Аналіз заданого часового ряду з дисципліни "Нелінійні часові ряди" Студента 2 курсу магістратугри групи "Статистика"

Горбунов Даніел

6 грудня 2022 р.

Зміст

1 Вступ.

Дана робота присвячена аналізу часового ряду, що репрезентує ціну акції Johnson & Johnson (скорочено JNJ) на момент відкриття фондового ринку (тобто торгів) [1]. Поведінка ціни досліджувалася у рамках взятого вікна з 2002 по 2012 роки. На основі відповідних статистичних тестів та деякої логіки була обрана найбільш доречна модель для прогнозування "локальної" поведінки ціни.
Інколи буває цікаво наперед вгадати ціну на ризиковий актив у наступний період торгів. Складність задачі полягає у тому, що поведінка ціни є недермінованим (випадковим) процесом, який може варіюватися з часом. Тому доречно підібрати певну часову модель локально та, відповідно, робити прогноз на декілька невеликих кроків вперед.

2 Хід роботи.

2.1 Опис даних.

Дані про часовий ряд містяться у таблиці, яку можна завантажити з джерела [1].

```
# Зчитуємо дані
data <- read.csv('data.csv', skip=14)
data.subset <- data[,c("date", "open")]
```

Окрім ціни "відкриття", у фреймі маємо відомості про ціну "закриття" акції (при завершенні торгів у заданий день) та її екстремальні значення протягом виконання торгів. Величина "об'єму" відповідає, наскільки гадає автор роботи, кількості виконаних торгів протягом дня.

```
> head(data)
        date   open   high    low  close  volume
1 1970-01-02 0.4079 0.4090 0.3977 0.3977 1195200
2 1970-01-05 0.3977 0.3988 0.3875 0.3880  964800
3 1970-01-06 0.3880 0.3954 0.3863 0.3954 1036800
4 1970-01-07 0.3954 0.3965 0.3926 0.3937  331200
5 1970-01-08 0.3937 0.3999 0.3937 0.3965  460800
6 1970-01-09 0.3988 0.4045 0.3988 0.3988  748800
> dim(data)
[1] 13347     6
```

2.2 Підготовча робота.

У наявній таблиці дні проведення торгів відрізняються у кожному місяці (перідоика така: або дані про торги вже наявні для наступного дня, або через декілька). Тому спроба "розмітити" ряд щоденно не здається коректною. Спробуємо спростити дослідження, беручи до уваги лише інформацію про ціну акції лише в перший день кожного місяця.

```
# Розбиваємо колонку з датою за роздільником '-'
date.split <- strsplit(data.subset$date, split='-')
...
```

```
...
# Виокремлюємо рік, місяць і день
data.subset["year"] <- unlist(lapply(date.split, function(u) u[1]))
data.subset["month"] <- unlist(lapply(date.split, function(u) u[2]))
data.subset["day"] <- unlist(lapply(date.split, function(u) u[3]))

# Проміжна змінна надалі не потрібна
date.split <- NULL

# Збираємо щомісячні дані (перший відомий день кожного місяця)
years <- unique(data.subset$year)
months <- unique(data.subset$month)

len.months <- length(months)
len.years <- length(years)
len.vals <- len.years * len.months
values <- numeric(len.vals)
for(i in 1:len.years)
{
  cur.year <- years[i]
  for(j in 1:len.months)
  {
    cur.month <- months[j]
    values[(i - 1) * len.months + j] <- data.subset[
      (data.subset$year == cur.year) & (data.subset$month == cur.month),
      ]$open[1]
  }
}

# За грудень 2022 року наразі немає інформації, вилучаємо пропуск
values <- as.numeric(na.omit(values))

# Остаточний часовий ряд
data.ts <- ts(
  data=values, start=c(1970, 1), end=c(2022, 11), frequency=12
)
```

Напевно, можна було б якось простіше зробити, без використання подвійних циклів. Втім, бажаний результат досягнуто, а саме можливість розглянути часовий ряд на місячній шкалі. Наступний рисунок покаже нелінійність досліджуваного ряду та його неоднорідність на всьому часовому проміжку.

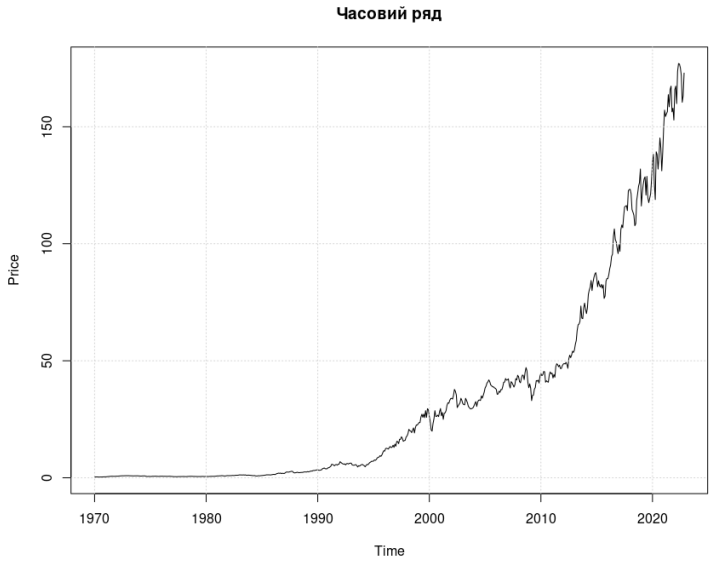

Рис. 1: Ціна ”відкриття” на акцію JNJ, щомісячні дані з 1970 по 2022 роки.

До середини дев’яностих років спостерігається повільне збільшення ціни на акцію, динаміка стає помірною протягом двохтисячних, а стрімко зростає з 2010 року. Загальна форма нагадує експоненту за характером зростання. Аби побороти ”екстремальні кроки” підйому на кінцях, прологарифмуємо спостережуваний ряд.

```
# Логарифмування та демонстрація перетвореного ряду
log.data.ts <- log(data.ts)
plot(log.data.ts, main="Часовий ряд після логарифмування", ylab="Price")
grid()
```

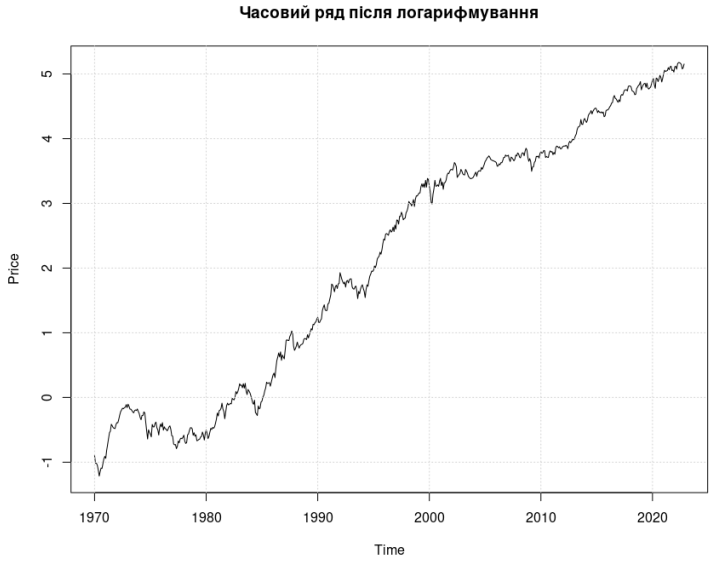

Рис. 2: Логарифмована ціна ”відкриття” на акцію JNJ, щомісячні дані з 1970 по 2022 роки.

Логарифмічне перетворення лінеаризувало шматками поведінку ряду.

2.3 Звуження по часу, створення вибірки.

До 2000-х років логарифмована ціна має неоднорідні коливання, що може ускладнити аналіз. Навпаки, починаючи з 2000-го року, логарифм зростає більш-менш помірно, якщо не брати до уваги (насправді треба) стрибкоподібне підвищення десь у 2013 році. Спробуємо підібрати таку модель, яка б добре описувала поведінку протягом 2002 та 2010 років (тобто локальну поведінку), використовуючи попередньо відомі моделі часових рядів, пов'язаних зі стаціонарністю.

```
# Цікавлять дані за 2002-2012 роки
sampled.ts <- window(log.data.ts, start=c(2002, 1), end=c(2012, 12))
plot(sampled.ts, main="Вибірка часового рядку", ylab="Price")
grid()
```

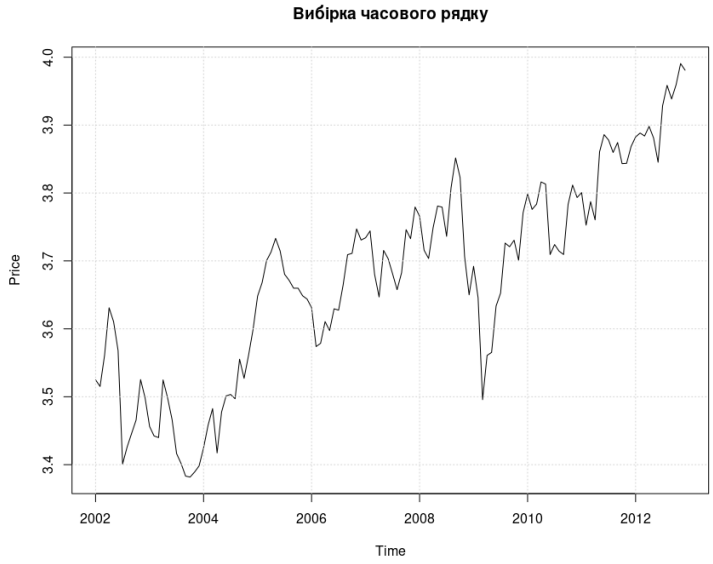

Рис. 3: Вибірка процесу ціни, щомісячні дані з 2002 по 2012 роки.

2.4 Дослідження ряду на стаціонарність.

Спробуємо витягти білий шум, що здається можливим після першого взяття дискретної похідної. Переконавшись у цьому, маємо наступний процес "залишків":

```
# Спроба прибрати тренд
delta.sampled.ts <- diff(sampled.ts)
plot(delta.sampled.ts, main="Вибірка часового рядку", ylab="Price")
grid()
abline(h=0, lty=2, col="red")
```

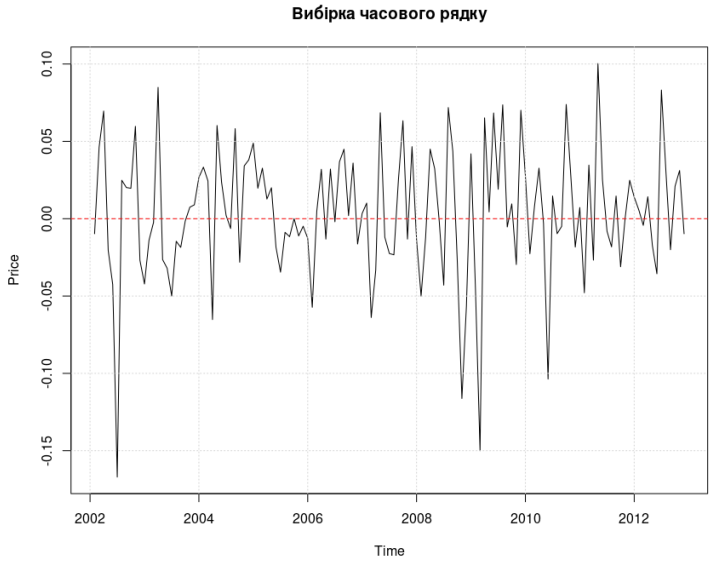

Рис. 4: Продиференційована вибірка процесу ціни, щомісячні дані з 2002 по 2012 роки.

Утворенням диференціюванням процес віддаленно нагадує процес білого шуму, за виключенням деяких викидих у відповідно місяці. Має сенс переглянути діаграми звичайної та частинної (вибіркових) автокореляцій.

```
# Диференціювання дало можливість позбутися тренду.
# Побудова діаграм для автокореляції та частинної автокореляції
acf(delta.sampled.ts)
pacf(delta.sampled.ts)
```

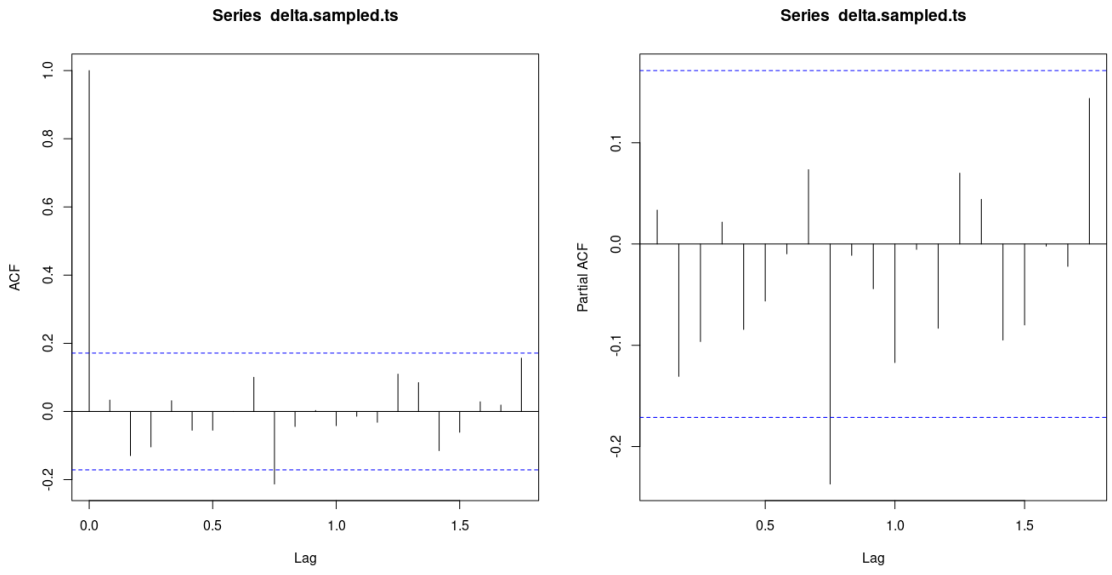

Рис. 5: Діаграми звичайної та частинної автокореляцій.

Про слабкий білий шум нам відомо про ортогональність його елементів та існування скінченного другого моменту. Зокрема частинна автокореляція для цього шуму має вийти нульовою внаслідок ортогональності компонент процесу.

Попередні факти майже вимальовується на діаграмах з вибірковими альтернативами функціоналів, якби не викиди, що спотворили значення при значенні лагу $h = 9$. Щоб більше переконатися у тому, що дискретну похідну можна вважати слабким білим шумом, застосуємо тест Ljung-Box [2]. Грубо кажучи, тест перевіряє наступні гіпотези на $X(t) := \Delta Y(t)$:

$$\mathbf{H}_0 : \ X(t) \text{ є слабким білим шумом}, \ \mathbf{H}_1 : \ X(t) \text{ не є слабким білим шумом}$$

Для перевірки цих гіпотез, у рамках тесту обичслюються h вибіркових автоковаріацій (відповідно із зсувами від 1 до h). Отримані значення збираються у так звану статистику Box-Ljung:

$$Q = T(T+2)\sum_{k=1}^{h}\frac{(\hat{\rho}_X(k))^2}{T-k}, \tag{1}$$

де $\hat{\rho}_X(k)$ – вибіркова автокореляція $X(t)$, T – довжина спостережуваної траєкторії процесу, h – кількість автоковаріацій для обчислення. У роботі [2] показано, що за умови гауссовості $X(t)$, розподіл статистики (1) наближається до розподілу χ^2 з h ступенями вільності.
Для $h = \overline{1, 20}$, обчислимо досягнуті рівні значущості попередньо зазначеного тесту (без додаткових правок на отримані значення).

```
# Є підозора, що диференційований процес є шумом. Застосуємо тест Ljung-Box
p.values.lb <- numeric(20)
for(h in 1:20)
{
  lb.test <- Box.test(delta.sampled.ts, lag=h, type="Ljung-Box")
  p.values.lb[h] <- lb.test$p.value
}
print(p.values.lb)
```

В результаті маємо масив з обчислених рівнів значущості для кожного "ступеня зсунутості":

```
> print(p.values.lb)
 [1] 0.6986811 0.2990619 0.2743471 0.4032668 0.4881557 0.5619855 0.6771322
 [8] 0.6155291 0.1738540 0.2215251 0.2907366 0.3480724 0.4232058 0.4897109
[15] 0.4317086 0.4279438 0.3651810 0.3946287 0.4516747 0.5129972
```

Якщо вважати стандартний рівень значущості $\alpha = 0.05$, то можна вважати, що продиференційований часовий ряд пройшов тест на відповідність слабкому білому шуму. Варто звернути увагу на спотворення p-value при $h = 9$ (базуючись на попередніх результатах).
Єдине зауваження в тому, що від процесу вимагалася умова гауссовості, а вона поки не була перевірена. Побудуємо квантильну діаграму.

```
# Побудова QQ-діаграми для приростів відносно гауссового розподілу
qqnorm(delta.sampled.ts)
qqline(delta.sampled.ts)
grid()
```

Далі перевіримо візуальну узгодженість процесу із теоретичними квантилями нормального розподілу:

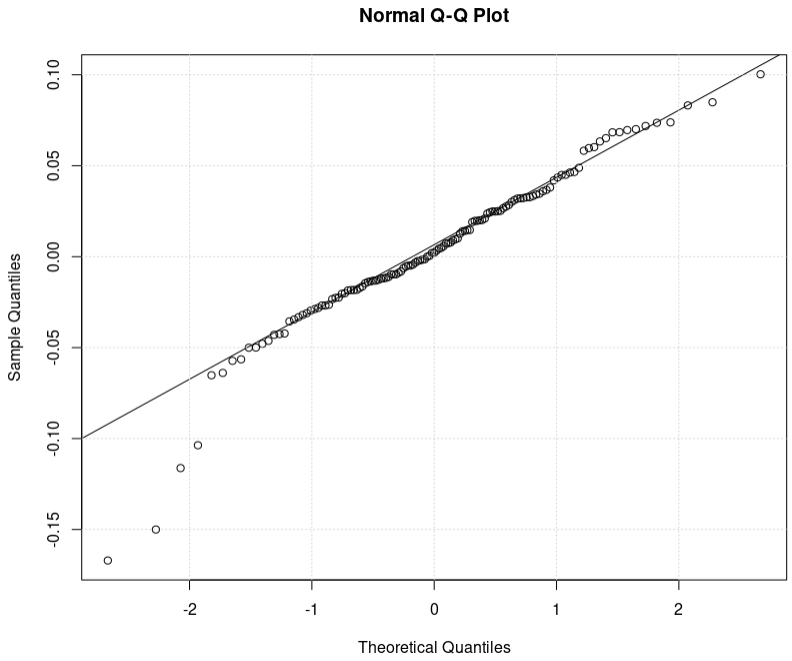

Рис. 6: Квантильна діаграма приростів процесу відносно нормального розподілу.

Добре видно як "ліві" викиди сильно тягнуть за собою нижні квантилі. Хоча верхні вибіркові квантилі теж здаються дещо зміщеними від теоретичних. Застосувавши один із тестів для перевірки гіпотези про нормальну рощзподіленість даних (виступає основною гіпотезою), наприклад тест Shapiro-Wilk [3], отримаємо бали на користь того, що висунуту гіпотезу слід відхилити:

```
> # Перевірка на гауссовість
> sw.test <- shapiro.test(delta.sampled.ts)
> print(sw.test)

        Shapiro-Wilk normality test

data:  delta.sampled.ts
W = 0.95286, p-value = 0.0001779
```

Даний тест досить чутливий до збурень. Якщо нехтувати екстремальними значеннями такого плану, тест працює "нормально":

```
> # Позбуваємося крайніх двох викидів зліва
> sw.test <- shapiro.test(delta.sampled.ts[abs(delta.sampled.ts) <= 0.12])
> print(sw.test)

        Shapiro-Wilk normality test

data:  delta.sampled.ts[abs(delta.sampled.ts) <= 0.12]
W = 0.9886, p-value = 0.3644
```

Два викиди відносно ≈ 130 спостережень часового ряду – таке число, яким можна на один раз закрити очі (відносно $2\hat{\rho}_X(0)$), тому автор вважає допустимим такі переходи. *Натягнуто.*

Незважаючи на те, що загальна картина моделі майже намальована, можна скористатися тестом Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [4], для перевірки основної гіпотези про стаціонарність відносно тренду досліджуваного ряду (тобто можна позбутися тренду, щоб отримати стаціонарний процес в результаті).

```
> # Чи є стаціонарність відносно тренду? Використаємо KPSS тест
> KPSS.test <- kpss.test(sampled.ts, null="Trend")
Warning message:
In kpss.test(sampled.ts, null = "Trend") :
  p-value greater than printed p-value
> print(KPSS.test)
        KPSS Test for Trend Stationarity
data:  sampled.ts
KPSS Trend = 0.10584, Truncation lag parameter = 4, p-value = 0.1
```

Результат тесту іде на користь отриманих міркувань раніше. Варто сказати про таке: подальший вибір моделі та аналіз якості дадуть розуміння того, чому початковий часовий ряд не задовольняє умову стаціонарності.

2.5 Вибір оптимальної моделі

З попередніх результатів виникає думка: можливо, випадкове блукання одиничного порядку $I(1) := ARIMA(0,1,0)$ зможе підійти для досліджуваного процесу? Можна також перевірити, що стаціонарний процес авторегресії $AR(1)$ не буде вдалим вибором з певних причин (одна з них – неузгодженість із емпіричним спостереженням поведінки процесу).

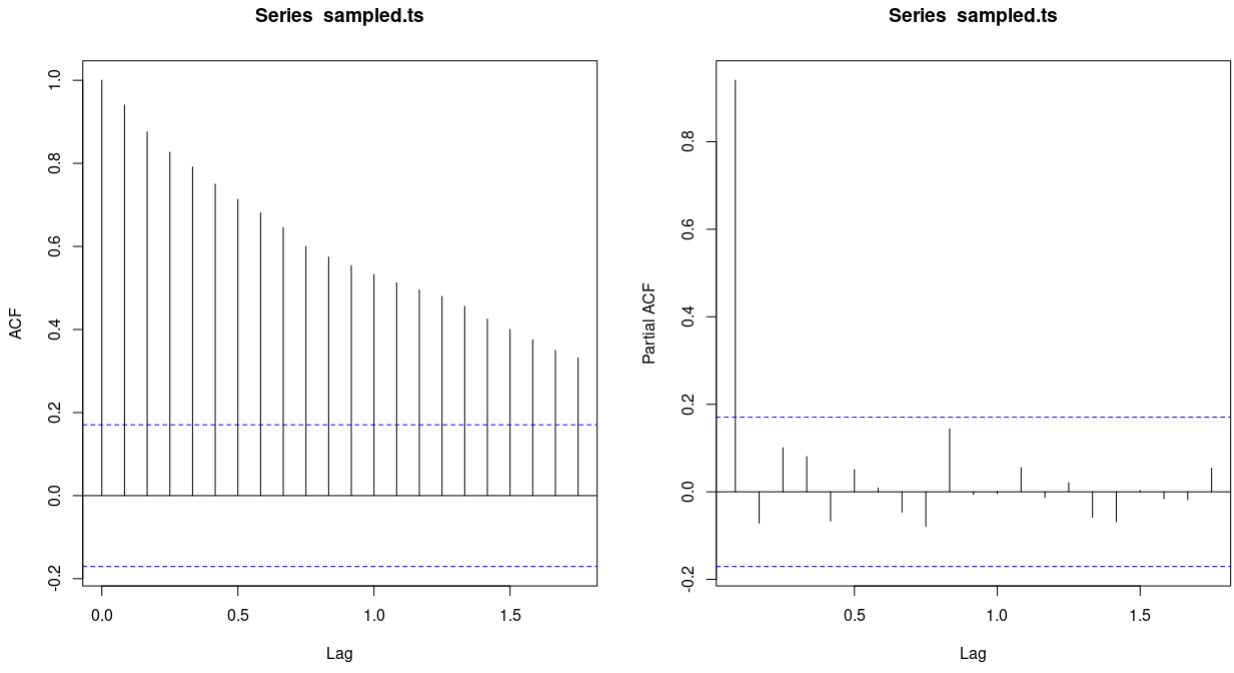

Рис. 7: Діаграми звичайної та частинної автокореляцій.

За формою ситуація нагадує випадок авторегресійної моделі. Чому немає стаціонарності?

Спочатку цікаво буде подивитися на форму залежності між $X(t)$ та $X(t-1)$ – виходить чітка лінійна залежність, де на око можна накинути коефіцієнт нахилу рівним одиниці.

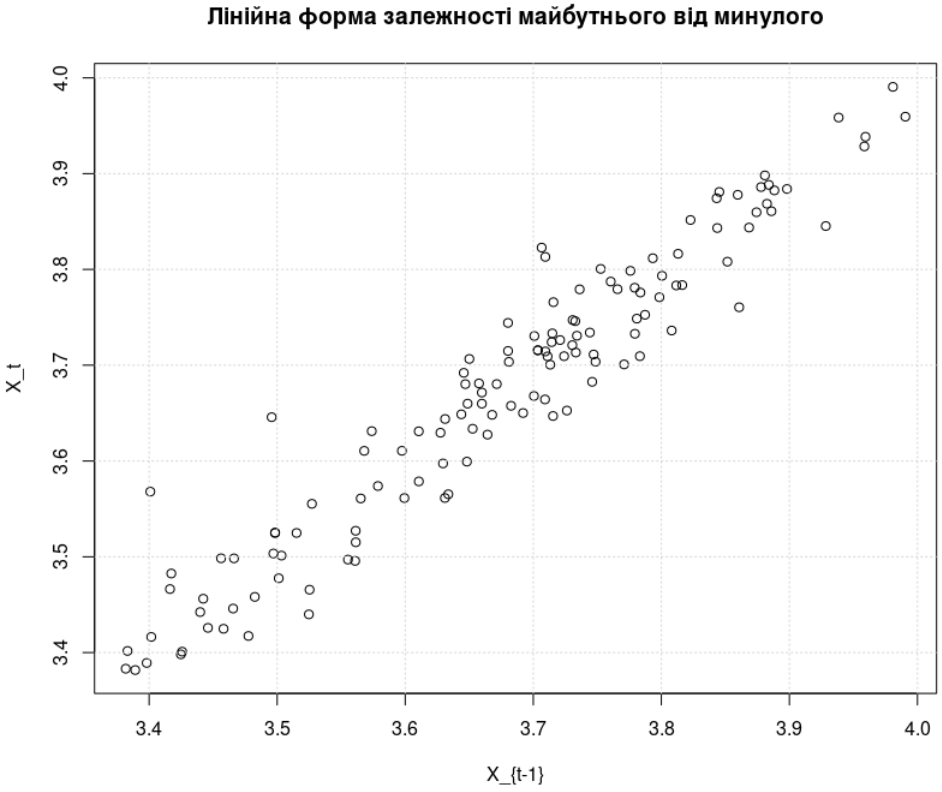

Рис. 8: Діаграма розсіювання майбутніх значень на попередні.

За нашим припущенням – досліджується модель авторегресії. Виходить, що відповідний параметр виходить близьким до одиниці, а це вже псує ідею стаціонарності спостережуваного ряду.
Застосуємо тест Dickey-Fuller [5] для перевірки гіпотези на коефіцієнт авторегресії.

```
> # Augmented Dickey-Fuller test
> df.test <- adf.test(sampled.ts)
> print(df.test)

        Augmented Dickey-Fuller Test

data:  sampled.ts
Dickey-Fuller = -3.3791, Lag order = 5, p-value = 0.06113
alternative hypothesis: stationary
```

Для заданого рівня значущості $\alpha = 0.05$, альтернативна гіпотеза про стаціонарність ряду відхиляється. Для розуміння цього треба підігнати параметри у відповідній моделі. У даному випадку мається на увазі підгонка параметрів в $AR(1)$.

```
> # Моделювання AR(1) = ARIMA(1,0,0)
> model.ar1 <- arima(sampled.ts, order=c(1,0,0))
> print(model.ar1)
Call:
arima(x = sampled.ts, order = c(1, 0, 0))
Coefficients:
         ar1  intercept
      0.9687     3.7006
s.e.  0.0218     0.0982
sigma^2 estimated as 0.001796:  log likelihood = 228.58,  aic = -451.16
```

Значення оцінки коефіцієнта при лаговому операторі близьке до одиниці. Для гарантування стаціонарності процесу авторегресії треба вимагати асимптотичну стійкість відповідного різницевого рівняння, а це можливе лише у тому разі, коли його корені лежить всередині одиничного кола. Можна очікувати, що коефіцієнт може бути дорівнювати одиниці, тому говорити про стаціонарність ряду справді не можна.
Неважко переконатися, що використання $MA(1)$ збільшить втрати Акаїке:

```
> # Моделювання MA(1) = ARIMA(0,0,1)
> model.ma1 <- arima(sampled.ts, order=c(0,0,1))
> print(model.ma1)
Call:
arima(x = sampled.ts, order = c(0, 0, 1))
Coefficients:
         ma1  intercept
      0.8163     3.6751
s.e.  0.0377     0.0143
sigma^2 estimated as 0.008248:  log likelihood = 128.8,  aic = -251.61
```

На основі показаних результатів, моделювати локальну поведінку зазначеними моделями не є коректним рішенням.

Варто зауважити, що альтернативою у тесті Dickey-Fuller є припущення про те, що спостережуваний процес є випадковим блуканням $I(1)$. Спадає на думку зробити підгонку у цій моделі:

```
> # Моделювання I(1) = ARIMA(0,1,0)
> model.rw1 <- arima(sampled.ts, order=c(0,1,0))
> print(model.rw1)
Call:
arima(x = sampled.ts, order = c(0, 1, 0))
sigma^2 estimated as 0.001818:  log likelihood = 227.42,  aic = -452.84
```

Попередньо було показано (за допомогою KPSS тесту), що доречно використати модель випадкового блукання. Можна побачити, що модель набуває найменших втрат Акаїке серед трьох розглянутих моделей часових рядів.

2.6 Прогнозування ціни

Тепер до веселої частини – прогнозування на основі підігнаних моделей. Буде зроблено прогноз ціни на рік вперед (тобто прогноз охопить весь 2013 рік, $h = 12$) на основі "оптимальної" моделі – випадкове блукання, та "поганої" моделі – процес авторегресії $AR(1)$. Прогноз вийде досить віддаленим (робиться дванадцять кроків вперед), якісь автором не гарантується з двох причин:

1. Прогноз в обох моделях буде лінійним,
2. Конкретно 2013 рік запам'ятався стрибкоподібним підйомом ціни на акцію. Історичні дані мають лише поведінку до цього "зламу", тому аномалія ніяк не врахована.

```
# Прогнозування
library('forecasting')
# Зробимо прогноз ціни на рік вперед
years.next <- 1

# ARIMA(1,0,0)
Arima.model <- Arima(sampled.ts, order=c(1,0,0))
pred.test <- forecast(Arima.model, h=years.next * 12)
true.test <- window(log.data.ts, start=c(2013, 1), end=c(2013, 12))
plot(pred.test, ylim=c(min(sampled.ts), max(pred.test$mean, true.test)))
lines(window(log.data.ts, start=c(2012, 12), end=c(2013, 12)), lty=2)
grid()

# ARIMA(0,1,0)
Arima.model <- Arima(sampled.ts, order=c(0,1,0))
pred.test <- forecast(Arima.model, h=years.next * 12)
true.test <- window(log.data.ts, start=c(2013, 1), end=c(2013, 12))
plot(pred.test, ylim=c(min(sampled.ts), max(pred.test$mean, true.test)))
lines(window(log.data.ts, start=c(2012, 12), end=c(2013, 12)), lty=2)
grid()
```

Продемонструємо далі як форми прогнозу кожної з моделей.

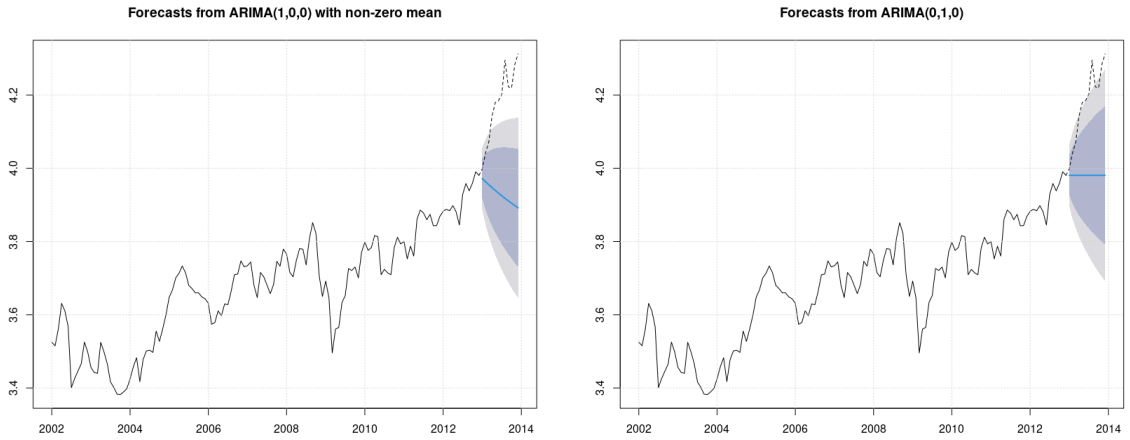

Рис. 9: Лінійний прогноз ціни на рік (на $h = 12$ кроків вперед) із довірчими смугами рівнів 0.8 (темно-синя) та 0.95 (сіра).

Як видно, результат не вражаючий. Прогноз авторегресії є суто затухаючим (внаслідок відсутності адитивної стохастичної компоненти, яка нам невідома), а для випадкового блукання майже незмінним (з аналогічної причини). Використання такого прогнозу на далеке майбутнє (як-от розглянуте) не буде хорошим рішенням. Випадкове блукання підійде для прогнозування "смуги пересування" процесу для коротких змін по часу. Занепокоєння з приводу неможливості вхопити зміни в 2013 році підтвердилися з очевидних (попередньо зазначених) причин.
Буде цікаво просто змоделювати продовження процесу випадкового блукання, використовуючи попередні результаті про "підходящу" модель для похибок.

```
# Моделювання на рік вперед

set.seed(121)
model.test <- c(sampled.ts[length(sampled.ts)], numeric(length(pred.test)))
eps.test <- rnorm(length(pred.test), mean=0, sd=sqrt(model.rw1$sigma2))
for(j in 1:length(pred.test))
{
  model.test[j+1] <- model.test[j] + eps.test[j]
}
model.test <- ts(model.test, start=c(2012, 12), end=c(2013, 12), frequency=12)

plot(window(log.data.ts, start=c(2002, 1), end=c(2013, 12)), ylim=c(
  min(sampled.ts, model.test, true.test),
  max(sampled.ts, model.test, true.test)
), ylab="Price", main="Modeled future values using ARIMA(0,1,0)")
lines(model.test, col="red", lty=2)
grid()
```

Продемонструємо 10 змодельованих випадковим блуканням траєкторій.

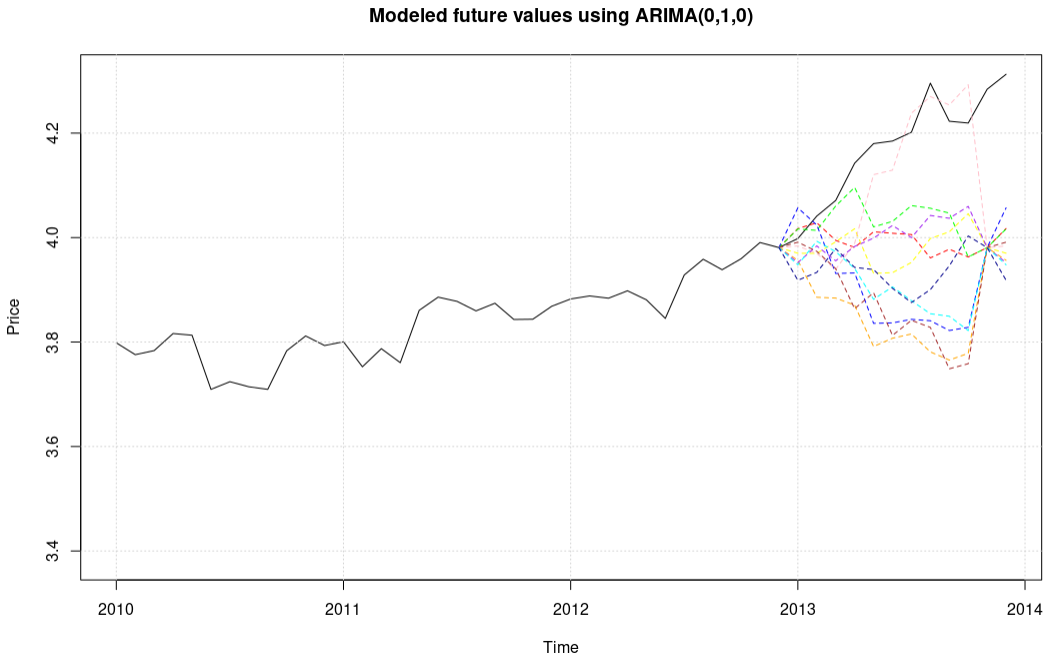

Рис. 10: Десять змодельованих траєкторій ціни на рік вперед.

Детермінований прогноз не вхоплює зміни, хоча реалізувати траєкторію, наближену до фактичної в 2013 році, цілком можливо. Такий експеримент наївний, але підтверджує узгодженість даних із обраною моделлю.

3 Висновки.

Аналіз локальної поведінки щомісячної ціни на акцію JNJ з 2002 по 2012 роки дав вагомі результати для вибору моделі, що найкраще відповідає даним на обраному часовому проміжку. Логарифмічна трансформація дала змогу зробити часовий ряд кусково-лінійним. Виявилося, що дискретна похідна ряду непогано імітує білий шум, за виключенням незначної кількості викидів (які, ймовірно, мають природний характер). Було виявлено, що характер залежності значень процесу на себе є лінійним, тобто використання процесу з авторегресійною залежністю має місце. Інше питання, чи доречно використовувати стаціонарну модель авторегресії? Як виявили тести Dickey-Fuller та KPSS, досліджуваний ряд не має характерні ознаки стаціонарності (на відміну від його приростів). Лінійний прогноз підігнаної моделі випадкового блукання можна використовувати лише для малих кроків вперед, внаслідок збільшення дисперсії прогнозу. Модельовані траєкторії за обраною моделлю частково нагадують справжню поведінку часового ряду в рамках визначеного, короткого проміжку часу. Додатково, частково повторюючись, зауважено, що прогноз на основі обранох моделі має лише локальний характер.

Література

[1] Дані про ціну на акцію JNJ: https://www.macrotrends.net/stocks/charts/JNJ/johnson-johnson/stock-price-history

[2] Ljung, G. M. and Box, G. E. P. (1978), On a measure of lack of fit in time series models. Biometrika, 65, 297–303. doi:10.2307/2335207.

[3] Patrick Royston (1995). Remark AS R94: A remark on Algorithm AS 181: The W test for normality. Applied Statistics, 44, 547–551. doi:10.2307/2986146.

[4] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin (1992): Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. Journal of Econometrics 54, 159–178.

[5] Тест Dickey Fuller сформульовано у матеріалах лекцій 10-11.