Practicum One

Josh Virene

2023-10-13

Problem Three: Least Squares Estimate of Beta

Now, download the daily S&P 500 indices (^GSPC) and the daily yields of the 13-week treasury yield (^IRX) for the same period. Combine your data into a single file where you have dates, stock prices for your "risky" and "safe" assets, and data on your market index, and risk free rates. What are we using as the risk-free rate and the market return rate? Compute the returns on each of your stocks. Divide your sample in two, the first half spanning January 2010-December 2014, and the second half covering January 2015–December 2019. Choose the half you want to work with.

- The file is available as "p2_dataset.xlsx"
- The risk free rate is given by the 13-week treasury yield (^IRX), and the market return rate is given by the S&P500 index (^GSPC)

Do the estimates of beta correspond well with your prior intution or beliefs? Why or why not? The regression output is provided in the summary statistics above, though to give the equations;

$$RiskPremium_{AAPL} = 0.00010 + 0.96910(r_m - r_f)$$

$$RiskPremium_{LUV} = 0.00154 + 1.04897(r_m - r_f)$$

These results are not consistent with what I would have expected prior to running these regression models. Given that I had set the airline industry to be much more risky compared to the tech industry, I anticipated that the airline beta coefficient would have been much larger compared to that of the tech industry's, though both are ~ 1.

Table 1: Apple Regression Summary Table

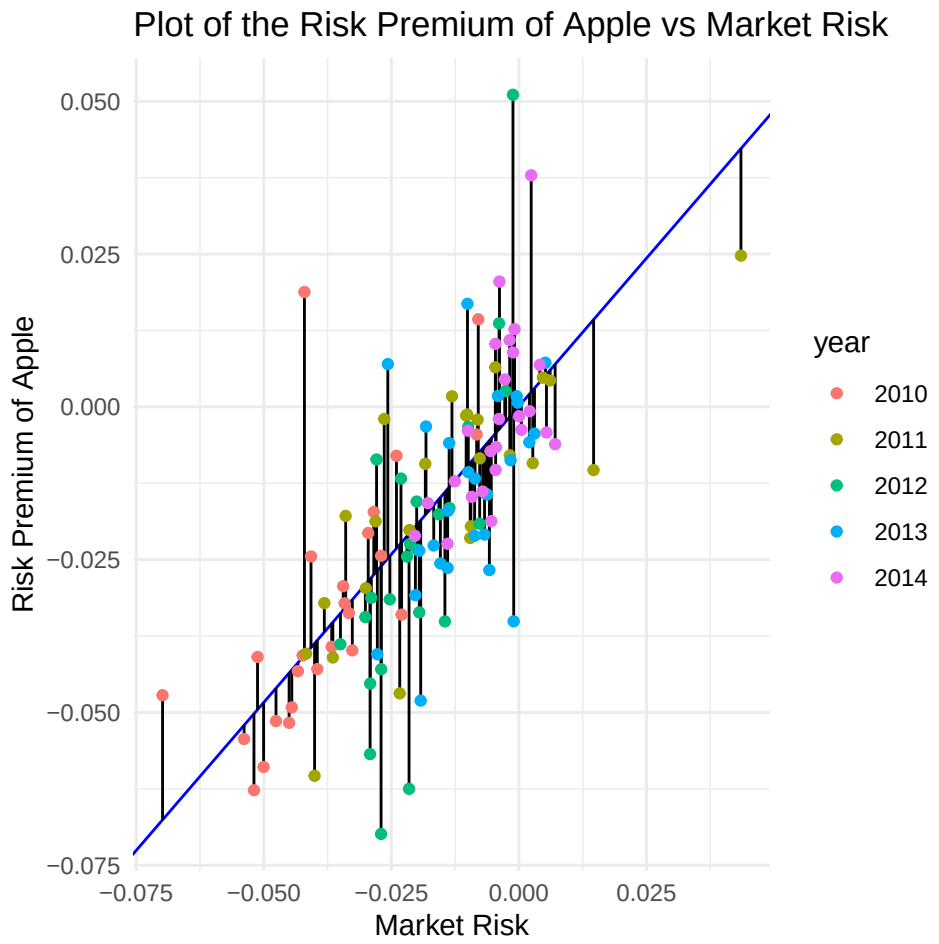
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0001032	0.0005842	0.176725	0.8597529
X	0.9690971	0.0242647	39.938566	0.0000000

Table 2: Southwest Regression Summary Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.001536	0.0005843	2.628877	0.0086715
X	1.048972	0.0242660	43.228029	0.0000000

For one of your companies, make a time plot of the historical company risk premium, the company risk premium predicted by the regression model, and the associated residuals. Are there any episodes or dates that appear to correspond with unusually large residuals? If so, attempt to interpret them.

From the graph, there appear to be a couple incidents where the residuals are very large relative to the rest of the data; these are in the years 2010 and 2012.



For each of the companies, test the null hypothesis that alpha = 0 against the alternative hypothesis that alpha != 0 using a significance level of 95%. Would rejection of this null hypothesis imply that the CAPM has been violated? Why or why not?

Testing the alpha coefficient for the AAPL model: In this model the summary statistics from the regression tell us that alpha *is not* statistically significant at the 95% confidence level (as indicated by P>0.05, and t <2.00). For this alpha parameter, we reject the null hypothesis at the 95% decision threshold.

Testing the alpha coefficient for the LUV model: In this model the summary statistics from the regression tell us that alpha *is* statistically significant at the 95% confidence level (as indicated by P<0.05, and t >2.00)

Rejecting the null hypothesis does not mean that we reject our CAPM model in this case. Looking at the alpha coefficient for Apple, the statistical decision should be to reject the null hypothesis that alpha = 1.000 in favor of the alternative that alpha != 1.000, we think that it is different enough from 1.000 at the 95% threshold. Based on the confidence interval constructed for alpha below however, and data provided in the graph above, I'll argue that unless the alpha is very unusual (i.e., a <= 10.00) we can say that the CAPM model still holds.

Table 3: Apple Regression Summary Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0001032	0.0005842	0.176725	0.8597529
X	0.9690971	0.0242647	39.938566	0.0000000

Table 4: Southwest Regression Summary Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.001536	0.0005843	2.628877	0.0086715
X	1.048972	0.0242660	43.228029	0.0000000

```
## [1] "0.9205677195 1.0176264987"

## [1] "The confidence interval for AAPL is:"

## [1] "0.9205677195 1.0176264987"

## [1] "The confidence interval for LUV is:"

## [1] "1.0004399728 1.0975040332"
```

For each company, construct a 95% confidence interval for beta. Then test the null hypothesis that the company's risk is the same as the average risk over the entire market, that is, test that beta = 1 against the alternative hypothesis that beta != 1. Do you find any surprises?

Formula: CI = point estimate +/- 2SE, these are calculated and obtained in the code above.

The confidence interval for AAPL is: (0.9205677195 1.0176264987)

The confidence interval for LUV is: (1.0004399728 1.0975040332)

Testing the beta coefficient for the AAPL model: In this model the summary statistics from the regression tell us that beta *is* statistically significant at the 95% confidence level (as indicated by P<0.05, and t >2.00). For this alpha parameter, we reject the null hypothesis at the 95% decision threshold.

Testing the beta coefficient for the LUV model: In this model the summary statistics from the regression tell us that beta *is* statistically significant at the 95% confidence level (as indicated by P<0.05, and t >2.00)

There are no surprises here, based on the figure created above for Apple, we clearly see the slope for the line of best fit is approximately equal to one, which makes sense given the data points that it runs through. Surely this would also hold if I made the same plot for Southwest instead of Apple.

For each of the two companies, compute the proportion of total risk that is market risk, also called systematic and nondiversifiable. William F. Sharpe [1985, p. 167] states that "Uncertainty about the overall market... accounts for only 30% of the uncertainty about the prospects for a typical stock." Does evidence from the two companies you have chosen correspond to Sharpe's typical stock? Why or why not? What is the proportion of total risk that is specific and diversifiable? Do these proportions surprise you? Why?

Formula to calculate this:

$$\beta = Risk_{market}/Risktotal$$

Rearranging, we can solve for the proportion of market risk using the formula:

$$Risk_{market} = \beta/Risk_{total}$$

Total risk = sd(returns) where returns are the returns for each company

Proportion of Market Risk for Apple = 0.969 / sd(returns_apple) **= 57.7**

Proportion of Market Risk for Southwest = 1.048 / sd(returns_southwest) **= 54.9**

Based on Sharpe's claim that uncertainty about the overall market accounts for only 30% of the uncertainty about the prospects for a typical stock makes good sense in the context of the estimates that I have provided above. Looking at Apple, the risk attributed to apple that is systematic and nondiversifiable is 57.7%, and thus the remaining 42.3% is specific and diversifiable. This relatively low proportion of specific risk is inconsistent with my hypothesis that it would be the more stable of the two companies when comparing to Southwest. A potential reason for this is that Apple had started branching out into Fintech in 2014, which may have introduced higher volatility in the latter portion of the study period. Looking at Southwest, it had a systematic and nondiversifiable risk of 54.9%, then the remaining 44.1% is specific and diversifiable. This makes pretty good sense, it is likely that the specific risk would have been higher if I examined the period during the COVID-19 infection because this introduced high volatility into the company's returns.

In your sample, do large estimates of beta correspond with higher R2 values? Would you expect this always to be the case? Why or why not?

From the regressions I ran, the Apple model has a beta coefficient of 0.969, and a R^2 of 0.56, while the Southwest model has a beta coefficient of 1.048, and an R^2 of 0.598. This shows that higher values of beta correspond to a higher R^2. A note about this however, I only ran these two regression models, so to make this conclusion without a larger sample size (i.e., more regressions) means we cannot put too much weight on this conclusion. Furthermore, as discussed in the lecture, there are exceptions to this: a. If a model has a high beta coefficient, the R^2 can be low if the (sigma_j)^2 parameter is large b. If a model has a high

$$\sigma_{jm}/\sigma_m^2$$

, the R^2 is large, while the beta coefficient is small.

Problem 4: Is January Different?

There is some weak evidence supporting the notion that stock returns in the month of January are, all else equal, higher than in other months, especially for smaller companies. Why this might be the case is not clear, since even if investors sold losing stocks during December for tax reasons, the expectation that January returns would be higher would shift supply and demand curves and thereby would tend to equilibrate returns

over the year via the possibility of inter-temporal arbitrage. However, the "January is different" hypothesis does seem worth checking out empirically. In this exercise you will investigate how this hypothesis might be tested.

First, if the "January premium" affected the overall market return rm and the risk-free return rf by the same amount, say, jm, show that the market risk premium would be unaffected. In this case, could the "January is different" hypothesis be tested within the CAPM framework? Would it not make more sense, however, to assume that the "January is different" hypothesis referred only to risky assets? Why or why not?

The market risk premium would be unaffected, which is shown by the following:

$$\delta = r_j$$

$$r_j - r_f = \alpha_j + \beta_j(r_m + \delta - r_f + \delta)$$

(1)

$$\delta(r_j - r_f) = \alpha_j + B_j * \delta(r_m - r_f)$$

(2)

In the above set of equations, we see that the January premium does not affect the market risk premium. Looking at the risky asset premium however, equation two illustrates that this is affected- it is multiplied by r_j.

Answer to the first part of this question (would the market risk premium be affected?: Yes, this market risk premium would be affected, as indicated by the following:

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f)$$

$$r_j - r_f = \alpha_j + \beta_j(r_m + j_m - r_f)$$

We can see from the second equation that the parameter that we are regressing the risk premium on (the term that beta_j is multiplied by) changes, so it follows that the market risk premium is affected.

Answer to the second part of this question: first, we use the equations:

$$rm_{prime} = r_m + j_m$$

and:

$$rp_{prime} = r_p + \beta_1 * j_m$$

Substituting these into the original CAPM model, we get:

$$r_p + \beta_j m = \alpha_j + \beta_j m(r_m - r_f)$$

Subtracting Beta*J_M, we reach

$$r_p - r_f = \alpha_j + (r_m - r_f)$$

This indicates that within the CAPM framework, it is not possible to test the January is different hypothesis due to an incomplete model specification.

Table 5: Motorola January Premium Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0190497	0.0078153	2.4375021	0.0178761
DUMJ	-0.0082804	0.0270729	-0.3058566	0.7608089

Table 6: Exxon January Premium Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0070827	0.0062510	1.1330497	0.2618555
DUMJ	0.0012262	0.0216541	0.0566254	0.9550382

Table 7: Bank of America January Premium Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0023806	0.0137945	0.1725766	0.8635851
DUMJ	0.0701030	0.0477854	1.4670369	0.1477677

I will define a "reasonable level of significance" as the 95% threshold, so statistical significance is satisfied under $p < 0.05$, $t > 2$"

Assessing the statistical significance of all three models (the Motorola, Exxon, and Bank of America), P>0.05, and t<2, from this decision, we fail to reject the null hypothesis that the beta coefficients for DUMJ on these variables are not equal to zero at the 95% confidence level.

Answering the question, is January different? Yes, it is different based on the conclusion provided above. To elaborate on this, a beta coefficient of 0 implies that that variable of interest tied to the coefficient has no effect on the model because we get 0*(variable), thus y is unchanged. Because our statistical test for the estimates, which in all cases gave coefficients that are so small (-0.008, 0.001, and 0.07) these coefficients can be approximated by zero. Since the statistical tests told us to reject he null hypothesis at the 95% confidence level; we cannot say with certainty that January is not different.

Table 8: Motorola January Premium Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0190497	0.0078153	2.4375021	0.0178761
DUMJ	-0.0082804	0.0270729	-0.3058566	0.7608089

Table 9: Exxon January Premium Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0070827	0.0062510	1.1330497	0.2618555
DUMJ	0.0012262	0.0216541	0.0566254	0.9550382

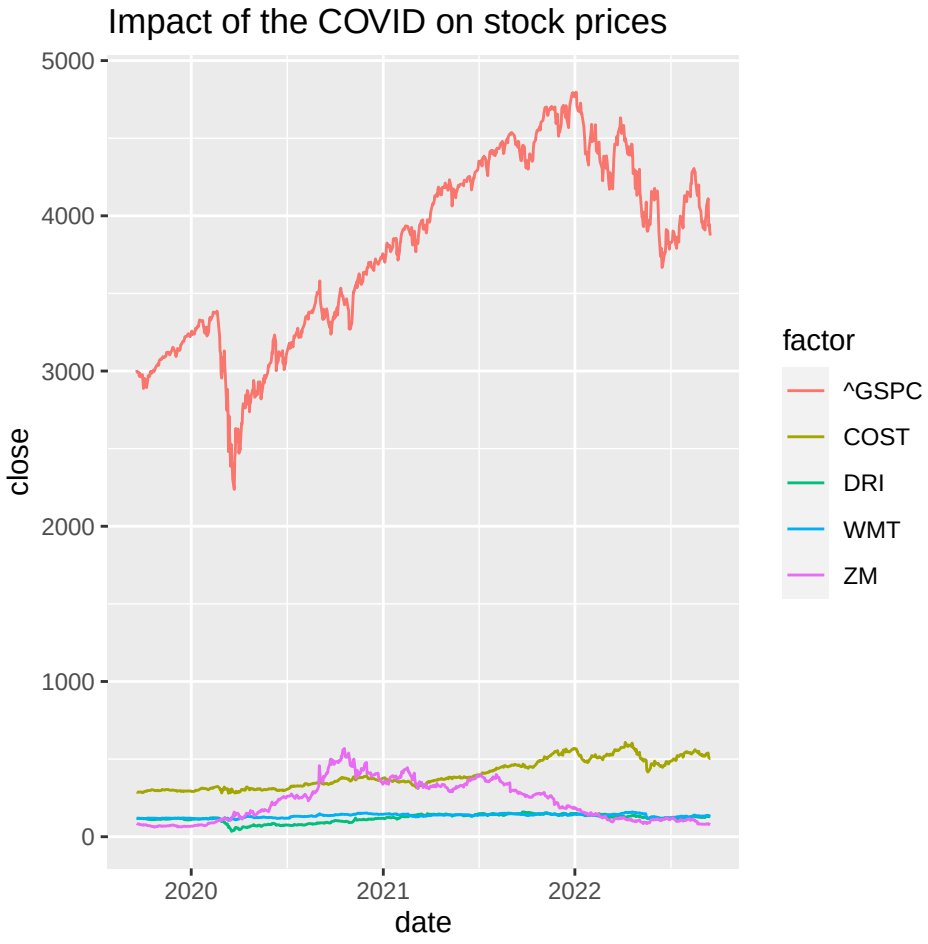
Table 10: Bank of America January Premium Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0023806	0.0137945	0.1725766	0.8635851
DUMJ	0.0701030	0.0477854	1.4670369	0.1477677

In running this new model where the intercept is different based on whether the month is January or not, the regression tables lead to the same conclusion to reject the null hypothesis at the 95% confidence level; we cannot say with certainty that January is not different.

Based on these model runs, it is clear that we cannot say with certainty that January is not different than the other months at the 95% significance level. *Plainly put, these results don't give evidence that January is not different.* Though, it is important to note that the small scale in terms of the stocks that we are evaluating, our only metric being returns, and other assumptions such as setting the price premium for the intercept in part d at 0.5 might play a role in this. In order to really get at the question on if January is better or not, further testing using other response variables, a more inclusive set of stocks, and some other regression tools / techniques could allow for a deeper dive into this question and a more complete analysis.

Problem 5: COVID lockdown: Event Studies



DO NOT EDIT: THIS IS THE TEMPLATE FOR ADDITIONAL CODE CHUNKS

Appendix

```
library(knitr)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 5,
    fig.height = 5, tidy = TRUE)
# obtaining all of the needed packages
require(pacman)
p_load(tidyquant, tidyverse, data.table, dplyr, writexl, stargazer, formatR, tinytex)
```

```r
# establish working directory:
mydir <- getwd()

# code for problem 3a download the data:

# timeframes
p1_start <- "2010-01-01"
p1_end <- "2014-12-31"

p2_start <- "2015-01-01"
p2_end <- "2019-12-31"
# list of all industries to download data for
combined <- c("AAPL", "LUV", "^GSPC", "^IRX")
# tq_get will download the data
full_dataset <- tq_get(combined, get = "stock.prices", from = p1_start, to = p2_end)

P1_dataset <- tq_get(combined, get = "stock.prices", from = p1_start, to = p1_end)

# Write the datasets to an excel file, as required in the question
write_xlsx(P1_dataset, "p2_dataset.xlsx")

# Get rid of unnecessary columns, we only need date, price (close), company
prices <- P1_dataset %>%
    select(c(symbol, date, close))

# reshape data, remove NAs from the data
prices <- prices %>%
    spread(., symbol, close)
prices <- na.omit(prices)

# calculate the risk free rate:
prices <- prices %>%
    mutate(r_f = 100 * (((1/(1 - ((`^IRX` * 88)/360)))^(1/88)) - 1))

prices <- prices %>%
    tq_mutate(select = AAPL, mutate_fun = periodReturn, period = "daily", col_rename = "r_AAPL") %>%
    tq_mutate(select = LUV, mutate_fun = periodReturn, period = "daily", col_rename = "r_LUV") %>%
    tq_mutate(select = "^GSPC", mutate_fun = periodReturn, period = "daily", col_rename = "r_^GSPC")

# end of the code to setup the data for question three

# code for problem 3b

# setup:need to create new columns for the dependent variables (risk free
# rates) & independent variable (r_m - r_f) dependent variables
prices$riskfree_AAPL = prices$r_AAPL - prices$r_f  #This is the dependent variable (market risk for App
prices$riskfree_LUV = prices$r_LUV - prices$r_f  #This is the dependent variable (market risk for South
prices$X = prices$`r_^GSPC` - prices$r_f  # This is the independent variable (Risk of overall market)

# running the CAPM model:
AAPL_reg <- lm(riskfree_AAPL ~ X, data = prices)
summarystat <- summary(AAPL_reg)
kable(summarystat$coefficients, caption = "Apple Regression Summary Table")
```

```r
LUV_reg <- lm(riskfree_LUV ~ X, data = prices)
summarystat2 <- summary(LUV_reg)
kable(summarystat2$coefficients, caption = "Southwest Regression Summary Table")

# code for problem 3c

# creating regression line
intercept <- 1e-04
slope <- 0.9691
prices$f <- intercept + slope * prices$X

# Note on this plot: I only select every tenth value from our predictor, this
# gives a subset the data that is representative in that it randomly selects
# every tenth value, and displays residuals more clearly (there are now only
# 126 data points instead of 1256)

selected_data <- prices[seq(5, nrow(prices), by = 10), ]
selected_data$year <- substr(format(selected_data$date, format = "%Y"), 1, 4)
prices$year <- substr(format(prices$date, format = "%Y"), 1, 4)

# plot
p <- ggplot(selected_data, aes(x = X, y = riskfree_AAPL)) + geom_abline(slope = slope,
    intercept = intercept, color = "blue") + geom_segment(aes(xend = X, yend = f)) +
    geom_point(aes(x = X, y = riskfree_AAPL, color = year)) + theme_minimal() + xlab("Market Risk") +
    ylab("Risk Premium of Apple") + ggtitle("Plot of the Risk Premium of Apple vs Market Risk")
p
# code for problem 3d
kable(summarystat$coefficients, caption = "Apple Regression Summary Table")
kable(summarystat2$coefficients, caption = "Southwest Regression Summary Table")
# code for problem 3e

# constructing the confidence intervals for each beta. Formula, CI = pt
# estimate +/- 2(SE) AAPL
lower_bound = 0.9690971091 - 2 * (0.0242646948)
upper_bound = 0.9690971091 + 2 * (0.0242646948)
print(paste0(lower_bound, " ", upper_bound))
print("The confidence interval for AAPL is:")
print(paste0(lower_bound, " ", upper_bound))

lower_bound = 1.048972003 - 2 * (0.0242660151)
upper_bound = 1.048972003 + 2 * (0.0242660151)
print("The confidence interval for LUV is:")
print(paste0(lower_bound, " ", upper_bound))

# code for problem 3f

# calculating the risk for apple
tr_apple = sd(prices$r_AAPL)
beta_appl <- 0.969
r_appl = beta_appl/tr_apple

# calculating the risk for southwest
tr_southwest = sd(prices$r_LUV)
```

```r
beta_southwest <- 1.048
r_southwest = beta_appl/tr_southwest
# code for problem 3g

# code for problem 4a

# code for problem 4b

# code for problem 4c

# download dataset for the new industries- Exxon (XOM), Motorola (MSI), and
# Chase (FXCB) timeframe
p1_start <- "2010-01-01"
p1_end <- "2014-12-31"
# list of industries to get data for
combined <- c("XOM", "MSI", "BAC")
# tq_get to download data
q4_dataset <- tq_get(combined, get = "stock.prices", from = p1_start, to = p1_end)

# change the price data to monthly
monthly_prices <- q4_dataset %>%
    na.omit() %>%
    group_by(symbol) %>%
    tq_transmute(select = close, mutate_fun = to.monthly, col_rename = "monthly_price")

monthly_prices <- monthly_prices %>%
    spread(., symbol, monthly_price)

# calculating the risk premium
monthly_prices <- monthly_prices %>%
    tq_mutate(select = XOM, mutate_fun = periodReturn, period = "daily", col_rename = "r_XOM") %>%
    tq_mutate(select = MSI, mutate_fun = periodReturn, period = "daily", col_rename = "r_MSI") %>%
    tq_mutate(select = BAC, mutate_fun = periodReturn, period = "daily", col_rename = "r_BAC")

# creating the dummy variable:
monthly_prices <- monthly_prices %>%
    mutate(DUMJ = ifelse(str_detect(monthly_prices$date, "Jan"), 1, 0))


# running the regression models:
MSI_mod <- lm(r_MSI ~ DUMJ + 1, data = monthly_prices)
MSI_summary <- summary(MSI_mod)
kable(MSI_summary$coefficients, caption = "Motorola January Premium Summary Statistics")

XOM_mod <- lm(r_XOM ~ DUMJ + 1, data = monthly_prices)
XOM_summary <- summary(XOM_mod)
kable(XOM_summary$coefficients, caption = "Exxon January Premium Summary Statistics")

BAC_mod <- lm(r_BAC ~ DUMJ + 1, data = monthly_prices)
BAC_summary <- summary(BAC_mod)
kable(BAC_summary$coefficients, caption = "Bank of America January Premium Summary Statistics")

# code for problem 4d
```

```r
# some basic if else logic allows us to change our intercept based on the value
# for DUMJ we say if month == January, we get an intercept that is 0.5 higher
# than that for any other month
monthly_prices$int1 <- ifelse(monthly_prices$DUMJ == 1, 0.5, 0)

MSI_mod <- lm(r_MSI ~ DUMJ + int1, data = monthly_prices)
MSI_summary <- summary(MSI_mod)
kable(MSI_summary$coefficients, caption = "Motorola January Premium Summary Statistics")

XOM_mod <- lm(r_XOM ~ DUMJ + int1, data = monthly_prices)
XOM_summary <- summary(XOM_mod)
kable(XOM_summary$coefficients, caption = "Exxon January Premium Summary Statistics")

BAC_mod <- lm(r_BAC ~ DUMJ + int1, data = monthly_prices)
BAC_summary <- summary(BAC_mod)
kable(BAC_summary$coefficients, caption = "Bank of America January Premium Summary Statistics")

# code for problem 5 - setup only

# SETUP getting the eventstudy package: install.packages('digest')
# install.packages('githubinstall') library(githubinstall)
# githubinstall('eventstudies',force=TRUE) library(eventstudies)

# Downloading the data:
start_date <- "2019-09-19"
end_date <- "2022-09-18"

combined <- c("DRI", "ZM", "^GSPC", "COST", "WMT")

full_dataset <- tq_get(combined, get = "stock.prices", from = start_date, to = end_date)


# reformatting the data: *** This is the dataset for the value of the stocks
stock_value <- full_dataset %>%
    select(c(symbol, date, close))

# reshape data, remove NAs from the data
stock_value <- stock_value %>%
    spread(., symbol, close)
stock_value <- na.omit(stock_value)

##########################################################################################################

# Get rid of unnecessary columns, we only need date, price (close), company
stock_return <- full_dataset %>%
    select(c(symbol, date, close))

# reshape data, remove NAs from the data
stock_return <- stock_return %>%
    spread(., symbol, close)
stock_return <- na.omit(stock_return)

stock_return <- stock_return %>%
```

```
    tq_mutate(select = DRI, mutate_fun = periodReturn, period = "daily", col_rename = "r_DRI") %>%
    tq_mutate(select = ZM, mutate_fun = periodReturn, period = "daily", col_rename = "r_ZM") %>%
    tq_mutate(select = COST, mutate_fun = periodReturn, period = "daily", col_rename = "r_COST") %>%
    tq_mutate(select = WMT, mutate_fun = periodReturn, period = "daily", col_rename = "r_WMT") %>%
    tq_mutate(select = "^GSPC", mutate_fun = periodReturn, period = "daily", col_rename = "r_^GSPC")

# code for problem 5a

# Establishing the estimation window, the event window, and the lockdown date:
ca_lockdown <- as_date("2019-03-19")
event_window_start <- ca_lockdown - ddays(30)
event_window_end <- ca_lockdown + ddays(30)

estimation_window_start <- ca_lockdown - ddays(120)
estimation_window_end <- ca_lockdown + ddays(120)

three_months <- ca_lockdown - ddays(90)

# Now we can plot to answer this question, working with the prices dataset:

# for this first plot, we need the data that isn't in the reshaped form so that
# we can color our trendlines by converting the company to a factor

stock_value1 <- full_dataset %>%
    select(c(symbol, date, close))

start_date = as_date(three_months)
end_date = as_date(event_window_end)

filtered_data <- subset(stock_value1, date >= start_date & date <= end_date)


# plotting the value of stocks during the event window:

factor <- as.factor(stock_value1$symbol)
p <- ggplot(stock_value1) + ggtitle("Impact of the COVID on stock prices") + geom_line(aes(x = date,
    y = close, color = factor)) + geom_vline(xintercept = as.numeric(as.Date("2019-03-19")),
    color = "red", lwd = 0.5)
p


# code for problem 1x
```