



POLITECHNIKA ŚLĄSKA
WYDZIAŁ AUTOMATYKI, ELEKTRONIKI I INFORMATYKI

Laboratorium Identyfikacji Procesów

Ćw. nr 1:
Identyfikacja statyczna
Sprawozdanie

Data laboratorium: 05.11.2021 r.

AiR S2-I/Rob

Sekcja 2, podsekcja 1:

Klaudia Bończyk

Paweł Kaźmieruk

Maksymilian Skibiński

1 Wstęp

Tematem ćwiczeń laboratoryjnych była identyfikacja modeli statycznych metodą najmniejszych kwadratów. Na podstawie otrzymanych danych pomiarowych wyznaczane były oceny współczynników dla kilku rozpatrywanych modeli, a następnie modele te były ze sobą porównywane przy pomocy innych narzędzi statystyki takich jak test t Studenta czy wariancja resztkowa.

Ogólnie rzecz biorąc, rozpatrywane są modele następującej postaci:

$$\hat{y} = b_0 + b_1 u_1 + \dots + b_s u_s = \mathbf{u}^T \mathbf{b}, \quad (1)$$

gdzie \mathbf{u} to wektor wejść modelu, a \mathbf{b} to wektor jego współczynników. Na podstawie N pomiarów, możemy zapisać pomiary wejść w macierzy \mathbf{U} , a wyjść w wektorze \mathbf{y} :

$$\mathbf{U} = \begin{bmatrix} 1 & u_{11} & u_{12} & \dots & u_{1s} \\ 1 & u_{21} & u_{22} & \dots & u_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{N1} & u_{N2} & \dots & u_{Ns} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad (2)$$

Kolumnom w macierzy \mathbf{U} odpowiadają kolejne wejścia modelu, a wierszom (w wektorze \mathbf{y} także) kolejne dane pomiarowe.

Choć model zapisany równaniem 1 jest liniowy to badany model wcale taki być nie musi, wtedy należy odpowiednio przygotować macierz \mathbf{U} , np. dla modelu:

$$\hat{y} = b_0 + b_1 u_1 + b_2 u_2^2, \quad (3)$$

macierz \mathbf{U} przyjmie postać:

$$\mathbf{U} = \begin{bmatrix} 1 & u_{11} & (u_{12})^2 \\ 1 & u_{21} & (u_{22})^2 \\ \vdots & \vdots & \vdots \\ 1 & u_{N1} & (u_{N2})^2 \end{bmatrix} \quad (4)$$

By wyznaczyć ocenę współczynników modelu, czyli wektora \mathbf{b} , minimalizując wskaźnik w postaci sumy kwadratów błędów (między modelem, a danymi pomiarowymi), wykorzystujemy wzór:

$$\hat{\mathbf{b}} = \left(\mathbf{U}^T \mathbf{U} \right)^{-1} \mathbf{U}^T \mathbf{y} \quad (5)$$

Jedynym koniecznym warunkiem by istniało rozwiązanie podanego wyżej wzoru jest by macierz \mathbf{U} miała rząd $s + 1$. Wzór zatem nie jest szczególnie wybredny i prawie zawsze nam zwróci pewien wynik, natomiast potrzeba ocenić czy ma on sens czy nie i tutaj z pomocą przychodzi pewne narzędzia statystyki.

Pierwszym z nich jest stosunek korelacyjny η^2 , dany wzorem:

$$\eta^2 = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^N e_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (6)$$

Wartości tego wskaźnika zawierają się pomiędzy 0, a 1, gdzie przez 1 oznaczamy model doskonały (błędy wtedy są zerowe), a przez 0 model zupełnie beznadziejny. Poza tym gdy wartość współczynnika jest pomiędzy 0, a 1 (co raczej będzie zawsze) trudno powiedzieć na podstawie samej wartości liczbowej czy model jest wystarczająco dobry czy nie. Korzystając z tego wskaźnika, chodzi o to, by wykorzystać kilka modeli, a następnie porównać ich współczynniki korelacji – w ten sposób możemy ocenić który z nich jest lepszy, a który gorszy.

Drugim z wykorzystywanych wskaźników była wariancja resztowa, która jest oceną wariancji błędów identyfikacji uzyskiwaną przy pomocy następującego nieobciążonego estymatora:

$$\sigma^2 = \frac{1}{N - s - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (7)$$

Im mniejszą wartość ma ta wariancja tym lepiej działa badany przez nas model. Gdy wartość jest naprawdę mała można przypuszczać, że wybrany model naprawdę pokrywa się z „modelem faktycznym”, a różnice wynikają jedynie z powodu błędów losowych.

Do oceny modelu można skorzystać także z testu t Studenta, z którego skorzystaliśmy by sprawdzić hipotezę zerowania się współczynników. Zacząć musimy od wyznaczenia macierzy kowariancji współczynników modelu:

$$\text{cov}(\hat{\mathbf{b}}) = \sigma^2 (\mathbf{U}^T \mathbf{U})^{-1} \quad (8)$$

Na jej diagonalu znajdują się wariancje współczynników równania modelu:

$$\text{var}(\hat{b}_i) = d_{ii} \sigma^2, \quad (9)$$

które właśnie wykorzystujemy do otrzymania liczb t potrzebnych do wykonania wspomnianego testu. Są one wyznaczone według wzoru:

$$t_i = \frac{b_i}{\hat{\sigma} \sqrt{d_{ii}}} \quad (10)$$

W ten sposób każdy ze współczynników modelu otrzymuje swoją liczbę t_i , a następnie korzystając z tabeli wartości krytycznych rozkładu t-Studenta, możemy dokonać porównania. Wpierw, znajdujemy istotną dla nas liczbę krytyczną t_{kr} , a by to zrobić potrzebujemy:

- założyć pewien poziom istotności α ,
- obliczyć liczbę stopni swobody $\nu = N - s - 1$,

zatem:

$$t_{kr} = t_{\nu}^{\alpha} \quad (11)$$

Następnie, korzystając ze znalezionej liczby krytycznej dokonujemy porównania wyznaczonych wartości t_i z nią. Jeśli $t_i > t_{kr}$ to możemy odrzucić hipotezę, że y nie zmienia się wraz ze zmianą u_i .

Drugi z wykorzystanych testów to test F Snedecora. Wyznaczamy liczbę F:

$$F = \frac{N - s - 1}{s} \frac{R^2}{1 - R^2}, \quad (12)$$

gdzie przez R^2 oznaczamy współczynnik korelacji wielowymiarowej, który jest równy stosunkowi korelacyjnemu η^2 (wzór 6).

Tak jak w przypadku test t Studenta, dla wyznaczonej liczby testowej F , dokonujemy porównania z wartością krytyczną F_{kr} , którą należy znaleźć w tabeli wartości krytycznych rozkładu F Snedecora – dla założonego poziomu istotności α wartość krytyczną znajdujemy korzystając z wartości: ilości wejść modelu s oraz liczby stopni swobody $(N - s - 1)$:

$$F_{kr} = F_{N-s-1}^s \quad (13)$$

Jeśli $F > F_{kr}$ to odrzucamy hipotezę o nieistotności funkcji regresji (czyli po prostu modelu) z ryzykiem błędu określonym poziomem istotności – czyli, bardziej po ludzku, badany model ma sens. W przeciwnym przypadku nie można niczego stwierdzić o funkcji regresji na podstawie tego testu.

2 Kod

Dla celów wykonania ćwiczenia napisany został następujący kod:

```

1 clear all; close all; clc;
2
3 % Wczytywanie danych.
4 dane = load('Dane do identyfikacji-20211102/WM5_18.TXT');
5
6 % Ilość pomiarów.
7 N = size(dane, 1);
8
9 % Wyjście (przyjmujemy ostatnią kolumnę danych).
10 y = dane(:, 3);
11
12 % Macierz U: pierwsza kolumna to jedyńki (współczynnik b_0)
13 % pozostałe kolumny pochodzą z danych -- są to wejścia (odpowienie
14 % kolumny).
15 % Odkomentowany pozostawiamy jeden wariant -- testujemy jeden model.
16
17 % U = [ones(N, 1), dane(:, 1) dane(:, 2)];
18 % U = [ones(N, 1), dane(:, 1) dane(:, 2) dane(:, 1).^2 dane(:, 2).^2];
19 U = [ones(N, 1), dane(:, 1) dane(:, 2) dane(:, 1).^2 dane(:, 2).^2 dane(:,
    1).*dane(:, 2)];
20 % U = [ones(N, 1), dane(:, 1) dane(:, 2) dane(:, 1).^2 dane(:, 2).^2 dane
    (:, 1).*dane(:, 2) dane(:, 1).^3 dane(:, 2).^3];
21
22 % Ilość wejść modelu.
23 s = size(U, 2) - 1;
24
25 % Ocena parametrów modelu -- wektor b.
```

```

26 b = (U' * U)^-1 * U' * y;
27
28 % Wyjście uzyskane z modelu.
29 y_hat = U * b;
30
31 % Wektor błędów.
32 e = y - y_hat;
33
34 % Stosunek korelacyjny.
35 eta_sqr = sum((y_hat - mean(y_hat)).^2) / sum((y - mean(y)).^2);
36
37 % Wariancja resztowa.
38 sigma_sqr = 1 / (N - s - 1) * e' * e;
39
40 % Test F Snedecora.
41 F = (N - s - 1) / s * eta_sqr / (1 - eta_sqr);
42
43 % Macierz kowariancji współczynników.
44 cov_b = sigma_sqr * (U' * U)^-1;
45
46 % Test t Studenta -- wektor wartości t.
47 t = b ./ sqrt(diag(cov_b));
48
49 % Wybrany poziom istotności dla testów t i F.
50 alfa = 0.05;
51
52 % Ilość stopni swobody.
53 ni = N - s - 1;

```

3 Uzyskane wartości

Wykorzystywane dane pomiarowe to 3 kolumnowe macierze, o następującej budowie:

$$[\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{y}],$$

ale liczba ich wierszy – czyli ilość pomiarów – jest już zmienna, tzn. rozpatrywane są 3 sytuacje: $N = 9$, $N = 18$ albo $N = 27$. Spośród umieszczonych na platformie PZE danych pomiarów wybieramy zestaw nr 5, zatem wykorzystujemy pliki: WM5_9.TXT, WM5_18.TXT, WM5_27.TXT.

Zbadaliśmy 4 modele:

$$y = b_0 + b_1u_1 + b_2u_2 \quad (\text{model 1})$$

$$y = b_0 + b_1u_1 + b_2u_2 + b_3u_1^2 + b_4u_2^2 \quad (\text{model 2})$$

$$y = b_0 + b_1u_1 + b_2u_2 + b_3u_1^2 + b_4u_2^2 + b_5u_1u_2 \quad (\text{model 3})$$

$$y = b_0 + b_1u_1 + b_2u_2 + b_3u_1^2 + b_4u_2^2 + b_5u_1u_2 + b_6u_1^3 + b_7u_2^3 \quad (\text{model 4})$$

Na następnych stronach znajdują się tabele z wartościami liczbowymi ocen współczynników oraz odpowiednich wykorzystanych statystyk.

		$N = 9$	$N = 18$	$N = 27$
model 1	b_0	4.52	3.71	2.79
	b_1	0.266	0.226	0.208
	b_2	-0.255	-0.252	-0.231
model 2	b_0	2.79	1.95	1.21
	b_1	0.311	0.277	0.281
	b_2	-0.255	-0.252	-0.231
	b_3	0.0230	0.0256	0.0361
	b_4	0.0306	0.0466	0.0567
model 3	b_0	2.79	1.95	1.21
	b_1	0.311	0.277	0.281
	b_2	-0.342	-0.346	-0.325
	b_3	0.0230	0.0256	0.0361
	b_4	0.0306	0.0466	0.0567
	b_5	-0.0870	-0.0945	-0.0944
model 4	b_0	2.47	1.76	0.971
	b_1	-0.911	0.0909	0.0499
	b_2	0.0580	-0.331	-0.163
	b_3	0.0375	0.0403	0.0553
	b_4	0.0306	0.0466	0.0567
	b_5	0.0870	-0.0945	-0.0944
	b_6	0.0256	0.00491	0.00640
	b_7	-0.00230	-0.000379	-0.00415

Tabela 1: Ocena współczynników b

	$N = 9$	$N = 18$	$N = 27$
model 1	0.306	0.246	0.179
model 2	0.360	0.357	0.425
model 3	0.919	0.881	0.853
model 4	0.922	0.887	0.873

Tabela 2: Stosunek korelacji η^2

	$N = 9$	$N = 18$	$N = 27$
model 1	15.0	10.4	8.67
model 2	20.8	10.2	6.62
model 3	3.51	2.06	1.78
model 4	66.9	2.33	1.69

Tabela 3: Wariancja resztowa σ^2

		$N = 9$	$N = 18$	$N = 27$
model 1	t_0	3.44	4.78	4.78
	t_1	1.17	1.48	1.53
	t_2	-1.13	-1.65	-1.70
model 2	t_0	0.828	1.37	1.68
	t_1	1.05	1.63	2.14
	t_2	-0.958	-1.66	-1.94
	t_3	0.349	0.672	1.29
	t_4	0.465	1.22	2.03
model 3	t_0	2.01	3.06	3.24
	t_1	2.55	3.63	4.12
	t_2	-3.08	-5.00	-5.18
	t_3	0.849	1.50	2.50
	t_4	1.13	2.72	3.93
	t_5	-4.55	-7.25	-7.81
model 4	t_0	$3.33 \cdot 10^{-8}$	2.42	2.42
	t_1	$-1.28 \cdot 10^{-8}$	0.356	0.292
	t_2	$1.09 \cdot 10^{-9}$	-1.22	-0.899
	t_3	$8.10 \cdot 10^{-9}$	1.53	2.87
	t_4	0.259	2.56	4.02
	t_5	-1.04	-6.81	-8.00
	t_6	$1.66 \cdot 10^{-8}$	0.769	1.46
	t_7	$-2.13 \cdot 10^{-9}$	-0.0594	-0.948

Tabela 4: Wartości t dla testu t Studenta

	$N = 9$	$N = 18$	$N = 27$
model 1	2.45	2.13	2.06
model 2	2.78	2.16	2.07
model 3	3.18	2.18	2.08
model 4	12.7	2.23	2.09

Tabela 5: Wartości krytyczne t_{kr} (test t Studenta)

	$N = 9$	$N = 18$	$N = 27$
model 1	1.32	2.44	2.61
model 2	0.563	1.81	4.06
model 3	6.81	17.7	24.3
model 4	1.69	11.2	18.7

Tabela 6: Liczby testowe F Snedecora

	$N = 9$	$N = 18$	$N = 27$
model 1	5.14	3.68	3.40
model 2	6.39	3.18	2.83
model 3	9.01	3.11	2.70
model 4	236	3.14	2.54

Tabela 7: Wartości krytyczne F_{kr} (test F Snedecora)

4 Analiza otrzymanych danych

4.1 Obserwacje

- Tabela 1: Dla wszystkich modeli wartość wyrazu wolnego b_0 maleje wraz ze wzrostem N . Dla modelu 1, wartości wszystkich współczynników maleją wraz ze wzrostem N .

Zwiększanie liczby pomiarów N zawsze wpływa na ocenę współczynników modelu, ale nigdy nie są to zmiany bardzo dużo – jest to pewne dostrajanie, tego co udało się uzyskać już dla małej ilości danych. Pewien wyjątek może stanowić model 4, który jest najbardziej skomplikowany.

- Tabela 2: Modele 3 i 4 wypadają znacznie lepiej niż modele 1 i 2, tzn. modele 3 i 4 mają znacznie większe wartości stosunku korelacji w porównaniu do pozostałych.

Zwiększanie liczby pomiarów (N) powoduje głównie spadek wartości stosunku korelacji (poza modelem 2).

- Tabela 3: Wariancja resztowa maleje wraz ze wzrostem N .

Podobnie jak poprzednio, modele 3 i 4 wypadają znacznie lepiej niż 1 i 2, tzn. ich wariancje resztowe są mniejsze, ale trzeba zaznaczyć, że pewien wyjątek stanowi model nr 4 dla $N = 9$ – tutaj wariancja resztowa jest bardzo duża, znacznie większa od wszystkich zarejestrowanych wartości.

- Tabele 4 i 5: Wartości liczb t wzrastają (mamy na myśli ich wartości bezwzględne) wraz ze wzrostem liczby pomiarów, choć znów pewien wyjątek stanowi model nr 4, który dla liczby pomiarów $N = 9$ ma wartości t znacznie mniejsze od wszystkich zarejestrowanych podczas badań.

Natomiast, wartości krytyczne t_{kr} maleją ze wzrostem N , a rosną ze wzrostem liczby stopni swobody ν .

Dla najmniejszej liczby pomiarów ($N = 9$) mało który model przechodzi w całości, tzn. wszystkie jego współczynniki, test t Studenta weryfikujący hipotezę zerową, tzn. nie możemy jej dla większości współczynników odrzucić. Dla $N = 18$ model nr 3 wypada najlepiej, bo tylko dla jednego jego współczynnik t jest mniejsze od wartości krytycznej t_{kr} , model 4 trochę gorzej, a 1 i 2 jeszcze gorzej. Dla $N = 27$ można wysunąć podobne wnioski, choć jest jeszcze lepiej: wszystkie t dla modelu 3 są większe t_{kr} , a dla modelu 4 jest ich więcej niż poprzednio.

- Tabele 6 i 7: Liczby testowe F rosną ze wzrostem N , a wartości krytyczne F_{kr} maleją.

Dla $N = 9$ żadna z wartości F nie jest większa od F_{kr} , a dla $N = 18$ i $N = 27$ wartości F modeli 3 i 4 są większe od ich wartości testowych. Dla $N = 27$ także model 2 przechodzi ten test. Analiza tych tabel przemawia ponownie na korzyść modeli 3 i 4.

Trzeba jednak znów zauważyć jak model 4 wypada bardzo słabo dla $N = 9$.

4.2 Wnioski

Poza tym co zostało zauważone i opisane wyżej trzeba jeszcze raz pochylić się nad modelem nr 4, który można powiedzieć zachowywał się „dziwacznie” w porównaniu do pozostałych, tzn. wartości współczynników b oraz wartości wykorzystywanych narzędzi statystyki przy ocenie modeli, miały inne wartości i zmieniały się one w trochę inny sposób.

Macierz U dla modelu czwartego jest gorzej uwarunkowana niż te odpowiadające mniej skomplikowanym modelom. Gdy tego typu problemy rozwiązywane są ręcznie na papierze bez pomocy komputera nie ma to większego znaczenia, ale tak raczej nie jest. Zmiennoprzecinkowe typy danych używane przez komputery mają swoją określoną dokładność i te 8 bajtów nie wystarczy nam w takich sytuacjach do uzyskania poprawnych wartości liczbowych, np. w trakcie odwracania macierzy.

Przykładowo wskaźnik unormowania macierzy¹ U dla $N = 9$ wynosi:

$$\text{cond}(U) = 7.7545 \cdot 10^{17}$$

Czym mniej tym lepiej, a tutaj wartość jest bardzo duża, znacznie większa niż np. dla modelu nr 3 dla którego jest to 117. Dla $N = 18$ wskaźnik ten dla modelu nr 4 spada już do ok. 300, co wystarcza dla naszych obliczeń.

Naszym celem jest wybór modelu na podstawie danych pomiarowych. Modele 1 i 2 wypadły znacznie gorzej, a modele 3 i 4 już lepiej. Spośród tych dwóch wybieramy model nr 3, bo po pierwsze dla $N = 9$ model 4 jest po prostu beznadziejny, a dla pozostałych N radzi sobie równie dobrze co model 3, zatem nie ma potrzeby wybierać bardziej skomplikowanego modelu, skoro i tak radzi sobie podobnie.

Zatem, dla wszystkich serii pomiarowych wybieramy model nr 3.

$$N = 9 : \quad y = 2.79 + 0.311 u_1 - 0.342 u_2 + 0.0230 u_1^2 + 0.0306 u_2^2 - 0.0870 u_1 u_2$$

$$N = 18 : \quad y = 1.95 + 0.277 u_1 - 0.346 u_2 + 0.0256 u_1^2 + 0.0466 u_2^2 - 0.0945 u_1 u_2$$

$$N = 27 : \quad y = 1.21 + 0.281 u_1 - 0.325 u_2 + 0.0361 u_1^2 + 0.0567 u_2^2 - 0.0944 u_1 u_2$$

¹Polecenie `cond` w MATABie