MPiS30 W11: *ANALIZA PROSTEJ REGRESJI I KORELACJI*

1. **Modele deterministyczne i probabilistyczne**
   Przykład 1
2. **Założenia prostej regresji liniowej**
3. **Estymacja parametrów modelu metodą najmniejszych kwadratów**
   Przykład 2
4. **Własności estymatorów parametrów modelu**
5. **Estymator wariancji błędu modelu**
   Przykład 3
6. **Ocena adekwatności modelu**
   Przykład 4
7. **Przedział ufności dla parametru nachylenia**

1. Modele deterministyczne i probabilistyczne

Regresja oznacza zespół metod pozwalających na zbadanie związku pomiędzy zmiennymi i wykorzystanie tej wiedzy do przewidywania nieznanych wartości jednych zmiennych na podstawie znajomości wartości innych.

W praktyce poszukuje się związku między domniemaną jedną (lub więcej) zmienną objaśniającą lub niezależną X (ang. explanatory variable) a zmienną objaśnianą lub zależną Y (ang. response variable). Praktycy często chcą znać związek między zmiennymi, np. wielkością środków przeznaczanych na reklamę a uzyskiwanymi obrotami. Można postawić pytania:

Czy istnieje dokładny związek pomiędzy tymi zmiennymi?

Czy wartość obrotów można dokładnie przewidzieć dla ustalonych wydatków na reklamę?

Odpowiedzi są negatywne, bowiem obroty firmy zależą nie tylko od środków wydatkowanych na reklamę, lecz od wielu innych zmiennych wpływających na wielkość obrotu, np. pory roku, koniunktury gospodarczej, struktury cen, asortymentu itp.

Wymienione zmienne objaśniające można również uwzględnić w modelu. Niestety, bez względu na to, jak wiele zmiennych byłoby uwzględnionych w modelu, wciąż można narazić się na popełnienie błędu. Przyczyną niedokładności są zjawiska losowe nie ujęte w modelu lub jeszcze nierozpoznane.

Czy zawsze praktycy są skazani na popełnianie błędów?

Nie. Na przykład, przy jednoczesnych pomiarach temperatury w skalach Fahrenheita i Celsjusza można znaleźć dokładny związek postaci: $T^{\circ\mathrm{F}} = 32 + 1{,}8T^{\circ\mathrm{C}}$.

http://pl.wikipedia.org/wiki/Skala_Fahrenheita

Równanie to można przyjąć za definicję skali Fahrenheita za pomocą skali Celsjusza. Znajomość wartości jednej zm. wyznacza automatycznie wartość drugiej zm.

Model wyznaczający dokładne związki między zmiennymi nazywamy ***modelem deterministycznym***.

Model niedający jednoznacznego związku między zmiennymi nazywamy ***modelem probabilistycznym***.

Ogólna postać modelu probabilistycznego jest następująca:

zmienna zależna = składnik deterministyczny + błąd losowy

Błąd losowy modelu odgrywa kluczową rolę w analizie regresji.

Wszystkie rozważania dotyczą modeli parametrycznych, czyli takich, w których ogólna postać modelu jest założona z góry, a celem jest takie wyznaczenie parametrów modelu, aby powstała funkcja dobrze była dopasowana do danych uczących.

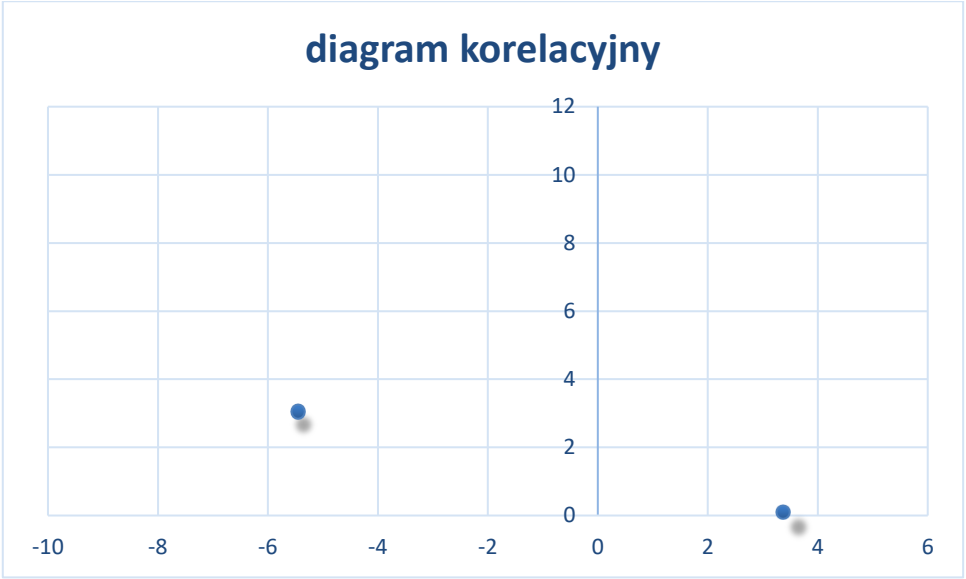
Przykład 1. Podać eksperyment pozwalający prognozować koszty ponoszone na energię cieplną, potrzebną do ogrzania mieszkania, w zależności od temperatury zewnętrznej.

Rozwiązanie. Eksperyment polega na dokonaniu n obserwacji średnich dziennych temperatur i średnich wielkości kosztów ponoszonych na ogrzanie mieszkania w miesiącach zimowych. Obserwacje prowadzone są od października do kwietnia, a więc przez 7 miesięcy. Wyniki są podane w tablicy danych.

Tablica. Dane źródłowe.

Miesiąc	Średnia temperatura powietrza (x_i [°C])	Dzienny koszt ogrzewania (y_i [zł])
październik	5	4
listopad	0	5
grudzień	–5	7
styczeń	–9	11
luty	–6	10
marzec	–2	4
kwiecień	2	3

Wstępne rozpoznanie związków pomiędzy badanymi zmiennymi przeprowadza się na podstawie ***diagramu korelacyjnego***.



Z wykresu można zauważyć, że przy wzroście temperatury otoczenia koszty ogrzewania maleją. Jeśli modelem zależności jest odcinek prostej przechodzącej przez diagram korelacyjny, to widać, że bez względu, jak będzie przeciągnięta prosta przez punkty, co najmniej jeden z nich zawsze będzie leżał poza dopasowaną prostą. Nie można więc w klasie modeli prostoliniowych określić modelu deterministycznego.

W takich sytuacjach można zastosować model prob., uwzględniający losowe odchylenia punktów od ustalonej linii.

Najprostszym modelem prob. jest warunkowa zm. l.:

$(Y|x) = \beta_0 + \beta_1 x + \varepsilon$ dla której stosowany jest zapis

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

gdzie

Y – zmienna zależna (*response variable*),

x – zmienna objaśniająca (*predictor variable*),

ε – błąd losowy (*random error*),

β_0 – parametr przesunięcia (*intercept*),

β_1 – parametr nachylenia (*slope*).

Jeżeli wartość oczekiwana błędu ε wynosi 0, to

$$\mathbb{E}Y = \beta_0 + \beta_1 x.$$

Jest to deterministyczne równanie prostej, którą zapisujemy

$$y = \beta_0 + \beta_1 x$$

β_0, β_1 są nieznanymi parametrami deterministycznej części modelu.

Ich estymatory są wyznaczane na podstawie próby (x_i, Y_i), $i = 1, 2, \ldots, n$, wówczas

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Rozkłady estymatorów zależą od rozkładu losowego składnika błędu ε. Aby je wyznaczyć należy poczynić pewne założenia dotyczące błędu ε.

2. Założenia prostej regresji liniowej

1. Losowe składniki ε_i modelu regresji mają wartość oczekiwaną 0 i stałą wariancję równą σ^2 (homoskedastyczność) bez względu na wartości x_i zmiennej objaśniającej X.
2. Błędy ε_i obserwacji zmiennej Y są nieskorelowane, czyli $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0, \forall i \neq j$.

Hipotetycznie przyjmowane założenia są sprawdzane za pomocą testów statystycznych.

3. Estymacja parametrów modelu metodą najmniejszych kwadratów (MNK)

Znajdowanie prostej „*najlepszego dopasowania*” do próby losowej

$$(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n),$$

sprowadza się do estymacji nieznanych parametrów β_0, β_1 modelu.

Estymatory $\hat{\beta}_0$, $\hat{\beta}_1$ parametrów β_0, β_1 są wyznaczane *metodą najmniejszych kwadratów* (method of least squares).

Istotę metody najmniejszych kwadratów można dostrzec na wykresie prostej nałożonej na diagram korelacyjny. Pionowe odcinki reprezentują *odchylenia* (deviations) punktów diagramu korelacyjnego od prostej.

Można wykazać, że istnieje wiele prostych, dla których *suma odchyleń* jest równa 0. Natomiast dla dokładnie jednej prostej ***suma kwadratów odchyleń*** osiąga minimum.

Suma kwadratów odchyleń zwie się również ***sumą kwadratów dla błędu*** (sum of squares for error) i jest oznaczana SSE.

Wyznaczone metodą najmniejszych kwadratów równanie nazywamy ***równaniem regresji*** (regression equation).

Estymatorem probabilistycznego modelu prostoliniowego jest prosta regresji

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

Predykcję $\hat{y}_i$ wartości zmiennej Y otrzymamy poprzez podstawienie wartości x_i do równania regresji, tj.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \; i = 1, 2, \ldots, n.$$

Odchylenie wartości y_i od jej predykcji $\hat{y}_i$ wynosi

$$y_i - \hat{y}_i = \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right).$$

Suma kwadratów odchyleń SSE dla wszystkich n par (x_i, y_i), $i = 1, 2, \ldots, n$ wyraża się wzorem:

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2.$$

Metoda najmniejszych kwadratów polega na takim oszacowaniu nieznanych parametrów modelu probabilistycznego, aby funkcja SSE osiągnęła minimum.

Estymatory $\hat{\beta}_0$, $\hat{\beta}_1$, dla których suma SSE osiąga minimum nazywamy ***estymatorami najmniejszych kwadratów*** parametrów β_0, β_1.

Estymatory te są wyznaczane z warunku koniecznego istnienia minimum, który dla funkcji SSE jest zarazem warunkiem dostatecznym. Obliczamy pochodne cząstkowe

$$\partial \mathrm{SSE}/\partial \hat{\beta}_0 = -2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right),$$
$$\partial \mathrm{SSE}/\partial \hat{\beta}_1 = -2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i.$$

Otrzymany układ równań normalnych przyjmuje postać:

$$\begin{cases} n\hat{\beta}_0 + \left(\sum x_i\right)\hat{\beta}_1 = \sum y_i \\ \left(\sum x_i\right)\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1 = \sum x_i y_i \end{cases}$$

lub w zapisie macierzowym

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Przy założeniu, że macierz odwrotna istnieje

$$\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix}\sum y_i \\ \sum x_i y_i\end{bmatrix}.$$

Wyznaczone z układu równań estymatory parametrów β_0 i β_1 wyrażają się wzorami:

$$\hat{\beta}_1 = \frac{(n\sum x_i y_i) - (\sum x_i)(\sum y_i)}{(n\sum x_i^2) - (\sum x_i)^2}$$

$$\hat{\beta}_0 = \bar{\boldsymbol{y}} - \hat{\beta}_1 \bar{\boldsymbol{x}}$$

Tak otrzymane estymatory są najefektywniejszymi i równocześnie nieobciążonymi estymatorami parametrów regresji liniowej. Parametr β_1 (ang. slope) jest współczynnikiem kierunkowym. Odpowiada on na pytanie, jaki jest przeciętny przyrost

wartości zmiennej objaśnianej na jednostkę przyrostu zmiennej objaśniającej.

Jeśli przyjmiemy oznaczenia

$$SS_{xy} = \left(\sum x_i y_i\right) - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$SS_{xx} = \left(\sum x_i^2\right) - \frac{(\sum x_i)^2}{n}$$

to:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Wielkości SS_{xx} i SS_{xy} nazywamy odpowiednio ***centrowaną sumą kwadratów*** dla x i ***centrowaną sumą*** *iloczynów* x i Y.

Przykład 2. Wyznaczyć prostą regresji kosztów ogrzewania mieszkania w zależności od temperatury otoczenia.

Rozwiązanie. Współczynniki równania regresji są wyznaczane komputerowo.

```
: to
5 0 –5 –9 –6 –2 2
: ko
4 5 7 11 10 4 3
: A GETS 2 2 RESHAPE (7,(SUM to),(SUM to),(SUM to^2))
: A
  7 –15
 –15 175
: (INVERSE A) MATMULT TRANSPOSE ((SUM ko),(SUM to*ko))
5.06 –0.572
```

Stąd $\hat{\beta}_0 = 5{,}06$, $\hat{\beta}_1 = -0{,}572$, czyli równanie regresji

$$\hat{y} = 5{,}06 - 0{,}572x.$$

4. Własności estymatorów parametrów modelu

Jeżeli błąd modelu ma rozkład normalny, to estymatory $\hat{\beta}_0$, $\hat{\beta}_1$ mają rozkłady normalne.

W szczególności

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma / \sqrt{SS_{xx}}\right)$$

Ponadto $\hat{\beta}_1$ jest estymatorem zgodnym i nieobciążonym parametru β_1 modelu.

5. Estymator wariancji błędu modelu

W praktyce wariancja σ^2 błędu modelu jest zwykle nieznana i trzeba ją estymować na podstawie próby losowej.

Wyznaczenie nieobciążonego estymatora wariancji jest oparte na następującym twierdzeniu.

Twierdzenie. Jeżeli są spełnione założenia modelu regresji oraz

$$S^2 = \mathrm{SSE}/(n-2),$$

to

$$\chi^2 = \mathrm{SSE}/\sigma^2 = (n-2)S^2/\sigma^2 \sim \mathrm{chis}(n-2)$$

tj. statystyka χ^2 ma rozkład *chi-kwadrat* z $n-2$ stopniami swobody.

Z podanego twierdzenia wynika, że

$$S^2 = \chi^2\sigma^2/(n-2),$$

więc wartość oczekiwana

$$\mathbb{E}(S^2) = \big(\sigma^2/(n-2)\big)\mathbb{E}(\chi^2) = \sigma^2,$$

czyli statystyka S^2 jest nieobciążonym estymatorem wariancji σ^2 błędu modelu ε. Do estymacji wariancji są wykorzystane $n-2$ stopnie swobody. Dwa stopnie swobody są pozostawione do estymacji dwóch parametrów modelu.

Przykład 3. Ocenić wariancję modelu dla danych z przykładu 1.

Rozwiązanie. Z obliczeń przeprowadzonych w przykładzie 2 wiemy, że oceną modelu regresji jest prosta $\hat{y} = 5{,}06 - 0{,}572x$. Ponieważ $n = 7$, więc do oceny wariancji σ^2 błędu jest wyko-

rzystane 5 stopni swobody. Obliczenia przeprowadzone pod interpreterem EXEC przedstawiono w ramce.

```
: SSyy GETS SUM(ko – AVERAGE ko)^2
: SSyy
59.4286
: SSxy GETS SUM (to – AVERAGE to)*(ko – AVERAGE ko)
: SSxy
–81.7143
: SSE GETS SSyy + 0.572*SSxy
: SSE
12.688
: MSE GETS SSE/5
: MSE
2.5376
: SQRT MSE
1.59298
```

Najpierw jest obliczana suma kwadratów dla błędu SSE.

Z przeprowadzonych obliczeń wynika, że $\mathrm{SSE} = 12{,}688$ oraz $\mathrm{MSE} = 2{,}5376$, natomiast ocena s odchylenia standardowego σ błędu ε wynosi $s = \sqrt{\mathrm{MSE}} = 1{,}59298$.

Na podstawie reguły sigm wiadomo, że ponad 95% obserwacji leży wewnątrz otoczenia $2s$ od linii regresji $\hat{y}$.

W przedstawionym przykładzie $2s = 3{,}18$ i wszystkie 7 punktów próby leżą w otoczeniu $2s$ od wyznaczonej prostej regresji.

6. Ocena adekwatności modelu

Oceny adekwatności modelu prostoliniowego dokonujemy poprzez ocenę parametru nachylenia.

Istota oceny jest oparta na spostrzeżeniu, że jeżeli zmienna x nie wnosi żadnych informacji do predykcji wartości zmiennej Y, to deterministyczna część modelu jest stała przy zmianie wartości zmiennej x.

W modelu prostoliniowym oznacza to, że parametr nachylenia β_1 jest równy 0 i ocena adekwatności modelu sprowadza się do testowania hipotezy zerowej

$$\mathrm{H}_0\colon \beta_1 = 0,$$

przeciw hipotezie alternatywnej, że zmienne są liniowo zależne.

Odrzucenie hipotezy zerowej na rzecz hipotezy alternatywnej oznacza, że zmienna x może być wykorzystana do predykcji wartości zmiennej losowej Y na podstawie oszacowanego modelu. Jeżeli założenia o błędzie losowym ε są spełnione, to otrzymany metodą najmniejszych kwadratów estymator $\hat{\beta}_1$ parametru nachylenia ma rozkład normalny

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma/\sqrt{SS_{xx}}\right).$$

Ponieważ parametr σ jest zwykle nieznany, więc odpowiednią statystyką testową jest statystyka t określona wzorem:

$$t = \frac{\hat{\beta}_1\sqrt{SS_{xx}}}{s}.$$

Tak określona statystyka t ma rozkład t–Studenta z $n-2$ stopniami swobody.

Przykład 4. Na poziomie istotności $\alpha = 0{,}05$ ocenić adekwatność modelu prostoliniowego wyznaczonego w przykładzie 2.

Rozwiązanie. Weryfikujemy hipotezę zerową $H_0: \beta_1 = 0$, przeciw lewostronnej hipotezie alternatywnej $H_0: \beta_1 < 0$. Z obliczeń w przykładzie 3 wiemy, że $s = 1{,}593$. Do obliczenia wartości statystyki testowej oraz sumy kwadratów SS_{xx} występującej we wzorze na statystykę testową korzystamy z interpretera EXEC.

```
: SSxx GETS SUM((to – AVERAGE(TO))*to)
: SSxx
142.857
: T GETS – 0.572*(SQRT(SSxx))/1.593
: T
–4.29172
: 5 STUDENT T
3.88818E–3
```

Stąd $SS_{xx} = 142{,}857$, natomiast wartość statystyki testowej $T = -4{,}29$

Zaobserwowana istotność testu $\alpha_0 = F_t(-4{,}29172)$, gdzie F_t jest dystrybuantą rozkładu *t-Studenta* z pięcioma stopniami swobody.

Ponieważ $\alpha_0 = 0{,}00389$, więc odrzucamy hipotezę zerową, na rzecz hipotezy alternatywnej orzekającej, że współczynnik nachylenia jest istotnie ujemny w przyjętym modelu.

Wyniki prowadzonych obliczeń w przykładach 1 – 4 można uzyskać korzystając z procedury REG.

Ekran akwizycji danych tej procedury składa się z pól podanych w ramce.

Dependent variable:
Independent variable:
Model: Linear
Confidence limits: 95.00
Prediction limits: 95.00
Point labels:

Arkusz wynikowy procedury *REG* składa się z trzech części.

Część górna zawiera oceny parametrów modelu (*estimate*), błędy standardowe (*standard error*), wartości *T* statystyk (*T value*), istotności testów *t-Studenta* (*prob. level*).

Część środkowa dotyczy analizy wariancji (*analysis of variance*) i zawiera sumy kwadratów, stopnie swobody, średnie kwadraty, statystykę *F* i istotność testu *F*.

Tablica. Arkusz wynikowy procedury *REG*

Regression Analysis – Linear model: Y = a + bX

Dependent variable: ko Independent variable: to

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	5.06	0.666393	7.59311	.00063
Slope	–0.572	0.133279	–4.29176	.00778

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	46.740571	1	46.740571	18.41920	.00778
Residual	12.688000	5	2.537600		
Total (Corr.)	59.428571	6			

Correlation Coefficient = –0.886848 R-squared = 78.65 percent

Stnd. Error of Est. = 1.59298

Dolna część arkusza wynikowego zawiera informacje dotyczące współczynnika korelacji (*correlation coefficient*), współczynnika determinacji (*R-squared*) oraz błędu standardowego estymacji (*stnd. error of est*.).

Sposób wykorzystania błędu standardowego do estymacji przedziałowej będzie pokazany dla parametru nachylenia.

7. Przedział ufności dla parametru nachylenia

Jeżeli założenia analizy regresji są spełnione, to końce $100(1-\alpha)$-procentowego przedziału ufności dla parametru nachylenia modelu prostoliniowego wyrażają się wzorem:

$$\hat{\beta}_1 \pm \frac{t_{\frac{\alpha}{2};n-2}\, s}{\sqrt{SS_{xx}}}$$

Przykład 5. Na podstawie danych z przykładu 1 wyznaczyć 95-procentowy przedział ufności dla parametru nachylenia modelu prostoliniowego.

Rozwiązanie. Z dotychczasowych obliczeń wiemy, że:

$\hat{\beta}_1 = -0{,}572,\ n-2=5,\ \alpha = 0{,}05,\ s = 1{,}593,\ SS_{xx} = 142{,}857.$

Formuła obliczeniowa oraz wynik obliczeń:

```
: –0.572+(5 INVSTUDENT 0.25 0.975)*1.593/SQRT(142.857)
–0.9146 –0.2293
```

Stąd 95-procentową realizacją przedziału ufności dla nieznanego parametru nachylenia β_1 jest przedział $(-0{,}915, -0{,}229)$.

Ponieważ końce przedziału są ujemne, więc można wnosić, że parametr nachylenia jest ujemny i wartość oczekiwana kosztów ogrzewania ma tendencję do malenia wraz ze wzrostem temperatury otoczenia.

Szeroki przedział ufności jest wynikiem małej liczebności danych i w konsekwencji niedostatku informacji.

8. Empiryczny współczynnik korelacji

Podstawowymi środkami do badania statystycznych relacji zachodzących między zmiennymi mierzalnymi są regresja i korelacja. Regresja pozwala badać kształt zależności, a korelacja siłę zależności liniowej. Współczynnik korelacji ρ jest jednym z parametrów wektora losowego (X, Y).

Celem tego punktu jest estymacja nieznanego parametru ρ wektora losowego (X, Y), będącego modelem populacji, w której interesują nas dwie cechy mierzalne.

Estymacja ta przeprowadzana jest na podstawie n par powiązanych zmiennych losowych (X_i, Y_i), tworzących próbę prostą. Realizację tej próby można przedstawić graficznie w postaci diagramu korelacyjnego.

Estymatorem zgodnym współczynnika korelacji liniowej jest empiryczny współczynnik korelacji liniowej r określony wzorem:

$$r_{XY} = \frac{\sum(X_i - \overline{\mathbf{X}})\,(Y_i - \overline{\mathbf{Y}})}{(n-1)S_X S_Y}$$

Mając obliczone wielkości SS_{xy}, SS_{xx}, SS_{yy} można dokonać oceny współczynnika korelacji korzystając ze wzoru:

$$r_{XY} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}.$$

W przeciwieństwie do estymatora parametru nachylenia, współczynnik korelacji jest wielkością niemianowaną.

Wartość współczynnika r mieści się zawsze w przedziale $[-1, 1]$, bez względu na jednostki zmiennych X i Y.

W poprzednim rozdziale adekwatność modelu prostoliniowego była oceniana na podstawie parametru nachylenia. Teraz jest pokazane, jak mierzyć siłę liniowego związku za pomocą współczynnika korelacji. Istnieje podobieństwo we wzorach na współczynnik r oraz estymator $\hat{\beta}_1$.

Suma SS_{xy} występuje w licznikach obydwu statystyk, a ponieważ ich mianowniki są zawsze dodatnie, więc będą zawsze tego samego znaku.

Wartość współczynnika korelacji bliska 0 oznacza mały lub żaden liniowy związek pomiędzy zmiennymi Y i X. Duża wartość bezwzględna współczynnika korelacji świadczy o dużej współzależności liniowej badanych cech.

Wysoki poziom korelacji nie musi oznaczać przyczynowości, tj. zmiany zmiennej X nie muszą być przyczyną zmian zmiennej Y.

Przykład 6. Ocenić współczynnik korelacji na podstawie danych z przykładu 1.

Rozwiązanie. Ponieważ $SS_{xx} = 142{,}857$, $SS_{yy} = 59{,}429$, $SS_{xy} = -81{,}714$, więc

$$r_{XY} = SS_{xy}/\sqrt{SS_{xx}SS_{yy}} = 81{,}714/\sqrt{(142{,}857)\cdot(59{,}429)} = -0{,}8868$$

Ocena współczynnika korelacji liniowej wynosi $-0{,}8868$. Wynik świadczy o dość silnej korelacji ujemnej.

Jeżeli badane cechy X i Y populacji mają dwuwymiarowy rozkład normalny o nieznanym współczynniku korelacji ρ, to hipoteza o braku korelacji liniowej między badanymi cechami $H_0: \rho = 0$ jest równoważna hipotezie zerowej $H_0: \beta_1 = 0$.

Różnicą pomiędzy parametrem nachylenia a współczynnikiem korelacji jest skala.

Parametr nachylenia daje dodatkową informację o wielkości zmian zmiennej Y na jednostkę zmian zmiennej x. Z tego powodu, do badania zależności liniowych pomiędzy zmiennymi zaleca się stosowanie testu dotyczącego parametru nachylenia.

Ocenę współczynnika korelacji można również wyznaczyć dla cech niemierzalnych. W tym celu należy najpierw uszeregować obserwacje według jakiegoś kryterium porządkującego i nadać numery miejsc zajmowanych przez obserwacje w uporządkowanym ciągu. Numery te nazywają się ***rangami***.

Dysponując dwoma ciągami rang, można ocenić współczynnik korelacji.

Rangi można nadać również cechom mierzalnym. Sposób rangowania danych mierzalnych jest podany w przykładzie.

Przykład 7. Ocenić współczynnik korelacji dla rang danych z przykładu 1 uszeregowanych w naturalnym porządku.

Rozwiązanie. Dane wyjściowe oraz ich rangi są zestawione.

Miesiąc	Temperatura otoczenia (w °*C*)	Rangi temp. otoczenia	Koszty ogrzewania (w zł)	Rangi kosztów ogrzewania
październik	5	7	4	2,5
listopad	0	5	5	4
grudzień	–5	3	7	5
styczeń	–9	1	11	7
luty	–6	2	10	6
marzec	–2	4	4	2,5
kwiecień	2	6	3	1

Ponieważ dzienne koszty ogrzewania w październiku oraz w marcu są równe, więc rangi przypadające na te miesiące są ich średnią arytmetyczną i wynoszą 2,5. Wyniki rankingu danych można otrzymać, korzystając z operatora *RANK*. Do obliczenia oceny macierzy korelacyjnej jest wykorzystany operator *CORRMAT* z biblioteki *REGSGRP*.

```
: TOR GETS RANK TO
: TOR
7 5 3 1 2 4 6
: KOR GETS RANK KO
: KOR
2.5 4 5 7 6 2.5 1
```

```
: CORRMAT TOR WITH KOR
 1        –0.882919
–0.882919  1
: CORRMAT TO WITH KO
 1        –0.886848
–0.886848  1
```

Oceny współczynników korelacji liniowej dla danych rangowych oraz nierangowych wynoszą:

$$r(TOR; KOR) = -0{,}882919,\ r(TO; KO) = -0{,}886848.$$

Mimo małej próby różnica ocen współczynnika korelacji jest nieznaczna.

9. Test dla współczynnika korelacji

Jeśli odrzucimy hipotezę o braku korelacji w populacji

$$(X, Y) \sim \mathcal{N}(m_1, m_2, \sigma_1, \sigma_2, \rho),$$

to możemy sprawdzić hipotezę, że współczynnik korelacji liniowej ρ jest równy liczbie ρ_0, tj.

$$\mathrm{H}_0 : \rho = \rho_0$$

Przy założeniu, że $|r| \neq 1$ do sprawdzenia tej hipotezy korzystamy ze statystyki

$$Z = (U - u_0)\sqrt{n - 3},$$

gdzie

$$U = \frac{1}{2}\ln\frac{1+r}{1-r}$$

$$u_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2}$$

której rozkład dla $n \geq 7$ ma dość dobre przybliżenie standardowym rozkładem normalnym.

Przykład 8. Sprawdzić hipotezę, że współczynnik korelacji dla badanych cech w przykładzie 1 jest mniejszy od $-0{,}5$.

Rozwiązanie. Zakładamy, że badana populacja ma dwuwymiarowy rozkład normalny. Podaną hipotezę ustawiamy jako lewostronną hipotezę alternatywną. Jako hipotezę zerową przyjmujemy, że współczynnik korelacji wynosi co najmniej $-0{,}5$.

Czyli

$$\mathrm{H}_0: \rho \geq -0{,}5,$$

$$\mathrm{H}_1: \rho < -0{,}5.$$

Do sprawdzenia hipotezy zerowej korzystamy z podanej statystyki Z. W przykładzie 6 wyznaczona została ocena współczynnika korelacji $r = -0{,}8868$. Podstawiając dane otrzymujemy

$$U = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{0{,}1132}{1{,}8868} = -1{,}4067,$$

$$u_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = \frac{1}{2}\ln\frac{0{,}5}{1{,}5} - \frac{0{,}5}{12} = -0{,}5910.$$

Wartość statystyki testowej Z wynosi

$$Z = (U - u_0)\sqrt{n-3} = (-1{,}4067 + 0{,}5910)\sqrt{7-3} = -1{,}6314,$$

stąd przybliżona istotność testu (p-value)

$$\alpha_0 \approx \Phi(-1{,}6314) = 0{,}0514 > 0{,}05,$$

gdzie Φ jest dystrybuantą rozkładu $\mathcal{N}(0;1)$.

Wniosek. Nie ma podstaw do odrzucenia hipotezy zerowej. Współczynnik korelacji nie jest istotnie mniejszy od $-0{,}5$.

10. Współczynnik determinacji

Jako miarę wkładu zmiennej X do predykcji zmiennej Y można przyjąć zmniejszenie się błędu predykcji spowodowane wykorzystaniem informacji dostarczonej przez zmienną X.

Przy założeniu, że zmienna X nie wnosi informacji dla predykcji zmiennej Y, najlepszą predykcją dla wartości zmiennej Y jest średnia arytmetyczna $\overline{\mathbf{Y}}$. Suma kwadratów odchyleń dla modelu $\hat{y} = \overline{\mathbf{Y}}$ jest określona wzorem:

$$SS_{YY} = \sum (Y_i - \overline{\boldsymbol{Y}})^2$$

Jeżeli zmienna X nie wnosi żadnych informacji lub wnosi mało do predykcji wartości zmiennej Y, to sumy kwadratów odchyleń SS_{yy} i SEE $= \sum \left(Y_i - \hat{y}_i\right)^2$ będą prawie sobie równe.

Jeżeli zmienna X wnosi informację do predykcji zmiennej Y, to suma SSE będzie mniejsza niż suma SS_{yy}.

W szczególności, gdy wszystkie pary (x_i, y_i), $i = 1, \ldots, n$ spełniają równanie regresji, to $\text{SSE} = 0$.

Ponieważ suma SS_{yy} jest miarą rozrzutu zaobserwowanej próby od średniej $\bar{\mathbf{y}}$, natomiast SSE jest miarą pozostałego niewyjaśnionego rozrzutu danych od dopasowanej prostej $\hat{y}$, więc różnica $(SS_{yy} - \text{SSE})$ jest miarą wyjaśnienia rozrzutu wnoszonego do liniowego związku przez zmienną X.

Udział zmiennej X w zmniejszaniu rozrzutu wyraża się wzorem:

$$(SS_{yy} - \text{SSE})/SS_{yy}$$

i nazywa się ***empirycznym współczynnikiem determinacji***.

Można wykazać, że w przypadku regresji prostoliniowej empiryczny współczynnik determinacji jest równy kwadratowi empirycznego współczynnika korelacji, stąd oznaczamy go symbolem r^2, czyli

$$r^2 = \frac{(SS_{yy} - \mathrm{SSE})}{SS_{yy}} = 1 - \frac{\mathrm{SSE}}{SS_{yy}}.$$

Oczywiście $r^2 \in [0, 1]$. Przykładowo, jeżeli $r^2 = 0{,}6$, to suma kwadratów odchyleń wartości zmiennej Y od jej prognozowanych wartości, na podstawie modelu, zmniejsza się o 60%, przy wykorzystaniu równania regresji $\hat{y}$ zamiast średniej $\bar{y}$ do predykcji wartości zmiennej Y.

Utrzymuje się, że gdy co najmniej połowa zmienności jednej zmiennej losowej jest wyjaśniona zmiennością drugiej, to warto wyznaczać współczynniki równania regresji.

Przykład 9. Obliczyć empiryczny współczynnik determinacji dla danych z przykładu 1 i zinterpretować wynik.

Rozwiązanie. Z wcześniejszych obliczeń wiadomo, że

$$\mathrm{SEE} = \sum(y_i - \hat{y}_i)^2 = 12{,}689.$$

Z kolei

$$SS_{yy} = \sum \hat{y}_i^2 - \frac{1}{7}(\sum y_i)^2 = 336 - 1936/7 = 59{,}42857.$$

Stąd empiryczny współczynnik determinacji

$$r^2 = 1 - \mathrm{SSE}/SS_{yy} = 0{,}786.$$

Empiryczny współczynnik korelacji $r = -0{,}8868$, więc empiryczny współczynnik determinacji można policzyć jako jego kwadrat.

Wniosek. Wykorzystanie temperatury otoczenia jako zmiennej X do predykcji kosztów ogrzewania jako zmiennej Y, na podstawie prostej najmniejszych kwadratów

$$\hat{y} = 5{,}06 - 0{,}572x,$$

zmniejsza o 78,9% całkowitą sumę kwadratów odchyleń próby od średniej arytmetycznej $\bar{y}$.

11. Elementy analizy regresji – podsumowanie

1. Przyjęcie modelu probabilistycznego jako hipotetycznego modelu, wiążącego badane zmienne.
2. Wyznaczenie ocen nieznanych parametrów deterministycznej części modelu na podstawie danych empirycznych dotyczących badanych zmiennych.
3. Sprawdzenie założeń analizy regresji i estymacja wariancji błędu modelu.
4. Sprawdzenie adekwatności modelu. W tym celu testowana jest hipoteza zerowa o braku zależności pomiędzy badanymi zmiennymi - przeciw jednej z hipotez alternatywnych. Jeśli założenia analizy regresji są spełnione, statystyka testowa

$$T = \frac{\hat{\beta}_1 \sqrt{SS_{xx}}}{s}$$

ma rozkład t-Studenta z $n-2$ stopniami swobody. Obliczana jest istotność testu i podejmowana decyzja statystyczna, co do hipotezy zerowej.

5. Dodatkową informację uzyskuje się na podstawie $(1-\alpha)100$-procentowego przedziału ufności dla parametru nachylenia.
6. Miarą adekwatności modelu liniowego są również empiryczne współczynniki korelacji i determinacji. Współczynnik determinacji pozwala na podstawie próby ustalić stopień wyjaśnienia zmiennej zależnej przez zmienną niezależną.
7. Zastosowanie modelu do estymacji i predykcji punktowej i przedziałowej.

13. Zestaw zadań

1. (KA) Sporządzić diagram rozrzutu, wyznaczyć oceny współczynników korelacji i determinacji, wyznaczyć równania prostych regresji (Y względem X, X względem Y), błędy standardowe estymacji oraz wykreślić równanie regresji dla podanych prób:

a) [x; y] = {[5.5, 1.5], [8.5, 4.0], [4.0, 2.0], [8.0, 7.5], [2.5, 0.5], [8.0, 5.0], [8.5, 8.5], [3.5, 1.0], [6.5, 2.5], [9.0, 8.0], [0.5, 1.0], [8.5, 6.5], [7.5, 3.5], [1.5, 1.0], [8.5, 9.5], [2.0, 1.5], [8.0, 9.0], [7.5, 5.5], [9.0, 5.5], [7.0, 1.5], [7.5, 7.0], [5.0, 0.5], [4.5, 1.5], [5.5, 2.5], [6.5, 4.0]}.

b)[x; y] = {[3.4, 3.7], [2.7, 4.7], [4.4, 4.6], [2.6, 2.5], [5.2, 5.3], [3.1, 4.6], [2.2, 3.5], [3.3, 4.1], [6.0, 5.3], [4.0, 5.4], [2.0, 2.7], [3.9, 5.0], [2.5, 1.5], [2.5, 4.3], [3.6, 3.0], [6.4, 5.1], [2.8, 3.7], [4.3, 5.8], [5.7, 5.5], [2.5, 3.2], [4.9, 5.0], [3.0, 1.8], [3.6, 4.3], [5.7,4.9], [3.0, 1.0], [4.1, 4.1], [5.0, 4.8], [2.2, 2.0], [3.7, 3.4], [5.0, 5.7], [3.1, 4.4], [3.4,

5.4], [3.4, 2.3], [2.5, 2.9], [5.3, 5.0], [4.1,4.6], [3.0, 5.0], [2.8, 2.3], [3.0, 3.9], [2.4, 3.9], [4.5, 5.5], [3.5, 5.0], [4.8, 5.3], [3.1, 2.5], [2.7, 4.1], [3.0, 3.3], [4.2, 5.0], [3.3, 2.2], [3.6, 3.9], [3.4, 4.7]}.

Odp.: a) $r = 0.7884$, $y = -1.5810 + 0.9152x$,
$y = 4.9916 + 1.4725x$.

2. Odnotowano miesięczne dochody przypadające na jednego członka rodziny (w *zł*) – cecha X oraz wyrażoną w procentach część budżetu rodzinnego przeznaczoną na zakup artykułów żywnościowych i utrzymanie mieszkania – cecha Y.

Wyniki

X	200	300	150	225	175	350	150	250	325	250
Y	70	80	95	75	90	60	60	65	85	90

Sporządzić diagram rozrzutu, wyznaczyć oceny współczynników korelacji i determinacji między dochodem przypadającym na jednego członka rodziny a wydatkami na artykuły żywnościowe i utrzymanie mieszkania.

Odp. Empiryczny współczynnik korelacji $r = -0.1965$. Statystyka testowa wynosi -0.5658, brak podstaw do odrzucenia hipotezy zerowej. Nie istnieje ujemna korelacja między badanymi cechami badanej populacji rodzin.

3. Naturalne jest przekonanie, że powinna być silna korelacja pomiędzy miesięcznymi obrotami firmy a jej liczebnością personelu handlowego. Dla pewnej firmy zostały zebrane dane dotyczące liczby sprzedawców w ostatnich 10 kwartałach oraz osiągane średniomiesięczne obroty (w mln *zł*) w tym czasie. Wynoszą one: [15, 1.35], [18, 1.63], [24, 2.33], [22, 2.41], [25, 2.63], [29, 2.93], [30, 3.41],

[32, 3.26], [35, 3.63], [38, 4.15]. Sprawdzić, czy to przekonanie potwierdziło się dla badanej firmy. Odp. $r = 0.99$.

4. Sformułować problem regresyjny i na podstawie danych przeprowadzić analizę regresji i korelacji.