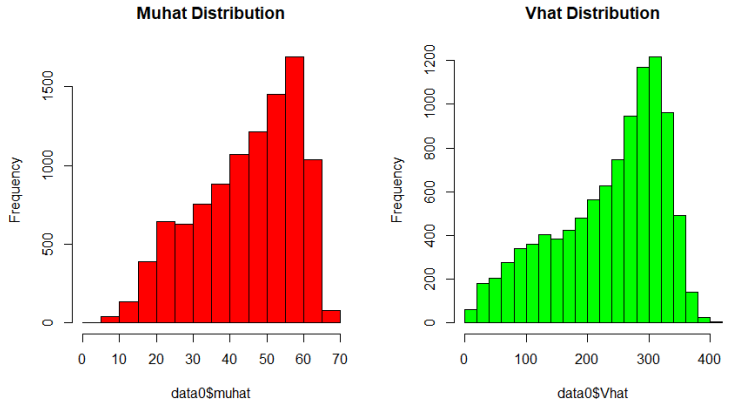
ISYE 7406 Take Home Final

Name: Nathan Kai Ching Chan (698)
Email: nchan64@gatech.edu
Date: 4/24/2025

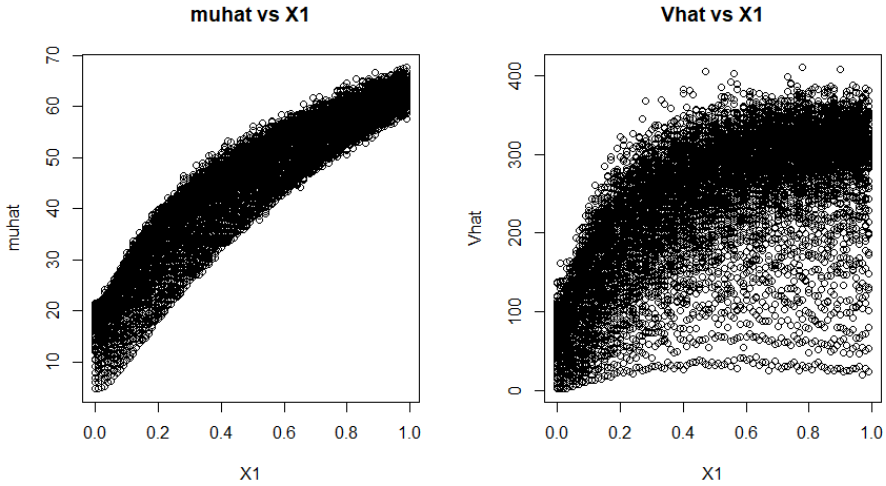
Introduction

In this final, the goal is to predict the mean and variance of a random variable, *Y*, based on two predictors: X1 and X2. Exploratory data analysis is performed on the dataset to explore the variables' relationships and guide model selection. The models chosen will undergo parameter tuning, and given that the true mean and variance values are not provided, the model performance is evaluated using Monte-Carlo Cross-Validation (MCCV) over 100 iterations to compare mean squared error.

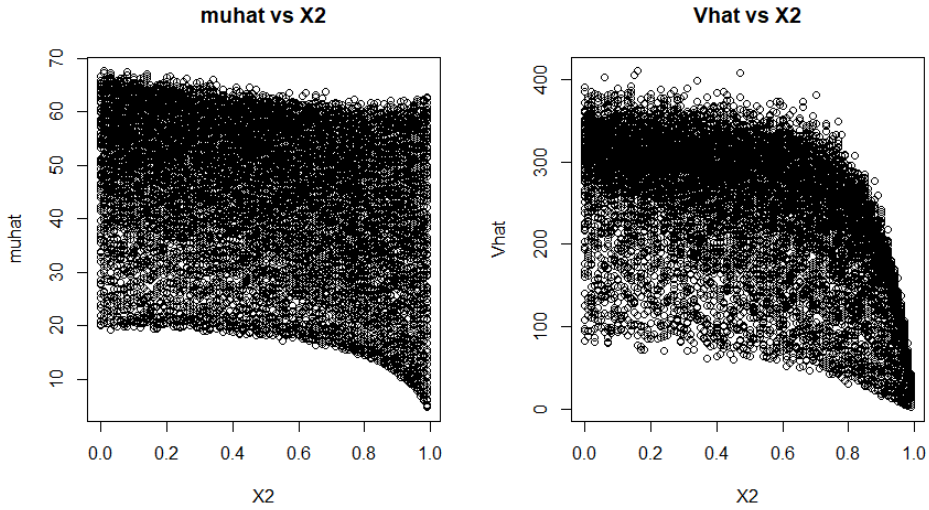
Exploratory Data Analysis



Beginning with histograms of Muhat and Vhat, the visuals above showcase skewness in the distribution of both the mean and variance. In particular, Vhat has a significantly right-skew showing that the most frequent variances are between 275 and 325.



The next visualizations show the paired visualizations of X1 when plotted against Muhat and Vhat. The relationship between X1 and muhat is mostly linear and with a slightly positive curvature, whereas the relationship to variance is nonlinear and concave.



Next, the relationship of X2 plotted against muhat and vhat. Muhat has a slight negative relationship with a relatively stable spread of mean values across X2. However, variance showcases a clear downward curvature where larger X2 values indicate significantly lower variance.

The combination of these visuals informs us of nonlinear patterns among the variables. Hence, beyond using linear regression as a baseline, the other selected models will be models that are better at handling nonlinear relationships, such as ensemble models, additive and kernel methods.

Methodology

Seven models were explored: Linear Regression, Polynomial Regression, Random Forest, GBM, SVM, Neural Network, and Generalized Additive Model (GAM).

When applicable, the models were tuned, and features were scaled. I had performed MCCV to compare MSE with and without scaling. Due to the number of models and computational limitations concerning parameter tuning, the first set of MCCV does not include a wide range of parameter choices and scaling. Nonetheless, the MCCV step provides some preliminary concept of model performance.

Afterwards, the number of models is narrowed down based on the models with the greatest minimization of mean and variance. Three nonlinear models (RF, GBM, and SVM) demonstrated better performances, and further parameter tuning and scaling were performed for a second round of MCCV.

Results

Table 1: Round 1 MCCV (100 iterations)

Model	MSE (Mean)	MSE (Var)
Linear Regression	9.16	2171.38
Polynomial Regression	3.56	962.41
Random Forest	1.46	575.46
Gradient Boosting Machine	1.31	544.88
Support Vector Machine	1.28	581.76
Neural Network	7.21	1415.51
Generalized Additive Model	3.20	698.77

As mentioned briefly in the methodology, the first round of MCCV was performed, and it was found that RF, GBM, and SVM had the lowest MSE among all models by far. Specifically, SVM had the lowest mean squared error for mean, but had the highest mean squared error for variance among the three models. GBM had the lowest MSE variance and ranked 2nd in MSE for mean. RF was included for the second round of MCCV because it was still respectably close in performance, and further tuning and additional information outweighed the cost of computing time.

Linear Regression's performance as a baseline is unsurprising, given the nonlinear relationships explored in the EDA section. Polynomial, NN, and GAM yielded MSE values that were unlikely to match the grading criteria, and therefore, further investigation was deemed unnecessary.

Table 2: Round 2 MCCV (100 iterations)

Model	MSE (Mean)	MSE (Var)
RF	1.46	570.57
GBM	1.31	543.91
SVM	1.23	541.37

With further parameter tuning, the MSE values were improved. Proper scaling implementation was also included at this stage, and this is reflected in the significant improvement in SVM's Variance MSE change. This gave me the confidence to select SVM as the model.

Conclusion

The analysis began with an exploratory data analysis to investigate the patterns and relationships of the predictor variables and their relationship with mean and variance. It was noted that there is skewness in the distribution of the mean and variance, and the predictor and response variables indicated nonlinear patterns. This helped shift the focus onto models better at tackling nonlinear relationships.

Several models were explored to predict the mean and variance of a response variable based on X1 and X2. After initially comparing seven models by their MSE in mean and variance using MCCV, it was identified that SVM, GBM, and Random Forest delivered the lowest MSE. These models were then selected for further tuning and scaling of response variables. In the end, support vector machine was selected for its robustness and yielded the lowest MSE of 1.23 for the predicted mean and 541.37 for the expected variance.

In order to find a better solution, I could have leveraged time with a wider range of tuning parameters. In addition, I could have given a fair and equal approach to all seven models in terms of tuning and/or scaling, so as to avoid having two rounds of MCCV.

Lessons I have learned

This process highlights the importance of repeated validation and justifying parameter tuning. It was challenging to discover coding mistakes or to omit essential methods when dealing with the tuning or 100 iterations of MCCV, both of which were time-consuming. A mistake I made early on was related to the scaling of the variables, which is why SVM had a much higher MSE variance in the first round. As embarrassing as it was, I included it as part of the methodology to showcase the issue and the steps I took to remedy it.