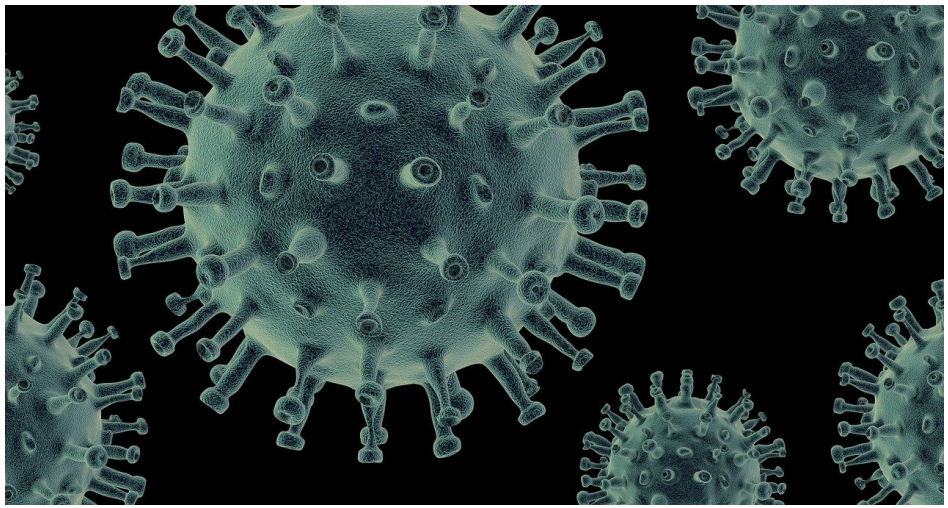
Weryfikacja założeń modelu, część 1

7 kwietnia 2021



Model regresji wielorakiej-przypomnienie

Rozważamy model liniowy w formie macierzowej:

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

Innymi słowy, model wyrażamy jako:

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

Założenia

Mamy więcej obserwacji niż parametrów, tzn. $k+1 \leq n$. Ponadto,

- Z1 Macierz $\mathbf{X}$ zmiennych objaśniających jest deterministyczna (nielosowa)), innymi słowy $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ nie są losowe dla wszystkich t;
- Z2 Rząd macierzy $\mathbf{X}$ jest $k+1$, innymi słowy kolumny 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 tych samej wariancji σ^2, (σ jest nieznane);
- Z5 Wszystkie zmienne ϵ_t mają rozkład normalny $N(0, \sigma^2)$.

Cel

Celem jest weryfikacja niektórych założeń:

Z1 Założenie jest niemożliwe do weryfikacji za pomocą narzędzi matematycznych (to inżynier musi być pewien czy ono pasuje do rzeczywistości);

Z2 Weryfikacja założenia to proste ćwiczenie z algebry;

Z3 Zakładając dodatkowo [Z4] i [Z5] możemy użyć standardowego testu t-Studenta.

Z4 Weryfikacja tego założenia prowadzi do sprawdzenia hipotez pomocniczych:

- ϵ_t nie posiada autokorelacji *test Durbina-Watsona*;
- ϵ_t jest ciągiem *homoskedastycznym*, tzn. $Var(\epsilon_t) = \sigma^2$ dla wszystkich t;

Poznamy te testy;

Z5 Poznamy tu kilka testów.

Macierz kowariancji i estymator

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

Estymator $\hat{\beta}$ jest zmienną losową której macierz kowariancji wyraża się wzorem (p. poprzednie wykłady):

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

Nieobciążony estymator σ^2 jest wyrażony wzorem

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

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 estymator

Zauważmy

- $D^2(\hat{\beta}) = [d_{ij}]$ jest macierzą $k \times k$, gdzie pozycja (i, j) zawiera $d_{ij} := Cov(\hat{\beta}_i, \hat{\beta}_j)$;
- Wtedy pozycja (i, i) zawiera

$$d_{ii} = Cov(\hat{\beta}_i, \hat{\beta}_i) = Var(\hat{\beta}_i) = E(\hat{\beta}_i - \beta_i)^2.$$

- Estymator $\hat{D}^2(\hat{\beta}) := [\hat{d}_{ij}]$ jest macierzą, gdzie $\hat{d}_{ij}$ wynosi $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$;
- Stąd

$$\hat{d}_{ii} \approx d_{ii} = E(\hat{\beta}_i - \beta_i)^2.$$

Test Studenta

Testujemy istotność zmiennej:

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

oddzielnie dla $i = 0, 1, 2, \ldots, k$. Statystyka testowa ma postać:

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

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

Interpretacja testu Studenta

Za pomocą testu t-Studenta testujemy $\beta_i = 0$ oddzielnie $i = 0, 1, 2, \ldots, k$:

- Przykładowo, przyjmujemy hipotezy $\beta_1 = 0$ oraz $\beta_2 = 0$, a odrzucamy pozostałe $\beta_0 = 0$, $\beta_3 = 0, \ldots, \beta_k = 0$:
  - znaczenie zmiennej $X_{t,1}$ w modelu jest niewielkie pod warunkiem, że zostawimy w modelu $X_{t,2}$ i vice versa;
  - jesteśmy uprawnieni do usunięcia $X_{t,1}$ **lub** $X_{t,2}$ **ale nie obu**;
- **Innymi słowy nie możemy usuwać wszystkich zmiennych nieistotnych na raz**.

Interpretacja testu Studenta

Ogólniej, jeśli przyjmiemy wszystkie hipotezy: $\beta_0 = 0$, $\beta_1 = 0$, ..., $\beta_k = 0$, jesteśmy uprawnieni:

- usunąć *stałą* z modelu pod warunkiem zatrzymania w modelu $X_{t,1}, X_{t,2}, ..., X_{t,k}$;
- usunąć $X_{t,1}$ z modelu pod warunkiem zatrzymania w modelu $X_{t,2}, X_{t,3}, ..., X_{t,k}$ oraz *stałą*;
- lub usunąć dowolną inną $X_{t,i}$ z modelu pod warunkiem zatrzymania w modelu $X_{t,1}, X_{t,2}, ..., X_{i-1}, X_{i+1}, X_{t,k}$ oraz *stałą*;

Nie możemy usuwać wszystkich nieistotnych zmiennych na raz .

Istotność każdej pojedynczej zmiennej, a istotność całej grupy zmiennych (**test Studenta**) -ilustracja

Tłum i rowerzysta:

- **Rowerzysta jest istotny:** fotografia bez **rowerzysta** byłaby inna niż z nim;
- **Każda pojedyncza osoba z tłumu jest nieistotna:** gdybyśmy usunęli jedną dowolną osobę z tłum, fotografia nadal przedstawiałaby ten sam tłum w mieście;

- **Ale cały tłum jest istotny:** fotografia bez tłumu przypominałaby wymarłe miasto.

Test Walda

- Rozważmy originaly model

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_k X_{t,k} + \epsilon_t;$$

- Dla pewnego ustalonego m takiego, że $1 \leq m \leq k$, rozważmy możliwość usunięcia wszystkich zmiennych $X_{t,m}, X_{t,m+1}, \ldots, X_{t,k}$ na raz:

$$H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0 \quad \text{VS} \quad H_1 : \exists_{m+1 \leq j \leq k} \beta_j \neq 0.$$

Test Walda

- Przyjęcie hipotezy H_0 upoważnia nas do przyjęcia **modelu obciętego**:

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_m X_{t,m} + \epsilon_t;$$

- W przypadku odrzucenia hipotezy H_0 pozostajemy przy **oryginalnym modelu**:

$$Y_t = \underbrace{\beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_m X_{t,m}}_{\textbf{część bazowa}} + \underbrace{+ \beta_{m+1} X_{t,m+1} + \beta_{m+2} X_{t,m+2} + \ldots + \beta_k X_{t,k}}_{\textbf{część rozszerzona}} + \epsilon_t.$$

Test Walda - konstrukcja

Wprowadzamy następujący algorytm:

- Za pomocą znanej nam metody najmniejszych kwadratów wyznaczamy $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ z modelu oryginalnego

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_k X_{t,k} + \epsilon_t,$$

oraz reszty:

$$\hat{Y}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_{t,1} - \hat{\beta}_2 X_{t,2} - \ldots - \hat{\beta}_k X_{t,k};$$

- Podobnie znajdujemy estymator $\hat{\beta}_0^b, \hat{\beta}_1^b, \ldots, \hat{\beta}_m^b$ parametru $\beta_0, \beta_1, \ldots, \beta_m$ w modelu obciętym ($\hat{\beta}_j^b$ oraz $\hat{\beta}_j$ $(j = 0, 1, \ldots, m)$ mogą się różnić choć, jeśli H_0 jest prawdziwa, przybliżają te same wartości)

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \ldots + \beta_m X_{t,m} + \epsilon_t,$$

oraz reszty:

$$\hat{Y}_t^b = Y_t - \hat{\beta}_0^b - \hat{\beta}_1^b X_{t,1} - \hat{\beta}_2^b X_{t,2} - \ldots - \hat{\beta}_m^b X_{t,m}.$$

Test Walda - konstrukcja

Jeśli hipoteza H_0 jest prawdziwa

- $\hat{Y}_t$ oraz $\hat{Y}_t^b$ nie różnią się zbyt wiele;
- Ponieważ,

$$\mathbf{e} := \sum_{j=1}^{k} \hat{Y}_t^2 = \hat{Y}^T \hat{Y}$$

oraz

$$\mathbf{b} := \sum_{j=1}^{m} (\hat{Y}_t^b)^2 = \hat{Y^b}^T \hat{Y^b},$$

$\mathbf{e}$ oraz $\mathbf{b}$ różnią się niewiele;

Test Walda - konstrukcja

- Statystyka testu Walda jest proporcjonalna do ró.znicy $\mathbf{b} - \mathbf{e}$ w stosunku do $\mathbf{e}$ i wyraża cię wzorem:

$$F^* = \frac{\frac{\mathbf{b}-\mathbf{e}}{k-m}}{\frac{\mathbf{e}}{n-(k+1)}}.$$

- Jeśli hipoteza H_0 jest prawdziwa, F ma rozkład *F-Snedecora* (lub *Fishera-Snedecora*) z $k - m$ oraz $n - (k + 1)$ stopniami swobody, który oznaczamy jako $\mathcal{F}(k - m, n - (k + 1))$;

Test Walda - konstrukcja

- Odrzucamy hipotezę H_0 na poziomie istotności κ (domyślnie $\kappa = 0.05$) wg następujących równoważnych procedur:
  - jeśli F^* przekroczy poziom krytyczny $f^*(1-\kappa, k-m, n-(k+1))$, gdzie $f^*(\cdot, k-m, n-(k+1))$ oznacza kwantyl rozkładu *F-Snedecora* - $\mathcal{F}(k-m, n-(k+1))$;
  - lub równoważnie

$$p - value = 1 - CDF_{\mathcal{F}(k-m,n-(k+1))}(F^*) \leq \kappa,$$

jak zwykle $CDF_{\mathcal{F}(k-m,n-(k+1))}$ oznacza dystrybuantę rozkładu $\mathcal{F}(k-m, n-(k+1))$.

Przykład - ceny mieszkań w Polsce

Przykład 1 (Ceny mieszkań w polskich miastach)

Ceny mieszkań w wybranych miastach podsumowuje tabela na następnej stronie. 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 - *cena mieszkania na* m^2 *w mieście* t*;*
- F_t - *populacja miasta* t*;*
- D_t - *gęstość zaludnienia w mieście* t*;*
- S_t - *powierzchnia województwa w* km^2 *w skład którego wchodzi miasto* t*.*

Odpowiemy na pytanie, czy cena mieszkania w mieście na jednostkę powierzchnii jest zależna od liczby jego mieszkańców, jego gęstości zaludnienia i powierzchnii całego regionu.

Przykład - dane

Województwo	Wybrane miasto	P_t	F_t	D_t	S_t
Podlaskie	Białystok	3933	295 459	2893	20187
Kuj.-Pomorskie	Bydgoszcz	3692	357 652	2032	17972
Pomorskie	Gdańsk	5552	461 489	1762	18310
Pomorskie	Gdynia	5037	247 820	1834	18310
Śląskie	Katowice	3636	301 834	1833	12333
Świętokrzyskie	Kielce	3617	198 857	1814	11711
Małopolskie	Kraków	5820	761 873	2331	15183
Lubelskie	Lublin	4571	341 722	2317	25122
Łódzkie	Łódź	3284	706 004	2408	18219
Warm.-Mazurskie	Olsztyn	4181	173 831	1968	24173
Opolskie	Opole	3910	119 574	1238	9412
Wielkopolskie	Poznań	5157	545 680	2083	29826
Podkarpackie	Rzeszów	4514	185 123	1591	17846
Zachodniopomorskie	Szczecin	4065	407 180	1355	22892
Mazowieckie	Warszawa	7300	1 735 442	3355	35558
Dolnośląskie	Wrocław	5332	634 487	2167	19947
Lubuskie	Zielona Góra	3267	138 512	498	13988

The solution

Za pomocą obliczeń w Gretlu mamy następujące obliczenia:

Zmienna	$\hat{\beta}_i$	$\sqrt{\hat{d}_{ii}}$	T_i	p-value
stała	3141.71	766.089	4.101	0.0013
F_t	0.00176187	0.000743498	2.370	0.0340
D_t	-0.0556650	0.420512	-0.1324	0.8967
S_t	0.0359841	0.0384350	0.9362	0.3662

Zatem mamy następujący model:

$$P_t \approx 3141.71 + 0.00176187 * F_t - 0.0556650 * D_t + 0.0359841 * S_t + \epsilon_t.$$

Za pomocą testu t-Studenta sprawdzamy istotność *pojedynczych* zmiennych $\beta_0, \beta_1, \beta_2$ oraz β_3 za pomocą wyników z poprzedniej tabelki przy poziomie istotności $\kappa = 0.05$:

Hipoteza	p-value	decyzja	czy jest istotna?
$H_0 : \beta_0 = 0$	0.0013	odrzucić H_0	tak
$H_0 : \beta_1 = 0$	0.0340	odrzucić H_0	tak
$H_0 : \beta_2 = 0$	0.8967	przyjąć H_0	nie
$H_0 : \beta_3 = 0$	0.3662	przyjąć H_0	nie

Rozwiązanie - C.D.

Uwaga 1

- *Przyjęcie hipotez* $H_0 : \beta_2 = 0$, *oraz* $H_0 : \beta_3 = 0$ *nie oznacza, że przyjmujemy hipotezę*

$$H_0 : \beta_2 = \beta_3 = 0 \quad VS \quad H_1 : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0.$$

- *Żeby zweryfikować powyższą hipotezę musimy użyć testu Walda.*

Rozwiązanie - CD

- Zatem za pomocą testów Studenta, D_t i S_t są nieistotne jako oddzielne zmienne;
- Za pomocą testu Walda sprawdzimy, czy obie zmienne są *łącznie* nieistotne.
  - Tu $k = 4$, $m = 2$, a jeśli hipoteza H_0 jest prawdziwa, statystyka F^* ma rozkład $\mathcal{F}(2, 13)$;
  - Obliczanie w Gretlu dają wyniki:

$$F^* = 0.44411 \quad \text{oraz} \quad p - value = 0.650772 > 0.05$$

;

  - Zatem przyjmujemy hipotezę

$$H_0 : \beta_3 = \beta_4 = 0$$

i tym samym **jesteśmy upoważnieni do usunięcia z modelu obu zmiennych D_t i S_t na raz.**

Rozwiązanie - CD

Po usunięciu obu zmiennych D_t i S_t obliczamy

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

Rozważamy zatem model obcięty i szacujemy w nim β_0, β_1. Testy t-Studenta istotności na poziomie $\kappa = 0.05$:

Hipoteza	p-value	decyzja	czy istotny?
$H_0 : \beta_0 = 0$	1.1610^{-09}	odrzucić H_0	tak
$H_0 : \beta_1 = 0$	0.0004	odrzucić H_0	tak

Obie zmienne są istotne. Model jest ostateczny.