EKONOMETRIA – SPRAWOZDANIE 2
MARTA SOMMER – BSMAD

Zadanie 1.

Wczytajmy nasze dane:

```
a <- read.csv2("C:\\Users\\Marta\\Desktop\\Marta\\studia\\rok4\\Ekonometria\\spr2\\zad1.csv",
    header = TRUE, sep = ";")
names(a) <- c("wydatki", "dochod")
attach(a)
head(a)

##   wydatki dochod
## 1    19.9   22.3
## 2    31.2   32.3
## 3    31.8   36.6
## 4    12.1   12.1
## 5    40.7   42.3
## 6     6.1    6.2
```

Zbudujmy model liniowy (zależność zmiennej *wydatki* od zmiennej *dochod*) stosując metodę MNK:

```
l <- lm(wydatki ~ dochod, data = a)
summary(l)

##
## Call:
## lm(formula = wydatki ~ dochod, data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3679 -0.9205  0.0096  1.2252  1.8838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8921     0.6852     1.3     0.21
## dochod        0.8966     0.0247    36.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 18 degrees of freedom
## Multiple R-squared:  0.987,Adjusted R-squared:  0.986
## F-statistic: 1.32e+03 on 1 and 18 DF,  p-value: <2e-16
```

Z *summary* widać, że model jest dobrze dopasowany ($p-value$ testu F jest małe – równe $2e-16$). Współczynnik R^2 równy $0,9858$ również świadczy o dobrym dopasowaniu modelu.

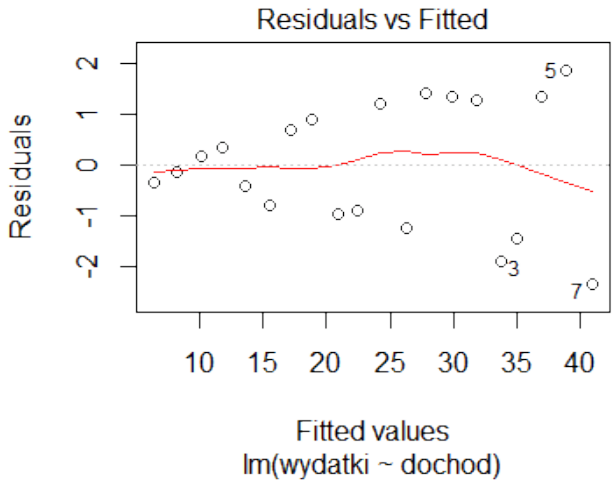
Współczynniki α i β w naszym modelu

$$wydatki = \alpha + \beta \cdot dochod$$

są równe odpowiednio $0,89$ i $0,9$. Z testu t, wynika również, że zmienna dochód jest rzeczywiście istotna w modelu ($p-value < 2e-16$).

Przedstawmy reszty z modelu na wykresie:

```
plot(l, 1)
```



Widać wyraźnie, że rezidua nie są losowo rozrzucone wokół zera, tylko jest między nimi jakaś zależność. Podejrzewamy więc występowanie heteroskedastyczności w modelu.

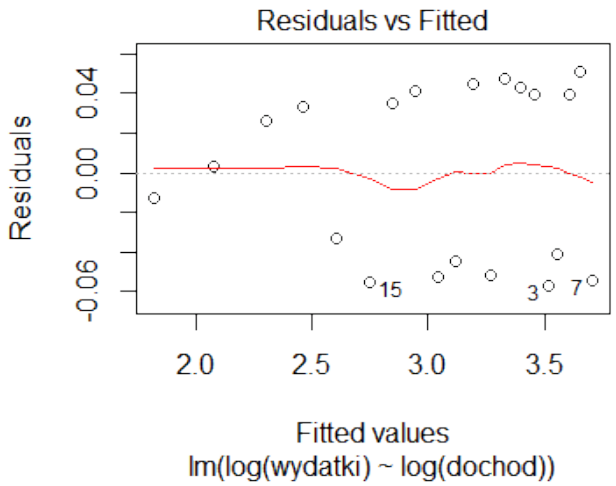
Powtórzmy rozumowanie dla modelu potęgowego tzn.

$$\ln(wydatki) = \alpha + \beta \cdot \ln(dochod)$$

```
l2 <- lm(log(wydatki) ~ log(dochod), data = a)
summary(l2)

##
## Call:
## lm(formula = log(wydatki) ~ log(dochod), data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.0573 -0.0468  0.0146  0.0398  0.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0789     0.0565     1.4     0.18
## log(dochod)   0.9549     0.0180    53.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.045 on 18 degrees of freedom
## Multiple R-squared:  0.994,Adjusted R-squared:  0.993
## F-statistic: 2.82e+03 on 1 and 18 DF,  p-value: <2e-16
```

I tutaj również, podobnie jak w modelu liniowym, model ma sens ($p - value$ testu F małe) oraz jest dość dobrze dopasowany ($R^2 = 0,9937$). Z testu t wynika, że zmienna *dochod* jest istotna. Przyjrzyjmy się jeszcze wykresowi reszt z modelu:

```
plot(l2, 1)
```



Widać, że rezidua wydlądają nieco lepiej niż w modelu liniowym (są bardziej spłaszczone). Heteroskedastyczność jest jakby mniejsza, niemniej jednak i tak jest zauważalna.

Przejdźmy teraz do formalnego sprawdzenia heteroskedastyczności dla obu modeli. Wykorzystamy do tego testy White'a, Breuscha-Pagana oraz Goldfelda-Quandta.

Zacznijmy od modelu liniowego:

```
library("lmtest")
library("bstats")

white.test(l)

##
##  White test for constant variance
##
## data:
## White = 16.73, df = 2, p-value = 0.0002323

bptest(l)

##
##  studentized Breusch-Pagan test for homoscedasticity
##
## data:  l
## BP = 15.39, df = 1, p-value = 8.766e-05

gqtest(wydatki ~ dochod, fraction = 0.33, order.by = ~dochod)

##
##  Goldfeld-Quandt test
##
## data:  wydatki ~ dochod
## GQ = 19.38, df1 = 5, df2 = 4, p-value = 0.006608
```

$P-value$ każdego z trzech testów jest mniejsze niż $0,05$, więc w każdym z testów hipotezę o równości wariancji odrzucamy. Mamy więc do czynienia z heteroskedastycznością. A w przypadku testu Breuscha-Pagana, znamy nawet charakter heteroskedastyczności, a mianowicie, wraz ze wzrostem dochodu, rośnie wariancja wydatków.

Zróbmy analogiczne testy dla modelu potęgowego:

```
white.test(l2)

##
##  White test for constant variance
##
## data:
## White = 19.87, df = 2, p-value = 4.837e-05

bptest(l2)

##
##  studentized Breusch-Pagan test for homoscedasticity
##
## data:  l2
## BP = 10.57, df = 1, p-value = 0.001148

gqtest(log(wydatki) ~ log(dochod), fraction = 0.33, order.by = ~dochod)

##
##  Goldfeld-Quandt test
##
## data:  log(wydatki) ~ log(dochod)
## GQ = 2.248, df1 = 5, df2 = 4, p-value = 0.2263
```

Analizując $p - value$ testów widzimy, że hipotezę odrzucimy w przypadku testów White'a i Breuscha-Pagana, zaś w przypadku testu Goldfelda-Quandta hipotezę przyjmiemy. Jako że większość testów (dwa z trzech) wskazuje na obecność heteroskedastyczności, więc w modelu tym również będziemy skłonni uważać, że rzeczywiście jest ona obecna. Widać jednak, że będzie już w takim razie mniejsza niż w przypadku modelu liniowego.

Jakie więc wyciągniemy wnioski? Przede wszystkim stwierdzamy, że model potęgowy będzie lepiej dopasowany niż liniowy. Po drugie, że w modelu obecna jest heteroskedastyczność (wariancja nie jest stała, a nawet jest funkcją monotoniczną). Widzimy więc, że wraz ze wzrostem dochodów, wariancja wydatków również rośnie. Interpretować należy to w ten sposób, że osoby zarabiające mało, bardziej pilnują i planują wydatki, zaś osoby zamożne mniej przejmują się tym, czy w danym miesiącu wydadzą mniej, czy więcej, gdyż mogą sobie na to finansowo pozwolić.

Zadanie 2.

Zbudujmy następujący model regresji:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \qquad t = 1, \ldots, 500,$$

gdzie $\alpha, \beta \in \mathbb{R}$, x_t są dowolnymi obserwacjami (w moim przypadku z rozkładu jednostajnego $U[1, 6]$) oraz zaburzenie ε_t jest zdefiniowane, jako:

$$\varepsilon_t = \eta_t \cdot \sqrt{a_0 + a_1 \varepsilon_{t-1}^2},$$

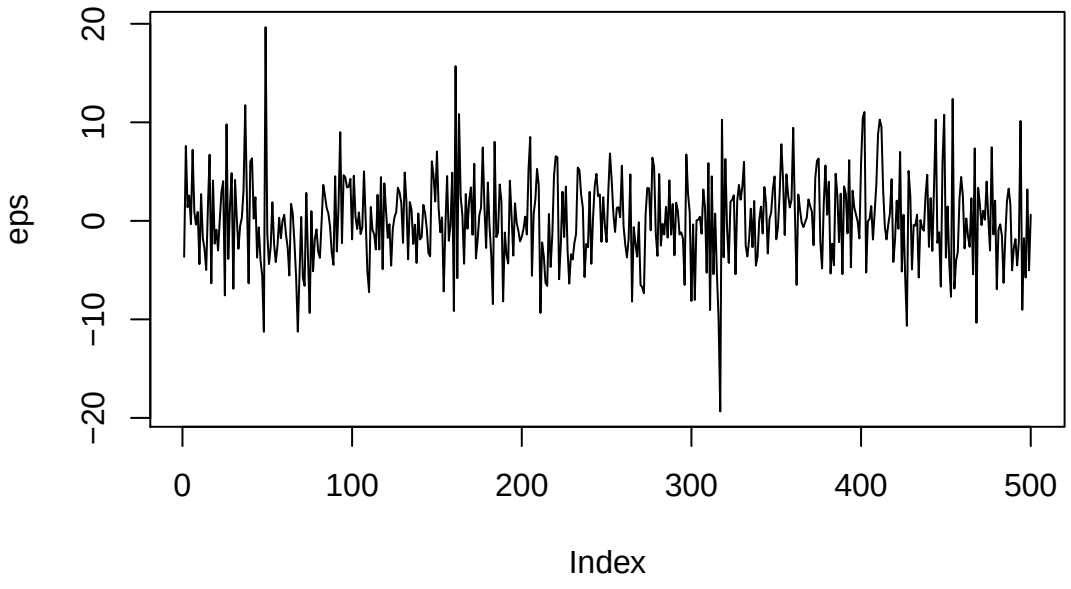
gdzie $\eta_t \sim \mathrm{N}(0, 1)$, a za a_0 i a_1 przyjmiemy odpowiednio liczby $12,5$ oraz $0,5$.

W związku z tym, że wzór na ε_t jest rekurencyjny bez podanej wartości początkowej, przyjmiemy $\varepsilon_1 = \eta_1$, ale żeby uniknąć błędu wynikającego z takiego założenia, zamiast 500 obserwacji, wygenerujemy ich 600, a następnie obetniemy 100 pierwszych, jako tych obarczonych błędem powyższego założenia.

```
x <- runif(500, 1, 6)
a0 <- 12.5
a1 <- 0.5
eta <- rnorm(600)
eps <- numeric(600)

eps[1] <- eta[1]
for (i in 2:600) {
    eps[i] <- eta[i] * sqrt(a0 + a1 * (eps[i - 1])^2)
}
eps <- eps[101:600]
```

Przyjrzyjmy się, jak wyglądają nasze ε_t na wykresie:

```
plot(eps, type = "l")
```

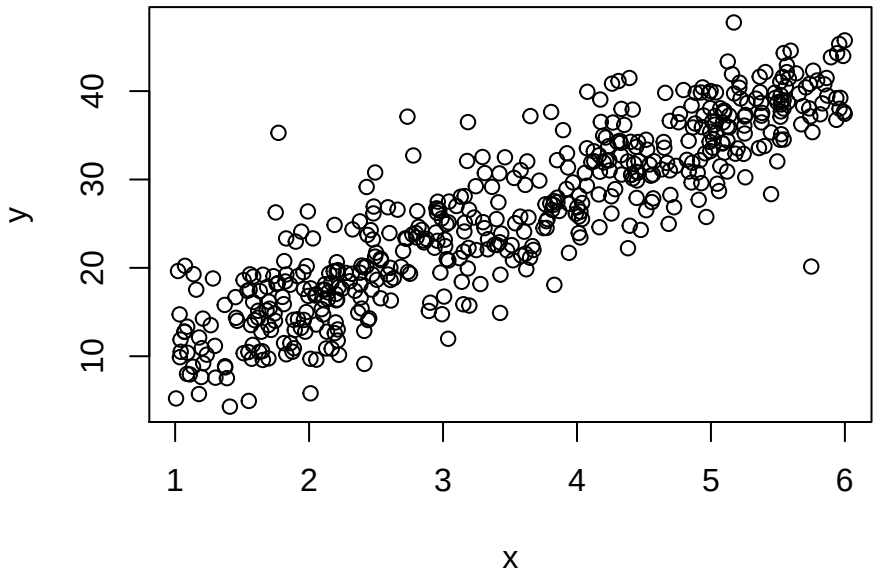


Rzeczywiście, widać, że wariancja „dziś", zależy od tej „wczoraj".

Stosując MNK, będziemy teraz chcieli oszacować parametry α i β. Żeby to jednak zrobić, musimy wygenerować

sobie obserwacje. Przyjmijmy więc $\alpha = 5$ i $\beta = 6$, wygenerumy dane, dopasujmy model liniowy, a następnie sprawdźmy, czy nasz model podał zbliżone do prawdziwych wartości α i β.

```
b0 <- 5
b1 <- 6
y <- b0 + b1 * x + eps
plot(y ~ x)
```



Na powyższym rysunku widać, że waraiancja y_t jest zmienna. Dopasujmy model liniowy:

```
l <- lm(y ~ x)
summary(l)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.572  -2.739   0.018   2.688  19.755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.749      0.524    9.07   <2e-16 ***
## x              6.083      0.138   43.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.41 on 498 degrees of freedom
## Multiple R-squared:  0.795,Adjusted R-squared:  0.795
## F-statistic: 1.93e+03 on 1 and 498 DF,  p-value: <2e-16
```

Z *summary* widać, że model jest dobrze dopasowany oraz zmienna x jest istotna. Model oszacował nam parametry następująco: $\alpha = 4,749, \beta = 6,083$. Ich wartości są więc bardzo zbliżone do wartości prawdziwych.

Obliczmy jeszcze błąd średniokwadratowy dla tego modelu:

```
mean((l$coefficients - c(5, 6))^2)

## [1] 0.03499
```

Zbudujmy teraz model metodą największej wiarogodności. Zmaksymalizujmy więc logarytm funkcji wiarogodności dany wzorem:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{n}\ln\left(a_0 + a_1(y_{t-1} - \alpha - \beta x_{t-1})^2\right) - \frac{1}{2}\sum_{t=1}^{n}\frac{(y_t - \alpha - \beta x_t)^2}{a_0 + a_1(y_{t-1} - \alpha - \beta x_{t-1})^2}$$

W tym celu użyjemy funkcji *optim()*, korzystającej z metody optymalizacji quasi-Newtona.

```
n <- 500

f <- function(a) {

    a0 <- a[1]
    a1 <- a[2]
    alfa <- a[3]
    beta <- a[4]

    m <- numeric(n - 1)
    for (t in 2:n) {
        m[t - 1] <- 1/2 * log(a0 + a1 * (y[t - 1] - alfa - beta * x[t - 1])^2) +
            (1/2 * (y[t] - alfa - beta * x[t])^2)/(a0 + a1 * (y[t - 1] - alfa -
                beta * x[t - 1])^2)
    }
    sum(m) + n/2 * log(2 * pi)
}

op <- optim(c(1, 1, 4.749, 6.083), f, method = "BFGS")$par
op

## [1] 11.9748  0.3979  4.9489  6.0025
```

Dzięki metodzie największej wiarogodności, otrzymujemy więc następujące rezultaty: $a_0 = 11,97$, $a_1 = 0,4$, $\alpha = 4,95$, $\beta = 6,00$. Na pierwszy rzut oka, rezultaty są więc lepsze od tych otrzymanych metodą najmniejszych kwadratów. Przekonajmy się jeszcze o tym formalnie, licząc błąd średniokwadratowy metody największej wiarogodności:

```
mean((op[3:4] - c(5, 6))^2)

## [1] 0.001306
```

Widzimy zatem, że dla MNK błąd śreniokwadratowy wynosi $0,03499$, zaś dla MNW 0.001306. Jest on o rząd wielkości mniejszy w przypadku metody największej wiarogodności, więc to ona w naszym modelu lepiej estymuje parametry α i β.

Zróbmy jeszcze krótką symulację. Wybierzmy 5 losowych ciągów zmiennych i oszacujmy parametry α i β dwiema metedami. Dla każdej z metod obliczmy błąd średniokwadratowy, a następnie policzmy jego średnią:

```
blad_mnk <- numeric(5)
blad_mnw <- numeric(5)

for (i in 1:5) {
```

```
    x2 <- runif(500, 1, 6)
    eta2 <- rnorm(600)
    eps2 <- numeric(600)

    eps2[1] <- eta2[1]
    for (j in 2:600) {
        eps2[j] <- eta2[j] * sqrt(a0 + a1 * (eps2[j - 1])^2)
    }
    eps2 <- eps2[101:600]
    y2 <- b0 + b1 * x2 + eps2
    l2 <- lm(y2 ~ x2)

    blad_mnk[i] <- mean((l2$coefficients - c(5, 6))^2)
    op2 <- optim(c(1, 1, l2$coefficients[1], l2$coefficients[2]), f, method = "BFGS")$par
    blad_mnw[i] <- mean((op2[3:4] - c(5, 6))^2)
}

mean(blad_mnk)

## [1] 0.1331

mean(blad_mnw)

## [1] 0.001256
```

Widać wyraźnie, że błąd średniokwadratowy metody największej wiarogodności $(0,001256)$ jest dużo mniejszy, niż błąd średniokwadratowy metody najmniejszych kwadratów $(0,1331)$. Metoda ta jest więc efektywniejsza.

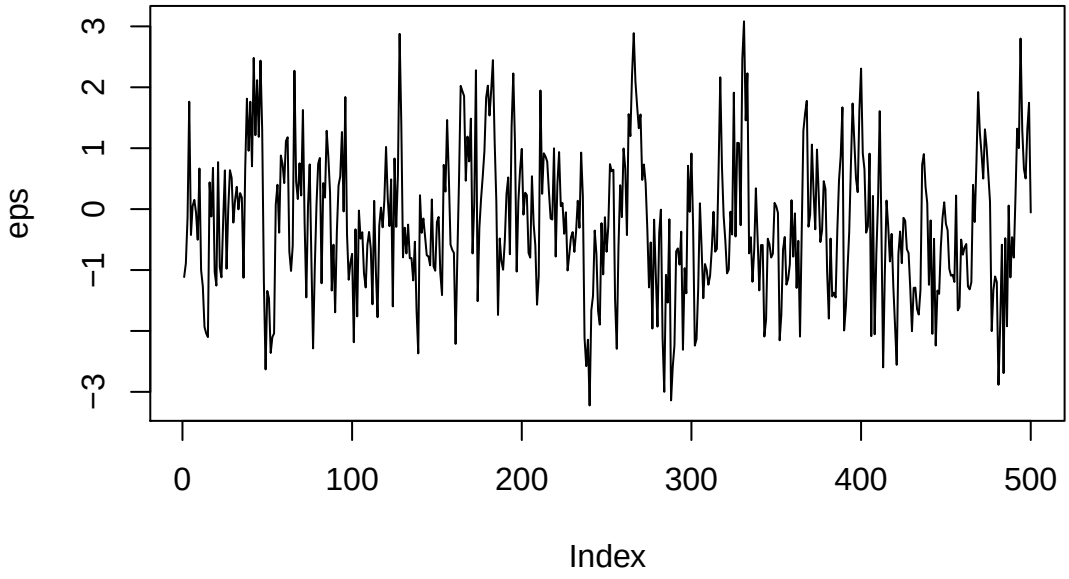
Przejdźmy teraz do trochę innego modelu. Będzie to model autoregresyjny $AR(1)$:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t,$$

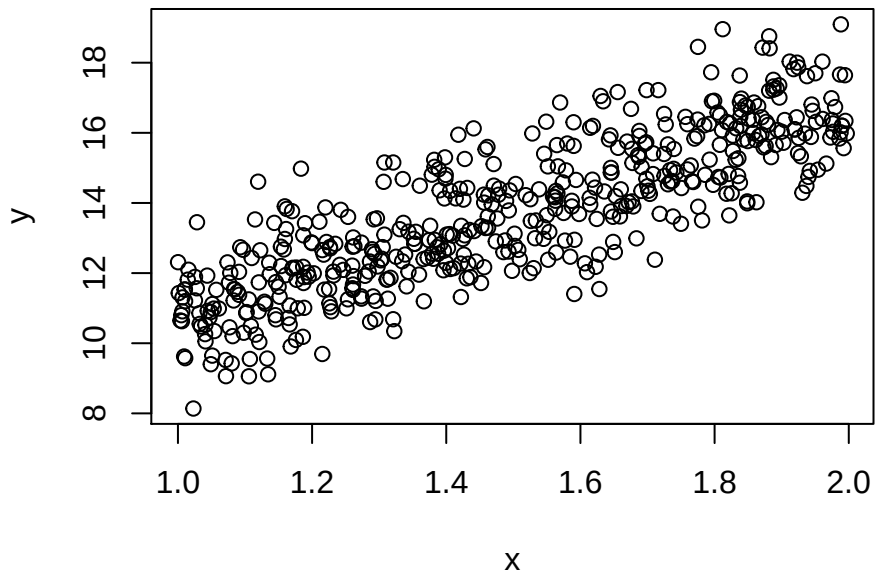
gdzie $\rho \in (0,1)$, $x_t \sim \mathrm{U}[1,2]$ a y_t oraz η_t są zdefiniowane tak, jak w poprzednim modelu.

Oszacujmy parametry α i β stosując MNK:

```
x <- runif(500, 1, 2)

rho <- 0.5
eta <- rnorm(600)

eps <- numeric(600)
eps[1] <- eta[1]
for (i in 2:600) {
    eps[i] <- rho * eps[i - 1] + eta[i]
}

eps <- eps[101:600]

plot(eps, type = "l")
```




```
b0 <- 5
b1 <- 6

y <- b0 + b1 * x + eps
plot(y ~ x)
```




```
l <- lm(y ~ x)
summary(l)

##
## Call:
```

```
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.011 -0.800 -0.077  0.755  3.279
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.629      0.272    17.0   <2e-16 ***
## x              6.097      0.179    34.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.17 on 498 degrees of freedom
## Multiple R-squared:   0.7,Adjusted R-squared:   0.7
## F-statistic: 1.16e+03 on 1 and 498 DF,  p-value: <2e-16
```

Estymatory α i β w tym modelu są zatem równe, odpowiednio, $4,629$ i $6,097$. A ich odchylenie standardowe to odpowiednio: 0.272 i 0.179. Policzmy błąd średniokwadratowy naszego dopasowania, przeprowadzając krótką symulację:

```
blad_mnk <- numeric(50)

for (i in 1:50) {
    x3 <- runif(500, 1, 2)
    eta3 <- rnorm(600)
    eps3 <- numeric(600)

    eps3[1] <- eta3[1]
    for (j in 2:600) {
        eps3[j] <- rho * eps3[j - 1] + eta3[j]
    }
    eps3 <- eps3[101:600]
    y3 <- b0 + b1 * x3 + eps3
    l3 <- lm(y3 ~ x3)

    blad_mnk[i] <- mean((l3$coefficients - c(5, 6))^2)
}

mean(blad_mnk)

## [1] 0.05251
```

Błąd średniokwadratowy wynosi 0.05251. Nie jest więc dostatecznie mały, ale dużo mniejszy, niż błąd dla MNK z poprzedniego modelu.

Z symulacji przeprowadzonych w tym ćwiczeniu widać więc, że MNK nie zachowuje się efektywnie, gdy mamy do czynienia z autokorelacją błędów, czyli pewnym odstępstwem od założeń. Lepszą metodą szacowania parametrów wydaje się zatem metoda największej wiarogodności. Jest ona jednak dużo bardziej skomplikowana obliczeniowo i numerycznie niż metoda najmniejszych kwadratów.

Zadanie 3.

Będziemy w tym zadaniu rozważać kursy zamknięcia spółki PZU na giełdzie. W tym celu ściągnęliśmy z internetu potrzebne dane – kursy zamknięcia spółki PZU w latach 12.05.2010 – 13.03.2014 i odpowiednie notowania WIG-u, również w tych latach.

Dla naszej spółki utwórzmy logarytmiczne stopy zwrotu:

```
wig <- read.csv2("C:\\Users\\Marta\\Desktop\\Marta\\studia\\rok4\\Ekonometria\\spr2\\wig_d.csv",
    header = TRUE, sep = ",")
head(wig, 3)
```

```
##         Date    Open    High     Low   Close   Volume
## 1 2010-05-12   41522 41971.3 41440.2 41896.6 58682559
## 2 2010-05-13 42283.1 42390.1 41741.5 41914.8 83412680
## 3 2010-05-14 41781.8 41855.6 41075.4 41075.4 57561466
```

```
pzu <- read.csv2("C:\\Users\\Marta\\Desktop\\Marta\\studia\\rok4\\Ekonometria\\spr2\\pzu_d.csv",
    header = TRUE, sep = ",")
head(pzu, 3)
```

```
##         Date      Open      High       Low     Close  Volume
## 1 2010-05-12 263.59874 271.90696 262.91891 271.90696 9566251
## 2 2010-05-13 270.39636 273.34201 267.37516 269.64106 2394317
## 3 2010-05-14 268.35704 271.90696 266.84653 268.13046 1399531
```

```
n <- nrow(pzu)
attach(pzu)
kurs_zamkn <- as.numeric(as.vector(pzu$Close))

stop_zwr <- numeric(n - 1)
for (i in 2:length(kurs_zamkn)) {
    stop_zwr[i] <- log(kurs_zamkn[i]/kurs_zamkn[i - 1])
}

kurs_zamkn_wig <- as.numeric(as.vector(wig$Close))

stop_zwr_wig <- numeric(n - 1)
for (i in 2:length(kurs_zamkn_wig)) {
    stop_zwr_wig[i] <- log(kurs_zamkn_wig[i]/kurs_zamkn_wig[i - 1])
}

mean(stop_zwr)
```

```
## [1] 0.0004248
```

Średnia logarytmiczna stopa zwrotu w tych latach wynosi 0.0004248. Widać więc, że średnio PZU zyskało.

Policzmy teraz współczynnik agresywności dla naszej spółki, korzystając z modelu:

$$(R_t - r_f) = \alpha + \beta(RM_t - r_f) + \eta_t,$$

gdzie RM_t jest logarytmiczną stopą zwrotu WIG-u, R_t logarytmiczną stopą zwrotu naszej spółki, r_f wynosi 5% w skali roku, a η_t to błąd losowy. Oszacujemy więc współczynnik agresywności β metodą najmniejszych kwadratów:

```r
rf <- 0.05/365
y <- stop_zwr - rf
x <- stop_zwr_wig - rf
l <- lm(y ~ x)

summary(l)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03530 -0.00721 -0.00026  0.00683  0.04487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.000247   0.000358    0.69     0.49
## x           0.909118   0.032770   27.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0111 on 959 degrees of freedom
## Multiple R-squared:  0.445,Adjusted R-squared:  0.445
## F-statistic:  770 on 1 and 959 DF,  p-value: <2e-16
```

Widać więc, że współczynnik agresywności $\beta = 0.91$, jest więc mniejszy niż 1. Oznacza to, że spółka nie zachowywała się agresywnie na giełdzie, czyli grała raczej asekuracyjnie – niewiele ryzykowała, ale też niewiele traciła.

Przeprowadźmy test Engle'a, żeby sprawdzić, czy w modelu występuje heteroskedastyczność drugiego rodzaju:

```r
library("FinTS")
ArchTest(l$residuals, 2)
```

```
##
##  ARCH LM-test; Null hypothesis: no ARCH effects
##
## data:  l$residuals
## Chi-squared = 6.045, df = 2, p-value = 0.04869
```

$P - value$ testu jest małe, zatem odrzucamy hipotezę, czyli występuje heteroskedastyczność drugiego rodzaju.

Zróbmy jeszcze test na występowanie heteroskedastyczności pierwszego rodzaju:

```r
white.test(l)
```

```
##
##  White test for constant variance
##
## data:
## White = 28.97, df = 2, p-value = 5.131e-07
```

$P - value$ testu White'a jest małe, zatem heteroskedastyczność pierwszego rodzaju występuje.

Nie mamy zatem jednorodności wariancji (zależy ona od dnia) oraz to co dziś zależy od tego, co było wczoraj.

Sprawdźmy teraz, czy w modelu występuje autokorelacja:

```
library("stats")
Box.test(l$residuals, type = "Ljung-Box")

##
##  Box-Ljung test
##
## data:  l$residuals
## X-squared = 6.981, df = 1, p-value = 0.008237

dwtest(l, alternative = "two.sided")

##
##  Durbin-Watson test
##
## data:  l
## DW = 2.168, p-value = 0.009223
## alternative hypothesis: true autocorrelation is not 0
```

Zarówno test Ljunga-Boxa, jak i test Durbina-Watsona wskazuje na obecność autokorelacji. Wprowadźmy więc poprawkę Newey'a-Westa na wariancję estymatorów MNK i sprawdźmy testem t, czy wtedy parametry są istotne:

```
library("sandwich")
nw <- NeweyWest(l)

co <- l$coefficients

T1 <- (co[1] - 0)/sqrt(nw[1, 1])
2 * min(pt(T1, n - 2), pt(T1, n - 2, lower.tail = FALSE))

## [1] 0.3765
```

$P-value$ testu na istotność parametru α jest więc małe i wskazuje na nieistotność parametru. Sprawdźmy, czy istotne jest β:

```
T2 <- (co[2] - 0)/sqrt(nw[2, 2])
2 * min(pt(T2, n - 2), pt(T2, n - 2, lower.tail = FALSE))

## [1] 1.197e-70
```

$P-value$ testu na istotność parametru β jest małe i wskazuje na istotność parametru β (współczynnik agresywności). Jako, że w modelu wyszło nam, że $\beta = 0.91$, sprawdźmy jeszcze, czy nasz parametr jest istotnie mniejszy od 1:

```
T3 <- (co[2] - 1)/sqrt(nw[2, 2])
pt(T3, n - 2)

##       x
## 0.02671
```

$P-value$ tego testu znowu jest małe, zatem rzeczywiście przyjmujemy hipotezę, że nasz współczynnik agresywności jest istotnie mniejszy od jedynki. Potwierdzają się więc wnioski, że spółka PZU nie grała agresywnie na giełdzie – zachowywała się raczej asekuracyjnie.