Linear Regression of Business News Headlines Sentiment and Stock Market Performance

Jordan Tapke

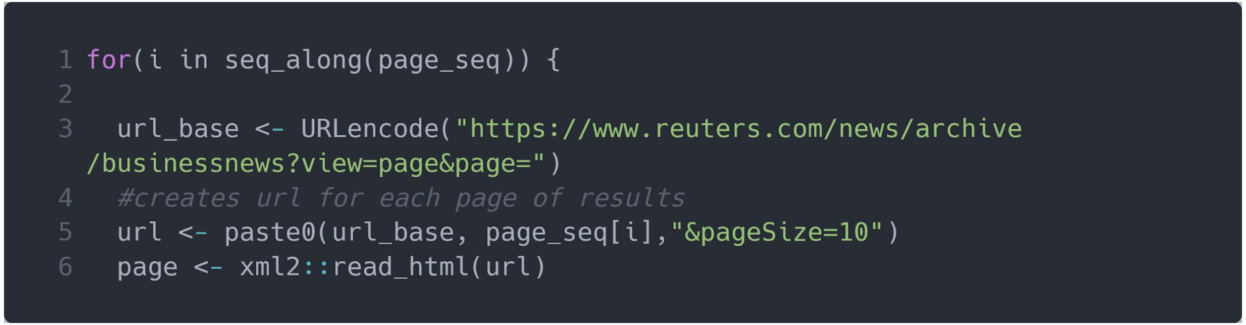
Fall 2020

Introduction

- Goal of project is to model the relationship between business news and stock performance via a linear regression
- **Independent Variable:** Average daily sentiment analysis score on business news headlines
  - Scraped from Reuters News
- **Dependent Variable:** The percent difference in closing and opening price of various index ETFs
  - Used quantmod library which gets finance data from Yahoo Finance

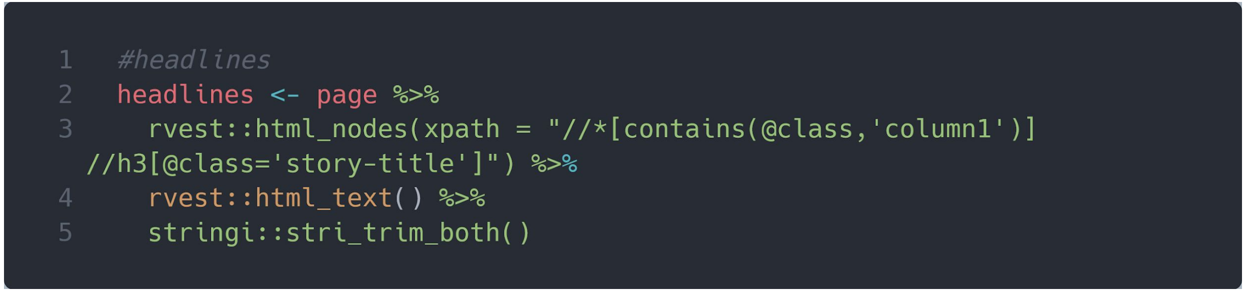
News Headlines

- *seq_along* allows for page # to increase by set increments



```r
for(i in seq_along(page_seq)) {

  url_base <- URLencode("https://www.reuters.com/news/archive/businessnews?view=page&page=")
  #creates url for each page of results
  url <- paste0(url_base, page_seq[i],"&pageSize=10")
  page <- xml2::read_html(url)
```

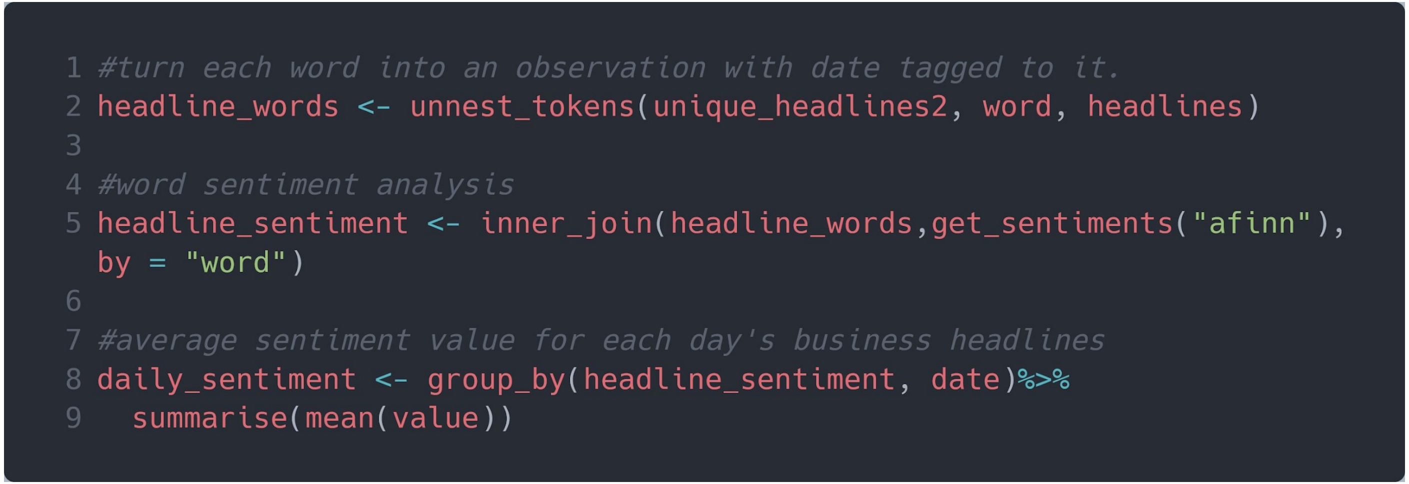
- Example of using xpath to identify correct nodes to pull data from



```r
  #headlines
  headlines <- page %>%
    rvest::html_nodes(xpath = "//*[contains(@class,'column1')]//h3[@class='story-title']") %>%
    rvest::html_text() %>%
    stringi::stri_trim_both()
```

Sentiment Analysis

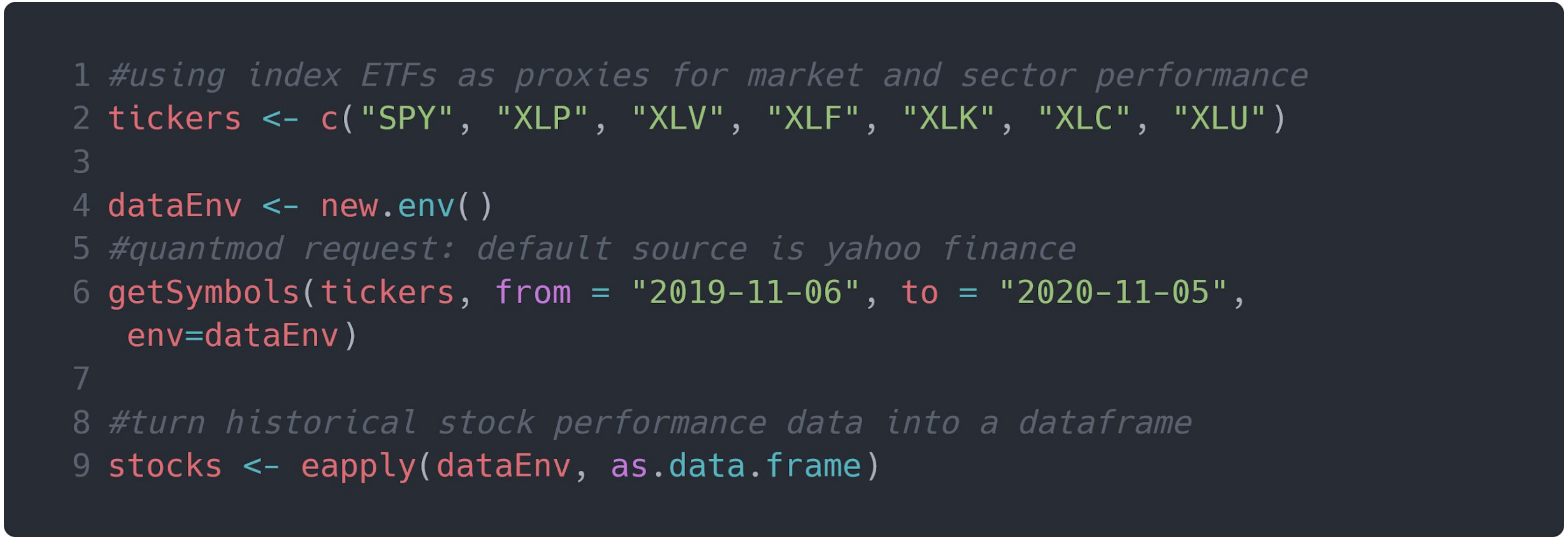
- *afinn* sentiment lexicon rates word from -5 (negative) to 5 (positive)



```r
#turn each word into an observation with date tagged to it.
headline_words <- unnest_tokens(unique_headlines2, word, headlines)

#word sentiment analysis
headline_sentiment <- inner_join(headline_words,get_sentiments("afinn"), by = "word")

#average sentiment value for each day's business headlines
daily_sentiment <- group_by(headline_sentiment, date)%>%
  summarise(mean(value))
```

Historical Stock Data

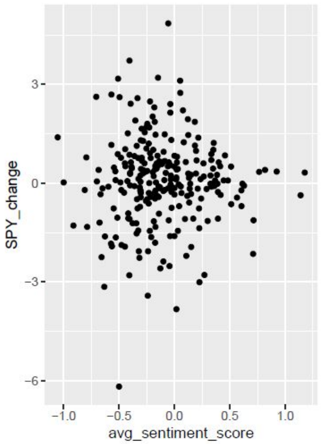
- Use ETFs as proxies for S&P 500 and sectors of the market.
- Output is multiple xts objects

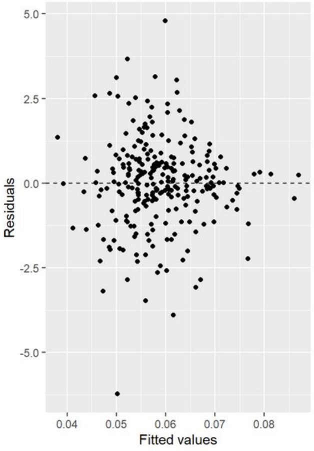


```r
#using index ETFs as proxies for market and sector performance
tickers <- c("SPY", "XLP", "XLV", "XLF", "XLK", "XLC", "XLU")

dataEnv <- new.env()
#quantmod request: default source is yahoo finance
getSymbols(tickers, from = "2019-11-06", to = "2020-11-05",
 env=dataEnv)

#turn historical stock performance data into a dataframe
stocks <- eapply(dataEnv, as.data.frame)
```

Model Results

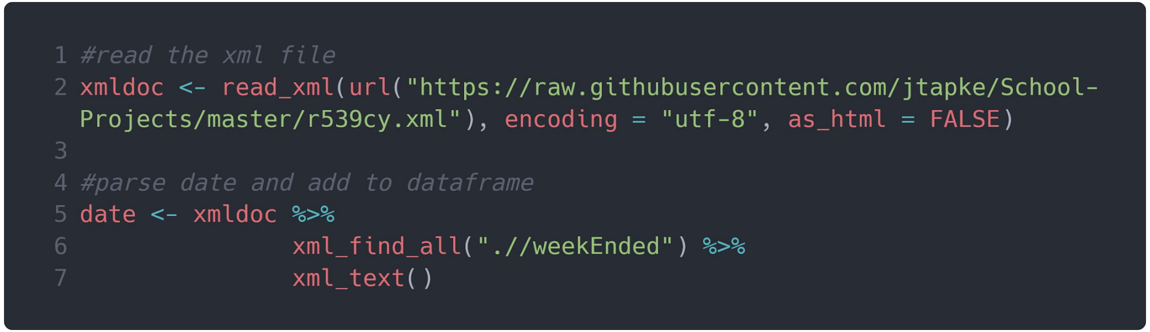
```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.06139  0.07871   0.780     0.436
avg_sentiment_score  0.27143        0.21004   1.292     0.197

Residual standard error: 1.199 on 250 degrees of freedom
Multiple R-squared:  0.006636,    Adjusted R-squared:  0.002662
F-statistic:  1.67 on 1 and 250 DF,  p-value: 0.1975
```





Addition of Unemployment Data

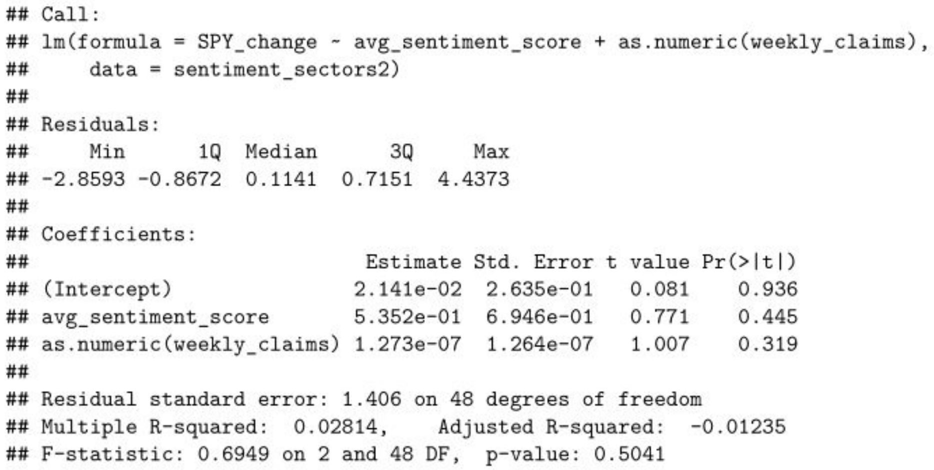
- Add another independent variable to data to see if the relationship improves in a Multiple Linear Regression.
- Use *xml2* library to read in xml data, then parse date and weekly unemployment claims



```r
#read the xml file
xmldoc <- read_xml(url("https://raw.githubusercontent.com/jtapke/School-Projects/master/r539cy.xml"), encoding = "utf-8", as_html = FALSE)

#parse date and add to dataframe
date <- xmldoc %>%
               xml_find_all(".//weekEnded") %>%
               xml_text()
```

Addition of Unemployment Data

- Regression resulted in higher p-values and higher R-squared value but, still had poor performance



```
## Call:
## lm(formula = SPY_change ~ avg_sentiment_score + as.numeric(weekly_claims),
##     data = sentiment_sectors2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8593 -0.8672  0.1141  0.7151  4.4373
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.141e-02  2.635e-01   0.081    0.936
## avg_sentiment_score       5.352e-01  6.946e-01   0.771    0.445
## as.numeric(weekly_claims) 1.273e-07  1.264e-07   1.007    0.319
##
## Residual standard error: 1.406 on 48 degrees of freedom
## Multiple R-squared:  0.02814,    Adjusted R-squared:  -0.01235
## F-statistic: 0.6949 on 2 and 48 DF,  p-value: 0.5041
```

Conclusion

- Issues with the Model:
  - More variables needed
  - Linear regression is most likely not the correct model for such a complicated dependent variable
- Proxies are not only a good way to get data that is not available but, also to find aggregated versions of the data.