Comparison of Linear, Polynomial, and Gaussian Modelling in Regression Analysis Using Stock Market Data

Michael G. Bemus[1]

[1]Purdue Universtiy Fort Wayne

introduction

Motivation:

Creating regression models is a useful tool in statistical analysis and prediction. Along with linear, polynomial, and other models, Gaussian process models are a special type of regression analysis give a pliable method of fitting line models to data without using point-estimations. The utilization of this function could be very useful in regression fields, which include chemistry, business, actuarial science, computer science, and medical science.

Goals of Study:

This study aims to compare Gaussian process regression to other models. Using this method of regression along with linear and polynomial regression, I attempted to predict the volume of Tesla stock sold based on price factors. In this poster, I will focus on the comparative fits of the models rather than their actual applicability.

Data and Software

- **Data**: For this project, I used Tesla stock market data provided on Kaggle (Chauhan, 2022). The data set contains 3,077 observations, each containing the opening price, closing price, highest price, lowest price, and volume sold of Tesla stock during a day range from 2012 and 2022.
- **Models**: To run the regression tests, I used a combination of R and Python to create regression models. R was used to create the linear and polynomial models. Python was used to create the Gaussian process model, along with the *sklearn* and *pandas* packages.
- **Graphs**: All graphs for this project were made using R and the *ggplot2* package.

Fitness Tests

For this analysis, I used Out-of-Sample Validation to compare the generated models. This method uses two data sets to measure the performance of a regression model (Frees, pg. 173, 2010). The first set, the training set, is used to create a candidate model designed by the selected regression techniques (Frees, pg. 173, 2010). The second set, the testing set, is used to determine how effective the model is at making predictions about the data set (Frees, pg. 173, 2010). For this test, the training set was composed of the first 80% of observations, and the testing set contained the remaining 20%.

To assess the training fit, I used the Mean Square Error (MSE) of the model. The MSE is an estimator of the variance, σ^2, of a distribution and can be calculated by:

$$MSE = (n-(p+1))^{-1}\sum_{i=1}^{n} e_i^2.$$

where n is the number of observations, p is the number of variables used, and e is the error of each estimate (Frees, pg. 84, 2010). MSE is a useful tool for fitness testing because it does not rely on the parameters of a model.

To assess the test fits, I used the candidate model to predict the test data point. To measure the fits of these, I calculated the sum of squared predicted errors, SSPE, through the following formula:

$$SSPE = \sum_{i=n_1+1}^{n_1+n_2}(\hat{y}_i - y)^2$$

where n_1 is the number of entries of the training set, and n_2 is the number of entries in the testing set (Frees, pg. 173, 2010).

Ideally, both measures should be minimized for the best overall fit.

Linear Regression

According to Abidoye et. al. (2022), the multiple linear regressions model takes the form of:

$$Y = X\beta + e,$$

where Y is a vector with n rows of single observations of the response variable (y), X is a matrix with the same n rows and p columns of explanatory variable observations (x_1, x_2, ..., x_p), β is a vector with the same p rows containing unknown coefficients (β_1, β_2, ..., β_p), and e is a vector with n entries that represents the residual errors of the estimations.

In this model, the most common way to find β is to calculate:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y),$$

with $\hat{\beta}$ representing the p by 1 vector of β estimates (Abidoye et. al., 2022). This method provides an unbiased estimator of our response variable with the lowest variance compared to other linear estimators (Abidoye et. al., 2022).

Polynomial Regression

The polynomial model of regression is a specialized version of the multiple regressions model. According to Ostertagová (2012), it adds extra explanatory variables computed by the polynomial powers of given observations. Commonly used in univariate regression, polynomial regression typically defines the vector X as $x_1 = x$, $x_2 = x^2$, ..., $x_p = x^p$ (Ostertagová, 2012). Otherwise, the calculations are the same (Ostertagová, 2012).

For this model, I will be using the second power of each explanatory variable for our polynomial model. In practice, the polynomial model should be applied in stages, testing the applicability of each additional power for each variable one at a time. However, for this analysis, a simple square model will be used as a curved comparison against the Gaussian Process model.

Gaussian Process Regression

The Gaussian process model of regression applies an approach different from the point-estimations shown above. According to Winarni et. al. (2022), the Gaussian approach utilizes the mean function and covariance matrix of explanatory variables to predict the response variable. The model does this using the Gaussian (normal) probability distribution, which can utilize Gaussian sets within a theoretically non-Gaussian space to create accurate predictions (Winarni et. al., 2022). It produces a more pliable interval prediction of y values that can curve with the shape of the data to make predictions (Winarni et. al., 2022).

As stated by Rasmussen & Williams (pg. 13, 2006), the basic mean and covariance functions are defined as follows:

$$m(x) = \mathbb{E}[f(x)]$$

$$k(x, x^T) = \mathbb{E}[(f(x) - m(x))(f(x^T) - m(x^T)].$$

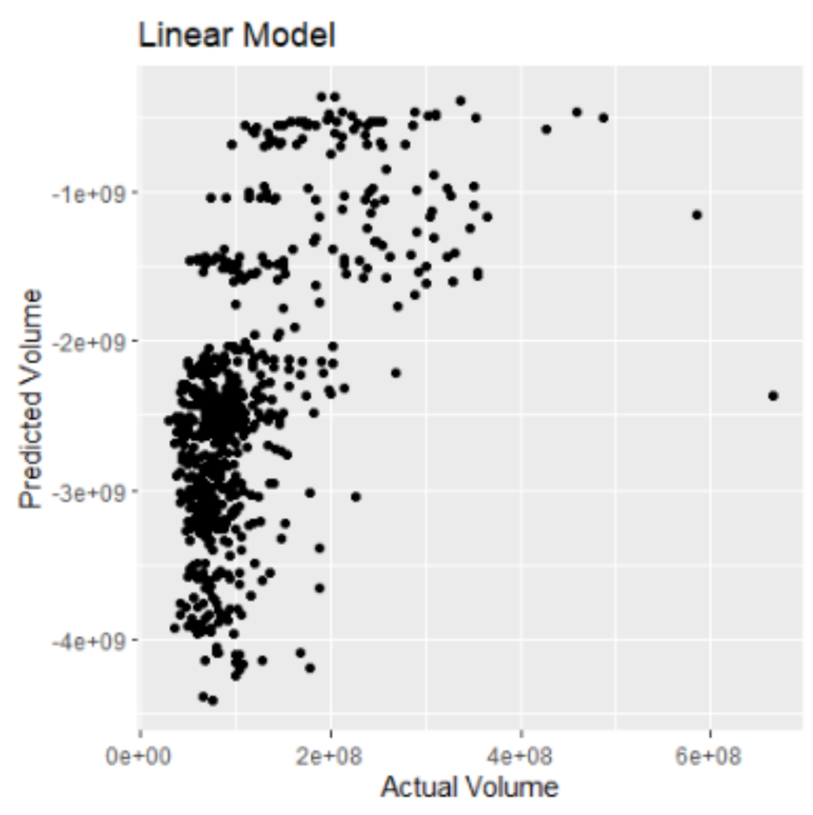
In the above, x represents a row of observations of the explanatory variables, X, that result in the response variable, y, or $f(x)$ (Rasmussen & Williams, pg. 13, 2006). $m(x)$, then, represents the mean function, and $k(x, x^T)$ represents the function to compute the covariance matrix (Rasmussen & Williams, 2006).

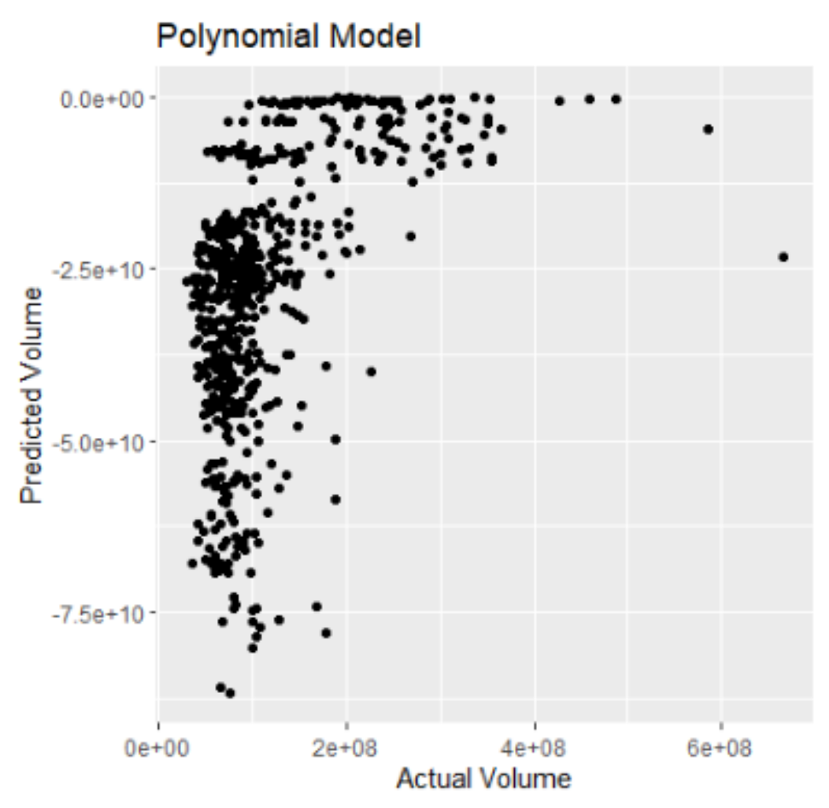
The shape of a model is determined by its covariance function (Rasmussen & Williams, pg. 6, 2006). For this test, I used the squared exponential function, which Rasmussen and Williams (pg. 14, 2006) define as:

$$cov(f(x_p), f(x_q)) = exp(-1/2|x_p - x_q|^2).$$

Results and Discussion

As shown by the table, the Gaussian Process model provided both the lowest MSE and the lowest SSPE of the three techniques. This was expected due to the model's ability to follow a curve. Of the other two models, the polynomial had a lower MSE, but the linear had a lower SSPE. The first was expected to occur because more variables allows the function to vary more along with the response variable. However the latter observation likely represents n over-fitting of the data.





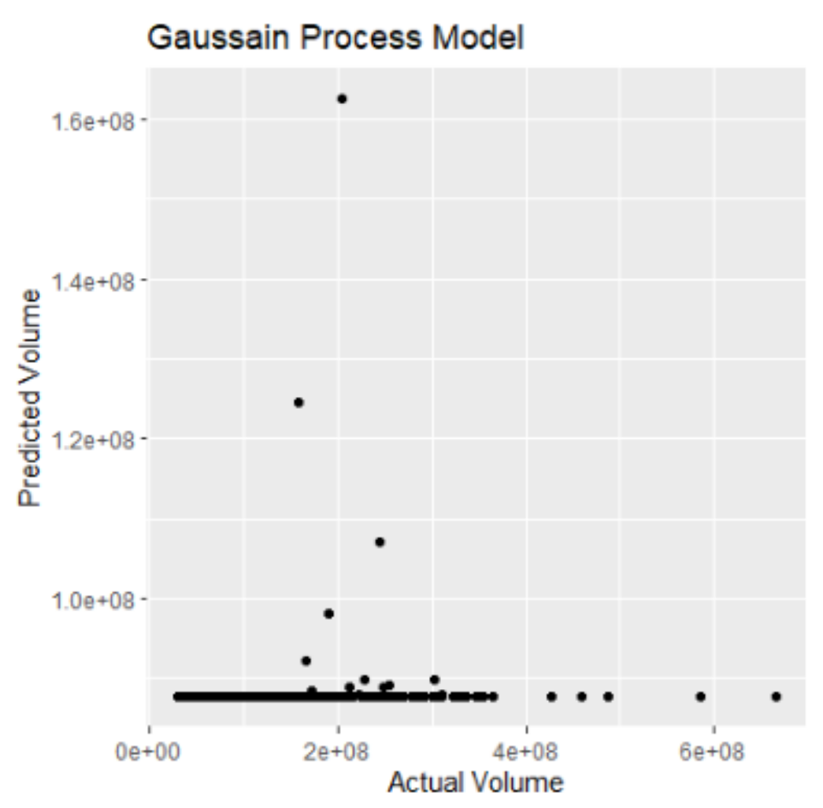


Table 1. Model Results

Technique	MSE	SSPE
Linear	$2.96 * 10^{15}$	$4.40 * 10^{21}$
Polynomial	$2.54 * 10^{15}$	$7.24 * 10^{23}$
Gaussian Process	$4.34 * 10^{14}$	$4.27 * 10^{18}$

One explanation for these results is the time-series nature of the data. One variable present in the set was the date a set of observations was taken. The test set was the last 20% of the set, meaning it was the latest values. This was done intentionally to test these model's efficacy in actual prediction.

Another explanation could be the correlations between the individual variables and y. In the full data set, the correlations to Volume were around 0.6.

In addition, for the Gaussian Model, we see that most predictions were quite flat, remaining near the mean. This likely represents a non-optimal covariance matrix being chosen for the model.

Conclusion

As shown by this analysis, over-fitting is a pervasive issue in standard models of linear regression. Especially in regression analysis of stock market data, it is often difficult to fit a model which can accurately move witht he shape of the data. Even so, by using Gaussian Process regression, I was able to provide the best fit for the data.

References

[1] A. O. Abidoye, I. M. Ajayi abd F. L. Adewale, and J. O. Ogunjobi. Unbiased modified two-parameter estimator for the linear regression model. *Journal of Scientific Research*, 14(3):785–795, 2022.

[2] A. Chauhan. Tesla inc. | stock market analysis | founding years. Kaggle, Accessed October 16, 2022 [Online].

[3] E. W. Frees. *Regression modeling with actuarial and financial applications*. Cambridge University Press, One Liberty Plaza, 20th Floor, New York, NY, 10006, USA, 2010.

[4] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 55 Hayward Street, Cambridge, MA, USA, 2006.

[5] S. Winarni, S. W. Indratno, and B. Hsu. Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Computing & Applications*, 32(10):5461–5469, 2019.