Stat 3303 Final Project

Jessica Li.12986

2025-04-24

Introduction

It is difficult to predict the performance of stocks in the volatile U.S. stock market. However, millions of people invest in stocks, and the total amount of stocks in the U.S. are valued at over 60 trillion dollars. As such, it is highly useful to be able to understand and predict if certain stocks increase in value or not and under what conditions. External factors such as inflation, trade policy, and disasters can all influence the stock market drastically. During disasters, there may be decreased demand for luxury items and increased demand items in the food or medicine sectors, so stocks tend to vary around a sector mean.

Data Description

The data are individual stocks modeled as coin flips, with a value of 1 indicating that the stock had a positive return and 0 for a negative return. There are 10 stocks, each measured over 30 periods, across 5 sectors for a total of 1500 observations. Stock returns are assumed to be independently and identically distributed from one period to the next, but stock movements as a whole and across sectors are assumed to be not independent. While these assumptions are helpful in simplifying model building, assuming that stock returns are independent and identically distributed across periods is a limitation of the model as well. In reality, changes in interest rates or policies would also affect individual stocks at specific points in time. From the box plots in Figure 1, we can see that the different sectors tend to vary around a proportion of approximately 50% positive returns, with sector 4 having a much lower proportion and sectors 2 and 5 having the highest proportion. Sectors 4 and 5 have the highest variance and the most outliers.

Exploratory Data Analysis

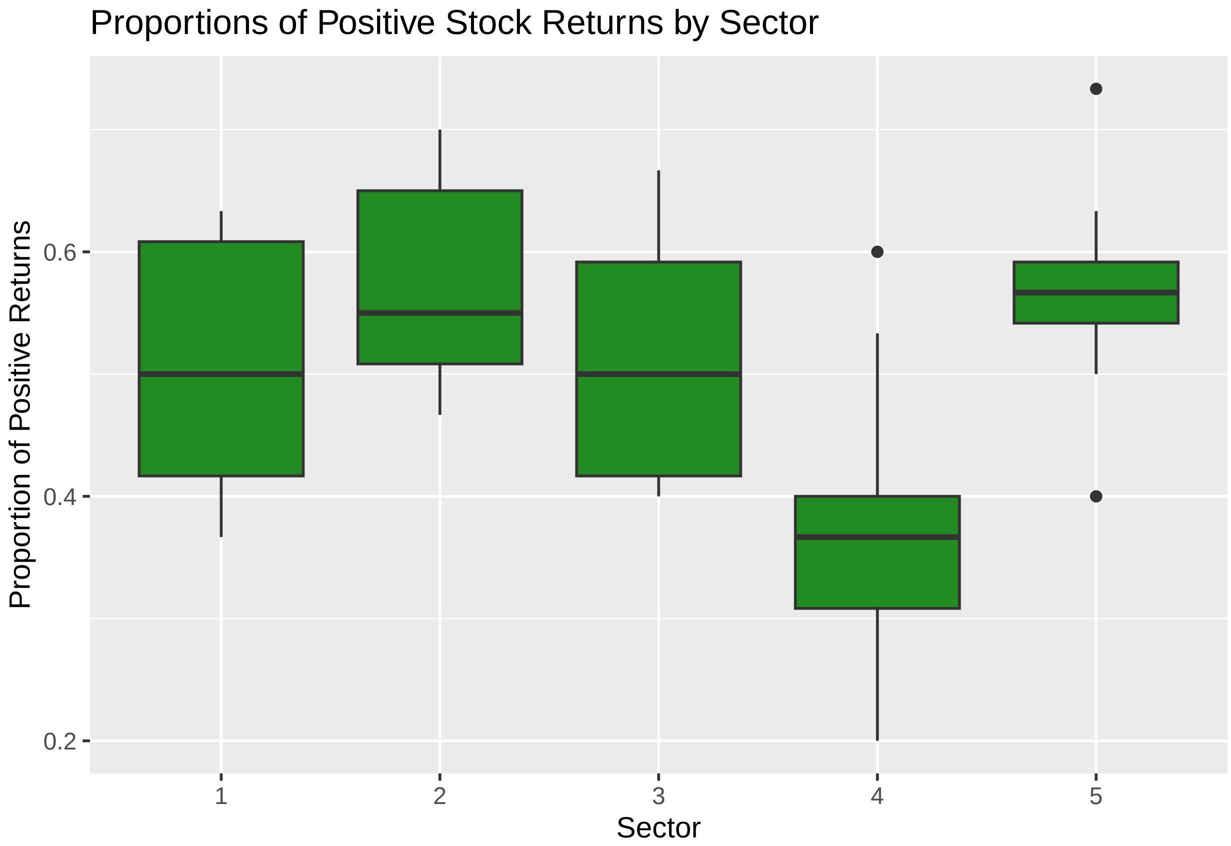

Figure 1: Box plots showing proportions of positive stock returns for each sector

Model Definition

We let y_{ij} be the number of positive returns for each stock j in sector i out of n_{ij} periods.

i = 1,2,3,4,5 and j = 1,2,. . . ,10. n_{ij} = 30.

We assume that each observation y_{ij} is independent across sector and stocks, conditional on the probability θ_{ij}. Since the stock returns are modeled as coin flips with a value of 0 or 1, we let y_{ij} be a realization from $Y_{ij}|\theta_{ij} \sim$ Binomial(n_{ij}, θ_{ij}) with θ_{ij} being the true probability that stock j in sector i goes up.

We can observe the log-odds of θ_{ij} as $logit(\theta_{ij}) = log(\frac{\theta_{ij}}{1-\theta_{ij}}) = \eta_{ij}$,

which follows a normal distribution $\eta_{ij}|\mu_i, \tau_i \sim Normal(\mu_i, \tau_i)$ for all sectors i.

Each sector mean μ_i follows a normal distribution, varying around the market mean and variance, $\mu_i \sim Normal(\mu_0, \tau_0)$. The market mean and variance and sector variance all follow weakly informative priors:

$\mu_0 \sim Normal(0,1)$, $\tau_0 \sim Gamma(0.01, 0.01)$, $\tau_i \sim Gamma(0.01, 0.01)$.

Model Fitting

We set the initial values for the market mean, μ_0, as coming from a randomized normal distribution with a mean of 0 and a variance of 1. We set the initial market variance, τ_0, coming from a randomized gamma distribution with the parameters being 0.01. The sector mean and variances drew from the same distributions as the market mean and variances, respectively. Sector means and variances were drawn for each sector 1 through 5. Using one chain, we ran the model for 60,000 iterations. 20,000 of these iterations were used as adaptive samples, and they stabilized the sampler. Half of the remaining samples were discarded as burn-in (20,000 samples). The remaining 20,000 iterations were then analyzed. Convergence was diagnosed using trace plots, as seen in Figure 2 for the θ parameter, where the trace plots show steady and random noise around a common center. The effective sample sizes of the posteriors were all high and well above 1000. The ECDF plots plots also demonstrate convergence to the same posterior distribution. The density plot showed a unimodal, approximately normal distribution. All of these plots can be used to safely assume convergence of the model.

Figure 2: Trace Plot, Autocorrelation Plot, ECDF Plot, Density Plot

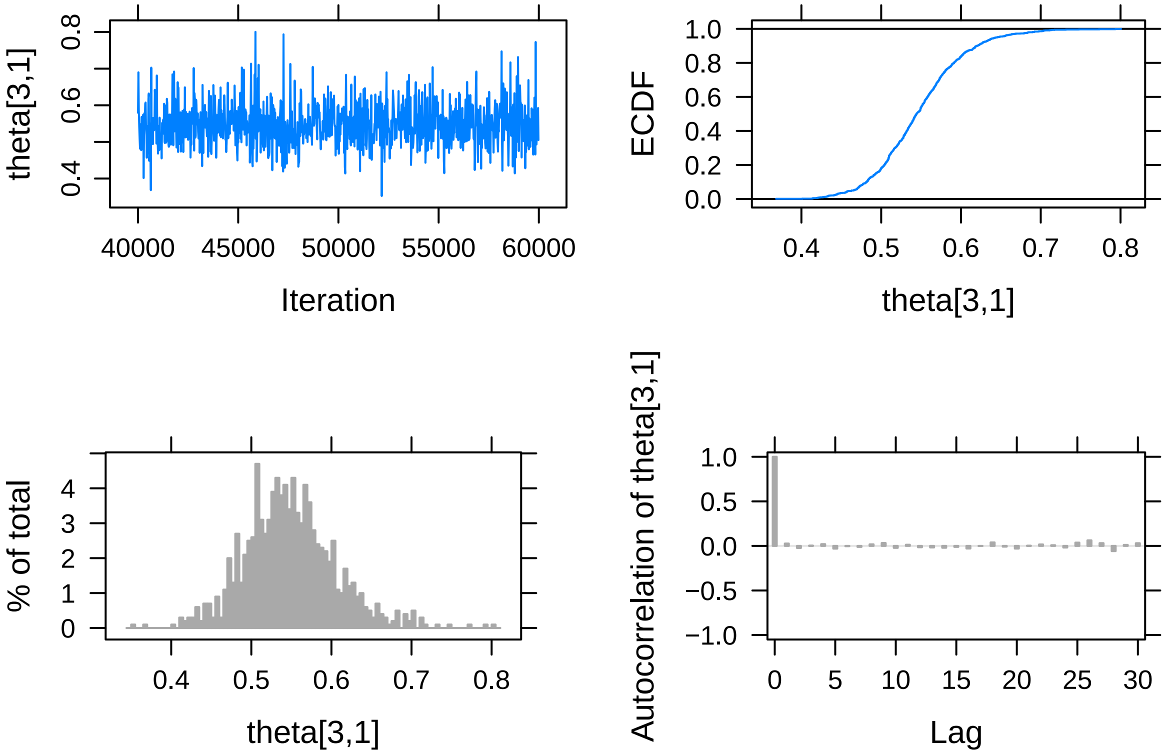

Table 1: Posterior Probabilities for Best and Worst Sectors

Sector	Posterior_Prob_Best	Posterior_Prob_Worst
1	0.0396	0.0408
2	0.4938	0.0023
3	0.0536	0.0326
4	0.0035	0.9218
5	0.4095	0.0026

Table 2: Posterior Probabilities for Best and Worst Stocks within Each Sector

Sector	Best_Stock_ID	Best_Stock_Prob	Worst_Stock_ID	Worst_Stock_Prob
1	1	0.2278	5	0.2473
2	6	0.2266	2	0.2016
3	9	0.2734	7	0.2054
4	9	0.4880	3	0.3747
5	6	0.3043	8	0.2974

Results and Interpretations

The best sector, which had the highest posterior probability of having positive stock returns, as seen in Table 1, was Sector 2 with Sector 5 as a close second best. Both sectors had a posterior probability of around 47.4% and 41.9%, respectively, which means that on average they are more likely to have positive returns than other sectors. The sector with the worst probability of having positive stock returns was sector 4, with a posterior probability of 92.5%. This means that it is much less likely than the other sectors to have positive returns. From Table 2, we can see that within Sector 1, stock 3 had the highest posterior probability (22.4%) of outperforming the other stocks in that sector. Stock 5 had the highest posterior probability (25.2%) of performing worse when compared to the other stocks in Sector 2. Interestingly, stock 3 was also the worst performing stock in Sector 4, with a posterior probability of 36.9% of performing worse than the other stocks.

Conclusion

Through our model using JAGS, we have analyzed 1500 returns on stocks across 10 different stocks across 5 different sectors, finding the log odds of how they perform. We have identified that Sector 2 had the highest performing stocks, so they are the best investment as they are the most likely have positive stock returns. Sector 4 had the lowest performing stocks.

Appendix

```r
stockReturns <- read.csv("/Users/jessicali/Documents/STAT 3303/Final Project/dataset127.csv")

# create data frame for y values
sum_df <- stockReturns %>%
  group_by(sector, stock) %>%
  summarise(y = sum(flip),.groups = 'drop') %>%
  arrange(sector,stock)

# create matrix for y values
y_matrix <- sum_df %>%
  pivot_wider(names_from = stock, values_from = y) %>%
  select(-sector) %>%
  as.matrix()
n <- 30

# proportion matrix for positive stock returns
prop <- y_matrix / n
prop_df <- melt(prop)
names(prop_df) <- c("Sector", "Stock", "Proportion")

# create box plots
ggplot(prop_df, aes(x = factor(Sector), y = Proportion)) +
  geom_boxplot(fill = "forestgreen") +
  labs(title = "Proportions of Positive Stock Returns by Sector",
       x = "Sector",
       y = "Proportion of Positive Returns")

#specify model
mod <- "model {
  # likelihood
  for (i in 1:n_sectors) {
    for (j in 1:n_stocks) {
      y[i, j] ~ dbin(theta[i, j], n)
      logit(theta[i, j]) <- eta[i, j]
      eta[i, j] ~ dnorm(mu[i], tau[i])
    }
  }

    # priors
    for (k in 1:n_sectors) {
    mu[k] ~ dnorm(mu_0, tau_0)
    tau[k] ~ dgamma(0.01,0.01)
  }

    mu_0 ~ dnorm(0, 1)
    tau_0 ~ dgamma(0.01, 0.01)

    # sigma
    sig <- sqrt(1/tau_0)

  }
"

n_sectors <- 5
n_stocks <- 10
n <- 30


list <- list(
  y = y_matrix,
  n = n,
  n_sectors = n_sectors,
  n_stocks = n_stocks
)

# initial values
myinit <- function() {
  list(
    mu_0 = rnorm(1, 0, 1),
    tau_0 = rgamma(1, 0.01, 0.01),
    mu = rnorm(n_sectors, 0, 1),
    tau = rgamma(n_sectors, 0.01, 0.01)
  )
}
posterior = run.jags(
  model = mod,
  monitor = c("theta","eta","mu","sig"),
  data = list,
  n.chains = 1,
  inits = myinit,
  adapt = 20000,
  burnin = 20000,
  sample = 20000
)
effectiveSize(posterior)
invisible(capture.output(plot(posterior,vars="theta[3,1]")))

# extract posterior samples for sector means
posterior_mu = as.data.frame(posterior$mcmc[[1]][, grep("^mu\\[", colnames(posterior$mcmc[[1]]))])

# best and worst sector for each sample
best_sector_each_draw = apply(posterior_mu, 1, which.max)
worst_sector_each_draw = apply(posterior_mu, 1, which.min)
```

```r
# posterior probabilities
best_sector_prob = table(best_sector_each_draw) / length(best_sector_each_draw)
worst_sector_prob = table(worst_sector_each_draw) / length(worst_sector_each_draw)

# theta samples
posterior_theta = as.data.frame(posterior$mcmc[[1]][, grep("^theta\\[", colnames(posterior$mcmc[[1]]))]

# find best and worst stock within each sector
find_best_worst_stock = function(sector_number) {
  sector_thetas = colnames(posterior_theta)[grep(paste0("^theta\\[", sector_number, ","), colnames(poste

  best_stock_each_draw = apply(posterior_theta[, sector_thetas], 1, which.max)
  worst_stock_each_draw = apply(posterior_theta[, sector_thetas], 1, which.min)

  best_stock_prob = table(best_stock_each_draw) / length(best_stock_each_draw)
  worst_stock_prob = table(worst_stock_each_draw) / length(worst_stock_each_draw)

  return(list(best = best_stock_prob, worst = worst_stock_prob))
}

# get best for all sectors
sector_best_worst = lapply(1:5, find_best_worst_stock)

# table summarizing best and worst stocks within each sector
sector_summary = data.frame(
  Sector = 1:5,
  Best_Stock_ID = NA,
  Best_Stock_Prob = NA,
  Worst_Stock_ID = NA,
  Worst_Stock_Prob = NA
)
```

```r
for (i in 1:5) {
  best = sector_best_worst[[i]]$best
  worst = sector_best_worst[[i]]$worst

  # stock number with highest probability
  best_id = as.numeric(names(which.max(best)))
  best_prob = max(best)

  worst_id = as.numeric(names(which.max(worst)))
  worst_prob = max(worst)

  sector_summary$Best_Stock_ID[i] = best_id
  sector_summary$Best_Stock_Prob[i] = round(best_prob, 4)
  sector_summary$Worst_Stock_ID[i] = worst_id
  sector_summary$Worst_Stock_Prob[i] = round(worst_prob, 4)
}

# table summarizing best and worst sector probabilities
sector_best_worst_summary = data.frame(
  Sector = as.numeric(names(best_sector_prob)),
  Posterior_Prob_Best = round(as.numeric(best_sector_prob), 4),
  Posterior_Prob_Worst = round(as.numeric(worst_sector_prob), 4)
)

# make tables look nice
kable(sector_best_worst_summary,
      caption = "Posterior Probabilities for Best and Worst Sectors",
      digits = 4,
      align = 'c')

kable(sector_summary,
      caption = "Posterior Probabilities for Best and Worst Stocks within Each Sector",
      digits = 4,
      align = 'c')
```