Laboratorium 8 – Analiza korelacyjna i regresyjna

0) Widomości podstawowe

0.1) Zmienna losowa wielowymiarowa, rozkłady brzegowe, dystrybuanta, funkcja gęstości

Niech $(\Omega, \mathcal{F}, \Pr)$ przestrzeń probabilistyczna.

Def. Zmienną losową $X: \Omega \to \mathbb{R}^n$ nazywamy zmienną losową (n-wymiarową), jeżeli

$$\forall A \in \mathcal{B}(\mathbb{R}^n)\ \ X^{-1}(A) \in \mathcal{F}$$

gdzie $\mathcal{B}(\mathbb{R}^n)$ jest rodziną zbiorów borelowskich.

Def. Rozkładem prawdopodobieństwa zmiennej losowej $X: \Omega \to \mathbb{R}^n$ nazywamy miarę μ_X na $\mathbb{R}^n$ taką, że

$$\mu_X(B) = \Pr\left(X^{-1}(B)\right), \qquad \forall B \in \mathcal{B}(\mathbb{R}^n)$$

Def. Jeżeli istnieje funkcja $f_X: \mathbb{R}^n \to \mathbb{R}$ taka, że

$$\mu_X(B) = \int_B f_X(x)\,dx, \qquad \forall B \in \mathcal{B}(\mathbb{R}^n)$$

to f_X nazywamy *gęstością zmiennej losowej* X. Zmienną losową posiadająca gęstość nazywamy *ciągłą*.

0.2) Zmienne losowe niezależne i warunkowo niezależne

Def. Zmienne losowe $X_1, \dots, X_k: \Omega \to \mathbb{R}$ nazywamy *niezależnymi*, jeżeli dla każdego ciągu zbiorów borelowskich $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ mamy

$$\Pr(X_1 \in B_1, \dots, X_k \in B_k) = \Pr(X_1 \in B_1), \dots, \Pr(X_k \in B_k).$$

Tw. Niech $X_1, \dots, X_k: \Omega \to \mathbb{R}$ zmienne losowe. Następujące warunki są równoważne:

1. Zmienne losowe są niezależne.
2. $\mu_{X_1,\dots,X_k} = \mu_{X_1} \otimes, \dots, \otimes \mu_{X_k}$.
3. $\forall (t_1, \dots, t_k) \in \mathbb{R}^k$ mamy $F_{X_1,\dots,X_k}(t_1, \dots, t_k) = F_{X_1}(t_1), \dots, F_{X_k}(t_k)$.

Tw. Zmienne losowe o wartościach dyskretnych, rzeczywistych $\{X_i: \Omega \to S_i \subset \mathbb{R}\}_{i=1,..,k}$ są niezależne wtedy i tylko wtedy gdy $\forall (x_1, \dots, x_k) \in S_1 \times, \dots, \times S_k$ mamy

$$\Pr(X_1 = x_1, \dots, X_k = x_k) = \Pr(X_1 = x_1), \dots, \Pr(X_k = x_k).$$

0.3) Zmienne losowe skorelowane, kowariancja i współczynnik korelacji

Def. Kowariancją współrzędnych $X_i, X_j,\ i,j = 1,\dots,n$ zmiennej losowej $X: \Omega \to \mathbb{R}^n$ nazywamy funkcję

$$COV(X_i, X_j) = E\left(\left(X_i - E(X_i)\right)\left(X_j - E(X_j)\right)\right), \qquad i,j = 1,\dots,n$$

oraz $C = \{c_{ij}\} = \{COV(X_i, X_j)\},\ i,j = 1,\dots,n$ nazywamy macierzą kowariancji dla tej zmiennej, podczas gdy

$$\rho(X_i, X_j) = \frac{COV(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}, \qquad i,j = 1,\dots,n$$

nazywamy współczynnikiem korelacji (korelacji liniowej Pearsona) tych współrzędnych.

Uwagi:

1. Zmienne losowe X_i, X_j niezależne są nieskorelowane, tj. $COV(X_i, X_j) = \rho(X_i, X_j) = 0$.
2. Zmienne nieskorelowane mogą być zależne.
3. Zmienne losowe X_i, X_j skorelowane $COV(X_i, X_j) \neq 0$ są zależne.
4. $\rho(X_i, X_j) \in [-1,1]$ jest miarą unormowaną.

0.4) Funkcja regresji

Def. Funkcją regresji (funkcją regresji I rodzaju) współrzędnej X_i względem współrzędnej X_j dla pewnych $i,j = 1,\dots,n$ zmiennej losowej $X: \Omega \to \mathbb{R}^n$ nazywamy funkcję

$$g_{ij}(x) = E(X_i | X_j = x), i,j = 1,\dots,n$$

Zmienną X_i nazywamy *zmienną opisywaną* i traktowana jest ona jako zmienna losowa, natomiast zmienna X_j nazywana jest *zmienną objaśniającą* i traktowana jest jako przewidywalna (deterministyczna) wartość pomiaru, obserwacji czy eksperymentu.

Zwyczajowo zakłada się, że wariancja zmiennej opisywanej $\mathrm{Var}(X_i)$ nie zależy od wartości $X_j = x$.

Regresją nazywamy każda metodę aproksymacji funkcji regresji $g_{ij}(x)$.

Regresją parametryczną nazywamy przypadek aproksymacji $g_{ij}(x)$ w skończenie wymiarowej przestrzeni„ której wymiary nazywamy *parametrami regresji*.

0.5) Dyskretne rozkłady dwuwymiarowe

Niech $X: \Omega \to \{a_1,\dots,a_r\},\ Y: \Omega \to \{b_1,\dots,b_k\},\ a_i, b_j \in \mathbb{R}$ dwie zmienne losowe.

Rozkład prawdopodobieństwa pary (X,Y) stanowi tablica $\{p_{ij}\} = \{\Pr(X = a_i,\ Y = b_j)\}, i = 1,\dots,r,\ j = 1,\dots,k.$

Rozkłady brzegowe będą wektorami $\{p_{\cdot j}\} = \{\sum_{i=1}^{r} p_{ij}\},\ j = 1,\dots,k$, $\{p_{i\cdot}\} = \{\sum_{j=1}^{k} p_{ij}\},\ i = 1,\dots,r.$

Próba losowa takiego rozkładu ma najczęściej postać „tablicy wielodzielczej” (contigency, correlation table) postaci $\{n_{ij}\}, i = 1, \dots, r,\ j = 1, \dots, k$ takiej, że n_{ij} jest ilością elementów w próbie spełniających $X = a_i,\ Y = b_j$. Liczność takiej próby oznaczamy typowo $n = \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}$.

Test niezależności chi-kwadrat Pearsona

Układ hipotez jest następujący

H_0: X, Y są niezależne

H_1: X, Y są zależne

Stosujemy statystykę:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{{n_{ij}}^2}{\hat{n}_{ij}} - n, \qquad \hat{n}_{ij} = \frac{n_{i\,\cdot}\, n_{\cdot j}}{n}$$

gdzie $n_{i\,\cdot}\, n_{\cdot j}$ są współrzędnymi empirycznych częstości brzegowych $\{n_{\cdot j}\} = \left\{\sum_{i=1}^{r} n_{ij}\right\},\ j = 1, \dots, k$, $\{n_{i\cdot}\} = \left\{\sum_{j=1}^{k} n_{ij}\right\},\ i = 1, \dots, r$.

Statystyka ta ma przy prawdziwości H_0 rozkład χ^2 z $(r-1)(k-1)$ stopniami swobody. Zbiór krytyczny dla poziomu ufności α jest równy $C = [\chi^2_{1-\alpha}, +\infty)$, gdzie $\chi^2_{1-\alpha}$ jest kwantylem rozkładu χ^2 dla $(r-1)(k-1)$ stopni swobody.

Uwagi:

1. Liczebność próby n powinna być dostateczne duża, w szczególności wszystkie wartości $n_{ij} \geq 5$.
2. Bardzo duże wartości χ^2 oznaczają dużą różnicę pomiędzy częstościami obserwowanymi a oczekiwanymi, jest to argument za istnieniem zależności. Przeciwnie małe bliskie zera wartości świadczą o niezależności zmiennych.

Zadanie 1. Produkt można wytwarzać trzema metodami. Wysunięto hipotezę, że wadliwość produkcji nie zależy od metody produkcji. Wylosowano próbę 270 produktów i sporządzono dla niej tablice wielodzielczą.

Jakość	Metoda 1	Metoda 2	Metoda 3	$n_{i\cdot}$
dobra	40	80	60	180
zła	10	60	20	90
$n_{\cdot j}$	50	140	80	270

Należy na poziomie istotności $\alpha = 0.05$ zweryfikować tę hipotezę.

Zadanie można rozwiązać przy pomocy funkcji R:

```
chisq.test(matrix(c(40,10,80,60,60,20), nrow=2))
```

```
qchisq(0.95,2)
```

Wynik:

```
       Pearson's Chi-squared test

data:  matrix(c(40, 10, 80, 60, 60, 20), nrow = 2)
X-squared = 12.214, df = 2, p-value = 0.002227

[1] 5.991465
```

Lub inną funkcją R:

```
TeachingDemos::chisq.detail(matrix(c(40,10,80,60,60,200), nrow=2))
qchisq(0.95,2)
```

Wynik:

```
observed
expected

                        Total
      40     80    60      180
      33.33  93.33 53.33

      10     60    20       90
      16.67  46.67 26.67

Total 50    140    80      270


Cell Contributions

 1.33 + 1.90 + 0.83 +
 2.67 + 3.81 + 1.67 = 12.21

df = 2  P-value = 0.002

[1] 5.991465
```

Miary korelacji dwóch zmiennych losowych

Współczynnik korelacji liniowej Pearsona

Mając próbę n – elementową (np. w postaci tablicy wielodzielczej) dla jego oceny możemy stosować estymator

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Wynik obliczonej statystyki można interpretować następująco:

Przedział do którego należy $\|r\|$	Klasyfikacja współzależności zmiennych X, Y
{0.0}	Brak korelacji, zmienne nieskorelowane
(0.0, 0.3)	Słaby stopień współzależności
[0.3, 0.5)	Średni stopień współzależności
[0.5, 0.7)	Znaczny stopień współzależności
[0.7, 0.9)	Wysoki stopień współzależności
[0.9, 1.0)	Bardzo wysoki stopień współzależności
{1.0}	Współzależność całkowita (funkcyjna)

Estymator współczynnika korelacji dla wektorów `x` i `y` oblicza funkcja `cor(x,y)`.

Uwagi

1. Po obliczeniu estymatora *r* wskazane jest sporządzenie wykresu rozrzutu zależności zmiennych (wykres kropkowy). Zmienne mogą być od siebie zależne w sposób nieliniowy, co nie objawia się wysoką wartości estymatora *r*.
2. Na wielkość współczynnika Pearsona nie ma wpływu przeskalowanie lub przesunięcie.
3. Estymator jest wrażliwy na wielkość i jakość próby. Najlepiej stosować próby duże *n* >50.
4. Wartość estymatora zależy od zakresu (przeciwdziedziny) zmiennych. Im większa dziedzina, tym większa jest wartość estymatora.

Zależność korelacyjna dwóch lub większej ilości zmiennych może być wizualizowana przy pomocy funkcji R

`ellipse::plotcorr`

przedstawia macierz korelacji za pomocą elips – im elipsa jest bardziej zbliżona do koła, tym korelacja mniejsza

`ellipse::heatmap`

rysuje „mapę ciepła” dla macierzy korelacji.

Test istotności dla współczynnika korelacji liniowej Pearsona

Jeżeli założymy, że dane (populacja generalna) mają rozkład dwuwymiarowy normalny dla dwóch wybranych cech (zmiennych losowych), to możemy wykorzystać następujące przedziały ufności dla poziomu ufności $1-\alpha$:

$$\left(r - z_{1-\frac{\alpha}{2}} \frac{1-r^2}{\sqrt{n-3}} < \rho < r + z_{1-\frac{\alpha}{2}} \frac{1-r^2}{\sqrt{n-3}}\right), \qquad n \geq 100$$

$$\left(w - z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}} < \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) < w + z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}\right), w = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right), \qquad n \geq 10$$

Stosując jeden z powyższych estymatorów możemy dokonać ocenę istotności estymatora r współczynnika korelacji ρ. Mamy hipotezy:

$H_0: \rho = 0,$

$H_1: \rho \neq 0$.

Jeżeli próba losowa o liczności n pochodzi z dwuwymiarowego rozkładu normalnego, to możemy zastosować statystykę

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

Która przy prawdziwości H_0 ma rozkład *t*-Studenta o *n* - 2 stopniach swobody. Zbiór krytyczny ma postać

$$C = \left(-\infty, -t_{1-\frac{\alpha}{2}}\right] \cup \left[t_{1-\frac{\alpha}{2}}, +\infty\right)$$

Gdzie $t_{1-\frac{\alpha}{2}}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu *t*-Studenta o *n* - 2 stopniach swobody.

Zadanie 2. Plony czarnej porzeczki zależą od wieku plantacji. Zebrano dane z 7 plantacji (tabela)

Wiek plantacji x_i	1	3	2	3	4	3	5
Plony y_i	85	105	100	110	125	115	130

Zbadać stopień skorelowania plonów czarnej porzeczki z wiekiem plantacji poprzez obliczenie współczynnika korelacji liniowej.

Zadanie można również rozwiązać przy pomocy funkcji R:

```
x=c(1,3,2,3,4,3,5)
y=c(85,105,100,110,125,115,130)
cor(x,y)
```

Test można wykonać funkcją

```
cor.test(x,y)
```

Wynik:

```
        Pearson's product-moment correlation

data:  x and y
t = 9.237, df = 5, p-value = 0.0002498
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8164260 0.9959973
sample estimates:
      cor
0.9719274
```

Przedział ufności nie zawiera wartości estymatora współczynnika korelacji, zatem hipotezę zerową należy odrzucić.

Związek cech niemierzalnych (opuszczamy)

Regresja liniowa

Niech X będzie zmienna objaśniającą, Y zmienną opisywaną. Regresja liniowa poszukuje zależności $E(Y) = \alpha_0 + \alpha_1 x$, gdzie x jest wartością zmiennej losowej X, czyli w 2-wymiarowej przestrzeni wielomianów 1 stopnia na $\mathbb{R}$.

W praktyce posługujemy się następującym modelem regresji liniowej

$$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$$

gdzie $\varepsilon_i; E(\varepsilon_i \varepsilon_j) = 0, E(\varepsilon_i) = 0$ są nazywane składnikami losowymi.

Mając próbę losową wartości zmiennych $x_i, y_i,\ i = 1, \dots, n$ możemy skonstruować

$\hat{Y}_i = \alpha_0 + \alpha_1 x_i$; $U_i = y_i - \hat{Y}_i$ zwany składnikiem resztowym.

Najlepszym, nieobciążonym estymatorem α_0, α_1 jest estymator średniokwadratowy:

$$\alpha_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\alpha_0 = \bar{y} - \alpha_1 \bar{x}$$

Błąd takiej estymacji mierzy odchylenie standardowe składnika resztowego

$$S(U) = \sqrt{\frac{\sum_{i=1}^{n} U_i}{n-2}}$$

Szacunek średniego błędu oceny każdego z parametrów regresji α_0, α_1 można mierzyć ich odchyleniami standardowymi

$$S(\alpha_1) = \frac{S(U)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$S(\alpha_0) = \sqrt{\frac{S^2(U)\ \sum_{i=1}^{n} x_i^2}{n\ \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Oszacowanie błędów elementów struktury dają możliwość oszacowania błędów względnych

$$\frac{S(\alpha_i)}{|\alpha_i|}\ 100\%, \qquad i = 1,2$$

Wartości te nie mogą przekraczać 50% aby model był poprawny.

Prawdziwa jest również „równość wariacyjna”

$$S^2(Y) = S^2(\hat{Y}) + S^2(U)$$

Daje ona możliwość obliczenia kolejnej miary jakości, współczynnika zbieżności"

$$\varphi^2 = \frac{S^2(U)}{S^2(Y)} = \frac{S^2(U)}{S^2(\hat{Y}) + S^2(U)}$$

oraz współczynnika determinacji i poprawionego współczynnika determinacji

$$R^2 = \frac{S^2(\hat{Y})}{S^2(Y)} = 1 - \varphi^2\,, \quad R^2_{\text{popr}} = 1 - \frac{n-1}{n-2}\,(1 - R^2)$$

Dla poprawnej weryfikacji modelu $R^2_{\text{popr}} > 60\%$.

Przedziały ufności dla współczynników regresji liniowej dla α_1, α_0:

$$\left(\alpha_1 - t_{1-\frac{\alpha}{2}}\,\frac{s(u)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\,, \qquad \alpha_1 + t_{1-\frac{\alpha}{2}}\,\frac{s(u)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

$$\left(\alpha_0 - t_{1-\frac{\alpha}{2}}\,\sqrt{\frac{s^2(u)\,\sum_{i=1}^n x_i^2}{n\,\sum_{i=1}^n (x_i - \bar{x})^2}}\,, \qquad \alpha_0 + t_{1-\frac{\alpha}{2}}\,\sqrt{\frac{s^2(u)\,\sum_{i=1}^n x_i^2}{n\,\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

gdzie u jest obserwacją (wartością obliczoną) statystyki U, $t_{1-\frac{\alpha}{2}}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu t-Studenta dla $n-2$ stopni swobody, α_1, α_0 obliczone wartości estymatorów średniokwadratowych.

Test (nie)istotności dla współczynnika regresji

Mamy zestaw hipotez:

H_0: $\alpha_1 = 0$,

H_1: $\alpha_1 \neq 0$.

W przypadku przyjęcia hipotezy H_0 uznajemy, że regresja jest funkcją stała, czyli zmienne nie są skorelowane. W przypadku jej odrzucenia liniowa (afiniczna) zależność regresyjna jest istotna.

Stosujemy statystykę testową:

$$T = \frac{\alpha_1}{S(\alpha_1)}$$

która ma w przypadku prawdziwości H_0 rozkład t-Studenta dla $n-2$ stopni swobody. Zbiór krytyczny jest postaci $C = \left(-\infty,\, t_{1-\frac{\alpha}{2}}\right] \cup \left[\, t_{1-\frac{\alpha}{2}}, +\infty\right)$.

Nie prowadzi się weryfikacji istotności wyrazu wolnego α_0 .

Prognoza i predykcja stochastyczna zmiennej Y

Funkcje regresji liniowej możemy stosować do prognozowania wartości średniej zmiennej losowej Y dla pewnej wartości zmiennej objaśniającej $X = x$. *Prognoza* $E(Y)$ ma postać $\alpha_0 + \alpha_1 x$, gdzie

α_1, α_0 obliczone wartości estymatorów średniokwadratowych. Przedział ufności dla poziomu ufności α takiej prognozy jest dany wzorem

$$\left(\alpha_0 + \alpha_1 x - t_{1-\frac{\alpha}{2}} s_e\,, \qquad \alpha_0 + \alpha_1 x + t_{1-\frac{\alpha}{2}} s_e\right)$$

$$s_e^2 = s^2(u)\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

gdzie $s^2(u)$ jest obliczoną wariancją dla składnika resztowego U, $t_{1-\frac{\alpha}{2}}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu t-Studenta dla $n-1$ stopni swobody.

Jeżeli chcemy uzyskać prognozę dla Y dla $x \in [a,b]$. Dla odcinka prostej $\alpha_0 + \alpha_1 x;\ x \in [a,b]$ można obliczyć dwie krzywe będące górnymi i dolnymi granicami przedziałów ufności dla $x \in [a,b]$ ograniczającymi zakres $E(Y)$ dla tego przedziału zmiennej objaśniającej, przy założonym poziomie ufności α.

Predykcją zmiennej losowej Y jest jej przybliżona wartością dla ustalonej wartości zmiennej objaśniającej $X = x$. *Predykcja* Y ma także postać $\alpha_0 + \alpha_1 x$, gdzie α_1, α_0 obliczone wartości estymatorów średniokwadratowych. Przedział ufności jest dany podobnym wzorem , przy czym wartość błędu s_e^2 powinna być obecnie zastąpiona przez

$$s_p^2 = s^2(u)\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

Prognoza i predykcja może być obliczana dla $X = x$ leżących w przedziale wyznaczonym przez zakres obserwowanych wartości zmiennej objaśniającej $[\min_{i=1,\ldots,n}\{x_i\}, \max_{i=1,\ldots,n}\{x_i\}]$.

Zadanie 3. W tabeli zestawiono liczbę zachorowań na gruźlicę (na 100 000 ludności) w latach 1995 – 2005.

Rok x_i	1995	1996	1997	1998	1999	2000	2001	2002
Zachorowania y_i	39.7	38.2	34.7	33.1	30.1	28.4	26.3	24.7

Zakładając liniową zależność liczby zachorowań od roku przeprowadzić wszech stronną analizę regresji.

Zadanie możemy rozwiązać przy pomocy funkcji R o nazwie `lm`.

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Opis danych:

`formula` – opis formuły regresji oraz danych, dla liniowej regresji opis modelu ma postać `y~x` gdzie `x,y` – obserwacje zmiennej objaśniającej i objasnianej,

```
x=c(1:8)
y=c(39.7, 38.2, 34.7, 33.1, 30.1, 28.4, 26.3, 24.70)
m=lm(y~x)
summary(m)
```

Wynik:

```
Call:
lm(formula = y ~ x)

Residuals:
  Min       1Q   Median       3Q      Max
-0.69048 -0.26071 -0.00952  0.20952  0.75238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.88571    0.41289  101.44 6.18e-11 ***
x           -2.21905    0.08177  -27.14 1.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5299 on 6 degrees of freedom
Multiple R-squared:  0.9919,    Adjusted R-squared:  0.9906
F-statistic: 736.5 on 1 and 6 DF,  p-value: 1.654e-07
```

Zadanie 4. Bazując na danych i wynikach poprzedniego zadania narysuj wykresy funkcji regresji oraz krzywe ufności na poziomie ufności $1-\alpha=0.95$ dla przedziału zmiennej objaśniającej odpowiadającej zakresowi jej obserwacji.