Correlation between SHEL and other securities

Andrea Lisco, Stefano Mauloni

December 2023

Abstract

This research aims to analyse the behaviour of Shell's stock price and determine its interaction with other securities and financial indices. In the following sections, a detailed account of the variables taken into consideration can be found, together with some insightful plots that hint towards the results which will ultimately be presented in the section on regression. Finally, the last section before the conclusions sheds light on possible investment strategies relative to this stock.

Dataset

Variable name	**Description**
`SHEL`	Shell plc. stock traded at the NYSE and NASDAQ.
`Refiners`	VanEck Oil Refiners ETF. It traces small-cap refiner companies worldwide (CRAK).
`Brent`	Brent Crude Oil Future (LCOH4)
`Crude_WTI`	Crude Oil WTI Future (CLG4)
`Natural_Gas`	Natural Gas Future (NGG4)
`Clean_Energy`	iShares Global Clean Energy ETF, traded at NASDAQ (ICLN).
`S&P500`	Standard & Poor 500 index (SPX)
`FedRates`	Federal Funds Effective Rate (DFF)
`HenryGas`	Henry Hub Natural Gas

Table 1: Variables considered in the final dataset, each one divided into Price and % Change, Daily.

Data cleaning

First, we computed the daily changes for the variables in which they were not present, namely `HenryGas`, `Refiners`, and `FedRates`. Then, we formatted in the right way some of the variables in the data frames (e.g. removing the % sign in the "Change" columns). In the end, we joined all the relevant variables by date: Price and Change of all securities. In doing so, we restricted our data in the period 2020-2023, by lack of information on the iShares ETF prices (`Clean_Energy`).

Visual exploration

We begin our analysis by developing an intuition of what variables will be relevant in the regression. A visual exploration of the data best fits this purpose. Hence, we proceed by plotting separately some of the variables against `SHEL`. Of course, the high-dimensionality of the dataset excludes the possibility of plotting all the variables together.

The plots of the prices (1) are not very informative. This could be caused by the strong influence of short-term events happening in the markets on the daily change in the price of a security. On the other hand, a prolonged high price for a stock is mostly due to mid-term macroeconomic conditions, and most of the time the correlation between the two differs strongly, the second being more complicated to model statistically (in fact, stochastic processes are applied to try to better explain those movements). Therefore, we take a look at the corresponding plots for the daily changes (2).

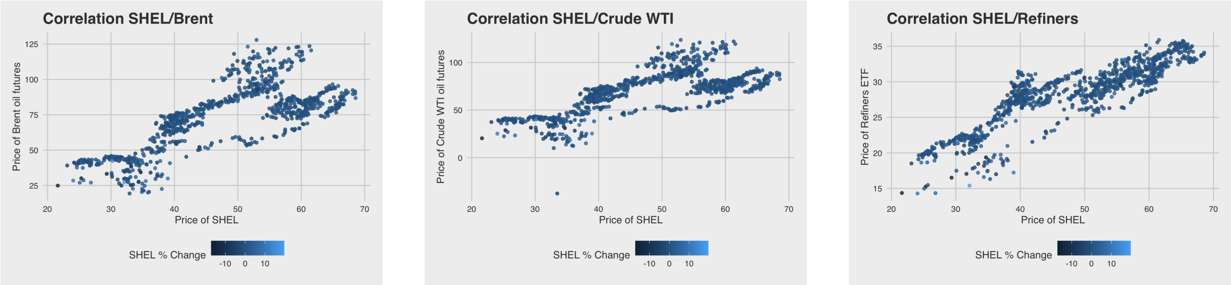


Figure 1: Price of SHEL against some of the proposed explanatory variables

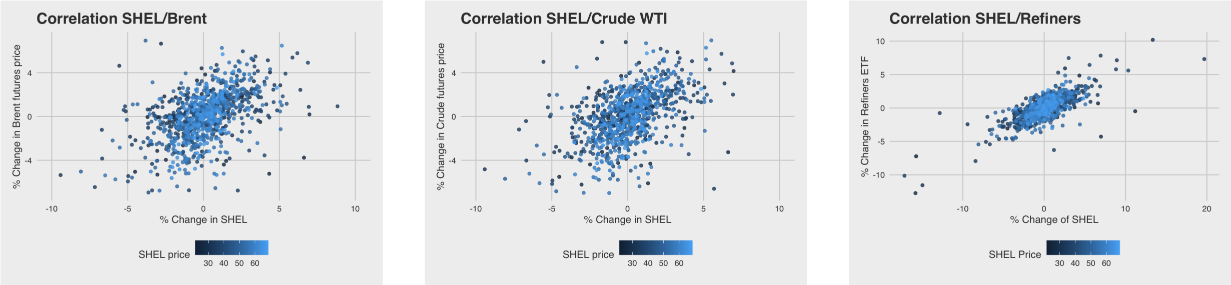
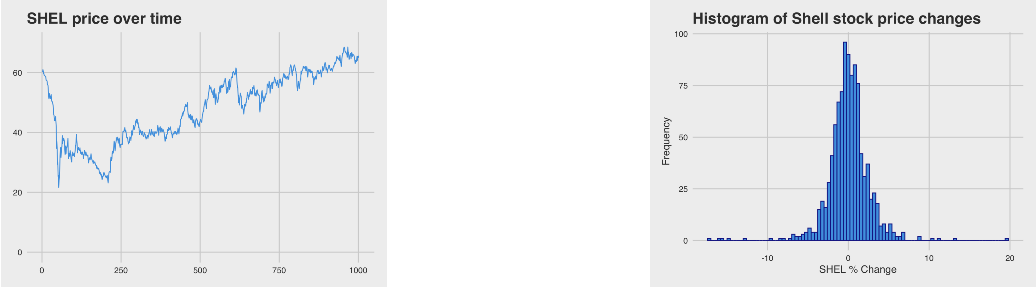


Figure 2: %Change of SHEL against some of the proposed explanatory variables

These separate graphs do not capture the correlation between the covariates themselves thus hiding any possible overlap of information. A more careful approach to variable selection will be explored later. However, we notice a pattern in all of these plots: the more sparse data points that escape the general trend tend to coincide with low prices of SHEL in the range 20-40.



(a) SHEL price from 2020 onwards

(b) Histogram of SHEL price

Figure 3: Plots of SHEL

As we can see from plot 3a, the data points with low SHEL prices are the ones corresponding to the Covid-19 outbreak period (indices 0-250), when prices were unstable, considerably more volatile and hence harder to include in a statistical model.

We conclude this preliminary exploration by noting the regularity of the distribution of the daily changes (3b). It is quite symmetric around 0 and does not exhibit any abnormality at the tails.

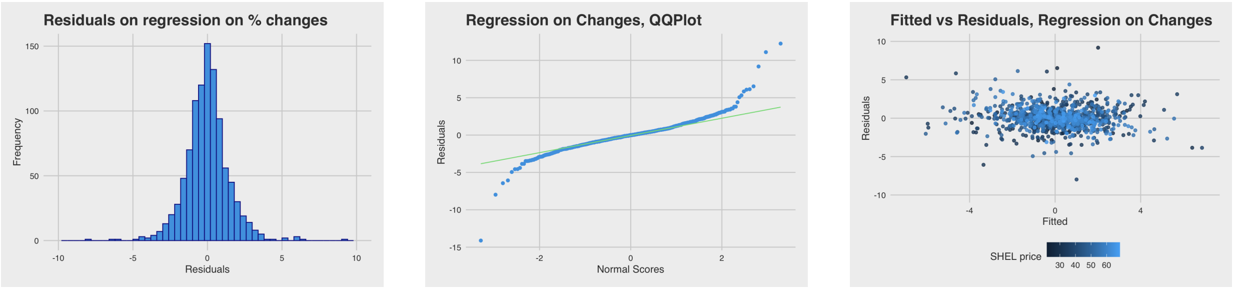
Regression on other securities

Full model

After this preliminary exploration of the dataset we can proceed with the regression: we want to predict the value of SHEL price and its daily change through the other 8 explanatory variables we described above. As a first attempt, we can try to include all the variables.

As for the daily change, the R output shows that Brent futures and Refiners ETF are deemed relevant by the t-test with p-values below the threshold of 0.05. As for the other variables, instead, there is not enough evidence to reject the null hypothesis that they are not relevant (i.e. that their coefficient in the regression vanishes). The same can be said about the intercept. The p-value corresponding to the F-test is $< 2.2 \times 10^{-16}$, providing very strong evidence in favour of this model versus a model containing only the intercept. Finally, we report an R^2 coefficient of 61%, meaning that this percentage of the variance of `SHEL`'s daily change is explained by the covariates. The adjusted R^2, taking into account also the model size, is only slightly smaller. Overall the model provides a good fit, however, a more precise selection of covariates can be found in the next section.

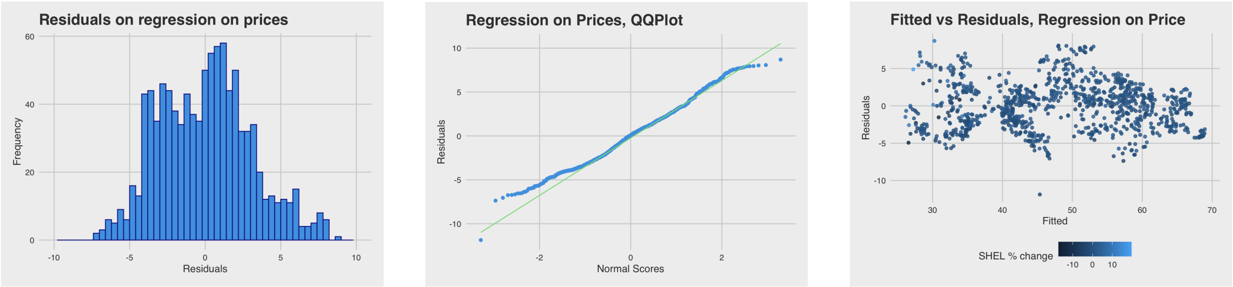
Next, we check the assumptions of normality and homoscedasticity of the residuals. Although a Kolmogorov-Smirnov test detects a significant deviation from normality, one could argue that this is due to the excessive sensitivity of the test and that the deviation is not significant in practice especially given the size of the dataset. Arguing with graphical methods, instead, we can see from the histogram 4a that the distribution of residuals is symmetric and that, in fact, the normality assumption is not evidently violated, as it can be deduced from the QQ-plot 4b. Lastly, we can see from the scatter plot 4c of residuals against fitted values that there is no evident violation of the homoscedasticity assumption either.



(a) Histogram of the residuals
(b) QQ-plot of the residuals
(c) Scatterplot of the residuals

Figure 4: Plots for the regression on daily changes

When the regression is done on prices instead of daily changes, the predictive power of the model increases notably. The R output shows that all covariates are deemed relevant by the t-test with p-values below the 5% threshold. The p-value corresponding to the F-test does not change compared to the last regression, however both the multiple and adjusted R^2 coefficients rise to a value of 93%. The normality of the residuals now is even more evident, as it is confirmed also by the Kolmogorov-Smirnov test as well as by the usual plots (5a, 5b). Also, in the scatterplot 5c we do not see any trend violating homoscedasticity.



(a) Histogram of the residuals
(b) QQ-plot of the residuals
(c) Scatterplot of the residuals

Figure 5: Plots for the regression on stock prices

The fact that the model explains almost all variance in `SHEL`'s stock price but only a modest 61% of its daily change may come as a surprise. We refer to the discussion in the ”Visual exploration” section for a plausible explanation.

Stepwise variable selection

Trying to improve on the previous results we resort to a stepwise approach to select suitable covariates. Starting with the regression on the daily change, the Forward algorithm ends up selecting only `Brent` futures and `Refiners`. As a matter of fact, the t-tests provided significant evidence in favour of their relevance. Both the multiple and adjusted R^2 coefficient for this simpler model slightly drop to 61%. Despite this slight decrease in performance, we consider this model preferable to the complete one considered at the beginning of the previous section for explanation purposes because of its simplicity. The Backward algorithm produces an empty model with only the intercept, while the mixed algorithm, which we may denote by "Both", agrees with the simpler Forward algorithm.

As we can see from plot 6 showing predicted values against observed values, the two models are both quite close to the identity line and the predictions often overlap. Hence, as far as the regression on daily change goes, we retain the model provided by the stepwise approach.

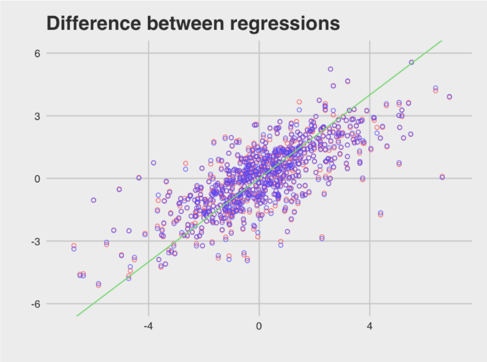


Figure 6: Predicted against observed values.
Blue: stepwise selected model.
Red: full model.

When it comes to the regression on prices, the Forward and Both algorithms agree with the complete model including all 8 covariates, while the Backward algorithm produces the model containing only the intercept.

Possible investment strategies

We want to check if the returns of `SHEL` are substantially higher in one period of a year or not. To do that, we compute the daily log returns in each semester for the years 2010-2023 and store them in a different dataframe.

We want to test if the location of the two samples coincides through a two-sample t-test. Not having enough information to assume the equality of the two variances and taking into account the large size of the observations we decide to rely on an asymptotic t-test. The computation of the log returns was done precisely so that the normality assumption on the samples would be reasonable. The histograms below (6) show that there is an evident symmetry to their distribution.

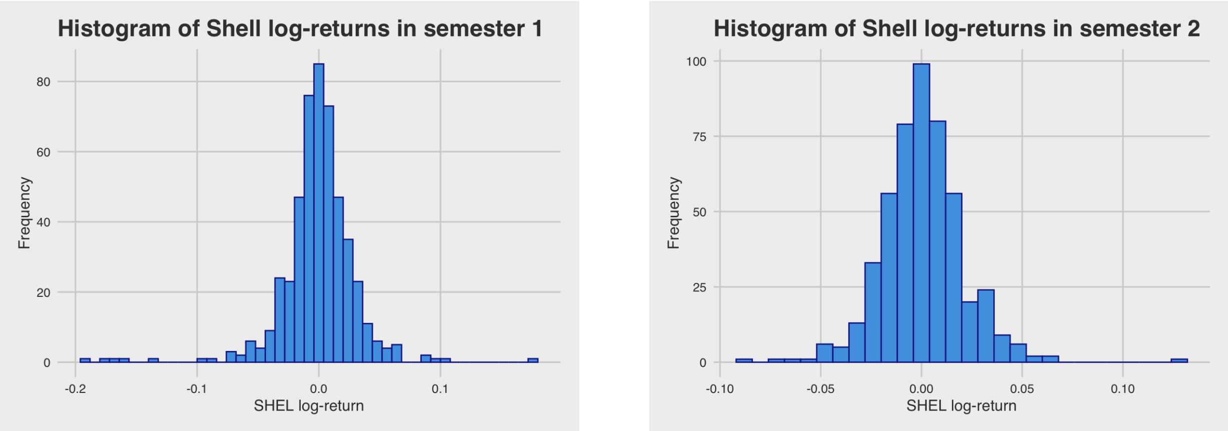


Figure 6: Histograms of log-returns in both semesters

```
data:  sem1$log_return and sem2$log_return
t = -0.71689, df = 991, p-value = 0.4736
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.004434864  0.002061587
sample estimates:
    mean of x     mean of y
-0.0005080029  0.0006786357
```

With a p-value above the usual threshold of 5%, we do not have sufficient statistical evidence to conclude that there is a significant difference between the mean values of the log-returns in the two semesters. In particular, a day trading strategy could be implemented indifferently in either of the two semesters since the data does not provide reasonable evidence to prefer one over the other.

Conclusion

In this project, we tried to explain changes in a stock price with some covariates that may seem correlated in some sense. Of course, the research is not completely satisfactory. Some critical points that show room for improvement are the following:

- *The accuracy of the model strongly varies if the regression is made on prices or percentage changes in price.* As explained before (see the section on visual exploration), this is because the correlation between a price level and its change is not due to the same variables.

- *For predictive purposes, much more complex models are put into work*, and even those may fail under some conditions. Our purpose was not to predict something, but just to try to explain changes in the prices of a stock with other changes in related commodities, interest rates, and so on.

- *The choice of the covariates may not be optimal*: other variables may potentially explain movements of `SHEL` in a more satisfactory way but they are definitely more complex (Options, Bond Derivatives, etc...) and even finding datasets with daily frequency for them is difficult, provided they even exist (if it is not market data, they may be not published at all).

- *Outliers are a problem*: the model cannot explain daily changes greater than 3% in absolute value with adequate precision. This is because stock prices encapsulate information about macroeconomic and geopolitical conditions, that here are not taken into consideration because it is not possible to introduce them effectively with a simple model (e.g. an announcement on the BBC that one of Shell's unlisted competitors is introducing a disruptive technology in the industry would cause the value of the stock to drop, but since the other company is not publicly traded, we cannot consider its valuation increase in the model). Of course, there are some workarounds but they are not easy to implement and their effectiveness could still be questioned. In our case, we tried to partially include macroeconomic information by including `S&P500` and `DFF`, which ideally would serve as market beta, but it turned out that they were both uncorrelated to `SHEL`.

Bibliography

- Fetsje Bijma, Marianne Jonker, Aad van der Vaart, *An Introduction to Mathematical Statistics* (Amsterdam University Press B.V., Amsterdam 2017).
- investing.com, Fusion Media Limited, https://www.investing.com/commodities/crude-oil-historical-data.
- investing.com, Fusion Media Limited, https://www.investing.com/commodities/brent-oil-historical-data.
- investing.com, Fusion Media Limited, https://www.investing.com/commodities/natural-gas-historical-data.
- investing.com, Fusion Media Limited, https://www.investing.com/indices/us-spx-500-historical-data.
- investing.com, Fusion Media Limited, https://www.investing.com/etfs/ishares-s-p-global-clean-energy-historical-data.
- investing.com, Fusion Media Limited, https://www.investing.com/equities/royal-dutch-shell-a-plc-historical-data.
- finance.yahoo.com, Yahoo, https://https://finance.yahoo.com/quote/CRAK/history?p=CRAK.
- eia.gov, U.S. Department of Energy, https://www.eia.gov/dnav/ng/hist/rngwhhdD.htm.
- fred.stlouisfred.org, Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org/series/DFF.