BMI – rozkład Normalny/ Weibulla

Maciej Odziemczyk, nr albumu: 388581

2020-01-20

Abstrakt

Celem pracy jest sprawdzenie dopasowania rozkładu BMI do rozkładów Normalnego i dwuparametrowego Weibulla (wybranego ze względu na swoją elastyczność). Jako drugi cel postawiono sprawdzenie stanu zdrowia społeczeństwa według standardów Światowej Organizacji Zdrowia. Wyniki analiz rozkładów wskazały na zbliżone wyniki dopasowania, w zależności od metody estymacji, lepszy okazywał się jeden lub drugi rozkład teoretyczny.

Jako metody estymacji parametrów posłużyły: Metoda Największej Wiarygodności (MNW) oraz Uogólniona Metoda Momentów (UMM). Hipotezy złożone poddane zostały testom: LR – MNW, Walda – UMM, natomiast hipotezy proste weryfikowano testem z.

Wprowadzenie

Wskaźnik BMI to stosunek wagi wyrażonej w kilogramach do kwadratu wzrostu wyrażonego w metrach. Na jego podstawie Światowa Organizacja Zdrowia klasyfikuje osoby jako cierpiące na niedowagę, otyłe lub posiadające wagę w normie. Widełki są następujące: >18,5 – niedowaga, 18,5-24,9 – norma, 25-29,9 – stan przed otyłością, 30-34,9 – otyłość pierwszego stopnia, 35-39,9 – otyłość drugiego stopnia, <40 – otyłość trzeciego stopnia[1]. Wskaźnik ten oparty jest na normach tkanki tłuszczowej, która w nadmiarze może spowodować przedwczesną śmierć. Wraz z rosnącym odchyleniem BMI powyżej normy, rośnie ryzko przypadłości takich jak choroby sercowo-naczyniowe, wysokie ciśnienie krwi, zapalenie kości i stawów, niektóre nowotwory czy cukrzyca. Z uwagi na konsekwencje nadmiaru tłuszczu w organiźmie ważne jest aby wiedzieć jak modelować to zjawisko. Hipotezy badawcze niniejszej pracy to średni wskaźnik BMI na poziomie 37,5 wraz z odchyleniem 2,5 co należy do klasy otyłości drugiego stopnia. Wnioski mogą przydać się organizacjom dbającym o zdrowie społeczeństwa, którymi *de facto* powinny być wszystkie organizacje, którym zależy również na wzroście gospodarczym, gdyż przedwczesna śmierć i choroby uniemożliwiające normalne funkcjonowanie wydają się być jedną z naturalnych przyczyn spadków w rozwoju państw.

Dane

Dane pochodzą z serwisu „kaggle.com" i w oryginale zawierają informację o płci, wzroście, wadze oraz klasie, do której należy dana osoba. Ze względu na nieznane pochodzenie znajdującej się w danych klasyfikacji posłużono się policzonym na nowo wskaźnikiem BMI według wzoru:

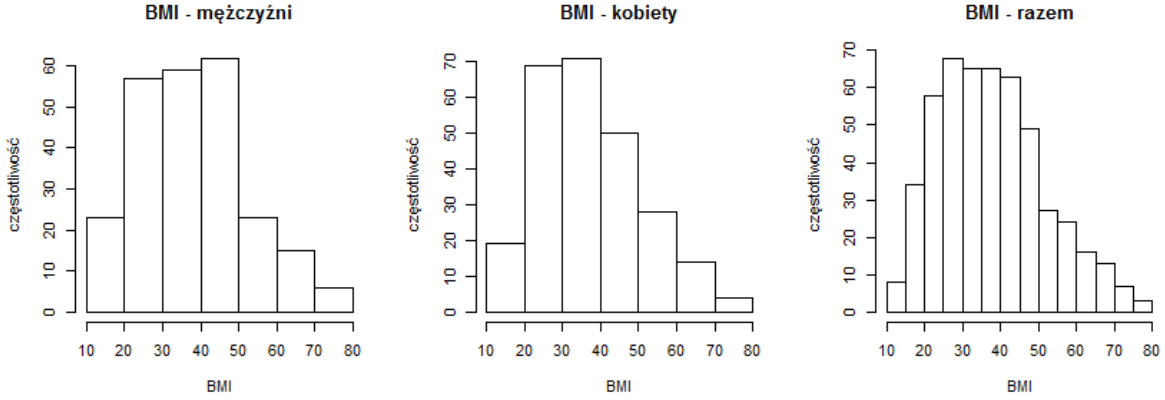
$$BMI = \frac{waga[kg]}{(wzrost[m])^2}$$

W celu uzyskania większej precyzji zdecydowno się podzielić próbę ze względu na płeć. Statystki opisowe prezentują się następująco:

	min	max	średnia	mediana	dominanta	odchylenie std.
Mężczyźni	12.7538	77.55102	38.15161	38.38831	29.72652	14.08098
Kobiety	12.7538	78.85340	37.39410	35.73960	27.35043	13.87141
Razem	12.7538	78.85340	37.76528	36.95694	37.34559	13.96562

[1] On-line: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi, [Dostęp: 19-01-2020]

W wyodrębnionych grupach znalazły się zarówno osoby ze skrajnie niskim BMI oraz znacznie powyżej granicy otyłości trzeciego stopnia. Jednocześnie 50% kobiet charakteryzuje się BMI poniżej granicy między pierwszą, a drugą klasą otyłości. Mężczyzn opisuje natomiast mediana ze wskaźnikiem większym o 2,5 oraz prawie o 1 większa średnia. Grupy są wyraźnie zróżnicowane, co opisuje odchylenie standardowe, stanowiące około 36% średniej. Dominanta niższa od średniej, wskazuje na asymetrię prawostronną każdej z grup, co można dodatkowo zauważyć na poniższych histogramach.



Metoda

Estymacja parametrów rozkładów Normalnego i Weibulla została dokonana za pomocą Metody Największej Wiarygodności (MNW) oraz Uogólnionej Metody Momentów (UMM). Dopasowanie zostało natomiast określone na podstawie wykresów kwantylowych (QQ-plots) dla otrzymanych rozkładów teoretycznych oraz rozkładu empirycznego. Weryfikacja hipotez odbyła się za pomocą testu LR (MNW) i testu Walda (UMM), dla rozkładu dwuparametrowego Weibulla oraz statystyki z, dla hipotez prostych.

Metoda Największej Wiarygodności polega na maksymalizacji funkcji wiarygodności, która w ogólnej postaci, dla ciągłej zmiennej losowej wygląda następująco:

$$L(X|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

Gdzie X stanowi próbę losową, θ to wektor parametrów rozkładu, P to funkcja gęstości prawdopodobieństwa, a x_i to obserwacja z próby. W niniejszej pracy zdecydowano się dla uproszczenia obliczeń zmaksymalizować logarytmy funkcji wiarygodności.

Gęstość rozkładu Weibulla:

$$f(x;k,\lambda) = \frac{k}{\lambda^k} x^{k-1} e^{-(x/\lambda)^k} \qquad x > 0$$

Funkcja log-wiarygodności Weibulla:

$$ln\, L(x_i;k,\lambda) = N\, ln(k) - Nk\, ln(\lambda) + (k-1)\sum_{i=1}^{N} ln\,(x_i) - \sum_{i=1}^{N} \left(\frac{x_i}{\lambda}\right)^k$$

Gęstość rozkładu Normalnego:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Funkcja log-wiarygodności rozkładu Normalnego:

$$\ln L(x_i;\mu,\sigma) = -N\ln(\sigma) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

Uogólniona Metoda Momentów polega natomiast na minimalizacji różnic pomiędzy momentami próbkowymi, a estymatorami momentów teoretycznych. Jeżeli momenty teoretyczne to: E[X] = μ, E[(X - E[X])^2) = σ^2, wówczas ich estymatory to: $\frac{1}{N}\sum_{i=1}^{N} x_i = \hat{\mu}$ oraz $\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 = \hat{\sigma}^2$.

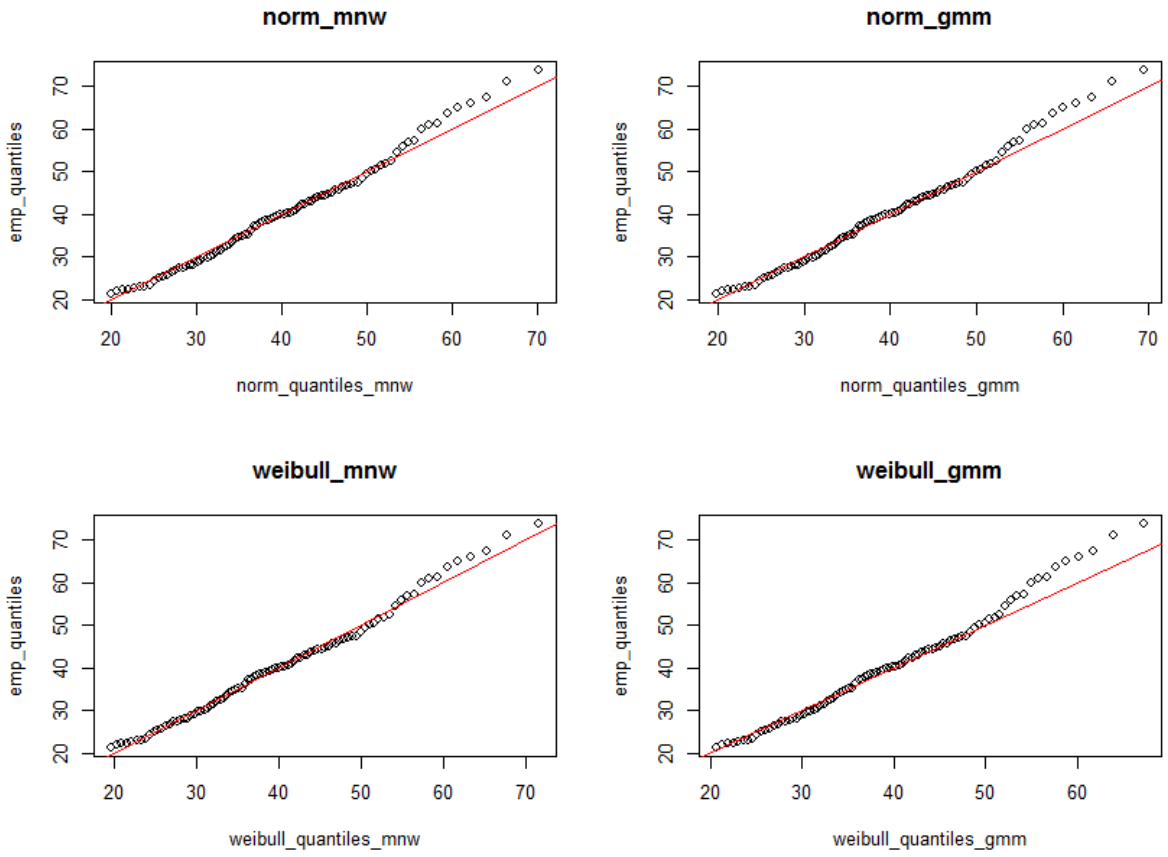
Wyniki

Wyniki estymacji:

	Mężczyźni	Kobiety	Razem
mu_MNW	38.151614	37.394104	37.765284
sigma_MNW	14.052212	13.844189	13.951648
mu_GMM	37.870830	37.065965	37.449414
sigma_GMM	13.830119	13.671200	13.754360
k_MNW	2.924959	2.898161	2.910302
lambda_MNW	42.828129	42.003062	42.408591
k_GMM	3.227974	3.167969	3.195486
lambda_GMM	42.045427	41.424814	41.732964

QQ-plots:

Z uwagi na podobieństwo wykresów w grupach, poniżej znajdują się wykresy kwantylowe dla całej próby tj. łącznie dla kobiet i mężczyzn.



Z analizy wykresów wynika, że rozkład BMI można modelować za pomocą rozkładów Normalnego i Weibulla z wyestymowanymi parametrami (Metoda Największej Wiarygodności okazała się jednak dokładniejsza - lepsze dopasowanie do linii 45^{o} w obydwu przypadkach).

Badanie hipotez (tabele p value):

Hipoteza złożona H0: $\mu = 37,5$ i $\sigma = 2,5$

	Mężczyźni	Kobiety	Razem
LR_test	0	0	0
test Walda	0	0	0

Hipoteza prosta: H0: k = 3

	Mężczyźni	Kobiety	Razem
z test MNW	0.000000e+00	0.000000e+00	0.0000000000
z test UMM	4.181532e-09	2.894101e-06	0.0001958469

Zerowe p value dla hipotezy złożonej w przypadku każdej grupy wyodrębnionej z próby oznacza konieczność jej odrzucenia, oznacza to, że pesymistyczny scenariusz społeczeństwa otyłego w stopniu drugim nie sprawdził się. Analizując powtórnie statystyki opisowe można wysnuć wniosek, że przyczyną odrzucenia badanej hipotezy jest znacznie większe od testowanego zróżnicowanie, objawiające się w odchyleniu standardowym zbliżonym do 14, co można zaobserwować również na histogramach.

Hipoteza o parametrze skali równym 3 w wyestymowanym modelu Weibulla również musi zostać odrzucona ze względu na znikome wartości p value (same 0 w MNW i wartości znacznie poniżej 0,05 w UMM). Powtórna analiza wykresów kwantylowych w kontekście wyników testów z pozwala wyciągnąć podobne wnioski. Zerowa wartość p, dla modeli estymowanych MNW objawia się w bardzo dobrym pokryciu kwantyli teoretycznych z empirycznymi na wykresach QQ-plots, jednocześnie w tabeli p value można zauważyć wyższe wyniki dla estymacji Uogólnionej Metody Momentów, co również odnajduje swoje odzwierciedlenie w gorszym dopasowaniu na wykresie kwantylowym.

Wnioski

Dopasowanie rozkładu BMI, na które składa się wzrost oraz waga, do rozkładu Normalnego nie jest niczym nadzwyczajnym, gdyż takie cechy zazwyczaj modelowane są za jego pomocą. Interesującym natomiast jest fakt, że owa cecha w świetle niniejszego badania może być z powodzeniem modelowana za pomocą dwuparametrowego rozkładu Weibulla, co oznacza niezwykłą elastyczność tego rozkładu (za jego pomocą dokonuje się analiz przeróżnych zjawisk, np. magnitudy trzęsień ziemi, awaryjności czy siły wiatru). Kolejnym pozytywnym aspektem badania okazał się fakt odrzucenia hipotezy o otyłości społeczeństwa w stopniu drugim według standardów Światowej Organizacji Zdrowia, jednocześnie w dysponowanej próbie mężczyźni okazali się grupą o gorszej kondycji fizycznej. Ze względu na fakt zróżnicowania stylu życia w różnych częściach świata, warto powyższe badanie przeprowadzić dla różnych grup społecznych w celu estymacji, o ile to możliwe, jednego wspólnego modelu, co zdecydowanie ułatwiłoby procesy decyzyjne dot. szeroko pojętego zdrowia fizycznego.

Aneks z kodem

```
#dane: https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex
library("maxLik") #Biblioteka do metody NR

# Instrukcja, po ustawieniu ścieżki dostępu, można uruchomić cały skrpyt
# wyświetlone zostaną wyniki i wykresy, które opisane są w pracy
# do regulacji wykresów kwantylowych służy zmienna wiersz, mogąca przyjmować parametry
# od 1 do 3 (1-Mężczyźni, 2-Kobiety, 3-Cała próbka)

getwd()
setwd('F:/Studia/II stopień/WNE/Programowanie narzędzi analitycznych/Projekt')


#**********************************************************************
#                    Dane: zmienna - BMI
#**********************************************************************
dane = read.csv(file = "BMI.csv", header = TRUE, sep = ";" , dec = ",")

indeksy = which(dane$Gender=="Male")

# Mężczyźni
x1 = dane$BMI[indeksy]
# Kobiety
x2 = dane$BMI[-indeksy]
# Razem
x3 = dane$BMI
# lista próbek
X = list(x1, x2, x3)

# ***Histogramy***
par(mfrow=c(1,3))
hist(x1, main = "BMI - mężczyźni", xlab = "BMI", ylab = "częstotliwość")
hist(x2, main = "BMI - kobiety", xlab = "BMI", ylab = "częstotliwość")
hist(x3, main = "BMI - razem", xlab = "BMI", ylab = "częstotliwość")

#***Tabele do zapisywania wyników (macierze)***
# macierz statystyk populacji
statystyki = matrix(nrow = 3, ncol = 6)
row.names(statystyki) = c("Mężczyźni", "Kobiety", "Razem")
colnames(statystyki) = c("min", "max", "średnia", "mediana", "dominanta", "odchylenie std.")

# macierz wyników estymacji
estymacje = matrix(nrow = 3, ncol = 8)
row.names(estymacje) = c("Mężczyźni", "Kobiety", "Razem")
colnames(estymacje) = c("mu_MNW", "sigma_MNW", "mu_GMM", "sigma_GMM", "k_MNW",
"lambda_MNW", "k_GMM", "lambda_GMM")

# macierz wynikóW hipotez złożonych
zlozone = matrix(nrow = 2, ncol = 3)
row.names(zlozone) = c("LR_test ","test Walda")
colnames(zlozone) = c("Mężczyźni", "Kobiety", "Razem")
```

```
# macierz wynikóW hipotez prostych
proste = matrix(nrow = 2, ncol = 3)
row.names(proste) = c("z test MNW","z test UMM")
colnames(proste) = c("Mężczyźni", "Kobiety", "Razem")
proste

#***Obliczanie statystyk opisowych***
Dominanta = function(x) {
  ux = unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
# zapis do tabeli wyników
i = 1
for(x in X){
  statystyki[i,1] = min(x)          # min
  statystyki[i,2] = max(x)          # max
  statystyki[i,3] = mean(x)         # średnia
  statystyki[i,4] = median(x)       # mediana
  statystyki[i,5] = Dominanta(x)    # dominanta
  statystyki[i,6] = sd(x)           # odchylenie standardowe
  i = i+1
}
#**********************************************************************
#                    Estymacje
#**********************************************************************
#----------------------ROZKŁAD NORMALNY---------------------------
#--------------------------MNW-----------------------------------
# funkcja log-wiarygodności
lnl_norm = function(parms){
  mu = parms[1]
  s = parms[2]
  ll = -N/2*log(2*pi)-N*log(s)-1/2*sum((x-mu)^2/(s^2))
  return(ll)
}
# gradient
gr_norm = function(params){
  mu = params[1]
  s = params[2]
  g = rep(0, times=2)
  g[1] = sum((x-mu)/(s^2))
  g[2] = -N/s+sum((x-mu)^2/(s^3))
  return(g)
}
# hessian
he_norm = function(params){
  mu = params[1]
  s = params[2]
  h = matrix(0, nrow=2, ncol=2)
  h[1,1] = -N/(s^2)
  h[2,2] = N/(s^2)-3*sum((x-mu)^2/(s^4))
  h[1,2] = -2*sum((x-mu)/(s^3))
```

```
  h[2,1] = h[1,2]
  return(h)
}
# zapisanie do tabeli wyników
i = 1
for(x in X){
  N = length(x)
  wynik = maxNR(fn=lnl_norm, grad=gr_norm, hess=he_norm, start=c(20, 1))
  estymacje[i,1] = wynik$estimate[1]
  estymacje[i,2] = wynik$estimate[2]
  i = i+1
}

#---------------------------GMM-----------------------------------
Qmin_norm = function(parms){
  mu = parms[1]
  s = parms[2]
  m = rbind(x-mu, (x-mu)^2-s^2, (x-mu)^4-3*s^4)
  W = m%*%t(m)/N
  M = rowMeans(m)
  val = -t(M)%*%solve(W)%*%M
  return(val)
}
# zapisanie do tabeli wyników
i = 1
for(x in X){
  N = length(x)
  wynik = maxNR(fn = Qmin_norm,  start = c(mean(x), sd(x)))
  estymacje[i,3] = wynik$estimate[1]
  estymacje[i,4] = wynik$estimate[2]
  i = i+1
}
#---------------------ROZKŁAD WEIBULLA----------------------------
#---------------------------MNW-----------------------------------
# funkcja log-wiarygodności
lnl_wbl = function(parms){
  k = parms[1]
  lambda = parms[2]
  ll = N*log(k)-N*k*log(lambda)+(k-1)*sum(log(x))-sum((x/lambda)^k)
  return(ll)
}
# gradient
gr_wbl = function(parms){
  k = parms[1]
  lambda = parms[2]
  g = rep(0, times = 2)
  g[1] = N/k-N*log(lambda)+sum(log(x))-sum((x/lambda)^k*log(x/lambda))
  g[2] = -N*k/lambda+sum(k*(x^k)/(lambda^(k+1)))
  return(g)
}
# hessian
he_wbl = function(parms){
```

```
  k = parms[1]
  lambda = parms[2]
  h = matrix(0, nrow = 2, ncol = 2)
  h[1,1] = -N/(k^2)-sum((x/lambda)^k*log(x/lambda)*log(x/lambda))
  h[2,2] = N*k/(lambda^2)+sum(k*lambda^(-k-2)*x^k*(-k-1))
  h[1,2] = -N/lambda-sum(-k*(x^k)/(lambda^(k+1))*log(x/lambda)+((x/lambda)^k)*(-1/lambda))
  h[2,1] = h[1,2]
  return(h)
}
# zapisanie do tabeli wyników
i = 1
for(x in X){
  N = length(x)
  wynik = maxNR(fn=lnl_wbl, grad=gr_wbl, hess=he_wbl, start=c(0.1, 0.1))
  estymacje[i,5] = wynik$estimate[1]
  estymacje[i,6] = wynik$estimate[2]
  i = i+1
}

#--------------------------GMM-----------------------------------
Qmin_wbl = function(parms){
  k = parms[1]
  lambda = parms[2]
  m = rbind(x-(lambda*gamma(1+1/k)),
        (x-(lambda*gamma(1+1/k)))^2-(lambda^2*(gamma(1+2/k)-(gamma(1+1/k))^2)),
        (x-(lambda*gamma(1+1/k)))^4-(lambda*gamma(1+3/k))^3
  )
  W = m%*%t(m)/N
  M = rowMeans(m)
  val = -t(M)%*%solve(W)%*%M
  return(val)
}
# zapisanie do tabeli wyników
i = 1
for(x in X){
  N = length(x)
  wynik = maxNR(fn = Qmin_wbl,  start = c(3, 40))
  estymacje[i,7] = wynik$estimate[1]
  estymacje[i,8] = wynik$estimate[2]
  i = i+1
}

#**********************************************************************
#                  Wykresy kwantylowe
#**********************************************************************
# kwantyle empiryczne
emp_quantiles = quantile(x = x1, probs = seq(0.1, 0.99, 0.01))
# kwanyle teoretyczne
wiersz = 3 # 1 - wykresy dla mężczyzn, 2 - dla kobiet, 3 - cała próba
norm_quantiles_mnw = qnorm(p=seq(0.1, 0.99, 0.01), mean = estymacje[wiersz,1], sd =
estymacje[wiersz,2])
```

```
weibull_quantiles_mnw = qweibull(p=seq(0.1, 0.99, 0.01), shape = estymacje[wiersz,5], scale =
estymacje[wiersz,6])
norm_quantiles_gmm = qnorm(p=seq(0.1, 0.99, 0.01), mean = estymacje[wiersz,3], sd =
estymacje[wiersz,4])
weibull_quantiles_gmm = qweibull(p=seq(0.1, 0.99, 0.01), shape = estymacje[wiersz,7], scale =
estymacje[wiersz,8])

# rysowanie wykresów
par(mfrow=c(2,2))
plot(norm_quantiles_mnw, emp_quantiles, main = "norm_mnw")
abline(0,1, col='red')
plot(norm_quantiles_gmm, emp_quantiles, main = "norm_gmm")
abline(0,1, col='red')
plot(weibull_quantiles_mnw, emp_quantiles, main = "weibull_mnw")
abline(0,1, col='red')
plot(weibull_quantiles_gmm, emp_quantiles, main = "weibull_gmm")
abline(0,1, col='red')

#*********************************************************************
#                    Hipotezy
#*********************************************************************
# złożone
i = 1
for(x in X){
 # LR_test
 wynik = maxNR(fn=lnl_norm, grad=gr_norm, hess=he_norm, start=c(20, 1))
 summary(wynik)
 lnl_U = wynik$maximum            # model bez ograniczeń
 lnl_R = lnl_norm(c(37.5, 2.5))    # model z ograniczeniami
 LR_test = 2*(lnl_U-lnl_R)        # należy do X^2 df=2
 p = 1-pchisq(q = LR_test, df = 2) # p_value
 zlozone[1,i] = p
 # test walda
 wynik = maxNR(fn = Qmin_norm,  start = c(mean(x), sd(x)))
 R = diag(2)
 theta = wynik$estimate
 q = rbind(37.5, 2.5)
 S = R%*%theta-q
 vcov = -solve(wynik$hessian)/N
 w_test = t(S)%*%solve(R%*%vcov%*%t(R))%*%S
 p = 1-pchisq(q = w_test, df = 2)
 zlozone[2,i] = p
 i = i+1
}
# proste
i = 1
for(x in X){
 # z_test MNW
 k_0 = 3
 wynik = maxNR(fn=lnl_wbl, grad=gr_wbl, hess=he_wbl, start=c(0.1, 0.1))
 vcov = -solve(wynik$hessian)/N                # macierz wariancji kowariancji
 std.err.k = sqrt(vcov[1,1])                  # błąd standardowy lambda
```

```
  z_test = (wynik$estimate[1]-k_0)/std.err.k       # z_test ~N(0,1)
  p = 2*(1-pnorm(q=abs(z_test), mean = 0, sd = 1))  # p_value
  proste[1,i] = p
  # z_test GMM
  wynik = maxNR(fn = Qmin_wbl,  start = c(3, 40))
  vcov = -solve(wynik$hessian)/N                   # macierz wariancji kowariancji
  std.err.k = sqrt(vcov[1,1])                    # błąd standardowy lambda
  z_test = (wynik$estimate[1]-k_0)/std.err.k       # z_test ~N(0,1)
  p = 2*(1-pnorm(q=abs(z_test), mean = 0, sd = 1))  # p_value
  proste[2,i] = p
  i = i+1
}

#**********************************************************************
#                    Wyniki
#**********************************************************************
statystyki
estymacje
zlozone
proste
```