STAT 3005 – Nonparametric Statistics

ASSIGNMENT 3

Due: 23 October (Wednesday) at 1800

Fall 2024

Exercise 3.1 (★★☆ — Strength of correlation (10%)). Let X, Y be two continuous RVs. Prove that it is NOT possible to have

$$\rho_S \neq 0 \qquad \text{and} \qquad \rho_C = 0.$$

Hints: See Remark 3.1. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

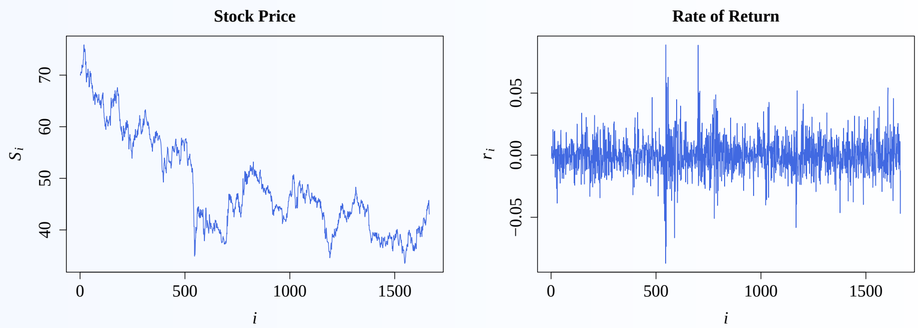
Exercise 3.2 (★☆☆ — Volatility clustering (90%)). Volatility clustering is a commonly observed phenomenon in stock data. It means that "large changes tend to be followed by large changes", so the stock returns are usually serially dependent, i.e., current return depends on the past returns. The aim of this exercise is to use nonparametric correlation tests to detect the existence of such serial dependence.

1. Data processing and data visualization.
   (a) (10%) Download the adjusted closing prices of CK Hutchison Holdings Limited (`0001.HK`) during all trading days between 2 Jan 2018 and 10 Oct 2024 ($n = 1665$). Denote them by $S_1, \ldots, S_n$. Compute the rates of return

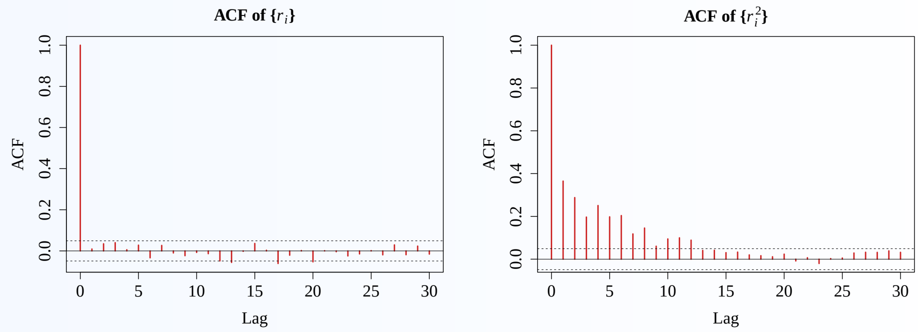
   $$r_i = \frac{S_i}{S_{i-1}} - 1, \qquad i = 2, \ldots, n.$$

   Does tie exist in $r_2, \ldots, r_n$?

   (b) (10%) Reproduce the following trace plots of S_i's and r_i's.



   (c) (10%) Autocorrelation of Z_i's means the correlation of Z_i and Z_j. E.g., $\rho_k = \mathsf{Corr}(Z_i, Z_{i-k})$ is the autocorrelation of Z_i's at time lag k. The `R` function `acf` plots values of $\rho_0, \rho_1, \rho_2, \ldots$. Use `acf` to plot the autocorrelation of r_i's and the autocorrelation of r_i^2's as follows.



2. Let

   $$(y_1, \ldots, y_{n-2}) = (r_2, \ldots, r_{n-1}) \qquad \text{and} \qquad (x_1, \ldots, x_{n-2}) = (r_3, \ldots, r_n).$$

   This part aims to test whether x_i's and y_i's are independent. *Note: You may use the asymptotic distributions of $\widehat{\rho}_P$, $\widehat{\rho}_S$, $\widehat{\rho}_K$, $\widehat{\rho}_{BD}$ and $\widehat{\rho}_C$ upon assuming the data are continuous; see the summary table of Chapter 4.*

   (a) (10%) Estimate ρ_P, ρ_S, ρ_K, ρ_{BD} and ρ_C between x_i's and y_i's.
   (b) (10%) Perform Pearson's, Spearman's, Kendall's, Bergsma–Dassios', and Chatterjee's correlation tests.
   (c) (10%) Report the computation time for each of the testing procedures in part (b).

	ρ_{P}	ρ_{S}	ρ_{K}	ρ_{BD}	ρ_{C}
(a) Estimate of correlation					
(b) p-value for testing independence					
(c) Time taken for testing					

3. (10%) Compare and interpret the results. Which test do you prefer? Why? (Use $\lesssim 50$ words.)
4. The generalized autoregressive conditional heteroskedasticity (GARCH(1, 1)) model is one of the standard ways to statistically model volatility clustering. It is defined as follows:

$$r_i = \sigma_i \varepsilon_{i-1},$$
$$\sigma_i^2 = \omega + \alpha r_{i-1}^2 + \beta\sigma_{i-1}^2,$$

where $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, and ε_i's are IID $\mathrm{N}(0, 1)$. Denote the maximum likelihood estimators of (ω, α, β) by $(\widehat{\omega}, \widehat{\alpha}, \widehat{\beta})$. Then the residuals of the fitted model can be computed as $\widehat{\varepsilon}_i := r_i/\widehat{\sigma}_i$, where $\widehat{\sigma}_i = \widehat{\omega} + \widehat{\alpha} r_{i-1}^2 + \widehat{\beta}\sigma_{i-1}^2$. We are going to check whether the residuals $\widehat{\varepsilon}_i$'s of the fitted model are sufficiently uncorrelated serially. If true, it means the GARCH model does its job. The following R code compute the residuals of the fitted GARCH(1, 1) model.

```
install.packages("tseries")    # install the package if you haven't
residuals = tseries::garch(r, order=c(1,1), trace=0)$residuals[-1] # r = rate of return
```

(a) (10%) Compute the residuals $\widehat{\varepsilon}_i$'s and produce a trace plot of $\widehat{\varepsilon}_i$'s
(b) (10%) Test whether $\widehat{\varepsilon}_i$'s are correlated at lag one by your preferred test. What is your conclusion?

Hints: See Remark 3.2. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 3.3 (★★★ — Property of different correlations (✄)). Let A be a continuous RV. Let f and g be two increasing functions. Define $X = f(A)$ and $Y = g(A)$.

1. (✄) Prove that the Pearson correlation between X and Y is $\rho_{\mathrm{P}} \geq 0$.
2. (✄) Prove that the Spearman correlation between X and Y is $\rho_{\mathrm{S}} = 1$.
3. (✄) Prove that the Kendall correlation between X and Y is $\rho_{\mathrm{K}} = 1$.
4. (✄) Prove that the Chatterjee's correlation between X and Y is $\rho_{\mathrm{C}} = 1$.
5. (✄) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be IID copies of (X, Y). Set $A \sim \mathrm{Exp}(1)$, $f(a) = a^3$, and $g(a) = \exp(a)$. Compute $\widehat{\rho}_{\mathrm{P}}$, $\widehat{\rho}_{\mathrm{S}}$, $\widehat{\rho}_{\mathrm{K}}$ and $\widehat{\rho}_{\mathrm{C}}$ for one realized dataset when $n = 5, 50, 500$. Comment. (Use $\lesssim 50$ words.)

	$\widehat{\rho}_{\mathrm{P}}$	$\widehat{\rho}_{\mathrm{S}}$	$\widehat{\rho}_{\mathrm{K}}$	$\widehat{\rho}_{\mathrm{C}}$
$n = 5$				
$n = 50$				
$n = 500$				

Hints: See Remark 3.3. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Remark 3.1 (Hints for Exercise 3.1). Note that

- ρ_{C} is a strong correlation, and
- some properties of ρ_{S} when $X \perp\!\!\!\perp Y$.

Remark 3.2 (Hints for Exercise 3.2). The goal of Exercise 3.2 is to train you to perform statistical analysis on problems that we may not be familiar with. After completing this exercise, you should be able to filter out unimportant materials and draw insightful conclusions that are useful to other experts.

1. You do not need to follow the graphical settings (e.g., colors and font size) of the displayed figures. You may follow the following structure to write your answers.

```
# Step 1: Download the stock prices
```

```r
#-----------------------------------------------------------------------------------
# install.packages("tseries")   # install the package if necessary
stock = c("0001.hk")
from  = as.Date("2018-01-02")
to    = as.Date("2024-10-10")
data  = tseries::get.hist.quote(stock,from,to,quote="Adjusted",quiet=TRUE)
# Step 2: Compute the rate of return
#-----------------------------------------------------------------------------------
s = c(as.matrix(data))
s =                             # <<<< Remove NA on non-trading days. Try na.omit
n = length(s)
r = ...                         # <<<< Complete this line
# Step 3: Check if tie exists
#-----------------------------------------------------------------------------------
sum(duplicated(...))            # <<<< Complete this line
# Step 4: Plot the prices and returns
#-----------------------------------------------------------------------------------
par(mfrow=c(1,2))
ts.plot(s, main="Stock Price", xlab=expression(i),  ylab=expression(S[i]))
ts.plot(...)             # <<<< Complete this line
# Step 5: Produce autocorrelation plot
#-----------------------------------------------------------------------------------
par(mfrow=c(1,2))
acf(r, main=expression("ACF of {"*r[i]*"}"))
acf(...)                 # <<<< Complete this line
```

2. Example 4.15 is similar to this exercise. The correlation/covariance tests can be found in the file `chp4.R`. It may take several minutes to compute some of the p-values. Be patient. You may follow the following structure to write your answers.

```r
# Step 1: define x and y
#---------------------------------------------------------------
x = r[-1]
y = ...                          # <<<< Complete this line
# Step 2: define space for storing the result table
#---------------------------------------------------------------
result = array(NA, dim=c(3,5))
rownames(result) = c("Estimate", "p-value", "time")
colnames(result) = c("P","S","K","BD","C")
# Step 3: (a) estimate of correlation
#---------------------------------------------------------------
for(i in 1:3){
    result[1,i] = cor(x,y,method=c("p","s","k")[i])
}
result[1,4] =                       # <<<< Compute BD correlation (NOT covariance); see chp4.R
result[1,5] =                       # <<<< Compute Chatterjee's correlation; see chp4.R
# Step 4: (b-c) p-value and time taken
#---------------------------------------------------------------
for(i in 1:5){
    t0=Sys.time()
    result[2,i] = cor.test0(...)$p.value     # <<<<  cor.test0 is a function in chp4.R.
    t1=Sys.time()
    result[3,i] = difftime(..., ..., units="secs")    # <<<< Complete this line
}
result
```

3. What's your assumption? What do you want to detect? Do you care about computational time?
4. It is nearly the same as part 3 except that you need to replace the data with ...

Takeaway: Nonparametric test can be used to assist parametric procedures.

Remark 3.3 (Hints for Exercise 3.3)**.** You may follow the following steps.

1. Let A_1, A_2, A_3 be IID copies of A.
   - Prove that $\mathsf{Cov}\{f(A), g(A)\} = \mathsf{E}[f(A_1)\{g(A_1) - g(A_2)\}]$.
   - Argue that $\mathsf{Cov}\{f(A), g(A)\} = \mathsf{E}[-f(A_2)\{g(A_1) - g(A_2)\}]$.

- Argue that

$$\mathsf{Cov}\{f(A), g(A)\} = \frac{1}{2}\mathsf{E}\Big[\{f(A_1) - f(A_2)\}\{g(A_1) - g(A_2)\}\Big].$$

- Note that $A_1 > A_2$ implies $f(A_1) - f(A_2) \geq 0$ and $g(A_1) - g(A_2) \geq 0$.

2. Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$ be IID copies of (X, Y). So, we can write $X_i = f(A_i)$ and $Y_i = g(A_i)$ for $i = 1, 2, 3$, where A_1, A_2, A_3 be IID copies of A. Note the following facts:
   - ρ_{S} can be defined according to equation (4.5) of the lecture note.
   - $X_1 > X_2$ is equivalent to $A_1 > A_2$.
   - $\mathsf{P}\{(A_1 > A_2), (A_1 > A_3)\} = \mathsf{E}[\mathsf{P}\{(A_1 > A_2), (A_1 > A_3) \mid A_1\}]$.
   - $\mathsf{P}(A_1 > A_2 \mid A_1) = \mathsf{P}(A_1 > A_3 \mid A_1)$.
   - $F(A_1) \sim \mathrm{Unif}(0, 1)$, where $F(\cdot)$ is the CDF of A.
   - If $U \sim \mathrm{Unif}(0, 1)$, then $\mathsf{E}(U^2) = 1/3$.
3. ρ_{K} can be defined according to Proposition 4.4. You may use the following fact: $\mathsf{P}(A_1 > A_2) = 1/2$.
4. See Definition 4.8 of the lecture note. The following remarks are helpful.
   - Note that knowing the value of X is equivalent to knowing the value of Y.
   - You may prove that

   $$\mathsf{E}\{\mathbb{1}(Y \geq y) \mid X\} = \cdots = \mathbb{1}(A \geq a) \qquad \text{and} \qquad \mathsf{E}\{\mathbb{1}(Y \geq y)\} = \cdots = 1 - F(a),$$

   where $a = g^{-1}(y)$ and $F(\cdot)$ is the CDF of A.
   - $\mathbb{1}(A \geq a) \sim \mathrm{Bern}(1 - F(a))$.
   - If $I \sim \mathrm{Bern}(p)$, then $\mathsf{Var}(I) = p(1 - p)$.
5. You may also perform the experiment several times. Do you obtain exactly the same results?

INSTRUCTIONS: Please follow the instructions stated on the last page of assignment 1.