zestaw zadań nr 7
Cel: analiza regresji – regresja prosta i wieloraka

Przebieg regresji liniowej:

1. Znaleźć funkcję y=f(x) (dopasowanie modelu)
2. Sprawdzić:
   a) Wsp. determinacji R^2
   b) Test istotności dla wsp. Kierunkowego b
      H: B=0
      K: $\neg$ H
   c) analiza wariancji
      H: nie istnieje zależnosc miedzy X i Y
      K: $\neg$ H
   d) czy resety mają rozkład normalny

MODELE

Wykładniczy:

$$\boldsymbol{Y = e^{a+bx} + \varepsilon}$$

$$Y = e^{a+bx} + \varepsilon \mid log$$
$$logY = a + bx + log\ \varepsilon$$
$$Y' = a + bx + \varepsilon'$$

lny = log(y)
model = lm(lny~x)

Odwrotnościowy (względem Y)

$$Y = \frac{1}{a + bx} \rightarrow \frac{1}{y} = a + bx$$
$$Y' = a + bx$$

oy = 1/y
model = lm(oy~x)

Odwrotnościowy (względem X)

$$\boldsymbol{Y = a + \frac{b}{x} + \varepsilon}$$

$$x' = \frac{1}{x}$$

$$Y = a + bx' + \varepsilon$$

ox = 1/x
model = lm(y~ox)

model multiplikatywny(potęgowy)

$$Y = aX^b\varepsilon -> \ln Y =$$
$$\ln a + b \ln X + \ln \varepsilon -> $$
$$Y' = a' + b' + \varepsilon'$$

lnx = log(x)
lny=log(y)

model = lm(lny~lnx)

Zadanie 1

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

Roczny dochód ($ 1000)	36	64	49	21	28	47	58	19	32
Wartość domu ($ 1000)	129	310	260	92	126	242	288	81	134

a) Wyznaczyć prostą regresji wartości domu względem dochodu.

dochod= c(36,64,49,21,28,47,58,19,32)
dom = c(129, 310, 260, 92, 126, 242, 288, 81, 134)

#Sprawdzam na oko czy da sie poprowadzić prostą

plot(dochod,dom)

model = lm(dom ~ dochod) # dopasowanie modelu linniowego (linnear model, szacuje metoda najmniejszych kwadratow)

abline(model,col="red") # rysowanie linni

I linia powinna w miarę przechodzić przez punkty, to będzie ok.

summary(model)
$\hat{Dom}$= -30.344 + 5.466 * dochod

b) Przeanalizować dopasowanie modelu.

summary(model)

Wsp. determinacji R^2

Multiple R-squared: 0.9612
96,12 % obserwacji jest "wyjaśnionych" przez regresje. W takim stopniu model wyjaśnia zmienność Y

Test istotności dla wsp. Kierunkowego b
H: B=0 (czyli ze x(dochod) nie jest istotny)
K: ¬ H
p-valu (Pr(>\|t\|) dla dochod) < ALFA -> odrzucamy H, dochód jest istotny

Test Anovy
H: nie istnieje zależnośc miedzy X i Y
K: ¬ H

F-statistic: 173.5 on 1 and 7 DF, p-value: 3.394e-06

p-vale < *ALFA* -> *istnieje zależność*

czy reszty mają rozkład normalny

H : reszty mają rozkłady normalne
K : ¬ H

shapiro.test(resid(model))

p-value > ALFA -> reszty mają rozkład normalny

Wszystkie Hipotezy OK., R^2 duże -> model OK

c) Oszacować wartość domu rodziny, której roczny dochód wynosi $40000.
d) Wyznaczyć 95% przedział ufności dla szacowanej wartości domu tej rodziny.

nowe=data.frame(dochod=40)
predict(model,nowe,interval='conf', level=0.95)

fit – oszacowanie
lwr, upr – przedział ufności

Zadanie 3
W poniższej tabeli podano liczbę ludności USA (w mln) w latach 1890-2007:

Rok	1890	1900	1910	1920	1930	1940	1950	1960	1970
Ludność	62.947	75.994	91.972	105.710	122.775	131.669	150.697	179.323	203.235

Rok	1980	1990	2000	2007	2008	2009
Ludność	226.542	248.718	281.422	301.140	305.529	309.237

a) Przyjmując wykładniczy model wzrostu populacji, oszacować parametry tego modelu i zweryfikować jego dopasowanie.

$Y = e^{a+bx} + \varepsilon$

R ma tylko lm wiec tarza przekształcić

$Y = e^{a+bx} + \varepsilon \mid log$
$logY = a + bx + log\ \varepsilon$
$Y' = a + bx + \varepsilon'$

Y'-log lud
Y-lud
x-rok

rok=scan()
1890 1900 1910 1920 1930 1940 1950 1960 1970
1980 1990 2000 2007 2008 2009

lud=scan()
62.947 75.994 91.972 105.710 122.775 131.669 150.697 179.323 203.235
226.542 248.718 281.422 301.140 305.529 309.237
loglud=log(lud)

plot(rok,loglud) # widac, że jest całkiem OK
model=lm(loglud~rok)

abline(model) ## jak widac model jest ok

model
summary(model)

$\hat{Y}' = -20.03761 + 0.01284\ x$

$\hat{Y} = e^{-20.03761 + 0.01284\ x}$

Wsp. determinacji R^2

R-squared: 0.9936

Test istotności dla wsp. Kierunkowego b
H: B=0 (czyli ze x(rok) nie jest istotny)
K: $\neg H$

p-val< ALFA -> odrzucamy

Test Anovy
H: nie istnieje zależnośc miedzy X i Y
K: ¬ H

F-statistic: 2024 on 1 and 13 DF, p-value: 1.175e-15

p-val < ALFA -> odrzucamy H

czy reszty mają rozkład normalny
H : reszty mają rozkłady normalne
K : ¬ H

shapiro.test(resid(model))

p-val > ALFA -> przyjmujemy H -> reszty mają rozkład normalny

plot(fitted(model),resid(model)) ## można sobie popatrzec
abline(h=0)

model OK

b) Oszacować przewidywaną wielkość populacji USA w 2015 i w 2020 roku.

nowe=data.frame(rok=c(2015,2020))
predict(model,nowe,interval='conf') ## ale jako, że to wartosci zlogarytmowane to
exp(predict(model,nowe,interval='conf'))

Zadanie 4
Niech *X* oznacza przeciętną liczbę samochodów poruszających się autostradą w ciągu dnia, natomiast *Y* liczbę wypadków samochodowych, która ma miejsce w ciągu miesiąca na autostradzie. Na podstawie danych zamieszczonych w poniższej tabeli wyznaczyć następujący model regresji
$\sqrt{Y} = a + bX$,
opisujący zależność liczby wypadków od natężenia ruchu na autostradzie. Oszacować liczbę wypadków, jakiej można się spodziewać przy natężeniu ruchu odpowiadającemu 3500 samochodom poruszającym się autostradą w ciągu dnia.

X	2000	2300	2500	2600	2800	3000	3100	3400	3700	3800	4000	4600	4800
Y	15	27	20	21	31	26	22	23	32	39	27	43	53

x=scan()
2000 2300 2500 2600 2800 3000 3100 3400 3700 3800 4000 4600 4800
y=scan()
15 27 20 21 31 26 22 23 32 39 27 43 53

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

$Y' = a + bX + \varepsilon$
$Y' = \sqrt{Y} \rightarrow \hat{Y} = (\hat{Y}')^2$

py=sqrt(y)

model =lm(py~x)
summary(model)

R-squared: 0.7206,~> taki sobie, anova i istotność b OK.

$\hat{Y}' = 2.3250882 + 0.0009152\, x$

shapiro.test(resid(model)) ## test normalnosci dla reszt

p-val > ALFA -> reszty maja rozklad normalny

MODEL w miarę ok., lipa trochę to R^2

nowe=data.frame(x=3500)
(predict(model,nowe,interval='conf'))^2 ## kwadrat bo taki model

Lepiej patrzeć na przedział, bo słabe R^2

Zadanie 5
Dokonano osiem niezależnych pomiarów wielkości drgań pionowych gruntu powstałych w wyniku trzęsienia ziemi w różnej odległości od epicentrum trzęsienia. Otrzymano następujące wyniki:

Odległość od epicentrum (km)	20	30	40	50	80	140	200	250
Wielkość drgań pionowych (cm)	4.8	3.2	2.5	2.5	1.5	1.8	1.2	0.8

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

$Y=a+ b/x + \varepsilon$
x' = 1/x
$Y=a+bx' + \varepsilon$
y~x'

odl=scan()
20 30 40 50 80 140 200 250

wielk=scan()
1: 4.8 3.2 2.5 2.5 1.5 1.8 1.2 0.8

oodl=1/odl

model =lm(wielk~oodl)

$\hat{Y} = 0.7552 +78.0908\ x' + \varepsilon$

b) Zweryfikować dopasowanie modelu.

plot(oodl,wielk)
abline(model)

Wsp. determinacji R^2

R-squared: 0.9577

Test istotności dla wsp. Kierunkowego b
H: B=0 (czyli ze x(oodl) nie jest istotny)
K: $\neg H$

p-val< ALFA -> odrzucamy

Test Anovy
H: nie istnieje zależnośc miedzy X i Y
K: $\neg H$

F-statistic: 135.7 on 1 and 6 DF, p-value: 2.411e-05

p-val < ALFA -> odrzucamy H

czy reszty mają rozkład normalny
H : reszty mają rozkłady normalne

K : ¬H

shapiro.test(resid(model))

p-val > ALFA -> przyjmujemy H -> reszty mają rozkład normalny

plot(fitted(model),resid(model)) ## można sobie popatrzec
abline(h=0)

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

nowe=data.frame(oodl=c(1/100))
predict(model,nowe,interval='conf')

Zadanie 7
W pewnej firmie postanowiono zbadać zależność między wielkością tygodniowej sprzedaży produktów chemicznych tej firmy, a wydatkami poniesionymi na reklamę radiowo-telewizyjną oraz wydatkami poniesionymi na pokazy w sklepach. Oto dane (w tyś. $) pochodzące z 10 tygodni:

Wartość tygodniowej sprzedaży	72	76	78	70	68	80	82	65	62	90
Wydatki na reklamę radiowo-telewizyjną	12	11	15	10	11	16	14	8	8	18
Wydatki na pokazy w sklepach	5	8	6	5	3	9	12	4	3	10

a) Wyznaczyć liniową funkcję regresji opisującą badaną zależność.

$Y= a_0 + a_1x_1 + a_2x_2 + \varepsilon$

sprzed - Y
rtv -x_1
pok- x_2

sprzed=scan()
72 76 78 70 68 80 82 65 62 90

rtv=scan()
12 11 15 10 11 16 14 8 8 18

pok=scan()
5 8 6 5 3 9 12 4 3 10

model=lm(sprzed~rtv+pok)

model
summary(model)

$\hat{Y} = 47.165 + 1.599\ x_1 + 1.149\ x_2$

test istotności dla wyrazu wolnego

H: a_0=0
K: a_0=/=0

Pv <ALFA, odrzucamy H

b) Zweryfikować dopasowanie modelu.

H : a1=a2=0(test F czyli anova) *## bo sa 2 zmienne a nie 1*
K ¬ H (czyli przynajmniej 1 zmienna jest istotna)

jeśli nie odrzucimy tej hipotezy, to znaczy ze obie zmienne sa nie istotne i niema co robić zadania

pval < ALFA -> odrzucamy

$\overline{R^2}$ ***= R^2 adj = 0.9499***
testy istotności dla wsp kierunkowych
H1: a_1=0
K1: a_1=/=0

pv1 0.000742 odrzucamy H1 *(Pr(>|t|)na prawo od zmiennej*

H2: a_2=0
K2: a_2=/=0

pv2 0.007044 odrzucamy H2

czy reszty mają rozkład normalny
H : reszty mają rozkłady normalne
K : ¬H

shapiro.test(resid(model))

p-value >ALFA -> przyjmujemy H

plot(fitted(model),resid(model)) ## można popatrzec
abline(h=0)

Model OK.

c) Wykorzystać uzyskane równanie regresji do prognozy wielkości sprzedaży, gdy wydatki na reklamę radiowo telewizyjną wyniosą 8000$, natomiast wydatki na pokazy w sklepach 12000$.

nowe=data.frame(rtv=8,pok=12)
predict(model,nowe,interval="conf")

Zadanie 8
Pośrednik w handlu nieruchomościami jest zainteresowany oszacowaniem wpływu powierzchni budynku i jego odległości od centrum miasta na wartość budynku. Poniższa tabela zawiera informacje o dziewięciu losowo wybranych budynkach.

Wartość budynku (tys.$)	345	320	452	422	328	375	660	466	290
Powierzchnia (m_2)	150	180	200	160	175	180	300	170	135
Odległość Od centrum (km)	5.6	1.2	2.4	7.2	2.9	2.5	5.5	4.8	1.6

a) Wyznaczyć liniową funkcję regresji opisującą zależność, którą interesuje się ów pośrednik.

$Y = a_0 + a_1x_1 + a_2x_2 + \varepsilon$

Wart - Y
pow – x_1
odl – x_2

wart=scan()
345 320 452 422 328 375 660 466 290

pow=scan()
150 180 200 160 175 180 300 170 135

odl=scan()
5.6 1.2 2.4 7.2 2.9 2.5 5.5 4.8 1.6

model = lm (wart~pow+odl)
summary(model)

$\hat{Y} = -19.873 + 1.929\, x_1 + 19.405\, x_2$

test istotności dla wyrazu wolnego

H: a_0=0
K: a_0=/=0

Pv >ALFA, przyjmujemy H

wiec tworzymy nowy model

model2 = lm (wart~pow+odl-1) #-1 czyli bez wyrazu wolnego
summary(model2)

$\hat{Y} = 1.839\, x_1 + 18.776 x_2$

b) Zweryfikować dopasowanie modelu.

H : a1=a2=0(test F czyli anova) *## bo sa 2 zmienne a nie 1*
K ¬ H (czyli przynajmniej 1 zmienna jest istotna)

jeśli nie odrzucimy tej hipotezy, to znaczy ze obie zmienne sa nie istotne i niema co robić zadania

pval < ALFA -> odrzucamy

$\overline{R^2}$ ***= R^2 adj = 0.9915***

testy istotności dla wsp kierunkowych
H1: a_1=0
K1: a_1=/=0

pv1 4.29e-06 odrzucamy H1 *(Pr(>|t|)na prawo od zmiennej*

H2: a_2=0
K2: a_2=/=0

pv2 0.0227 odrzucamy H2

czy reszty mają rozkład normalny
H : reszty mają rozkłady normalne
K : ¬H

shapiro.test(resid(model2))

p-value >ALFA -> przyjmujemy H

plot(fitted(model2),resid(model2)) *## można popatrzec*
abline(h=0)

c) Podać przewidywaną wartość domu o powierzchni 160 m2, położonego w odległości 3 km od centrum miasta.

nowe = data.frame(odl=3,pow=160)
predict(model2,nowe,interval="conf")

KOLOWIUM

Grupa A

Zadanie 2
Badano zależność wzrostu niemowląt (w cm) od kilku czynników: wieku(w dniach), wzrostu w chwili urodzenia (w cm), oraz wagi w chwili urodzenia(w kg). Otrzymano następujące dane:

Wzrost	57.5	52.8	61.3	67	53.5	62.7	56.2	68.5	69.2
Wiek	78	69	77	88	67	80	74	94	102
Wzrost w chwili urodzenia	42.8	45.5	46.3	49	43	48	48	53	58
Waga w chwili urodzenia	2.75	2.15	4.41	5.52	3.21	4.32	2.31	4.30	3.71

a) Wyznaczyć liniową funkcję regresji opisującą zależność wzrostu niemowląt od wymienionych czynników

$Y= a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \varepsilon$

wzrost - Y
wiek -x_1
wzrost_ur- x_2
waga- x_3

```
wzrost=scan()
 57.5 52.8 61.3 67 53.5 62.7 56.2 68.5 69.2
wiek=scan()
 78 69 77 88 67 80 74 94 102
wzrost_ur=scan()
 42.8 45.5 46.3 49 43 48 48 53 58
waga=scan()
2.75 2.15 4.41 5.52 3.21 4.32 2.31 4.30 3.71

model=lm(wzrost~wiek+ wzrost_ur +waga)
summary(model)
```

$Y =18.1472 +0.3661\ x_1 + 0.1137\ x_2 + 2.1168\ x_3$

Ale ponieważ p-value x_2 (wzrost_ur) > alfa -> czyli jest nie istotny

więc

```
model2=lm(wzrost~wiek+waga)
summary(model2)
```

$Y =20.1085 + 0.4136\ x_1 + 2.0253\ x_3$

b) Przeanalizować dopasowanie modelu

$$\overline{R^2} = R\text{^}2\ adj = 0.9843$$

H : a1=a3=0(test F czyli anova) *## bo sa 2 zmienne a nie 1*
K $\neg$ H (czyli przynajmniej 1 zmienna jest istotna

p-val< ALFA

testy istotności dla wsp kierunkowych
H1: $a_1=0$
K1: a_1=/=0

pv1 < ALFA odrzucamy H1 *(Pr(>\|t\|)na prawo od zmiennej*

H2: a_3=0
K2: a_3=/=0

pv2 <ALFA odrzucamy H2

czy reszty mają rozkład normalny
H : reszty mają rozkłady normalne
K : ¬H

shapiro.test(resid(model2))

p-value >ALFA -> przyjmujemy H

plot(fitted(model2),resid(model2)) ## można popatrzec
abline(h=0)

c) Oszacować wzrost niemowlęcia mającego 90 dni, jeśli wiadomo, ze w chwili urodzenia miało 50 cm i ważyło 4,2 kg

nowe=data.frame(wiek=90, waga=4,2)
predict(model2,nowe,interval="conf")

Grupa B

Badano zależność czasu działania akumulatora od czasu eksploatacji tego urządzenia. Otrzymano następujące wyniki:

Czas działania – Y	3.9	4	4.3	4.5	6	7.5	10	18.3	30.5
Czas eksploatacji – X	60	55	50	45	40	35	30	20	15

A) przyjmując model odwrotnościowy (względem Y) model regresyjny oszacować parametry tego modelu i zweryfikować jego dopasowanie.

dzialanie = scan() ## Y
3.9 4 4.3 4.5 6 7.5 10 18.3 30.5

ekspl=scan() ## X
60 55 50 45 40 35 30 20 15

$$Y = \frac{1}{a + bx} \rightarrow \frac{1}{y} = a + bx$$
Y'=a+bx

odzialanie = 1/ dzialanie
model = lm(odzialanie ~ ekspl)

plot(ekspl,odzialanie)
abline(model)

Y' = -0.052032 + 0.005477x
$Y = \frac{1}{-0.052032 + 0.005477x}$

Wsp. determinacji R^2

Multiple R-squared: 0.9742

Test istotności dla wsp. Kierunkowego b
H: B=0 (czyli ze x(czas eksploatacji) nie jest istotny)
K: ¬ H
p-valu (Pr(>\|t\|) dla ekspl) < ALFA -> odrzucamy H, czas eksploatacji jest istotny

Test Anovy
H: nie istnieje zależnośc miedzy X i Y
K: ¬ H

F-statistic: 264.1 on 1 and 7 DF, p-value: 8.13e-07

p-vale < *ALFA -> istnieje zależność*

czy reszty mają rozkład normalny

H : reszty mają rozkłady normalne
K : ¬ H

shapiro.test(resid(model))

p-value > ALFA -> reszty mają rozkład normalny

plot(fitted(model),resid(model)) ## można sobie popatrzec
abline(h=0)

widać, że może być zależność funkcyjna to minus

Wszystkie Hipotezy OK., R^2 duże -> model OK

B) Oszacować czas działania akumulatora po 25 mies. eksploatacji.

nowe=data.frame(ekspl=25)
predict(model,nowe,interval="conf")