Ćwiczenia:

Pierwsze modele regresyjne. Regresja liniowa, wieloraka.

Ocena jakości modeli.

Przykład: regresja prosta liniowa

```
#podział na zbiór uczący i testowy

Wybór celowy
train<-cereal[-c(5, 15, 25, 35, 55),]
test<-cereal[c(5, 15, 25, 35, 55),]

Losowy
#sets <- sample(1:nrow(cereals), 0.9 * nrow(cereals))
#train2<- cereals [sets,]
#test2<- cereals[-sets,]

#model 1: Regresja liniowa
m1<-lm(rating~sugars, train)
summary(m1)

##
## Call:
## lm(formula = rating ~ sugars, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.002  -5.879  -1.481   5.145  34.356
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.3494     2.0779  28.562  < 2e-16 ***
## sugars       -2.3915     0.2552  -9.371 5.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 70 degrees of freedom
## Multiple R-squared:  0.5564, Adjusted R-squared:  0.5501
## F-statistic: 87.81 on 1 and 70 DF,  p-value: 5.549e-14

#ocena poprawności modelu: wszystkie parametry istotne statystycznie
```

Ocena jakości modelu:

Samego modelu:

1. istotność statystyczna modelu,

2. współczynnik determinacji,

3. błąd standardowy estymacji,

4. dopasowanie modelu do danych (wykres).

Reszt (wizualizacje lub testy statystyczne):

1. normalność reszt,

2. autokorelacja reszt *,

3. wariancja reszt.

Prognoz:

1. MAPE, RMSPE…,

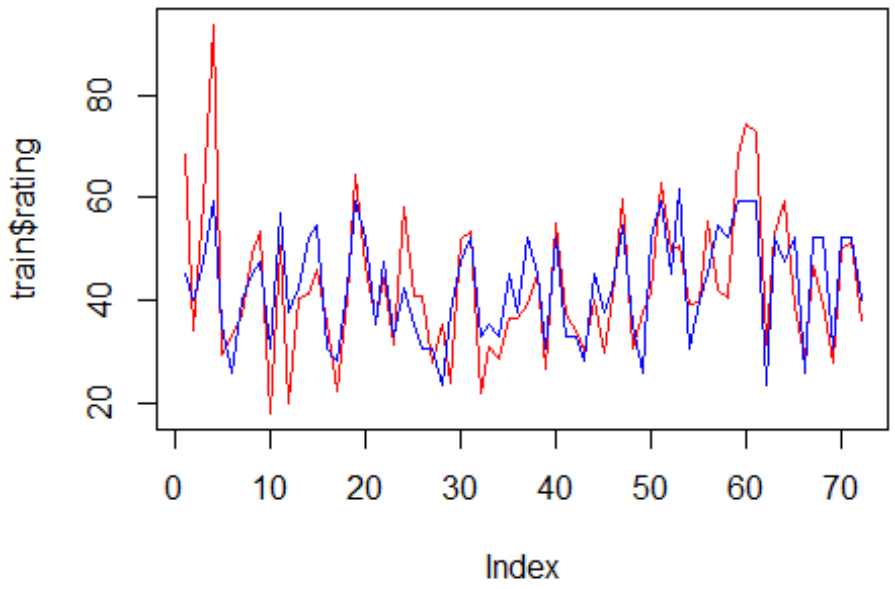
2. wsp. Theil'a (składowe: szczególnie istotne dla osób, które zaczynają pracę z modelowaniem danych),

3. wsp. Janussowy.

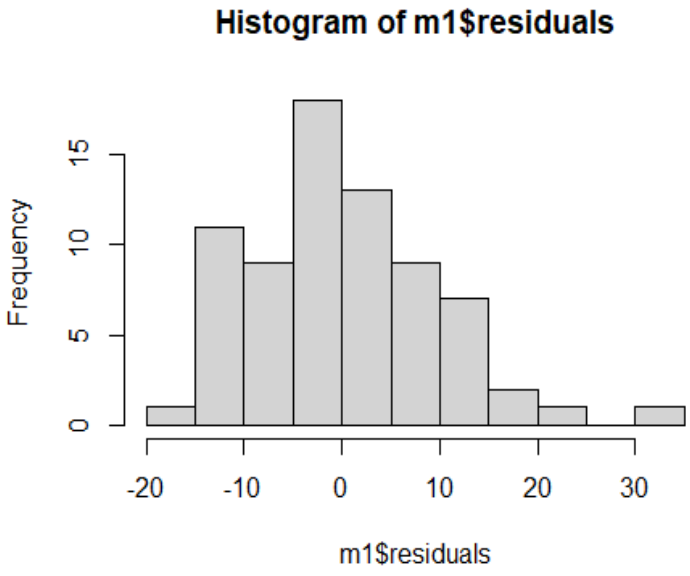
Porównanie między modelami:

1. Kryterium AIC, BIC,
2. Kryterium interpretowalności,
3. Kryterium złożoności obliczeniowej,
4. Kryterium K.I.S.S. (keep it simple...),
5. Własne.

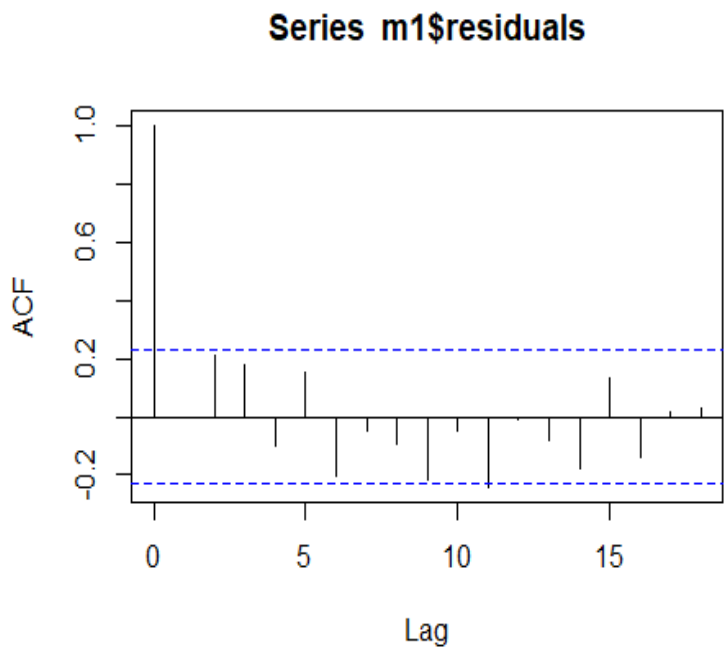
```
# ocena samego modelu:
summary(m1)

##
## Call:
## lm(formula = rating ~ sugars, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.002  -5.879  -1.481   5.145  34.356
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.3494     2.0779  28.562  < 2e-16 ***
## sugars       -2.3915     0.2552  -9.371 5.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 70 degrees of freedom
## Multiple R-squared:  0.5564, Adjusted R-squared:  0.5501
## F-statistic: 87.81 on 1 and 70 DF,  p-value: 5.549e-14

plot(train$rating, type="l", col="red")
lines(m1$fitted.values, type="l", col="blue")
```



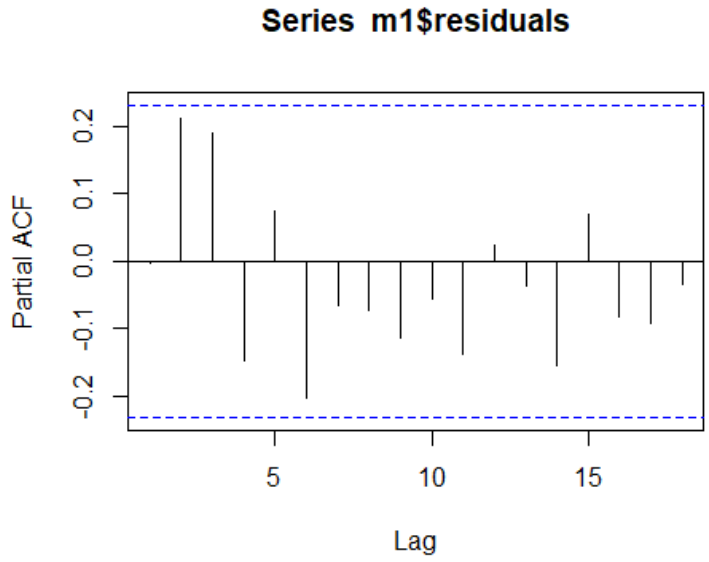

```
#kryterium porównawcze z innymi modelami
aic_m1<-AIC(m1)
bic_m1<-BIC(m1)

#reszty
hist(m1$residuals)
```



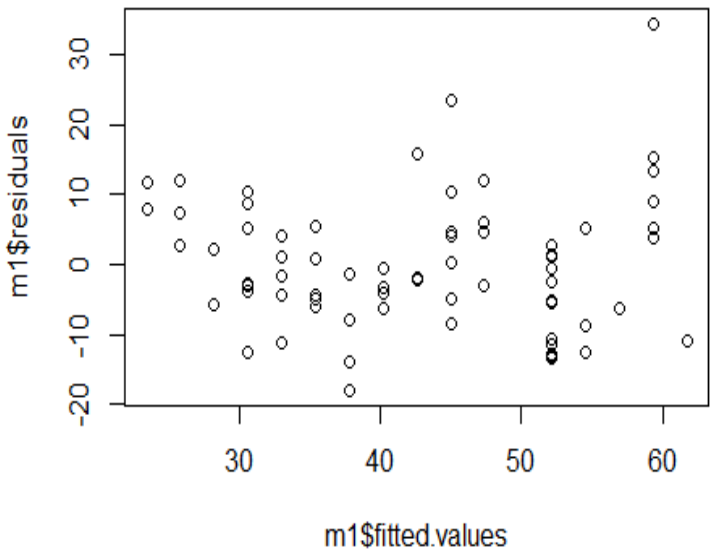

```
acf(m1$residuals)
```




```
pacf(m1$residuals)
```




```
plot(m1$residuals~m1$fitted.values)
```




```
#prognozy
library(MLmetrics)

##
## Dołączanie pakietu: 'MLmetrics'

## Następujący obiekt został zakryty z 'package:base':
##
##     Recall

pred_m1<-predict(m1, newdata = test)
mape_m1<-MAPE(pred_m1, test$rating)
# w kolejnych krokach wyliczane są współczynniki Theila i wsp. Janussowy
```

Zadanie 1.

Wykonaj model regresji wielorakiej dla danych **cereal**. Do zbioru testowego (do predykcji przenieś 5 obserwacji). Zmienną zależną jest rating. Zmienne niezależne wybierz na podstawie analizy danych. Oceń jakość powstałego modelu, jakość reszt oraz prognoz. Uzasadnij kolejne kroki analizy danych i modelowania, opracuj wnioski o możliwości wykorzystania wykonanego modelu do predykcji danych/możliwości jego poprawy. Porównaj jakość modelu z wykonanym na poprzednich zajęciach.

Zadanie 2.

Dla danych **Series_G/ AirPassengers** (kanał ogólny zajęć) wykonaj analizę danych oraz model endogenny regresji wielorakiej (struktury danych). Do zbioru testowego do predykcji (prognozy wygasłe) przenieś 3 obserwacje. Oceń jakość modelu, reszt i predykcji.

Podpowiedź: Najpierw sprawdź występowanie trendu i sezonowości. Dodaj odpowiednie zmienne do danych informujące o trendzie i sezonowości jeśli one występują (dummies variables). Przydatne funkcje **ifelse(), one_hot()** z pakietu **mltools**, lub **dummyVars()** z pakietu **caret.**

Zadanie 3.

Wykonaj model regresji prostej liniowej oraz wielorakiej i prognozy na 5% obserwacji ze zbioru testowego do predykcji dla danych **real_estate**. Oceń jakość powstałych modeli i porównaj ze sobą.