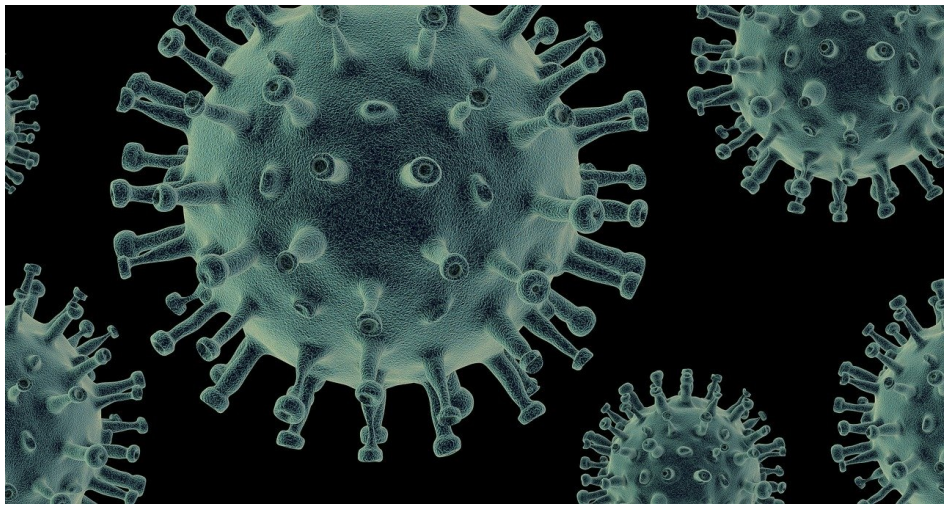
The verification of assumptions of this model (V)

10 maja 2021



Model regresji liniowej-przypomnienie

Ponownie rozważamy układ równań liniowych:

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

lub w formie macierzowej:

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

Założenia - powtórzenie

Przede wszystkim mamy więcej obserwacji niż parametrów, czyli $k + 1 \leq n$. Ponadto,

- Z1 Macierz $\mathbf{X}$ zmiennych objaśniających jest deterministyczna (nielosowa), tzn. macierz zmiennych objaśniających $[X_{t1}, X_{t2}, \ldots, X_{tk}]^T$ jest nielosowa dla t;
- Z2 Rząd macierzy $\mathbf{X}$ jest $k + 1$, zatem kolumny są liniowo niezależne;
- Z3 $E(\epsilon) = \mathbf{0}$, czyli $E\epsilon_t = 0$ dla t;
- Z4 $D^2(\epsilon) = \sigma^2\mathbf{I}$, tzn. ϵ_t jest ciągiem nieskorelowanych zmiennych losowych o wariancji σ^2, (σ jest nieznane);
- Z5 Wszystkie ϵ_t mają rozkład normalny $N(0, \sigma^2)$.

Testy weryfikujące założenie [Z4]

Zweryfikujemy założenie [Z4], że wszystkie ϵ_t mają tą samą wariancję. Robimy to w następującym celu.

- Jeśli to jest prawda, estymator $\hat{\beta}$ otrzymany metodą najmniejszych kwadratów jest:
  - nieobciążony, i.e. $E\hat{\beta} = \beta$ - przy wielokrotnym powtarzaniu eksperymentu, $\hat{\beta}$ będzie oscylował wokół prawdziwego β;
  - zgodny, tzn. dla dostatecznie dużych n, z dużym prawdopodobieństwem $\hat{\beta}$ będzie blisko β;
  - efektywny, tzn. najlepszy w sensie **metody najmniejszych kwadratów** spośród wszystkich liniowych estymatorów β (matematycznie $\hat{\beta}$ stanowią współrzędne rzutu ortogonalnego wektora Y na przestrzeń kolumn macierzy $\mathbf{X}$ ze standardowym iloczynem skalarnym w $\mathbb{R}^n$)

Testy weryfikujące założenie [Z4]

- Jeśli hipoteza jest fałszywa, można zmodyfikowoać metodę **ważonych najmniejszych kwadratów** np.

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{t=1}^{n} \underbrace{\sqrt{Var(\epsilon_t)}}_{w_t}(Y_t - \beta_0 - \beta_1 x_{t,1} - \ldots - \beta_k x_{t,k})^2.$$

Otrzymany estymator również daje współrzędne rzutu $\mathbf{Y}$ na przestrzeń generowaną przez kolumny $\mathbf{X}$ z iloczynem skalarnym postaci

$$(u_1, u_2, \ldots, u_n) \circ (v_1, v_2, \ldots, v_n) = \sum_{t=1}^{n} \sqrt{w_t} u_t v_t.$$

Intuicyjne własności estymatora $\hat{\beta}$

- **Nieobciążony**: podobnie jak rzut monetą: nagroda za orła jest 1, a za reszkę to -1
  - przy wielu rzutach ok. 50% wyników da nam 1 a pozostałe 50% da nam -1, zatem wyniki oscylują wokół średniej, czyli 0;
- **Zgodny**: przy wielu rzutach

$$\frac{\text{suma nagród}}{\text{liczba wszystkich rzutów}}$$

  - będzie blisko 0, np. $1/N$ or $-1/N$, dla dużej liczby rzutów N;

Terminologia

Homoskedastyczność i heteroskedatyczność

Ciąg zmiennych losowych $Z_1, Z_2, \ldots, Z_n$ jest

- **Homoskedastyczny:** jeśli wariancja nie zależy od indeksu, tzn.
$$Var(Z_1) = Var(Z_2) = \ldots = Var(Z_n);$$
- **Heteroskedastyczny:** jeśli wariancja zależy od indeksu, tzn..
$$Var(Z_i) \neq Var(Z_j) \quad \text{dla pewnego } i \neq j.$$

Test homoskedastyczności

Jak zwykle, reszy oznaczamy jako

$$\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta},$$

oraz zakładamy, że $\hat{\epsilon}_t$ jest realizacją zmiennej losowej ϵ_t. Testujemy hipotezę:

$H_0 : Var(\epsilon_1) = Var(\epsilon_2) = \ldots = Var(\epsilon_n)$ VS $H_1 : H_0$ jest fałszywa.

W tym celu opiszemy testy:

- **Harrisona-McCabe'a**;
- **White'a**.

Test Harrison-McCabe'a

Za pomocą reszt $\hat{\epsilon}_t$ konstruujemy **statystykę testową** jako:

$$HM = \frac{\sum_{t=1}^{m} \hat{\epsilon}_t^2}{\sum_{t=1}^{n} \hat{\epsilon}_t^2}$$

gdzie m jest w zasadzie dowolną liczbą taką, że $k + 1 < m < n - k - 1$, ale w praktyce, zależy monotonicznie od $|\hat{\epsilon}_t|$:(VERTE)

Test Harrison-McCabe'a

- Domyślna opcja to

$$m = \begin{cases} \frac{n}{2} & \text{jeśli } n \text{ jest parzyste} \\ \frac{n-1}{2} & \text{jeśli } n \text{ jest nieparzyste.} \end{cases}$$

- Jeśli widzimy, że $|\hat{\epsilon}_t|$ mają wyraźnie rosnącą tendencję, a następnie malejącą, wtedy m jest indeksem obserwacji t taki, że $|\hat{\epsilon}_t|$ jest największy;
- Jeśli widzimy, że $|\hat{\epsilon}_t|$ mają wyraźnie malejącą tendencję, a następnie rosnącą, wtedy m jest indeksem obserwacji t taki, że $|\hat{\epsilon}_t|$ jest najmniejszy;

Test Harrison-McCabe'a - obszar odrzucenia

Dla poziomu istotności α (domyślnie $\alpha = 0.05$) znajdujemy poziomy krytyczne:

$$b_L = \frac{1}{1 + \frac{(n-m)F_1}{n-(k+1)}} \quad b_U = \frac{1}{1 + \frac{[n-m-(k+1)]F_2}{m}},$$

gdzie:

- F_1 - jest kwantylem rozkładu $\mathcal{F}(n-m, m-(k+1))$ (Snedecora) rzędu $1-\alpha$;
- F_2 - jest kwantylem rozkładu $\mathcal{F}(n-m-(k+1), m)$ (Snedecora) rzędu $1-\alpha$;

Test Harrison-McCabe'a - obszar odrzucenia

Decyzja

- Jeśli $HM < d_L$, to odrzucamy hipotezę H_0, że ϵ_t jest **homoskedastyczny** na rzecz alternatywnej hipotezy, ϵ_t jest **heteroskedastyczny**;
- Jeśli $d_L < HM < d_U$, problem jest nierozstrzygnięty;
- Jeśli $HM > d_U$, to przyjmujemy hipotezę H_0, że ϵ_t jest **homoskedastyczny**.

Test Harrison-McCabe'a - obszar odrzucenia

Uwaga

- Idea testu: jeśli hipoteza homoskedastyczności jest prawdziwa, to

$$HM = \frac{\sum_{t=1}^{m} \hat{\epsilon}_t^2}{\sum_{t=1}^{n} \hat{\epsilon}_t^2} \approx \frac{m}{n}$$

  a odrzucimy hipotezę, gdy HM jest zbyt małe.
- Podobnie jak w przypadku testu Durbina-Watsona nie można wykluczyć, że test nie rostrzynie problemu.

Test White'a

Konstruujemy test **White'a**, ktŕy również weryfikuje kipotezę o homoskedastyczności reszt.

- Dla ilustracji załóżmy, że liczba parametrów wynosi $k = 2$:

$$Y_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \epsilon_t,$$

- Testujemy hipotezę homoskedastyczności dopuszczając heteroskedastyczność jako alternaltywę w następującej formie:

$$\epsilon_t^2 = a_0 + a_1 x_{t,1} + a_2 x_{t,2} + a_3 x_{t,1}^2 + a_4 x_{t,2}^2 + a_5 x_{t,1} x_{t,2} + \eta_t,$$

gdzie η_t jest prawdziwym szumem spełniającym założenie [Z4] i [Z5], oraz $a_0, a_1, a_2, a_3, a_4, a_5$ to nieznane parametry;

- Wtedy $Var(\epsilon_t) = E\epsilon_t^2 = \sigma_t^2$ ma następującą formę:

$$\sigma_t^2 = a_0 + a_1 x_{t,1} + a_2 x_{t,2} + a_3 x_{t,1}^2 + a_4 x_{t,2}^2 + a_5 x_{t,1} x_{t,2}.$$

Test White'a - konstrukcja

Testujemy homoskedastyczność ϵ_t, poprzez testowanie hipotezy pomocniczej:

$$H_0 : a_1 = a_2 = a_3 = a_4 = a_5 = 0 \quad \text{VS} \quad H_1 : \exists_{j=1,2,3,4,5} a_j \neq 0.$$

- Jeśli H_0 jest prawdziwa $Var(\epsilon_t) = a_0$ dla t, oraz $a_0 = \sigma^2$ tak jak [Z4];
- Zakładamy, że $\hat{\epsilon}_t$ jest realizacją ϵ_t;
- Pomocniczy model

$$\hat{\epsilon}_t^2 = a_0 + a_1 x_{t,1} + a_2 x_{t,2} + a_3 x_{t,1}^2 + a_4 x_{t,2}^2 + a_5 x_{t,1} x_{t,2} + \eta_t$$

jest modelem liniowym z $k = 5$ zmiennymi objaśniającymi;

  - rolę **zmiennej objaśnianej** gra $Y_t := \hat{\epsilon}_t^2$;
  - rolę **zmiennych objaśniających** grają: $x_{t,1}, x_{t,2}, x_{t,1}^2, x_{t,2}^2$, oraz $x_{t,1} x_{t,2}$.

Test White'a - konstrukcja

Znajdujemy **statystykę testową** w następującej formie:

- Szukamy $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$, $\hat{a}_3$, $\hat{a}_4$ i $\hat{a}_5$ za pomocą **standardowej metody najmniejszych kwadratów**:

$$\min_{a_0,a_1,a_2,a_3,a_4,a_5} \sum_{t=1}^{n} (\hat{\epsilon}_t^2 - a_0 - a_1 x_{t,1} - a_2 x_{t,2} - a_3 x_{t,1}^2 - a_4 x_{t,2}^2 - a_5 x_{t,1} x_{t,2})^2.$$

- Szukamy oszacowań reszt:

$$\hat{\eta}_t = \hat{\epsilon}_t^2 - \hat{a}_0 - \hat{a}_1 x_{t,1} - \hat{a}_2 x_{t,2} - \hat{a}_3 x_{t,1}^2 - \hat{a}_4 x_{t,2}^2 - \hat{a}_5 x_{t,1} x_{t,2};$$

Test White'a - konstrukcja

- Szukamy "prognoz" $\hat{\epsilon}_t^2$ wg wzoru:

$$\tilde{\epsilon}_t^2 := \hat{a}_0 + \hat{a}_1 x_{t,1} + \hat{a}_2 x_{t,2} + \hat{a}_3 x_{t,1}^2 + \hat{a}_4 x_{t,2}^2 + \hat{a}_5 x_{t,1} x_{t,2}$$

- Bazując na resztach obliczamy **współczynnik determinacji** (przypomnij sobie definicję):

$$R^2 = \frac{\sum_{t=1}^{n} \left(\tilde{\epsilon}_t^2 - \frac{1}{n} \sum_{j=1}^{n} \hat{\epsilon}_j^2 \right)^2}{\sum_{t=1}^{n} \left(\hat{\epsilon}_t^2 - \frac{1}{n} \sum_{j=1}^{n} \hat{\epsilon}_j^2 \right)^2}.$$

Test White'a - statystyka testowa

- Statystyka testowa ma następującą formę:

$$W = nR^2$$

- Jeśli hipoteza H_0 jest prawdziwa, W ma rozkład $\chi^2(5)$ (chi-kwadrat z 5 stopniami swobody);
- Odrzucamy hipotezę H_0 gdy:
  - $W > q_{\chi^2(5)}(1-\alpha)$ - kwantyl $\chi^2(5)$ rzędu $1-\alpha$,
  - lub równoważnie,

$$\alpha > p - value = 1 - CDF_{\chi^2(5)}(W),$$

gdzie $CDF_{\chi^2(5)}(\cdot)$ oznacza dystrybuantę rozkładu $\chi^2(5)$.

Test White'a - statystyka testowa

Uwagi

- Uogólnienie testu **White'a** na przypadek k zmiennych objaśnianych jest również znane. W tym przypadku, rozkład statystyki przy hipotezie H_0 ma rozkład $\chi^2(2k+1)$.
- Gdy test White'a odrzuci H_0 na rzecz H_1 możemy rozważyć zmodyfikację estymatora $\hat{\beta}$ w kierunku ważonej metody najmniejszych kwadratów

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

  z wagą $w_t = \tilde{\epsilon}_t$;
- Metodę ważonych najmniejszych kwadratów można stosować również gdy test Harrisona Mc-Cabe'a odrzuci H_0, ale w odróżnieniu od testu White'a, procedura testowa nie daje nam żadnej wskazówki co do wag w_t.

rok	miesiąc	stopa bezrobocia
2010	styczeń	13,7
	luty	14
	marzec	14
	kwiecień	13,8
	maj	13,6
	czerwiec	13,6
	lipiec	13,8
	sierpień	13,9
	wrzesień	14
	październiki	14,1
	listopad	14,5
	grudzień	15,1
2011	styczeń	15,7
	luty	15,9
	marzec	16,1
	kwiecień	16
	maj	15,9
	czerwiec	15,9
	lipiec	16
	sierpień	16,2