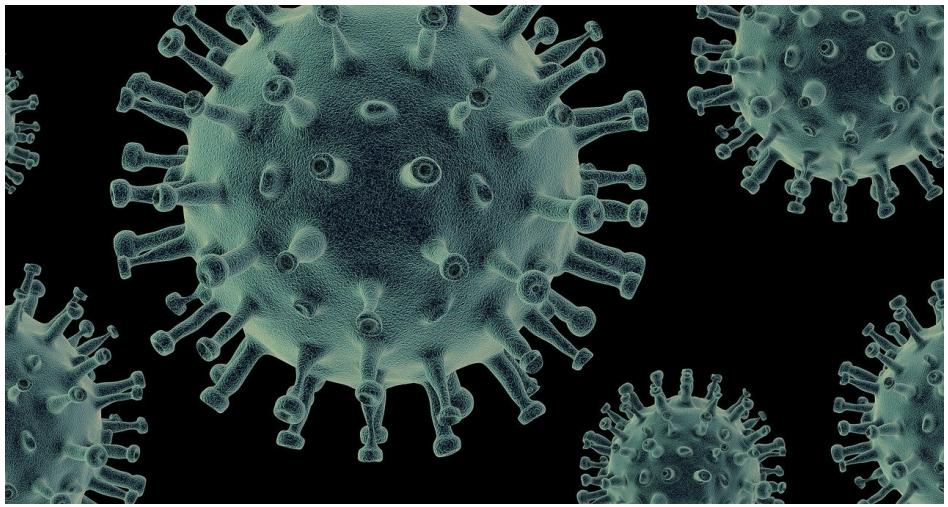
Weryfikacja założeń modelu (II)

19 kwietnia 2021



Model regresji wielorakiej-przypomnienie

Przypominam model:

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

Może być wyrażony w formie macierzowej:

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

Założenia

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

- Z1 Macierz zmiennych objaśniających $\mathbf{X}$ jest deterministyczna, tzn. macierz zmiennych objaśniających $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ nie jest losowa (dla wszystkich t);
- Z2 Rząd macierzy $\mathbf{X}$ jest równy $k+1$, zatem 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
  - wszystkie mają tą samą wariancję σ^2, (σ jest nieznaną liczbą);
- Z5 Wszystkie zmienne ϵ_t mają rozkład normalny $N(0, \sigma^2)$.

Autokorelacja składnika losowego ϵ_t

- Jak poprzednio zakładamy:

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

- W celu przetestowania hipotezy ϵ_t jest ciągiem nieskorelowanych zmiennych losowych, dopuszczamy hipotezę alternatywną, że ϵ_t **posiada autokorelację**, tzn.

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

gdzie η_t to *prawdziwy* składnik losowy spełniający założenia [Z3,Z4] and [Z5], gdzie dla t

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

oraz $\rho \in (-1, 1)$ jest nieznanym parametrem, a σ_0 jest parametrem *zakłócającym*.

Testowanie hipotezy - powtórzenie

Problem testowania hipotezy, że ciąg ϵ_t jest nieskorelowany prowadzi do następującej hipotezy pomocniczej:

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

- **Brak autokorelacji:** jeśli przyjmiemy hipotezę H_0, przypuszczamy, że $\epsilon_t = \eta_t$, stąd ϵ_t to *prawdziwy* szum;
- **Obecność autokorelacji:** jeśli odrzucimy hipotezę H_0 przestajemy przypuszczać, że ϵ_t jest *prawdziwym* szumem;

Test Durbina - Watsona powtórzenie

- Estymator ρ wyraża się wzorem:

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

- Do zweryfikowania hipotezy H_0 służy test *Durbina-Watsona*, którego statystyka testowa wyraża się wzorem:

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

gdzie $\hat{\epsilon}_t$ są *resztami* modelu.

Konstrukcja metody *Cochrana-Orcutta*

Jeśli test Durbina-Watsona nie przyjmie hipotezy $\rho = 0$ na rzecz $\rho > 0$ lub $\rho < 0$ (tzn. odrzuci hipotezę lub pozostawi problem nieroztrzygnięty) wtedy:

- Nie możemy zakładać ϵ_t spełnia założenie [Z4] i nie możemy ignorować parametru ρ;
- Za pomocą metody *Cochrana-Orcutta* można odpowiednio zmodyfikować model;

Konstrukcja metody *Cochrana-Orcutta*

Gdy test Durbina-Watsona nie przyjmie hipotezy $\rho = 0$. Wtedy dla $t = 1, 2, ..., n-1$ mamy (duże litery Y_t oraz X_t są zarezerwowane dla modelu, podczas gdy y_t oraz x_t to realizacje Y_t oraz X_t)

$$y_{t+1} = \beta_0 + \beta_1 x_{t+1,1} + \ldots + \beta_k x_{t+1,k} + \epsilon_{t+1} \tag{1}$$

oraz

$$y_t = \beta_0 + \beta_1 x_{t,1} + \ldots + \beta_k x_{t,k} + \epsilon_t \tag{2}$$

gdzie ϵ_t spełnia

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

a η_t spełnia [Z4]. Mnożymy obie strony (2) przez ρ i dodajemy do (1).

Konstrukcja metody *Cochrana-Orcutta*

Mnożąc (2) przez ρ otrzymamy

$$\rho y_t = \rho\beta_0 + \rho\beta_1 x_{t,1} + \ldots + \rho\beta_k x_{t,k} + \rho\epsilon_t$$

i dodajemy to wyrażenie do

$$y_{t+1} = \beta_0 + \beta_1 x_{t+1,1} + \ldots + \beta_k x_{t+1,k} + \epsilon_{t+1}$$

mamy

$$y_{t+1} - \rho y_t = \beta_0(1-\rho) + \beta_1(x_{t+1,1} - \rho x_{t,1}) + \ldots + \beta_k(x_{t+1,k} - \rho x_{t,k}) + \epsilon_{t+1} - \rho\epsilon_t.$$

Ponieważ nie znamy ρ, zastępujemy go przez $\hat{\rho}$:

$$\underbrace{y_{t+1} - \hat{\rho} y_t}_{y_t^*} = \underbrace{\beta_0(1-\hat{\rho})}_{\beta_0^*} + \underbrace{\beta_1}_{\beta_1^*}\underbrace{(x_{t+1,1} - \hat{\rho} x_{t,1})}_{x_{t,1}^*} + \ldots + \underbrace{\beta_k}_{\beta_k^*}\underbrace{(x_{t+1,k} - \hat{\rho} x_{t,k})}_{x_{t,k}^*} + \underbrace{\epsilon_{t+1} - \hat{\rho}\epsilon_t}_{\epsilon_t^* := \eta_{t+1}}.$$

Konstrukcja metody *Cochrana-Orcutta*

Innymi słowy, dla $t = 1, 2, ..., n-1$ mamy nowy model:

$$y_t^* = \beta_0^* + \beta_1^* x_{t,1}^* + \ldots + \beta_k^* x_{t,k}^* + \epsilon_t^*,$$

gdzie

- nieznanym **wektorem parametrów** jest

$$\beta_j^* = \begin{cases} \beta_0(1-\hat{\rho}) & if \quad j = 0 \\ \beta_j & if \quad j = 1, 2, \ldots, k \end{cases}$$

- rolę zmiennych **objaśniających** pełni

$$X_{t,k}^* = x_{t+1,k} - \hat{\rho} x_{t,k}$$

- rolę zmiennych **objaśnianych** pełni

$$y_t^* = y_{t+1} - \hat{\rho} y_t.$$

- Rolę **składnika losowego** pełni $\epsilon_t^* := \eta_{t+1}$.

Konstrukcja metody *Cochrana-Orcutta*

Podsumowując, metoda Cohrana -Orcutta sprowadza model gdzie **składnik losowy ma autokorelację** do modelu gdzie **składnik losowy nie ma autokorelacji**

$$\underbrace{\begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_{n-1}^* \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & X_{11}^* & X_{12}^* & \cdots & X_{1k}^* \\ 1 & X_{21}^* & X_{22}^* & \cdots & X_{2k}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n-1,1}^* & X_{n-1,2}^* & \cdots & X_{n-1,k}^* \end{bmatrix}}_{\mathbf{X}^*} \underbrace{\begin{bmatrix} \beta_0^* \\ \beta_1^* \\ \vdots \\ \beta_k^* \end{bmatrix}}_{\beta^*} + \underbrace{\begin{bmatrix} \epsilon_1^* \\ \epsilon_2^* \\ \vdots \\ \epsilon_{n-1}^* \end{bmatrix}}_{\epsilon^*}.$$

Można go wyrazić w formie macierzowej:

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

Iteracyjna metoda Cochrana-Orcutta

Dalsze kroki:

- Za pomocą znanych wzorów można obliczyć β^* za pomocą znanych wzorów

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

  i zrekonstruować estymator oryginalnego parametru β:

$$\beta_0 \to \frac{\beta_0^*}{1-\hat{\rho}} \text{ and } \quad \beta_i \to \beta_i^* \text{ and for } i \geq 1.$$

- Zatem obliczymy "nowe" reszty

$$\hat{\epsilon}_t^* = y_t^* - \hat{\beta}_0^* - \hat{\beta}_1^* x_{t,1}^* - \ldots - \hat{\beta}_k^* x_{t,k}^*;$$

- Przypuszczamy, że $\hat{\epsilon}_t^*$ jest szumem spełniającym założenie [Z4] o braku korelacji, ale jeśli nie jesteśmy pewni, ponownie stosujemy test Durbina-Watsona na bazie $\hat{\epsilon}_t^*$ i w razie potrzeby powtarzamy algorytm Cochrana-Orcutta.

Iteracyjna metoda Cochrana-Orcutta

Powtarzając metodę Cochrana-Orcutta można zastosować algorytm iteracyjny dla oszacowania modelu

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

1. **ARGUMENTY WEJŚCIOWE: Y, X** z modelu regresji wielorakiej i inicjujemy $s = 1$;
2. Jeśli $s = n$ zatrzymujemy algorytm. W przeciwnym razie obliczamy

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

3. Bazując na resztach, $\hat{\epsilon}$ stosujemy test Durbina-Watsona. Obliczamy i zapamiętujemy $\hat{\rho}_{\boldsymbol{s}} := \hat{\rho}$.
4. Jeśli test Durbina Watsona zastosowany w poprzednim punkcie przyjmie hipotezę o braku autokorelacji, **ZATRZYMUJEMY ALGORYTM** i zwracamy **ARGUMENTY WYJŚCIOWE**: estymatory: $\beta_{\boldsymbol{j}} \to \hat{\beta}_{\boldsymbol{j}}$ dla $j = 1, 2, \ldots, k$ oraz

$$\beta_{\mathbf{0}} \to \frac{\hat{\beta}_{\mathbf{0}}}{\prod_{\boldsymbol{r}=\mathbf{1}}^{\boldsymbol{s}-\mathbf{1}} (1 - \hat{\rho}_{\boldsymbol{r}})} \quad \text{if } s > 1 \quad \text{oraz } \beta_{\mathbf{0}} \to \hat{\beta}_{\mathbf{0}} \text{ jeśli } s = 1.$$

5. W przeciwnym razie wracamy do KROKU 2 z nowymy ARGUMENTAMI WEJŚCIOWYMI:

$$\mathbf{Y} \to \mathbf{Y}^*, \quad \mathbf{X} \to \mathbf{X}^*, \quad s \to s + 1.$$

Algorytm Cochrane'a-Orcutta

Uwaga

Można próbować stosować algorytm Cochrane'a-Orcutta do momentu gdy test Durbina-Watsona przyjmie hipotezę o braku korelacji. Stosowanie algorytmu Cochrana-Orcutta redukuje długość wektora $\mathbf{Y}$ z n do $n-1$, stąd algorytm iteracyjny Cochrane'a -Orcutta ma niewięcej niż $n-1$ kroków.

Przykład - jesienna fala zachorowań w Polsce

Przykład

COVID w Polsce - jesienna fala zachorowań

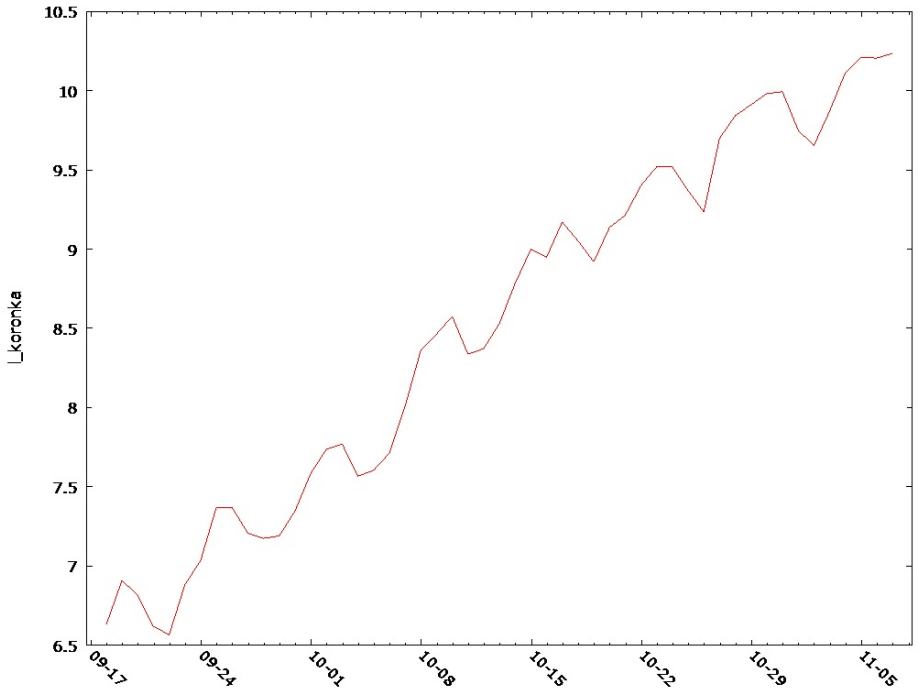
Analizujemy krzywą infekcji COVID-19 w okresie 18 września do 07 listopada 2020. Niech I_t będzie liczbą dziennych infekcji $Y_t = \ln(I_t)$. Analizujemy

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

dla $t = 1, 2, \ldots, 51$ oznaczającego indeks dnia począwszy od 18 września, a skończywszy na 7 listopada.

Rozwiązanie

Wkres Y_t trochę przypomina prostą.



Rozwiązanie

Za pomocą standardowej metody najmniejszych kwadratów mamy model:

$$Y_t = 6.52413 + 0.0767324 * t + \epsilon_t,$$

ale zastanawiamy się, czy ϵ_t to klasyczny szum, poprzez rozważenie szerszego modelu:

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

gdzie η_t jest prawdziwym szumem, natomiast $\rho \in (-1, 1)$, a na końcu zweryfikowanie hipotezy, że $\rho = 0$.

Rozwiązanie

- W tym celu obliczamy reszty $\hat{\epsilon}_t$ i bazując na nim obliczamy statystykę Durbina-Watsona

$$DW = 0.728864 < 2,$$

zatem testujemy

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

- Ponieważ liczba obserwacji to $n = 51$, a parametr jest jeden $k = 1$, stąd wartości krytyczne dla $\kappa = 0.05$ to

$$d_L = 1.5086 \quad \text{oraz} \quad d_U = 1.5884 \quad (\text{ wartości stablicowane}).$$

- Zatem stosując algorytm z poprzedniego wykładu $DW = 0.728864 < d_L = 1.5086$, odrzucamy hipotezę $H_0 : \rho = 0$ na rzecz $H_1 : \rho > 0$.
- Modyfikujemy model za pomocą metody Cochrana-Orcutta z oszacowaniem autokorelacji

$$\hat{\rho}_1 = 0.6246.$$

Rozwiązanie

- Zatem rozważamy nowy model $t = 1, 2, .., n-1$ with $n - 1 = 50$ gdzie

$$y_t^* = \beta_0^* + \beta_1^* x_t^* + \epsilon_t^*$$

gdzie ϵ_t^* oznaczają szum w nowym modelu oraz

$$y_t^* = y_{t+1} - \hat{\rho} y_t \quad \text{oraz } x_t^* = x_{t+1} - \hat{\rho} x_t.$$

- Przyjmując $\hat{\rho} = 0.6246$ oraz $x_t = t$ mamy

$$y_t^* = y_{t+1} - 0.6246 * y_t$$

jako nową zmienną objaśnianą oraz

$$x_t^* = (t+1) - 0.6246 * t = 0.3754 * t + 1$$

nową zmienną objaśniającą.

Rozwiązanie

- Stosując ponownie metodę najmniejszych kwadratów

$$y_t^* = 2.45280 + 0.0760954 * x_t^* + \epsilon_t^*$$

$t = 1, 2, .., 50$ gdzie ϵ_t^* oznacza szum w nowym modelu.

- Na podstawie reszt w nowym modelu, stosujemy test Durbina-Watsona.

Rozwiązanie

- Na bazie nowych reszt otrzymujemy $\hat{\rho}_2 = 0.3031$ i otrzymujemy:

$$DW = 1.346007 < 2$$

stąd testujemy autokorelację w nowym modelu:

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

- Krytyczne wartości dla $n = 50$ oraz $k = 1$ to

$$d_L = 1.5035 \quad \text{oraz} \quad d_U = 1.5849.$$

- Ponieważ

$$DW = 1.346007 < d_L = 1.5035$$

ponownie odrzucamy hipotezę $\rho = 0$ na rzecz $\rho > 0$.

- Powtarzamy algorytm Cohrana-Orcutta z obliczonym $\rho_2 = 0.3031$.

Rozwiązanie

- Zmienna objaśniana to:

$$y_t^{**} = y_{t+1}^* - \hat{\rho}_2 y_t^*,$$

a objaśniająca to

$$x_t^{**} = x_{t+1}^* - \hat{\rho}_2 x_t^*.$$

- Przy $\hat{\rho}_2 = 0.3031$, oraz $t = 1, 2, ..., 49$ otrzymujemy

$$y_t^{**} = y_{t+1}^* - 0.3031 y_t^*$$

oraz

$$x_t^{**} = x_{t+1}^* - 0.3031 x_t^*.$$

Rozwiązanie

- Nowy model daje $\beta_0^{**} = 1.69091$ oraz $\beta_1^{**} = 0.0778724$.
- Stosując test Durbina-Watsona dla tego modelu $(n = 49, k = 1)$ mamy

$$DW = 1.727780, \quad d_L = 1.4982 \quad d_U = 1.5813.$$

Ponieważ $DW < 2$, ponownie testujemy hipotezę $H_0 : \rho = 0$ przeciwko $H_1 : \rho > 0$.

- Otrzymujemy

$$DW = 1.727780 > d_U = 1.5813.$$

- Tym razem przyjmujemy hipotezę, że reszty tworzą biały szum.

Rozwiązanie

- Teraz podsumujemy model:

$$y_t = \beta_0 + \beta_1 * t + \epsilon_t$$

gdzie

$$\epsilon_t = 0.6246\epsilon_{t-1} + \eta_t,$$

ale η_t również ma autokorelację

$$\eta_t = 0.3031\eta_{t-1} + \xi_t$$

gdzie ξ_t jest prawdziwym szumem.

- Pamiętamy aproksymacje $\hat{\beta}_0^{**} = 1.69091$ oraz $\beta_1^{**} = 0.0778724$ i rekonstuujemy estymator β_0 oraz β_1 jako:

$$\hat{\beta}_0 \rightarrow = \frac{1.69091}{(1 - 0.6246)(1 - 0.3031)} = 6.43322 \quad \text{oraz} \quad \beta_1 \rightarrow 0.0778724.$$

Rozwiązanie

- Zatem ostateczny model

$$Y_t = 6.43322 + 0.0778724 * t + \epsilon_t.$$

- Ignorując ϵ_t i wracając do $Y_t = \ln(I_t)$ gdzie I_t - liczba dziennych infekcji

$$\ln(I_t) \approx 6.43322 + 0.0778724 * t$$

- Zatem

$$I_t \approx e^{6.43322+0.0778724*t} = e^{6.43322}(e^{0.0778724})^t,$$

hence

$$I_t \approx e^{6.43322+0.0778724*t} = 622.1741 * (1.080984716)^t.$$