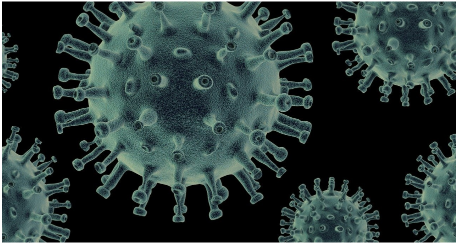
The verification of assumptions of this model (I)

13 kwietnia 2021



Model regresji wielorakiej-powtórzenie

Rozważamy madele regresji liniowej, które można zapisać w formie macierzowej:

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

W skrócie możemy więc zapisać:

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

Założenia

Przypuśćmy, że jest więcej obserwacji niż parametrów tzn: $k+1 \leq n$. Ponadto,

- Z1 Macierz $\mathbf{X}$ zmiennych objaśniających jest **deterministic** (nielosowa), tzn. $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ są nielosowe dla wszystkich t;
- Z2 Rząd macierzy $\mathbf{X}$ wynosi $k+1$, co oznacza, że **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 tej samej, ale nieznanej wariancji** σ^2;
- Z5 All ϵ_t has normal distribution $N(0, \sigma^2)$.

Plan na najbliższe wykłady

Mając oszacowanie $\hat{\beta}$ i reszt $\hat{\epsilon}$ wiemy tylko

- Jaki model jest najbliższy rzeczywistości;
- Mamy tylko przymiarkę do doboru zmiennych, które pasują do modelu bardziej, a które mniej;

Nie mając pewności czy założenia Z1-Z5 nadal nie wiemy:

- jak dokładne są oszacowania;
- czy poziomy krytyczne w testach istotności są adekwatne;

W tym celu wykonujemy analizujemy weryfikację niektórych założeń.

Weryfikacja założenia Z1

Założenie Z1, że $\mathbf{X}$ jest nielosowa jest niemożliwa do wykonania przez statystyka:

- Przykładowo gdy inżynier zleca statystykowi zbadanie opracowanie modelu wytrzymałości belki w zależności od rozkładu włókien, to on musi wskazać które zmienne mogą być losowe, a które nie;
- Czasami Z1 mamy za darmo, np. $X_{j,t}$ jest zadana jako funkcja czasu np. $X_{t,j} = t^2$;

Weryfikacja założeń Z2

Aby sprawdzić czy rząd macierzy $\mathbf{X}$ wynosi $k+1$:

- Wystarczy sprawdzić czy kolumny macierzy $\mathbf{X}$ są liniowo niezależne;
- Jest to proste ćwiczenie z algebry i prosta komenda dla przeciętnego pakietu matematycznego.

Weryfikacja założenia Z3

Aby sprawdzić $E\epsilon_t = 0$ wystarczy zauważyć:

- W zasadzie to reszty mają średnią równą 0;
- Pakiet zwróci najbliższą liczbę maszynową zeru;
- Tradycyjny test Studenta dla próby zweryfikuje pozytywnie hipotezę, że Z3 jest spełnione;

Weryfikacja założenia Z4-Z5

Z4 Weryfikacja założenia Z4 prowadzi do poniższych hipotez pomocniczych:

- ϵ_t jest ciągiem nieskorelowanych zmiennych losowych, w tym celu używamy testu o braku autokorelacji *test Durbina-Watsona*;
- ϵ_t jest ciągiem *homoskedastycznym*, tzn. $Var(\epsilon_t) = \sigma^2$ dla t (test Harrisona McCabe'a, test White);

Z5 Jest wiele testów normalności, np. test *chi-kwadrat zgodności*, ale poznamy również inne testy:

- *test Jarque-Bera* - łatwy do przeprowadzenia i popularny rozkład statystyki testowej (rozkład χ^2);
- *test Shapiro-Wilka* - bardziej skomplikowany, ale mocy i **odporny na autokorelację reszt**;

Założenia [Z4] i [Z5]

Znaczenie założeń [Z4] i [Z5]

Potrzebujemy założeń [Z4] and [Z5] dla testów istotności

- **pojedynczych zmiennych**, w przeciwnym razie p-value nie będzie wyrażona przez dystrybuantą *rozkładu t-Studenta* (rozkład statystyki testowej może być inny);
- **grup zmiennych**, w przeciwnym razie p-value *testu Walda* nie jest poprawnie wyznaczona (rozkład statystyki nie musi być F-Snedecore) (VERTE).

Założenia [Z4] i [Z5]

Znaczenie założeń [Z4] i [Z5] - C.D.

Ponadto,

- Założenie normalności [Z5] wzmacnia zasadność używania estymatora $\hat{\beta}$, ponieważ jest uzyskany nie tylko za pomocą metody *najmniejszych kwadratów*, ale również za pomocą metody *największej wiarogodności*;
- Jeśli założenie homoskedastyczności [Z4] nie jest spełnione, zamiast metody *najmniejszych kwadratów* można zastosować metodę *ważonych najmniejszych kwadratów*:

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{t=1}^{n} w_t \left(y_t - \beta_0 - \beta_1 x_{t,1} -, \ldots, -\beta_k x_{t,k}\right)^2,$$

dla wag w_t zależnych od wariancji.

Zakładamy jak zwykle

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

ale pozwalamy, żeby ϵ_t nie był nieskorelowany poprzez **istnienie autokorelacji**, tzn.

$$\epsilon_t = \rho\epsilon_{t-1} + \eta_t,$$

gdzie η_t jest *prawdziwym* składnikiem losowym spełniającym założenia [Z3,Z4] i [Z5], gdzie dla wszystkich t

$$Var(\eta_t) = \sigma_0, \quad \text{and} \quad Cov(\epsilon_t, \eta_{t+1}) = 0,$$

gdzie $\rho \in (-1, 1)$ jest nieznanym parametrem, a σ_0 jest *parametrem zakłócającym*, który jest również nieznany.

Testowanie hipotezy ϵ_t jest ciągiem nieskorelowanych zmiennych prowadzi do problemu testowania hipotezy pomocniczej:

$$H_0 : \rho = 0 \quad \text{VS} \quad \rho \neq 0;$$

- **Brak autokorelacji:** przyjmując hipotezę H_0 przyjmujemy, że $\epsilon_t = \eta_t$, stąd jest to *prawdziwy* składnik losowy (szum);
- **Istnienie autokorelacji:** gdy odrzucimy hipotezę H_0 przestajemy wierzyć, że ϵ_t jest *prawdziwym* szokiem;

Test Durbina - Watsona (konstrukcja statystyki testowej)

W przypadku istnienia autokorelacji

$$\epsilon_t = \rho\epsilon_{t-1} + \eta_t$$

możemy zauważyć,że:

$$Var(\epsilon_t) = Var(\rho\epsilon_{t-1} + \eta_t) = Var(\rho\epsilon_{t-1} + \eta_t) = \underbrace{\rho^2 Var(\epsilon_{t-1}) + \sigma_0^2}_{\text{ponieważ } Cov(\epsilon_{t-1},\eta_t)=0} = \sigma^2,$$

zatem ($Var(\epsilon_t) = \sigma^2$)

$$\sigma^2 = \frac{1}{1-\rho^2}\sigma_0^2 = \frac{1}{1-\rho^2}Var(\eta_t).$$

Test Durbina - Watsona (konstrukcja statystyki testowej)

Zatem,

$$Cov(\epsilon_t, \epsilon_{t-1}) = Cov(\rho\epsilon_{t-1} + \eta_t, \epsilon_{t-1})$$

$$= \underbrace{\rho Var(\epsilon_{t-1}) + Cov(\eta_t, \epsilon_{t-1})}_{\text{Kowariancja jest dwuliniowa}} = \frac{\rho\sigma_0^2}{1-\rho^2}.$$

Stąd korelacja pomiędzy ϵ_t i ϵ_{t-1} spełnia

$$Corr(\epsilon_t, \epsilon_{t-1}) = \frac{Cov(\epsilon_t, \epsilon_{t-1})}{\sqrt{Var(\epsilon_t) Var(\epsilon_{t-1})}} = \frac{\frac{\rho\sigma_0^2}{1-\rho^2}}{\frac{1}{1-\rho^2}\sigma_0^2} = \rho.$$

Dla testu hipotezy dotyczącej ϵ_t, potrzebujemy odpowiednika *reszt*, gdzie $\hat{\epsilon}$ spełnia jak zwykle:

$$\hat{\epsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t,1} - \ldots - \beta_k x_{t,k},$$

gdzie

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

Test Durbina - Watsona (konstrukcja statystyki testowej)

Estymator ρ jest wyrażony wzorem:

$$\hat{\rho} = \frac{\sum_{\tau=2}^{n} \hat{\epsilon}_{\tau}\hat{\epsilon}_{\tau-1}}{\sqrt{\sum_{\tau=2}^{n} \hat{\epsilon}_{\tau}^{2} \sum_{\tau=2}^{n} \hat{\epsilon}_{\tau-1}^{2}}} \approx \frac{Cov(\epsilon_t, \epsilon_{t-1})}{\sqrt{Var(\epsilon_t)Var(\epsilon_{t-1})}} = \rho.$$

Statystyka testowa:

$$DW := \frac{\sum_{t=2}^{n} (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\epsilon}_t^2}.$$

Test Durbina - Watsona (konstrukcja statystyki testowej)

Zauważmy

$$DW := \frac{\sum\limits_{t=2}^{n} \hat{\epsilon}_t^2 + \sum\limits_{t=2}^{n} \hat{\epsilon}_{t-1}^2 - 2\sum\limits_{k=2}^{2} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum\limits_{t=1}^{n} \hat{\epsilon}_t^2}$$

i dla dostatecznie dużych n mamy

$$\sum_{t=1}^{n} \hat{\epsilon}_t^2 \approx \sum_{t=2}^{n} \hat{\epsilon}_t^2 \approx \sum_{t=2}^{n} \hat{\epsilon}_{t-1}^2 .$$

Test Durbina - Watsona (konstrukcja statystyki testowej)

Mamy więc

$$DW \approx 2(1 - \hat{\rho}).$$

Wnioskujemy:

- DW zmienia się w przedziale $[0, 4]$;
- Jeśli hipoteza H_0 jest prawdziwa, oczekujemy $\hat{\rho} \approx 0$, stąd oczekujemy $DW \approx 2$;
- W przeciwnym razie odrzucimy hipotezę H_0 na rzecz $\rho > 0$ lub $\rho < 0$.

Jeśli $H_0 : \rho = 0$ jest prawdziwa:

- Rozkład statystyki testowej DW jest stablicowany;
- Zależy on od liczby obserwacji n i liczby parametrów k;
- Realizacja statystyki spełnia $DW \approx 2$.

Statystyka testu Durbina - Watsona - wartości krytyczne

- Jeśli wartość statystyki DW spełnia $DW > 2$ wtedy testujemy hipotezę
$$H_0 : \rho = 0 \quad \text{VS} \quad H_1 : \rho < 0$$
- Jeśli realizacja DW spełnia $DW < 2$ wtedy testujemy hipotezę
$$H_0 : \rho = 0 \quad \text{VS} \quad H_1 : \rho > 0$$

Statystyka testu Durbina - Watsona - wartości krytyczne

Niech $\kappa \in (0, 1)$ będzie ustalonym poziomem istotności (zwykle $\kappa = 0.05$).

- W odróżnieniu od większości testów, zczytujemy **dwie** wartości krytyczne d_U i d_L z tablic rozkładu DW, gdzie $d_L < d_U < 2$;
  - Jeśli hipotezą alternatywną jest $H_1 : \rho > 0$, co ma miejsce przy realizacji $DW < 2$ wtedy:
    1. Jeśli $DW < d_L$, wtedy **odrzucamy hipotezę** H_0 **na rzecz** $H_1 : \rho > 0$.
    2. Jeśli $d_L < DW < d_U$, wtedy **problem jest nierozstrzygnięty**.
    3. Jeśli $DW > d_U$, wtedy przyjmujemy hipotezę $H_0 : \rho = 0$.
  - Jeśli hipotezą alternatywną jest $H_1 : \rho < 0$, co ma miejsce przy realizacji $DW > 2$ wtedy:
    1. Jeśli $DW > 4 - d_L$, wtedy **odrzucamy hipotezę** H_0 **na rzecz** $H_1 : \rho < 0$.
    2. Jeśli $4 - d_U < DW < 4 - d_L$, wtedy **problem jest nierozstrzygnięty**.
    3. Jeśli $DW < 4 - d_U$, wtedy przyjmujemy hipotezę $H_0 : \rho = 0$.

Statystyka testu Durbina - Watsona - wartości krytyczne

Poniższa tabelka ilustruje procedurę decyzyjną w teście Durbina-Watsona. Wprowadzamy DW, d_U, d_L przy $d_L < d_U < 2$.

$DW < 2$ i $H_1 : \rho > 0$		
$DW < d_L$	$d_L < DW < d_U$	$DW > d_U$
odrzucamy H_0 na rzecz H_1	problem jest nierozstrzygnięty	przyjmujemy $\rho = 0$
$DW > 2$ and $H_1 : \rho < 0$		
$DW > 4 - d_L$	$4 - d_U < DW < 4 - d_L$	$DW < 4 - d_U$
odrzucamy H_0 na rzecz H_1	problem jest nierozstrzygnięty	przyjmujemy $\rho = 0$

Example

Znajdziemy związek między poziomem cen w Stanach Zjednoczonych, a

- kursem obligacji w Stanach Zjednoczonych;
- kursem obligacji w Australii;
- poziomem

Przykład (dane z GRETLa)

Znajdziemy model poziomu cen w Stanach Zjednoczonych w stosunku do cen w Australi oraz cen obligacji w Stanach Zjednoczonych i Australii. Na podstawie 77 obserwacji, model ma postać

$$Y_t = \beta_0 + \beta_1 * X_{t,1} + \beta_2 * X_{t,2} + \beta_3 * X_{t,3} + \epsilon_t,$$

gdzie:

- Y_t poziom cen w Stanach Zjednoczonych;
- $X_{t,1}$ ceny obligacji w Stanach Zjednoczonych;
- $X_{t,2}$ poziom cen w Australii;
- $X_{t,3}$ ceny obligacji w Australii;

Rozwiązanie

Na podstawie obliczeń w Gretlu mamy następujący model

$$Y_t = 6.52 + 1.33 * X_{t,1} + 0.45 * X_{t,2} + 0.89 * X_{t,3} + \epsilon_t$$

którego reszty dają następującą wartość statystyki Durbina-Watsona:

$$DW = 0.539680 < 2.$$

Rozwiązanie-C.D.

Zatem testujemy hipotezy:

$$H_0 : \rho = 0 \quad \text{VS} \quad H_1 : \rho > 0$$

gdzie $n = 77$ oznacza liczbę obserwacji i $k = 3$ oznacza liczbę parameterów. Z tablic rozkładu Durbina-Watsona zczytujemy

$$d_L = 1.5502 \quad \text{oraz} \quad d_U = 1.7117.$$

Rozwiązanie-C.D.

Mamy więc,

$$DW = 0.539680 < 1.5502 = d_L,$$

czyli odrzucamy hipotezę $H_0 : \rho = 0$ na rzecz $H_1 : \rho > 0$. Zatem mamy podstawy do odrzucenia hipotezy, że reszty są realizacą ciągu nieskorelowanych zmiennych losowych.