ADPS 21Z — Ćwiczenie 4 - rozwiązania

Ola Jaglińska

Zadanie 1

Treść zadania

Korzystając z metod analizy wariancji (przy założeniu normalności rozkładów oraz bez tego założenia), dla wybranej spółki notowanej na GPW zweryfikuj hipotezę o równości wartości średnich procentowych zmian cen zamknięcia

- porównując średnie w ostatnich sześciu miesiącach,
- porównując średnie w ostatnich trzech miesiącach.

Wskazówki:

- obliczenie procentowych zmian cen zamknięcia:

dane$close_ch = with(dane, c(NA, 100*diff(close)/close[-length(close)]))

- przykładowy sposób wczytania danych dot. np. czerwca 2021:

x1 = with(dane, close_ch[format(date, '%Y-%m') == '2021-06'])

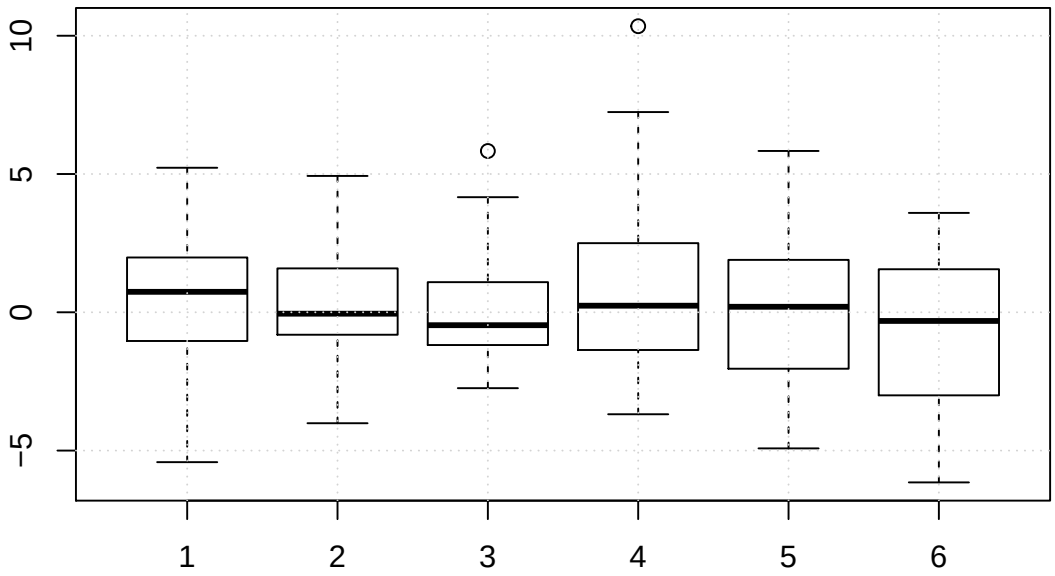
Rozwiązanie

W zadaniach zakładam poziom istotności na poziomie 0.05

```
unzip('mstall.zip', 'DATAWALK.mst')
datawalk = read.csv('DATAWALK.mst')
names(datawalk) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
datawalk$date = as.Date.character(datawalk$date, format ='%Y%m%d')
datawalk$close_ch = with(datawalk, c(NA, 100*diff(close)/close[-length(close)]))
z1_x1 = with(datawalk, close_ch[format(date, '%Y-%m') == '2021-06'])
z1_x2 = with(datawalk, close_ch[format(date, '%Y-%m') == '2021-07'])
z1_x3 = with(datawalk, close_ch[format(date, '%Y-%m') == '2021-08'])
z1_x4 = with(datawalk, close_ch[format(date, '%Y-%m') == '2021-09'])
z1_x5 = with(datawalk, close_ch[format(date, '%Y-%m') == '2021-10'])
z1_x6 = with(datawalk, close_ch[format(date, '%Y-%m') == '2021-11'])
```

- porównanie średnich w ostatnich sześciu miesiącach

```
boxplot(z1_x1, z1_x2, z1_x3, z1_x4, z1_x5, z1_x6)
grid()
```




```
datawalk_anova1 = data.frame( datawalk = c(z1_x1, z1_x2, z1_x3, z1_x4, z1_x5, z1_x6),
          proba1 = rep( c('z1_x1', 'z1_x2', 'z1_x3', 'z1_x4', 'z1_x5', 'z1_x6'),
          times1 = c(length(z1_x1), length(z1_x2), length(z1_x3), length(z1_x4), length(z1_x5), leng
```

Tu mam błąd z powodu różnej ilości wierszy w poszczególnych kolumnach. Nie zdążyłam znaleźć rozwiązania w R, ale wyrównałabym ilość wierszy przy pomocy uzupełniania braków danych 'NA'

Analiza wariancji dla sześciu miesięcy przy założeniu normalności rozkładów:

```
z1_aov_res1 = aov(datawalk~proba1, data = datawalk_anova1)
summary(z1_aov_res1)
```

Zakładam wartość alfa = 0.05. Wysoka p-wartość = 0.565 - nie mogę odrzucić hipotezy zerowej o równości średnich procentowych zmian cen zamknięcia.

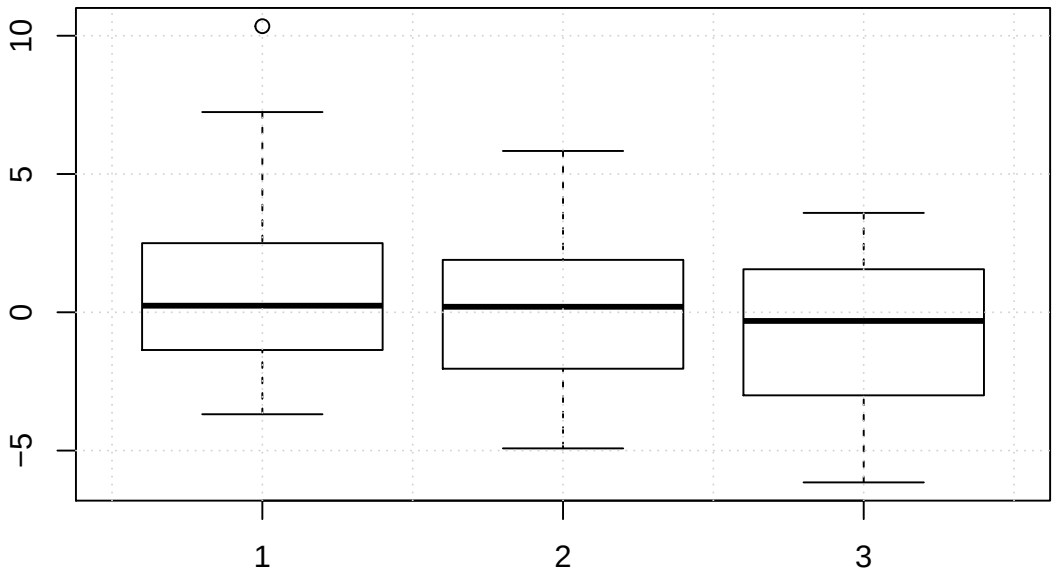
Analiza wariancji dla sześciu miesięcy bez zakładania normalności rozkładów - test Kruskala-Wallisa:

```
kruskal.test(datawalk~proba1, data = datawalk_anova1)
```

Wysoka p-wartość = 0.7518 - nie mogę odrzucić hipotezy zerowej.

- porównanie średnich w ostatnich trzech miesiącach

```
boxplot(z1_x4, z1_x5, z1_x6)
grid()
```




```
datawalk_anova2 = data.frame(datawalk = c(z1_x4, z1_x5, z1_x6),
          proba2 = rep( c('z1_x4', 'z1_x5', 'z1_x6'),
          times2 = c(length(z1_x4), length(z1_x5), length(z1_x6))) )
```

Analiza wariancji dla trzech miesięcy przy założeniu normalności rozkładów:

```
z1_aov_res2 = aov(datawalk~proba2, data = datawalk_anova2)
summary(z1_aov_res2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## proba2       2   24.1  12.043   1.264   0.29
## Residuals   60  571.5   9.525
```

Zakładam wartość alfa = 0.05. Wysoka p-wartość = 0.29 - nie mogę odrzucić hipotezy zerowej o równości średnich procentowych zmian cen zamknięcia.

Analiza wariancji dla trzech miesięcy bez zakładania normalności rozkładów - test Kruskala-Wallisa:

```
kruskal.test(datawalk~proba2, data = datawalk_anova2)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  datawalk by proba2
## Kruskal-Wallis chi-squared = 1.9223, df = 2, p-value = 0.3825
```

Wysoka p-wartość = 0.3825 - nie mogę odrzucić hipotezy zerowej.

Zadanie 2

Treść zadania

- Korzystając z regresji liniowej wyznacz zależność indeksu WIG20 (na zamknięciu notowań) od kursów zamknięcia spółek AMICA, COMARCH, GETIN, PEKAO, PGNIG, PZU dla danych z trzeciego kwartału roku 2021.
- Oceń istotność poszczególnych zmiennych objaśniających w tak skonstruowanym modelu.
- Przeprowadź analogiczne analizy w przypadku uwzględnienia w modelu mniejszej ilości spółek: AMICA, COMARCH, GETIN.

Rozwiązanie

```
read_mst = function(plik_zip, plik_mst) {
unzip(plik_zip, plik_mst)
dane = read.csv(plik_mst)
names(dane) = c('ticker', 'date', 'open', 'high', 'low', 'close', 'vol')
dane$date = as.Date.character(dane$date,format = '%Y%m%d')
dane }
```

```
if(!file.exists('mstall.zip')) {
  download.file('http://info.bossa.pl/pub/metastock/mstock/mstall.zip','mstall.zip')
}
```

Wczytuje dane spółek:

```
dane = read_mst('mstall.zip', 'WIG20.mst')
WIG20_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(WIG20_df) = c('date','WIG20')

dane = read_mst('mstall.zip', 'AMICA.mst')
AMICA_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(AMICA_df) = c('date','AMICA')

dane = read_mst('mstall.zip', 'COMARCH.mst')
COMARCH_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(COMARCH_df) = c('date', 'COMARCH')

dane = read_mst('mstall.zip', 'GETIN.mst')
GETIN_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(GETIN_df) = c('date', 'GETIN')

dane = read_mst('mstall.zip', 'PEKAO.mst')
PEKAO_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(PEKAO_df) = c('date', 'PEKAO')

dane = read_mst('mstall.zip', 'PGNIG.mst')
PGNIG_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(PGNIG_df) = c('date', 'PGNIG')

dane = read_mst('mstall.zip', 'PZU.mst')
PZU_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(PZU_df) = c('date', 'PZU')
```

Łączę dane w jedną ramkę:

```
WIG20_all = merge(WIG20_df, AMICA_df, by = 'date')
WIG20_all = merge(WIG20_all, COMARCH_df, by = 'date')
WIG20_all = merge(WIG20_all, GETIN_df, by = 'date')
WIG20_all = merge(WIG20_all, PEKAO_df, by = 'date')
WIG20_all = merge(WIG20_all, PGNIG_df, by = 'date')
WIG20_all = merge(WIG20_all, PZU_df, by = 'date')
```

Metoda regresji liniowej za pomocą funkcji lm:

```
lm_res = lm(WIG20 ~ AMICA + COMARCH + GETIN + PEKAO + PGNIG + PZU, data = WIG20_all)
summary(lm_res)
```

```
##
## Call:
## lm(formula = WIG20 ~ AMICA + COMARCH + GETIN + PEKAO + PGNIG +
##     PZU, data = WIG20_all)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.915 -14.229  -1.112  18.326  40.622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1419.9903   167.4704   8.479 8.51e-12 ***
## AMICA         -0.1172     0.7783  -0.151 0.880800
## COMARCH       -0.8534     0.2429  -3.513 0.000857 ***
## GETIN         65.6516    35.3550   1.857 0.068316 .
## PEKAO          4.5913     0.9079   5.057 4.42e-06 ***
## PGNIG         53.1010    27.5740   1.926 0.058957 .
## PZU            5.8267     3.1532   1.848 0.069640 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.71 on 59 degrees of freedom
## Multiple R-squared:  0.8144, Adjusted R-squared:  0.7955
## F-statistic: 43.15 on 6 and 59 DF,  p-value: < 2.2e-16
```

W przypadku spółki CORMACH i PEKAO p-wartość jest bardzo niska hipoteza zerowa powinna zostać odrzucona. Wartość ich współczynników jest istotna. Dla spółek AMICA, GETIN, PGNNIG i PZU p-wartość jest większa od 0.05 i nie możemy odrzucić hipotezy zerowej o zależności WIG20 od kursów zamknięcia tych spółek.

Metoda regresji liniowej za pomocą funkcji lm dla mniejszej ilości spółek:

```
lm_res = lm(WIG20 ~ AMICA + COMARCH + GETIN, data = WIG20_all)
summary(lm_res)
```

```
##
## Call:
## lm(formula = WIG20 ~ AMICA + COMARCH + GETIN, data = WIG20_all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.215  -18.117    5.724   19.982   64.116
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2048.1890   131.3686  15.591  < 2e-16 ***
## AMICA          1.3635     0.6995   1.949   0.0558 .
## COMARCH       -1.2462     0.2991  -4.166 9.76e-05 ***
```

```
## GETIN        258.6593    28.2511   9.156 4.02e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.6 on 62 degrees of freedom
## Multiple R-squared:  0.6458, Adjusted R-squared:  0.6287
## F-statistic: 37.69 on 3 and 62 DF,  p-value: 5.467e-14
```

W przypadku spółki CORMACH i GETIN p-wartość jest bardzo niska hipoteza zerowa powinna zostać odrzucona. Wartość och współczynników jest istotna. Dla spółkI AMICA p-wartość jest większa od 0.05 i nie możemy odrzucić hipotezy zerowej.

Zadanie 3

Treść zadania

Korzystając z regresji liniowej dla danych z trzeciego kwartału roku 2021 zbadaj:

- zależność kursu zamknięcia CHF od kursów zamknięcia EUR, USD, GBP, JPY (dane w pliku mstnbp.zip),
- zależność kursu zamknięcia CDPROJEKT (mstall.zip) od kursów zamknięcia CHF, EUR, USD, GBP, JPY (mstnbp.zip),
- zależność kursu zamknięcia kontraktu terminowego (futures) FPZUZ21 (PZU na grudzień 2021, dane w pliku mstfut.zip) od kursów zamknięcia PZU, PEKAO, ALIOR (mstall.zip).

Wskazówka:

- pobranie plików mstnbp.zip i mstfut.zip:

```
if(!file.exists('mstnbp.zip')) {
  download.file('https://info.bossa.pl/pub/metastock/waluty/mstnbp.zip','mstnbp.zip')
}
if(!file.exists('mstfut.zip')) {
  download.file('https://info.bossa.pl/pub/metastock/futures/mstfut.zip','mstfut.zip')
}
```

Rozwiązanie

- zależność kursu zamknięcia CHF od kursów zamknięcia EUR, USD, GBP, JPY (dane w pliku mstnbp.zip)

```
kurs = read_mst('mstnbp.zip', 'CHF.mst')
CHF_df = subset(kurs, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(CHF_df) = c('date','CHF')

kurs = read_mst('mstnbp.zip', 'EUR.mst')
EUR_df = subset(kurs, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(EUR_df) = c('date','EUR')

kurs = read_mst('mstnbp.zip', 'USD.mst')
USD_df = subset(kurs, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(USD_df) = c('date', 'USD')

kurs = read_mst('mstnbp.zip', 'GBP.mst')
GBP_df = subset(kurs, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(GBP_df) = c('date', 'GBP')
```

```
kurs = read_mst('mstnbp.zip', 'JPY.mst')
JPY_df = subset(kurs, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(JPY_df) = c('date', 'JPY')
```

Wspólna ramka danych:

```
CHF_all = merge(CHF_df, EUR_df, by = 'date')
CHF_all = merge(CHF_all, USD_df, by = 'date')
CHF_all = merge(CHF_all, GBP_df, by = 'date')
CHF_all = merge(CHF_all, JPY_df, by = 'date')
```

Metoda regresji liniowej za pomocą funkcji lm:

```
lm_res = lm(CHF ~ EUR + USD + GBP + JPY, data = CHF_all)
summary(lm_res)
```

```
##
## Call:
## lm(formula = CHF ~ EUR + USD + GBP + JPY, data = CHF_all)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.054176 -0.009100  0.002471  0.012858  0.031861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.19900    0.46676  -0.426   0.6714
## EUR          0.37329    0.25080   1.488   0.1418
## USD         -0.29039    0.16501  -1.760   0.0835 .
## GBP          0.20416    0.09488   2.152   0.0354 *
## JPY          0.78105    0.14026   5.569 6.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01843 on 61 degrees of freedom
## Multiple R-squared:  0.8311, Adjusted R-squared:   0.82
## F-statistic: 75.02 on 4 and 61 DF,  p-value: < 2.2e-16
```

P-wartość przy kursach GBP i JPY jest niska - odrzucam hipotezę zerową. Wartość ich współczynników jest istotna. Dla kursów EUR i USD p-wartość jest większa od 0.05 i nie możemy odrzucić hipotezy zerowej o zależności kursu zamknięcia CHF od kursów zamkniecia EUR i USD.

- zależność kursu zamknięcia CDPROJEKT (mstall.zip) od kursów zamknięcia CHF, EUR, USD, GBP, JPY (mstnbp.zip),

```
dane = read_mst('mstall.zip', 'CDPROJEKT.mst')
CDPROJEKT_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close')
names(CDPROJEKT_df) = c('date','CDPROJEKT')
```

```
CDPROJEKT_all = merge(CDPROJEKT_df, CHF_df, by = 'date')
CDPROJEKT_all = merge(CDPROJEKT_all, EUR_df, by = 'date')
CDPROJEKT_all = merge(CDPROJEKT_all, USD_df, by = 'date')
CDPROJEKT_all = merge(CDPROJEKT_all, GBP_df, by = 'date')
CDPROJEKT_all = merge(CDPROJEKT_all, JPY_df, by = 'date')
```

```
lm_res = lm(CDPROJEKT ~ CHF+ EUR + USD + GBP + JPY, data = CDPROJEKT_all)
summary(lm_res)
```

```
##
## Call:
## lm(formula = CDPROJEKT ~ CHF + EUR + USD + GBP + JPY, data = CDPROJEKT_all)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7835  -4.7243  -0.3818   3.6015  14.0542
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -564.454    166.430  -3.392  0.00124 **
## CHF         -280.656     45.585  -6.157 6.71e-08 ***
## EUR          711.555     90.902   7.828 9.64e-11 ***
## USD         -268.769     60.222  -4.463 3.62e-05 ***
## GBP          -54.406     35.038  -1.553  0.12574
## JPY            3.488     61.329   0.057  0.95483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.56 on 60 degrees of freedom
## Multiple R-squared:  0.6251, Adjusted R-squared:  0.5939
## F-statistic: 20.01 on 5 and 60 DF,  p-value: 1.101e-11
```

P-wartość przy kursach CHF, EUR i USD jest niska - odrzucam hipotezę zerową. Wartość ich współczynników jest istotna. Dla kursów GBP i JPY p-wartość jest większa od 0.05 i nie możemy odrzucić hipotezy zerowej o zależności kursu zamknięcia spółki CDPROJEKT od kursów zamkniecia GBP i JPY.

- zależność kursu zamknięcia kontraktu terminowego (futures) FPZUZ21 (PZU na grudzień 2021, dane w pliku mstfut.zip) od kursów zamknięcia PZU, PEKAO, ALIOR (mstall.zip).

```
dane = read_mst('mstfut.zip', 'FPZUZ21.mst')
FPZUZ21_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(FPZUZ21_df) = c('date','FPZUZ21')

dane = read_mst('mstall.zip', 'PZU.mst')
PZU_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(PZU_df) = c('date', 'PZU')

dane = read_mst('mstall.zip', 'PEKAO.mst')
PEKAO_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(PEKAO_df) = c('date', 'PEKAO')

dane = read_mst('mstall.zip', 'ALIOR.mst')
ALIOR_df = subset(dane, date >= '2021-07-01' & date <= '2021-09-30', select = c('date', 'close'))
names(ALIOR_df) = c('date','ALIOR')
```

```
FPZUZ21_all = merge(FPZUZ21_df, PZU_df, by = 'date')
FPZUZ21_all = merge(FPZUZ21_all, PEKAO_df, by = 'date')
FPZUZ21_all = merge(FPZUZ21_all, ALIOR_df, by = 'date')
```

```
lm_res = lm(FPZUZ21 ~ PZU + PEKAO + ALIOR , data = FPZUZ21_all)
summary(lm_res)
```

```
##
## Call:
## lm(formula = FPZUZ21 ~ PZU + PEKAO + ALIOR, data = FPZUZ21_all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96602 -0.24326 -0.01607  0.17101  1.01702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 17.51613    1.45591  12.031  < 2e-16 ***
## PZU          0.18852    0.04381   4.303 6.31e-05 ***
## PEKAO        0.01659    0.03470   0.478    0.634
## ALIOR        0.24248    0.03885   6.242 4.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3684 on 60 degrees of freedom
## Multiple R-squared:  0.943,  Adjusted R-squared:  0.9401
## F-statistic: 330.6 on 3 and 60 DF,  p-value: < 2.2e-16
```

P-wartość przy kursach PZU i ALIOR jest niska - odrzucam hipotezę zerową. Wartość ich współczynników jest istotna. Dla kursu PEKAO p-wartość jest większa od 0.05 i nie możemy odrzucić hipotezy zerowej o zależności kursu zamknięcia kontraktu terminowego od kursu zamkniecia PEKAO.

Zadanie 4

Treść zadania

W pliku sprzedaz.txt znajdują się dane dotyczące wydatków na reklamę pewnej firmy (w tys. zł) i wartości sprzedaży jej produktów (w mln zł) w poszczególnych kwartałach.

- Metodą regresji liniowej wyznacz zależność pomiędzy wartością sprzedaży a wydatkami na reklamę. Na jednym wykresie narysuj punkty odpowiadające danym oraz prostą regresji.
- Oblicz prognozowane wartości sprzedaży, jeśli wydatki na reklamę będą wynosiły: 300, 500, 700 tys. zł.
- Oszacuj odchylenie standardowe błędu z jakim wyznaczono prognozowane wartości sprzedaży dla poszczególnych wartości wydatków na reklamę.

Rozwiązanie

```
sprzedaz = read.csv('sprzedaz.txt')
```

- Metodą regresji liniowej wyznacz zależność pomiędzy wartością sprzedaży a wydatkami na reklamę. Na jednym wykresie narysuj punkty odpowiadające danym oraz prostą regresji.

```
lm_res_sprzedaz = lm(sprzedaz$Income ~ sprzedaz$Advert)
summary(lm_res_sprzedaz)
```

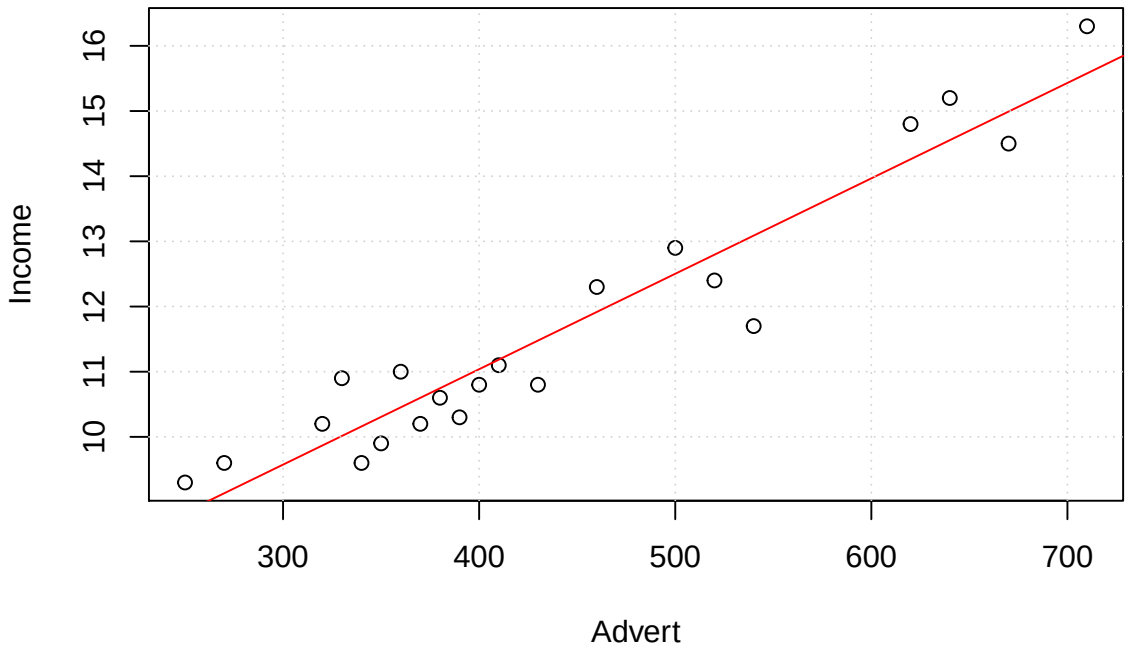
```
##
## Call:
## lm(formula = sprzedaz$Income ~ sprzedaz$Advert)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38840 -0.40633 -0.08488  0.46507  0.88652
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.18145    0.47316   10.95 1.20e-09 ***
## sprzedaz$Advert  0.01464    0.00103   14.22 1.41e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6078 on 19 degrees of freedom
## Multiple R-squared:  0.9141, Adjusted R-squared:  0.9095
## F-statistic: 202.1 on 1 and 19 DF,  p-value: 1.41e-11
```

```
R2 = summary(lm_res_sprzedaz)$r.squared
```

Bardzo niska p-wartość wskazuje na odrzucenie hipotezy zerowej.

```
plot(sprzedaz$Advert, sprzedaz$Income, xlab = 'Advert', ylab = 'Income')
abline(lm_res_sprzedaz, col = 'red')
grid()
```



* Prognozowane wartości sprzedaży i szacowane odchylenie standardowe błędu z jakim wyznaczono prognozowane wartości sprzedaży dla wydatków na reklamę wynoszących 300 tys zł

```
x_300 = 300
y_300 = coef(lm_res_sprzedaz)[2]*x_300 + coef(lm_res_sprzedaz)[1]
```

Prognozowana wartość sprzedaży przy wydatkach na reklamę 300 tys zł jest równa 10.

Odchylenie standardowe błędu prognozy

```
n_sprzedaz = length(sprzedaz$Advert)
s2_300 = 1/(n_sprzedaz - 2)*sum((sprzedaz$Income - y_300)^2)
s2_y_300 = s2_300*( 1 + 1/n_sprzedaz + (mean(sprzedaz$Advert) - x_300)^2/(n_sprzedaz*(mean(sprzedaz$
s_y_300 = sqrt(s2_y_300)
```

Błąd predykcji wynosi: 3

- Prognozowane wartości sprzedaży i szacowane odchylenie standardowe błędu z jakim wyznaczono prognozowane wartości sprzedaży dla wydatków na reklamę wynoszących 500 tys zł

```
x_500 = 500
y_500 = coef(lm_res_sprzedaz)[2]*x_500 + coef(lm_res_sprzedaz)[1]
```

Prognozowana wartość sprzedaży przy wydatkach na reklamę 500 tys zł jest równa 13.

Odchylenie standardowe błędu prognozy

```
n_sprzedaz = length(sprzedaz$Advert)
s2_500 = 1/(n_sprzedaz - 2)*sum((sprzedaz$Income - y_500)^2)
s2_y_500 = s2_500*( 1 + 1/n_sprzedaz + (mean(sprzedaz$Advert) - x_500)^2/(n_sprzedaz*(mean(sprzedaz$
s_y_500 = sqrt(s2_y_500)
```

Błąd predykcji wynosi: 2

- Prognozowane wartości sprzedaży i szacowane odchylenie standardowe błędu z jakim wyznaczono prognozowane wartości sprzedaży dla wydatków na reklamę wynoszących 700 tys zł

```
x_700 = 700
y_700 = coef(lm_res_sprzedaz)[2]*x_700 + coef(lm_res_sprzedaz)[1]
```

Prognozowana wartość sprzedaży przy wydatkach na reklamę 700 tys zł jest równa 15.

Odchylenie standardowe błędu prognozy

```
n_sprzedaz = length(sprzedaz$Advert)
s2_700 = 1/(n_sprzedaz - 2)*sum((sprzedaz$Income - y_700)^2)
s2_y_700 = s2_700*( 1 + 1/n_sprzedaz + (mean(sprzedaz$Advert) - x_700)^2/(n_sprzedaz*(mean(sprzedaz$
s_y_700 = sqrt(s2_y_700)
```

Błąd predykcji wynosi: 5

Z powyższych wyników wnioskuję, że najlepiej wypada średni model tj. wydatkina reklamę wynoszące 500 tys zł -> najniższy błąd.

Zadanie 5

Treść zadania

- Dla danych z pliku sprzedaz.txt zbadaj czy lepszym modelem zależności między wartością wydatków na reklamę (w tys. zł) a wartością sprzedaży (w mln zł) byłaby zależność kwadratowa.
- Nanieś odpowiednią linię przedstawiającą tę zależność na rysunek z danymi oraz prostą regresji wyznaczoną w poprzednim punkcie.

Rozwiązanie

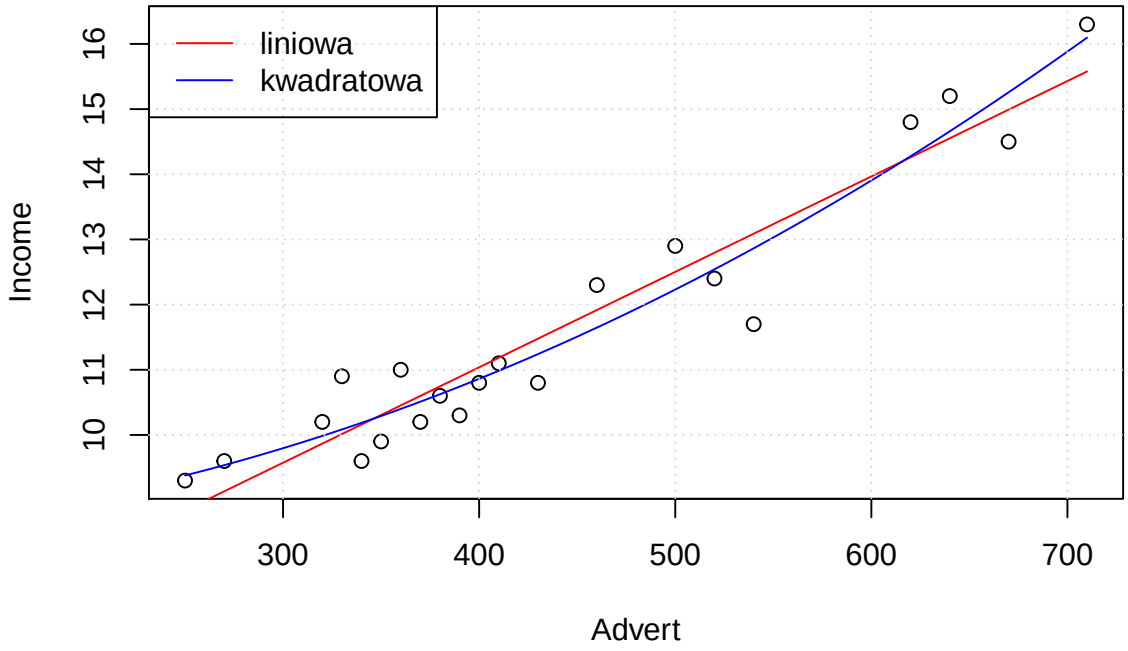
```
lm_res_sprzedaz2 = lm(sprzedaz$Income ~ I (sprzedaz$Advert^2))
summary(lm_res_sprzedaz2)
```

```
##
## Call:
## lm(formula = sprzedaz$Income ~ I(sprzedaz$Advert^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16413 -0.39129 -0.02449  0.52393  0.81564
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.427e+00  2.304e-01   36.58  < 2e-16 ***
## I(sprzedaz$Advert^2) 1.521e-05  9.391e-07   16.20 1.41e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5387 on 19 degrees of freedom
## Multiple R-squared:  0.9325, Adjusted R-squared:  0.9289
## F-statistic: 262.5 on 1 and 19 DF,  p-value: 1.409e-12
```

```
R22 = summary(lm_res_sprzedaz2)$r.squared

plot(sprzedaz$Advert, sprzedaz$Income, xlab = 'Advert', ylab = 'Income')
arg = seq(min(sprzedaz$Advert), max(sprzedaz$Advert), by = 1)
y_est = coef(lm_res_sprzedaz)[2]*arg + coef(lm_res_sprzedaz)[1]
y_est2 = coef(lm_res_sprzedaz2)[2]*arg^2 + coef(lm_res_sprzedaz2)[1]
lines(arg, y_est, col = 'red')
lines(arg, y_est2, col = 'blue')
grid()
legend('topleft', c('liniowa', 'kwadratowa'), col = c('red', 'blue'), lwd = 1)
```



W tym przypadku lepszym modelem jest model kwadratowy.
