Rozdział 8

Regresja

Definiowanie modelu

Analizę korelacji można traktować jako wstęp do analizy regresji. Jeżeli wykresy rozrzutu oraz wartości współczynników korelacji wskazują na istniejąca współzmienność między analizowanymi zmiennymi można pokusić się o skwantyfikowanie tej relacji.

Jeżeli zależność między zmiennymi ma charakter liniowy (dodatkowo zmienne mają rozkład normalny oraz nie posiadają obserwacji odstających) można pokusić sie o dopasowanie funkcji f która powiąże zmienną wyjaśnianą y (zwana również predyktantą , lub zmienną zależną) ze zmienną niezależną x (predyktor, zmienna wyjaśniająca).

$$\hat{y} = f(x)$$

Postać funkcji f w przypadku liniowości można zapisać w sposób następujący;

$$\hat{y} = a + bx$$

W przypadku zależności nieliniowych może to wskazywać na konieczność użycia funkcji wyższego rzędu np. $\hat{y} = a+bx+bx^2$ lub $\hat{y} = ae^{bx}$ Zawsze staramy się poszukiwać funkcji w najprostszej postaci, ponieważ dla skomplikowanego równania trudno znaleźć odpowiednią interpretację fizyczną.

Dopasowanie parametrów linii regresji dokonujemy metodą najmniejszych kwadratów (istnieją również inne metody - ta jest jedną z bardziej oczywistych), zakładając takie dobranie współczynników, aby wartość błędu S była najmniejsza

$$S = \sum_{i=1}^{n}[y_i - f(x_i)]^2, S = S(a,b) = \sum_{i=1}^{n}[y_i - a - bx_i]^2$$

Wartości współczynników a i b otrzymujemy, wykorzystując poniższe wzory.

$$b = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_1 \sum_{i=1}^{n} x_1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} xi\right)^2} \quad a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

Analiza jakości modelu

SST zmienność całkowita	=	SSR zmienność wyjaśniana równaniem regresji	+	SSE zmienność pozostała poza regresją

gdzie: SS-sum of squares, T-total, R-regression, E-error

Można to również zapisać jako

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	Suma kwadratów	**Stopnie swobody**	**Wartość średnia kwadratów**
Regresja	SSR	$k - 1 = 1$	$MSR = \frac{SSR}{1}$
Resztkowy	SSE	$n - k = n - 2$	$MSE = \frac{SSE}{n-2}$
Razem	SST	$n - 1$	

Wartość średnia kwadratów reszt (wariancja składnika losowego) **MSE** mówi o zgodności z danymi obserwowanymi w modelu. (informuje o zmienności składnika losowego)

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k} = \frac{\sum_{i=1}^{n} e_i^2}{n-k} = \frac{SSE}{n-k}$$

Odchylenie standardowe reszt (**RMSE** - standardowy błąd estymacji) informuje o ile średnio wartości obserwowane Y odchylają się od wartości prognozowanych w modelu.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k}}$$

Do oceny istotności statystycznej współczynnika kierunkowego linii regresji (b) stosujemy test t przy $H_0 : b = 0$ i hipotezie alternatywnej $b \neq 0$.

$$t = \frac{b}{S_e}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}}$$

Do oceny jaka część wariancji zmiennej określona jest poprzez równanie trendu wykorzystujemy test *F-Snedecora*.

$$F = \frac{\frac{\sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2}{k-1}}{\frac{\sum_{i=1}^{n}[\hat{y}_i - y_i]^2}{n-k}} = \frac{MSR}{MSE}$$

Używany jest również współczynnik determinacji R^2, który określa stosunek zmienności wynikającej z trendu do zmienności całkowitej.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST}$$

Wyznaczanie przedziałów ufności dla linii trendu, gdzie s^2 (MSE) – oszacowanie wariancji resztowej przy $n-2$ stopniach swobody $s = \mathbf{RMSE}$, $t_{(\alpha,n-2)}$ wartość krytyczna rozkładu t przy poziomie ufności α i $n-2$ stopniach swobody.

Górne oszacowanie

$$\eta_1(x) = \hat{y} + t_{\alpha,n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

oraz dolne

$$\eta_2(x) = \hat{y} - t_{\alpha,n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Zadanie

Korzystając z tablicy zamieszczonej poniżej narysuj wykres rozrzutu a następnie oblicz współczynniki równania regresji liniowej.

Zweryfikuj hipotezę o istotności współczynnika kierunkowego jak również całości modelu. Oblicz R^2, **RMSE** oraz **przedziały ufności** dla linii regresji.

Lp	x	y	$x \cdot y$	x^2	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$(\hat{y} - \bar{y})^2$	$(y - \bar{y})^2$	$(x - \bar{x})^2$	η_1	η_2
1	5,1	3,7										
2	1,2	-1,4										
3	6,9	3,7										
4	1,7	0,5										
5	0,7	-1,4										
6	1,1	-1,0										
7	8,2	3,7										
8	2,6	3,4										
9	7,1	4,1										
10	3,5	1,1										
11	8,6	4,9										
12	0,5	2,3										
13	5,2	2,2										
14	6,7	5,1										
15	6,0	5,4										
16	0,9	0,6										
17	6,1	2,5										
18	3,7	0,4										
19	6,9	5,1										
20	8,3	6,1										
$\sum$												

Rozkład F-Snedecora

Tablice podają wartości kwantyla rzędu 0.95 centralnego rozkładu F Snedecora z n_1 i n_2 stopniami swobody. Wiersze odpowiadają wartościom n_1, kolumny wartościom n_2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	161,45	18,51	10,13	7,71	6,61	5,99	5,59	5,32	5,12	4,96	4,84	4,75	4,67	4,6	4,54	**1**
2	199,5	19	9,55	6,94	5,79	5,14	4,74	4,46	4,26	4,1	3,98	3,89	3,81	3,74	3,68	**2**
3	215,71	19,16	9,28	6,59	5,41	4,76	4,35	4,07	3,86	3,71	3,59	3,49	3,41	3,34	3,29	**3**
4	224,58	19,25	9,12	6,39	5,19	4,53	4,12	3,84	3,63	3,48	3,36	3,26	3,18	3,11	3,06	**4**
5	230,16	19,3	9,01	6,26	5,05	4,39	3,97	3,69	3,48	3,33	3,2	3,11	3,03	2,96	2,9	**5**
6	233,99	19,33	8,94	6,16	4,95	4,28	3,87	3,58	3,37	3,22	3,09	3	2,92	2,85	2,79	**6**
7	236,77	19,35	8,89	6,09	4,88	4,21	3,79	3,5	3,29	3,14	3,01	2,91	2,83	2,76	2,71	**7**
8	238,88	19,37	8,85	6,04	4,82	4,15	3,73	3,44	3,23	3,07	2,95	2,85	2,77	2,7	2,64	**8**
9	240,54	19,38	8,81	6	4,77	4,1	3,68	3,39	3,18	3,02	2,9	2,8	2,71	2,65	2,59	**9**
10	241,88	19,4	8,79	5,96	4,74	4,06	3,64	3,35	3,14	2,98	2,85	2,75	2,67	2,6	2,54	**10**
11	242,98	19,4	8,76	5,94	4,7	4,03	3,6	3,31	3,1	2,94	2,82	2,72	2,63	2,57	2,51	**11**
12	243,91	19,41	8,74	5,91	4,68	4	3,57	3,28	3,07	2,91	2,79	2,69	2,6	2,53	2,48	**12**
13	244,69	19,42	8,73	5,89	4,66	3,98	3,55	3,26	3,05	2,89	2,76	2,66	2,58	2,51	2,45	**13**
14	245,36	19,42	8,71	5,87	4,64	3,96	3,53	3,24	3,03	2,86	2,74	2,64	2,55	2,48	2,42	**14**
15	245,95	19,43	8,7	5,86	4,62	3,94	3,51	3,22	3,01	2,85	2,72	2,62	2,53	2,46	2,4	**15**
16	246,46	19,43	8,69	5,84	4,6	3,92	3,49	3,2	2,99	2,83	2,7	2,6	2,51	2,44	2,38	**16**
17	246,92	19,44	8,68	5,83	4,59	3,91	3,48	3,19	2,97	2,81	2,69	2,58	2,5	2,43	2,37	**17**
18	247,32	19,44	8,67	5,82	4,58	3,9	3,47	3,17	2,96	2,8	2,67	2,57	2,48	2,41	2,35	**18**
19	247,69	19,44	8,67	5,81	4,57	3,88	3,46	3,16	2,95	2,79	2,66	2,56	2,47	2,4	2,34	**19**
20	248,01	19,45	8,66	5,8	4,56	3,87	3,44	3,15	2,94	2,77	2,65	2,54	2,46	2,39	2,33	**20**
21	248,31	19,45	8,65	5,79	4,55	3,86	3,43	3,14	2,93	2,76	2,64	2,53	2,45	2,38	2,32	**21**
22	248,58	19,45	8,65	5,79	4,54	3,86	3,43	3,13	2,92	2,75	2,63	2,52	2,44	2,37	2,31	**22**
23	248,83	19,45	8,64	5,78	4,53	3,85	3,42	3,12	2,91	2,75	2,62	2,51	2,43	2,36	2,3	**23**
24	249,05	19,45	8,64	5,77	4,53	3,84	3,41	3,12	2,9	2,74	2,61	2,51	2,42	2,35	2,29	**24**
25	249,26	19,46	8,63	5,77	4,52	3,83	3,4	3,11	2,89	2,73	2,6	2,5	2,41	2,34	2,28	**25**
26	249,45	19,46	8,63	5,76	4,52	3,83	3,4	3,1	2,89	2,72	2,59	2,49	2,41	2,33	2,27	**26**
27	249,63	19,46	8,63	5,76	4,51	3,82	3,39	3,1	2,88	2,72	2,59	2,48	2,4	2,33	2,27	**27**
28	249,8	19,46	8,62	5,75	4,5	3,82	3,39	3,09	2,87	2,71	2,58	2,48	2,39	2,32	2,26	**28**
29	249,95	19,46	8,62	5,75	4,5	3,81	3,38	3,08	2,87	2,7	2,58	2,47	2,39	2,31	2,25	**29**
30	250,1	19,46	8,62	5,75	4,5	3,81	3,38	3,08	2,86	2,7	2,57	2,47	2,38	2,31	2,25	**30**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
1	4,49	4,45	4,41	4,38	4,35	4,32	4,3	4,28	4,26	4,24	4,23	4,21	4,2	4,18	4,17	**1**
2	3,63	3,59	3,55	3,52	3,49	3,47	3,44	3,42	3,4	3,39	3,37	3,35	3,34	3,33	3,32	**2**
3	3,24	3,2	3,16	3,13	3,1	3,07	3,05	3,03	3,01	2,99	2,98	2,96	2,95	2,93	2,92	**3**
4	3,01	2,96	2,93	2,9	2,87	2,84	2,82	2,8	2,78	2,76	2,74	2,73	2,71	2,7	2,69	**4**
5	2,85	2,81	2,77	2,74	2,71	2,68	2,66	2,64	2,62	2,6	2,59	2,57	2,56	2,55	2,53	**5**
6	2,74	2,7	2,66	2,63	2,6	2,57	2,55	2,53	2,51	2,49	2,47	2,46	2,45	2,43	2,42	**6**
7	2,66	2,61	2,58	2,54	2,51	2,49	2,46	2,44	2,42	2,4	2,39	2,37	2,36	2,35	2,33	**7**
8	2,59	2,55	2,51	2,48	2,45	2,42	2,4	2,37	2,36	2,34	2,32	2,31	2,29	2,28	2,27	**8**
9	2,54	2,49	2,46	2,42	2,39	2,37	2,34	2,32	2,3	2,28	2,27	2,25	2,24	2,22	2,21	**9**
10	2,49	2,45	2,41	2,38	2,35	2,32	2,3	2,27	2,25	2,24	2,22	2,2	2,19	2,18	2,16	**10**
11	2,46	2,41	2,37	2,34	2,31	2,28	2,26	2,24	2,22	2,2	2,18	2,17	2,15	2,14	2,13	**11**
12	2,42	2,38	2,34	2,31	2,28	2,25	2,23	2,2	2,18	2,16	2,15	2,13	2,12	2,1	2,09	**12**
13	2,4	2,35	2,31	2,28	2,25	2,22	2,2	2,18	2,15	2,14	2,12	2,1	2,09	2,08	2,06	**13**
14	2,37	2,33	2,29	2,26	2,22	2,2	2,17	2,15	2,13	2,11	2,09	2,08	2,06	2,05	2,04	**14**
15	2,35	2,31	2,27	2,23	2,2	2,18	2,15	2,13	2,11	2,09	2,07	2,06	2,04	2,03	2,01	**15**
16	2,33	2,29	2,25	2,21	2,18	2,16	2,13	2,11	2,09	2,07	2,05	2,04	2,02	2,01	1,99	**16**
17	2,32	2,27	2,23	2,2	2,17	2,14	2,11	2,09	2,07	2,05	2,03	2,02	2	1,99	1,98	**17**
18	2,3	2,26	2,22	2,18	2,15	2,12	2,1	2,08	2,05	2,04	2,02	2	1,99	1,97	1,96	**18**
19	2,29	2,24	2,2	2,17	2,14	2,11	2,08	2,06	2,04	2,02	2	1,99	1,97	1,96	1,95	**19**
20	2,28	2,23	2,19	2,16	2,12	2,1	2,07	2,05	2,03	2,01	1,99	1,97	1,96	1,94	1,93	**20**
21	2,26	2,22	2,18	2,14	2,11	2,08	2,06	2,04	2,01	2	1,98	1,96	1,95	1,93	1,92	**21**
22	2,25	2,21	2,17	2,13	2,1	2,07	2,05	2,02	2	1,98	1,97	1,95	1,93	1,92	1,91	**22**
23	2,24	2,2	2,16	2,12	2,09	2,06	2,04	2,01	1,99	1,97	1,96	1,94	1,92	1,91	1,9	**23**
24	2,24	2,19	2,15	2,11	2,08	2,05	2,03	2,01	1,98	1,96	1,95	1,93	1,91	1,9	1,89	**24**
25	2,23	2,18	2,14	2,11	2,07	2,05	2,02	2	1,97	1,96	1,94	1,92	1,91	1,89	1,88	**25**
26	2,22	2,17	2,13	2,1	2,07	2,04	2,01	1,99	1,97	1,95	1,93	1,91	1,9	1,88	1,87	**26**
27	2,21	2,17	2,13	2,09	2,06	2,03	2	1,98	1,96	1,94	1,92	1,9	1,89	1,88	1,86	**27**
28	2,21	2,16	2,12	2,08	2,05	2,02	2	1,97	1,95	1,93	1,91	1,9	1,88	1,87	1,85	**28**
29	2,2	2,15	2,11	2,08	2,05	2,02	1,99	1,97	1,95	1,93	1,91	1,89	1,88	1,86	1,85	**29**
30	2,19	2,15	2,11	2,07	2,04	2,01	1,98	1,96	1,94	1,92	1,9	1,88	1,87	1,85	1,84	**30**