STA457: Data Analysis Report

1. Data preview

For the dataset, I computed r_t for each of the 310 months and stored the series as a monthly time series indexed from 2000-01-01 to 2025-10-01 with frequency "month".

Basic statistics for the monthly excess return series are:

- Sample size: 310 months
- Mean: 0.00533 (0.533% per month)
- Standard deviation: 0.0431 (4.31% per month)
- Minimum: -0.1435 (-14.35%)
- Maximum: 0.1469 (14.69%)

Over the 310 months, 187 months indicate positive excess returns while 123 months indicate negative ones. As a result, the distribution is slightly left skewed, i.e. it is not exactly normal. The returns are more heavy-tailed than a Gaussian distribution.

Also, I observed that the series fluctuates around a small positive mean with several periods of unusually high volatility. There are no obvious long-run trends or structural breaks in the mean level.

2. Stationarity and autocorrelation analysis

I split the sample into three sub-periods and computed the sample mean of the monthly excess return r_t in each:

Period	$\boldsymbol{n}$	$\boldsymbol{E(r_t)}$
2000-2007	96	0.0082
2008-2015	96	0.0095
2016-2025	118	-0.0004

For the same three sub-periods, the sample variance of r_t is:

Period	**Variance**
2000-2007	0.0024
2008-2015	0.0011
2016-2025	0.0020

Let $r(h) = Cov(r_t, r_{t+h})$. The autocorrelation function $\rho(h) = \frac{r_h}{\sigma^2}$.

Hence, the sample autocorrelations for lags 1-12 are:

$$\rho(1, \cdots, 12)$$
$$= (-0.033, -0.123, 0.050, 0.049, 0.003, -0.086, -0.078, 0.065, -0.085, -0.045, 0.021, 0.049).$$

Using the definition of weak stationarity, I examined the monthly excess returns' sample mean, variance and covariance. The mean and variance are fairly stable across sub-periods, and the estimated autocorrelation function depends only on lag and is small for all lags. This provides mathematical evidence that the excess-return series can reasonably be treated as stationary.

In addition, I inspected the sample and partial autocorrelation function (ACF and PACF) of r_t for 24 lags. The results show trivial autocorrelations, incidentally, only a few individual lags are marginally outside the 95% interval. I also applied the Ljung–Box test for overall autocorrelation up to lag 12. The p-values are 0.18, so I fail to reject the null hypothesis that the first 12 autocorrelations are zero. However, the excess returns resemble white noise regardless of such small pattern inconsistency: they are centered around a constant mean with little serial dependence.

3. Model construction

I model the data series using the ARIMA because it is a standard and flexible modes for time-series forecasting. ARIMA models express the current observation as a linear function of its own past values and past shocks, plus a random innovation.

Model comparison was based on the Akaike Information Criterion (AIC). The AIC values I obtained are:

- ARIMA(0,0,0): AIC = -1066.26
- ARIMA(1,0,0): AIC = -1064.59
- ARIMA(0,0,1): AIC = -1064.70
- ARIMA(1,0,1): AIC = -1063.76
- ARIMA(2,0,0): AIC = -1067.46
- ARIMA(0,0,2): AIC = -1066.92

Since the ARIMA(2,0,0) has the smallest AIC, the estimated AR(2) model is $r_t = \mu + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \varepsilon_t$, with estimates:

- Intercept: $\hat{\mu}_0 = 0.00623$
- $\widehat{\phi_1} = -0.0377$
- $\widehat{\phi_2} = -0.1252$

Thus, the implied AR(2) process is $\hat{\mu} = \frac{\hat{\mu}_0}{1-\widehat{\phi_1}-\widehat{\phi_2}} \approx 0.00536$, which is almost equal to the sample mean. It is clear that AR(2) model captures essentially all linear dependence. Finally, I chose ARIMA(0,0,0) with constant, i.e. a white noise process as the forecasting model.

4. *Forecasts for November and December 2025*

Now the model is defined as $r_t = \mu + \varepsilon_t, \varepsilon_t \sim white\ noise\ (0, \sigma^2)$, where μ is the constant mean monthly return and ε_t is the uncorrelated innovation. Note that the model is stationary with no autoregressive terms, its optimal forecast is simply the estimated mean $\hat{\mu}$.

Estimating this model on the dataset gives:

- $\hat{\mu} = 0.00533\ (0.533\%\ per\ month)$
- $\hat{\sigma}^2 = 0.00186 \rightarrow \hat{\sigma} = 0.0431\ (4.31\%\ per\ month)$

Using the ARIMA(0,0,0) model and by R code, I computed the forecasts for November and December 2025, with 95% confidence intervals.

The forecasted monthly excess returns are:

- November 2025: $\hat{r}_{11} = 0.00533\ (0.53\%)$
- December 2025: $\hat{r}_{12} = 0.00533\ (0.53\%)$

The corresponding confidence intervals are $\hat{r}_t \pm 1.96 \cdot \mathrm{SE}(\hat{r}_t)$, which give:

- November 2025: $[-0.0791, 0.0897]$
- December 2025: $[-0.0791, 0.0897]$

The wide confidence intervals represent high instability of monthly returns in the stock market. Moreover, the best forecast is a relatively small positive excess return, but outcomes could be extremely negative or positive depending on the market.

5. Conclusion

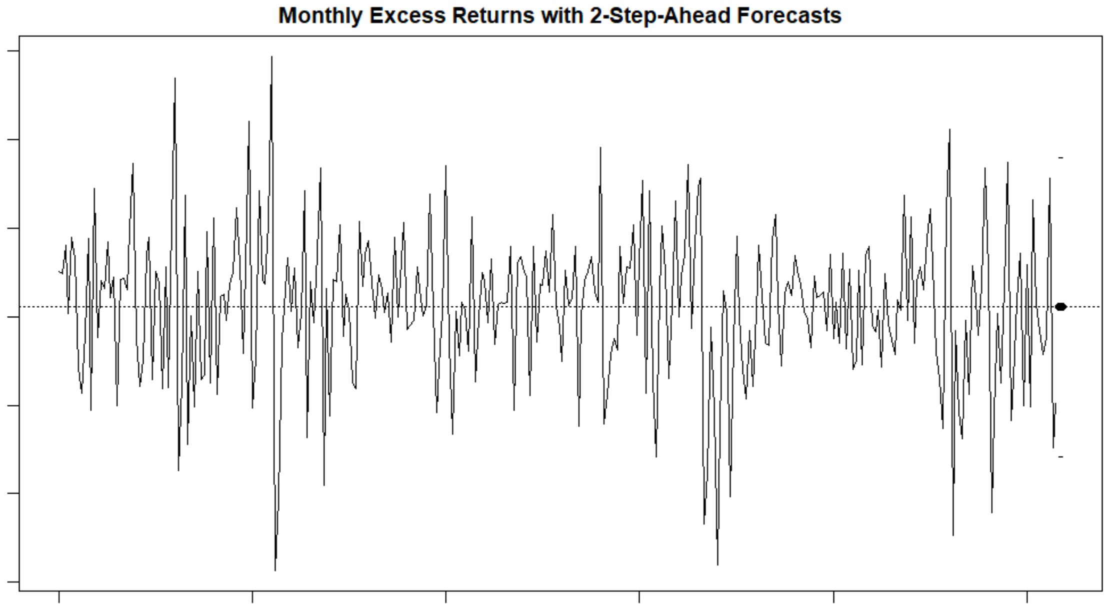
To summarize, the monthly excess returns over 2000–2025 behave very close to a white-noise process with a small positive mean per month. Formal tests and ACF/PACF show very weak serial dependence. For this reason, the most appropriate time-series model for forecasting the next two months is an ARIMA(0,0,0) with constant. Speaking of the results, the analysis demonstrates that the returns in November and December 2025 are both 0.53%, lying in the confidence interval $[-0.0791, 0.0897]$.

6. R code reference

```r
## STA457 – Dow Jones Monthly Excess Returns
## R code reference

library(zoo)
library(forecast)
library(moments)


## 1. Read and prepare data ---------------------------------------------

## Path
data_path <- "Dow Jones Industrial Average Historical Data.csv"

dj_raw <- read.csv("C:/Users/chris/Downloads/Dow Jones Industrial Average Historical Data.csv", stringsAsFactors = FALSE)

## Convert Date
dj_raw$Date <- as.Date(dj_raw$Date, format = "%b %d, %Y")

## Sort chronologically
dj_raw <- dj_raw[order(dj_raw$Date), ]

## Make numeric price/open
dj_raw$Price_num <- as.numeric(gsub(",", "", dj_raw$Price))
dj_raw$Open_num  <- as.numeric(gsub(",", "", dj_raw$Open))

## 2. Compute monthly excess returns and time series ---------------------

## Excess return r_t = (Price - Open) / Open
dj_raw$excess_ret <- (dj_raw$Price_num - dj_raw$Open_num) / dj_raw$Open_num

## Check sample size
length(dj_raw$excess_ret)

## Create monthly time series from 2000-01
ret_ts <- ts(dj_raw$excess_ret,
             start     = c(2000, 1),
             frequency = 12)

## 3. Descriptive statistics & basic plots -------------------------------

n_obs    <- length(ret_ts)
mean_ret <- mean(ret_ts)
sd_ret   <- sd(ret_ts)
min_ret  <- min(ret_ts)
max_ret  <- max(ret_ts)

## Counts of positive vs negative months
pos_months <- sum(ret_ts > 0)
neg_months <- sum(ret_ts <= 0)

## Shape of distribution
skew_ret  <- skewness(ret_ts)
kurt_ret  <- kurtosis(ret_ts)
```

```r

## Time series plot
plot(ret_ts,
     main = "Dow Jones Monthly Excess Returns (2000-2025)",
     xlab = "Year", ylab = "Monthly excess return")

## Histogram
hist(ret_ts, breaks = 20,
     main = "Histogram of Monthly Excess Returns",
     xlab = "Monthly excess return")


## 4. Stationarity: sub-period means/variances & rolling stats ------------

## Sub-periods: 2000-2007 (96), 2008-2015 (96), 2016-2025 (118)
r1 <- window(ret_ts, end          = c(2007, 12))
r2 <- window(ret_ts, start = c(2008, 1), end = c(2015, 12))
r3 <- window(ret_ts, start = c(2016, 1))

subperiod_stats <- data.frame(
  Period = c("2000-2007", "2008-2015", "2016-2025"),
  n      = c(length(r1), length(r2), length(r3)),
  mean   = c(mean(r1),   mean(r2),   mean(r3)),
  var    = c(var(r1),    var(r2),    var(r3)),
  sd     = c(sd(r1),     sd(r2),     sd(r3))
)

subperiod_stats

## 60-month rolling mean & variance
roll_window <- 60

roll_mean <- rollapply(ret_ts, width = roll_window,
                       FUN = mean, align = "right", fill = NA)
roll_var  <- rollapply(ret_ts, width = roll_window,
                       FUN = var,  align = "right", fill = NA)

rm_ts <- ts(roll_mean, start = start(ret_ts), frequency = 12)
rv_ts <- ts(roll_var,  start = start(ret_ts), frequency = 12)

plot(rm_ts,
     main = "Rolling 60-Month Mean of Monthly Excess Returns",
     xlab = "Year", ylab = "Rolling mean")

plot(rv_ts,
     main = "Rolling 60-Month Variance of Monthly Excess Returns",
     xlab = "Year", ylab = "Rolling variance")


## 5. ACF/PACF and Ljung-Box on returns ----------------------------------

## ACF and PACF of r_t
Acf(ret_ts, lag.max = 24,
```

```r
    main = "ACF of Monthly Excess Returns")

Pacf(ret_ts, lag.max = 24,
     main = "PACF of Monthly Excess Returns")

## Sample autocorrelations up to lag 12
acf_ret <- acf(ret_ts, lag.max = 12, plot = FALSE)
rho_hat <- as.numeric(acf_ret$acf)  # includes lag 0 at position 1
round(rho_hat, 3)

## Ljung-Box test on returns
lb_ret <- Box.test(ret_ts, lag = 12, type = "Ljung-Box")
lb_ret


## 6. ARMA/ARIMA model comparison -----------------------------------------

## Candidate models: ARMA(1,0), (0,1), (1,1), (2,0), (0,2) and white noise

fit_000 <- arima(ret_ts, order = c(0, 0, 0),
                 include.mean = TRUE, method = "ML")   # white noise with mean
fit_100 <- arima(ret_ts, order = c(1, 0, 0),
                 include.mean = TRUE, method = "ML")
fit_010 <- arima(ret_ts, order = c(0, 0, 1),
                 include.mean = TRUE, method = "ML")
fit_110 <- arima(ret_ts, order = c(1, 0, 1),
                 include.mean = TRUE, method = "ML")
fit_200 <- arima(ret_ts, order = c(2, 0, 0),
                 include.mean = TRUE, method = "ML")
fit_020 <- arima(ret_ts, order = c(0, 0, 2),
                 include.mean = TRUE, method = "ML")

aic_table <- data.frame(
  Model = c("ARIMA(0,0,0)", "ARIMA(1,0,0)", "ARIMA(0,0,1)",
            "ARIMA(1,0,1)", "ARIMA(2,0,0)", "ARIMA(0,0,2)"),
  AIC   = c(AIC(fit_000), AIC(fit_100), AIC(fit_010),
            AIC(fit_110), AIC(fit_200), AIC(fit_020))
)

aic_table

## AR(2) model details
summary(fit_200)

phi1   <- coef(fit_200)["ar1"]
phi2   <- coef(fit_200)["ar2"]
mu0    <- coef(fit_200)["intercept"]
sigma2 <- fit_200$sigma2

## Unconditional mean of AR(2): mu / (1 - phi1 - phi2)
mu_hat <- as.numeric(mu0 / (1 - phi1 - phi2))
mu_hat
```



```r
## Residual diagnostics for AR(2)
tsdisplay(residuals(fit_200),
          main = "Residuals from AR(2) Fit")

lb_res_ar2 <- Box.test(residuals(fit_200), lag = 12, type = "Ljung-Box")
lb_res_ar2


## 7. Final white-noise-with-mean model and forecasts ---------------------


final_model <- fit_000

## Residual diagnostics for final model
tsdisplay(residuals(final_model),
          main = "Residuals from ARIMA(0,0,0) with Mean")

lb_res_final <- Box.test(residuals(final_model), lag = 12, type = "Ljung-Box")
lb_res_final

## 2-step-ahead forecasts (Nov and Dec 2025)
fc_final <- predict(final_model, n.ahead = 2)

point_forecast <- as.numeric(fc_final$pred)
se_forecast    <- as.numeric(fc_final$se)

lower95 <- point_forecast - 1.96 * se_forecast
upper95 <- point_forecast + 1.96 * se_forecast

forecast_results <- data.frame(
  Horizon  = c("2025-11", "2025-12"),
  Forecast = point_forecast,
  SE       = se_forecast,
  Lower95  = lower95,
  Upper95  = upper95
)

forecast_results

## Plot: historical series + 2-step-ahead forecasts ----------------------

last_time <- time(ret_ts)[length(ret_ts)]
fc_times  <- seq(from = last_time + 1/12, by = 1/12, length.out = 2)

plot(ret_ts,
     xlim = c(start(ret_ts)[1], fc_times[2]),
     main = "Monthly Excess Returns with 2-Step-Ahead Forecasts",
     xlab = "Year", ylab = "Monthly excess return")

lines(fc_times, point_forecast)
lines(fc_times, lower95, lty = 2)
lines(fc_times, upper95, lty = 2)
points(fc_times, point_forecast, pch = 19)

abline(h = mean_ret, lty = 3)

```