Instrukcja - lista 3

12 listopada 2024

Zadanie 1

Rozpatrzymy klasyczny model regresji dany następującym wzorem:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{1}$$

gdzie $\varepsilon_i, i = 1, 2, ..., n$ są niezależnymi zmiennymi losowymi o rozkładzie $N(0, \sigma^2)$. Skonstruuj przedziały ufności dla parametrów β_0 i β_1 na danym poziomie ufności α. Wyniki wykonaj dla różnych długości prób n, $\alpha \in \{0.01, 0.05\}$ oraz $\sigma \in \{0.01, 0.5, 1\}$. Przy konstrukcji przedziałów ufności zakładamy, że σ jest wielkością znaną. Za pomocą metody Monte Carlo, sprawdź jakie jest prawdopodobieństwo, że teoretyczne wartości parametrów należą do wyznaczonych przedziałów ufności dla wybranych parametrów β_0 i β_1. W symulacjach przyjmij, że $x_i = i$ dla każdego i = 1, 2, $\cdots$, n.

1. **Konstrukcja przedziału ufności dla β_0:**

$$P\Big(\underbrace{\hat{\beta}_0 - z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}}_{a} \leqslant \beta_0 \leqslant$$

$$\leqslant \underbrace{\hat{\beta}_0 + z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}}_{b} \Big) = 1 - \alpha$$

$z_{1-\alpha/2}$ - kwantyl na poziomie $1 - \alpha/2$ ze standardowego rozkładu normalnego.

2. **Konstrukcja przedziału ufności dla β_1:**

$$P\Big(\underbrace{\hat{\beta}_1 - z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}}_{c} \leqslant \beta_1 \leqslant$$

$$\leqslant \underbrace{\hat{\beta}_1 + z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}}_{d} \Big) = 1 - \alpha$$

$z_{1-\alpha/2}$ - kwantyl na poziomie $1 - \alpha/2$ ze standardowego rozkładu normalnego.

3. **Estymacja parametru $\hat{P}U = 1 - \alpha$.**
Równania z poprzednich slajdów mówią nam, że prawdopodobieństwo tego, że estymowany parametr β_0 znajdzie się w skonstruowanym przedziale ufności $[a, b]$ wynosi $1 - \alpha$. Tak samo: prawdopodobieństwo tego, że estymowany parametr β_1 znajdzie się w skonstruowanym przedziale ufności $[c, d]$ wynosi $1 - \alpha$. Zatem możemy obliczyć estymatory parametru $1 - \alpha$ w następujący sposób:

$$\hat{P}U_{\beta_0} = \frac{\#\{\beta_0 \in [a, b]\}}{M}, \quad \hat{P}U_{\beta_1} = \frac{\#\{\beta_1 \in [c, d]\}}{M}$$

gdzie M to liczba powtórzeń Monte Carlo.

x

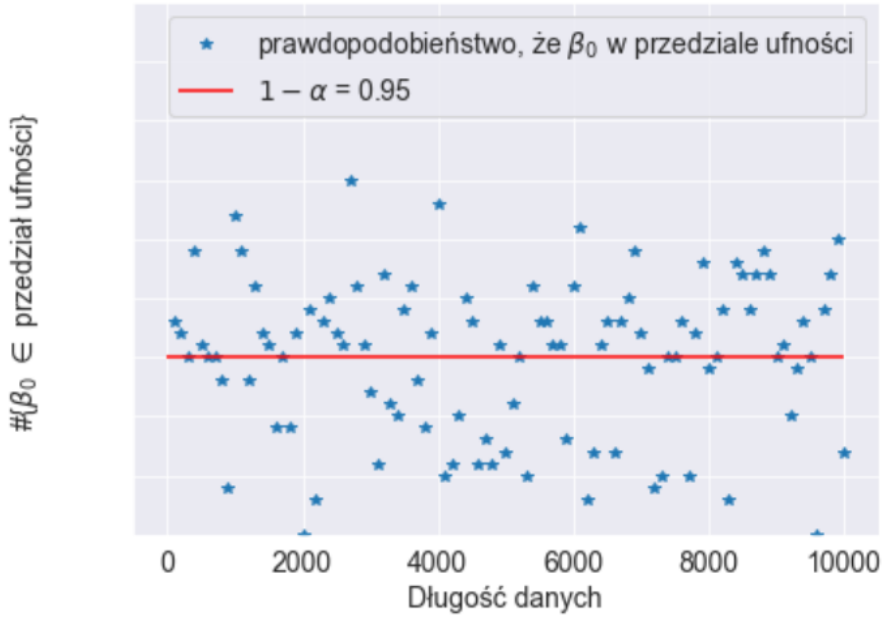
Zadanie 1

Algorytm

1. Weź $\beta_1 = 4, \beta_0 = 2$. Ustal α, σ tak jak w zadaniu. Generuj $X = 1, 2, ..., n$.
2. Generuj Y_i dla $i = 1, 2, ..., n$.
3. Wyznacz $\hat{\beta}_0, \hat{\beta}_1$, a, b, c, d.
4. Sprawdź czy $\beta_0 \in [a, b]$ i czy $\beta_1 \in [c, d]$.
5. Powtórz krok 2-4 MC=1000 razy. Wyznacz $\hat{PU}_{\beta_0}$ oraz $\hat{PU}_{\beta_1}$.

Zadanie 1

Oczkiwany wynik



Rysunek: Oczekiwany wynik dla β_0

Zadanie 2

Wykonaj zad. 1 przy założeniu, że σ nie jest znane. Jakie są różnice pomiędzy skonstruowanymi przedziałami ufności uzyskanymi w zad.1 i zad. 2? Wyniki porównaj w zależności od długości próby, wielkości α oraz σ. Jakie możesz wyciągnąć wnioski na podstawie uzyskanych wyników.

1. **Konstrukcja przedziału ufności dla β_0:**

$$P\Big(\underbrace{\hat{\beta}_0 - t_{n-2,1-\alpha/2} \cdot S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}}_{a} \leqslant \beta_0 \leqslant$$

$$\leqslant \underbrace{\hat{\beta}_0 + t_{n-2,1-\alpha/2} \cdot S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}}_{b} \Big) = 1 - \alpha$$

$t_{n-2,1-\alpha/2}$ - kwantyl na poziomie $1 - \alpha/2$ z rozkładu t-Studenta o n-2 stopniach swobody.

$$S^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

2. **Konstrukcja przedziału ufności dla β_1:**

$$P\Big(\underbrace{\hat{\beta}_1 - t_{n-2,1-\alpha/2} \cdot S\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}}_{c} \leqslant \beta_1 \leqslant$$

$$\leqslant \underbrace{\hat{\beta}_1 + t_{n-2,1-\alpha/2} \cdot S\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}}_{d} \Big) = 1 - \alpha$$

$t_{n-2,1-\alpha/2}$ - kwantyl na poziomie $1 - \alpha/2$ z rozkładu t-Studenta o n-2 stopniach swobody.

$$S^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

Zadanie 3

Wysymuluj dwuwymiarowy wektor (x, y) opisany ogólnym modelem regresji liniowej:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

gdzie $\varepsilon_i \sim N(0, \sigma)$, oraz $\{\varepsilon_i\}$ i.i.d. Wybierz dowolne wartości β_0, β_1 oraz σ. Niech $x_1, x_2, ..., x_n$ będą zdefiniowane tak jak w zad. 1. Wyznacz przedziały ufności dla wartości średniej zmiennej $Y(x_0)$ dla $x_0 = x + \gamma$ dla pewnej wielkości γ dla różnych wielkości n przy założeniu, że:

a. σ jest wielkością znaną,

b. σ jest wielkością nieznaną.

Wyniki przedstaw w zależności od n, σ oraz γ. Przyjmij $\alpha = 0.05$.

1. **Konstrukcja przedziału ufności w zależności od długości wektora danych (σ znana):**
   W pierwszej części zadania zajmiemy się wyznaczaniem przedziału ufności dla $Y(x_0)$ na podstawie wyprowdzonej na wykładzie zależności. Zakładając, że parametr σ jest znany, przedział ufności ma postać:

$$P\Big(\underbrace{\hat{\mu}_{Y(x_0)} - z_{1-\alpha/2} \cdot \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}}_{a} \leqslant \mu_{Y(x_0)} \leqslant$$

$$\leqslant \underbrace{\hat{\mu}_{Y(x_0)} + z_{1-\alpha/2} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}}_{b} \Big) = 1 - \alpha$$

$z_{1-\alpha/2}$ - kwantyl na poziomie $1 - \alpha/2$ ze standardowego rozkładu normalnego.

2. **Konstrukcja przedziału ufności w zależności od długości wektora danych (σ nieznana):**
   Zakładając, że parametr σ jest nieznany, przedział ufności ma postać:

$$P\Big(\underbrace{\hat{\mu}_{Y(x_0)} - t_{1-\alpha/2,n-2} \cdot S\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}_{a} \leqslant \mu_{Y(x_0)} \leqslant$$

$$\leqslant \underbrace{\hat{\mu}_{Y(x_0)} + t_{1-\alpha/2,n-2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}_{b} \Big) = 1-\alpha,$$

gdzie $S^2 = \frac{1}{n-2}\sum(Y_i - \hat{Y}_i)^2$, $t_{1-\alpha/2,n-2}$ - kwantyl na poziomie $1-\alpha/2$ rozkładu t-Studenta z $n-2$ stopniami swobody.

3. **Estymacja parametru $\hat{P}U = 1 - \alpha$.**
   Równanie z poprzedniego slajdu mówi nam, że prawdopodobieństwo tego, że estymowany parametr znajdzie się w skonstruowanym przedziale ufności $[a, b]$ wynosi $1 - \alpha$. Zatem możemy obliczyć estymator parametru $1 - \alpha$, który ma następującą postać:

$$\hat{P}U = \frac{\#\mu_{Y(x_0)} \in (a, b)}{M},$$

gdzie M to liczba powtórzeń Monte Carlo.

$$x$$

Zadanie 3

Algorytm

Niech $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, gdzie $\varepsilon_i \sim N(0, \sigma)$, oraz $\{\varepsilon_i\}$ i.i.d.

$E[Y(x_0)] = \mu_{Y(x_0)} = E[\beta_0 + \beta_1 x_0 + \varepsilon_0] = \beta_0 + \beta_1 x_0$.
Estymator: $\hat{\mu}_{Y(x_0)} = \beta_0 + \beta_1 x_0$.

1. Ustal $\alpha = 0.05, \sigma = 1, \beta_0 = 2, \beta_1 = 4, x = linspace(0, 10, n)$, $\gamma = 0.5$.
2. Generuj $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, dla $i = 1, ..., n$.
3. Wyznacz $\hat{\beta}_0$ i $\hat{\beta}_1$, *a* oraz *b*.
4. Powtarzaj kroki 1-3 dla $n = 100 : 100 : 2000$.

Zadanie 3

Oczekiwane wyniki

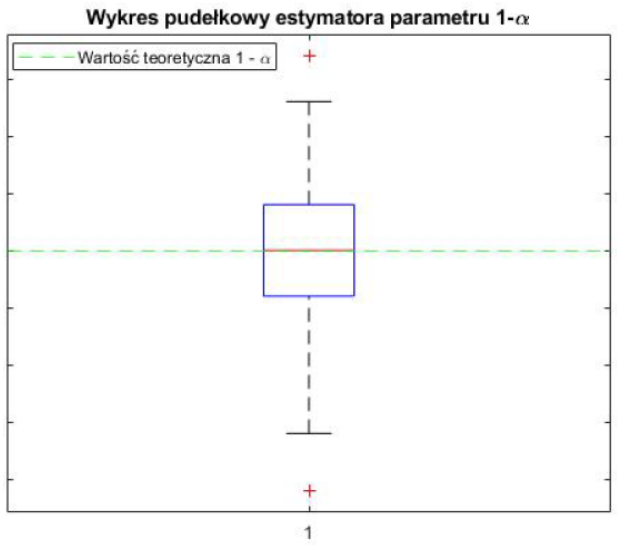
Ustal $\beta_0 = 2, \beta_1 = 4, \sigma = 1, \gamma = 0.5, \alpha = 0.05, x = \mathit{linspace}(0, 10, n)$.

1) wynik dla pojedynczej symulacji MC

n	$\mu_{Y(x_0)}$	a	b	Czy $\mu_{Y(x_0)} \in [a, b]$ (jeśli 1 to tak, jeśli 0 to nie)
100				
200				
300				
400				
500				
600				
700				
800				
900				
1000				
1100				
1200				
1300				
1400				
1500				
1600				
1700				
1800				
1900				
2000				

Tabela 1: Zależność przedziału ufności od n.

2) Wynik dla wielu powtórzeń MC



rys. 1: Wykres pudełkowy $\hat{PU}$ gdy teoretyczne $\alpha = 0.05$.

Zadanie 4

Wysymuluj dwuwymiarowy wektor (x, y) o długości n = 1000 opisany ogólnym modelem regresji liniowej

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

dla wybranych wielkosci parametrów β_0, β_1 oraz σ oraz $x_1, x_2, ..., x_n$ zdefiniowanych jak zad.1. Skonstruuj prostą regresji na podstawie 990 najmniejszych obserwacji wielkości x. Skonstruuj przedział ufności dla prognozy w modelu dla ostatnich 10 największych obserwacji i porównaj z danymi. Zadanie wykonaj przy założeniu, że σ jest znana i nieznana.

Zadanie 4

Podstawy teoretyczne

Z treści zadania: $Y(x_0) = \beta_0 + \beta_1 x_0 + \varepsilon_0$. Zgodnie z przyjętą notacją: $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

1. **Konstrukcja przedziału ufności w zależności od długości wektora danych (σ znana):**
   Zakładając, że parametr σ jest znany, przedział ufności ma postać:

$$P\Big(\underbrace{\hat{Y}(x_0) - z_{1-\alpha/2} \cdot \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}}_{a} \leqslant Y(x_0) \leqslant$$

$$\leqslant \underbrace{\hat{Y}(x_0) + z_{1-\alpha/2} \cdot \sigma \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}}_{b} \Big) = 1 - \alpha$$

$z_{1-\alpha/2}$ - kwantyl na poziomie $1 - \alpha/2$ ze standardowego rozkładu normalnego.

2. **Konstrukcja przedziału ufności w zależności od długości wektora danych (σ nieznana):**
   Zakładając, że parametr σ jest nieznany, przedział ufności ma postać:

$$P\Big(\underbrace{\hat{Y}(x_0) - t_{1-\alpha/2,n-2} \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}}_{a} \leqslant Y(x_0) \leqslant$$

$$\leqslant \underbrace{\hat{Y}(x_0) + t_{1-\alpha/2,n-2} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}}_{b} \Big) = 1 - \alpha$$

gdzie $S^2 = \frac{1}{n-2}\sum(Y_i - \hat{Y}_i)^2$, $t_{1-\alpha/2,n-2}$ - kwantyl na poziomie $1 - \alpha/2$ rozkładu t-Studenta z $n - 2$ stopniami swobody.

Zadanie 4

Oczekiwane wyniki

Ustal $\beta_0 = 2, \beta_1 = 4, \sigma = 2, \alpha = 0.05, x = \mathit{linspace}(0, 10, 1000)$. Wykonaj symulacje, które umożliwią stworzenie poniższego wykresu: