ModeleLiniowe4

Katarzyna Stasińska

2023-12

Zadanie 1

a)

```
library(MASS)
sigma = matrix( c(1,0.9,0.9,1),2,2)
mi = c(0,0)
dane <- mvrnorm(n = 100, mi, sigma/100)
eps = rnorm(100,0,1)
Y = 3 * dane[,1] + eps
```

b)

prosty model

```
## Przedział ufności B1 = ( -0.3851523 , 3.692105 )
```

Ustalmy poziom istotności $\alpha = 0,05$

Hipotezy: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

```
## statystyka testowa: 1.609548 , p-wartosc: 0.1107133
```

Możemy zauważyć, że $p > \alpha$, zatem nie możemy odrzucić hipotezy H_0.

dwie zmienne objaśniające

```
## Przedział ufności B1 = ( -2.685149 , 6.253965 )
```

Ustalmy poziom istotności $\alpha = 0,05$

Hipotezy: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

```
## statystyka testowa: 0.7923728 , p-wartosc: 0.430077
```

Możemy zauważyć, że $p > \alpha$, zatem nie możemy odrzucić hipotezy H_0.

W przypadku żadnego z tych dwóch modeli nie możemy odrzucić hipotezy 0. Prosty model wyznacza krótszy przedział ufności dla parametru β_1. Oba przedziały ufności zawierają ustaloną wartość $\beta_1 = 3$.

c)

```
## Model pierwszy: Odchylenie standardowe estymatora B1 wynosi 1.027293
```

```
## Moc identyfikacji X1, to jest moc B1 wynosi 0.8492287
```

```
## Model drugi: Odchylenie standardowe estymatora B1 wynosi 2.25198
```

```
## Moc identyfikacji X1, to jest moc B1 wynosi 0.26129
```

d)

```
## model1, wyestymowane sd: 1.098682 wyestymowana moc: 0.767
```

```
## model2, wyestymowane sd: 2.356873 wyestymowana moc: 0.226
```

Wyniki są bliskie wartościom teoretycznym.

Możemy zwrócić uwagę, że przy uwzględnieniu w modelu dodatkowej zmiennej objaśniającej mocno skorelowanej ze zmienną, której istotność badamy, moc β_1 znacząco maleje.

Zadanie 2

a)

```
X = rnorm(950000,0,0.1)
X = matrix(X,1000,950)
B = c(3,3,3,3,3,rep(0,945))
eps = rnorm(1000,0,1)
Y = X %*% B + eps
```

b)

```
## k= 1
## liczba istotnych regresororów o indeksach <=5: 1
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1409.911
## MSE= 0.7689056
## AIC= 345.5267
## pvalue1= 1.511328e-13
##
## k= 2
## liczba istotnych regresororów o indeksach <=5: 2
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1286.466
## MSE= 4.390876
## AIC= 255.899
## pvalue1= 2.598342e-13 pvalue2= 1.174395e-21
##
## k= 5
## liczba istotnych regresororów o indeksach <=5: 5
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1054.176
## MSE= 5.246194
## AIC= 62.75955
## pvalue1= 4.828709e-16 pvalue2= 2.995634e-24
##
## k= 10
## liczba istotnych regresororów o indeksach <=5: 5
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1052.567
```

```
## MSE= 6.854848
## AIC= 71.2324
## pvalue1= 5.448309e-16 pvalue2= 2.779755e-24
##
## k= 50
## liczba istotnych regresororów o indeksach <=5: 5
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 4
##
## SSE= 998.2638
## MSE= 61.15853
## AIC= 98.26229
## pvalue1= 3.193309e-15 pvalue2= 2.633491e-23
##
## k= 100
## liczba istotnych regresororów o indeksach <=5: 5
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 6
##
## SSE= 957.2688
## MSE= 102.1535
## AIC= 156.329
## pvalue1= 1.802196e-15 pvalue2= 1.16943e-23
##
## k= 500
## liczba istotnych regresororów o indeksach <=5: 5
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 23
##
## SSE= 531.2964
## MSE= 528.126
## AIC= 367.5647
## pvalue1= 3.124083e-07 pvalue2= 2.292575e-12
##
## k= 950
## liczba istotnych regresororów o indeksach <=5: 2
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 37
##
## SSE= 45.27554
## MSE= 1014.147
## AIC= -1194.988
## pvalue1= 0.2989777 pvalue2= 0.01975346
```

Model biorący 5 pierwszych kolumn macierzy planu ma najmniejszą wartość AIC, zatem jest najlepszy biorąc pod uwagę to kryterium. Nie uwzględniamy modelu biorącego pod uwagę wszystkie kolumny, bo nie jest on zgodny z założeniami AIC (próbka musi być znacząco większa niż liczba estymowanych zmiennych).

c)

```
## k= 1
## liczba istotnych regresororów o indeksach <=5: 0
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1486.662
## MSE= 122.2475
## AIC= 398.5335
## pvalue1= 0.201998
##
```

```
## k= 2
## liczba istotnych regresororów o indeksach <=5: 0
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1486.656
## MSE= 218.0359
## AIC= 400.5296
## pvalue1= 0.2016962 pvalue2= 0.9504791
##
## k= 5
## liczba istotnych regresororów o indeksach <=5: 1
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1349.607
## MSE= 410.8723
## AIC= 309.8135
## pvalue1= 0.1585796 pvalue2= 0.8478164
##
## k= 10
## liczba istotnych regresororów o indeksach <=5: 1
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 0
##
## SSE= 1344.446
## MSE= 416.0336
## AIC= 315.9819
## pvalue1= 0.1407138 pvalue2= 0.8462107
##
## k= 50
## liczba istotnych regresororów o indeksach <=5: 1
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 3
##
## SSE= 1276.306
## MSE= 484.1733
## AIC= 343.9701
## pvalue1= 0.07835728 pvalue2= 0.7617413
##
## k= 100
## liczba istotnych regresororów o indeksach <=5: 3
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 11
##
## SSE= 1043.113
## MSE= 717.3664
## AIC= 242.2095
## pvalue1= 0.01727761 pvalue2= 0.7819373
##
## k= 500
## liczba istotnych regresororów o indeksach <=5: 5
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 309
##
## SSE= 302.5687
## MSE= 1457.911
## AIC= -195.447
## pvalue1= 2.955474e-14 pvalue2= 1.448752e-08
##
```

```
## k= 950
## liczba istotnych regresororów o indeksach <=5: 2
## liczba istotnych regresororów o pozostałych indeksach (liczba fałszywych odkryć): 37
##
## SSE= 45.27554
## MSE= 1715.204
## AIC= -1194.988
## pvalue1= 0.0002559 pvalue2= 0.006382929
```

Model biorący 500 pierwszych kolumn macierzy planu ma najmniejszą wartość AIC, zatem jest najlepszy biorąc pod uwagę to kryterium. Nie uwzględniamy modelu biorącego pod uwagę wszystkie kolumny, bo nie jest on zgodny z założeniami AIC (próbka musi być znacząco większa niż liczba estymowanych zmiennych).

Porównując oba przykłady możemy zauważyć, że gdy nie mamy żadnych informacji o tym, jakie regresory są istotne (podpunkt c)), test AIC wskazał model, który zawiera bardzo dużo potencjalnie istotnych regresorów (314), a wiemy, że jest ich jedynie 5.

W obu przypadkach SSE maleje wraz ze wzrostem k, a MSE rośnie.

d)

```
## [1] "Średnia liczba fałszywych odkryć"

## [1] "b)"

##      1      2      5     10     50    100    500    950
##  0.000  0.000  0.000  0.246  2.266  4.734 24.480 43.917

## [1] "c)"

##       1       2       5      10      50     100     500     950
##   0.041   0.107   0.367   0.791   5.050  11.863 288.541  43.917

## [1] "Moc identyfikacji X1"

## [1] "b)"

##     1     2     5    10    50   100   500   950
## 1.000 1.000 1.000 1.000 1.000 1.000 1.000 0.585

## [1] "c)"

##     1     2     5    10    50   100   500   950
## 0.547 0.550 0.552 0.555 0.594 0.631 1.000 0.998

## [1] "Moc identyfikacji X2"

## [1] "b)"

##     1     2     5    10    50   100   500   950
## 0.000 1.000 1.000 1.000 1.000 1.000 1.000 0.391

## [1] "c)"

##     1     2     5    10    50   100   500   950
## 0.000 0.372 0.378 0.391 0.433 0.490 1.000 0.992

## Średni rozmiar modelu b) 5.093

## Średni rozmiar modelu c) 293.186
```