MULTIPLE LINEAR REGRESSION REPORT

S&P 500

Group 21 Members:
Patrick Luo
Antonio Melacini
Kris Zhang

Contents

Abstract

We explore the use of multiple linear regression to model the relationship between the S&P500 price (the response variable) against variables such as the price of a high yield bond index, treasury bond yields, and the weights of certain sectors in the S&P500. We carefully select variables in our model to avoid multicollinearity and discuss competing models to select a final best model. Our final model shows statistically significant linear associations for all covariates, although we find that the model does not satisfy all assumptions. We propose a transformation to address this, although we are not successful. Finally, we discuss model interpretation and study limitations.

Introduction

The primary goal of our analysis is to investigate potential linear associations between financial and macroeconomic variables against the response variable of the S&P500 price. We first explore the relationship between the S&P U.S. High Yield Corporate Bond Index and the price of S&P 500. Is there any linear relationship between the price of this bond index and the price of the S&P500? We are interested in potential relationships among the variates which are logically sound in the underlying financial context of the data. Next, we consider the possible relationships with the two sector weight variates (information technology sector weighting and industrial sector weighting), with particular interest in possible interaction effects. What sectors have positive or negative associations with the S&P500 price? We then explore the presence of any linear relationships among selected macroeconomic variables and fixed income securities, such as the federal funds rate and the yields of different maturity treasury bonds. What are the associations between the macroeconomic variables and the S&P500 price? We set out to answer these questions by going through several different models and iteratively constructing larger models before finally selecting the best one.

Dataset Description

The dataset consists of 9 variates with values recorded monthly over time from November 30, 1994 to December 31, 2023. Data was collected from Bloomberg. If the frequency of recording for the variate in the original source was more frequent than monthly, this data was aggregated by taking the last value of the month. Our dataset contains variate values recorded across 350 time points (months), so our sample size is $n = 350$. Our dataset is composed of 9 quantitative variates.

The response variate is “sp500_price” (y), the closing price of the S&P500 in $USD. The covariate in our dataset with the strongest correlation with the response is “corp_high_yield_index” (the closing price of the S&P U.S. High Yield Corporate Bond Index). We also have the covariates “tech_weight” and “industrials_weight”, which represent the market-cap weighted proportion of the S&P500 that is composed of the information technology and industrials sectors, respectively. Several macroeconomic variates are included, such as “fed” (U.S. Federal Funds Effective Rate, in percentage points) and “yield_curve_spread” (the 10-year treasury yield minus the 2-year treasury yield, in basis points). Also included are the yields in percentage points for U.S. treasury bonds of 2-year, 5-year, and 10-year maturities, which are the variates “tbill_2y”, “tbill_5y”, “tbill_10y”.

Note: for variates which are not available at a constant daily frequency (e.g. data only available on trading days), the “Date” value in the dataset may be off by one or two days. The “Date” value is intended to indicate the end of the month and not necessarily the exact date.

Multicollinearity

As can be observed by inspecting the correlation matrix in the appendix, the covariates in our dataset exhibit multicollinearity. We must choose covariates with low multicollinearity in our model. We

observe a very strong negative correlation between the covariates “tech_weight” and “industrials_weight”. Out of these two sector weight variates, the industrials weight is more strongly correlated with the response, however it also has a higher correlation with “junkbonds_price”. Due to this, we explore a linear model with the only sector weight being tech as well as a model with the only sector weight being industrials. We also find a large amount of multicollinearity among the macroeconomic variates. The manner in which we deal with the presence of multicollinearity for each scenario is addressed in our main analysis.

Initial Analysis of Response Variate

We begin by investigating the sample distribution of the response variate (y), the end of month closing price of the S&P500. Selected plots and summary statistics are shown below. The histogram in Figure 2 is superimposed with the probability density function of a random variable which follows a Normal distribution with a mean of the sample mean and a variance of the sample variance. The distribution is very clearly not Normal. The variate has a median of 1362.5, observed from Figure 1, and the distribution is extremely unsymmetric. The boxplot in Figure 3 also shows the positive skew of the distribution. The distribution appears unimodal, concentrated around the median of 1362.5, with a large positive tail.

Figure 1

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
453.7  1104.3  1362.5  1784.4  2170.3  4769.8
```

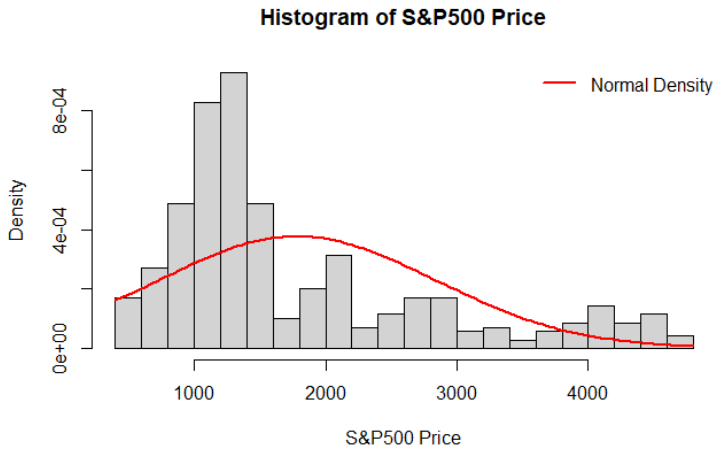


Figure 2

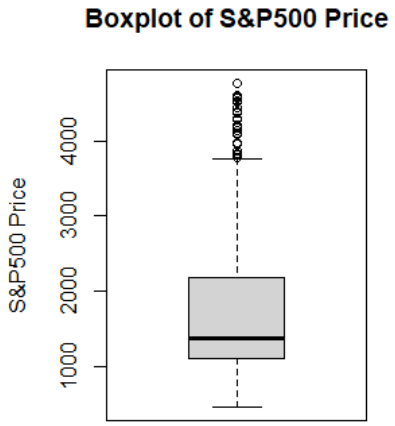
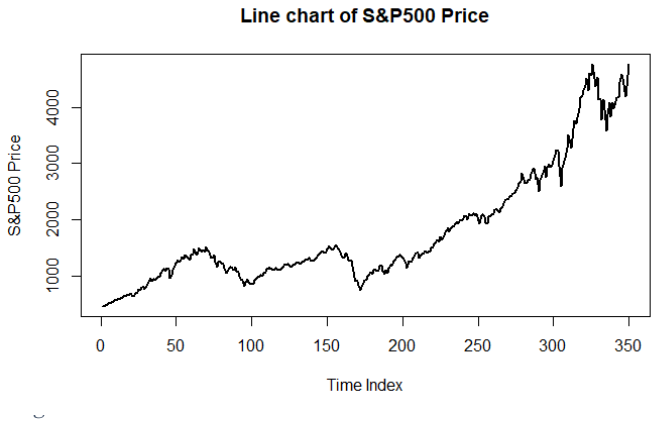


Figure 3



It is important to acknowledge that due to the nature of the underlying variate, the distribution of the raw price of the S&P500 index over time is not very useful in this context, and largely depends on the time frame selected for the sample. We can see from Figure 4, the S&P500 price variate is clearly time-dependent, i.e. the variable is non-stationary, meaning that we observe a clear trend over time. Our response variate, as well as some of our covariates such as the junk bonds index, display clear positive trends over time. This is a key limitation to our study (which we discuss further in our conclusion), since this means that these variates may display misleading correlations with unrelated variates. If interest lied in the S&P500 price trends alone, it would be more meaningful to analyze the distribution of the returns of the S&P500 (the period over period growth values) instead. Although for the purposes of this study we are only interested in the relationships between the price of the S&P500 and the selected covariates discussed earlier.

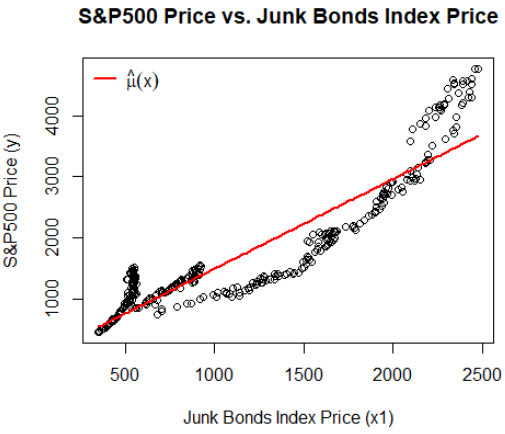
Main Analysis

We set our type 1 error tolerance to $\alpha = 5\%$. We find that the variate in our dataset holding the greatest correlation with the response variate is the price of the S&P U.S. high yield corporate bond index (a.k.a. junk bonds index). We first discuss the simple linear regression with the single explanatory variate being the price of the junk bonds index (call it x_1) and the response variate y. After considering this model, we explore other covariates in our dataset, and eventually discuss a final combined model for which we will analyze the validity of the model assumptions.

We assume the model

$$y = \mu(x_1) + \epsilon = \beta_0 + \beta_1 x_1 + \epsilon$$

where y is the price of the S&P500 and x_1 is the price of the junk bonds index. Let this model be referred to as M1. This model has $R^2_{adj} = 0.8344$, indicating a strong linear association, and $\hat{\beta}_1 > 0$ indicates a positive association. The price of the junk bonds index has a statistically significant linear association with the S&P500 price at the 5% level, i.e. the p-value for testing $H_0: \beta_1 = 0 \ \ vs. \ \ H_A: \beta_1 \neq 0$ is a value <5%.



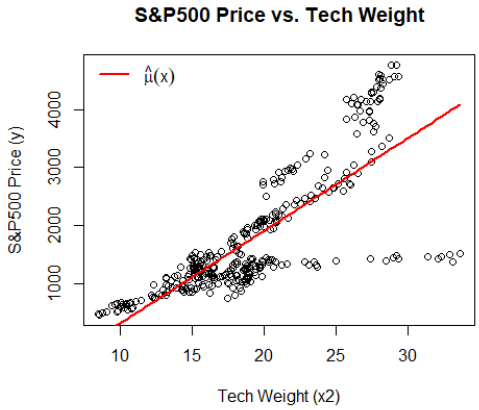

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.50518   47.94645   0.449    0.654
x1           1.46940    0.03503  41.942   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 431.6 on 348 degrees of freedom
Multiple R-squared:  0.8348,    Adjusted R-squared:  0.8344
```

We consider M1 (this model) our initial model. From here on we search for other statistically significant covariates to answer our questions of interest and potentially include in a combined multiple regression model.

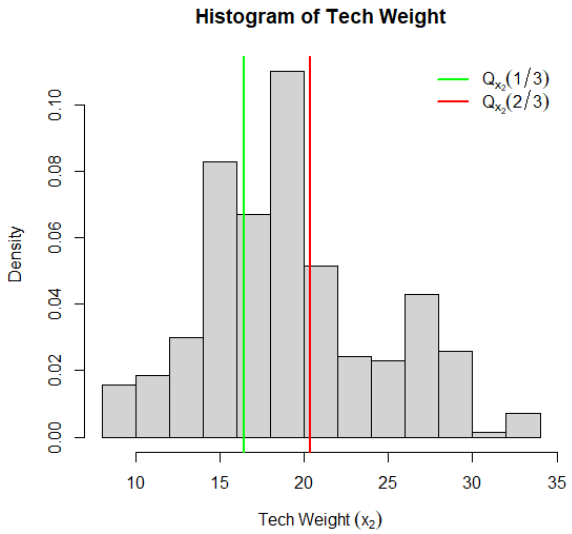
Figure 6

Moving on, we explore the weights of certain sectors in the S&P500 as potential covariates. As previously mentioned, the two sector weight variates are strongly correlated with each other. We first investigate a potential model using only tech weight because interest lies in the isolated relationship of the response with the tech sector. We also eventually consider a model using the industrial sector weight. Let x_2 be the tech weight variate. We fit a simple linear regression using the response y with the covariate x_2. This model has a relatively high adjusted R-squared of $R^2_{adj} = 0.6213$, and the slope parameters are significant. We observe a moderately strong positive linear association. As an alternative to the model presented above, we consider quantifying the tech sector weight variate as a categorical variate by defining



$$x_{\text{tech level}} = \begin{cases} 2 \text{ if } x_2 > Q_{x_2}\left(\frac{2}{3}\right) \\ 1 \text{ if } Q_{x_2}\left(\frac{1}{3}\right) \leq x_2 \leq Q_{x_2}\left(\frac{2}{3}\right) \\ 0 \text{ if } x_2 < Q_{x_2}\left(\frac{1}{3}\right) \end{cases}$$

So for each case $i = 1, \ldots, n$, where $Q_{x_2}(q)$ is the q^{th} quantile of the x_2. This newly defined categorical variate splits up the values of x_2 into 3 different levels based on its (1/3) and (2/3) quantiles. We are interested in fitting the model with this categorically transformed variate so that we can explore whether a broadly defined 'level' of tech weight has a stronger linear relationship than the exact percentage value (which could possibly include unnecessary noise). Figure 10 shows the histogram of x_2 with vertical lines separating the levels by the quantiles and Figure 11 shows a scatterplot with the fitted line.



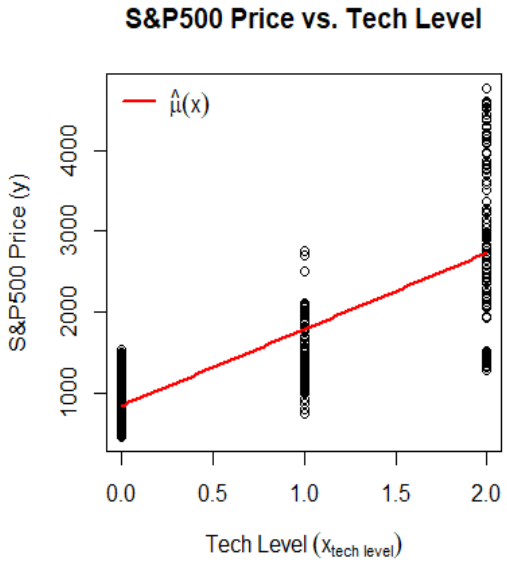


Figure 11

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       838.40      61.24   13.69   <2e-16 ***
sectors$tech_cat  946.04      47.41   19.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 725.2 on 348 degrees of freedom
Multiple R-squared:  0.5336,    Adjusted R-squared:  0.5323
```

Covariate in Model	AIC of Model
x_2	5533.885
$x_{\text{tech level}}$	5607.766

When we fit the simple linear regression with the response y against the covariate $x_{\text{tech level}}$, we find that this categorical variate is a statistically significant covariate. However, this model performs worse than our previous model since $R^2_{adj} = 0.5336 < 0.6213$. That is, the R^2_{adj} from using $x_{\text{tech level}}$ is less than that of the previous model which left tech weight as a quantitative variate. We also find that the AIC of the model using the categorical covariate is higher than the one using the original quantitative covariate, which is suggesting that the model using the quantitative variate is better. Therefore, we conclude that there is important detail in the original data x_2 which is lost when we use $x_{\text{tech level}}$, so we continue by using the quantitative variate x_2.

Figure 10

The table below summarizes correlations between the sector weights, the junk bonds index, and the response.

Correlation	S&P500 Price (y)	Junk Bonds Price
Tech Weight	0.7889169	0.7069426
Industrial Weight	−0.8457221	−0.7932552

As previously mentioned, we also know that the tech weight and industrial weight are highly correlated with each other, so we only consider them in separate models. The VIF between the tech weight and industrial weight variates is $4.933399 \approx 4.94$, which is technically below the threshold of 5, although just barely. For this reason, we make the conservative decision to only use one of these variates in a single model. We are faced with the tradeoff between the industrial weight's stronger correlation with the response combined with it's stronger collinearity with x_1 versus the tech weight's weaker correlation with the response combined with it's weaker collinearity with x_1.

Our interest lies more with the tech weight variate, so we proceed by first attempting to construct a model using x_2 (tech weight), and eventually we compare this larger model with a model using the industrial weight. The VIF between x_1 and x_2 is $1.9992 \approx 2$, which does not indicate high multicollinearity, so we continue by including x_2 in our combined model with x_1.

Our combined model is now $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, which we refer to as M2. This model has $R^2_{adj} = 0.8757$, indicating a strong linear association. The adjusted R-squared value for M2 is higher than that of M1, showing that we obtain a stronger model by transitioning from M1 to M2 while accounting for the number of parameters in the model. The p-value obtained for the F-test $H_0: \beta_1 = \beta_2 = 0$ $vs.$ $H_A: \beta_1 \neq 0$ or $\beta_2 \neq 0$ as well as the individual t-tests for the slope parameters are statistically significant.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -702.02465   79.51513  -8.829   <2e-16 ***
x1             1.14443    0.04303  26.595   <2e-16 ***
tech          57.93757    5.42377  10.682   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 374.9 on 347 degrees of freedom
Multiple R-squared:  0.8757,    Adjusted R-squared:  0.875
F-statistic:  1223 on 2 and 347 DF,  p-value: < 2.2e-16
```

In an attempt to incorporate some macroeconomic logic into our model, we examine the relationships between the response and the yields of different maturity U.S. treasury bonds as well as the fed rate. The correlation matrix for these variates along with the response is provided in the appendix, and we summarise our findings here. All of these macroeconomic variates are highly correlated with one

another, however the yield of the 10-year treasury has the strongest correlation with the S&P500 price, so this is what we use as a covariate going forward.

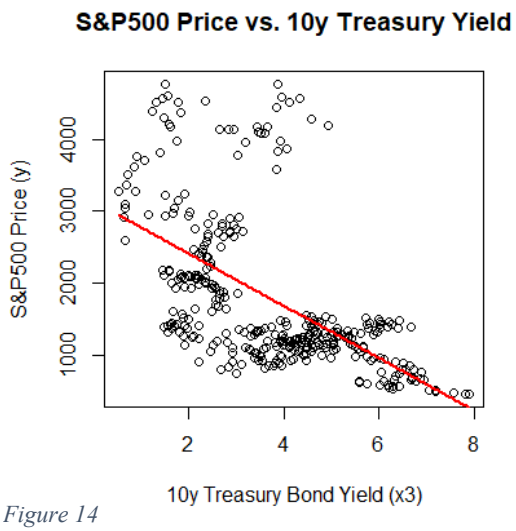


Figure 14

Let x_3 be the variate representing the yield of the 10-year treasury bond. We fit the simple linear regression model using the response y and the covariate x_3.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3134.98     115.43   27.16   <2e-16 ***
x3           -363.14      28.37  -12.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 875.6 on 348 degrees of freedom
Multiple R-squared:  0.3201,    Adjusted R-squared:  0.3181
```

We find that the linear relationship between the percentage yield of the 10-year treasury and the price of the S&P500 is significant. This model has an adjusted R-squared of $R^2_{adj} = 0.3181$, indicating a moderately weak linear association, and we see that it is a negative association by the scatterplot.

Before considering the addition of x_3 to M2, we look into the yield curve spread variate.

The yield curve spread is the difference in basis points between the 10-year treasury yield and the 2-year treasury yield. “The inverted curve reflects bond investors' expectations for a decline in longer-term interest rates, a view typically associated with recessions” (Investopedia, 2024). We note again that a key limitation of our study is that the response variate is the *price* of the S&P500, not the return, and in scenarios such as this one, we tend to care more about relative price changes (returns, or growth values). This limitation is mentioned now only to keep a cautious perspective on our findings.

The model fit using the response and the yield curve spread has $R^2_{adj} = 0.123$, which is quite weak, so we try creating a binary variate which acts as an indicator variable for the event that the yield curve is inverted (the yield curve spread is negative). We define

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1661.28      57.61  28.838  < 2e-16 ***
curve_inv    1002.41     164.35   6.099 2.84e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1009 on 348 degrees of freedom
Multiple R-squared:  0.09657,    Adjusted R-squared:  0.09397
```

$$x_{\text{yield inverted}} = \begin{cases} 1 & \text{if yield curve spread} \leq 0 \\ 0 & \text{if yield curve spread} > 0 \end{cases}$$

for each case $i = 1, \ldots, n$. Fitting the model of the response regressed against the covariate $x_{\text{yield inverted}}$ results in the even weaker value $R^2_{adj} = 0.09397$, and the estimated slope coefficient $\hat{\beta}_1 = 1002.41$. The slope parameter for this model is significant although the adjusted R-squared for this model is lower than the original quantitative model, so we select yield curve spread over the binary $x_{\text{yield inverted}}$ variate. Although, the slope coefficient is the opposite of what we would expect, and we attribute this to fact that our response variate is the raw price value of the S&P500 with values recorded over time (not capturing relative price changes). Hence, this variate is no longer of interest to us and we do not


```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1661.28      57.61  28.838  < 2e-16 ***
curve_inv    1002.41     164.35   6.099 2.84e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1009 on 348 degrees of freedom
Multiple R-squared:  0.09657,    Adjusted R-squared:  0.09397
```

consider further. Note that if the interest of our study was purely to find the best fitting model, we would include this term, since it does not show high multicollinearity with the other covariates in M2 and x_3, and the slope coefficient is significant.

We now consider the addition of the 10-year treasury yield (x_3) to the previous model M2. We calculate the VIF of x_3 with x_1 and x_2 to get $VIF(x_3) = \frac{1}{1-R_3^2} = \frac{1}{1-0.6285} \approx 2.692$, which does not indicate high multicollinearity. We add x_3 to M2 which gives us the model we call M3:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

M3 has a very strong adjusted R-squared of 0.9259, which is greater than that of M2. Moreover, M3 has an AIC value of 4,964.68, while M2 has a higher AIC value of 5,146.931. Both of these metrics show that M3 fits better than M2 even while penalizing the number of parameters in each model.

Returning to the sector weight variates, we now explore a possible interaction effect relationship between the tech sector weight and the junk bonds index, We use the categorical tech level variate instead of the quantitative tech weight for simplicity and clearer interpretation. The following plot shows the response plotted against the junk bonds price where each data point is colour coded based on the $x_{\text{tech level}}$ value corresponding to that case. This plot also includes the fitted lines for each level of $x_{\text{tech level}}$. We observe that as the tech level changes, the slope of junk bonds price vs S&P500 price changes as well. This indicates that there is an interaction effect, but we're unsure if main effects are present.

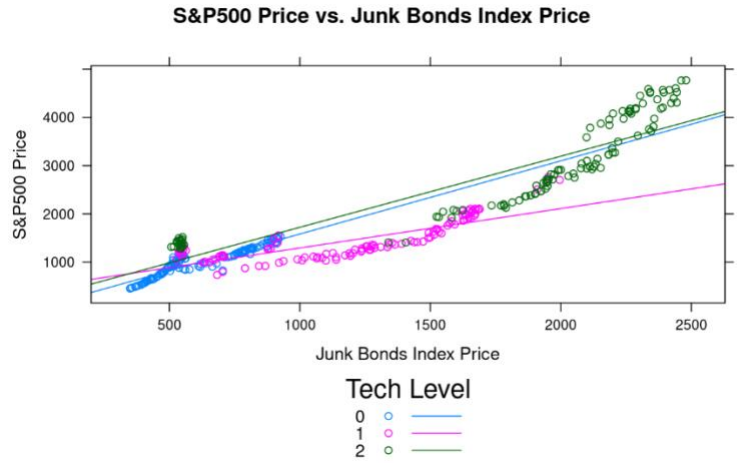


Figure 13

We fit the following multiple linear regression model which includes an interaction term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_{\text{tech level}} + \beta_3 x_1 x_{\text{tech level}} + \epsilon$$

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          743.93717   74.30767  10.012  < 2e-16 ***
x1                     0.35319    0.09507   3.715 0.000237 ***
sectors$tech_cat    -314.12496   55.22332  -5.688 2.73e-08 ***
x1:sectors$tech_cat    0.58412    0.05166  11.307  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 360.5 on 346 degrees of freedom
Multiple R-squared:  0.8855,    Adjusted R-squared:  0.8845
```

The t-test for the slope of the interaction term results in a significant p-value. The model summary from R also indicates that the junk bonds index and tech level are significant to the model, which tells us that the main effects are present. This model has a high adjusted R-squared of 0.8845. For interpreting the estimated slope coefficients, we note the following:

$$\text{Assuming } y = \beta_0 + \beta_1 x_1 + \beta_2 x_{\text{tech level}} + \beta_3 x_1 x_{\text{tech level}},$$

$$\Rightarrow \begin{cases} \dfrac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_{\text{tech level}} \\ \dfrac{\partial y}{\partial x_{\text{tech level}}} = \beta_1 + \beta_3 x_1 \end{cases}$$

The estimated slope coefficient of the interaction term represents the added change in the effect of tech weight on the S&P500 price as the junk bonds index price rises, and vice versa. We refer to this interaction model as Mi. The slope coefficients are significant for both Mi and M2, so to compare the two models we provide the following table which summarizes model selection criteria.

Model	R^2_{adj}	$PRESS$	AIC	BIC
M2	0.8750002	49,661,047	5,146.931	5,162.363
Mi	0.8844586	46,153,165	5,120.382	5,139.672

The higher adjusted R-squared, lower PRESS statistic, lower AIC, and lower BIC for Mi compared to M2 all suggest that our interaction model Mi is outperforming M2 on the four criteria selected.
Now we try to improve the interaction model Mi by adding x_3 to the model. We fit the model that we call M3i:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_{\text{tech level}} + \beta_3 x_3 + +\beta_4 x_1 x_{\text{tech level}} + \epsilon$$

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -758.23253  155.19607  -4.886 1.58e-06 ***
x1                    1.17538    0.11310  10.393  < 2e-16 ***
sectors$tech_cat   -147.17086   50.48170  -2.915  0.00379 **
x3                  197.89726   18.59235  10.644  < 2e-16 ***
x1:sectors$tech_cat   0.34115    0.05036   6.774 5.40e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 313.2 on 345 degrees of freedom
Multiple R-squared:  0.9138,    Adjusted R-squared:  0.9128
```

We find that all the covariates are s ignificant and this model has $R^2_{adj} \approx 0.913$. In this model, $VIF(x_3) \approx 3.357$, indicating no h igh multicollinearity.

Now, we finally fit a model using the industrial weight and compare it to our other models. Let x_4 represent the industrials weight variate. We fit the following model which we call M4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_3 + \epsilon$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1928.98278  237.69719   8.115 8.56e-15 ***
x1             1.56665    0.04323  36.237  < 2e-16 ***
x4          -288.43078   18.19002 -15.857  < 2e-16 ***
x3           269.28723   13.64622  19.733  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 259 on 346 degrees of freedom
Multiple R-squared:  0.9409,    Adjusted R-squared:  0.9403
F-statistic:  1835 on 3 and 346 DF,  p-value: < 2.2e-16
```

The t-tests for each slope coefficient as well as the F-test are significant, and the model has $R^2_{adj} \approx 0.94$. Thus, M4 shows a very strong, and significant, linear association between the covariates in the model and the response. To be wary of possible multicollinearity which was introduced when we substituted x_4 for x_2 in M3 to get M4, we calculate $VIF(x_4) \approx 2.644$, which indicates no high multicollinearity with the other covariates. Due to the strong model performance and significance, we consider M4 in our final model selection.

We are interested in whether any subset of M4 outperforms the M4, so we consider Mallow's C_p criteria on M4, and we find that the C_p values of all subset models are very far from $p+1$ (where $p=3$ for M4), which indicates that no subset model of M4 is performing better than M4. The R output from using the leaps library in R to calculate C_p values is shown on the left.

```
$which
      1     2     3
1  TRUE FALSE FALSE
1 FALSE FALSE  TRUE
1 FALSE  TRUE FALSE
2  TRUE  TRUE FALSE
2  TRUE FALSE  TRUE
2 FALSE  TRUE  TRUE
3  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"

$size
[1] 2 2 2 3 3 3 4

$Cp
[1]  620.1441 1319.8078 3631.4235  253.4299  391.4099 1315.0906    4.0000
```

As another way of checking whether we inefficiently included any unnecessary variates, we now run stepwise automatic model selection for the variates in M4. The model returned by the algorithm (implemented by R) contains all the same variates, as can be seen by running the code in the appendix. Performing the same task with the variates in M3, we are also returned with all the same variates.

Results

The three models we are left with for our final model selection are M3, M3i, and M4. The table below summarizes how these three models compare against one another.

Model	R^2_{adj}	$PRESS$	AIC	BIC
M3	0.9259481	29633665	4964.68	4983.97
M3i	0.9127695	34897984	5022.993	5046.141
M4	0.9403417	23792513	4889.033	4908.323

AIC and BIC are essentially measuring the same criteria, just penalizing the number of parameters in the model slightly differently. We observe that M4 achieves the highest and thus the strongest adjusted R-squared. M4 also has the lowest values for it's PRESS statistic, AIC, and BIC. This tells us that M4 performs the best when considering both model fit and the number of parameters. The PRESS statistic being the lowest among the three candidate models suggests that M4 is not overfitting the data relative to the other models (or, it is overfitting the least compared to the other models). Thus, M4 is our choice for the best final model.

To summarise our findings, we saw that there is a strong significant linear association between the junk bonds index price and the S&P500 price (a positive association), we saw moderately strong significant linear associations between the S&P500 price and sector weights (positive for tech, negative for industrial), and we saw a moderately strong negative association between the S&P500 price and the yield of the 10-year treasury (which was the strongest relationship out of all the macroeconomic variables). We also saw a significant interaction relationship between the sector weights. We revisit the practical interpretation of our final model, M4, after checking the model assumptions.

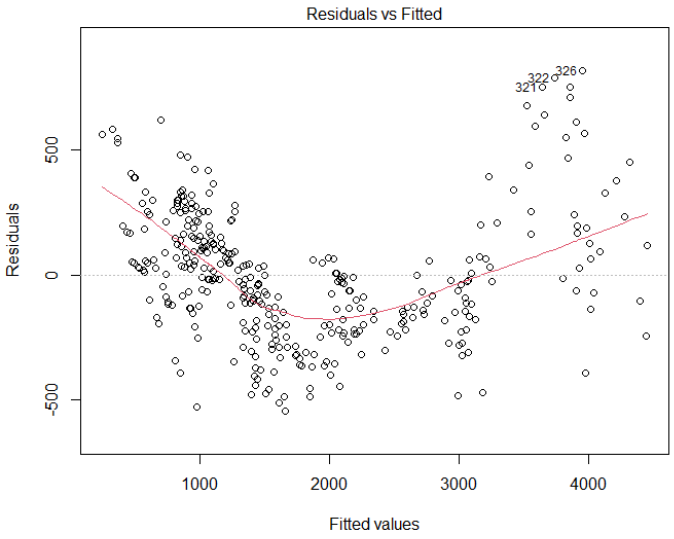
M4 assumes that:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{4i} + \beta_3 x_{3i} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \text{ indepdently for } i = 1, \dots, n$$

Put in words, we make the assumptions that the residuals are independent, Normally distributed, have a mean of zero, and all have the same constant variance.

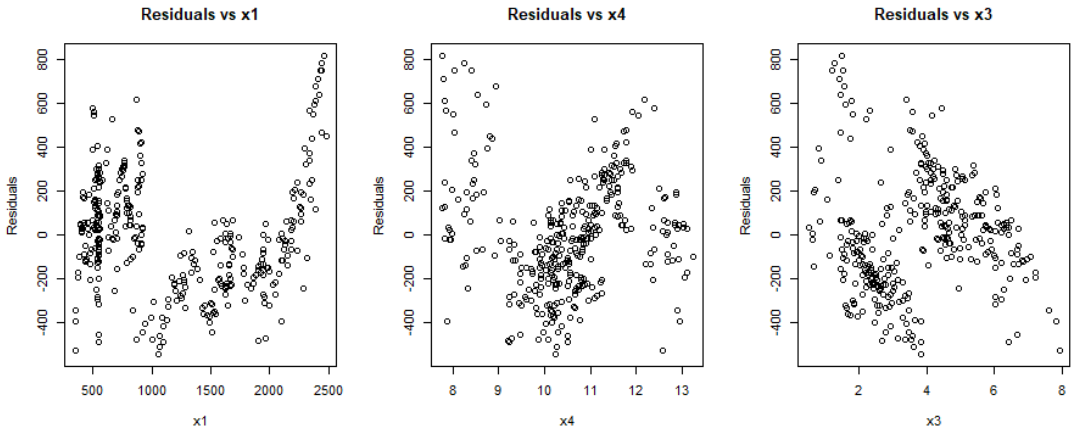
Several plots related to the observed residuals e_i are presented below as we discuss possible model assumption violations and potential outliers.

We begin by inspecting the residuals vs fitted values plot. We observe roughly the same amount of points scattered above and below the horizontal line at zero, suggesting that the mean of zero assumption is not violated. However, we do not observe points which appear very randomly scattered.

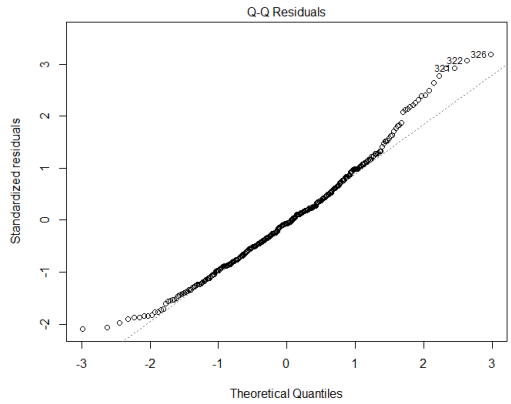


We notice a systematic quadratic trend, suggesting that the independence assumption is violated. Violation of the independence assumption makes sense in the context of the data, since some of our variates tend to follow trends over time.

Also, this plot suggests that the relationship may not be not linear (which aligns with the quadratic looking scatterplot of y vs x_1 at the start of our main analysis). To investigate this further we look at the plots of the residuals vs each covariate. These plots are shown below.



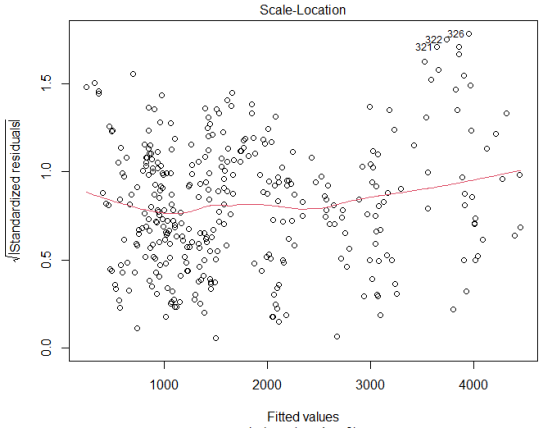
The plot for x_1 shows somewhat of a quadratic trend, the plot for x_4 shows somewhat of a positive linear trend, and the plot for x_3 shows somewhat of a negative linear trend. Therefore, we conclude that the linearity assumption appears violated.

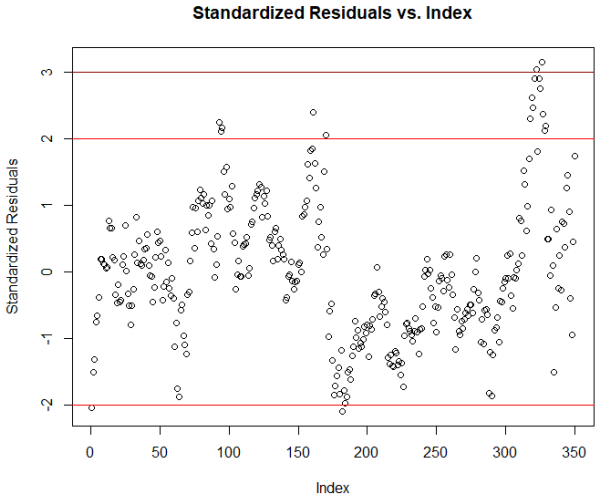


Next, we consider the Normal QQ plot. The points on this plot lie reasonably along a straight line, suggesting that the Normality assumption is not violated.

Moving on to the scale-location plot, we observe points that appear randomly scattered with no systematic trend, and a horizontal line can be drawn along the centre of the points. This suggests that the constant variance assumption is not violated.

The standardized residuals vs index plot shown below has two red lines for the values ± 2, and dark red lines for ± 3. Most of the points are contained within the horizontal bands ± 2 showing a random scatter around 0, again suggesting that the mean of zero assumption is not violated.

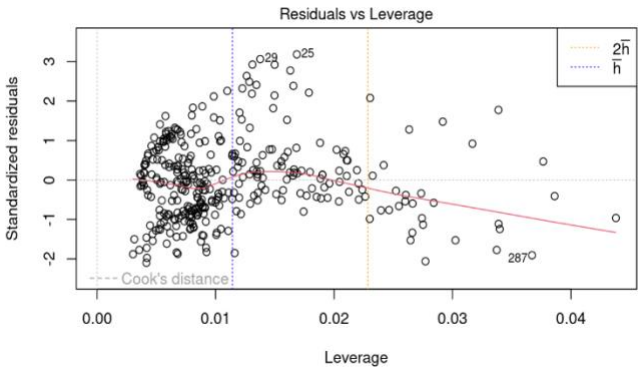




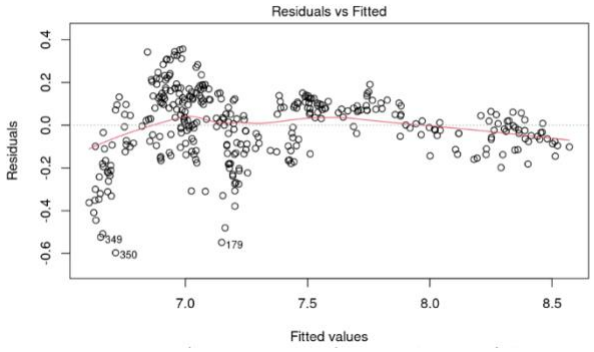
To summarise our findings for model checking (for our final model M4), we conclude that all assumptions seem reasonably satisfied except for the violated independence and linearity assumptions.

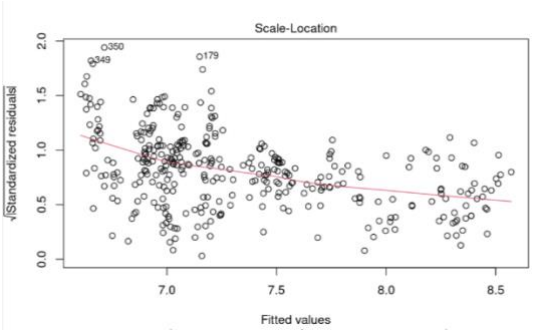
We observe from the residuals vs leverage plot that there are cases with leverage $> 2\bar{h}$, which are deemed high leverage cases. However, these cases are still within Cook's distance so there are no x-direction outliers.

Now, we perform the natural logarithm transformation on the response with the hope of fixing the independence violation, and present the diagnostic plots for this transformed model.

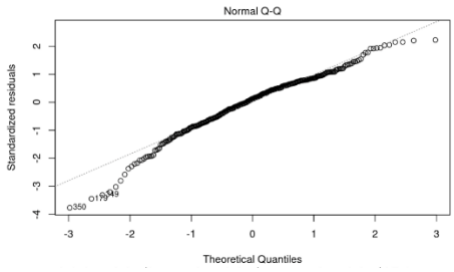


We do not observe the same quadratic trend in the residuals vs fitted plot, however we do observe fanning in, which suggests that the constant variance assumption is violated. The points seem to still show systematic trends, so we conclude that the independence assumption is still violated, and the transformation did not work.
The scale location plot shows a similar fanning in pattern, suggesting that the constant variance assumption is violated in the transformed model.





The normality and mean of zero assumption seem to still be satisfied in the transformed model. The transformed model seems to present a new violation of the constant variance assumption, and seems to retain the independence violation.



Now, we interpret the model parameters. As mentioned before, we came to the model with the following form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{4i} + \beta_3 x_{3i} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \text{ indepdently for } i = 1, \dots, n$$

where y is the S&P 500 price in \$USD, x_1 is the price of the junk bonds (high yield corporate bonds) index in \$USD, x_3 is the yield of the 10-year treasury bond (in percentage points), and x_4 is the market-weighted percentage of firms in the S&P500 which are in the industrial sector.

$\hat{\beta}_1 \approx 1.57$ suggests that a \$1 (USD) increase in the junk bonds index price is associated with a mean increase in the S&P500 price of \$1.57 (USD), holding x_4 and x_3 constant. Since this value is positive, it suggests that the prices of high yield U.S. corporate bonds and the prices of stocks in the S&P500 move in the same direction. This also suggests that stock market price movements (as measured by the S&P500 price) tend to be on average roughly 1.57 times the size of the high yield corporate bond market price movements (as measured by the junk bonds index price).

$\hat{\beta}_2 \approx -288.43$ suggests that a 1 percentage point increase in the weight of industrial firms in the S&P500 is associated with a mean *decrease* in the S&P500 price of roughly \$288.43 (USD) holding x_1 and x_3 constant. This suggests that perhaps the industrial sector is dying out.

$\hat{\beta}_3 \approx 269.29$ suggests that a 1 percentage point increase in the 10-year treasury yield is associated with a mean increase in the S&P500 price of roughly \$269.29 (USD), holding x_1 and x_4 constant. This is may seem contradictory, but recall that the response variate is simply the price of the S&P500, not the relative change in price. There is also the concern that our data is time-dependent, so there is an important variable (time) which is not included in our model.

Note that when we say "1 percentage point increase", we mean that the value increases from $k\%$ to $(k+1)\%$, we *do not* mean 1% relative growth.

Finally, we provide a prediction example using our model. We use the real values for the covariates at the end of October 2024, and note that we did not include these data in our sample. The values can be found in the appendix. Using these values, we get the point estimate for the mean S&P500 price $4798.97 USD, and a 95% prediction interval for this same value is ($4276.91 USD, $5321.04 USD), rounded to the nearest $0.01. For comparison, the true value that we attempt to predict here is 5705.45. One may interpret this prediction error as a sign that the U.S. stock market is overvalued relative to the 10-year U.S. treasury yield, the weight of industrial firms in the S&P500, and the price of the U.S. corporate high yield bonds index, although it's important to note that our model does not satisfy all the model assumptions (namely independence), so any further application should be preceded by further investigation.

Limitations of Study & Conclusion

As noted several times earlier, the variates in our dataset are time-dependant, introducing non-stationarity, which can severely hinder the validity of the model assumptions (as we saw with the independence violation). Despite an attempt to address this issue with a transformation, the assumption remains violated. The final model fails to meet all the model assumptions, which indicates that the model might not be practical in real life, and that further investigation must be carried out before any application, which should be done with caution. Furthermore, our data was aggregated using the last value of each month, and one may obtain different results when using different aggregation methods. Lastly, we note that our selected sample ranges from monthly time points between 1994 and 2023, so a sample from a different set of time points may lead to different results.

References

Liberto, Daniel. "Inverted Yield Curve: Definition, What It Can Tell Investors, and Examples." Investopedia. Accessed November 29, 2024. https://www.investopedia.com/terms/i/invertedyieldcurve.asp#:~:text=An%20inverted%20yield%20curve%20is,view%20typically%20associated%20with%20recessions.

Appendix

Correlation Matrix (Entire Dataset)

	sp500_price	fed	yield_curve_spread	junkbonds_price	tbond_2y	tbond_10y	tbond_5y	tech_weight	industrials_weight
sp500_price	1.0000000	-0.2558787	-0.3506643	0.9136995	-0.2875669	-0.5657711	-0.4276242	0.7889169	-0.8457221
fed	-0.2558787	1.0000000	-0.7061849	-0.5214371	0.9611967	0.8408172	0.9119119	-0.2567922	0.3774558
yield_curve_spread	-0.3506643	-0.7061849	1.0000000	-0.1200240	-0.6716429	-0.3049555	-0.4952758	-0.2462089	0.2049458
junkbonds_price	0.9136995	-0.5214371	-0.1200240	1.0000000	-0.5607585	-0.7882783	-0.6846232	0.7069426	-0.7932552
tbond_2y	-0.2875669	0.9611967	-0.6716429	-0.5607585	1.0000000	0.9102189	0.9748200	-0.2802686	0.4078745
tbond_10y	-0.5657711	0.8408172	-0.3049555	-0.7882783	0.9102189	1.0000000	0.9765688	-0.4977855	0.6381311
tbond_5y	-0.4276242	0.9119119	-0.4952758	-0.6846232	0.9748200	0.9765688	1.0000000	-0.3903819	0.5281849
tech_weight	0.7889169	-0.2567922	-0.2462089	0.7069426	-0.2802686	-0.4977855	-0.3903819	1.0000000	-0.8928991
industrials_weight	-0.8457221	0.3774558	0.2049458	-0.7932552	0.4078745	0.6381311	0.5281849	-0.8928991	1.0000000

Correlation Matrix of Macroeconomic Variates &

	twoyear	fiveyear	tenyear	fed	sp500
twoyear	1.0000000	0.9748200	0.9102189	0.9611967	-0.2875669
fiveyear	0.9748200	1.0000000	0.9765688	0.9119119	-0.4276242
tenyear	0.9102189	0.9765688	1.0000000	0.8408172	-0.5657711
fed	0.9611967	0.9119119	0.8408172	1.0000000	-0.2558787
sp500	-0.2875669	-0.4276242	-0.5657711	-0.2558787	1.0000000

Out of Sample Values Used for Prediction Example

Date	Junk Bond Index Price	10 Year Treasury Yield	Industrial Sector Weight
10/31/2024	2663.98	4.285	8.52

```
data <- read.csv("C:/Users/aamelaci/Downloads/Project Data 11.28.csv")
summary(data)
data <- as.data.frame(lapply(data, FUN = rev)) # reverse order s.t.
time increasing
data <- data[,-1] # removing date col

### Correlation Matrix
cor(data)

### Investigation of Response Variate

y <- data$sp500_price

summary(y)
hist(y, breaks = 30, probability = TRUE,
     main = "Histogram of S&P500 Price", xlab = "S&P500 Price")
curve(expr = dnorm(x, mean = mean(y), sd = sd(y)),
      add = TRUE, col = "red", lwd = 2)
legend("topright", legend = c("Normal Density"),
       col = c("red"), lty = c(1), lwd = c(2), bty = 'n')
boxplot(y, ylab = "S&P500 Price", main = "Boxplot of S&P500 Price")
plot(y, xlab = "Time Index", ylab = "S&P500 Price",
     main = "Line chart of S&P500 Price", type = "l", lwd = 2)

### Individual Model Selection

## Junk Bonds model
x1 <- data$junkbonds_price
m1 <- lm(y ~ x1)
summary(m1)
```

```
coefs_m1 <- coefficients(m1)
beta0_m1 <- coefs_m1[1]
beta1_m1 <- coefs_m1[2]
plot(x = x1, y = y, xlab = "Junk Bonds Index Price (x1)",
     ylab = "S&P500 Price (y)", main = "S&P500 Price vs. Junk Bonds
Index Price")
curve(expr = beta0_m1 + beta1_m1*x, add = TRUE, col = "red", lwd = 2)
legend("topleft", legend = c(bquote(hat(mu)(x))),
       col = c("red"), lty = c(1), lwd = c(2), bty = 'n')

## Sector Weights
sectors <- data.frame(tech = data$tech_weight,
                      indust = data$industrials_weight)
cor(sectors) # strong negative corr between tech and industrials
summary(sectors)
x2 <- sectors$tech
plot(x = x2, y = y, main = "S&P500 Price vs. Tech Weight",
     xlab = "Tech Weight", ylab = "S&P500 Price (y)")
tech_model <- lm(y ~ x2)
summary(tech_model)
beta0_tech_model <- coefficients(tech_model)[1]
beta1_tech_model <- coefficients(tech_model)[2]
curve(expr = beta0_tech_model + beta1_tech_model*x,
      add = TRUE, col = "red", lwd = 2)
legend("topleft", legend = c(bquote(hat(mu)(x))),
       col = c("red"), lty = c(1), lwd = c(2), bty = 'n')
cor(sectors$tech, data)

# trying categorical with tech weight
hist(x2, probability = TRUE, main = "Histogram of Tech Weight",
     xlab = bquote("Tech Weight "(x[2])))
tech_quantiles <- quantile(x2, probs = c(1/3, 2/3))
abline(v = tech_quantiles[1], col = "green", lwd = 2)
abline(v = tech_quantiles[2], col = "red", lwd = 2)
legend("topright", legend = c(bquote(Q[x[2]](1/3)),
bquote(Q[x[2]](2/3))),
       col = c("green", "red"), lty = c(1,1), lwd = c(2,2), bty='n')
# defining categorical variate
sectors$tech_cat <- NULL
sectors$tech_cat[x2 < tech_quantiles[1]] <- 0
l1_cond <- tech_quantiles[1] <= x2
l1_cond <- l1_cond & (x2 <= tech_quantiles[2])
sectors$tech_cat[l1_cond] <- 1
sectors$tech_cat[tech > tech_quantiles[2]] <- 2
# plotting categorical variate
plot(x = sectors$tech_cat, y = y, main = "S&P500 Price vs. Tech
Level",
     xlab = bquote("Tech Level "(x["tech level"])),
     ylab = "S&P500 Price (y)")
tech_cat_model <- lm(y ~ sectors$tech_cat)
summary(tech_cat_model)
```

```
curve(expr = coefficients(tech_cat_model)[1] +
coefficients(tech_cat_model)[2]*x,
      add = TRUE, col = "red", lwd = 2)
legend("topleft", legend = c(bquote(hat(mu)(x))),
       col = c("red"), lty = c(1), lwd = c(2), bty = 'n')

# comparing tech level with tech weight
aic_cat <- AIC(tech_cat_model)
aic_reg <- AIC(tech_model)
data.frame(Covariate = c("x2", "tech level"),
           AIC = c(aic_reg, aic_cat))


# tech weight VIF with industrial weight
summary(lm(data$tech_weight ~ data$industrials_weight))
1/(1-0.7973)

## M2 (junk bonds + tech weight)
m2 <- lm(y ~ x1 + tech)
summary(m2)

## Macro: FED, T-Bonds, yield curve spread
tbonds <- data.frame(twoyear = data$tbond_2y,
                     fiveyear = data$tbond_5y,
                     tenyear = data$tbond_10y)
macro <- cbind(tbonds, data.frame(fed=data$fed, sp500=y))
cor(macro) # 10y tbond has strongest (negative) correlation with
response
x3 <- tbonds$tenyear
bond_model <- lm(y ~ x3)
plot(x = x3, y = y, main = "S&P500 Price vs. 10y Treasury Yield",
     xlab = "10y Treasury Bond Yield (x3)", ylab = "S&P500 Price (y)")
summary(bond_model)
curve(expr = coefficients(bond_model)[1] +
coefficients(bond_model)[2]*x,
      add = TRUE, col = "red", lwd = 2)

cor(macro$tenyear, data$yield_curve_spread)
cor(data$yield_curve_spread, y)
plot(x = data$yield_curve_spread, y = y)
summary(lm(y~data$yield_curve_spread))
curve_inv <- NULL
curve_inv[data$yield_curve_spread <= 0] <- 1
curve_inv[data$yield_curve_spread > 0] <- 0
plot(x = curve_inv, y = y)
mean(y[curve_inv == 1])
mean(y[curve_inv == 0])
summary(lm(y ~ curve_inv))
summary(lm(x3~x1+tech))
1/(1-0.6285) # vif calculation

xyplot(data$sp500_price ~ data$junkbonds_price,
```

```
       groups = data$tech_cat,


       auto.key = list(space = "bottom",
                       points = TRUE,
                       lines = TRUE,
                       columns = 1,
                       title ="Tech Level"),

       xlab="Junk Bonds Index Price",
       ylab="S&P500 Price",
       main = "S&P500 Price vs. Junk Bonds Index Price",

       type=c("p","r"))

## Interaction Model
inter_model <- lm(y ~ x1 * sectors$tech_cat)
summary(inter_model)

get_four_criteria <- function(model) {
  # Get the summary of the model
  model_summary <- summary(model)
  result <- list(
    adjusted_r_squared = model_summary$adj.r.squared,
    press_statistic = press(model),
    AIC = AIC(model),
    BIC = BIC(model)
  )  # Return the adjusted R-squared value
  return(result)
}

plot(m4, which = 5)
abline(v = 4/350 * 2,lty = 3, col = "orange")
abline(v = 4/350 ,lty = 3, col = "blue")
legend("topright", c(expression(2*bar(h)), expression(bar(h))), col =
c("orange", "blue"), lty = c(3,3))

prediction_data =
read.csv("/home/paluo/Code/uwaterloo/stat371/Results/prediction2024.cs
v")
m4 = lm(sp500_price ~ junkbonds_price + tbond_10y +industrials_weight,
data = data)
pre_prediction =  prediction_data[, c("junkbonds_price",
                                      "tbond_10y",
                                      "industrials_weight",
                                      "sp500_price")]
prediction = predict(m4, newdata =  pre_prediction, interval =
"prediction", level = 0.95)
print(prediction)

## M3i
```

```
m3i <- lm(y~x1+sectors$tech_cat+x3+x1*sectors$tech_cat)
summary(m3i)

## M3
m3 <- lm(y~x1+tech+x3)
summary(m3)
AIC(m3)
AIC(m2)

# PRESS Function
press <- function(model, stat=TRUE) {
  PRESSreds<-residuals(model)/(1-lm.influence(model)$hat)
  PRESSstat<-sum(PRESSreds^2)
  par(mfrow=c(1,1))
  plot(x = PRESSreds, y = resid(model), xlab="PRESS Residuals",
ylab="Residuals")
  if (stat) {
    PRESSstat
  } else {
    PRESSreds
  }
}


### Final Best Model

x4 <- data$industrials_weight
m4 <- lm(y ~ x1 + x4 + x3)
summary(m4)

# Cp Criteria
leaps(m4, data$sp500_price)

# Auto Selection
m4_automatic_selection= step(lm(data$sp500_price ~ 1), scope = ~
                             data$junkbonds_price +
                             data$tbond_10y +
                             data$industrials_weight , direction =
"both")

m3_automatic_selection= step(lm(data$sp500_price ~ 1), scope = ~
                             data$junkbonds_price +
                             data$tbond_10y +
                             data$tech_weight , direction = "both")


plot(m4)
se <- 259
std_resids <- residuals(m4)/se
plot(std_resids, main = "Standardized Residuals vs. Index",
     ylab = "Standardized Residuals")
abline(h = -2, col = "red")
```

```
abline(h = 2, col = "red")
abline(h = 3, col = "darkred")
which(std_resids > 3)
par(mfrow = c(1,3))
plot(x = x1, residuals(m4), main="Residuals vs x1",
     ylab = "Residuals")
plot(x = x4, residuals(m4), main="Residuals vs x4",
     ylab = "Residuals")
plot(x = x3, residuals(m4), main="Residuals vs x3",
     ylab = "Residuals")

# log transform response
logy <- log(y)
m4log <- lm(logy ~ x1 + x4 + x3)
summary(m4log)
plot(m4log)
press(m4log)
```