Testowanie jakości dopasowania. Część 2.

Współczynnik korelacji wielokrotnej

Pojęcie współczynnika korelacji, który opisuje zależność między dwoma zmiennymi można rozszerzyć tak, żeby uwzględnić wielokrotne korelacje między zachodzące jednocześnie między wieloma zmiennych. Poprzednio otrzymaliśmy wzór, który wyraża współczynnik korelacji przez wariancje i kowariancję oraz współczynnik kierunkowy zależności liniowej obliczone dla zestawu danych pomiarowych

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = a\frac{s_{xy}}{s_y^2}.$$

Przez analogię zdefiniujemy współczynnik korelacji wielokrotnej (wielowymiarowej) R

$$R^2 \equiv \sum_{j=0}^{m}\left(a_j \frac{s_{jy}}{s_y^2} \right) = \sum_{j=0}^{m}\left(a_j \frac{s_j}{s_y} r_{jy} \right)$$

Współczynnik korelacji (liniowej) r_{jy} jest przydatny do testowania czy konkretna zmienna powinna być uwzględniona w modelu dopasowywanym do danych. Współczynnik korelacji wielokrotnej R charakteryzuje dopasowanie całego modelu (pełnej funkcji) do danych i można go używać do porównywania różnych dopasowywanych modeli (postaci funkcji).

Test F

Zmienna F (Fishera - Snedecora) jest obliczana dla prób (zestawów danych) dwóch zmiennych losowych i jest równa

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}.$$

gdzie: s_1^2 jest estymatorem wariancji pierwszej zmiennej σ_1^2, a s_2^2 estymatorem drugiej wariancji σ_2^2.
Zmienna F ma następujący rozkład gęstości prawdopodobieństwa:

$$p_F(x;\nu_1,\nu_2) = \frac{\Gamma\left[(\nu_1+\nu_2)/2\right]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{(1+x\nu_1/\nu_2)^{(\nu_1+\nu_2)/2}}$$

gdzie $x > 0$ a ν_1 i ν_2 są liczbami stopni swobody odpowiadającymi s_1^2 i s_2^2.
Jak wynika z definicji zmiennej F również stosunek zredukowanych zmiennych χ_ν^2 ma ten sam rozkład:

$$F = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2}$$

Wykorzystując zmienną F do testowania wartości stosunku χ_ν^2 korzystamy z tablic, w których podaje się wartości graniczne F_α

$$\alpha = \int_{F_\alpha}^{\infty} p_F(x; \nu_1, \nu_2)\, dx,$$

których przekroczenie może zdarzyć się z określonym prawdopodobieństwem α (zwykle 0,05 i 0,01).

Jeżeli uzyskana wartość stosunku $F \geq F_\alpha$ jest nie mniejsza od granicznej, czyli jest w tym przypadku bardzo mało prawdopodobna (na przykład <0,05 lub <0,01) to mamy prawo przypuszczać, że różnica między wartościami χ_ν^2 nie jest przypadkowa, że jest statystycznie istotna, i ich macierzyste rozkłady prawdopodobieństwa (wariancje) są różne.
Jeżeli wartość $F < F_\alpha$, to nie można wykluczyć, że obserwowana różnica jest przypadkowa.

Ze względu na konstrukcję tablic wartości granicznych rozkładu zmiennej F przy jej obliczaniu do licznika wstawiamy większą wartość.

Rozważmy sumę kwadratów odchyleń S_y^2 związaną z zakresem danych (rozrzutem względem ich średniej), pomijając dla uproszczenia czynniki wagowe

$$S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

S_y^2 jest statystyką o rozkładzie χ^2 z $n-1$ stopniami swobody. Załóżmy dalej, że

$$y(x_i) = \sum_{j=0}^{m} a_j f_j(x_i)$$

oraz oznaczmy

$$\bar{f}_j = \frac{1}{n}\sum_{i=1}^{n} f_j(x_i)$$

Sumę S_y^2 możemy przekształcić, drogą odpowiednich podstawień i przekształceń (wykorzystując przy tym fakt, że parametry dopasowania spełniają odpowiednie równania), do następującej postaci:

$$S_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left[(y_i - \bar{y})\sum_{j=0}^{m} a_j\left(f_j(x_i) - \bar{f}_j(x_i)\right)\right] + \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{m} a_j f_j(x_i)\right)^2$$

lub krócej

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left[(y_i - \bar{y})\sum_{j=0}^{m} a_j\left(f_j - \bar{f}_j\right)\right] + \sum_{i=1}^{n}\left[y_i - \sum_{j=0}^{m} a_j f_j\right]^2$$

i po zmianie kolejności sumowania

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{j=0}^{m}\left[a_j \sum_{i=1}^{n}\left[(y_i - \bar{y})\left(f_j - \bar{f}_j\right)\right]\right] + \sum_{i=1}^{n}[y_i - y(x_i)]^2$$

W statystyce udowadnia się twierdzenie dotyczące właściwości zmiennych χ^2, które mówi, że suma dwóch zmiennych χ_1^2 i χ_2^2 o ν_1 i ν_2 stopniach swobody jest też zmienną χ^2 o $\nu = \nu_1 + \nu_2$ stopniach swobody.

Jeżeli przyjrzymy się rozkładowi na składniki sumy S_y^2, która jest zmienną χ^2 o $n-1$ stopniach swobody, to zauważymy, że $\sum[y_i - y(x_i)]^2$ jest też zmienną χ^2, ale o $n-m-1$ stopniach swobody. Zatem pierwszy składnik też musi być zmienną χ^2 o m stopniach swobody.
Wróćmy teraz do definicji współczynnika korelacji wielokrotnej R:

$$R^2 \equiv \sum_{j=0}^{m}\left(a_j \frac{s_{jy}}{s_y^2}\right) = \frac{\sum_{j=0}^{m}\left(a_j s_{jy}\right)}{s_y^2}$$

Przez analogię możemy przedstawić pierwszy składnik S_y^2 jako

$$\sum_{j=0}^{m}\left[a_j \sum_{i=1}^{n}\left[(y_i-\bar{y})\left(f_j-\bar{f}_j\right)\right]\right] = R^2 S_y^2 = R^2\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Zatem

$$\sum[y_i - y(x_i)]^2 = R^2\sum[y_i - y(x_i)]^2 + (1-R^2)\sum[y_i - y(x_i)]^2$$

gdzie oba składniki po prawej stronie są w dalszym ciągu zmiennymi o rozkładach χ^2 z $n-m-1$ i m stopniami swobody.

Sumę S_y^2 (która jest miarą rozrzutu wartości zmiennej zależnej) nazywa się często początkową sumą kwadratów, a $\sum[y_i - y(x_i)]^2$ resztkową sumą kwadratów (pozostającą po dopasowaniu funkcji $y(x)$). W tym kontekście $R^2\sum[y_i - y(x_i)]^2$ jest częścią początkowego rozrzutu usuniętą przez dopasowania. Dopasowanie (w sensie użytego modelu) jest tym lepsze im większą część początkowego rozrzutu usuwa. Do sprawdzenia czy pozostały rozrzut jest istotnie mniejszy od usuniętego, tzn. czy dopasowanie ma w ogóle sens możemy wykorzystać statystykę F definiując nową wielkość

$$F_R = \frac{R^2/m}{(1-R^2)/(n-m-1)}$$

Testowanie wartości F_R jest w istocie testem, że wszystkie współczynniki a_j są różne od zera, czyli są znaczące w dopasowywanym modelu. Jeżeli wartość F_R nie przekracza granicznej wartości statystyki F, to oznacza to, że parametrów modelu jest za dużo (przynajmniej jeden ze współczynników powinien wynosić zero).

Testowanie zasadności dodatkowego parametru modelu

Jeżeli dopasujemy do n danych punktów funkcję o m parametrach, to pozostała suma kwadratów (reszt dopasowania) $\chi^2(m)$ ma rozkład o $n-m$ stopniach swobody. Po zwiększeniu liczby parametrów modelu (np. dodając kolejny składnik wielomianu) otrzymujemy sumę kwadratów reszt dopasowania $\chi^2(m+1)$ o $n-m-1$ stopniach swobody. Różnica $\chi^2(m)-\chi^2(m+1)$ ma zatem rozkład χ^2 o 1 stopniu swobody. Do sprawdzenia czy zmniejszenie sumy kwadratów w wyniku dodania no-

wego parametru jest statystycznie istotne możemy wykorzystać test F. Wielkość

$$F_\chi = \frac{\chi^2(m) - \chi^2(m+1)}{\chi^2(m+1)/(n-m-1)} = \frac{\Delta\chi^2}{\chi^2_\nu}$$

ma rozkład F z $\nu_1 = 1$ i $\nu_2 = n-m-1$.

Stosunek F_χ mierzy jak bardzo wprowadzenie nowego parametru poprawiło wartość zredukowanej χ^2_ν i będzie mały, jeżeli zmiana nie jest istotna. Podobnie jak w przypadku wielkości F_R, testujemy czy nowy parametr jest równy zero. Jeżeli wartość F_χ przekracza graniczną, wprowadzenie nowego parametru możemy uznać za uzasadnione.

Przykład

i	x_i	y_i	$u(y_i)$	w_i	$w_i(y_i-y_{sr})^2$
1	0,00	1,143	0,2	25	8,14
2	0,20	1,241	0,2	25	11,18
3	0,40	1,442	0,2	25	18,93
4	0,60	1,504	0,2	25	21,73
5	0,80	1,725	0,2	25	33,25
6	1,00	1,614	0,2	25	27,12
7	1,20	1,389	0,2	25	16,68
8	1,40	1,217	0,2	25	10,39
9	1,60	1,077	0,2	25	6,37
10	1,80	0,754	0,2	25	0,83
11	2,00	0,450	0,2	25	0,38
12	2,20	0,067	0,2	25	6,37
13	2,40	-0,351	0,2	25	21,30
14	2,60	-0,278	0,2	25	18,05
15	2,80	-0,439	0,2	25	25,58
16	3,00	-0,757	0,2	25	44,15
17	3,20	-0,840	0,2	25	49,83
18	3,40	-0,401	0,2	25	23,68
19	3,60	-0,606	0,2	25	34,71
20	3,80	-0,585	0,2	25	33,45
21	4,00	0,430	0,2	25	0,51
22	4,20	0,192	0,2	25	3,60
23	4,40	0,530	0,2	25	0,04
24	4,60	1,403	0,2	25	17,24
25	4,80	2,382	0,2	25	81,86

Przykład do „Testowanie jakości dopasowania. Część 2.”

Model

$Y(X) = A_0 + A_1 X + A_2 X^2 + A_3 X^3 + A_4 X^4 + \ldots$

$A_0 = 1{,}000$

$A_1 = 1{,}000$

$A_2 = -0{,}500$

$A_3 = -0{,}167$

$A_4 = 0{,}042$

$A_5 = 8{,}33 \cdot 10^{-3}$

$A_6 = -1{,}39 \cdot 10^{-3}$

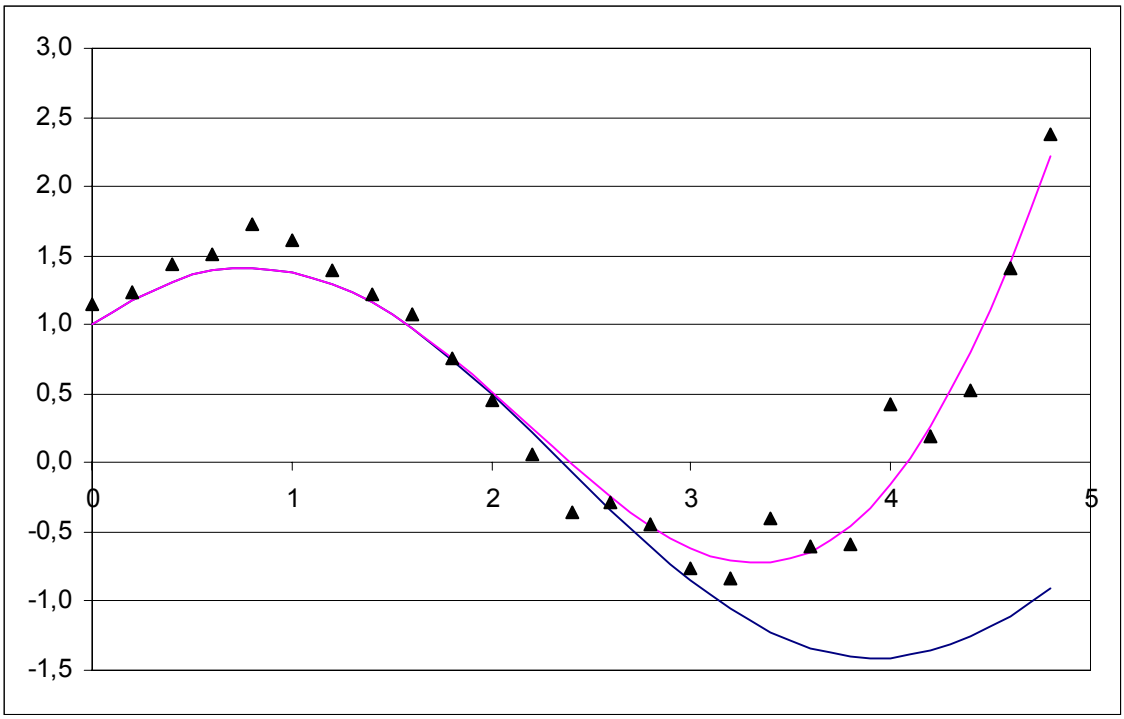
$A_7 = -1{,}98 \cdot 10^{-4}$

$\Delta Y = N(0,\ 0{,}2)$

Wygenerowane dane

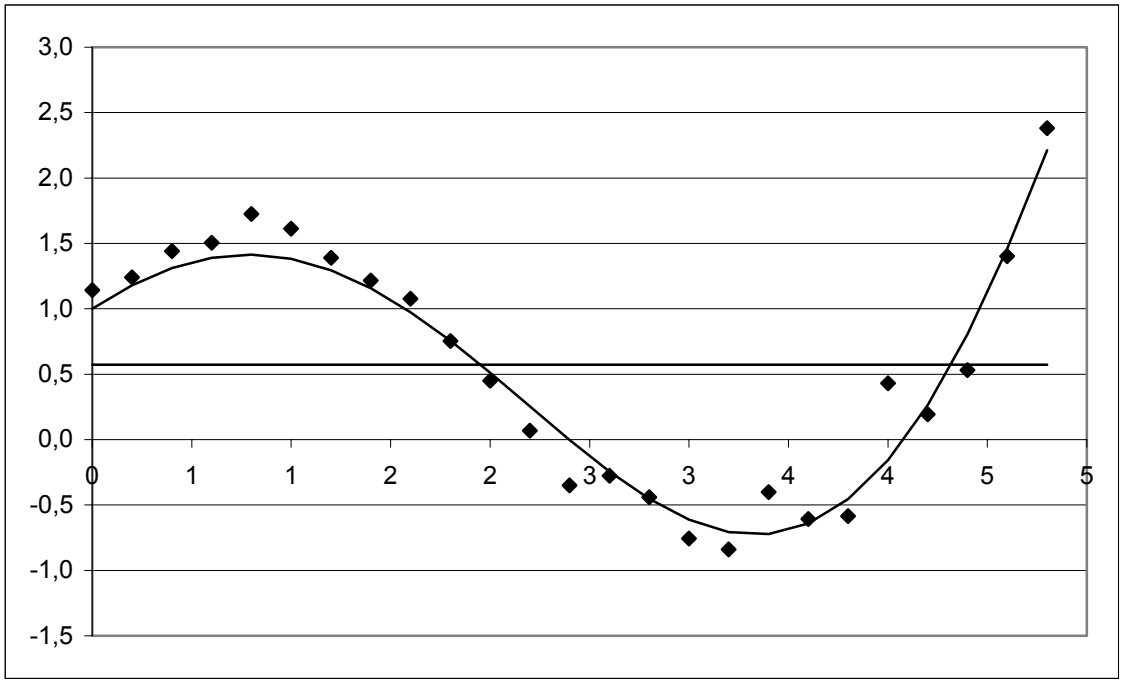
i	x_i	y_i	$u(y_i)$	w_i	$w_i(y_i-y_{sr})^2$
1	0,00	1,143	0,2	25	8,14
2	0,20	1,241	0,2	25	11,18
3	0,40	1,442	0,2	25	18,93
4	0,60	1,504	0,2	25	21,73
5	0,80	1,725	0,2	25	33,25
6	1,00	1,614	0,2	25	27,12
7	1,20	1,389	0,2	25	16,68
8	1,40	1,217	0,2	25	10,39
9	1,60	1,077	0,2	25	6,37
10	1,80	0,754	0,2	25	0,83
11	2,00	0,450	0,2	25	0,38
12	2,20	0,067	0,2	25	6,37
13	2,40	-0,351	0,2	25	21,30
14	2,60	-0,278	0,2	25	18,05
15	2,80	-0,439	0,2	25	25,58
16	3,00	-0,757	0,2	25	44,15
17	3,20	-0,840	0,2	25	49,83
18	3,40	-0,401	0,2	25	23,68
19	3,60	-0,606	0,2	25	34,71
20	3,80	-0,585	0,2	25	33,45
21	4,00	0,430	0,2	25	0,51
22	4,20	0,192	0,2	25	3,60
23	4,40	0,530	0,2	25	0,04
24	4,60	1,403	0,2	25	17,24
25	4,80	2,382	0,2	25	81,86

$y_{sr} = 0{,}572$ $\Sigma = 515{,}36$


3,0
2,5
2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
0
1
2
3
4
5

m = 1

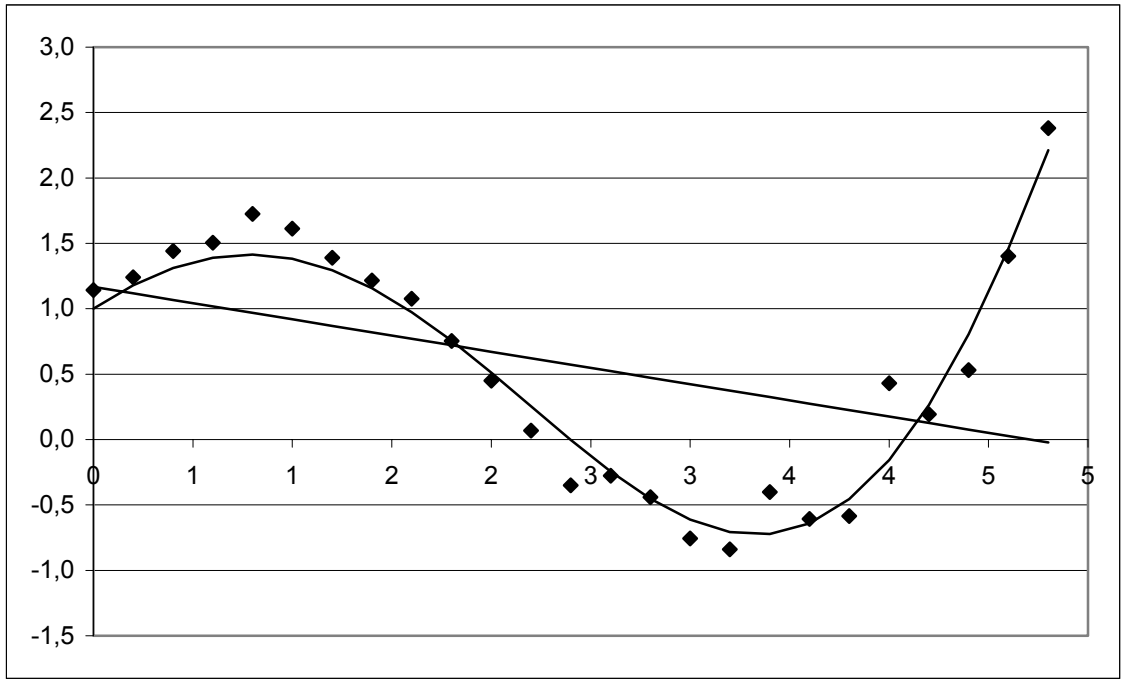
$y(x_i)$	$w_i(y_i-y(x_i))^2$
0,572	8,140
0,572	11,180
0,572	18,929
0,572	21,729
0,572	33,250
0,572	27,121
0,572	16,676
0,572	10,388
0,572	6,365
0,572	0,830
0,572	0,375
0,572	6,370
0,572	21,302
0,572	18,048
0,572	25,583
0,572	44,155
0,572	49,827
0,572	23,677
0,572	34,708
0,572	33,453
0,572	0,506
0,572	3,605
0,572	0,044
0,572	17,242
0,572	81,858
$\Sigma =$	515,3618



$\chi^2 = 515{,}36$

m = 2

$y(x_i)$	$w_i(y_i-y(x_i))^2$
1,167	0,014
1,117	0,382
1,068	3,509
1,018	5,913
0,969	14,320
0,919	12,065
0,869	6,745
0,820	3,937
0,770	2,347
0,721	0,028
0,671	1,228
0,622	7,682
0,572	21,302
0,523	16,004
0,473	20,816
0,423	34,829
0,374	36,818
0,324	13,155
0,275	19,402
0,225	16,398
0,176	1,615
0,126	0,110
0,077	5,139
0,027	47,303
-0,022	144,497
$\Sigma =$	435,5589

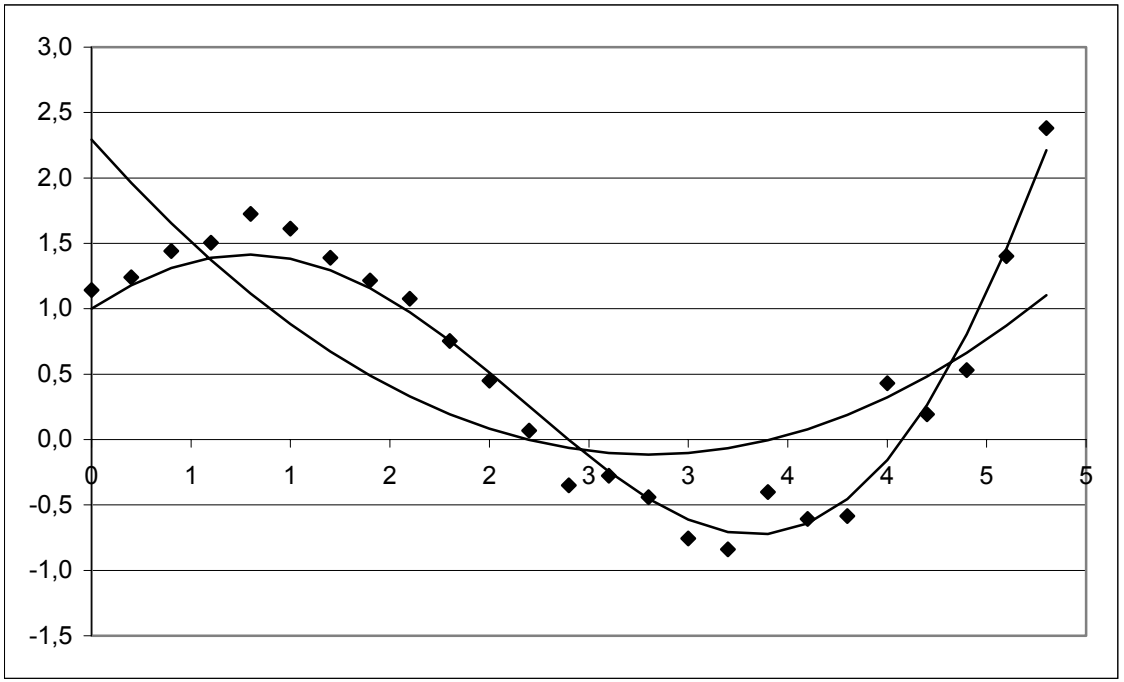


$\chi^2 = 435{,}56 \quad \Delta\chi^2 = 79{,}80 \quad \frac{\chi^2}{n-m} = 18{,}937 \quad F_\chi = 4{,}21 \quad P(F \geq F_\chi) = 0{,}0516$

m = 3

$y(x_i)$	$w_i(y_i-y(x_i))^2$
2,293	33,072
1,962	12,995
1,655	1,133
1,373	0,431
1,115	9,301
0,882	13,374
0,674	12,790
0,489	13,226
0,330	13,952
0,194	7,839
0,084	3,347
-0,003	0,122
-0,064	2,053
-0,102	0,773
-0,115	2,640
-0,103	10,692
-0,067	14,934
-0,006	3,899
0,079	11,734
0,189	14,945
0,323	0,288
0,481	2,084
0,664	0,450
0,872	7,047
1,104	40,831

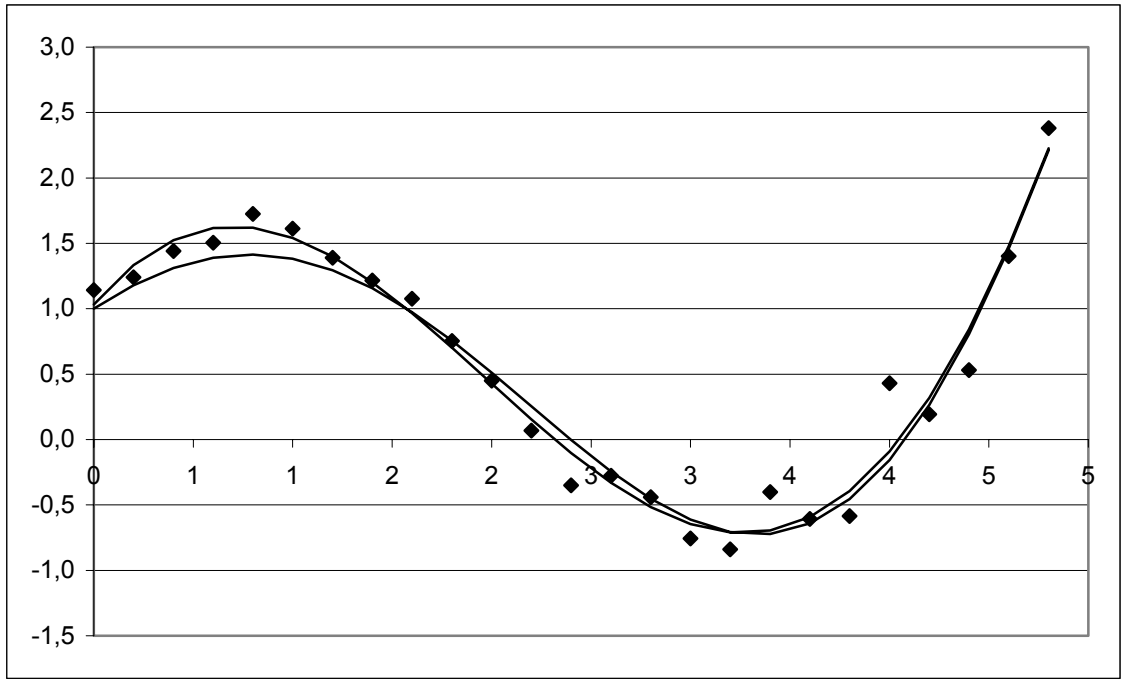
$\Sigma =$ 233,9511



$\chi^2 = 233,95 \quad \Delta\chi^2 = 201,61 \quad \frac{\chi^2}{n-m} = 10,634 \quad F_\chi = 18,96 \quad P(F \geq F_\chi) = 0,0003$

m = 4

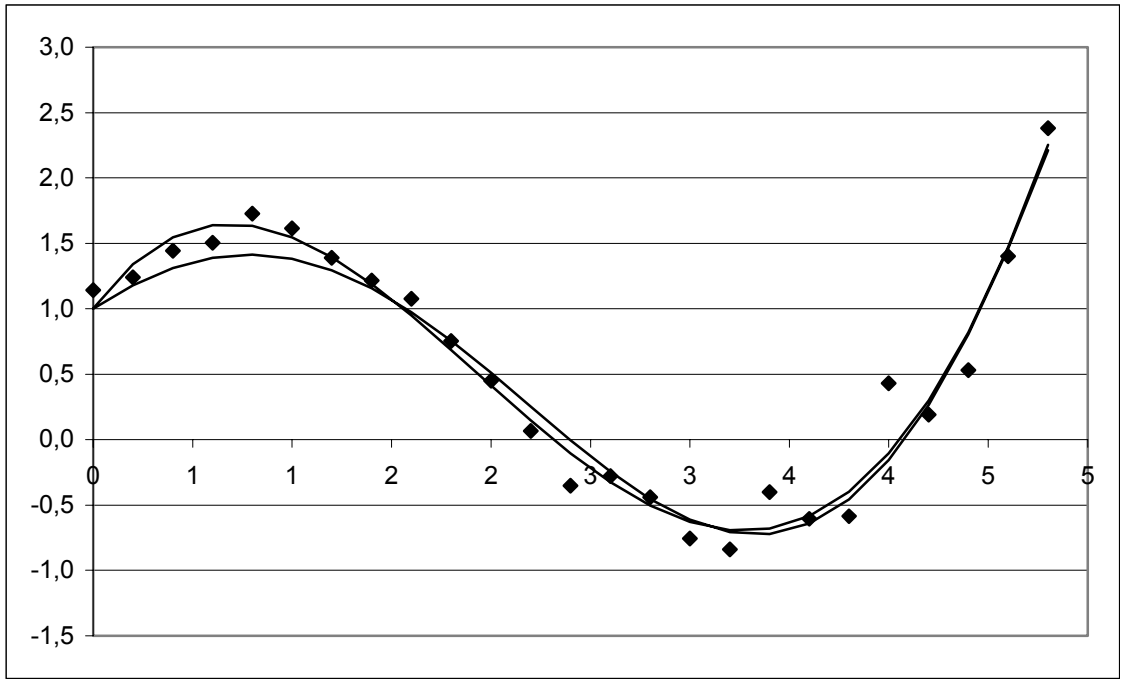
$y(x_i)$	$w_i(y_i-y(x_i))^2$
1,073	0,122
1,352	0,308
1,523	0,161
1,597	0,216
1,588	0,471
1,507	0,286
1,366	0,013
1,177	0,040
0,952	0,390
0,703	0,065
0,443	0,001
0,183	0,335
-0,064	2,053
-0,287	0,002
-0,474	0,030
-0,612	0,527
-0,689	0,568
-0,693	2,136
-0,613	0,001
-0,436	0,552
-0,150	8,408
0,257	0,104
0,797	1,779
1,482	0,156
2,324	0,084
$\Sigma =$	18,81033



$\chi^2 = 18{,}810 \quad \Delta\chi^2 = 215{,}14 \quad \frac{\chi^2}{n-m} = 0{,}896 \quad F_\chi = 240{,}18 \quad P(F \geq F_\chi) = 5{,}72 \cdot 10^{-13}$

m = 5

$y(x_i)$	$w_i(y_i-y(x_i))^2$
1,004	0,484
1,340	0,247
1,546	0,270
1,638	0,448
1,633	0,214
1,546	0,115
1,393	0,000
1,189	0,020
0,948	0,413
0,685	0,119
0,414	0,032
0,146	0,156
-0,103	1,531
-0,324	0,054
-0,503	0,101
-0,630	0,405
-0,692	0,542
-0,681	1,962
-0,586	0,010
-0,397	0,881
-0,105	7,160
0,298	0,278
0,820	2,108
1,470	0,114
2,255	0,404



$\Sigma =$ 18,06697

$\chi^2 = 18{,}067 \quad \Delta\chi^2 = 0{,}74 \quad \frac{\chi^2}{n-m} = 0{,}903 \quad F_\chi = 0{,}82 \quad P(F \geq F_\chi) = 0{,}375$