Electricity Trading Strategy Evaluation and Prediction Modeling

Brennan Hall

September 18, 2020

Energy Trading Strategy Evaluation

Descriptive Analysis

First, we present basic summary statistics for and plots of the price process in Table 1, Figure 1, and Figure 2, considering the entire time interval as well as each individual year. The obvious note is that there are massive jumps in the process; however, since the mean and median are relatively very close, we can conclude these jumps occur very rarely and do not heavily skew the distribution. For future reference, we define a *jump* as the event when the price is above \$150, or five times the largest yearly 3^{rd}-quartile.

	Min	1^{st}-Quartile	Median	Mean	3^{rd}-Quartile	Max	St. Dev.
2017	-13.61	18.73	20.89	26.84	26.18	861.72	32.77
2018	-4.22	18.92	22.59	31.89	30.91	2783.61	64.48
2019	-10.96	17.01	20.16	38.03	27.16	8999.33	230.28
2020	-24.18	12.80	16.45	20.85	19.80	984.14	34.057
Total	-24.18	17.54	20.46	30.63	27.28	8999.33	129.81

Table 1: Summary statistics of price process

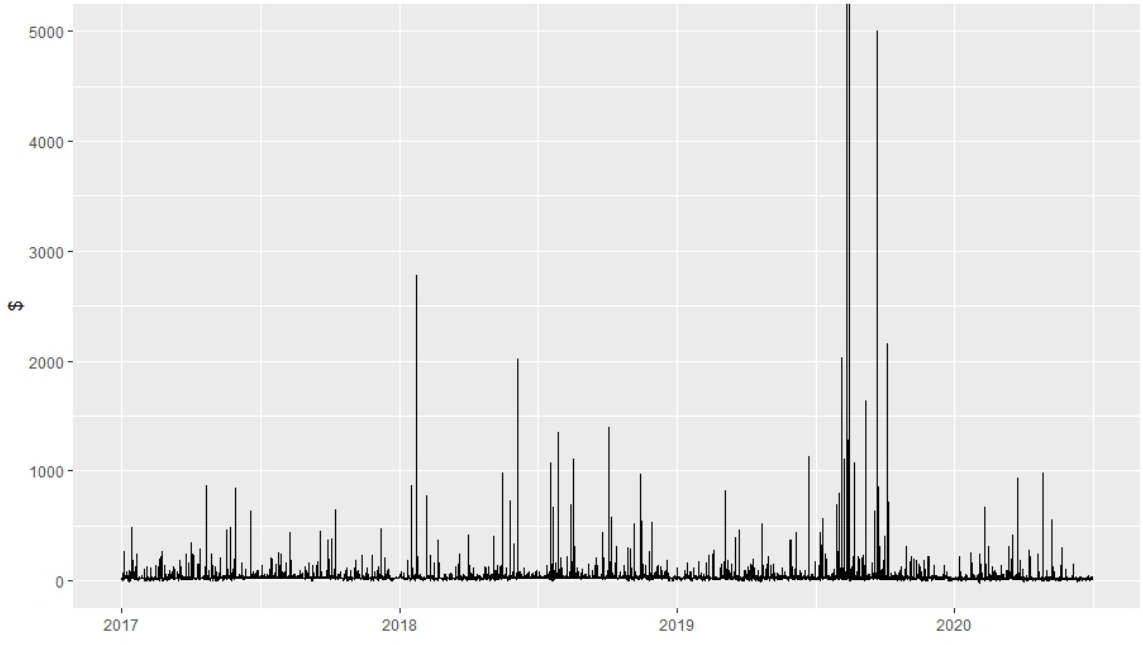

Figure 1: Price over entire time interval. Note: max price exceeds graph limits.

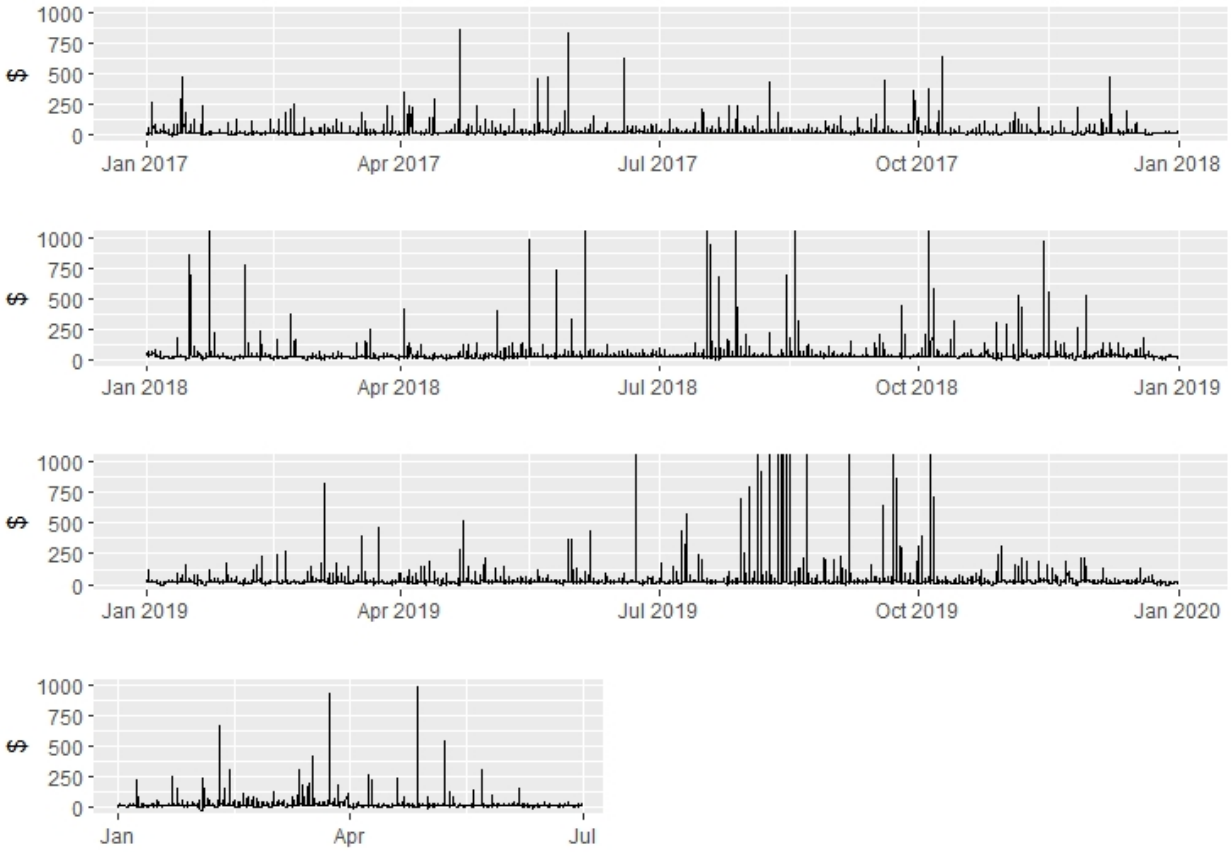

Figure 2: Price over each year. Note the change in vertical scale than above.

Following these observations, we investigate abnormalities in the price process, particularly the jumps and instances of negative prices.

Prob. of:	"Jump"	"Afternoon Jump"	"Negative Price"	"Morning Neg. Price"
2017	0.009	0.71	0.004	0.943
2018	0.012	0.72	0.003	0.88
2019	0.017	0.79	0.004	0.974
2020	0.008	0.59	0.014	0.918
Total	0.012	0.735	0.052	0.931

Table 2: Table of sample probability estimates.

As noted in Table 2, it was discovered that the majority of jumps in the price process occur during the "afternoon" (defined as hours 11, 12, 13, 14, 15, 16, 17). Specifically, about 70% of the time, a jump occurred within one particular quarter of the day.

Furthermore, it was discovered that about 90% of the events when the price was negative was during the "morning" (defined as hours 23, 0, 1, 2, 3, 4, 5).

Additionally, we investigate the price process' volatility process over the time intervals. In Figure 3 we see that most of the volatility occurs during the Summer months, but large spikes in volatility are noted in Winter months as well. It should also be noted that spikes occurr in relatively close proximity to other spikes. Meanwhile we should recognize from Table 3 that the Mean of the volatility process is at least double its Median, indicating a skewed distribution.

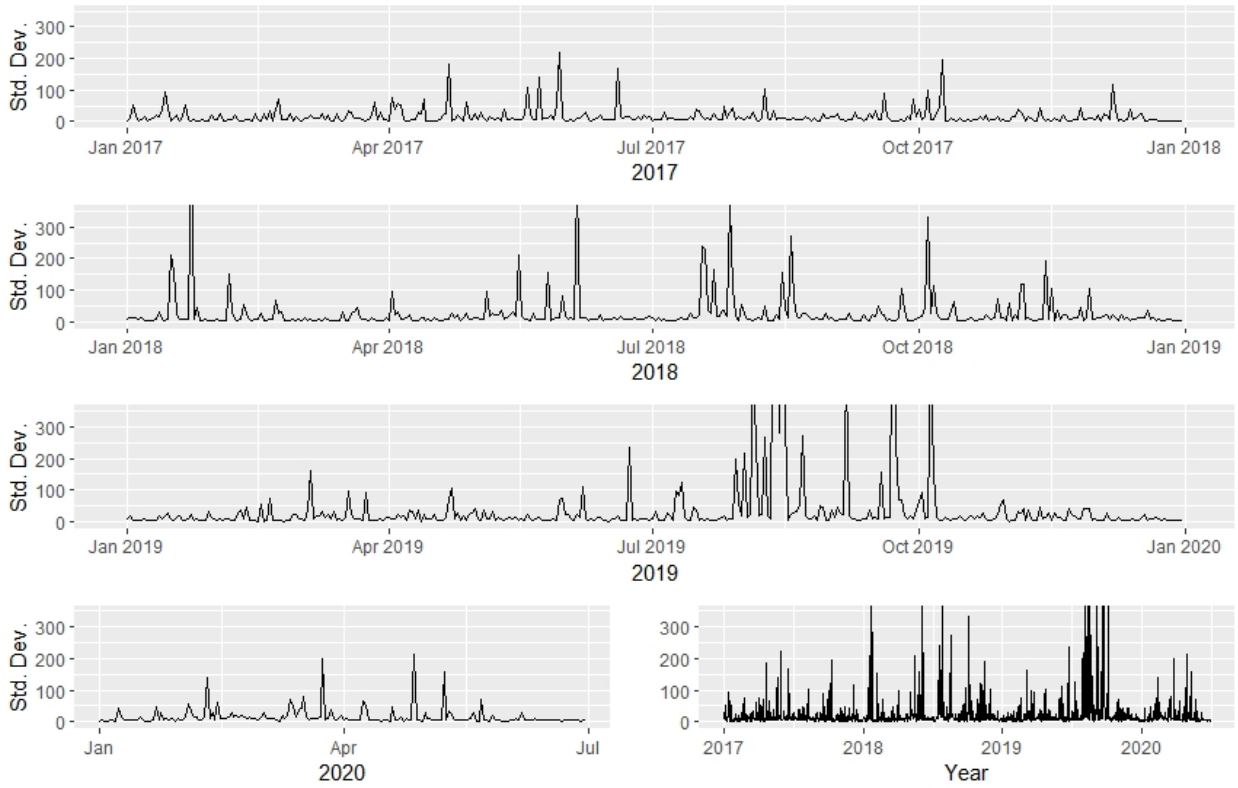

Figure 3: Standard deviation of price over each year.

	Min	1st-Quartile	Median	Mean	3rd-Quartile	Max	SD of SD
2017	0.546	4.668	8.019	16.474	17.664	220.956	26.279
2018	1.052	5.377	8.813	24.850	19.608	563.236	56.775
2019	0.932	5.669	9.491	46.902	23.011	2676.605	213.231
2020	1.470	4.112	6.26	15.54	13.04	215.39	28.54
Total	0.547	4.983	8.627	27.433	18.938	1676.605	119.863

Table 3: Summary statistics of the daily volatility process

Finally we present information on the autocorrelation of the price process in Figure 4. Based on the structure of the autocorrelation functions over each of the time intervals, we can hypothesize that the process follows closely to an AR(1) process with a 24-hour seasonal component *during normal periods*. Essentially, throughout a day the process is mostly dependent on the price of the previous hour. The dependence then resets every 24 hours. This could suggest a Markovian modeling approach to daily price changes. It should again be noted that this behavior should only be used to construct a strategy during normal periods, and periods of negative prices or jumps in price should be accounted for and modeled separately.

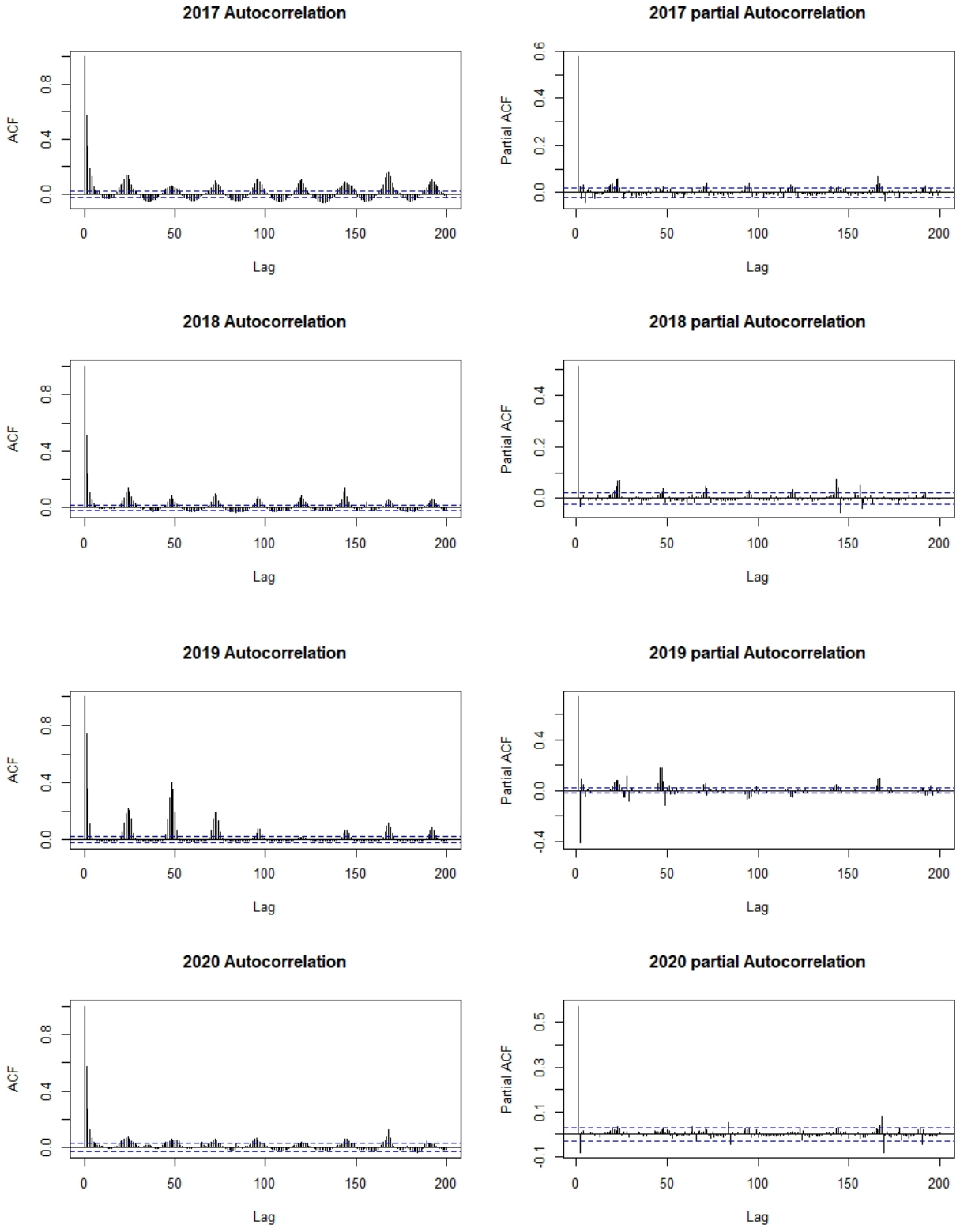

Figure 4: ACF and PACF of price process over 2019 and 2020.

Strategy Evaluation

The first evaluation of the strategy is a summary of the daily trading behavior.

Daily Returns	Daily Returns Std. Dev.	Daily Trade Volume
Min. :-169594.7	Min. : 45.43	Min. : 397.9
1st Qu.: -529.5	1st Qu.: 237.71	1st Qu.: 891.0
Median : 675.2	Median : 359.54	Median : 944.1
Mean : 3131.2	Mean : 1033.34	Mean : 944.4
3rd Qu.: 2640.2	3rd Qu.: 693.45	3rd Qu.: 997.7
Max. : 344308.3	Max. :43099.72	Max. :1193.2

The first note to make is that this is a highly volatile trading strategy with a single daily return range of -$169594.7 to $344308.3 and mean standard deviation of $1033.34. Interestingly enough, those minimum and maximum daily return values occurred within ten days of each other while executing roughly around the strategy's mean trading volume, indicating this was possibly just a result of an extremely volatile period of prices (as was the case, noted above) and no fault to the strategy.

Some of the positive notes on the strategy are that it appears to be consistently profitable strategy with 53.56% of trades being profitable, its positive average daily returns, seen clearly by the Cumulative Returns, displayed in Figure 5, whose final figure lies at $3,998,176.

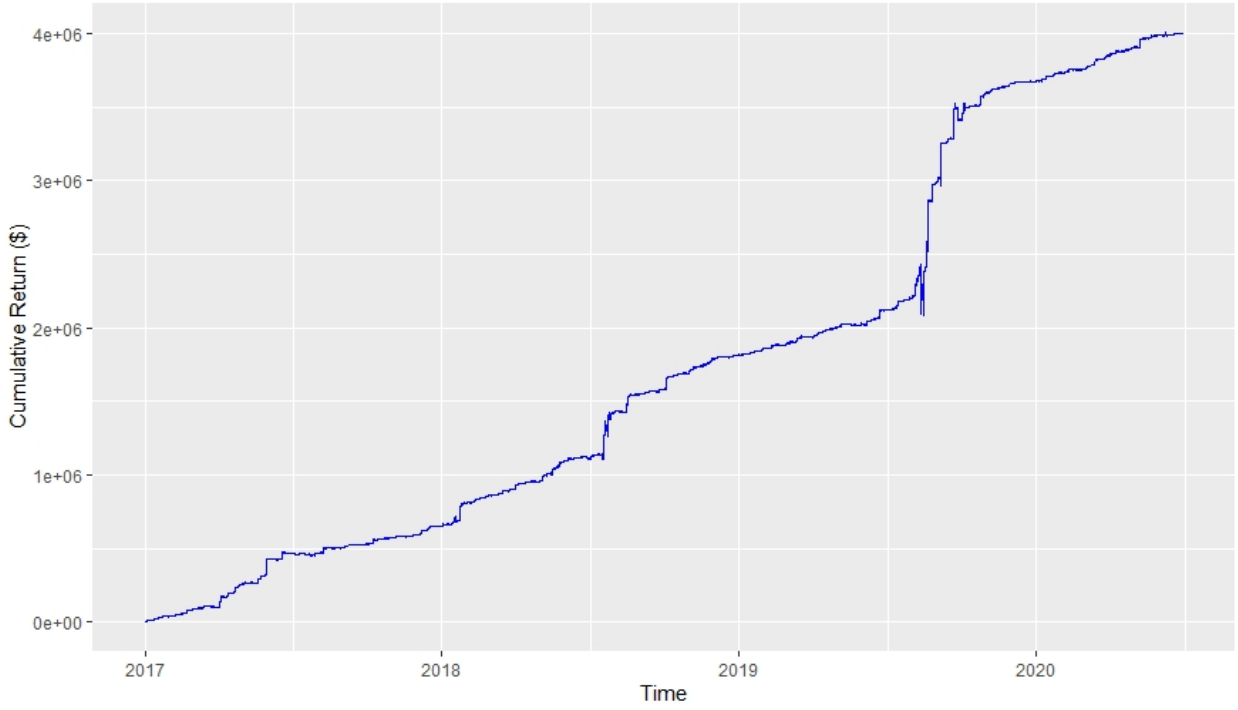

Figure 5: Cumulative daily return.

Some faults of the strategy can be egregious though. While the extreme volatility in latter part of 2019 may be excused, there is still a large degree of volatility throughout the entire time period with daily return standard deviations being (rudimentarily) classified as large by $1.5 \cdot (IQR)$ over 25% of the time, where IQR is the range between the 1^{st} and 3^{rd} quartiles.

This strategy may be profitable, but there appears to be a high amount of risk involved as well.

Prediction Modeling

Variable Selection

To start this problem, some initial data preparation was conducted. Before any of standard statistical methods are applied, the missing data was accounted for by imputing data via the *predictive mean matching* method, implemented in the mice package in R. This method constructs a linear regression to predict the missing values based on the non-missing values. This method is implemented on all data sets with missing values. Following this we attempt to reduce the dimensionality of the training data. We do this in two steps. First, we remove variables that have perfect collinearity with another variable. Second, we only include variables that have some marginal impact on the variance of the data.

The first step results in just five variables being removed, $var_3 3$, $var_3 4$, $var_3 5$, $var_8 7$, and $var_8 8$. The second step is executed using Principal Components Analysis (PCA). The eigenvalues computed from the PCA indicates that, of the 143 loadings, only the first 37 have any significant marginal impact on the variance of the data. This is clearly seen in Figure 7.

Therefore, of the remaining variables after the initial five collinear variables were removed, we only use the first 37 principal components for the following analysis.

To start the modeling selection, our initial thought was to use a Gaussian regression model based on the fit of the estimated density, shown in Figure 8.

However the extreme outliers in the data made initial attempts at fitting a Gaussian regression model seem ill advised. To correct for the non-Gaussian behavior, a Yeo-Johnson transformation was applied to the target variable and a ridge regression model was finally decided upon. Since there is no need for interpretation of this model, the induced bias from a regularized regression is acceptable.

Point Predictions and Prediction Intervals

To now make our point predictions, we first transform the test data to match the components found by the PCA conducted earlier. We then use the results from our ridge regression model to construct point predictions and prediction intervals based on the test data provided (after being put through the same PCA

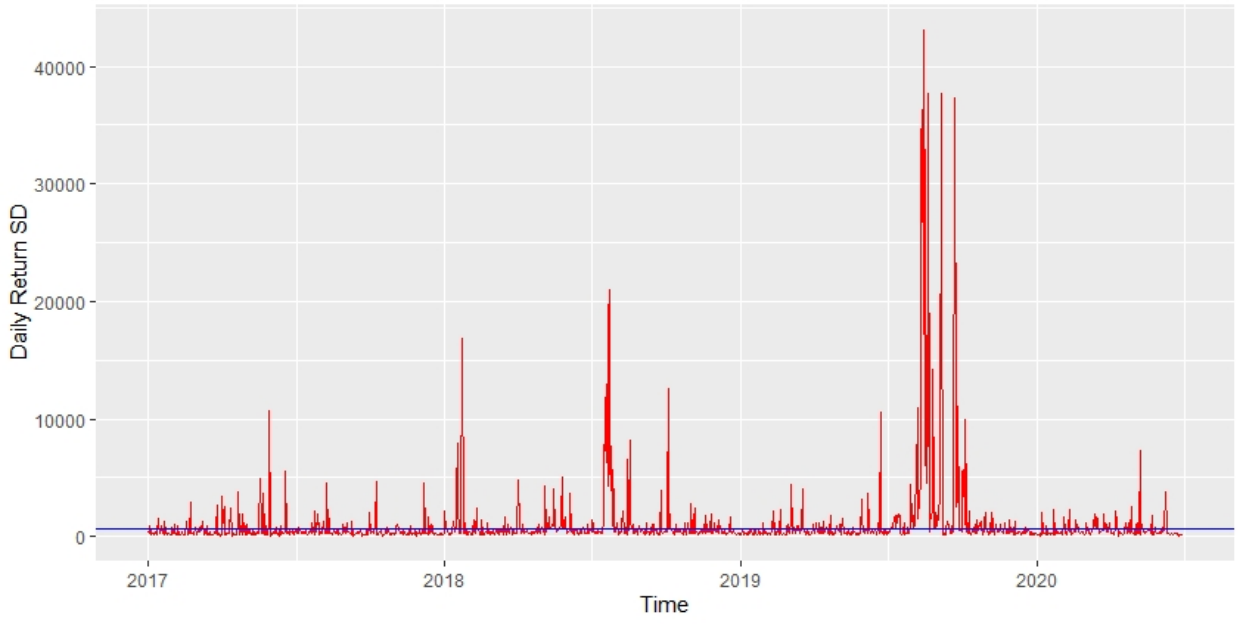

Figure 6: Daily return standard deviation bisected by an indicator measure for outliers.

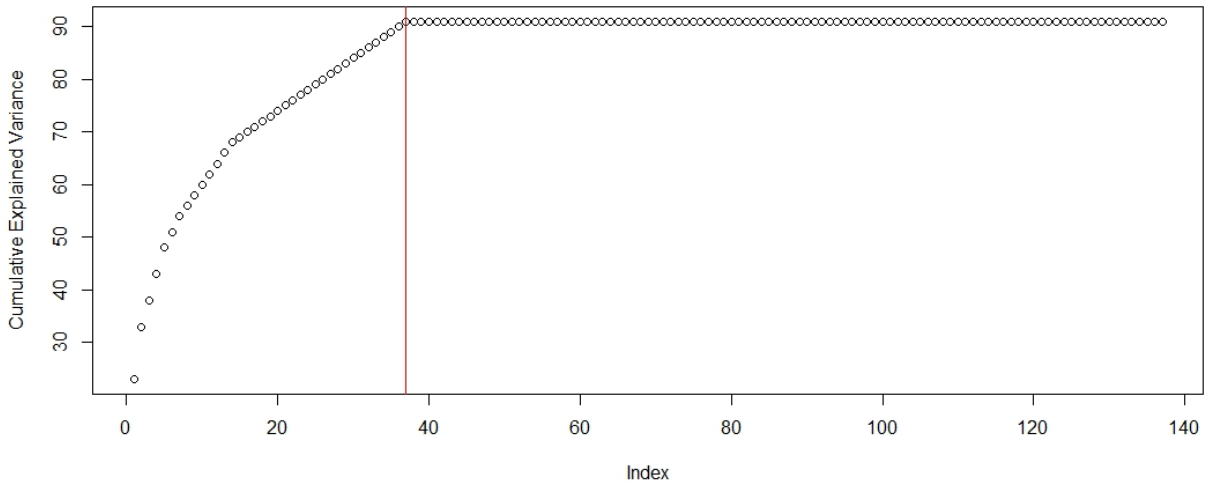

Figure 7: Cumulative variance explained by each consecutive principal component.

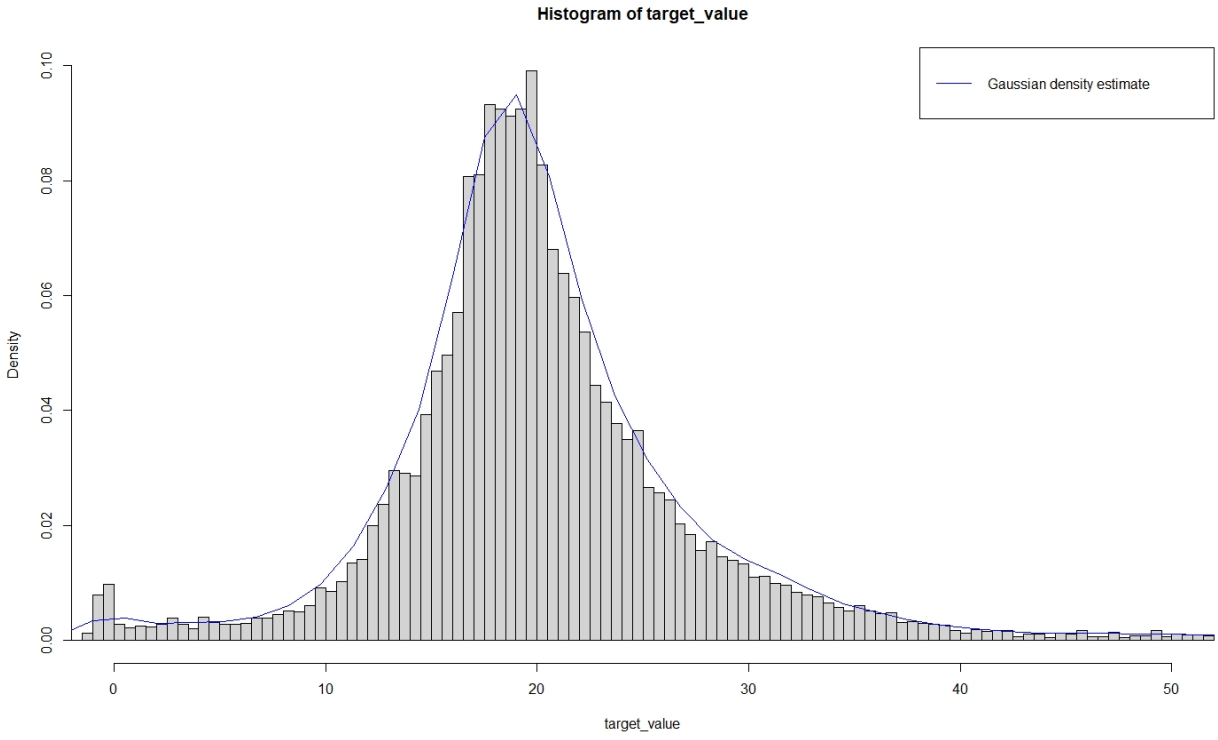

Figure 8: Histogram of target variable with large outliers removed from plot range. An estimated Gaussian density curve plotted over it deceptively appears to fit well.

process as the training data).

As it is difficult to attain accurate standard errors for ridge regression due to the penalized estimation method, we attempt to construct confidence intervals through a bootstrapping algorithm. However, it should be

noted that these intervals are likely much tighter due to the bias of the regression and the fact that the bootstrap-based intervals are only based on the variance of the estimates.

The results are presented below as well as provided as a .csv file in the predictions.csv file.

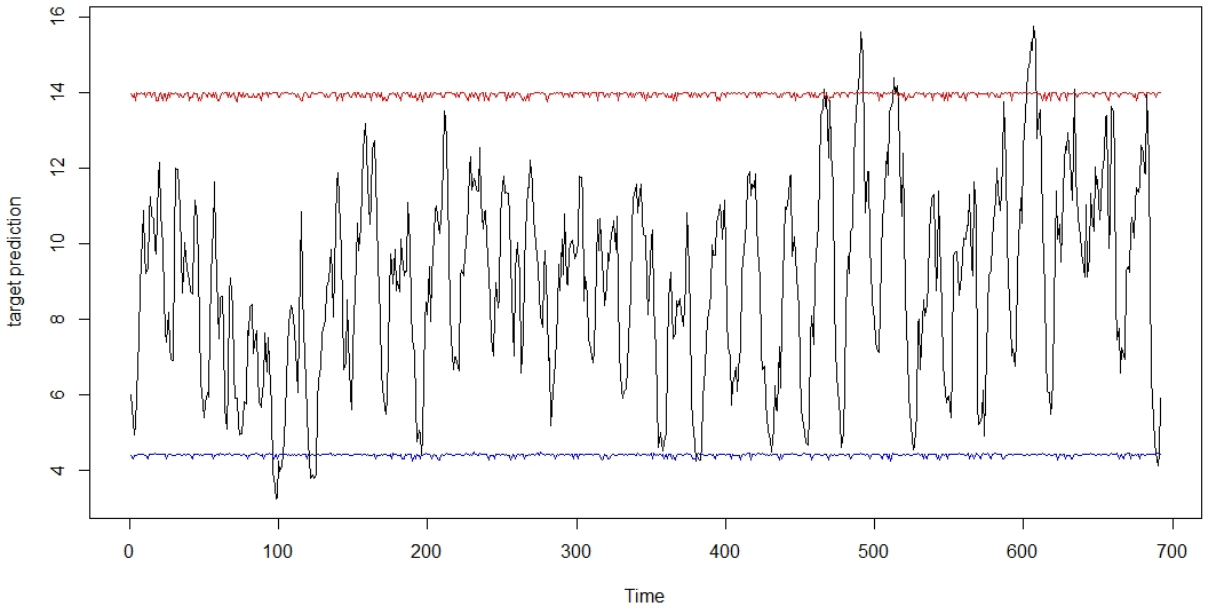

Figure 9: Point predictions with 95% prediction intervals.

Predictions from variable bounds

Given instead two matrices containing the upper and lower bounds of the independent variables, we attempt to construct point predictions of the target variable. Without having any information about the structure or nature of the bounds, we make the assumption that they are constructed in a symmetric fashion so that the point estimate is the midway value between the bounds. And so, using the mid points of the bounds as the new test variables, we can construct point predictions and prediction intervals of the target variable in the same fashion as above.

The results are presented below as well as provided as a .csv file in the bound_predictions.csv file.

Considering Figure 10, we see the point predictions from both methods produce similar results with a maximum difference of 0.155.

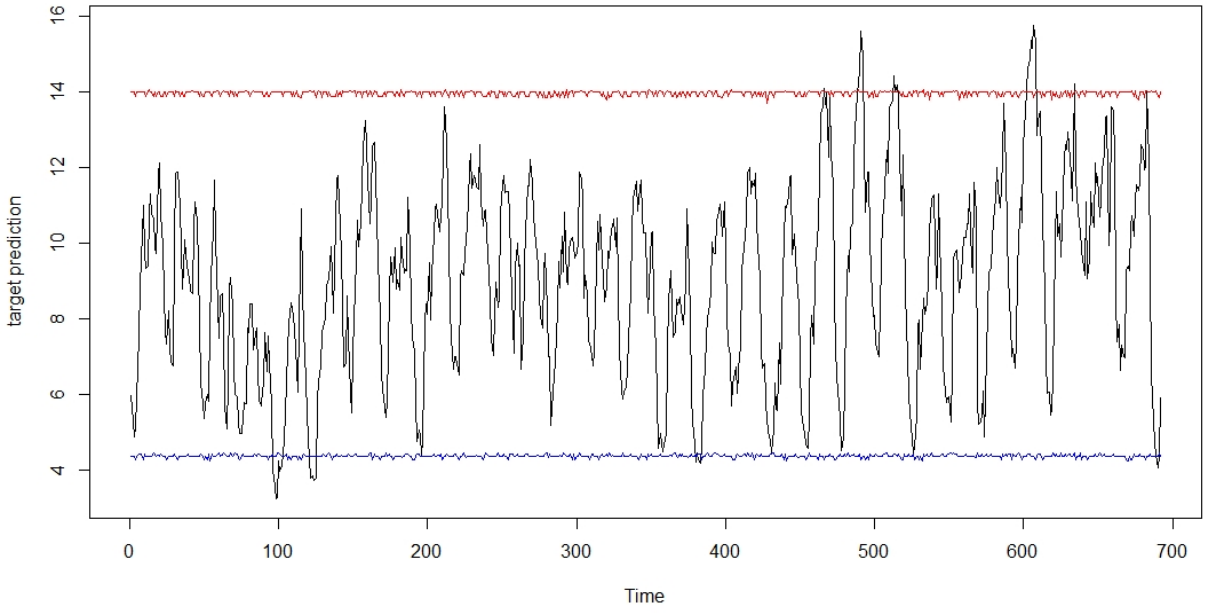

Figure 10: Point predictions with 95% prediction intervals from bounds of independent variables.

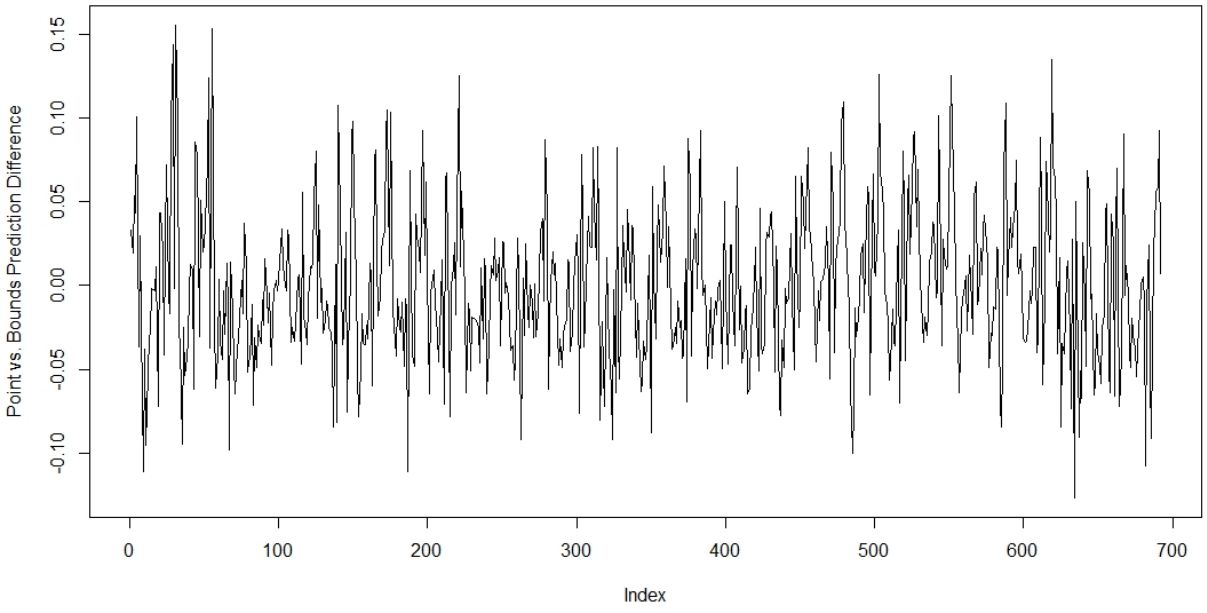

Figure 11: Predictions difference from point method and bounds method.