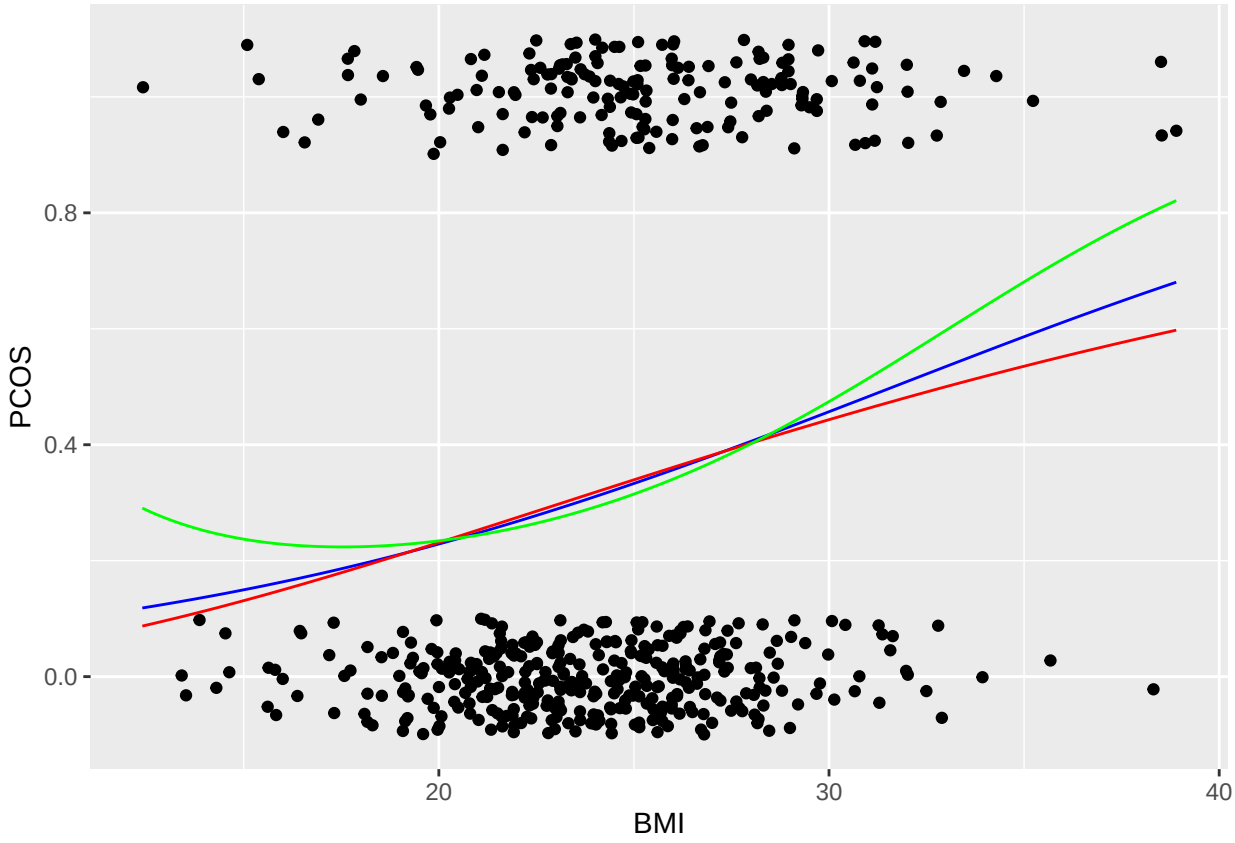
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, BMI) %>% drop_na())$PCOS, fitted(model.twoo)
## X-squared = 14.404, df = 8, p-value = 0.07183
```

Wniosek:

Test Hosmera-Lemeshowa mówi nam, że model nie jest najlepiej dopasowany, ponieważ pvalue wynosi 0.07183.

Graficzne porównanie trzech modeli:

```
ggplot(data, aes(x = BMI, y = PCOS)) +
  geom_point(position = position_jitter(height = 0.1)) +
  stat_function(fun = function(x) predict(object = model.bmi, newdata = data.frame(BMI = x),
      type = 'response'), color = 'blue') +
  stat_function(fun = function(x) predict(object = model.bmi.log, newdata = data.frame(BMI = x),
      type = 'response'), color = 'red') +
  stat_function(fun = function(x) predict(object = model.twoo, newdata = data.frame(BMI = x),
      type = 'response'), color = 'green')
```



Wniosek:

Zgodnie z wynikami testu Hosmera-Lemeshowa, model z dwoma predyktoramii wygląda na najlepiej dopasowany spośród wszystkich.

Badanie hipotezy pierwszej:

Z uwagi na to, że puls i oba rodzaje ciśnienia są ze sobą ściśle powiązane w organiźmie człowieka, sprawdzimy jak wpływa na model puls.

```
model.h1 <- glm(PCOS ~ TSH + FSH + Pulse.rate.bpm., binomial, data)
summary(model.h1)
```

```
##
## Call:
## glm(formula = PCOS ~ TSH + FSH + Pulse.rate.bpm., family = binomial,
##     data = data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.190399   2.465445  -2.105  0.03527 *
## TSH             -0.005356   0.044415  -0.121  0.90402
## FSH             -0.124776   0.044986  -2.774  0.00554 **
## Pulse.rate.bpm.  0.069001   0.033101   2.085  0.03711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.80  on 528  degrees of freedom
## Residual deviance: 651.52  on 525  degrees of freedom
## AIC: 659.52
##
## Number of Fisher Scoring iterations: 4
```

Wniosek:

Przy powyższym modelu współczynnik FSH jest istotny statystycznie (0.00554), puls również, natomiast współczynnik TSH. Puls również jest mniejszy od 0.05.

Odchylenie resztowe jest mniejsze niż odchylenie zerowe, zatem powyższy model jest lepiej dopasowany niż model zerowy.

Co wykaże test?

```
data$PCOS <- as.numeric(as.character(data$PCOS))

hoslem.test((data %>% select(PCOS, FSH, TSH, Pulse.rate.bpm.) %>% drop_na())$PCOS, fitted(model.h1))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, FSH, TSH, Pulse.rate.bpm.) %>% drop_na())$PCOS, fitted(model.h1)
## X-squared = 6.0978, df = 8, p-value = 0.6363
```

Wniosek:

Test wskazuje na bardzo dobre dopasowanie modelu, ze względu na p-value na poziomie 0.6363, czyli znacznie większym niż przyjęty przez nas poziom istotności.

Możemy zatem założyć, że hipoteza jest prawdziwa.

Badanie hipotezy dugiej:

Zbudujemy na początek model zawierający zmienne z hipotezy:

```
model.h2 <- glm(PCOS ~ Cycle.R.I. + Skin.darkening..Y.N. + Pimples.Y.N., binomial, data)
summary(model.h2)
```

```
##
## Call:
## glm(formula = PCOS ~ Cycle.R.I. + Skin.darkening..Y.N. + Pimples.Y.N.,
##     family = binomial, data = data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.6527     0.2292 -11.574  < 2e-16 ***
## Cycle.R.I.1             1.7132     0.2454   6.981 2.94e-12 ***
## Skin.darkening..Y.N.1   2.0377     0.2379   8.567  < 2e-16 ***
## Pimples.Y.N.1           1.1630     0.2345   4.960 7.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.8  on 528  degrees of freedom
## Residual deviance: 471.6  on 525  degrees of freedom
## AIC: 479.6
##
## Number of Fisher Scoring iterations: 5
```

Wniosek:

Wszystkie współczynniki użyte w powyższym modelu są istotne statystcznie. Ponadto odchylenie resztowe jest widocznie mniejsze niż zerowe (sugerując się wcześniejszymi modelami, ten model jest najlepszy względem przyjętego modelu zerowego).

Dodatkowo AIC również jest najmniejsze niż w którymkolwiek innym modelu.

Oznacza to, że występowanie nieregularnej miesiączki, ciemnienie skóry oraz wypryski są dobrymi predyktorami występowania zespołu policystycznych jajników.

Wykonujemy test:

```
data$PCOS <- as.numeric(as.character(data$PCOS))

hoslem.test((data %>% select(PCOS, Cycle.R.I., Skin.darkening..Y.N., Pimples.Y.N.) %>%
            drop_na())$PCOS, fitted(model.h2))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  (data %>% select(PCOS, Cycle.R.I., Skin.darkening..Y.N., Pimples.Y.N.) %>% , fitted(model.h2)
## X-squared = 1.8172, df = 8, p-value = 0.9861
```

Wniosek:

Wyniki powyższego testu mówią o bardzo dobrym dopasowaniu modelu do danych.

W związku z tym potwierdzamy naszą hipotezę, że ciemnienie skóry, pojawianie się wyprysków i nieregularnych miesiączek może świadczyć o chorowaniu na zespół policystycznych jajników.

Odstawmy jednak statystykę i testy na bok. Oczywistym jest, że występowanie powyższych objawów nie determinuje chorowania na PCOS. Jednak większość osób chorujących na omawiana przez nas chorobę, skarży się na te objawy - stąd tak dobre dopasowanie modelu.

Badanie hipotezy trzeciej:

Z uwagi na to, że hormon FSH u kobiet pobudza wzrost pęcherzyków jajnikowych w jajniku, sprawdźmy czy model to potwierdzi.

```
model.h3 <- lm(Follicle.No...L. ~ FSH, data)
summary(model.h3)
```

```
##
## Call:
## lm(formula = Follicle.No...L. ~ FSH, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6039 -3.1139 -0.8494  2.3672 15.7081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.56870    0.44610  16.967  < 2e-16 ***
## FSH         -0.28887    0.08195  -3.525  0.00046 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.203 on 527 degrees of freedom
## Multiple R-squared:  0.02303,    Adjusted R-squared:  0.02118
## F-statistic: 12.43 on 1 and 527 DF,  p-value: 0.0004605
```

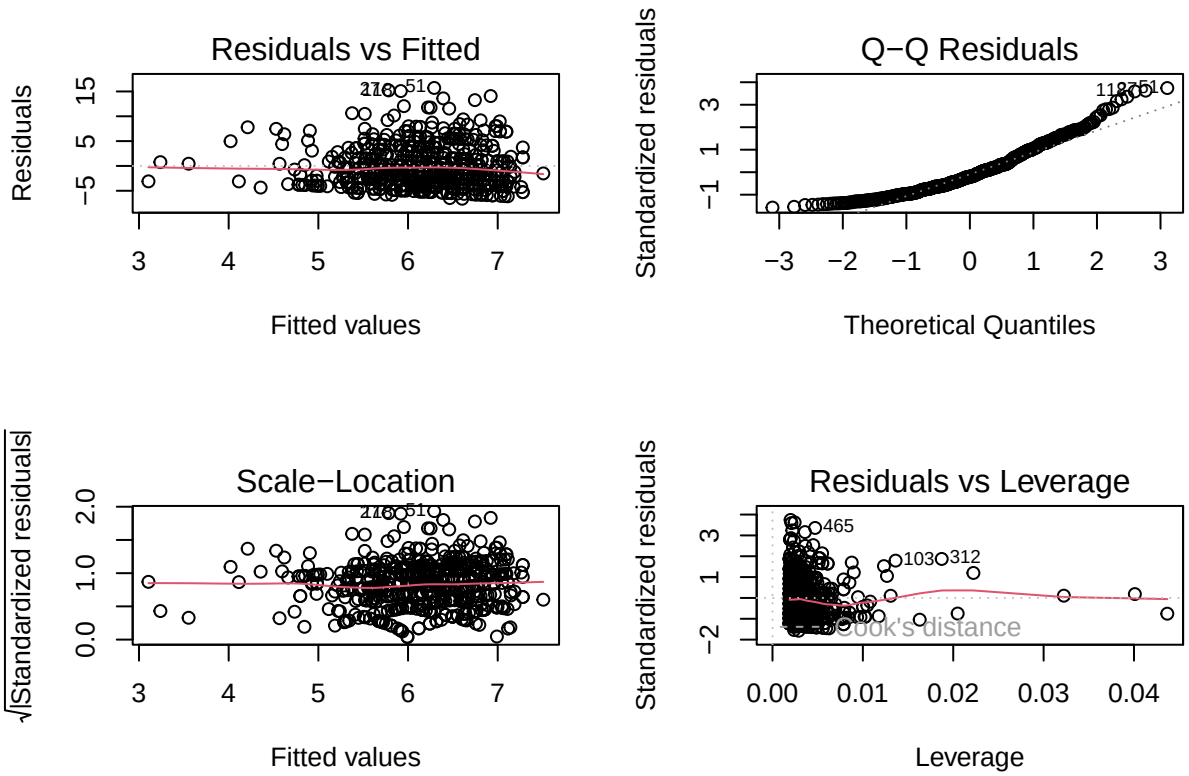
Wniosek:

Powyższy model liniowy potwierdza, że FSH jako zmienna objaśniająca jest istotna statystycznie. Oznacza to, że poziom FSH istotnie wpływa na ilość pęcherzyków w jajnikach.

Warto jednak zauważyć, że Multiple R-squared wynosi jedynie 0.02303, co oznacza, że FSH jako zmienna objaśniająca, wyjaśnia małą część powodów, dla których liczba pęcherzyków w jajnikach ulega zmianie.

Wykresy diagnostyczne:

```
par(mfrow = c(2, 2))
plot(model.h3)
```



wniosek:

Pomimo, że wykresy diagnostyczne już wyglądają bardzo dobrze, w związku z tym, sprawdzimy czy model regresji ważonej będzie lepszym dopasowaniem.

Na początek stworzymy model pomocniczy:

Użyjemy wartości bezwzględnej modelu reszt oraz wartości dopasowanych.

```
model.h3.resid <- lm(abs(resid(model.h3)) ~ fitted(model.h3))
```

Tworzymy wagi:

```
s <- fitted(model.h3.resid)
model.weights <- 1 / s^2
```

Model regresji ważonej:

```
model.h3.weighted <- lm(Follicle.No...L. ~ FSH, data, weights = model.weights)
summary(model.h3.weighted)
```

```
##
## Call:
## lm(formula = Follicle.No...L. ~ FSH, data = data, weights = model.weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9158 -0.9494 -0.2483  0.7403  4.6796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.61517    0.44312  17.185  < 2e-16 ***
## FSH         -0.29813    0.07782  -3.831 0.000143 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.257 on 527 degrees of freedom
## Multiple R-squared:  0.0271, Adjusted R-squared:  0.02525
## F-statistic: 14.68 on 1 and 527 DF,  p-value: 0.0001429
```

Wniosek:

Analiza modelu regresji ważonej potwierdziła, że poziom FSH jest istotny statystycznie dla ilości pęcherzyków w jajnikach.

P-value na poziomie 0.000143 dla FSH oznacza, że istnieją statystycznie istotne dowody na istnienie zależności między zmienną niezależną (FSH) a zmienną zależną (Follicle.No. . . L.).

Współczynnik determinacji R-squared wynosi 0.0271, co jest niewielką zmianą w porównaniu do oryginalnego modelu.

Wartość statystyki F - 14.68 i niska p-value sugerują, że istnieje statystycznie istotna zależność między zmiennymi FSH i Follicle.No. . . L.

Jednakże, ze względu na niski współczynnik determinacji, model nie wyjaśnia dużego odsetka zmienności zmiennej zależnej.

Podsumowaując: Zadana przez nas hipoteza jest prawdziwa i poziom FSH ma istotny wpływ na liczbę pęcherzyków w jajnikach. Trzeba jednak pamiętać, że nie jest to jedyna zmienna mająca wpływ na ilość pęcherzyków. FSH jako jedyny predyktor nie jest wystarczająco silny do przewidzenia ilości pęcherzyków w jajnikach i do podwyższenia współczynnika determinacji potrzebna by była bardziej dogłębna analiza.