Modele liniowe, Lista 4

Dominika Ochalik

2024-01-05

SPIS TREŚCI

Wstęp

Lista 4 rozpoczyna analizę regresji liniowej z więcej niż jedną zmienną objaśniającą, tzn. modele postaci:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon$$

Jest to przydatne uogólnienie dotychczas rozważanego modelu regresji liniowej, ponieważ często badana wielkość Y zależy od więcej niż jednego czynnika.

Zadania na liście skupiają się także wokół analizy, czy dodanie do równania kolejnej zmiennej objaśniającej sprawi, że model staje się lepszy.

Podstawowe elementy teorii

Macierzowa postać równania regresji

Rozważamy równanie:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_n X_{in} + \epsilon_i$$

Można je zapisać w postaci macierzowej:

$$Y = \mathbf{X}\beta + \epsilon$$

gdzie:

- $Y = (Y_1, Y_2, ...Y_n)^T$,
- $\beta = (\beta_0, \beta_1, ..., \beta_n)^T$,
- $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$,

oraz:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & ... & X_{1n} \\ 1 & X_{21} & ... & X_{2n} \\ ... & ... & ... & ... \\ 1 & X_{n1} & ... & X_{nn} \end{pmatrix}$$

Przedział ufności dla kombinacji liniowej $c'\beta$.

Odchylenie standardowe $c'\hat{\beta}$ obliczymy korzystając ze wzoru:

$$c'\hat{\beta} \pm t_c s(c'\hat{\beta})$$

gdzie:

- $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_n)^T$,
- c to wektor liczb,
- t_c, czyli kwantyl rzędu $1 - \frac{\alpha}{2}$ z rozkładu studenta z n-p stopniami swobody,
- $s(c'\hat{\beta})$ to odchylenie standardowe, które obliczamy ze wzoru:

$$s^2(c'\hat{\beta}) = s^2 c'(\mathbf{X}'\mathbf{X})^{-1} c$$

Testowanie dotyczące wektora parametrów β.

Chcemy testować istotność dowolnej kombinacji liniowej elementów wektora β.
Mamy wektor c = $(c_0, ..., c_{p-1})' \in R^p$ oraz stałą d $\in R$. Testujemy hipotezę zerową postaci:

$$H_0 : c'\beta - d = 0$$

Przeciwko alternatywie:

$$H_1 : c'\beta - d \neq 0$$

Statystyka testowa ma postać:

$$T = \frac{c'\hat{\beta} - d}{s(c'\hat{\beta})}$$

gdzie $s^2(c'\hat{\beta}) = s^2 c'(\mathbf{X}'\mathbf{X})^{-1}c$.

Przy założeniu prawdziwości H_0, statystyka T ma rozkład studenta z n-p stopniami swobody. Zatem będziemy odrzucać H_0 na poziomie istotności α, gdy |T| > t_c, gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n - p)$, czyli jest kwantylem rzędu 1 - $\frac{\alpha}{2}$ z rozkładu studenta z n-p stopniami swobody.

Kryterium AIC

Kryterium AIC bada dopasowanie modelu do danych wraz z uwzględnieniem liczby zmiennych objaśniających użytych w modelu (im więcej, tym gorzej, ponieważ chcemy, aby model był możliwie jak najprostszy). Im mniejszą wartość przyjmuje statystyka AIC, tym model jest lepszy.

$$AIC = nlog\left(\frac{SSE}{n}\right) + 2k$$

gdzie:

- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$,
- k to liczba zmiennych objaśniających użytych w modelu,
- n to wymiar zmiennej Y.

Zadanie 1: Wpływ korelacji

a) Wygenerowanie macierzy X i wektora Y

W podpunkcie a) jesteśmy proszeni o wygenerowanie macierzy $\mathbf{X}_{100\times 2}$ takiej, że jej wiersze są niezależnymi wektorami losowymi z rozkładu N(0, $\frac{\Sigma}{100}$), gdzie

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

Następnie należy wygenerować wektor $Y = \beta_1 X_1 + \epsilon$, gdzie:

- $\beta_1 = 3$,
- X_1 to pierwsza kolumna macierzy $\mathbf{X}$,
- $\epsilon \sim$ N(0, I).

b) t-test i 95% PU dla β_1

W tym podpunkcie należy przeprowadzić t-test na poziomie istotności $\alpha = 0.05$ oraz wyznaczyć 95% przedział ufności dla β_1 w dwóch przypadkach:

- przy użyciu modelu prostej regresji liniowej: $Y = \beta_0 + \beta_1 X_1 + \epsilon$,
- używając modelu z dwiema zmiennymi objaśnianymi: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, gdzie X_2 to druga kolumna macierzy $\mathbf{X}$ wygenerowanej w poprzednim podpunkcie.

Model $Y = \beta_0 + \beta_1 X_1 + \epsilon$.

95% Przedział ufności dla β_1 jest postaci: [1.7305; 4.8416].

Statystyka testowa T ma wartość 4.1922, natomiast kwantyl rzędu 0.975 z 98 stopniami swobody ma wartość 1.9845. Widzimy, że T> t_c, zatem odrzucamy H_0.

Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

W tym modelu, 95% przedział ufności dla β_1 jest postaci: [1.2217; 8.4243].

Wartość statystyki T ma wartość 2.658, natomiast kwantyl $t_c = 1.9847$. Widzimy, że statystyka T przyjmuje mniejszą wartość niż w przypadku wyżej rozważanego modelu, jednak nie zmienia to faktu, że i tak odrzucamy H_0, ponieważ T > t_c.

Porównajmy wyniki dla obu modeli w tabeli:

Model	$Y = \beta_0 + \beta_1 X_1 + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
$\hat{\beta}_1$	3.286	4.823
95% PU	[1.7305; 4.8416]	[1.2217; 8.4243]
Długość PU	3.1111	7.2026
Wartość statystyki T	4.1922	2.658
Wartość kwantyla t_c	1.9845	1.9847
Czy odrzucamy H_0?	TAK	TAK

W obu przypadkach z 95% prawdopodobieństwem odrzucamy H_0 mówiącą, że $\beta_1 = 0$. Dla modelu z dwoma zmiennymi objaśniającymi widzimy, że przedział ufności dla β_1 osiąga większą długość niż dla modelu z jedną zmienną, a także wartość $\hat{\beta}_1$ jest większa.

c) Odchylenie standardowe β_1 i moc identyfikacji X_1

Odchylenie standardowe β_1 obliczymy korzystając ze wzoru:

$$s^2(c'\hat{\beta}) = s^2 c'(\mathbf{X}'\mathbf{X})^{-1}c$$

dla wektora $c' = (0,1)$ w przypadku pierwszego modelu lub $c' = (0,1,0)$ w przypadku drugiego modelu.

Z racji, że interesuje nas teoretyczna wartość odchylenia standardowego, we wzorze zamiast s^2 użyjemy znane nam $\sigma^2 = 1$.

Moc identyfikacji X_1 to prawdopodobieństwo odrzucenia H_0 mówiącej, że $\beta_1 = 0$ wiedząc, że prawdziwa jest hipoteza alternatywna $H_1 : \beta_1 = 3$.

Statystyka testowa T jest postaci:

$$T = \frac{c'\hat{\beta} - d}{s(c'\hat{\beta})}$$

dla d = 0 oraz tak dobranego wektora c, aby $c'\hat{\beta} = \hat{\beta}_1$. Przy założeniu, że $\beta_1 = 3$, statystyka T ma niecentralny rozkład studenta z parametrem niecentralności $\delta = \frac{\beta_1}{\sigma(\hat{\beta}_1)}$.

$\sigma(\hat{\beta}_1)$ liczymy ze wzoru:

$$\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Moc testu to $P_{\beta_1=3}(|T| > t_c)$, gdzie t_c to odpowiedni kwantyl.

Model	$Y = \beta_0 + \beta_1 X_1 + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
$s(\hat{\beta}_1)$	0.9335	2.1596
Moc identyfikacji X_1	0.8892	0.2798

Odchylenie standardowe estymatora β_1 jest większe dla modelu z dwoma zmiennymi. Jest to spójne z wnioskiem z poprzedniego podpunktu mówiącym, że przedział ufności dla β_1 w tym modelu jest większy. Moc identyfikacji dla modelu z jedną zmienną objasniającą jest wysoka, natomiast dla drugiego modelu znacznie mniejsza. Może wynikać to z faktu, że zmienne X_1 i X_2 są skorelowane, zatem można podejrzewać, że da się przybliżyć model $Y = \beta_0 + \beta_1 X_1 + \epsilon$ zastępując X_1 zmienną X_2, a także zmieniając wartości parametrów β_0 i β_1.

d) Estymacja wartości parametrów z poprzedniego podpunktu.

Estymowaną wartość mocy obliczamy poprzez zsumowanie liczby zdarzeń, w których odrzucamy H_0, a następnie podzielenie wyniku przez liczbę wszystkich doświadczeń, czyli 1000.

Odchylenie standardowe estymujemy poprzez obliczenie odchylenia standarowego dla każdego z 1000 wygenerowanych zbiorów danych, a następnie obliczenie ich średniej arytmetycznej.

Zobaczmy wyniki dla modelu $Y = \beta_0 + \beta_1 X_1 + \epsilon$ w tabeli:

Badana wielkość	Teoretyczna wartość	Estymowana wartość
$\hat{\beta}_1$	3	3.0752
$s(\hat{\beta}_1)$	0.9335	0.9329
Moc	0.8892	0.918

Wyniki dla modelu $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$:

Badana wielkość	Teoretyczna wartość	Estymowana wartość
$\hat{\beta}_1$	3	3.1344
$s(\hat{\beta}_1)$	2.1596	2.1579
Moc	0.2798	0.307

Dla obu modeli możemy wyciągnąć wniosek, że estymowane wartości dobrze przybliżają teoretyczne.

Zadanie 2: wpływ wymiaru.

a) Wygenerowanie macierzy X i wektora Y

W tym zadaniu należy wygenerować macierz $X_{1000\times 950}$ tak, że jej elementy są niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma = 0.1)$. Wektor zmiennej odpowiedzi wyraża się wzorem:

$$Y = X\beta + \epsilon$$

gdzie $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$.

b) Analiza różnych postaci modelu i wybranie najlepszego z użyciem kryterium AIC.

Model jest zbudowany przy użyciu pierwszych k kolumn macierzy dla $k \in \{1, 2, 5, 10, 50, 100, 500, 950\}$. Dla każdego modelu obliczamy następujące wielkości:

- SSE,
- MSE,
- AIC,
- p-wartości dla dwóch pierwszych zmiennych objaśniających,
- liczbę fałszywych odkryć.

Poprzez fałszywe odkrycie rozumiemy odrzucenie H_0, gdy jest prawdziwa, albo nieodrzucenie H_0, gdy jest fałszywa.

k	SSE	MSE	AIC	p_1	p_2	Fałszywe odkrycia
1	393.7039	0.3941	-930.1562	2.8179×10^{-50}	NA	0
2	280.0688	0.2806	-1268.72	4.6826×10^{-63}	7.3526×10^{-76}	0
5	10.8927	0.0109	-4509.6633	0	0	0
10	10.8177	0.0109	-4506.5724	0	0	0
50	10.3135	0.0109	-4474.299	0	0	4
100	9.916	0.011	-4413.6019	0	0	5
500	5.4585	0.0109	-4210.5758	1.923×10^{-237}	1.282×10^{-259}	21
950	0.6453	0.0129	-5445.8159	1.9672×10^{-22}	1.1213×10^{-27}	35

Widzimy, że wraz ze wzrostem ilości zmiennych (wartości k), wartość SSE oraz MSE maleje. Bardzo dużą różnicę widać między k=2 a k=5. P-wartości dla każdej wartości k są bardzo małe i pozwalają na odrzucenie hipotez mówiących, że $\beta_1 = 0$ oraz $\beta_2 = 0$. Liczba fałszywych odkryć rośnie od k=50. Dla mniejszych k jest równa 0.

Na podstawie kryterium AIC należy wybrać model, dla którego wartość AIC jest najmniejsza. Analizując powyższą tabelę możemy zauważyć, że minimalną wartość AIC równą -5445.8159 osiąga model z liczbą zmiennych równą k = 950.

d) Powtórzenie zadania 2b z dodatkiem mocy identyfikacji X_1 i X_2.

W tym podpunkcie należy powtórzyć polecenia z zadania 2b 1000 razy. Wyniki ponownie zostaną przedstawione w tabeli, jednak każda wartość będzie średnią arytmetyczną poszczególnych wyników (np. wartość AIC dla k=1 będzie obliczana 1000 razy, zatem w tabeli znajdzie się średnia arytmetyczna otrzymanych wyników).

k	SSE	MSE	AIC	p_1	p_2	Fałszywe odkrycia
1	391.4062	0.3918	-936.0627	6.73802×10^{-49}	NA	0
2	281.0235	0.2816	-1265.3906	1.7347×10^{-60}	4.7358×10^{-72}	0
5	9.952	0.01	-4600.9123	0	0	0
10	9.8994	0.01	-4596.2152	0	0	0.292
50	9.4985	0.01	-4557.5891	0	0	2.314
100	9.0024	0.01	-4511.2881	0	0	4.72
500	5.0009	0.01	-4299.988	4.5242×10^{-232}	1.1963×10^{-243}	24.518
950	0.503	0.0101	-5717.5273	3.7042×10^{-22}	3.6641×10^{-25}	47.424

Z powyższej tabeli możemy wyciągnąć podobne wnioski, co dla tabeli z zadania 2b. Należy zwrócić uwagę, że już dla k=10 pojawiają się pojedyncze fałszywe odkrycia. Wartość SSE również maleje wraz ze wzrostem wartości k. Wartość MSE znacznie maleje między k=2 a k=5, po czym utrzymuje się na stałym poziomie. Jedynie dla k=950 jest ciut większa, jednak różnica pojawia się dopiero na 4. miejscu po przecinku, zatem decydujemy się na pominięcie tego wniosku.

Średni rozmiar modelu wybranego przez AIC to 950. W zadaniu 2b było to 950. Widzimy, że wynik jest taki sam.

W tym zadaniu należy jeszcze obliczyć moc identyfikacji X_1 i X_2, którą rozumiemy jako prawdopodobieństwo odrzucenia $H_0 : \beta_1 = 0$, pod warunkiem, że $\beta_1 = 3$. Analogicznie dla X_2.

Patrząc na to, jak bardzo małe są p-wartości w powyższej tabeli, można przypuszczać, że moce identyfikacji będą równe 1 lub prawie równe 1.

Liczba kolumn	Moc identyfikacji X_1	Moc identyfikacji X_2
1	1	NA
2	1	1
5	1	1
10	1	1
50	1	1
100	1	1
500	1	1
950	1	1

Wyniki nas nie zaskoczyły: moce identyfikacji są równe 1.