DATA 606 Data Project Proposal

Paul Britton

Data Preparation

We'll collect the data that we need from the internet (yahoo!finance) using the "quantmod" package:

```
startDate <- as.Date("2007-01-01")
endDate <- as.Date("2018-03-29")
tickers <- c("spy")

data <- tail(as.data.frame(getSymbols(tickers,
                                      env=NULL,
                                      src="yahoo",
                                      from=startDate,
                                      to=endDate)),-9)

write.csv(data,file = "SPY_data.csv")
```

Now we'll take a look and see what we've collected:

```
kable(head(data))
```

	SPY.Open	SPY.High	SPY.Low	SPY.Close	SPY.Volume	SPY.Adjusted
2007-01-17	142.85	143.46	142.73	143.02	50241400	113.4800
2007-01-18	143.17	143.26	142.31	142.54	68177300	113.0992
2007-01-19	142.54	143.10	142.46	142.82	56973000	113.3213
2007-01-22	143.07	143.10	141.93	142.38	60253600	112.9722
2007-01-23	142.26	143.08	142.06	142.80	54064400	113.3055
2007-01-24	142.97	143.98	142.91	143.95	55834700	114.2180

The data looks good. Note that I've frozen a copy on my github here

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Is the change (i.e. "higher" or "lower") of 3-month daily standard deviation of the S&P 500 stock index predictive of price-return for that index over the subsequent 1-month period?

Cases

What are the cases, and how many are there?

Each pairing of "3-month standard deviation" and "subsequent 1-month performance" represents a case. For simplicity, we'll assume that 1 month == 20 business days and thus, our standard-deviation lookback will be 60 periods.

Assuming that the inclusion of overlapping periods will not be allowed, we will have 47 periods. We need to allow for the fact that we need a 3-month lookback, and a one-month "out-sample" for each case, thus we will end up with 46 cases.

Data collection

Describe the method of data collection.

The data is collected in real-time by the NYSE and is downsampled and cleaned by CSI data, a vendor/provider of financial market data.

Type of study

What type of study is this (observational/experiment)?

This is an obvervational study

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

Collected from yahoo!finance and saved to github using the code in the "collectData" code chunk above.

Response

What is the response variable, and what type is it (numerical/categorical)?

1 month stock returns

Explanatory

What is the explanatory variable, and what type is it (numerical/categorival)?

3-month standard deviation of price returns - numerical.

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you're comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

The first thing that we'll do is compute percentage returns:

```
pct.returns <- ((data["SPY.Close"] / shift(data["SPY.Close"],1)) -1)*100
```

Then we'll compute the standard deviation of returns on a rolling 60-day basis, and extract non-overlapping periods for both standard-deviation and subsequent returns.

```
stdev <- rollapplyr(pct.returns,60,sd,fill=0)
stdev <- stdev * 16 #scale to "annual stdev #s"

#use only non-overlapping periods
df = as.data.frame(stdev[seq(61, nrow(stdev), 60), ])
df$returns = pct.returns[seq(81, nrow(pct.returns), 60), ]

#rename the cols
colnames(df) <- c("sd","r")
```

Now we'll look at the summary statistics and histograms for each variable:

```
describe(df$sd)
```

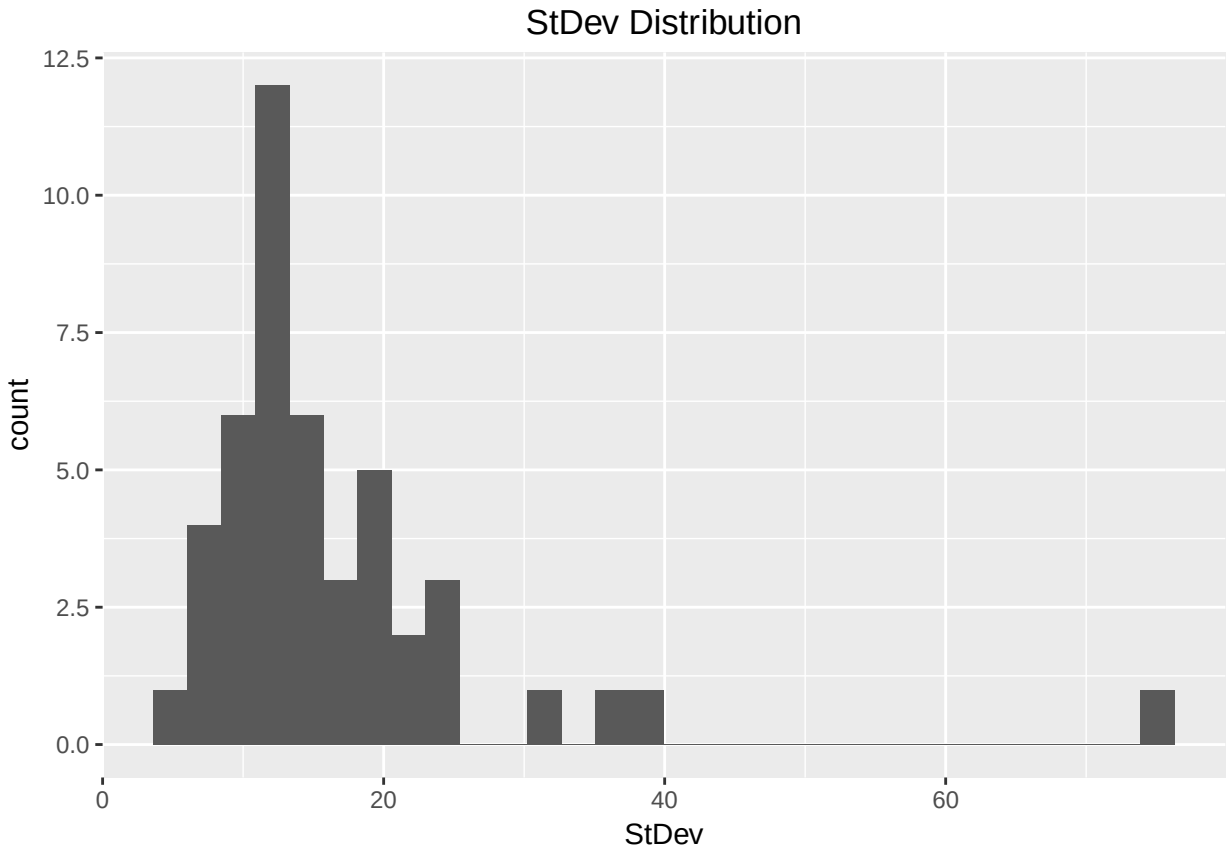
```
##    vars  n  mean    sd median trimmed  mad  min   max range skew kurtosis
## X1    1 46 16.79 11.43  13.36    14.8 5.58 5.63 75.85 70.22  3.2    13.09
##      se
## X1 1.69
```

```
describe(df$r)
```

```
##    vars  n mean   sd median trimmed  mad   min  max range  skew kurtosis
## X1    1 46 0.07 1.08   0.04     0.1 0.64 -2.73 3.29  6.02 -0.13     1.37
##      se
## X1 0.16
```

```
ggplot(df,aes(x=df$sd)) +
  geom_histogram() +
  xlab("StDev") +
  ggtitle("StDev Distribution")  +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df,aes(x=df$r)) +
  geom_histogram()+
  xlab("Returns")+
  ggtitle("Return Distribution")  +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```