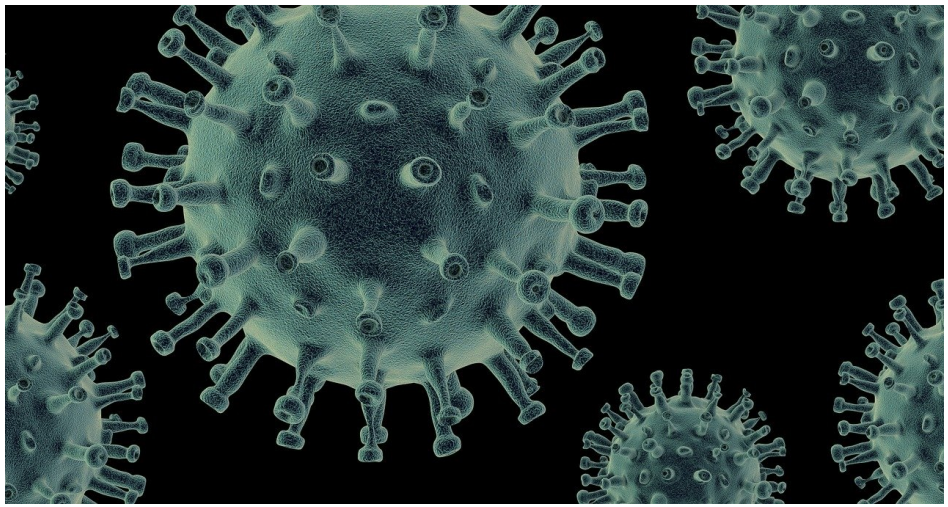
Model regresji wielorakiej

29 marca 2021



Model regresji wielorakiej-powtórzenie

Nadal rozważamy układ równań:

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\epsilon}.$$

Równoważnie w formie macierzowej:

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

Założenia (slajd 13, wykład 22.03.2021

Zakładamy, że dysponujemy większą liczbą obserwacji niż parametrów $k+1 \leq n$. Ponadto,

- Z1 Macierz $\mathbf{X}$ macierzy objaśniających jest deterministyczna (nielosowa), tzn. podane liczby $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ są nielosowe t;
- Z2 Rząd macierzy $\mathbf{X}$ jest $k+1$, innymi słowy wektory kolumnowe są liniowo niezależne;
- Z3 $E(\epsilon) = \mathbf{0}$, tzn. $E\epsilon_t = 0$ dla wszystkich t;
- Z4 $D^2(\epsilon) = \sigma^2\mathbf{I}$, tzn. ϵ_t jest ciągiem nieskorelowanych zmiennych losowych o tej samej wariancji σ^2, (σ jest nieznana);
- Z5 Wszystkie ϵ_t mają rozkład normalny $N(0, \sigma^2)$ (dodatkowo z Z4 są niezależne).

Estymatory

Estymator β może być wyrażony w formie macierzowej

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Estymator $\hat{\beta}$ jest zmienną losową o macierzy kowariancji (slajd 24, wykład 22.03.2021):

$$D^2(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Nieobciążony estymator σ^2 jest zdefiniowany jako

$$\hat{\sigma}^2 := \frac{1}{n-k-1}\sum_{t=1}^{n}\left(Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_{t1} - \hat{\beta}_2 X_{t2} - \ldots - \hat{\beta}_k X_{tk}\right)^2.$$

Zatem naturalny estymator $D^2(\hat{\beta})$ to

$$\hat{D}^2(\hat{\beta}) := \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Macierz kowariancji i jej estymacja

Zauważmy:

- $D^2(\hat{\beta}) = [d_{ij}]$ jest macierzą rozmiaru $k \times k$ którego komórka (i, j) zawiera
$$d_{ij} := Cov(\hat{\beta}_i, \hat{\beta}_j);$$
- Wtedy komórka na przekątnej (i, i) zawiera
$$d_{ii} = Cov(\hat{\beta}_i, \hat{\beta}_i) = Var(\hat{\beta}_i) = E(\hat{\beta}_i - \beta_i)^2.$$
- Estymator macierzy kowariancji wyrażamy jako
$$\hat{D}^2(\hat{\beta}) := [\hat{d}_{ij}],$$
$\hat{d}_{ij}$ jest (i, j) komórką macierzy $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$;
- Stąd
$$\hat{d}_{ii} \approx d_{ii} = E(\hat{\beta}_i - \beta_i)^2.$$

Konstruujemy miernik *błędu względnego* estymatora β_i jako:

$$T_i(\beta_i) = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{d}_{ii}}}.$$

Liczy różnicę między estymatorem $\hat{\beta}_i$, a prawdziwą wartością β_i w stosunku do *błędu standardowego* estymatora $\hat{\beta}_i$. Dla $\beta_i = 0$ oznaczmy po prostu T_i

$$T_i := T_i(0) = \frac{\hat{\beta}_i}{\sqrt{\hat{d}_{ii}}}.$$

Istotność zmiennych obajśniających

Definition 1 (Istotność zmiennych obajśniających)

Mówimy, że zmienne objaśniające $(X_{t,i})_{t=1}^{n}$ *w modelu*

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

jest ***istotna*** *jeśli* $\beta_i \neq 0$. *Jeśli* $\beta_i = 0$ *wtedy mówimy, że zmienna* $X_{t,i}$ *nie jest istotna lub* ***nieistotna***.

Istotność zmiennych obajśniających (**test Studenta**)

Test Studenta

Testujemy hipotezę $(X_{t,i})_t$ $(i = 1, 2, \ldots, k)$ jest nieistotna przeciwko hipotezie, że jest istotna. Formalnie:

$$H_0 : \beta_i = 0 \quad \text{VS} \quad H_1 : \beta_i \neq 0.$$

Statystyka testowa jest (slajd 6)

$$T_i = \frac{\hat{\beta}_i}{\sqrt{\hat{d}_{ii}}}.$$

Jeśli hipoteza H_0 jest prawdziwa, T_i ma rozkład **t-Studenta** z $n - k - 1$ **stopniami swobody** $(\mathcal{T}(n - k - 1))$. **Domyślny poziom istotności** $\kappa = 0.05$.

Test Studenta - obszar odrzucenia

Testujemy hipotezę H_0 o **poziomie istotności** $\kappa \in (0,1)$.

- **Pierwszy sposób.** Jeśli poniższa nierówność jest spełniona,

$$|T_i| > q\left(1-\frac{\kappa}{2}, n-k-1\right) \quad \text{- poziom krytyczny,}$$

gdzie $q(\cdot, n-k-1)$ jest kwantylem dystrybuanty rozkładu *t-Studenta* $\mathcal{T}(n-k-1)$, wtedy **odrzucamy** H_0 **na rzecz hipotezy alternatywnej** H_1. W przeciwnym razie przyjmujemy hipotezę H_0 (właściwie to stwierdzamy tylko brak podstaw do odrzucenia hipotezy H_0).

- **Drugi sposób.** Obliczamy **p-value** (lub p-wartość):

$$\text{p-value} = 2(1 - CDF_{n-k-1}(|T_i|)),$$

gdzie $CDF(\cdot, n-k-1)$ jest **dystrybuantą** t-Studenta $\mathcal{T}(n-k-1)$, oraz

  - jeśli p-value $\geq \kappa$ wtedy **odrzucamy hipotezę** H_0 **na rzecz hipotezy alternatywnej** H_1.
  - w przeciwnym razie odrzucamy H_0.

Remark 1

Drugi sposób jest bardziej praktyczny niż ten pierwszy:

- *p-value jest najmniejszym poziomem istotności prowadzącym do odrzucenia H_0:*
  - *Na przykład, jeśli p-value jest* 0.08 *wtedy odrzucamy hipotezę H_0 dla każdego poziomu istotności większego niż* 0.08*, a przyjmujemy hipotezę dla niższego poziomu niż* 0.08*;*
  - *Można powiedzieć, że licząc p-value możemy powiedzieć, że testujemy hipotezę H_0 na wszystkich poziomach jednocześnie (mamy wskazówkę postępowania w zależności od naszej tolerancji na błąd I rodzaju);*
- *Pierwszy sposób wymaga obliczenia poziomu krytycznego w zależności od przyjętego istotności (tolerancji na błąd I rodzaju):*
  - *Przykładowo, testując na poziomie $\kappa = 0.05$ znajdujemy poziom krytyczny. Gdy chcemy zmieni' $\kappa = 0.1$ musimy obliczyć nowy poziom krytyczny.*

Remark 2

Akceptacja hipotez $\beta_1 = 0$, ..., *oraz* $\beta_k = 0$ *w tym samym czasie* ***nie jest tożsama*** *z akceptacją hipotezy*

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad VS \quad \exists_{i=1,2,\ldots,k} \beta_k \neq 0.$$

Podobnie sytuacja jest w przypadku dowolnego zbioru zmiennych.

- *Oddzielna akceptacja hipotez* $\beta_1 = 0$, ..., *oraz* $\beta_k = 0$ *oznacza, że każda ze zmiennych jest nieistotna w modelu i możemy usunąć z modelu jedną z nich, bo być może jej rolę mogą przejąć pozostałe zmienne;*
- *Akceptacja łącznej hipotezy* $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ *oznacza, że cały zestaw zmiennych nie odgrywa żadnej roli w modelu i możemy od razu usunąć wszystkie zmienne.*

Remark 3

Akceptacja hipotez kolejnych hipotez $\beta_1 = 0$, $\beta_2 = 0, \ldots$ motywuje nas do stopniowego usuwania zmiennych z modelu. Przykładowo na początek usuwamy zmienną z największą p-value i obliczamy nowy model bez tej zmiennej.

Przykład -ceny mieszkań w Polsce

Przykład

Example 1 (Ceny mieszkań w miastach Polski)

Ceny mieszkań w miastach wojewódzkich w Polsce są opisane w następnej tabelce. Za pomocą danych dopasowujemy następujący model:

$$P_t = \beta_0 + \beta_1 F_t + \beta_2 D_t + \beta_3 S_3 + \epsilon_t,$$

gdzie

- P_t*-średnia cena na* m^2 *wyrażona w PLN w mieście* t*;*
- F_t*-liczba mieszkańców w mieście* t*;*
- D_t *- gęstość zaludnienia w mieście* t*;*
- S_t *- średnia powierzchnia mieszkania w mieście* t*.*

Odpowiemy na pytanie czy cena mieszkania w polskim mieście zależy od liczby mieszkańców, gęstości zaludnienia lub średniej powierzchni mieszkań w mieście.

Przykład i dane

Miasto	P_t	F_t	D_t	S_t
Białystok	3933	295 459	2893	102,13
Bydgoszcz	3692	357 652	2032	175,98
Gdańsk	5552	461 489	1762	1762
Gdynia	5037	247 820	1834	1834
Katowice	3636	301 834	1833	1833
Kielce	3617	198 857	1814	1814
Kraków	5820	761 873	2331	2331
Lublin	4571	341 722	2317	2317
Łódź	3284	706 004	2408	2408
Olsztyn	4181	173 831	1968	1968
Opole	3910	119 574	1238	1238
Poznań	5157	545 680	2083	2083
Rzeszów	4514	185 123	1591	116,36
Szczecin	4065	407 180	1355	300,55
Warszawa	7300	1 735 442	3355	517,24
Wrocław	5332	634 487	2167	292,82
Zielona Góra	3267	138 512	498	278,32

Rozwiązanie

Za pomocą obliczeń w programie Gretl mamy podane rozwiązania:

Zmienna	$\hat{\beta}_i$	$\sqrt{\hat{d}_{ii}}$	T_i	p-value
constant	3473.48	690.486	5.030	0.0002
F_t	0.00207999	0.000705706	2.947	0.0113
D_t	0.0209891	0.427145	0.04914	0.9616
S_t	0.0599937	0.215717	0.2781	0.7853

Rozwiązanie c.d.

Mamy więc przybliżony model:

$$P_t \approx 3473.48 + 0.00207999 * F_t + 0.0209891 * D_t + 0.0599937 * S_t + \epsilon_t.$$

Testujemy istotność współczynników $\beta_0, \beta_1, \beta_2$ and β_3 (i odpowiednich zmiennych) za pomocą wyników z poprzedniej tabelki przy poziomie istotności $\kappa = 0.05$:

Hipoteza	p-value	decyzja	czy istotna
$H_0 : \beta_0 = 0$	0.0002	odrzucić H_0	tak
$H_0 : \beta_1 = 0$	0.0113	odrzucić H_0	tak
$H_0 : \beta_2 = 0$	0.9616	przyjąć H_0	nie
$H_0 : \beta_3 = 0$	0.7853	przyjąć H_0	nie

Remark 4

Akceptacja hipotez $H_0 : \beta_2 = 0$ *oraz* $H_0 : \beta_3 = 0$ *nie jest identyczna z akceptacją hipotezy*

$$H_0 : \beta_2 = \beta_3 = 0 \quad VS \quad H_1 : \beta_2 \neq 0 \ lub \ \beta_3 \neq 0.$$

Rozwiązanie - C.D.

Innymi słowy

- realizacja $\hat{\beta}_2$ jest tak mała, że możemy rozważyć nowy model bez D_t:
$$P_t = \beta_0 + \beta_1 F_t + \beta_2 S_t + \epsilon_t;$$
- realizacja $\hat{\beta}_3$ jest tak mała, że możemy rozważyć nowy model bez S_t,
$$P_t = \beta_0 + \beta_1 F_t + \beta_2 D_t + \epsilon_t;$$
- p-value dla β_2 (dla hipotezy $H_0 : \beta_2 = 0$) jest 0.9616, większa niż 0.7853 - p-value dla β_3, stąd usuwamy D_t;
- Obliczamy nowy model
$$P_t = \beta_0 + \beta_1 F_t + \beta_2 S_t + \epsilon_t.$$

Rozwiązanie - c.d.

Dalsze obliczenia w Gretlu:

Zmienna	$\hat{\beta}_i$	$\sqrt{\hat{d}_{ii}}$	T_i	p-value
constant	3501.22	383.041	9.141	$2.81 * 10^{-7}$
F_t	0.00210458	0.000479542	4.389	0.0006
S_t	0.0620457	0.203957	0.3042	0.7654

Rozwiązanie - c.d.

Zatem model jest w przybliżeniu:

$$P_t \approx 3501.22 + 0.00210458 * F_t + 0.0620457 * S_t + \epsilon_t.$$

Testujemy istotność β_0, β_1 and β_2 wg. wyników z poprzedniej tabelki przy poziomie istotności $\kappa = 0.05$:

Hipoteza	p-value	decyzja	czy istotny
$H_0 : \beta_0 = 0$	$2.81 * 10^{-7}$	odrzucić H_0	tak
$H_0 : \beta_1 = 0$	0.0006	odrzucić H_0	tak
$H_0 : \beta_2 = 0$	0.7654	przyjąć H_0	nie

Jedynie współczynnik β_2 jest nieistotny. Rozważamy więc kolejny model

$$P_t = \beta_0 + \beta_1 F_t + \epsilon_t.$$

Rozwiązanie - c.d.

Kolejne obliczenia w Gretlu dają wynik:

Zmienna	$\hat{\beta}_i$	$\sqrt{\hat{d}_{ii}}$	T_i	p-value
constant	3580.80	271.226	13.20	$1.16 * 10^{-9}$
F_t	0.00210107	0.000464675	4.522	0.0004

Oba współczynniki są istotne, model końcowy

$$P_t = 3580.80 + 0.00210107 * F_t + \epsilon_t.$$

Cena mieszkania jest zdeterminowana przez wielkoćć populacji. Stwierdzamy brak zale:noći ceny z gęstością populacji miasta i średniej powierzchnii zajmowanych powierzchnii mieszkaniowych.