MA 541

The Course Project Description

Part 1: Meet the data

Data description – This data includes four columns/random variables: the daily ETF return; the daily relative change in the price of the crude oil; the daily relative change in the gold price; and the daily return of the JPMorgan Chase & Co stock. The sample size is 1000.

Requirements – Use any software to obtain the sample mean and sample standard deviation for each random variable (column) of the data; the sample correlations among each pair of the four random variables (columns) of the data.

Part 2: Describe your data

Requirements – Use any software to draw the following plots:

1) A histogram for each column (**hint**: four histograms total)
2) A time series plot for each column (**hint**: use the series "1, 2, 3, …, 1000" as the horizontal axis; four plots total)
3) A time series plot for all four columns (**hint**: one plot including four "curves" and each "curve" describes one column)
4) Three scatter plots to describe the relationships between the ETF column and the OIL column; between the ETF column and the GOLD column; between the ETF column and the JPM column, respectively

Part 3: What distribution does your data follow

Requirements – Propose an assumption/a hypothesis regarding the type of distribution each column of the data set may follow (i.e., the ETF, OIL, GOLD, and JPM column), based on the plots from **Part 2**. Then verify or object that assumption/hypothesis with appropriate tests (for example, normality test). You may use any software to perform those tests.

Part 4: Break your data into small groups and let them discuss the importance of the Central Limit Theorem

Requirements – Consider the ETF column (1000 values) as the population (x), and do the follows. Any software may be used.

1) Calculate the mean μ_x and the standard deviation σ_x of the population.
2) Break the population into 50 groups **sequentially** and each group includes 20 values.

3) Calculate the sample mean ($\bar{x}$) of each group. Draw a histogram of all the sample means. Comment on the distribution of these sample means, i.e., use the histogram to assess the normality of the data consisting of these sample means.
4) Calculate the mean ($\mu_{\bar{x}}$) and the standard deviation ($\sigma_{\bar{x}}$) of the data including these sample means. Make a comparison between μ_x and $\mu_{\bar{x}}$, between $\frac{\sigma_x}{\sqrt{n}}$ and $\sigma_{\bar{x}}$. Here, n is the number of sample means calculated from Item 3) above.
5) Are the results from Items 3) and 4) consistent with the Central Limit Theorem? Why?
6) Break the population into 10 groups **sequentially** and each group includes 100 values.
7) Repeat Items 3) ~ 5).
8) Generate 50 simple **random** samples or groups (**with replacement**) from the population. The size of each sample is 20, i.e., each group includes 20 values.
9) Repeat Items 3) ~ 5).
10) Generate 10 simple **random** samples or groups (**with replacement**) from the population. The size of each sample is 100, i.e., each group includes 100 values.
11) Repeat Items 3) ~ 5).
12) In **Part 3** of the project, you have figured out the distribution of the population (the entire ETF column). Does this information have any impact on the distribution of the sample mean(s)? Explain your answer.

Part 5: Construct a confidence interval with your data

Requirements –

1) Pick up one of the 10 simple random samples you generated in Step 10) of **Part 4**, construct an appropriate 95% confidence interval of the mean μ.
2) Pick up one of the 50 simple random samples you generated in Step 8) of **Part 4,** construct an appropriate 95% confidence interval of the mean μ.
3) In **Part 1**, you have calculated the mean μ of the population (the entire ETF column) using Excel function. Do the two intervals from 1) and 2) above include (the true value of) the mean μ? Which one is more accurate? Why?

Part 6: Form a hypothesis and test it with your data

Requirements –

1) Use the same sample you picked up in **Step 1) of Part 5** to test $\boldsymbol{H_0}$: $\boldsymbol{\mu = 100}$ ***vs.*** $\boldsymbol{H_a}$: $\boldsymbol{\mu \neq 100}$ at the significance level 0.05. What's your conclusion?
2) Use the same sample you picked up in **Step 2) of Part 5 to** test $\boldsymbol{H_0}$: $\boldsymbol{\mu = 100}$ ***vs.*** $\boldsymbol{H_a}$: $\boldsymbol{\mu \neq 100}$ at the significance level 0.05. What's your conclusion?
3) Use the same sample you picked up in **Step 2) of Part 5 to** test $\boldsymbol{H_0}$: $\boldsymbol{\sigma = 15}$ ***vs.*** $\boldsymbol{H_a}$: $\boldsymbol{\sigma \neq 15}$ at the significance level 0.05. What's your conclusion?
4) Use the same sample you picked up in **Step 2) of Part 5 to** test $\boldsymbol{H_0}$: $\boldsymbol{\sigma = 15}$ ***vs.*** $\boldsymbol{H_a}$: $\boldsymbol{\sigma < 15}$ at the significance level 0.05. What's your conclusion?

Part 7: Compare your data with a different data set

Requirements –

1) Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.
2) Subtract the entire Gold column from the entire Oil column and generate a sample of differences. Consider this sample as a random sample from the target population of differences between Gold and Oil. Form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.
3) Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal standard deviations in the significance level 0.05.

Part 8: Fitting the line to the data

Requirements –

Consider the data including the ETT column and Gold column only. Using any software,

1) Draw a scatter plot of ETF (Y) vs. Gold (X). Is there any linear relationship between them which can be observed from the scatter plot?
2) Calculate the coefficient of correlation between ETF and Gold and interpret it.
3) Fit a regression line (or least squares line, best fitting line) to the scatter plot. What are the intercept and slope of this line? How to interpret them?
4) Conduct a two-tailed *t*-test with $H_0\colon\ \beta_1 = 0$. What is the P-value of the test? Is the linear relationship between ETF (Y) and Gold (X) significant at the significance level 0.01? Why or why not?
5) Suppose that you use the coefficient of determination to assess the quality of this fitting. Is it a good model? Why or why not?
6) What are the assumptions you made for this model fitting?
7) Given the daily relative change in the gold price is 0.005127. Calculate the 99% confidence interval of the mean daily ETF return, and the 99% prediction interval of the individual daily ETF return.

Part 9: Does your model predict?

Requirements –

Consider the data including the ETF, Gold and Oil column. Using any software, fit a multiple linear regression model to the data with the ETF variable as the response. Evaluate your model with adjusted R^2.

Part 10: Checking residuals and model selection

Requirements –

Calculate the residuals of the model fitting you did in **Part 9**. Check the four assumptions made for the error terms of the multiple regression model using these residuals (mean 0; constant variance; normality; and the independence). You may draw some plots over the residuals to check these assumptions. For example, draw a Normal Probability Plot to check the normality assumption; draw a scatter plot of Residuals vs. Fitted Values to check the constant variance assumption and the independence assumption; and so on. You may refer to the following link https://www.youtube.com/watch?v=4zQkJw73U6I for some hints. In your project report, all the relevant plots and at least one paragraph of summary of checking the four assumptions using those plots must be included.

Discuss how you may improve the quality of your regression model according to the strategy of model selection.