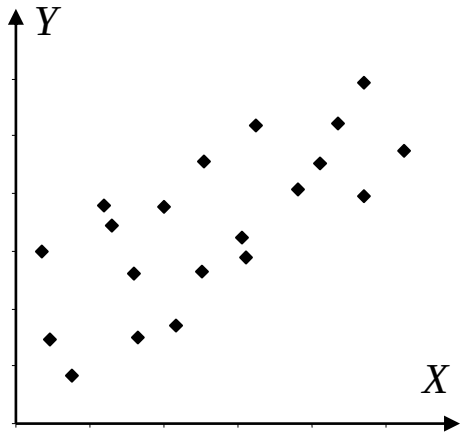
8. ANALIZA KORELACJI I REGRESJI

Do sprawdzenia czy wybrane cechy mierzalne populacji generalnej (dwie lub więcej) są ze sobą związane wykorzystywane są korelacja i regresja. Korelacja zajmuje się siłą tej zależności, regresja skupia się na jej kształcie. Równania regresji opisujące związek funkcyjny badanych cech powinny być wyznaczane gdy korelacja pomiędzy tymi cechami nie jest zbyt słaba.

W przypadku badania związku dwóch cech analizę zależności rozpoczyna się zwykle od sporządzenia *wykresu rozrzutu*. Poszczególne punkty wykresu odpowiadają wartościom uzyskanym z kolejnych obserwacji zmiennych analizowanych zmiennych.



8.1. Analiza korelacji

8.1.1. Miary zależności

Siła związku analizowanych cech może być opisywana na kilka różnych sposobów. W przypadku liniowej zależności dwóch cech siłę ich związku opisuje się za pomocą ***współczynnika kowariancji*** lub ***współczynnika korelacji liniowej Paersona***, w przypadku nieliniowej zależności cech wykorzystywany jest wskaźnik nazywany ***stosunkiem korelacyjnym***.

Kowariancja zmiennych X i Y jest oznaczana jako $\mathrm{cov}(X,Y)$.

- jeżeli każda z wartości cechy X może pojawić się z każdą wartością cechy Y to zmienne są niezależne a $\mathrm{cov}(X,Y)=0$,
- jeżeli wartości cechy X większe od średniej tej cechy pojawiają się najczęściej z wartościami cechy Y większymi od jej średniej to $\mathrm{cov}(X,Y)>0$,
- jeżeli wartości cechy X większe od średniej tej cechy pojawiają się najczęściej z wartościami cechy Y mniejszymi od jej średniej to $\mathrm{cov}(X,Y)<0$.

Wartość kowariancji zależy od wariancji zmiennych X i Y, w celu uniezależnienia miary zależności od wariancji wprowadzono standaryzowaną kowariancję, tzw. ***współczynnik korelacji liniowej Paersona***:

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \, \sigma_Y}$$

gdzie: σ_X, σ_Y – odchylenie standardowe zmiennej X i Y.

Współczynnik korelacji jest miarą unormowaną i przyjmuje wartości:

$$-1 \le \rho \le 1$$

Jeżeli $|\rho| = 1$ to pomiędzy zmiennymi X i Y istnieje ścisła zależność liniowa, gdy $\rho = 0$ zmienne nie są skorelowane, im wartość $|\rho|$ jest bliższa 1 tym korelacja zmiennych jest silniejsza.

Empiryczna kowariancja i ***korelacja*** wyznaczane są z próby:

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)$$

$$r = \frac{\operatorname{cov}(X,Y)}{s_X\, s_Y}$$

gdzie: x_i, y_i – wartości zmiennych X i Y, $\bar{x}$, $\bar{y}$, s_X, s_y – średnie i odchylenia standardowe zmiennych X i Y; dla małych prób, podobnie jak w przypadku wzorów na odchylenia standardowe s_X, s_y stosowane są wzory z dzieleniem przez $(n-1)$.

Współczynniki te zapisywane są również jako:

$$\operatorname{cov}(X,Y) = \overline{xy} - \bar{x}\,\bar{y} \quad (*),$$

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_X\, s_Y} \quad \text{lub} \quad r = \frac{\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - y\right)^2}},$$

gdzie: $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

(*) Wzór $\operatorname{cov}(X,Y) = \overline{xy} - \bar{x}\,\bar{y}$ otrzymuje się po przekształceniach:

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\,\bar{y}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\bar{x}\,\bar{y}\right)$$

$$\operatorname{cov}(X,Y) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{y}\,\bar{x} - n\,\bar{x}\,\bar{y} + n\,\bar{x}\,\bar{y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} = \overline{xy} - \bar{x}\,\bar{y}.$$

W przypadku badania wpływu wielu zmiennych niezależnych na pewną zmienną zależną, przy założeniu istnienia liniowego związku pomiędzy badanymi zmiennymi, miary zależności: kowariancja i korelacja są uogólniane i zapisywane w postaci ***macierzy kowariancji*** i ***macierzy korelacji***.

Macierz kowariancji jest uogólnioną na *n* wymiarów wariancją, zawiera informacje o sile związku każdej pary zmiennych nie uwzględniając wpływu pozostałych zmiennych.

Macierz korelacji jest standaryzowaną *macierzą kowariancji*, jej elementami są współczynniki korelacji liniowej Paersona wyznaczone dla każdej pary zmiennych.

Macierz kowariancji:

$$\begin{bmatrix} \sigma_1^2 & \operatorname{cov}(X_1,X_2) & \dots & \operatorname{cov}(X_1,X_n) \\ \operatorname{cov}(X_2,X_1) & \sigma_2^2 & \dots & \operatorname{cov}(X_2,X_n) \\ \dots & \dots & \dots & \dots \\ \operatorname{cov}(X_n,X_1) & \operatorname{cov}(X_n,X_2) & \dots & \sigma_n^2 \end{bmatrix}$$

Macierz korelacj:i

$$\begin{bmatrix} 1 & \frac{\operatorname{cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}} & \dots & \frac{\operatorname{cov}(X_1,X_n)}{\sigma_{X_1}\sigma_{X_n}} \\ \frac{\operatorname{cov}(X_2,X_1)}{\sigma_{X_2}\sigma_{X_1}} & 1 & \dots & \frac{\operatorname{cov}(X_2,X_n)}{\sigma_{X_2}\sigma_{X_n}} \\ \dots & \dots & \dots & \dots \\ \frac{\operatorname{cov}(X_n,X_1)}{\sigma_{X_n}\sigma_{X_1}} & \frac{\operatorname{cov}(X_n,X_2)}{\sigma_{X_n}\sigma_{X_2}} & \dots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho(X_1,X_2) & \dots & \rho(X_1,X_n) \\ \rho(X_2,X_1) & 1 & \dots & \rho(X_2,X_n) \\ \dots & \dots & \dots & \dots \\ \rho(X_n,X_1) & \rho(X_n,X_2) & \dots & 1 \end{bmatrix}$$

gdzie: $X_1, X_2, \dots, X_n$ – badane cechy, σ_i – wariancja zmiennej X_i; $\operatorname{cov}(X_i, X_j)$ – kowariancja pomiędzy parą zmiennych X_i i X_j, $\rho(X_i, X_j)$ – współczynnik korelacji zmiennych X_i i X_j, $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$ i $\rho(X_i, X_j) = \rho(X_j, X_i)$ – macierze kowariancji i korelacji są więc symetryczne.

Przykład 1.

Podczas badania zależności dwóch wymiarów pewnego detalu otrzymano z próby n=20 następujące wyniki:

x_i	3.3	2.4	3.2	10.5	5.1	2.6	1.5	9.4	6.2	0.7	4	8.6	9.4	6.1	4.3	0.9	8.2	6.5	5	8.7
y_i	15.1	38	26	31.6	45.6	34.4	8.5	26.8	28.9	30	53.4	40.6	69.4	32.3	18.1	14.6	45.2	51.9	26.6	52.3

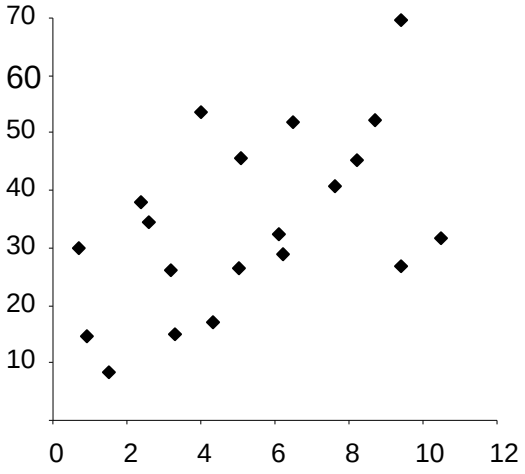
Narysować wykres rozrzutu, wyznaczyć współczynnik korelacji.

$\bar{x} \approx 0{,}5669$, $s_x \approx 2{,}9973$,

$\bar{y} \approx 34{,}415$, $s_y \approx 15{,}3756$,

$\operatorname{cov}(X, Y) \approx 24{,}3619$,

$r \approx 0{,}5286$



8.1.2. Istotność współczynnika korelacji

Skrajne wartości współczynnika korelacji jednoznacznie wskazują na istnienie ścisłej zależności liniowej ($r = \pm 1$) lub braku korelacji ($r = 0$) pomiędzy badanymi cechami. W przypadkach pośrednich, sprawdzenie czy analizowane cechy są skorelowane przeprowadza się testując hipotezę H_0: $\rho = 0$ (cechy nie są skorelowane) wobec hipotezy alternatywnej H_1: $\rho \neq 0$ (istnieje korelacja). Korelację uznaje się za istotną jeżeli test wykaże, że hipotezę H_0 należy odrzucić, tzn. wartość statystyki testowej leży w obszarze krytycznym, lub graniczny poziom istotności jest mniejszy od założonego.
Jeżeli badane cechy mają rozkład normalny lub zbliżony do normalnego to do zbadania istotności ich korelacji wykorzystuje się statystykę testową:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}\,,$$

gdzie: n to rozmiar próby a r to współczynnik korelacji wyznaczony na jej podstawie.
Przy założeniu prawdziwości hipotezy H_0 statystyka ta ma rozkład t – *Studenta* o (n–2) stopniach swobody.

Przykład 2.

Na poziomie istotności $\alpha = 0{,}05$ zbadać istotność współczynnika korelacji z przykładu 1.

Obliczona na podstawie wyników z próby wartość statystyki testowej wynosi:

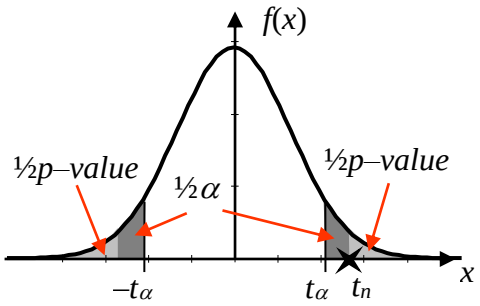
$$t_n = \frac{0{,}5286}{\sqrt{1-0{,}5286^2}}\sqrt{20-2} \approx 2{,}6421\,,$$

Graniczny poziom istotności otrzymuje się jako:

$$\frac{1}{2}\,p\text{–}value = F_{t(18)}(-t_n) = F_{t(18)}(-2{,}6421) \approx 0{,}0083\,, \rightarrow\ p\text{–}value \approx 0{,}0166\,,$$

a granica obszaru krytycznego wynosi w tym przypadku:

$$t_\alpha = -F^{-1}_{t(18)}\left(\frac{0{,}05}{2}\right) \approx 2{,}1009\,.$$



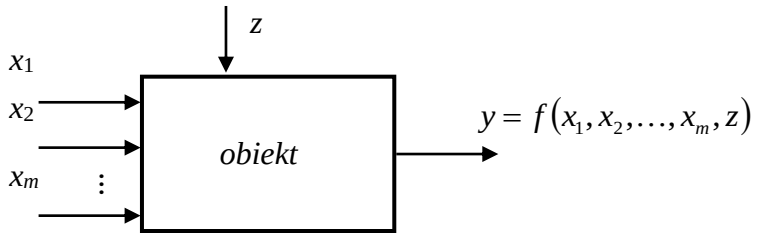
Hipotezę H_0 należy odrzucić na korzyść hipotezy alternatywnej H_1 ($|t_n| > |t_\alpha|$, $\alpha > p\text{–}value$) – co oznacza, że z prawdopodobieństwem błędu mniejszym od 0,05 można twierdzić, że korelacja wymiarów detalu jest istotna.

8.2. Analiza regresji

Analiza regresji pozwala na ilościowe określenie związków pomiędzy kilkoma zmiennymi niezależnymi a zmienną zależną – metoda nazywana jest *analizą regresji wielorakiej (wielowymiarowej lub wielokrotnej)*, w przypadku dwóch zmiennych (jedna zmienna niezależna i jedna zależna) nazywana jest *analizą regresji**.

*Termin *regresja* wprowadził Sir Francis Galton w pracy „Naturalna dziedziczność" (1899). Opisał on zjawisko „regresji do średniej": dzieci rodziców uzdolnionych są przeciętnie mniej uzdolnione, rozmiary nasion groszku w kolejnych pokoleniach wracają do średniego rozmiaru, itp.

Analiza regresji pozwala na wyznaczenie parametrów teoretycznego modelu obiektu na podstawie obserwacji jego wejść i wyjść.



Schemat obiektu badań (x_1, x_2, ..., x_m – wielkości wejściowe, z – zakłócenie, y – wielkości wyjściowa).

Regresja liniowa to jeden z najbardziej popularnych modeli regresji zakładający, że związek pomiędzy zmienną zależną a zmiennymi niezależnymi ma charakter liniowy.

Celem *analizy regresji* jest wyznaczenie parametrów modelu matematycznego obiektu w postaci:

$$\hat{y} = f(x_1, x_2, \ldots, x_m, b_0, b_1, \ldots, b_k) \qquad (*)$$

gdzie: $\hat{y}$ – przybliżona wartość wyjścia obiektu; x_1, x_2, ..., x_m – wejścia obiektu; b_i – nieznane parametry; i=0, 1,..., k.

Funkcja (*) nazywana jest *funkcją regresji*. W przypadku *regresji liniowej*, *funkcja regresji* jest funkcją liniową:

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$$

a w przypadku *regresji nieliniowej* zakłada się, że *funkcja regresji* jest liniową kombinacją *funkcji bazowych*:

$$\hat{y} = b_0 \varphi_0(x) + b_1 \varphi_1(x) + \ldots + b_k \varphi_k(x)$$

gdzie: $x = (x_1, x_2, \ldots, x_m)^T$ – wektor wejść obiektu; $\varphi_i(x)$ – z góry zadane funkcje bazowe.

Współczynniki b_i *funkcji regresji* wyznaczane są zgodnie z ***zasadą najmniejszej sumy kwadratów błędów***, tzn. wyznaczane są w taki sposób, aby suma kwadratów odchyleń pomiędzy zaobserwowanymi wartościami wyjść a wyznaczonymi w oparciu o współczynniki b_i była minimalna:

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

8.2.1. Metoda najmniejszych kwadratów

Zadanie.

Zaleźć współczynniki b_i funkcji postaci:

$$\hat{y} = b_0\,\varphi_0(x) + b_1\,\varphi_1(x) + \ldots + b_k\,\varphi_k(x) \qquad (*)$$

tak aby w najlepszy sposób (przy minimalnej sumie kwadratów błędów) funkcja ta odwzorowywała związek zmiennej zależnej *y* od zmiennych niezależnych $x = (x_1, x_{.2}, \ldots x_{.m})^T$.

Rozwiązanie.

Załóżmy, że wykonano *n* obserwacji:

$$\begin{array}{ccccc} x_{11} & x_{12} & \ldots & x_{1m} & y_1 \\ x_{21} & x_{22} & \ldots & x_{2m} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} & y_n \end{array}$$

gdzie: $x_{i1}, x_{i2}, \ldots x_{in}$ – wartości zmiennych niezależnych przyjęte w *i*–tej obserwacji, y_i – wartość zmiennej zależnej otrzymana w *i*–tej obserwacji.

Wprowadzając oznaczenia:

wektor wejść obiektu w *i*–tej obserwacji:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T,$$

wektor parametrów funkcji:

$$b = (b_1, b_2, \ldots, b_k)^T,$$

macierz wejść:

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_k(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \ldots & \varphi_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_k(x_n) \end{pmatrix},$$

wektor przybliżonych wartości zmiennej zależnej:

$$\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T,$$

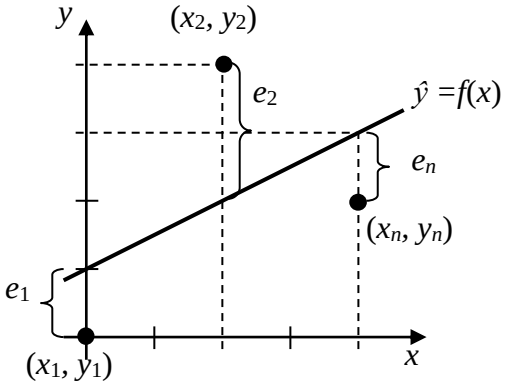
rezultat przybliżania funkcji obiektu funkcją aproksymującą (*) dla wykonanych obserwacji można zapisać w postaci równania macierzowego:

$$\hat{y} = X\,b\,.$$

Wartości poszukiwanych parametrów *b* należy wyznaczyć w taki sposób, żeby suma kwadratów błędów:

$$sse = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min$$

osiągała wartość minimalną. Wektor $e = (e_1, e_2, \ldots, e_n)^T$ jest nazywany wektorem błędów lub wektorem reszt ($e_i = y_i - \hat{y}_1$).



Zapisując zaobserwowane wartości wyjść obiektu w postaci wektora:

$$y = (y_1, y_2, \ldots, y_n)^T$$

błąd *sse* można zapisać jako:

$$sse = (y - \hat{y})^T (y - \hat{y}) = (y - X\,b)^T (y - Xb) = y^T y - y^T X\,b - b^T X^T y + b^T X^T X\,b$$

$$sse = y^T y - 2b^T X\,y + b^T X^T Xb$$

Zgodnie z teorią rachunku różniczkowego funkcji wielu zmiennych, warunkiem koniecznym istnienia ekstremum funkcji wielu zmiennych w określonym punkcie jest zerowanie się jej wszystkich pochodnych cząstkowych w tym punkcie. Ponieważ błąd aproksymacji musi być minimalny to:

$$\frac{\partial sse}{\partial b} = 0\,.$$

Pochodną błędu wyznacza się jako:

$$\frac{\partial sse}{\partial b} = -2X^T y + 2X^T Xb,$$

więc:

$$-2X^T y + 2X^T X\,b = 0\,,$$

czyli:

$$X^T X\,b = X^T y,$$

i ostatecznie poszukiwane parametry funkcji aproksymującej można wyznaczyć z zależności:

$$b = (X^T X)^{-1} X^T y\,.$$

Istnienie rozwiązania zależy od postaci macierzy $X^T X$ – macierz ta musi być macierzą nieosobliwą.

Przykład 3.

Podczas badania zależności kosztów produkcji od ilości produkowanych sztuk otrzymano z próby n=4 następujące wyniki:

x_i	1	2	3	4
y_i	3	4	6	8

Zakładając liniową postać funkcji regresji wyznaczyć jej parametry, obliczyć otrzymaną wartość błędu, narysować wykres rozrzutu, wykreślić znalezioną funkcję.

W zadaniu przyjęto założenie:

$$\hat{y} = b_0 + b_1 x .$$

Z porównania założonej postaci funkcji z jej ogólną postacią: $\hat{y} = b_0 \varphi_0(x) + b_1 \varphi_1(x)$, wynika, że funkcje bazowe $\varphi_0(x)$ i $\varphi_1(x)$ wynoszą w tym przypadku:

$$\varphi_0(x) = 1 \text{ i } \varphi_1(x) = x$$

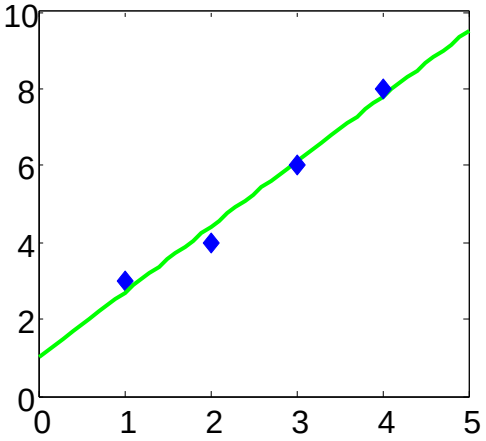
Dla otrzymanych danych pomiarowych macierz wejść X oraz wektor wyjść y wynoszą więc:

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) \\ \varphi_0(x_3) & \varphi_1(x_3) \\ \varphi_0(x_4) & \varphi_1(x_4) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \qquad y = \begin{pmatrix} 3 \\ 4 \\ 6 \\ 8 \end{pmatrix}.$$

Poszukiwane parametry funkcji otrzymuje się po obliczeniach:

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}, \qquad \left(X^T X\right)^{-1} = \begin{pmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{pmatrix},$$

$$b = \begin{pmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 6 \\ 8 \end{pmatrix} = \begin{pmatrix} 1 \\ 1{,}7 \end{pmatrix}.$$



Znaleziona funkcja regresji ma więc postać:

$$\hat{y} = 1 + 1{,}7\, x .$$

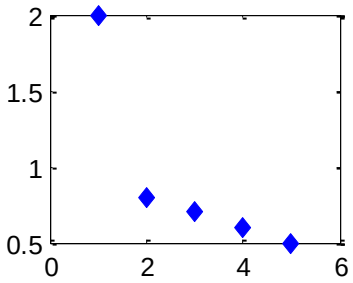
Wektor przybliżonych wartości zmiennej zależnej wynosi w tym przypadku $\hat{y} = (2{,}7,\ 4{,}4,\ 6{,}1,\ 7{,}8)^T$ a suma kwadratów błędów $sse = (3-2{,}7)^2 + \ldots + (8-7{,}8)^2 = 0{,}3$.

Przykład 4.

Podczas badania zależności zmian pewnego parametru procesu od czasu otrzymano następujące wyniki:

t_i	1	2	3	4	5
z_i	2	0,8	0,7	0,6	0,5

Na brak liniowej korelacji parametru od czasu wskazuje wykres rozrzutu oraz test istotności korelacji.



Obliczony na podstawie wyników z próby współczynnik korelacji wynosi: $r \approx -0{,}824$, wartość statystyki testowej w teście istotności korelacji otrzymuje się jako:

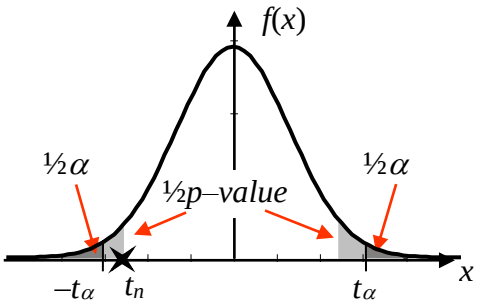
$$t_n = \frac{-0{,}824}{\sqrt{1-0{,}824^2}}\sqrt{5-2} \approx -2{,}5193$$

Graniczny poziom istotności wynosi w tym przypadku:

$$\frac{1}{2}\,p\text{–}value = F_{t(3)}(t_n) = F_{t(3)}(-2{,}5193) \approx 0{,}0431\,, \qquad \rightarrow\ p\text{–}value \approx 0{,}0862\,,$$

a granicę obszaru krytycznego dla $\alpha = 0{,}05$ stanowi:

$$t_\alpha = -F^{-1}_{t(3)}\left(\frac{0{,}05}{2}\right) \approx 3{,}1824\,.$$



Hipotezy o braku istotności korelacji nie można więc w tym przypadku odrzucić ($|t_n| < |t_\alpha|$, $\alpha < p\text{–}value$).

W dalszych badaniach przyjęto, że zależność zmian analizowanego parametru od czasu ta ma postać:

$$\hat{z} = a_0 e^{a_1/t}\,. \qquad (*)$$

Parametry a_0 i a_1 nie mogą być bezpośrednio wyznaczone poprzez zastosowanie analizy regresji, ponieważ przyjęta postać funkcji nie może być zapisana bezpośrednio w postaci liniowej kombinacji funkcji bazowych: $\hat{z} = b_0\,\varphi_0(t) + b_1\,\varphi_1(t) + \ldots + b_k\,\varphi_k(t)$. Funkcję tą można jednak przekształcić do takiej postaci:

$$\hat{z} = a_0 e^{a_1/t} \quad \Rightarrow \quad \ln(\hat{z}) = \ln\left(a_0 e^{a_1/t}\right) \quad \Rightarrow \quad \ln(\hat{z}) = \ln(a_0) + \ln\left(e^{a_1/t}\right)$$

$$\ln(\hat{z}) = \ln(a_0) + \frac{a_1}{t}.$$

Po wprowadzeniu oznaczeń:

$$\hat{y} = \ln(\hat{z}), \quad b_0 = \ln(a_0), \quad b_1 = a_1,$$

można ostatecznie funkcję (*) zapisać w postaci kombinacji liniowej funkcji bazowych $\varphi_0(t) = 1$ i $\varphi_1(t) = \frac{1}{t}$ jako:

$$\hat{y} = b_0 + b_1 \frac{1}{t}.$$

Test istotności korelacji zmiennych $\ln(z)$ i $1/t$ pokazuje, że korelacja ta jest istotna. W poniższej tabeli zestawiono wartości przekształconych zmiennych:

$\frac{1}{t_i}$	1	0,5	0,33	0,25	0,2
$\ln(z_i)$	0,69	-0,22	-0,36	-0,51	-0,69

Współczynnik korelacji dla nowych zmiennych wynosi: $r \approx 0.9934$ a wartość statystyki testowej przyjmuje w tym przypadku wartość: $t_n \approx 15{,}098$. Graniczny poziom istotności otrzymuje się jako: $p\text{–}value \approx 0{,}0006$, granica obszaru krytycznego wynosi $t_\alpha \approx 3{,}1824$. Wartości te pokazują, że hipoteza o braku istotności korelacji musi być w tym przypadku odrzucona ($|t_n| > |t_\alpha|$, $\alpha > p\text{–}value$) – korelacja zmiennych $\ln(z)$ i $1/t$ jest więc istotna.

Analiza regresji pozwala na wyznaczenie parametrów b_0 i b_1 funkcji $\hat{y}$. Dla przekształconych zmiennych macierz wejść *X* oraz wektor wyjść *y* wynoszą:

$$X = \begin{pmatrix} \varphi_0(t_1) & \varphi_1(t_1) \\ \varphi_0(t_2) & \varphi_1(t_2) \\ \varphi_0(t_3) & \varphi_1(t_3) \\ \varphi_0(t_4) & \varphi_1(t_4) \\ \varphi_0(t_5) & \varphi_1(t_5) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0{,}5 \\ 1 & 0{,}33 \\ 1 & 0{,}25 \\ 1 & 0{,}2 \end{pmatrix} \qquad y = \ln(z) = \begin{pmatrix} 0.69 \\ -0.22 \\ -0.36 \\ -0.51 \\ -0.69 \end{pmatrix}.$$

Parametry funkcji $\hat{y}$ otrzymuje się jako:

$$b = \left(X^T X\right)^{-1} X^T y \approx \begin{pmatrix} -0{,}9716 \\ 1{,}6499 \end{pmatrix}.$$

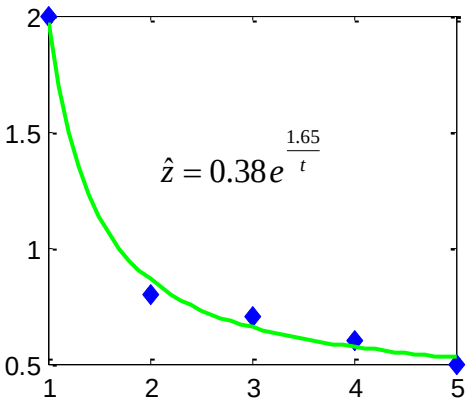
Funkcję $\hat{y}$ można więc zapisać jako:

$$\hat{y} \approx -0{,}9716 + 1{,}6499\frac{1}{t}.$$

Wykorzystując zależności: $\hat{z} = e^{\hat{y}}$, $a_0 = e^{b_0}$, $a_1 = b_1$,

otrzymuje się ostateczną postać poszukiwanej funkcji $\hat{z}(t)$:

$$\hat{z} \approx 0{,}3785\, e^{\frac{1{,}6499}{t}}.$$



Przykład 5.

W celu zbadania wpływu dwóch parametrów procesu obróbki na wysokość nierówności wykonano n=10 pomiarów i otrzymano następujące wyniki:

x_{i1}	2,1	1,1	3,1	1,1	2,4	4,4	4,1	1,7	1,1	1,1
x_{i2}	5,8	4,6	2,4	5,6	2,2	2,3	4,1	3,7	3,7	2,1
y_i	24,4	18,4	16,0	22,2	14,0	18,6	24,2	17,7	16,4	11,4

Zakładając następującą postać funkcji regresji:

$$\hat{y}(x) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2,$$

wyznaczyć jej parametry, narysować wykres rozrzutu, wykreślić znalezioną funkcję.

Z porównania założonej postaci funkcji z jej ogólną postacią:

$$\hat{y} = b_0 \varphi_0(x) + b_1 \varphi_1(x) + b_2 \varphi_2(x) + b_3 \varphi_3(x),$$

wynika, że funkcje bazowe $\varphi_0(x)$, $\varphi_1(x)$, $\varphi_2(x)$, $\varphi_3(x)$ wynoszą w tym przypadku:

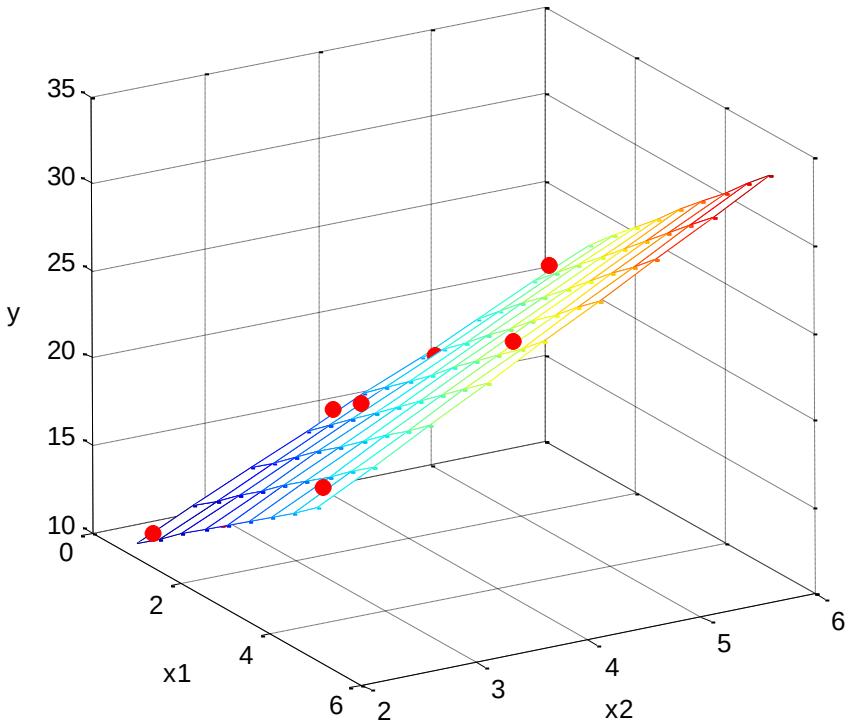
$$\varphi_0(x) = 1, \quad \varphi_1(x) = x_1, \quad \varphi_2(x) = x_2, \quad \varphi_3(x) = x_1 x_2.$$

Dla otrzymanych danych pomiarowych macierz wejść X oraz wektor wyjść y wynoszą odpowiednio:

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \varphi_2(x_1) & \varphi_3(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \varphi_2(x_2) & \varphi_3(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_{10}) & \varphi_1(x_{10}) & \varphi_2(x_{10}) & \varphi_3(x_{10}) \end{pmatrix} = \begin{pmatrix} 1 & 2{,}1 & 5{,}8 & 12.{,}8 \\ 1 & 1{,}1 & 4{,}6 & 5{,}06 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1{,}1 & 2{,}1 & 2{,}31 \end{pmatrix} \quad y = \begin{pmatrix} 24{,}4 \\ 18{,}4 \\ \vdots \\ 11{,}4 \end{pmatrix}.$$

Parametry funkcji $\hat{y}$ otrzymuje się jako: $b = (X^T X)^{-1} X^T y \approx (3{,}2173 \quad 1{,}6963 \quad 2{,}8484 \quad 0{,}1255)^T$, funkcję $\hat{y}$ można więc zapisać jako:

$$\hat{y}(x) = 3{,}2173 + 1{,}6963\, x_1 + 2{,}8484\, x_2 + 0{,}1255\, x_1 x_2.$$



8.2.2. Wybrane własności rozwiązania metody najmniejszych kwadratów

1. Macierz wejść jest nieskorelowana z wektorem błędów:

$$X^T e = 0 .$$

Mnożąc znaleziony wektor współczynników równania regresji b lewostronnie przez macierz $\left(X^T X\right)$ *otrzymuje się:* $b = \left(X^T X\right)^{-1} X^T y \quad \Rightarrow \quad X^T X\, b = X^T y .$

Korzystając z zależności $\hat{y} = X\, b$ *otrzymuje się następnie:*

$X^T \hat{y} = X^T y \quad \Rightarrow \quad X^T y - X^T \hat{y} = 0 \quad \Rightarrow \quad X^T \left(y - \hat{y}\right) = 0 \quad \Rightarrow \quad X^T e = 0 .$ ■

2. Suma elementów wektora e wynosi 0:

$$\sum_{i=1}^{n} e_i = 0 \ (*).$$

Jeśli funkcja regresji zawiera stałą to pierwsza kolumna macierzy X jest zbudowana z jedynek, i w konsekwencji pierwszy wiersz macierzy X^T *jest również zbudowany z jedynek. Na mocy poprzedniej własności otrzymuje się więc:* $\sum_{i=1}^{n} e_i = 0 .$ ■

(*) – zależność jest prawdziwa o ile funkcja regresji zawiera stałą.

3. Wektor przybliżonych wartości zmiennej zależnej $\hat{y}$ jest nieskorelowany z wektorem błędów:

$$\hat{y}^T e = 0 .$$

Wektor $\hat{y}$ otrzymuje się jako:

$$\hat{y} = X b .$$

Transpozycja i przemnożenie prawostronne powyższej zależności przez wektor błędów e prowadzi do:

$$\hat{y}^T e = (X b)^T e = b^T X^T e .$$

Z własności 1. otrzymuje się ostatecznie:

$$\hat{y}^T e = b^T X^T e = b^T 0 = 0 .$$ ■

4. Całkowita zmienność zmiennej zależnej może być dekomponowana na zmienność wyjaśnioną równaniem regresji i zmienność niewyjaśnioną:

$$sst = ssr + sse \ ^{(*)}.$$

gdzie:

sst – całkowita zmienność zmiennej zależnej jest mierzona jako suma kwadratów odchyleń zmiennej zależnej od średniej: $sst = \sum_{i=1}^{n} (y_i - \bar{y})^2$,

ssr – zmienność zmiennej zależnej wyjaśniona równaniem regresji: $ssr = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$,

sse – zmienność zmiennej zależnej niewyjaśniona równaniem regresji: $sse = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$,

(*) – zależność jest prawdziwa o ile funkcja regresji zawiera stałą.

Zależność tą można łatwo pokazać. Z definicji wektora błędu wynika, że:

$$y_i = \hat{y}_i + e_i .$$

Po odjęciu od obu stron powyższego równania średniej $\bar{y}$ i po podniesieniu do kwadratu powyższa zależność przyjmuje następującą postać:

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y} + e_i)^2 = (\hat{y}_i - \bar{y})^2 + 2(\hat{y}_i - \bar{y})e_i + e_i^2 .$$

Zsumowanie powyższego wyrażenia dla każdego $i = 1:n$ prowadzi do:

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

Wprowadzając błędy sst, ssr i sse otrzymuje się:

$$sst = ssr + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i + sse .$$

Teraz pozostaje jeszcze pokazać, że $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i = 0$. *Rzeczywiście:*

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})e_i = \sum_{i=1}^{n} \hat{y}_i e_i - \sum_{i=1}^{n} \bar{y} e_i = \sum_{i=1}^{n} \hat{y}_i e_i - \bar{y}\sum_{i=1}^{n} e_i = y^T e - \bar{y}\sum_{i=1}^{n} e_i .$$

Biorąc pod uwagę własności 2. i 3. otrzymuje się $\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)e_i = 0$. ■

Liczba niezależnych zmiennych (czyli liczba stopni swobody) potrzebna do wyznaczenia sum *sst, ssr i sse* została zestawiona w poniższej tabeli:

zmienność	*liczba stopni swobody*	*uwagi*
sst	$n-1$	do wyznaczania sumy $sst = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ potrzebnych jest n zmiennych y_i, jeden stopień swobody znika ze względu na ograniczenie $\bar{y} = 1/n \sum_{i=1}^{n} y_i$
ssr	k	do wyznaczania sumy $ssr = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ potrzebne są wartości $\hat{y}_i$, które otrzymuje się wykorzystując $k+1$ współczynników b_i z równania regresji, jeden stopień swobody znika ze względu na to, że można pokazać, że $b_0 = \bar{y} - \left(b_1\overline{\varphi_1(x)} + \ldots + b_k\overline{\varphi_k(x)}\right)$
sse	$n-k-1$	liczba zmiennych niezależnych po obu stronach równości $sst = ssr + sse$, czyli także $sse = sst - ssr$, jest taka sama

5. *Jakość dopasowania równania regresji*

Jakości dopasowania równania regresji do danych empirycznych ocenia się za pomocą *współczynnika determinacji*:

$$R^2 = \frac{ssr}{sst}.$$

Współczynnik ten określa więc jaka część zmiennej zależnej jest wyjaśniona równaniem regresji. Wartości współczynnika R^2 należą do przedziału $[0,1]$, dopasowanie funkcji regresji jest tym lepsze im wartość współczynnika jest bliższa 1.

Współczynnik R^2 dla funkcji regresji zawierającej stałą może być przekształcony do postaci:

$$R^2 = \frac{ssr}{sst} = \frac{sst - sse}{sst} = 1 - \frac{sse}{sst}.$$

W przypadku *regresji liniowej* współczynnik determinacji R^2 jest równy kwadratowi współczynnika korelacji liniowej Paersona: $R^2 = r^2$.

W celu uniezależnienia wartości wskaźnika R^2 od liczby stopni swobody wprowadzany jest tzw. *skorygowany współczynnik determinacji:*

$$\bar{R}^2 = 1 - \frac{sse/n-k-1}{sst/n-1} = 1 - \left(\frac{sse}{sst}\right)\left(\frac{n-1}{n-k-1}\right) = 1 - \left(1-R^2\right)\left(\frac{n-1}{n-k-1}\right).$$

Do oceny dopasowania równania regresji wykorzystywany jest również błąd standardowy estymacji:

$$s_e = \sqrt{\frac{sse}{n-k-1}}.$$

Wielkość błędu opisuje przeciętną różnicę pomiędzy wartościami empirycznymi a wartościami otrzymanymi z równania regresji, dopasowanie funkcji regresji jest tym lepsze im wartość błędu jest bliższa 0.

8.2.3. Istotność funkcji regresji

Badanie istotności regresji może być realizowane poprzez weryfikację hipotez statystycznych. Wyznaczone w wyniku *metody najmniejszych kwadratów* współczynniki b_i stanowią przybliżenie (są estymatorami) rzeczywistych współczynników β_i funkcji regresji dla całej populacji. *Funkcję regresji* uznaje się za *istotną* jeżeli przynajmniej jeden ze współczynników β_i funkcji regresji jest *istotnie* różny od zera. Istotność funkcji regresji bada się stawiając hipotezę zerową o braku wpływu zmiennych niezależnych na zmienną zależną, tzn. H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ wobec hipotezy alternatywnej, że przynajmniej jeden ze współczynników jest różny od zera, tzn. H_1: $\beta_i \neq 0$. Do przeprowadzenia tego testu wykorzystuje się, podobnie jak w analizie wariancji, statystykę F:

$$F = \frac{ssr/k}{sse/n-k-1}.$$

Wyrażenie w liczniku jest wariancją zmiennej zależnej wyjaśnioną równaniem regresji, wyrażenie w mianowniku jest wariancją zmiennej zależnej niewyjaśnioną równaniem regresji. Statystyka F może być także zapisana w postaci zależnej od współczynnika determinacji:

$$F = \frac{ssr}{sse}\frac{n-k-1}{k} = \frac{ssr/sst}{sse/sst}\frac{n-k-1}{k} = \frac{R^2}{1-R^2}\frac{n-k-1}{k}.$$

Jeżeli hipoteza H_0 jest prawdziwa statystyka F ma rozkład F *Snedecora (Fishera)* o $v_1 = k$ i $v_2 = n-k-1$ stopniach swobody. Funkcję regresji uznaje się za istotną jeżeli test wykaże, że hipotezę H_0 należy odrzucić, tzn. wartość statystyki testowej leży w obszarze krytycznym (obszar krytyczny w teście budowany jest jako prawostronny), lub graniczny poziom istotności jest mniejszy od założonego.

8.2.4. Istotność współczynników funkcji regresji

Istotność i-tego współczynnika funkcji regresji bada się stawiając hipotezę zerową o braku wpływu odpowiedniej zmiennej niezależnej na zmienną zależną, tzn. H_0: $\beta_i = 0$ wobec hipotezy alternatywnej, że przynajmniej zmienna wpływa na zmienną zależną, tzn. H_1: $\beta_i \neq 0$. Do przeprowadzenia tego testu wykorzystuje się statystykę t – *Studenta* o $(n-k-1)$ stopniach swobody:

$$t = \frac{b_i - \beta_i}{s_i} = \frac{b_i - 0}{s_i} = \frac{b_i}{s_i},$$

gdzie: s_i to estymator odchylenia standardowego parametru β_i obliczany jako: $s_i = \sqrt{\frac{sse}{n-k-1} c_{ii}}$ a współczynnik c_{ii} to *i*-ty element diagonalny macierzy $\left(X^T X\right)^{-1}$.

Przykład 6.

Dla funkcji regresji z przykładu 5. zweryfikować na poziomie istotności $\alpha = 0.05$ hipotezę o istotności funkcji regresji i istotności jej współczynników. Wyznaczyć współczynnik determinacji.

Zmienności zmiennej zależnej wyjaśnione i niewyjaśnione równaniem regresji wynoszą odpowiednio: $sse \approx 0{,}8364$, $ssr \approx 161{,}8446$, współczynnik determinacji wynosi więc:

$$R^2 = 1 - \frac{sse}{sst} \approx 0{,}9949 .$$

Otrzymana wartość współczynnika oznacza, że ponad 99% zmienności zmiennej zależnej jest wyjaśniona równaniem regresji.

Do przeprowadzenia testu istotności funkcji regresji należy obliczyć wartość statystyki testowej ($n = 10$ i $k = 3$):

$$F_n = \frac{R^2}{1-R^2} \frac{n-k-1}{k} \approx 386{,}9976 .$$

Graniczny poziom istotności wyznacza się korzystając z dystrybuanty rozkładu *F Snedecora (Fishera)* o $v_1 = 3$ i $v_2 = 6$ stopniach swobody:

$$p\text{–}value = 1 - F_{F(3,6)}\left(F_n\right) \approx 3e-7 ,$$

granica obszaru krytycznego dla $\alpha = 0{,}05$ wynosi w tym przypadku:

$$F_\alpha = F^{-1}_{F(3,6)}\left(1-0{,}05\right) \approx 4{,}7571 .$$

Wartości $p-value$ i F_α pokazują, że hipotezę o braku istotności funkcji regresji należy odrzucić ($F_n > F_\alpha$, $\alpha > p\text{–}value$) – znaleziona funkcja regresji musi więc być uznana za istotną.

Do przeprowadzenia testu istotności współczynników funkcji regresji należy wyznaczyć dla każdego współczynnika wartość statystyki testowej. Poniżej zostały zebrane wybrane wyniki cząstkowe:

$$\left(XX^T\right)^{-1} = \begin{bmatrix} 4{,}9562 & -1{,}7946 & -1{,}2101 & 0{,}4682 \\ -1{,}7946 & 0{,}8083 & 0{,}4623 & -0{,}2228 \\ -1{,}2101 & 0{,}4623 & 0{,}328 & -0{,}1338 \\ 0{,}4682 & -0{,}2228 & -0{,}1338 & 0{,}068 \end{bmatrix}$$

$$mse = \frac{sse}{n-k-1} \approx 0{,}1394 , \quad \sqrt{mse} \approx 0{,}3734 ,$$

$$t_\alpha = -F^{-1}_{t(6)}\left(\frac{0{,}05}{2}\right) \approx 2{,}4469 ,$$

i	0	1	2	3
b_i	$b_0 \approx 3{,}2173$	$b_1 \approx 1{,}6963$	$b_2 \approx 2{,}8484$	$b_3 \approx 0{,}1255$
c_{ii}	$c_{00} \approx 4{,}9562$	$c_{11} \approx 0{,}8083$	$c_{22} \approx 0{,}328$	$c_{33} \approx 0{,}068$
s_i	$s_0 \approx 0{,}3734\sqrt{4{,}9562}$ $s_0 \approx 0{,}8312$	$s_1 \approx 0{,}3734\sqrt{0{,}8083}$ $s_1 \approx 0{,}3357$	$s_2 \approx 0{,}3734\sqrt{0{,}328}$ $s_2 \approx 0{,}2138$	$s_3 \approx 0{,}3734\sqrt{0{,}068}$ $s_3 \approx 0{,}0974$
t_{in}	$t_{0n} \approx 3{,}8706$	$t_{1n} \approx 5{,}0534$	$t_{2n} \approx 13{,}3209$	$t_{3n} \approx 1{,}2890$
$p-value$	$p_0 \approx 0{,}0083$	$p_1 \approx 0{,}0023$	$p_2 \approx 0{,}0000$	$p_3 \approx 0{,}2449$

.

Test istotności dla współczynników funkcji regresji wskazuje, że wszystkie współczynniki z wyjątkiem ostatniego związanego z iloczynem zmiennych x_1x_2 są istotne (dla $i = 0,1,2$: $|t_{in}| > |t_\alpha|$ i $\alpha > p_i$, dla $i = 3$: $|t_{i3}| < |t_\alpha|$ i $\alpha < p_3$).

Z przeprowadzonej analizy wynika, że znalezioną funkcję regresji należy uznać za istotną, jednak tylko trzy spośród czterech współczynników zostały uznane za istotne. Czwarty współczynnik b_3 powinien zostać wyeliminowany z modelu.

8.2.5. Dobór funkcji regresji

Funkcja regresji powinna w możliwie jak największym stopniu wyjaśniać zmienność zmiennej zależnej, jednocześnie jednak powinna mieć możliwie najprostszą strukturę. Dobór modelu funkcji regresji nie jest sprawą prostą, stosowane są różne strategie. W selekcji postępującej konstruowanie modelu rozpoczyna się od jednej zmiennej niezależnej, w kolejnych krokach dodawane są kolejne zmienne.

W eliminacji wstecznej poszukiwanie optymalnego modelu rozpoczyna się od modelu maksymalnego a w kolejnych krokach kolejno usuwane są zmienne o najmniejszym wpływie na zmienną zależną. Proces eliminacji kończy się gdy w modelu występują wyłączne zmienne, które w istotny sposób wyjaśniają zmienność zmiennej zależnej. W każdym kroku eliminacji najmniejszą istotność współczynnika wskazuje największy graniczny poziomie istotności otrzymany dla tego współczynnika w teście istotności.

Przykład 8.

Stosując eliminację wsteczną uprościć funkcję regresji z przykładu 6.

Funkcja regresji w przykładzie 6. zawierała współczynnik, który został uznany za nieistotny – współczynnik ten, a właściwie skojarzona z nim interakcja zmiennych x_1 i x_2, zostanie usunięty z modelu, po wyeliminowaniu z funkcji składnika zawierającego iloczyn zmiennych x_1x_2 funkcja regresji przyjmuje następującą postać:

$$\hat{y}(x) = b_0 + b_1 x_1 + b_2 x_2 .$$

Zmiana postaci funkcji wiąże się ze zmianą postaci macierzy wejść *X*:

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \varphi_2(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \varphi_2(x_2) \\ \vdots & \vdots & \vdots \\ \varphi_0(x_{10}) & \varphi_1(x_{10}) & \varphi_2(x_{10}) \end{pmatrix} = \begin{pmatrix} 1 & 2{,}1 & 5{,}8 \\ 1 & 1{,}1 & 4{,}6 \\ \vdots & \vdots & \vdots \\ 1 & 1{,}1 & 2{,}1 \end{pmatrix} \quad y = \begin{pmatrix} 24{,}4 \\ 18{,}4 \\ \vdots \\ 11{,}4 \end{pmatrix}.$$

Parametry funkcji $\hat{y}$ otrzymuje się jako: $b = \left(X^T X\right)^{-1} X^T y \approx \left(2{,}3533 \quad 2{,}1075 \quad 3{,}0953\right)^T$, funkcję $\hat{y}$ można więc zapisać jako:

$$\hat{y}(x) = 2{,}3533 + 2{,}1075\, x_1 + 3{,}0953\, x_2 .$$

Zmienności zmiennej zależnej wyjaśnione i niewyjaśnione równaniem regresji wynoszą odpowiednio: $sse \approx 1{,}068$, $ssr \approx 161{,}613$, współczynnik determinacji wynosi więc:

$$R^2 = 1 - \frac{sse}{sst} \approx 0{,}9934 .$$

Wartość statystyki testowej wykorzystywanej do przeprowadzenia testu istotności funkcji regresji otrzymuje w tym przypadku wartość ($n = 10$ i $k = 2$):

$$F_n = \frac{R^2}{1 - R^2} \frac{n - k - 1}{k} \approx 529{,}6127 .$$

Graniczny poziom istotności wyznaczony z dystrybuanty rozkładu *F Snedecora (Fishera)* o $v_1 = 2$ i $v_2 = 7$ stopniach swobody wynosi:

$$p\text{–}value = 1 - F_{F(2,7)}(F_n) \approx 2e - 8 ,$$

a granicę obszaru krytycznego dla $\alpha = 0{,}05$ otrzymuj się jako:

$$F_\alpha = F^{-1}_{F(2,7)}(1 - 0{,}05) \approx 4{,}7374 .$$

Wartości $p - value$ i F_α pokazują, że hipotezę o braku istotności funkcji regresji należy odrzucić ($F_n > F_\alpha$, $\alpha > p\text{–}value$) – znaleziona funkcja regresji musi więc być uznana za istotną.

Do przeprowadzenia testu istotności współczynników funkcji regresji wyznacza się, podobnie jak poprzednio, wartości statystyk testowych. Poniżej zostały zebrane wybrane wyniki cząstkowe:

$$\left(XX^T\right)^{-1} = \begin{bmatrix} 1{,}7336 & -0{,}2608 & -0{,}2889 \\ -0{,}2608 & 0{,}0783 & 0{,}0238 \\ -0{,}2889 & 0{,}0238 & 0{,}0647 \end{bmatrix}$$

$$mse = \frac{sse}{n - k - 1} \approx 0{,}1526 , \quad \sqrt{mse} \approx 0{,}3906 ,$$

$$t_\alpha = -F^{-1}_{t(7)}\left(\frac{0{,}05}{2}\right) \approx 2{,}3646 ,$$

i	*0*	*1*	*2*
b_i	$b_0 \approx 2{,}3533$	$b_1 \approx 2{,}1075$	$b_2 \approx 3{,}0953$
c_{ii}	$c_{00} \approx 1{,}7336$	$c_{11} \approx 0{,}0783$	$c_{22} \approx 0{,}0647$
s_i	$s_0 \approx 0{,}3906\sqrt{1{,}7336}$ $s_0 \approx 0{,}5143$	$s_1 \approx 0{,}3906\sqrt{0{,}0783}$ $s_1 \approx 0{,}1093$	$s_2 \approx 0{,}3906\sqrt{0{,}0647}$ $s_2 \approx 0{,}0993$
t_{in}	$t_{0n} \approx 4{,}5758$	$t_{1n} \approx 19{,}2869$	$t_{2n} \approx 31{,}1643$
$p-value$	$p_0 \approx 0{,}0026$	$p_1 \approx 0{,}0000$	$p_2 \approx 0{,}0000$

Test istotności dla współczynników funkcji regresji wskazuje, że wszystkie współczynniki są istotne (dla $i = 0,1,2: |t_{in}| > |t_\alpha|$ i $\alpha > p_i$).

Proces eliminacji wstecznej można już zakończyć – znaleziona funkcja regresji i wszystkie jej współczynniki są istotne. Współczynnik determinacji dla znalezionej funkcji wynosi: $R^2 = 0.9934$, ma więc nieznacznie mniejszą wartość w stosunku do obliczonego dla modelu pełnego: $R^2 = 0.9949$.