Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

STATYSTYKA MATEMATYCZNA

z pakietem R

VII. Analiza regresji

Przemysław Grzegorzewski
Konstancja Bobecka-Wesołowska
Marek Gągolewski

Spis treści

1 Wprowadzenie

1.1 Współczynniki korelacji liniowej i rangowej

W programie `R` współczynniki korelacji wyznaczamy za pomocą funkcji `cor()`. Domyślnie wyliczany jest próbkowy współczynnik korelacji liniowej Pearsona. Wektory `x` i `y`, zawierające wartości badanych prób, podajemy jako pierwszy i drugi argument tej funkcji.

Jeśli chcemy obliczyć współczynnik korelacji rangowej Spearmana bądź Kendalla, jako kolejny argument funkcji `cor()`, podajemy, odpowiednio, `method="spearman"` albo `method="kendall"`.

Do weryfikacji hipotezy o istotności współczynnika korelacji liniowej ϱ, tzn.

$$\begin{aligned} H &: \varrho = 0, \\ K &: \varrho \neq 0, \end{aligned}$$

służy funkcja `cor.test()`. Wektory zawierające wartości prób, na podstawie których przeprowadzamy test, podajemy jako pierwszy i drugi argument tej funkcji. Jeśli chcemy testować istotność współczynnika korelacji rangowej Spearmana albo Kendalla, jako kolejny parametr podajemy `method="spearman"` bądź `"kendall"`.

1.2 Regresja prosta liniowa

Do budowy modelu regresji liniowej służy funkcja `lm()` (ang. *linear model*). Za jej pomocą wyznaczamy estymatory współczynników $\beta_0, \beta_1, \ldots, \beta_n$ w równaniu regresji postaci

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon, \tag{1}$$

gdzie $X_1, \ldots, X_n$ — zmienne niezależne (objaśniające), Y — zmienna zależna (objaśniana), ε — wyraz odpowiadający błędowi losowemu. Ponadto, procedura ta umożliwia sprawdzenie dopasowania modelu.

Pierwszym argumentem funkcji `lm()` jest tzw. *formuła* opisująca model, tzn. symboliczny opis zależności między zmiennymi. Jej składnia jest następująca:

$$\texttt{Y} \sim \texttt{X}_1 + \texttt{X}_2 + \cdots + \texttt{X}_\texttt{n}. \tag{2}$$

Więcej informacji na temat formuł znajdzie Czytelnik w dokumentacji: `?lm`.

Nas interesować będzie tutaj przypadek $n = 1$, zwany regresją liniową prostą. Zmienną zależną wyrażamy jako liniową funkcję jednej zmienne niezależnej:

$$Y = a + bX + \varepsilon. \tag{3}$$

Informację o dopasowanym modelu zwracanym przez `lm()` można zapisać jako zmienną. Możliwe jest wtedy na niej wykonywanie m.in. następujących funkcji: `summary()` — informacje szczegółowe, `predict()` — prognozowanie, `plot()` — diagnostyka za pomocą wykresów.

Np. przypuśćmy, że chcemy dopasować model regresji liniowej opisujący zależność wielkości brukselek (`y`) od wagi ich nasion (`x`).

```
> x <- c(4.2,5.4,6.4,7.3,8.2,9.1) # średnica brukselki w cm
> y <- c(2,2.8,3.1,3.8,3.9,4.8)   # waga nasion w g
> lm(y~x)
```

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

Coefficients:
(Intercept)            x
    -0.2037       0.5326
```

Jeśli zaobserwowane wartości `x` i `y` przechowywane są jako kolumny w pewnej ramce danych, to nazwę tej ramki podajemy jako argument parametru `data` funkcji `lm()`, np.

```
> lm(y~x, data=dane) # co jest równoważne następującemu zapisowi:
> lm(dane$y~dane$x)
```

Aby otrzymać pełen opis dopasowania modelu, używamy funkcji `summary()`.

```
> nasz_model <- lm(y~x)
> summary(nasz_model)
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5        6
-0.03306  0.12785 -0.10472  0.11596 -0.26336  0.15733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.20375    0.31736  -0.642 0.555817
x            0.53257    0.04556  11.689 0.000306 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1844 on 4 degrees of freedom
Multiple R-squared: 0.9716,        Adjusted R-squared: 0.9644
F-statistic: 136.6 on 1 and 4 DF,  p-value: 0.0003063
```

Możemy też odwoływać się bezpośrednio do wybranych elementów tego opisu, np.

```
> nasz_model$coefficients # oszacowania współczynników równania regresji
(Intercept)           x
 -0.2037459   0.5325733

> nasz_model$residuals    # reszty
          1           2           3           4           5           6
-0.03306189  0.12785016 -0.10472313  0.11596091 -0.26335505  0.15732899

> opis <- summary(nasz_model);
> opis$r.squared          # współczynnik determinacji R^2
[1] 0.9715563
```

Funkcja `predict()` służy do prognozowania, na podstawie dopasowanego wcześniej modelu, wartości zmiennej zależnej dla niezaobserwowanych jeszcze wartości zmiennej objaśniającej.

Np. gdy chcemy przewidzieć wielkość brukselki, która wyrośnie z nasionka o wadze 8 g, możemy napisać

```
> nowy <- data.frame(x=8)
> predict(nasz_model, nowy) # prognoza y dla nowy$x na podstawie modelu nasz_model
      1
4.05684
```

2 Zadania rozwiązane

Zadanie 7.1. Wygeneruj próbę $n = 200$-elementową z rozkładu wektora losowego (X, Y),

a) o rozkładzie dwuwymiarowym normalnym NN$(0, 1, 0, 1, 0)$,

b) o rozkładzie dwuwymiarowym normalnym NN$(0, 1, 2, 1, 0{,}6)$.

Dla uzyskanych próbek oszacuj współczynniki korelacji, zweryfikuj hipotezę o niezależności zmiennych X i Y oraz o istotności współczynnika korelacji liniowej.

Rozwiązanie.
Realizację próby z rozkładu NN$(0, 1, 0, 1, 0)$ można stworzyć, generując 2 nieskorelowane wektory danych z rozkładów N$(0, 1)$.

```
> n <- 200;
> x <- rnorm(n);
> y <- rnorm(n);
```

Współczynniki korelacji:

```
> cor(x, y);

[1] 0.0584777


> cor(x, y, method="pearson"); # to samo co wyżej

[1] 0.0584777


> cor(x, y, method="spearman");

[1] 0.0561689


> cor(x, y, method="kendall");

[1] 0.0358794
```

Test istotności współczynnika korelacji liniowej:

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

        Pearson's product-moment correlation

data:  x and y
t = 0.8243, df = 198, p-value = 0.4108
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08091984  0.19563150
sample estimates:
      cor
0.0584777
```

Macierz kowariancji rozkładu $\mathrm{NN}(\mu_1, \sigma_1, \mu_2, \sigma_2, \varrho)$, ma postać:

$$\mathbf{C} = \begin{bmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{4}$$

czyli w naszym przypadku

```
> (C <- matrix(c(1, 0.6, 0.6, 1), nrow=2))

     [,1] [,2]
[1,]  1.0  0.6
[2,]  0.6  1.0
```

Do wygenerowania n-elementowej próby z rozkładu $\mathrm{NN}(0, 1, 2, 1, 0{,}6)$ najprościej jest wykorzystać funkcję `mvrnorm()` z biblioteki `MASS()`.

```
> library("MASS"); # ładujemy bibliotekę
> proba <- mvrnorm(n, c(0, 2), C);
> head(proba); # wyświetlmy kilka pierwszych wierszy

           [,1]       [,2]
[1,] -2.2273015 -0.3932214
[2,]  0.6390456  1.2566023
[3,] -0.1373123  0.5763051
[4,]  0.5279900  3.9281826
[5,] -0.6789030  1.8853051
[6,]  1.3001541  4.5933153
```

Współczynniki korelacji:

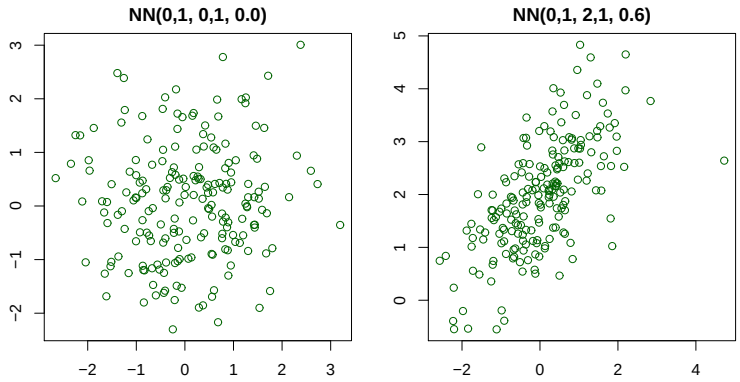
```
> cor(proba[,1], proba[,2]);

[1] 0.6392752


> cor(proba[,1], proba[,2], method="spearman");

[1] 0.6526888


> cor(proba[,1], proba[,2], method="kendall");

[1] 0.4686432
```

Test istotności współczynnika korelacji liniowej Pearsona:

```
> cor.test(proba[,1], proba[,2]);

        Pearson's product-moment correlation

data:  proba[, 1] and proba[, 2]
t = 11.6979, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5492491 0.7146328
sample estimates:
      cor
0.6392752
```

Porównajmy na koniec wykresy rozproszenia obydwu prób:

```
> par(mfrow=c(1,2))
> plot(x, y, main="NN(0,1, 0,1, 0.0)", col="darkgreen");
> plot(proba[,1], proba[,2], main="NN(0,1, 2,1, 0.6)", col="darkgreen");
```



□

Zadanie 7.2. Dwaj profesorowie postanowili dokonać oceny zdolności swoich 11 ulubionych dyplomantów. W tym celu każdy z belfrów uszeregował uczniów od najzdolniejszego do najmniej zdolnego:

Student	A	B	C	D	E	F	G	H	I	J	K
Profesor X	1,	7,	8,	3,	6,	10,	9,	2,	11	4,	5
Profesor Y	4,	8,	10,	1,	5,	9,	11,	3,	7,	2,	6

Czy można uznać, że istnieje zależność między opiniami obu profesorów? Przyjmij poziom istotności 0,05.

Rozwiązanie.

```
> x <- c(1, 7, 8, 3, 6, 10, 9, 2, 11, 4, 5)
> y <- c(4, 8, 10, 1, 5, 9, 11, 3, 7, 2, 6)
```

Mamy dane wektory, zawierające profesorskie rangi zdolności poszczególnych studentów. Obliczamy więc wartość próbkowego współczynnika korelacji rangowej (nie liniowej!):

```
> cor(x, y, method="spearman")

[1] 0.7909091
```

Test istotności dla współczynnika korelacji rangowej Spearmana:

```
> cor.test(x, y, method="spearman")

        Spearman's rank correlation rho

data:  x and y
S = 46, p-value = 0.00562
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7909091
```

Na poziomie istotności $\alpha = 0{,}05$ odrzucamy hipotezę o braku zależności między badanymi zmiennymi (test istotności współczynnika ρ Spearmana, $S = 46$, p-value $= 0{,}00562$). Można więc przyjąć, że istnieje istotny, dodatni związek między ocenami dokonanymi przez profesorów.
⊡

Zadanie 7.3. W zamieszczonej poniżej tabeli podano wysokość rocznego dochodu i wartość posiadanego domu dziewięciu rodzin wybranych w sposób losowy spośród mieszkańców pewnego osiedla:

Roczny dochód (tys. zł.)	360,	640,	490,	210,	280,	470,	580,	190,	320
Wartość domu (mln. zł.)	1.49,	3.10,	2.60,	0.92,	1.26,	2.42,	2.88,	0.81,	1.34

a) Wyznacz prostą regresji wartości domu względem dochodu.

b) Przeanalizuj dopasowanie modelu.

c) Oszacuj wartość domu rodziny, której roczny dochód wynosi 400 000 zł.

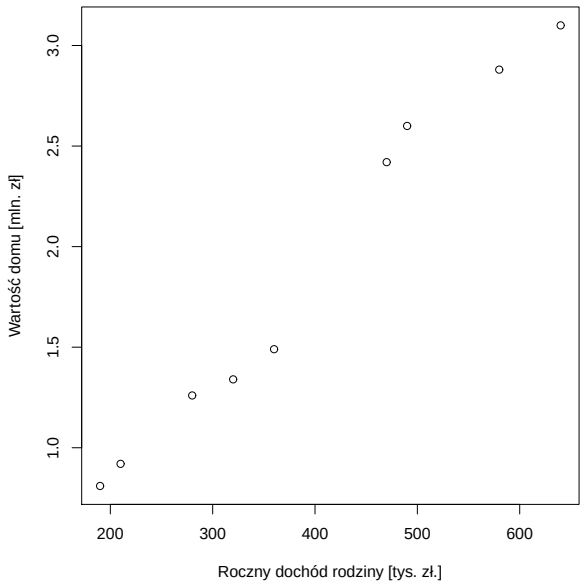
d) Wyznacz 95% przedział ufności dla szacowanej wartości domu tej rodziny.

Rozwiązanie.
Wprowadźmy dane do programu `R`:

```
> dochod <- c(360, 640, 490, 210, 280, 470, 580, 190, 320);
> dom    <- c(1.49, 3.10, 2.60, 0.92, 1.26, 2.42, 2.88, 0.81, 1.34);
```

Najpierw sprawdźmy na wykresie, czy zależność między zmienną niezależną X (dochód), a objaśnianą Y (wartość domu), jest typu liniowego:

```
> plot(dochod, dom, xlab="Roczny dochód rodziny [tys. zł.]",
       ylab="Wartość domu [mln. zł]");
```



Dopasowujemy więc model prostej regresji liniowej:

```
> modl <- lm(dom~dochod)
> (opis <- summary(modl))

Call:
lm(formula = dom ~ dochod)

Residuals:
      Min        1Q    Median        3Q       Max
-0.197759 -0.109247  0.006951  0.047323  0.205835

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2684395  0.1281678  -2.094   0.0745 .
dochod       0.0054339  0.0003042  17.863 4.25e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1378 on 7 degrees of freedom
Multiple R-squared: 0.9785,        Adjusted R-squared: 0.9755
F-statistic: 319.1 on 1 and 7 DF,  p-value: 4.254e-07
```

Wyestymowane parametry modelu możemy odczytać z kolumny `Estimate` tabelki `Coefficients`, bądź ręcznie:

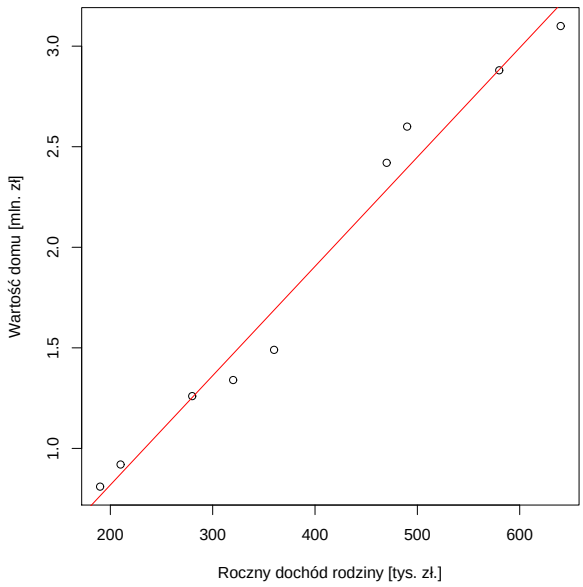
```
> modl$coefficients

 (Intercept)       dochod
-0.268439463  0.005433886
```

`(Intercept)` oznacza wyraz wolny $\hat{a}$, `dochod` — współczynnik $\hat{b}$ przy zmiennej X, zatem nasz model jest postaci:

$$Y = -0{,}2684 + 0{,}0054X + \varepsilon.$$

Wykres obserwacji wraz z prostą regresji:

```
> plot(dochod, dom, xlab="Roczny dochód rodziny [tys. zł.]",
       ylab="Wartość domu [mln. zł]");
> abline(modl, col="red")
```



Sprawdźmy jakość dopasowania. Weryfikację poprawności modelu możemy przeprowadzić za pomocą odczytania wyników zwracanych przez funkcję `summary()`.

Weryfikacja poprawności modelu

Analiza współczynnika determinacji. Współczynnik R^2 opisuje „frakcję zmienności zmiennej zależnej opisywanej przez model".

```
> opis$r.squared;

[1] 0.9785324
```

Test F (analiza wariancji w analizie regresji). Hipoteza zerowa $H : b = 0$ (nie ma zależności liniowej między zmiennymi) kontra $K : b \neq 0$.

```
> opis$fstatistic;

   value    numdf    dendf
319.0721   1.0000   7.0000


> anova(modl)

Analysis of Variance Table

Response: dom
          Df Sum Sq Mean Sq F value    Pr(>F)
dochod     1 6.0590  6.0590  319.07 4.254e-07 ***
Residuals  7 0.1329  0.0190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> anova(modl)$"Pr(>F)"[1] # p-value

[1] 4.254379e-07
```

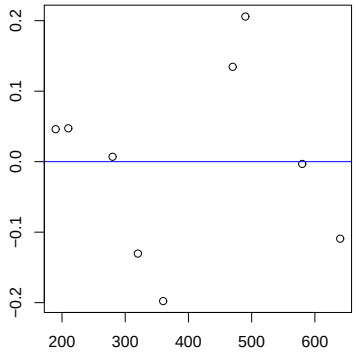
Test t (test istotności współczynników regresji). Testowanie hipotez $H_1 : a = 0, H_2 : b = 0$ przeciwko $K_1 : a \neq 0, K_2 : b \neq 0$ (współczynniki są istotnie różne od zera).

```
> opis$coefficients;

               Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -0.268439463 0.1281677764 -2.094438 7.448037e-02
dochod       0.005433886 0.0003042048 17.862589 4.254379e-07
```

Analiza reszt. Wykreślmy wykres reszt w zależności od X.

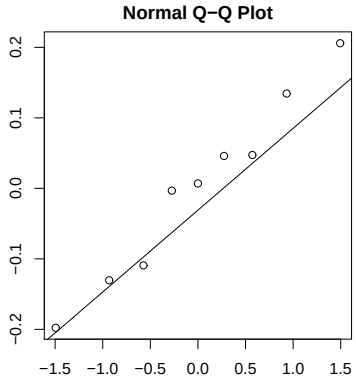
Analiza reszt

```
> plot(dochod, modl$residuals)
> abline(h=0, col="blue")
```



Narysujmy wykres normalności reszt.

```
> qqnorm(modl$residuals);
> qqline(modl$residuals);
```



Weryfikacja hipotezy o normalności rozkładu reszt testem Shapiro-Wilka:

```
> shapiro.test(modl$residuals);

        Shapiro-Wilk normality test

data:  modl$residuals
W = 0.9706, p-value = 0.9002
```

Test istotności dla współczynnika korelacji liniowej Pearsona:

```
> cor.test(dochod, dom)

        Pearson's product-moment correlation

data:  dochod and dom
t = 17.8626, df = 7, p-value = 4.254e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9476475 0.9978124
sample estimates:
     cor
0.989208
```

Prognozowanie:

Prognozowanie

```
> (nowe <- data.frame(dochod=400)) # uwaga na nazwę kolumny!

  dochod
1    400


> predict(modl, nowe)

       1
1.905115
```

95% przedział ufności dla szacowanej wartości:

```
> predict(modl, nowe, interval="prediction", level=0.95) # przedział dla predykcji
```

```
       fit      lwr      upr
1 1.905115 1.561606 2.248624
```

```
> predict(modl, nowe, interval="confidence", level=0.95) # przedział dla wartości średniej

       fit      lwr      upr
1 1.905115 1.796393 2.013837
```

Granice przedziałów odczytujemy z kolumn `lwr` (ang. *lower* — dolna) i `upr` (ang. *upper* — górna).

□

Zadanie 7.4. W poniższej tabeli podano liczbę ludności USA (w milionach) w latach 1890-2007:

Rok	1890,	1900,	1910,	1920,	1930,	1940,	1950,
Ludność	62.947,	75.994,	91.972,	105.710,	122.775,	131.669,	150.697,
Rok	1960,	1970,	1980,	1990,	2000,	2007	
Ludność	179.323,	203.235,	226.542,	248.718,	281.422,	301.140	

a) Przyjmując wykładniczy model wzrostu populacji, oszacuj parametry tego modelu i zweryfikuj jego dopasowanie.

b) Oszacuj przewidywaną wielkość populacji USA w 2010 i w 2020 roku.

Rozwiązanie.

```
> rok <- c(1890,1900,1910,1920,1930,1940,1950,1960,1970,1980,1990,2000,2007);
> pop <- c(62.947,75.994,91.972,105.710,122.775,131.669,150.697,
  179.323,203.235,226.542,248.718,281.422,301.140)
```

Model wykładniczy: $y = \exp(a + bx)$
Linearyzacja modelu wykładniczego: $z := \ln(y)$ i stąd $z = a + bx$.

```
> logpop <- log(pop);
> usa    <- lm(logpop~rok);
> summary(usa);

Call:
lm(formula = logpop ~ rok)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0886914 -0.0278510  0.0006202  0.0378818  0.0569246

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.045e+01  6.516e-01  -31.39 4.07e-12 ***
rok          1.306e-02  3.341e-04   39.09 3.72e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04464 on 11 degrees of freedom
Multiple R-squared: 0.9929,        Adjusted R-squared: 0.9922
F-statistic:  1528 on 1 and 11 DF,  p-value: 3.719e-13
```

Prognozowanie:

```
> nowyrok <- data.frame(rok=c(2010,2020));
> (nowylogpop <- predict(usa, nowyrok, interval="prediction"))

       fit      lwr      upr
1 5.798317 5.687156 5.909479
2 5.928929 5.814638 6.043219


> exp(nowylogpop)   # przekształcenie odwrotne do log()

       fit      lwr      upr
1 329.7443 295.0534 368.5140
2 375.7517 335.1700 421.2469
```

□

Zadanie 7.5. Dokonano ośmiu niezależnych pomiarów wielkości drgań pionowych (w cm) gruntu powstałych w wyniku trzęsienia ziemi w różnej odległości od epicentrum trzęsienia (w km). Otrzymano następujące wyniki:

Odległość	20,	30,	40,	50,	80,	140,	200,	250
Drganie	4.8,	3.2,	2.5,	2.5,	1.5,	1.8,	1.2,	0.8

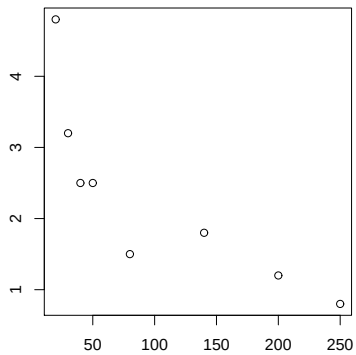
a) Wyznacz funkcję regresji wielkości drgań gruntu względem odległości od epicentrum.

b) Zweryfikuj dopasowanie modelu.

c) Oszacuj wielkość drgań w odległości 100 km od epicentrum.

Rozwiązanie.

```
> odl <- c(20,  30,  40,  50,  80,  140, 200, 250);
> wys <- c(4.8, 3.2, 2.5, 2.5, 1.5, 1.8, 1.2, 0.8);
> plot(odl,wys)
```



Szukamy najlepszego modelu.

Model wykładniczy: $y = \exp(a + bx)$. Linearyzacja modelu wykładniczego: $z := \ln(y)$ i stąd $z = a + bx$.

```
> yp <- log(wys);
> t1 <- lm(yp~odl);
> summary(t1);

Call:
lm(formula = yp ~ odl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41291 -0.09903  0.01523  0.10161  0.38665

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.303166   0.142135   9.169 9.48e-05 ***
odl         -0.006060   0.001099  -5.516  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2503 on 6 degrees of freedom
Multiple R-squared: 0.8353,        Adjusted R-squared: 0.8078
F-statistic: 30.42 on 1 and 6 DF,  p-value: 0.001493
```

Model multiplikatywny (potęgowy): $y = ax^b$. Linearyzacja modelu multiplikatywnego: $z := \ln(y)$, $u := \ln(x)$, $a' := \ln(a)$ i stąd $z = a' + bu$.

```
> xp <- log(odl);
> t2 <- lm(yp~xp);
> summary(t2)

Call:
lm(formula = yp ~ xp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21648 -0.12952 -0.01131  0.10843  0.29685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.21390    0.33119   9.704 6.87e-05 ***
xp          -0.59150    0.07607  -7.776 0.000238 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1853 on 6 degrees of freedom
Multiple R-squared: 0.9097,        Adjusted R-squared: 0.8947
F-statistic: 60.46 on 1 and 6 DF,  p-value: 0.0002382
```

Model odwrotnościowy względem zmiennej zależnej: $y = \frac{1}{a+bx}$. Linearyzacja modelu: $v := 1/y$ i stąd $v = a + bx$.

```
> yb <- 1/wys;
> t3 <- lm(yb~odl);
> summary(t3);

Call:
lm(formula = yb ~ odl)

Residuals:
```

```
      Min        1Q    Median        3Q       Max
-0.165680 -0.080000  0.003884  0.066478  0.166753

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2048174  0.0694234   2.950 0.025604 *
odl         0.0036887  0.0005366   6.874 0.000467 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1222 on 6 degrees of freedom
Multiple R-squared: 0.8873,        Adjusted R-squared: 0.8686
F-statistic: 47.25 on 1 and 6 DF,  p-value: 0.0004672
```

Model odwrotnościowy względem x: $y = a + \frac{b}{x}$ Linearyzacja modelu: $w := 1/x$ i stąd $y = a + bw$.

```
> xb <- 1/odl;
> t4 <- lm(wys~xb);
> summary(t4);

Call:
lm(formula = wys ~ xb)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26756 -0.21344 -0.05194  0.15094  0.48701

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7552     0.1656   4.559  0.00385 **
xb           78.0908     6.7037  11.649 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2847 on 6 degrees of freedom
Multiple R-squared: 0.9577,        Adjusted R-squared: 0.9506
F-statistic: 135.7 on 1 and 6 DF,  p-value: 2.411e-05
```

Spośród zbadanych modeli najlepiej dopasowany jest ostatni tzn. model odwrotnościowy względem x: $y = a + b/x$.

Prognozowanie w tym modelu:

```
> pw <- data.frame(xb=1/100);
> predict(t4, pw, interval="confidence");

       fit      lwr      upr
1 1.536108 1.243547 1.828670
```

⊡

3 Zadania do rozwiązania

Zadanie 7.6. Pewien przedsiębiorca zainteresowany jest oceną ryzyka planowanych inwestycji. Dwóch zatrudnionych przez niego analityków uszeregowało planowane inwestycje od tej o największym ryzyku (10) do tej o najmniejszym (1):

Inwestycja	A	B	C	D	E	F	G	H	I	J
Analityk I	1,	4,	9,	8,	6,	3,	5,	7,	2,	10
Analityk II	1,	5,	6,	2,	9,	7,	3,	10,	4,	8

Czy można uznać, że istnieje zależność między opiniami obu analityków? Przyjmij poziom istotności 0,05.

Zadanie 7.7. Wyznacz prostą regresji poziomu cholesterolu względem wieku dziesięciu losowo wziętych mężczyzn. Zweryfikuj dopasowanie modelu.

Wiek	58,	69,	43,	39,	63,	52,	47,	31,	74,	36
Poziom cholesterolu	189,	235,	193,	177,	154,	191,	213,	175,	198,	181

Zadanie 7.8. Niech X oznacza przeciętną liczbę samochodów poruszających się autostradą w ciągu dnia (w tys.), natomiast Y — liczbę wypadków samochodowych, która ma miejsce w ciągu miesiąca na autostradzie. Na podstawie danych zamieszczonych w poniższej tabeli wyznacz model regresji

$$\sqrt{Y} = a + bX,$$

opisujący zależność liczby wypadków od natężenia ruchu na autostradzie. Oszacuj liczbę wypadków, jakiej można się spodziewać przy natężeniu ruchu odpowiadającemu 3500 samochodom poruszającym się autostradą w ciągu dnia.

X	2.0,	2.3,	2.5,	2.6,	2.8,	3.0,	3.1,	3.4,	3.7,	3.8,	4.0,	4.6,	4.8
Y	15,	27,	20,	21,	31,	26,	22,	23,	32,	39,	27,	43,	53

Zadanie 7.9. Korzystając z danych zawartych w poniższej tabeli, wyznacz funkcję regresji, opisującą zależność między liczbą cykli do zniszczenia pewnego detalu (w mln.) a wywieranym na ten detal naprężeniem (w MPa):

Naprężenie	55,	50.5,	43.5,	42.5,	42,	41,	35.7,	34.5,	33,	32
Cykle	0.223,	0.925,	6.75,	18.1,	29.1,	50.5,	126,	215,	445,	420

Oszacuj liczbę cykli do zniszczenia detalu pracującego pod naprężeniem 40 MPa.

Zadanie 7.10. Badano wpływ dawki pewnego leku na puls pacjenta. Oto wyniki uzyskane dla 10 losowo wybranych osób:

Dawka leku	2,	2,	4,	4,	8,	8,	16,	16,	32,	32
Puls	68,	58,	63,	62,	67,	65,	70,	70,	74,	73

Dopasuj właściwy model regresji do powyższych danych.

Zadanie 7.11. Dla 15 wybranych losowo samochodów pewnej marki zbadano zależność zużycia paliwa (mile/galon) od mocy ich silnika. Wyniki przedstawia poniższa tabela:

Zużycie paliwa	43.1,	20.3,	17,	21.6,	16.2,	31.5,	31.9,	25.4,
Moc	48,	103,	125,	115,	133,	71,	71,	77,
Zużycie paliwa	27.2,	37.3,	41.5,	34.3,	44.3,	43.4,	36.4	
Moc	71,	69,	76,	78,	48,	48,	67	

Dopasuj najlepszy model regresji do powyższych danych. Zweryfikuj jego dopasowanie. Podaj przewidywane zużycie paliwa samochodu o mocy 150.

Zadanie 7.12. Zapytano dziesięciu losowo wybranych mężczyzn stojących pod osiedlową Pijalnią Jogurtu, ile litrów pewnego napoju mlecznego wypijają w ciągu tygodnia. Wyniki przedstawia poniższa tabela:

Wiek	37,	42,	25,	14,	48,	78,	18,	34,	20,	57
Spożycie	3,	2,	4,	5,	1,	0.3,	8,	2.5,	7,	0.5

Dopasuj wykładniczy model regresji do powyższych danych. Zweryfikuj jego dopasowanie. Podaj przewidywaną wielkość tygodniowego spożycia owego napoju przez 40-latka.

Zadanie 7.13. Badano zależność miesięcznych wydatków na rozrywki mieszkańców pewnego miasta od wysokości ich miesięcznych dochodów. Poniższa tabela zawiera informacje o 7 losowo wybranych osobach:

Wydatki	186,	700,	490,	385,	266,	357,	613
Dochody	2800,	4200,	3500,	3150,	2975,	3175,	3850

a) Wyznacz najlepszy model regresji opisujący badaną zależność.
b) Zweryfikuj dopasowanie modelu.
c) Podaj przewidywaną wysokość wydatków na rozrywki osoby o dochodach w wysokości 4000 zł.

Zadanie 7.14. Badano zależność między liczbą wypalanych dziennie papierosów a prawdopodobieństwem zachorowania na raka płuc w populacji 40-letnich palaczy, palących od 10 lat. Poniższa tabela zawiera informacje o 7 losowo wybranych osobach:

Liczba papierosów	5	10	20	30	40	50	60
Prawdopodobieństwo raka	0.061,	0.113,	0.192,	0.259,	0.339,	0.401,	0.461,

a) Wyznaczyć potęgowy model regresji opisujący badaną zależność.
b) Przeanalizuj dopasowanie modelu.
c) Oszacuj prawdopodobieństwo zachorowania na raka płuc przez palacza wypalającego 35 papierosów dziennie.