The analysis of the log returns for a stock market

Elliot Lawrence Candidate Number: 710046334

Summative Coursework Report March 2023

1 Introduction

We are analysing log returns for a stock market and trying to find a good model for them and use that model to create some predictions for how likely certain events are to unfold in the future.

2 Modelling with a normal distribution

We aim to model the log returns data we have for the stock market using a normal distribution. In order to do that, we need to calculate the sample mean and sample variance for our data, and then use the method of moments along with those values to get the parameters for the normal distribution μ and σ^2. In order to do that, the equations below allow us to calculate the parameters using those sample statistics.

$$\bar{y} = E[X] = \mu \qquad \bar{v} = Var(x) = \sigma^2$$

By using the R code, we can calculate the E[X] and Var[X] to be approx. 4.3199×10^{-4} and 6.3325×10^{-5} respectively.

By plotting a Quantile-Quantile plot of the normal distribution against the data, we can see that the data matches the model in the centre, which is to be expected as we created our model using the mean and the standard deviation of the data. However, we can see at the tails the data and model start to deviate considerably. This suggests that even though the model and data match up in their mean and standard deviation as constructed. The model still does not fit the data well in terms of shape. See figure 1 and figure 2 below for graphs of the normal distribution and Quantile-Quantile plot.

3 Modelling with a student T distribution

We attempt another model for our data, this time we attempt to model our data using a student T. In order to create a student T distribution fit, we need three parameters α, β and γ. We use the method of moments to calculate those. In order to do that we need the mean, variance and kurtosis of our data. Using those quantities, we can calculate our parameters for the student T distribution via these three equations.

$$\bar{\mu} = \alpha \qquad \bar{\sigma}^2 = \frac{\beta^2\gamma}{\gamma - 2} \qquad \bar{k} = \frac{3\gamma - 6}{\gamma - 4}$$

In order to calculate the parameters, we need to solve each of these equations for α, β and γ respectfully. We have already solved for α in the first case. Next we solve for γ using the third equation; the value we get is

$$\gamma = \frac{4\bar{k} - 6}{\bar{k} - 3}$$

Now we can solve for β, using the quantity we found for γ in the second equation, yielding

$$\beta = \pm\sqrt{\frac{(\gamma - 2)\bar{\sigma}^2}{\gamma}}$$

We get two solutions for β, however we know in the student T distribution β needs to be positive, so we use the positive solution.

By using the R code, we can calculate these three quantities to be $\alpha = 4.3199 \times 10^{-4}$, $\beta = 6.5690 \times 10^{-3}$ and $\gamma = 6.2782$.

By plotting a Quantile-Quantile plot of the student T distribution against the data, we can see that the data matches the model in the centre and also matches the data at the tails, which is an improvement compared to the normal distribution. The tails matching the data suggest that the student T distribution is a better distribution to model the log returns compared to the normal distribution. See figure 3 and figure 4 below for graphs of the student T distribution and Quantile-Quantile plot.

The student T distribution would be a better model to use, as the the Quantile-Quantile plot shows that it matches the data considerably better compared to the normal distribution.

4 Using the models for qualitative analysis

Using our preferred model, the student T distribution, we aim to compute two predictions of the stock market prices, the value at risk and the probability that the price reduces by at least 3% of its value in one day.

Part 1: The Value at Risk

We know value at risk is defined as the value, v, such that (-v, ∞) is a 99% prediction interval in log return.

As such, we know that in order to get (-v, ∞) to be a 99% interval, the CDF value of -v is going to be 1% of the entire distribution. Therefore we know v has a CDF given by this command:

$$\text{v} = \text{-qstudent}(\frac{1}{100},\alpha,\beta,\gamma)$$

Using this command with the values corresponding to the student T distribution, we can calculate v, v = 1.9910×10^{-2}.

Part 2: Probability that the price X_t reduces by at least 3% of its value in one day.

In order to calculate this value, we note that the price decreasing by 3% in one day means that the price tomorrow will be at most 97% of the price today, which means that the return tomorrow divided by the return today will be at most 97%. Therefore, if we take a log of both sides, the left side is now a log return, and the right side is log(0.97).

$$\frac{X_{t+1}}{X_t} \leq 97\%$$

$$Y_t \leq \log(0.97)$$

Our goal is to compute log(0.97) and find the CDF value at that point. This value is 1.4715×10^{-3}.

Part 3: Trustworthiness

The trustworthiness of the model can be analysed in different areas, for example this is a model, so we don't know that these values are completely accurate. We can see it is a good model based on the Quantile-Quantile plot. We can have confidence that the model will match the true data reasonably well based on our analysis. However these values are in the tail; the value at risk and the probability are both at the tail. Generally, models become less reliable closer to the tails. These values do seem trustworthy based on the accuracy of the fit, but one should be wary of taking too much stock in them, simply because they are at the tails.

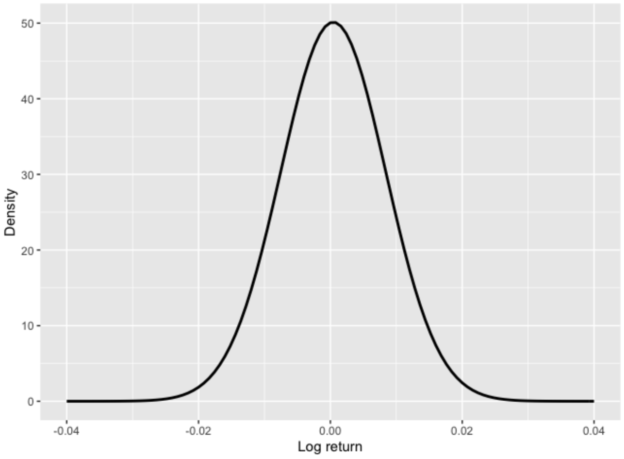

Figure 1: The pdf of our normal distribution model

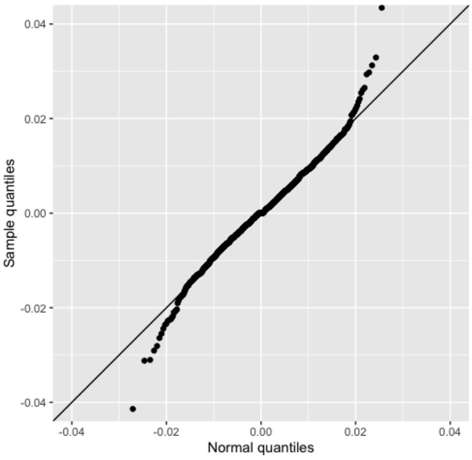

Figure 2: A Quantile-Quantile Plot showing the sample quantiles vs normal quantiles

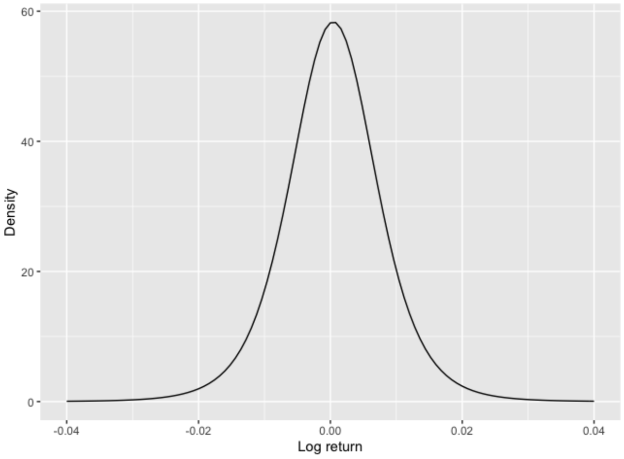

Figure 3: The pdf of our student T distribution model

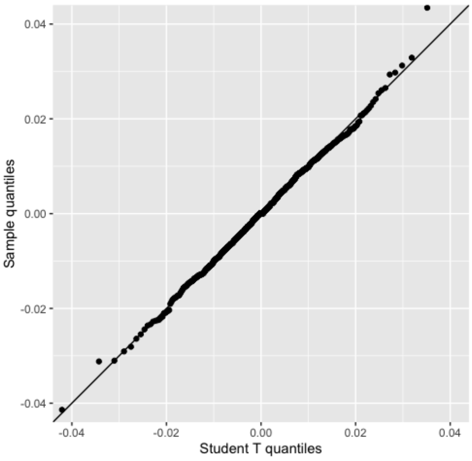

Figure 4: A Quantile-Quantile Plot showing the sample quantiles vs student T quantiles

Analysis of the effectiveness of two treatments for arthritis

Elliot Lawrence Candidate Number: 710046334

Summative Coursework Report March 2023

5 Strengths and weaknesses of arthritis study

This particular study for arthritis has a few strengths. For example the study undergoes simple random sampling. Therefore measurement bias is reduced. Also the study is measurable, as we get quantitative figures to work with, as opposed to qualitative questions, which are subjective and can yield apprehension bias, where participants may want to please the study group.

The study also had many flaws, for example, intervention group was twice the size of the control group. In addition to matching and Simpsons Paradox considerations, there is more uncertainty in smaller groups and thus having uneven group sizes doesn't spread the uncertainty evenly over both groups, which would be desirable. Also, three weeks many not be long enough to show results of the treatment working; six weeks might be more sensible. In addition, age groups are not taken into consideration. Younger adults might respond better to the treatment than older adults. It is important also to understand that arthritis isn't just flexibility but also pain. As such a patient having more flexibility could mean the treatment is working, but is pointless if it is still just as painful. The study also occurs in Devon, not the whole of the UK, therefore there could be generalisation bias. Moreover, the patients know whether they are getting the new treatment or not, so it is not blind.

6 Using point estimates and confidence intervals to assess the effects of the two treatments

We use point estimates and confidence intervals to assess the effectiveness of the intervention treatment against the control treatment.

To do so, we first calculate the differences in flexion for each subject. We then take the mean differences for the control and intervention group, which are respectively:

$$\mu_C = 3.375$$

$$\mu_I = 6$$

These mean differences are point estimates for the improvement of patients in each group after receiving the treatment. We can see using these point estimates μ_I is larger than μ_C, which on the surface of it seems to suggest that the intervention is more effective than the control. However, we are not too sure how reliable that conclusion is; to get a more conclusive verdict we look into a confidence intervals.

In order to compute confidence intervals, we compute the standard deviations for the control and intervention groups, which are respectively:

$$\sigma_C = 5.3156$$

$$\sigma_I = 6.4077$$

From that, we notice that even though the intervention group has twice as many patients compared to the control group, the standard deviations for the intervention and control groups were still fairly close, but the intervention group's was higher. In statistics, the default confidence interval approach is to build a confidence interval around μ_C and compare μ_I to that interval. This allows us to determine the likelihood that the mean of the intervention group could be atleast as large as μ_I in the case where the intervention is no better than the control (i.e. the null hypothesis)

We aim to find the smallest confidence interval centered at μ_C which contains μ_I. To do that, we compute the following:

$$\frac{\mu_C - \mu_I}{\sigma_C} \approx 0.4939$$

Therefore we want to compute a confidence interval based on these values. $(\mu_C - 0.4939\sigma_C, \mu_C + 0.4939\sigma_C)$ is a confidence interval containing μ_I. In order to decide what percentage confidence this is, we assume that our data is normally distributed, and we use a normal distribution table to determine what percentage confidence interval $(\mu_C - 0.4939\sigma_C, \mu_C + 0.4939\sigma_C)$ is.

This is a 37.8% confidence interval. Based on that, there is a 62.2% chance that the intervention mean would differ at least this much from the control group mean in the case of the null hypothesis.

We could also consider a single tail interval: $(-\infty, \mu_C + 0.4939\sigma_C)$. This is a 68.8% confidence interval, i.e. is a 31.2% chance that the intervention mean would be at least this large relative to the control group mean in the case of the null hypothesis.

Even 68.8% is not compelling enough to conclusively show that the treatment is better than the control.

Therefore, even though the treatment shows promise due to its higher point estimate, more study would be needed to have confidence that this improvement is statistically significant.

7 Appendix

```
library(tidyverse)

load('MTH1004T2CW.RData')

# Question 1

# PART A

# Calculating mean and variance of finance data

Logret <- finance$logret

mean(Logret)

var(Logret)

kurtosis(Logret)

# Plotting Normal Distribution for finance data

ggplot(finance) +
  stat_function(geom = ”line”, fun = dnorm, args = list(mean = mean(Logret),
    sd = sqrt(var(Logret))), lwd = 1) + lims(x = c(-0.04, 0.04)) +
      labs(x = ”Log return”, y = ”Density”)

# Using Q-Q Plots  to see if the Normal Distribution is a good estimate
# for the sample
ggplot(finance,aes(sample = Logret)) +
  stat_qq(distribution = qnorm, dparams = list(mean = mean(Logret),
    sd = sqrt(var(Logret)))) + geom_abline(intercept = 0, slope = 1) +
      coord_fixed(xlim = c(-0.04, 0.04), y = c(-0.04, 0.04)) +
        labs(x = ”Normal quantiles”, y = ”Sample quantiles”)

# PART  B

mu = mean(Logret)

gamma = ((4*kurtosis(Logret))-6)/(kurtosis(Logret) - 3)

beta = sqrt(((var(Logret)*(gamma-2))/gamma))

ggplot(finance) + stat_function(geom = ”line”, fun = dstudent,
                                 args = list(a = mu, b = beta, g = gamma)) +
  lims(x = c(-0.04, 0.04)) +
  labs(x = ”Log return”, y = ”Density”)

ggplot(finance,aes(sample = Logret)) +
  stat_qq(distribution = qstudent, dparams = list(a = mu, b = beta, g = gamma))+
       geom_abline(intercept = 0, slope = 1) +
          coord_fixed(xlim = c(-0.04, 0.04), y = c(-0.04, 0.04)) +
             labs(x = ”Student T quantiles”, y = ”Sample quantiles”)
```

```
# PART C

v = -qstudent(1/100,mu,beta,gamma)

pstudent(log(0.97),mu,beta,gamma)




# Question 2

# PART B


#Code below shows how to obtain difference, mean and standard deviation for
# control group

Control <- filter(arthritis, Group == "Control")

BeforeControl <- Control$Before

AfterControl <- Control$After

DifferenceControl <- AfterControl-BeforeControl

MeanControl <- mean(DifferenceControl)

VarianceDifferenceControl <- var(DifferenceControl)

StandardDeviationControl <- sqrt(VarianceDifferenceControl)

# Code below shows how to obtain difference, mean and standard deviation for
# intervention group.


Intervention <- filter(arthritis, Group == "Intervention")

BeforeIntervention <- Intervention$Before

AfterIntervention <- Intervention$After

DifferenceIntervention <- AfterIntervention-BeforeIntervention

MeanIntervention <- mean(DifferenceIntervention)

VarianceDifferenceIntervention <- var(DifferenceIntervention)

StandardDeviationIntervention <- sqrt(VarianceDifferenceIntervention)

# Code below shows how to smallest confidence interval centered at the mean of
# the control group containing mean of intervention group.

(MeanIntervention - MeanControl) / StandardDeviationControl
```