Time Series Modelling Technique Analysis for Enterprise Stress Testing

by

James Edbrooke (CID: 01290027)

Department of Mathematics
Imperial College London
London SW7 2AZ
United Kingdom

Thesis submitted as part of the requirements for the award of the MSc in Mathematics and Finance, Imperial College London, 2016-2017

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Signature and date:

Acknowledgements

I would like to thank Dr Mikko Pakkanen for his help and support throughout this project, especially for his insight and feedback on the many different modelling techniques that he has come across in his career so far. I would also like to extend my thanks to Ben Steiner, Jamila Mathias N'Diaye and Dylan Carew of the Quantitative Strategies team at CIT Bank, for their time, guidance and expertise offered over the course of the summer.

Contents

Abstract

The Federal Reserve Banks' Comprehensive Capital Analysis and Review exercise is a mandatory requirement for the largest bank holding companies, to assess the bank's capital adequacy in times of stress. With the Federal Reserve Bank pressuring bank holding companies to improve their modelling for Pre-Provision Net Revenue calculations, the need for sophisticated statistical modelling techniques in this area is becoming a topic of high focus for banks. Critical to the choice of model to be utilised is the requirement for strong statistical tests to be undertaken, for variable distributions, stationarity, serial correlation and out-of-sample performance. This paper analyses a set of potential models with an aim to forecasting time series based on stress scenarios. Relevant statistical tests are conducted, modelling assumptions are documented and tested for verification, and finally a relative comparison of models is provided.

1 Introduction

In the aftermath of the most recent financial crisis there has been a large push in the United States for enterprise wide stress testing through Comprehensive Capital Analysis and Review (CCAR). This is an annual exercise conducted by the Federal Reserve Bank (FRB) that is used to "assess whether the largest bank holding companies operating in the United States have sufficient capital to continue operations throughout times of economic and financial stress and that they have robust, forward-looking capital-planning processes that account for their unique risks."[1]

Stress testing models are designed to simulate how a bank holding company will respond under various scenarios forecasting adverse operating conditions, and are used to provide an idea of the potential losses that could arise from an economic downturn. One of the many modelled aspects of the balance sheet of a bank is the inputs to the Pre-Provision Net Revenue (PPNR) calculation, which is chiefly concerned with how variable bank earnings are, especially in a time of economic stress. These inputs include forecasting volume, pricing and prepayment rates throughout a bank, and they are used to calculate ending balances and net interest income throughout the adverse scenarios.

Classically, the majority of banks have employed simple regression models as the most common method to model the scenarios devised by the FRB, but with the FRB pushing banks to employ more sophisticated models, more research is being conducted into alternative techniques. A hindrance to date has been access to data, with many practitioners reporting problems obtaining good quality, reliable historic data[2], which can prove to be crucial in constructing effective models.

This paper details the research conducted into a set of modelling techniques, with a requirement to fit time series models for unspecified dependent variables using the standard list of 28 supervisory economic independent variables provided by the Federal Reserve Bank, and a smaller group of approximately 23 additional macroeconomic extension variables. Both the long run history and 9-quarter stress forecasts of the independent variables were provided, along with the historical time series data.

The first technique implemented is a multiple linear regression, which, due to its current widespread use, serves as a foundation for comparison against other techniques, and can produce effective results with prudent data analysis. The model assumes there is a linear relationship between the dependent and independent variables, and seeks to build a model this quantifies that relationship. Construction of a regression model requires initial data transformations and independent variable selection to be undertaken, and model outputs will be based on a set of assumptions,

which will be documented and tested in this paper. The results when applied to the empirical data will serve as a basis for other techniques, to compare model building, model postulations and model outcomes.

A popular branch of statistical modelling concerns the use of autoregressions, where the variable being modelled depends linearly on its own previous values in some guise. These include techniques such as Autoregressive Moving Average (ARMA), Vector Autoregression (VAR) and Generalised Autoregressive Conditional Heteroskedastic (GARCH) models. With a view to utilising the independent variables in the chosen model, the second technique that is implemented is a set of Autoregressive Integrated Moving Average (ARIMA) models, which is extended to ARIMA models with Exogenous variables. Seeking to expand the use of the initial ARIMA technique introduced by Box and Jenkins in 1976[3], these work on the basis of employing a form of autoregression with the additional ability of including other variables in the model for the dependent variable.

When dealing with a large number of independent variables, Principal Component Analysis (PCA) offers a potential method to identify patterns in the data. Invented in 1901 by Karl Pearson[4], the technique looks to compress the data by reducing the number of dimensions via appropriate transformations, without a substantial loss of information. This concept can be extended to utilise these fewer dimensions to fit a linear regression model, similar to the multiple linear regression model. Referred to as Principal Component Regression, this technique is the third to be analysed in this paper.

Machine learning has become an increasingly popular field in many financial (and non-financial) institutions, and it presents a large number of data-driven prediction algorithms that could be selected for use in this setting, including Neural Networks, Support Vector Machines and Hidden Markov Chains. Typically these techniques rely on large amounts of data to provide sufficient training cases to "learn" from, in conflict with one of the main challenges expressed by banks and regulators (that of a lack of reliable historical data). The machine learning algorithm that is implemented in this study is a K-nearest neighbours regression, which stores all available data in the form of training scenarios, to be used to predict future values of the series based on the data "nearest" the values at the time point.

2 Stress Testing

Following the financial crisis in 2008, regulatory reporting for banks was significantly increased, with a greater focus concentrated on the need for capital adequacy and stress testing. The 2010 Dodd-Frank Act was the leading source of this, which included a requirement for bank holding companies to maintain a minimum level of capital to convert to equity in times of financial stress. In 2011, banks were then required to submit documentation for CCAR, reporting on their capital management procedures, which must include a set of stress-tested scenarios. The requirement for evidence of stress testing is also prevalent in other regulatory mandates, such as the Volcker Rule and the global BASEL III.

In the United States, PPNR calculations fall under the annual CCAR carried out by the FRB. These calculations measure the net revenue forecast from asset-liability spreads and non-trading fees. They do not cover credit and market losses, which were previously the areas given the most focus by the FRB, and where banks were therefore utilising stress testing. Because the variability of PPNR estimations could exceed deviations in credit and market losses, the FRB have increased their focus on PPNR in recent times, specifically with a view to banks implementing more statistical time series methods.

Stress testing is a simulation technique designed to gauge how a response variable will react to different financial situations. When conducting CCAR, the FRB publish three supervisory scenarios for banks to perform forecasting under, which cover baseline, adverse and severely adverse conditions. The baseline scenario follows a profile similar to the average projection from a survey of economic forecasters, while the adverse scenario is characterised by weakening economic activity across all economies included in the scenario. The severely adverse scenario describes a severe global recession, accompanied by a period of heightened stress in corporate loan and commercial real estate markets. In this study, a further scenario, cstress, was created to simulate a sustained period of global economic recession.

The FRB provide quarterly historic data for a set of 20 macroeconomic variables, dating from 2000 to 2016, as well as quarterly scenario forecasts for the same set of variables, projecting values from 2017 to 2020. Table 1 provides a subset of the variables provided by the FRB, with calculated means for their historic values, and the 4 stress testing scenarios. This paper utilises these variables in the modelling techniques, but as this is primarily an exercise of statistical data fitting, less emphasis was given to the business analysis of these variables.

Scenario Mean	History	Baseline	Adverse	Severely Adverse	CStress
RealGDPGrowth	1.88	2.22	0.86	-0.33	-0.72
RealDispIncGrowth	2.43	2.44	0.95	-0.37	0.24
UnemploymentRate	6.21	4.56	6.78	8.88	8.88
CPIInflationRate	2.18	2.29	1.91	1.56	0.96
TreasuryRate3month	1.62	1.73	0.1	0.1	0.01
TreasuryRate10Year	3.73	3.12	2.6	1.28	1.28
MortgageRate	5.36	4.77	4.97	4.2	5.74
PrimeRate	4.89	4.85	3.22	3.22	3.13
DJStockMarketIndex	13,929.59	25,351.65	16,700.51	16,002.06	11,844.65
HousePriceIndex	150.41	191.65	167.74	151.43	151.43
VIX	27.84	19.66	25.2	32.86	32.86
EuroAreaRealGDPgrowth	1.18	1.53	0.16	-1.4	-3.59
EuroAreaInflation	1.75	1.59	0.88	-0.22	-0.22
EuroAreaXRate.USDtoEuro	1.22	1.04	0.97	0.96	0.96

Table 1: Selection of variables with means of historic values and the four 9 quarter forecast scenarios

The baseline scenario depicts a time of moderate economic expansion, illustrated by slow rises in Treasury yields, slow rises in equity prices, and average equity market volatility. The adverse scenario envisions a moderate recession followed by recovery, portrayed by weakening economic activity, steepening yield curves, falls in short term rates, rises in long term rates, and declining asset prices. As a heightened stress scenario, the severely adverse scenario forecasts a severe global recession, with declines in GDP growth, rises in unemployment rates, short term Treasury rates falling to near zero levels, and a severe drop in asset prices. Finally, the cstress scenario uses similar values to that of the severely adverse scenario, but with a prolonged period of severe global recession, where rates can even drop to negative values.

Stress testing scenarios put statistical models through these uncommon, but plausible, circumstances, to determine how the model deals with the likely out-of-range values, in order to mimic challenging operating environments that financial institutions could find themselves in. Subjecting models to these stress tests provides an effective gauge of the rigour of the model building process. Statistical soundness of models becomes crucial, with the need to prevent spurious correlations influencing model dynamics, and residual diagnostics requiring in depth scrutiny to ensure there is no bias affecting model output. With forecasting the primary goal of running stress tests, strong out-of-sample performance becomes essential when judging model success.

3 Modelling Techniques

3.1 Time Series Analysis

The first step to effective modelling is to analyse the time series being modelled, as well as its relationships with any independent variables that could be used in the models. In this section we will introduce some concepts that are vitally important to understanding how time series models work, which will be based on those from [5].

Definition 3.1. A univariate time series model is a discrete-time stochastic process $(Y_t)_{t\in\mathbb{Z}}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

This common probability space means that we can determine the joint distribution of the random variables $Y_{t_1}, ..., Y_{t_n}$ for any values of time $t_1, ..., t_n \in \mathbb{Z}$, for any number of time points $n \in \mathbb{N}$, which then means that it is possible to specify dependence in time.

Definition 3.2. If a process $(Y_t)_{t\in\mathbb{Z}}$ is square-integrable ($\mathbb{E}[Y_t^2] < \infty$ for all $t \in \mathbb{Z}$), then the mean function is defined by

$$\mu(t) := \mathbb{E}[Y_t], \quad t \in \mathbb{Z}$$

and its autocovariance is defined by

$$\gamma(s,t) := Cov[Y_s, Y_t] = \mathbb{E}[(Y_s - \mathbb{E}[Y_s])(Y_t - \mathbb{E}[Y_t])], \ s,t \in \mathbb{Z}$$

.

A concept that is crucial to a number of modelling techniques is the notion of stationarity. In layman's terms, stationarity means that the statistical properties (the mean, variance, and autocovariance) of the stochastic process do not change over time, but we formalise the definition below.

Definition 3.3. A process $(Y_t)_{t\in\mathbb{Z}}$ is considered strictly stationary if

$$(Y_{t_1}, ..., Y_{t_n}) \stackrel{d}{=} (Y_{t_{1+k}}, ..., Y_{t_{n+k}})$$

for any $t_1, ..., t_n \in \mathbb{Z}$, and $n \in \mathbb{N}$.

Definition 3.4. A process $(Y_t)_{t\in\mathbb{Z}}$ is considered covariance-stationary (also referred to weakly stationary), if it is square integrable and

$$\mu(t) = \mu(t+k), \ and \ \gamma(s,t) = \gamma(s+k, t+k),$$

for any $s, t \in \mathbb{Z}$ and $k \in \mathbb{Z}$.

Techniques such as multiple linear regression and ARIMAX require the modelled time series to be made stationary prior to commencing modelling, in order to avoid spurious relationships between variables impacting the models. The stationarity of a time series can be inspected by displaying time series plots, and examining autocorrelations of the given series.

Definition 3.5. The autocorrelation function of a covariance-stationary process $(Y_t)_{t\in\mathbb{Z}}$ is given by

$$\rho(k) := Cor[Y_k, Y_0] = \frac{Cov[Y_k, Y_0]}{\sqrt{Var[Y_k]Var[Y_0]}} = \frac{\gamma(k)}{\gamma(0)},\ k \in [0, 1, ...]$$

.

Graphically, the results for this computation for each value of the lag, k, can be displayed by generating autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The difference between the two is that the PACF includes a control for shorter lags, so that it shows autocorrelation that is not explained by previous shorter lags. If a time series is random, the autocorrelations would be near zero for all non-zero time lags. For a time series to be stationary, its autocorrelations would be expected to decay quickly to zero as the lag is increased, and its partial autocorrelations to show no significant values.

3.2 Multiple Linear Regression

With p distinct predictors for the dependent variable Y, multiple linear regression takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon,$$

where X_j is the jth predictor, β_j is the (unknown) coefficient of X_j (i.e. the association between that variable and Y), and ϵ is a zero mean error term, typically assumed to be independent of the X_j terms. In order to make predictions for values of Y, the coefficients of the regression $\beta_0, ..., \beta_p$ need to be estimated, to then be used in the prediction formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p.$$

These parameters can be estimated by a least squares approach, where the values are chosen to minimise the sum of squared residuals (RSS) $= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_p x_{ip})^2$.

A linear regression model is built with several underlying assumptions about the structure of the model and the variables being modelled. The first assumption is an obvious one: that the relationship between the predictors and response is linear. This refers to the belief that the change in the output variable Y due to a one unit change in a predictor variable X_j is constant, regardless of the value that X_j takes. Coupled with this is the assumption that the relationship is additive.

The understanding of this is that the effect of changes in a predictor, X_j, on the output variable, Y, is independent of the values of the other predictors. Assumptions are also made about the residuals of the constructed model. They are assumed to be normally distribution, display no autocorrelation, and show no heteroskedasticity, meaning that sub populations of the residuals do not have different variabilities to each other. An assumption common to many techniques, including models built by multiple linear regression, is that there should exist no multicollinearity in the model, which occurs when predictors used in a model are strongly correlated with each other. Models containing multicollinearity in predictors could lead to erratic changes in coefficient estimates through only small alterations in the model or the data. In practice, these assumptions are often broken, and these will be rigorously tested as part of the empirical application in this paper.

3.2.1 Variable Selection

As alluded to in the introduction, when dealing with a large number of potential predictors in a model, the application of a form of variable selection is desirable. This can help increase the likelihood of only significant variables being included as inputs to models, reducing potential noise from redundant predictors, while also assisting in removing any variables displaying collinearity. Before going into detail on the approaches followed in this paper, here we introduce concepts that can be used in model evaluation when determining variables for selection.

Definition 3.6. The adjusted R-squared ($\bar{R}^2$) is a modified version of the widely known R-squared (R^2) measure of the proportion of the variance of a dependent variable that is predictable from the independent variables in the model. The formula for R^2 is $R^2 = 1 - \frac{RSS}{TSS}$ where RSS is the sum of squared residuals defined earlier in this paper, and TSS is the total sum of squares $= \sum_i (y_i - \bar{y})^2$. The formula for $\bar{R}^2$ with p parameters in a model for n observations is $\bar{R}^2 = R^2 - (1 - R^2)\frac{p}{n-p-1}$.

Preference towards using $\bar{R}^2$ rather than R^2 is due to its attempt to take into account the tendency for R^2 to increase automatically (and potentially spuriously) with the addition of further variables, which is achieved with the fractional variables to observations term $\frac{p}{n-p-1}$ in $\bar{R}^2$. The $\bar{R}^2$ provides a value for direct comparison of a set of models for a dependent variable. There are also techniques designed to compare the relative quality of statistical models, using the concept of information theory, which provides an estimate for the information lost when a model is used to represent the process that generates the data.

Definition 3.7. For a model M of data x, with k estimated parameters in the model, then the Akaike Information Criteria (AIC)[6] value of the model is defined as $AIC = -2ln\hat{L} + 2k$, where $\hat{L}$ is the maximum value of the likelihood function for the model ($\hat{L} = P(x|\hat{\theta})$, where $\hat{\theta}$ are the parameter values that maximises the likelihood function).

When the number of observations is not many times larger than the number of parameters in the model, AIC can lead to choice of models that are overfitted[7, 8]. Many therefore argue AICc, which is AIC with a correction for finite sample sizes[9], should generally be used, such as Burnham and Anderson in [10].

Definition 3.8. Under the same notation as 3.7, $AICc = -2ln\hat{L} + 2k + \frac{2k(k+1)}{n-k-1} = AIC + \frac{2k(k+1)}{n-k-1}$.

A Bayesian statistic closely related to AIC also exists, namely the Bayesian Information Criteria, developed by Schwartz in [11]. The number of parameters in a model are more heavily penalised than with the AIC statistic.

Definition 3.9. Under the same notations as 3.7 and 3.8, Bayesian Information Criteria (BIC) is defined as $BIC = -2ln\hat{L} + ln(n)k$.

There are a wide ranging set of methods that have been documented as approaches for selecting the independent variables to be used in a model. These include subset selection techniques, which seek to identify a subset of predictors that are most related to the dependent variable (using criterion based procedures such as AIC and BIC), and regularisation methods, often concerned with shrinkage methods to reduce the impact of coefficients of insignificant variables.

Remark 3.10. A number of dimension reduction procedures also exist, but for this set of procedures, this paper will solely focus on PCA, which is discussed in Section 3.5.

The following passages describe the subset selection and regularisation methods employed in this study.

3.2.2 Subset Selection approaches

Best subset selection is a form of subset selection that is carried out by fitting least squares regressions for each possible combination of the candidate independent variables.

The algorithm for best subset selection is as follows:

- Begin with a null model, $\mathcal{M}_0$, which predicts the sample mean for each observation, and contains no independent variables.
- For the p number of total candidate variables, fit all $\binom{p}{k}$ models that contain exactly k predictors, for $k = 1, ..., p$, which produces a set of 2^p potential models.
- From this set of models, pick the best model, $\mathcal{M}_k$, for each value of k, where "best" is quantified as the model with the smallest RSS, resulting in $p + 1$ potential models.
- Pick a final model from the selected $\mathcal{M}_0, ..., \mathcal{M}_p$ via chosen criteria (such as AIC, cross-validation of prediction error, etc).

Note that RSS should not be used as the criteria to choose the final model, due to this decreasing monotonically as k increases, which would therefore leave us with a model with the highest number of predictors. This method would lead to a concept referred to as overfitting the model to the data, whereby the model has a very low training error, but may well produce a high test error due to the model being overly sensitive to small fluctuations in the training data.

An issue with best subset selection is that computationally it may not be possible to apply with a very large value for p. It can also suffer from statistical problems when p is large (the larger the search space, the higher the chance of finding models that could lead to overfitting).

Alternatively, stepwise methods are an automated set of procedures, in which the process is provided with a set of explanatory variables to begin with, and in each step of the procedure, a new variable is considered for addition to or subtraction from the set of variables in the model. Typically, the process can take three distinct forms: forward selection, backward elimination, or a hybrid bidirectional form that looks to combine forward and backward approaches.

The algorithm for forward selection is as follows:

- Begin with no variables in the model, i.e. the null model $\mathcal{M}_0$.
- For the p number of total candidate variables, fit models with the addition of the variable to $\mathcal{M}_0$, and determine which produces a better model, where better is again quantified by improved RSS, and name this model $\mathcal{M}_1$.
- With the remaining $p-1$ predictors, repeat the process, and do so until all predictors are selected in a model, to produce a set of models $\mathcal{M}_0, ..., \mathcal{M}_p$.
- Pick a final model from the selected $\mathcal{M}_0, ..., \mathcal{M}_p$ via chosen criteria (such as AIC, cross-validation of prediction error, etc).

The algorithm for backward elimination is as follows:

- Begin with all candidate variables in the model, i.e. the full least squares model with p predictors, namely $\mathcal{M}_p$.
- For each of the p variables, determine which of the models containing $p-1$ predictors produces a better model, where better is once again quantified by improved RSS, and name the model $\mathcal{M}_{p-1}$

- With the remaining $p-1$ predictors, repeat the process, and do so until no predictors are selected in the model to produce a set of models $\mathcal{M}_0, ..., \mathcal{M}_p$.
- Pick a final model from the selected $\mathcal{M}_0, ..., \mathcal{M}_p$ via chosen criteria (such as AIC, cross-validation of prediction error, etc).

The hybrid method is a combination of the above two algorithms, but with considerations for both addition and subtraction considered at the relevant steps. Forward selection requires fitting one null model initially, and then subsequently $p-k$ models for the kth iteration, i.e. $1+\sum_{k=0}^{p-1}(p-k) = 1+\frac{p(p+1)}{2}$ models. Backward elimination does the same in reverse, and therefore constructs the same number of total models. It is worth noting that this is a significantly smaller number of models that require fitting than the best subset selection, which can have substantial benefits computationally.

However, as noted by numerous authors including [12], forward selection and backward elimination are not guaranteed to find the best possible model out of all possible 2^p models. For example, if the candidate variables are $x_1, ..., x_p$ for $p > 2$, if the best two variable model contains (x_2, x_3), but the best single variable model uses x_1, then forward selection will pick the best model containing x_1 and one other variable, meaning it cannot choose the best model with x_2 and x_3 as predictors. A similar argument for backward elimination can be shown, with the danger of an important predictor for smaller models potentially being discarded early on in the process, as outlined in [13]. It is also worth noting that backward elimination requires the number of samples, n, to be larger than the number of candidate variables, p, so that the full model can be fit initially. Stepwise methods also come under criticisms of fitting final models followed by reporting estimates and confidence intervals without adjusting them to take the model building process into account[14].

3.2.3 Regularisation approaches

Regularisation approaches, also known as Shrinkage approaches, fit models containing all of the predictors, using methods that constrain the coefficient estimates, shrinking these coefficient estimates towards zero. A benefit of this is that this method of coefficient reduction can significantly reduce their variance. There are a number of potential regularisation techniques that can be applied, most notably Ridge Regression and Least Absolute Shrinkage and Selection Operator (Lasso).

The coefficient estimates in Ridge Regression are those that minimise the equation

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j{}^2 = RSS + \lambda \sum_{j=1}^{p} \hat{\beta}_j{}^2,$$

where $\lambda \geq 0$ is a tuning parameter that requires determining. The $\lambda \sum_{j=1}^{p} \hat{\beta}_j{}^2$ term is referred to as the shrinkage penalty, as it has the effect of reducing the coefficient estimates towards zero as it is small when the coefficients are close to zero. Note that with $\lambda = 0$, the equation amounts to nothing more than the least squares estimate, and as $\lambda \to \infty$, the coefficient estimates will approach (but not reach) zero. A benefit of Ridge over least squares is that as λ increases, the flexibility of the Ridge Regression fit decreases, leading to a decreased variance, but this can come with an increased bias, illustrating a case of what is known as a bias-variance trade-off.

One main criticism of using Ridge Regression is that the method will shrink coefficients, but not remove any entirely. This will lead to models with potentially small coefficients, but still a large number of variables in the model. Least Absolute Shrinkage and Selection Operator (Lasso), introduced by Tibshirani in 1996 [15], goes one step further than Ridge, and allows for coefficients to be shrunk to zero, removing them from the model entirely. The coefficients of Lasso are those that minimise the equation

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \left|\hat{\beta}_j\right| = RSS + \lambda \sum_{j=1}^{p} \left|\hat{\beta}_j\right|.$$

Comparing to the minimisation problem for Ridge, we can see that the penalty term in Lasso uses a ℓ_1 penalty term $\left|\hat{\beta}_j\right|$ rather than the ℓ_2 penalty term $\hat{\beta}_j{}^2$ used in Ridge. This way, the ℓ_1 penalty forces some coefficients to be zero with a sufficiently large value for λ. Consequently, due to the variable selection the Lasso performs, this allows the Lasso to produce sparse models that will be easier to interpret than models created by Ridge.

The commonly used technique chosen to determine the λ parameters in the Ridge and Lasso methods is k-fold cross-validation. This validation technique randomly divides the training data into k parts of approximately equal size, and the first of these parts (or "folds") is held back as a validation set (i.e. a set to be used for prediction tests), with the method fitted on the remaining folds. To assess the fit against the validation set, the mean squared error (MSE) is calculated, by computing

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ is the model prediction for the ith observation. This procedure is repeated for the k folds, where a different group of observations is used for the validation set in each run of the procedure. The k-fold cross-validation value is then computed as the average of each of the computed MSEs

$= \frac{1}{k}\sum_{i=1}^{k} MSE_i$. In the case of Ridge and Lasso, this technique is carried out by choosing a grid of λ values, computing the cross-validation error for each of the values of λ, and selecting the λ with the lowest cross-validation error as the tuning parameter.

Remark 3.11. Further regularisation techniques exist, such as Zou and Hastie's Elastic Net, introduced in 2005 [16]. This can address shortcomings of Lasso, such as when the number of candidate variables, p, exceeds the number of observations, n, where Lasso can only select a maximum of n variables even if more are associated with the dependent variable. In 2006, Yuan and Lin also introduced a technique called Group Lasso[17], which allows for pre-defined groups of independent variables to be selected together in a model, in an all-or-nothing manner. However, in this paper, the Lasso was deemed a sufficient technique for the data under observation.

3.3 Autoregressive Integrated Moving Average models with Exogenous variables

One of the most commonly used techniques for modelling time series are techniques based on the Autoregressive Integrated Moving Average (ARIMA) model that was first introduced by Box and Jenkins in 1976 [3]. This technique uses historical data of the time series to analyse its trend, and bases future predictions on this. As this paper stated earlier, this technique requires the time series to be stationary for meaningful results to be obtained. This enables long-term forecasts of the series to converge to the unconditional mean of the series. Before giving an explanation of the model, we define a type of series that is an integral part of ARIMA processes.

Definition 3.12. A covariance-stationary process $(Y_t)_{t\in\mathbb{Z}}$ is called white noise if its autocorrelation function $\rho(k) = 1$ for $k = 0$, and $\rho(k) = 0$ for $k > 0$. If the process has a zero-mean, and a variance of σ^2, then the notation $(Y_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ is used.

An ARIMA process is a process constructed from white noise by linear transformations, and can be broken into three parts: the autoregressive (AR) component refers to the use of the past values of the time series (Y_t) in the regression equation, the moving average (MA) component represents the error of the model as a combination of previous error terms (ϵ_t), and the integrated (I) component refers to the possible requirement of an initial differencing step due to the time series showing evidence of non-stationarity. The orders of the AR, I and MA terms are commonly given the notation (p, d, q) respectively. ARIMA models also provide the ability to include seasonal terms, which are terms that track any seasonal trends in the time series. The orders of the AR, I and MA terms in the seasonal component are commonly given the notation $(P, D, Q)_s$, where s is the length of the season cycle (e.g. for quarterly data, $s = 4$ would be a yearly cycle).

A seasonal ARIMA$(p,d,q)(P,D,Q)_s$ process can be written as

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{k=1}^{P} \Phi_k y_{t-ks} + \epsilon_t - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} - \sum_{m=1}^{Q} \Theta_m \epsilon_{t-ms},$$

with y_t having undergone d and D difference and seasonal difference operations respectively to ensure stationarity, where ϕ_i, Φ_k, θ_j and Θ_m are the parameters of the process. This can be written instead using backshift operations as

$$\phi_p(\mathcal{B})\Phi_P(\mathcal{B}^s)z_t = \theta_q(\mathcal{B})\Theta_Q(\mathcal{B}^s)\epsilon_t,$$

where $z_t = (1-\mathcal{B})^d(1-\mathcal{B}^s)^D ln(y_t)$.

Box-Jenkins followed an iterative approach to fitting ARIMA models, following three stages of: identification of the ARIMA terms through examination of ACFs and PACFs, estimation of models and their coefficients fitted by least squares or maximum likelihood (with comparisons between models using AIC and BIC), and diagnostic checking the goodness of fit of a model with tests of model assumptions. The assumptions of an ARIMA model are that the residuals exhibit no serial autocorrelation, they are homoskedastic and they demonstrate normality.

One of the limitations of ARIMA processes are that due to the nature of the process relying directly on historical data, they work best on long and stable series, which isn't always the case with typical financial data. Another drawback of the ARIMA methodology is it only seeks to approximate patterns in the historic data, and does not attempt to explain the structure of the data that say, PCA would. ARIMA models assume that the underlying pattern of a time series will continue to stay the same in the predictions generated by the model, which means that it will likely prove beneficial to re-evaluate the model on regular basis, in order to incorporate new information (and new patterns) into the model.

When extending this model to include the time series of an additional variable(s), known as ARIMA models with Exogenous variables (ARIMAX), there are two distinct approaches that have been identified. The first is to incorporate the additional variable(s) into the above ARIMA equation via transfer functions, such as those described by Pankratz in [18] and Bierens in [19]. However these can lead to complicated model fitting processes when seeking to include more than one exogenous variable. Writing this model using backshift operators (using an ARMAX non-seasonal model for simplicity) gives

$$\phi(\mathcal{B})y_t = \beta x_t + \theta(\mathcal{B})\epsilon_t,$$

or alternatively

$$y_t = \frac{\beta}{\phi(\mathcal{B})}x_t + \frac{\theta(\mathcal{B})}{\phi(\mathcal{B})}\epsilon_t.$$

The interpretation of the coefficients obtained in the model can be unintuitive, due to the value of the coefficient of the exogenous variable x_t no longer being the effect on y_t of a one-unit increase in x_t (as it would be with linear regression), as it needs to be interpreted in conjunction with the lagged previous values of y_t.

For these reasons, some practitioners instead build ARIMAX models as regression models with ARIMA errors, such as Athanasopoulos and Hyndman in [20]. Hyndman presents a clear explanation of the approach in [21], which we utilise here. The set-up for the model (again for the non-seasonal ARMAX model for simplicity) is

$$y_t = \sum_{k=1}^{b} \beta_k x_{kt} + \eta_t, \quad \eta_t = \sum_{i=1}^{p} \phi_i \eta_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j},$$

where β_k are the coefficients of the b exogenous variables $x_1, ..., x_b$, and the errors of the model η_t are now modelled as an ARMA process. Switching to backshift operators, this can be written as

$$y_t = \beta x_t + \frac{\theta(\mathcal{B})}{\phi(\mathcal{B})}\epsilon_t.$$

Now the βs can be interpreted in the same way they are for multiple linear regression. Note for ARIMA errors, the $\phi(\mathcal{B})$ term is replaced by $\nabla^d\phi(\mathcal{B})$, where ∇ is the difference operator.

3.4 Statistical Stationarity Tests

There exists a number of statistical tests that can be utilised to assist with determining whether a time series is stationary. In this paper we draw on the augmented Dickey-Fuller, which stems from the Dickey-Fuller test (Dickey and Fuller 1979, 1981[22][23], Said and Dickey 1984[24]) and the Kwiatkowski-Phillips-Schmidt-Shin (Kwiatkowski et al. 1992[25]) tests.

The Dickey-Fuller test is a unit root test, performed by constructing a null hypothesis of a unit root being present in an AR model. The formulation of the model is as follows:

For the series $(y_t)_{t\in\mathbb{Z}}$, take an AR(1) model, $y_t = \rho y_{t-1} + \epsilon_t$, where ρ is a coefficient, and $\epsilon_t \sim WN(0, \sigma^2)$. A unit root is present, indicating the series is non-stationary, if $\rho = 1$. Therefore the time series converges to a stationary series (as $t \to \infty$) if $|\rho| \leq 1$.
The regression equation used for the test is

$$\Delta y = (\rho - 1)y_{t-1} + \epsilon_t = \gamma y_{t-1} + \epsilon_t,$$

where Δ is the lag operator, and $\gamma = \rho - 1$, meaning that testing for a unit root can be carried out by testing if $\gamma = 0$.

In order to extend this test out to a series that doesn't necessarily follow the AR(1) process, below is a description of the augmented Dickey-Fuller test, which allows for general autoregressive

moving average models with unknown orders (ARMA(p,q)), the application of which enables the removal of any autocorrelation in the time series for testing.

The augmented Dickey-Fuller (ADF) test also tests the null hypothesis that there is a unit root present in the time series, with the alternative hypothesis being that the time series is stationary.
The regression equation used for the test is

$$\Delta y = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_{p-1} \Delta y_{t-(p-1)} + \epsilon_t,$$

where α is a constant, β is the coefficient of a time trend, and p is the lag order of the autoregressive process. This allows for a unit root test to be performed with a null hypothesis of $H_0 : \gamma = 0$ against a one-sided alternative hypothesis of $H_A : \gamma < 0$.

The basis for the test is to compute the t statistic for γ with the null hypothesis of the presence of a unit root. The t statistic is computed as $\frac{\hat{\gamma}}{SE(\hat{\gamma})}$, where $SE(\hat{\gamma})$ is the standard error of the estimate of γ, and this is then compared to the relevant critical value for the Dickey-Fuller test[24]. The null hypothesis is rejected if the test statistic is less than this critical value.

Alternatively, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests for either the stationarity of the series around a deterministic trend, known as "trend stationarity", or the stationarity of the series around a fixed level, known as "level stationarity".

The KPSS test assumes that the series $(y_t)_{t \in \mathbb{Z}}$ can be decomposed into a linear regression model as follows:

$$y_t = w_t + \beta_t + \epsilon_t$$

where w_t is a random walk, β_t is a deterministic trend, and ϵ_t is a stationary error. Note that w_t defined as a random walk infers that it can be modelled as $w_t = w_{t-1} + v_t$, where v_t is an independent, identically distributed $N(0, \sigma_v^2)$ error.
For the test with a null hypothesis of the series being trend stationary, this translates to $H_0 : \sigma_v^2 = 0$. This results in the intercept being a fixed value, so that the residuals, e_t are from a regression on y with an intercept (w_0) and a time trend (βt), i.e. $e_t = \epsilon_t$. The alternative hypothesis in this case would be a positive value for σ_v^2, $H_A : \sigma_v^2 > 0$.
When the test is carried out with a null hypothesis of level stationarity, the null hypothesis translates to $H_0 : \beta = 0$. This subsequently means there is no time trend in the series, and the residuals are from a regression on y with an intercept only, i.e. $e_t = y_t - \bar{y}$.

3.5 Principal Component Regression

Dimension reduction techniques seek to take a large data set of variables, and reduce the dimensions while losing the minimal amount of useful information from the data. One such technique is Principal Component Analysis (PCA), which is an unsupervised approach used to reduce the dimensions of an n x p data matrix X, by constructing a set of uncorrelated linear combinations of the p features that contain as much variation of the data as possible. The outline of the technique here follows that described in [26].

The first principal component of the p features is the normalised linear combination

$$Z_1 = \phi_{11}X_1 + ... + \phi_{p1}X_p$$

that has the largest variance, where normalised means $\sum_{i=1}^{p} \phi_{i1}^2 = 1$. The vector $(\phi_{11}, ..., \phi_{p1})^T$ is referred to as the loading vector ϕ_1. The terms $\phi_{11}, ..., \phi_{p1}$ are referred to as the loadings of the first principal component, and these are constrained so that their sum of squares is equal to one, to prevent arbitrarily large variances. Computationally, this is achieved as an optimisation problem:

$$\max_{\phi_{11},...,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \quad subject\ to \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1,$$

with the requirement for each variable to have been centred with mean zero, prior to being used in this computation. The second principal component is the normalised linear combination of the p features that has the highest variance out of all potential linear combinations that are uncorrelated with the first principal component. This is achieved by confining the direction of ϕ_2 to be orthogonal to ϕ_1. The solution to the optimisation problem can be obtained via an eigenvector decomposition, or a singular value decomposition. This paper does not detail the full workings of the decompositions, but the extensive explanations of which can be found in [27].

Principal Component Regression (PCR) extends this method to be used as a way to predict values of a variable based on the principal components generated by PCA. The technique uses the principal components as regressors in a linear regression model for a response variable y_t, fit by least squares. The overriding assumption with the technique is that often a small number of principal components are required to explain the majority of the variance in the data, and that this variability is in the same direction as the variability for the response variable. Therefore a dimension reduction is achieved by only requiring this small number of principal components in the model. The decision on the number of principal components to use as regressors in a model is typically carried out by cross-validation, following the same process as explained in Section 3.2.3.

PCR also has the benefit of avoiding a form of overfitting in models, because if the assumption holds, most of the information in the original independent variables is contained in the fewer principal components being used as regressors in the model. It also succeeds in preventing any multicollinearity between predictors, due to the nature of the construction of the principal components requiring them to be uncorrelated with each other. However, it is worth noting that PCR is not a feature selection method, as each of the principal components is a linear combination of the original variables, which in turn means the final model is more difficult to interpret. This also means a complete data set is required to be gathered and maintained to perform the technique.

3.6 K-Nearest Neighbours Regression

Techniques such as multiple linear regression and ARIMAX are examples of parametric methods, which make assumptions about the form of the function of the underlying data. An alternative set of nonparametric approaches that do not make these assumptions fall under the moniker machine learning. One of the most commonly used is the K-nearest neighbours (KNN) algorithm, which can be employed in a classification setting, or in a regression model.

The KNN regression method takes a value for K and a point it is looking to compute a prediction using, x_0, and determines the K training observations that are nearest to x_0, denoted by $\mathcal{N}_0$. The concept of nearest here can be computed in a number of different ways, including Manhattan, Minkowksi and Hamming distances, but the most widely used is the Euclidean distance formula. This is the straight line between two points (u, v) in Euclidean space, calculated in a p-dimensional space as

$$d(u, v) = \sqrt{(u_1 - v_1)^2 + ... + (u_p - v_p)^2}$$

. In order to remove bias from higher ranged variables, each of the features should be normalised before being input into the algorithm. The algorithm then estimates the value for y_0 by taking the average of the training responses in $\mathcal{N}_0$, i.e. its nearest K neighbour values:

$$\hat{y}_0 = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

Choosing the value for K is often the most difficult part of the technique, and a widely agreed approach is yet to be settled upon. A larger value of K will result in reduced variance in predictions, but can ignore smaller patterns in the data. Generally the value for K is set to be the square root of the number of observations, as first coined by [28], but utilising techniques such as k-folds cross-validation (note that $k \not\equiv K$ here) discussed in Section 3.2.3, for values of K around

this initial level can help identify the optimal value for K.

Though it is simple to understand and fit, and does not make any assumptions about the underlying data that say a parametric technique would, KNN regression can prove to be very expensive computationally, due to its requirement to log and store all the training cases. Additionally, as the method determines predictions using means of previously occurring values of the response variable, it is worth noting that this prevents the method predicting values outside of the range already witnessed in the training data. This can lead to issues if training data has a limited scope for extreme events.

4 Application to Empirical Data

This section demonstrates the application of the techniques described in this paper to the empirical data set provided for the stress testing scenarios devised by the FRB. This begins with a study of the two time series under consideration (Volume and Pricing), before appropriate data transformations are carried out to then construct the models for the techniques. Tests to assess the accuracy of the assumptions of each model are examined, before an analysis of the overall performance of each model is presented.

4.1 Time Series Analysis and Stationarity

Figures 1 and 2 display the time series plots of the two dependent variables. It is noted that the Volume time series contains consistent spikes every fourth quarter of the year, with a subsequent sharp drop in the first quarter of the following year, indicating that there is a seasonality (a seasonal trend or pattern) in the series. This can be confirmed by generating a seasonal plot of the series, which produces a plot of the quarterly values for each year of the series.

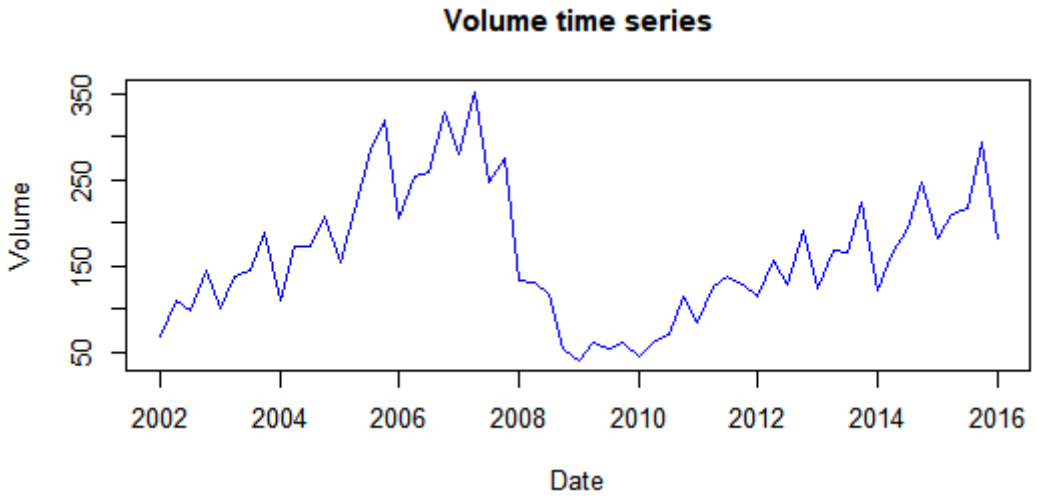

Figure 1: The Volume time series

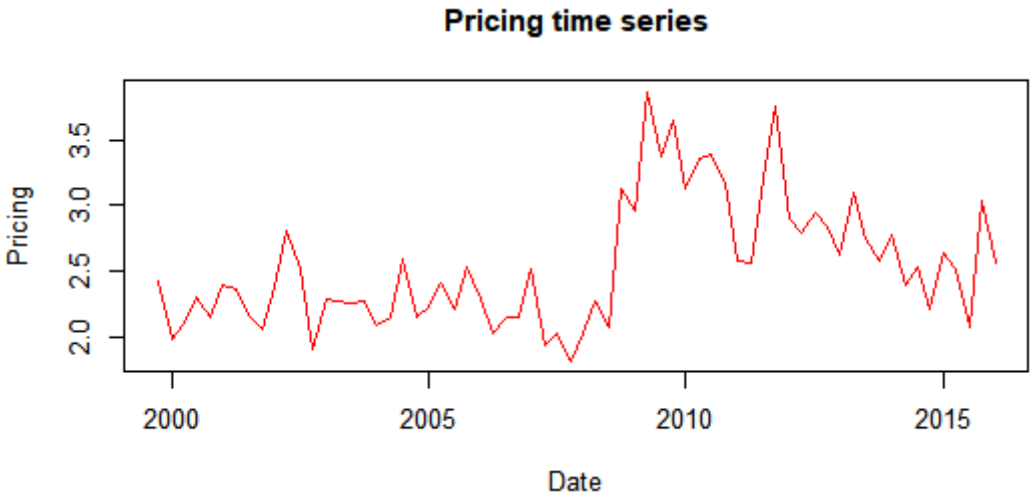

Figure 2: The Pricing time series

Figure 3 confirms the pattern of a series that generally increases over the course of the year, peaking in the fourth quarter.

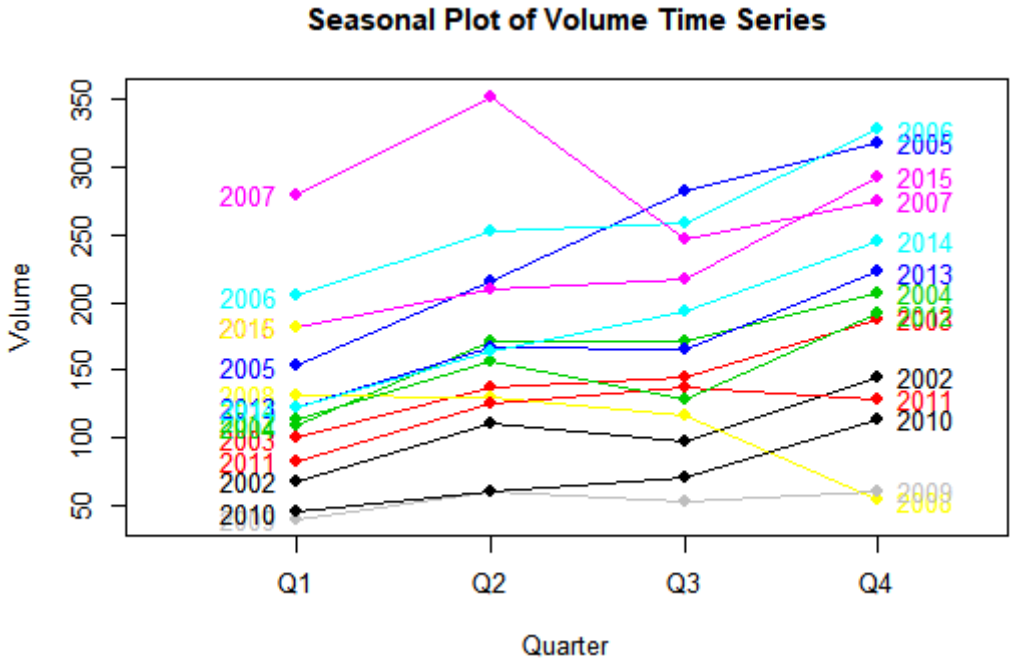

Figure 3: Seasonal Plot for Volume time series

From Figures 1, 2 and 3, visually it appears that neither series is stationary, but this can also be confirmed by examining their autocorrelation plots, and carrying out statistical tests described in this text.

Figures 4 and 5 show the ACF and PACF plots for the two series. For a stationary time series, we would expect the autocorrelation to decay quickly to zero as the lags increase. The Volume time series (left hand plots) displays slowly decaying autocorrelations, and significant correlations with recent previous values of the series. This is indicative of a nonstationary series. The Pricing time series (right hand plots) also shows a slow decaying autocorrelation function, and significant autocorrelations for smaller lags, again indicating the series is not stationary.

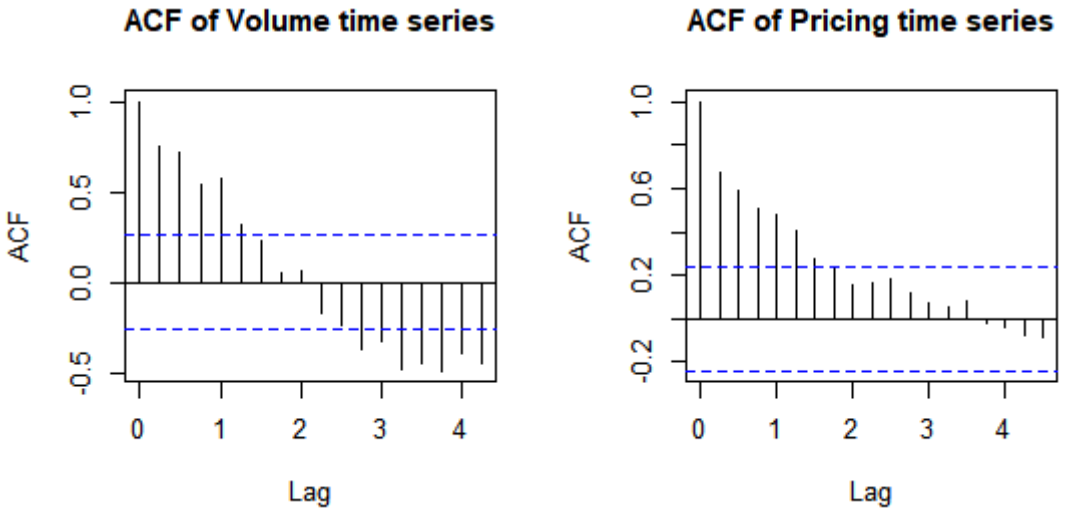

Figure 4: ACFs of Volume (Left) and Pricing (Right) time series

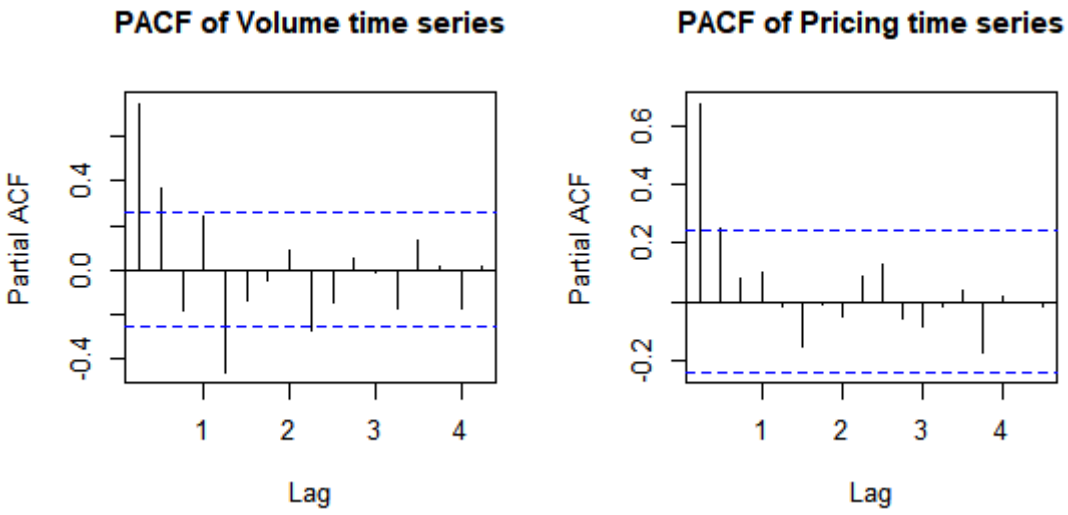

Figure 5: PACFs of Volume (Left) and Pricing (Right) time series

ADF tests were carried out in R using the function adf.test(), in the tseries package (v0.10-42, K. Hornik, A. Trapletti, 2017). This function computes p-values for the test using a simplified procedure by interpolating values from Table 4.2 in [29], where the null hypothesis (unit root) should be rejected if the p-value is significant. A lag order needs to be specified in the function, and based on the series being quarterly data, the lag was set to 4 to test yearly autocorrelation. The function kpss.test(), also in the tseries package, was used to conduct the KPSS tests. Both versions of the test (trend stationary or level stationary) were carried out, and the function interpolates the p-values from Table 1 of [25]. Unless a test returns a significant value, the null hypothesis of stationarity should not be rejected.

Series	**ADF (p-value)**	**KPSS Trend-Stat (p-value)**	**KPSS Level-Stat (p-value)**
Volume	0.4175	> 0.1	< 0.01
Pricing	0.5896	< 0.01	< 0.01

Table 2: Table of statistical test results for Volume and Pricing time series

Table 2 presents the results of the statistical tests for the Volume and Pricing time series. The Volume series passes the trend-stationarity test as expected (by the consistent trend in the majority of the series). However it fails the level-stationarity test (significant p-value so reject stationarity), and the ADF test (insignificant p-value so no reason to reject the null hypothesis of presence of unit root). The Pricing time series fails all three of the statistical tests, with the ADF test inferring there may be a unit root present, and the KPSS tests both producing small enough p-values to

reject the null hypotheses of trend and level stationarity. From the visual inspection, autocorrelation plots, and now statistical tests, it is clear that neither the Volume nor the Pricing time series are stationary in their current guise, so require transformations in order to obtain stationary series.

A number of different data transformations can be implemented in order to search for a stationary series. In this paper the following techniques were applied to the dependent variables:

- *Remove Seasonality* = calculate and remove the seasonal means from the data, i.e. $y_t - Q$ where Q is the quarterly seasonal averages.
- *First Difference* = calculate the series of changes from one period to the next, i.e. $y_t - y_{t-1}$.
- *Logarithms* = to remove any potential exponential properties in the data, take logarithms of each value of the time series, i.e. $log(y_t)$.
- *Percentage Change* = calculate the percentage growth of the time series from one period to the next, i.e. $\frac{y_t}{y_{t-1}} - 1$.
- *Second Difference* = as with the first difference, but calculating the series of changes from one first difference to the next, i.e. $(y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$.
- Combinations of the above, such as first difference with seasonality removed.

Here we present the results of the time series concluded to be stationary, which are a first difference with seasonality removed series for Volume, "VDiffX", and a first differenced series for Pricing, "PDiff".

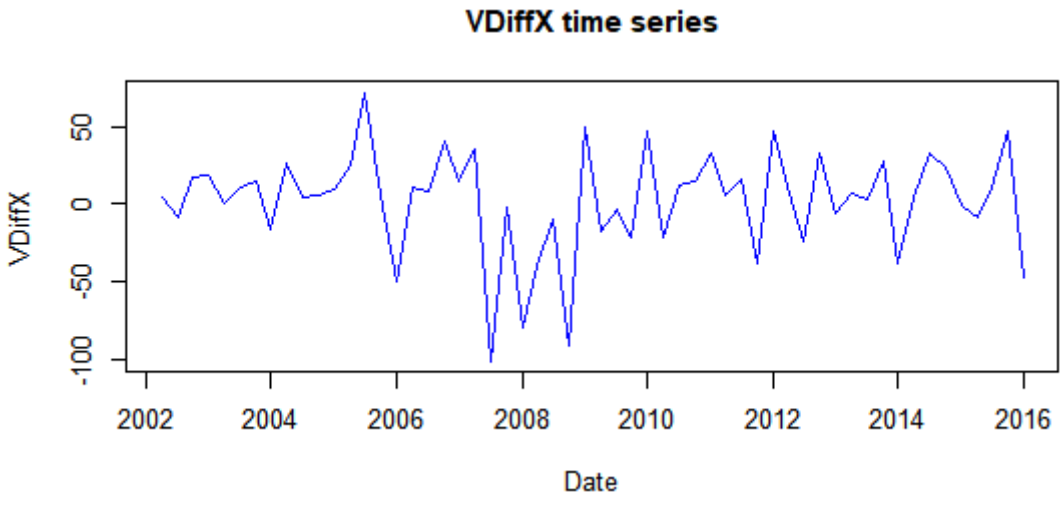

Figure 6: The Differenced Volume excluding Seasonality time series

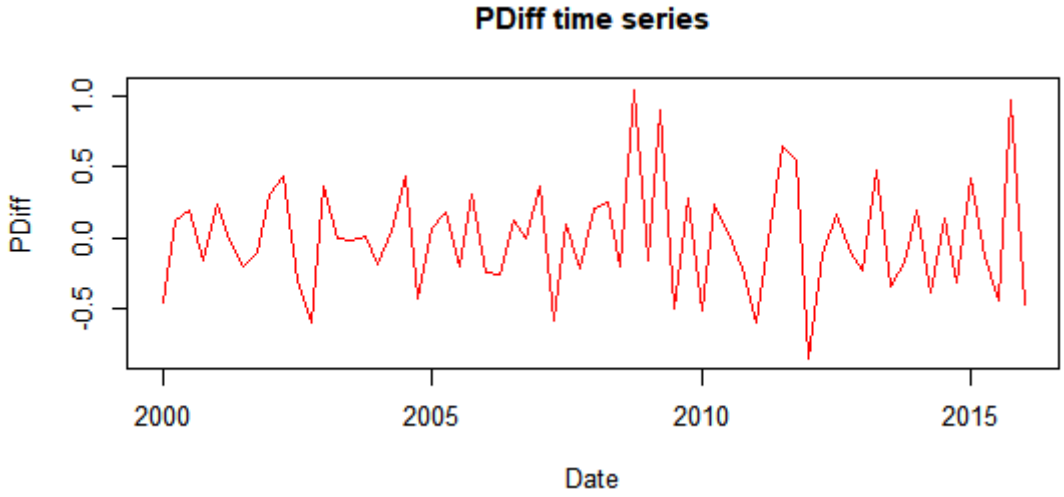

Figure 7: The Differenced Pricing time series

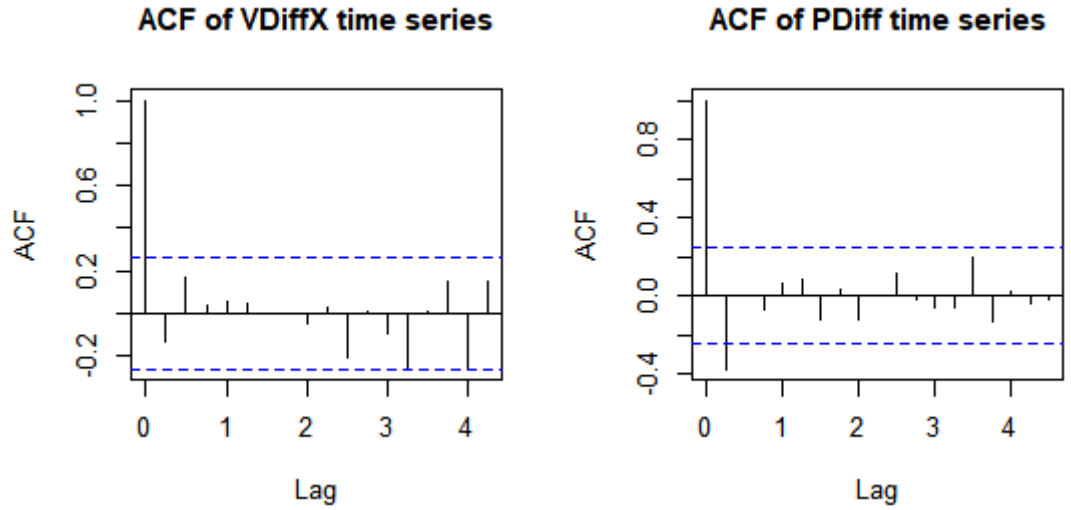

Figure 8: ACFs of Differenced Volume excluding Seasonality (Left) and Differenced Pricing (Right) time series

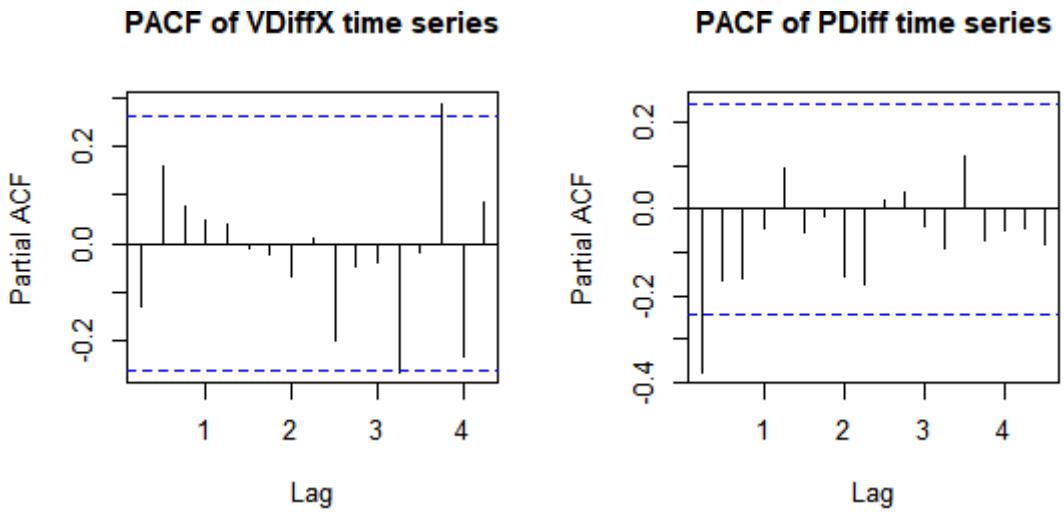

Figure 9: PACFs of Differenced Volume excluding Seasonality (Left) and Differenced Pricing (Right) time series

Figures 6, 8 and 9 demonstrate that the VDiffX time series visually appears stationary, and the autocorrelation plots confirm that the series shows no serial autocorrelation, as one would expect with a stationary series. Figures 7, 8 and 9 also depicts a series that stays around a constant mean with a fairly consistent variance, and the ACF again decays quickly to zero as expected for stationarity. It is however noted that there is a potential significant correlation remaining with the first lag of the PDiff series.

Series	**ADF (p-value)**	**KPSS Trend-Stat (p-value)**	**KPSS Level-Stat (p-value)**
VDiffX	0.3684	> 0.1	> 0.1
PDiff	0.04584	> 0.1	> 0.1

Table 3: Table of statistical test results for VDiffX and PDiff time series

Table 3 contains the results of the statistical tests for the two series. From the table, we can see that the VDiffX series passes both the KPSS tests for stationarity, but still does not have a small enough p-value in the ADF test to reject the null hypothesis of the presence of a unit root. However, as the series passes the majority of the stationarity tests, it was deemed to be sufficiently stationary to be used in the modelling techniques. The table also confirms that the PDiff series passes both KPSS stationarity tests, with sufficiently high p values not to reject the null hypothesis of stationarity. The series also passes the ADF at a 5% significance level, rejecting the unit root hypothesis. While more complicated transformations, such as second differencing with/without seasonality excluded, may produce more definitive stationarity test results, it should be noted that

it would require strong business rationale to utilise a time series with increasingly complicated transformations, due to the potential for unintuitive forecast results.

First differencing was also carried out to the set of independent variables, creating a further set of 51 variables. Macroeconomic data can take time to have an effect, so with this in mind, up to four quarters of lagged values of the differenced independent variables was also included in the candidate variable set, creating a total of 306 potential independent variables to use in the models. Any rows with missing data were removed prior to building the models.

4.2 Multiple Linear Regression

As described in Section 3.2.1, in order to obtain models fitted with only significant variables, application of variable selection techniques are required. Here the outcome of the implementation of the chosen methods for the two stationary dependent variables are presented. Note that for the Volume variable, a quarterly indicator variable was created and included as a predictor, due to the clear seasonality the series displayed.

Method	**Best Subset**	**Backward Elim**	**Lasso**
X1	JPNXRate.YentoUSD.D	JPNXRate.YentoUSD.D	JPNXRate.YentoUSD.D
X2	RealGDPGrowth	LIBOR1month .D.1L	CPIInflationRate.D.2L
X3	HYOAS.D	HYOAS.D	PPICapitalEquipment.D.2L
X4	PersConsumpDurable.D.1L	TreasuryRate3month.D.1L	TreasuryRate3month.D

Table 4: Variables selected by the methods for the VDiffX time series

Method	**Best Subset**	**Backward Elim**	**Lasso**
X1	RetailSales.D	RetailSales.D	RetailSales.D
X2	CorporateProfits.D.2L	ACorporateYield.D.2L	CorporateProfits.D.2L
X3	JPNRealGDPgrowth.D.3L	HYOAS.D.2L	JPNRealGDPgrowth.D.3L
X4	CapacityUtilization.D	NomDispIncGrowth.D.2L	NomDispIncGrowth.D.1L

Table 5: Variables selected by the methods for the PDiff time series

Tables 4 and 5 show the outcome of variable selection using the methods described in this paper. The notation ".D" refers to a differenced variable, and ".iL" refers to the ith lag of that variable. Note that best subset selection and forward selection chose the same variables for the models, so

only one is included here. Further note that all models contain an intercept (deemed significant by p-values), and all Volume models contain the quarterly indicator variable. Linear regression models were fitted using the respective variables selected by each technique. These models were compared against each other using a variety of model evaluation techniques.

The MSE (defined previously in Section 3.2.3) was computed for the fitted values of the model. A common technique in modelling is to leave a set of values out of the model fitting process to be used as out-of-sample (OoS) tests for the fitted model. Given PPNR models are used to make predictions over a nine quarter forecast, the stability of the model's predictions is crucial to assessing the soundness of the model. The final year's worth of data (4 quarters) were left out for this data set, and the out-of-sample MSE was computed for each of the models.

Model	**Method**	**Adj R-Sq**	**Fit MSE**	**OoS MSE**	**AIC**	**AICc**	**BIC**
vmlr1	Best Sub	0.7838	535.4086	345.483	528.7708	532.6839	546.999
vmlr2	Back Elim	0.7495	620.3657	676.5478	537.0184	540.9315	555.2466
vmlr3	LASSO	0.7559	579.3603	932.4078	537.1889	543.1889	559.4678
pmlr1	Best Sub	0.3328	0.0909	0.3256	40.5996	42.0479	53.646
pmlr2	Back Elim	0.2468	0.1026	0.2614	48.4766	49.92485	61.5229
pmlr3	LASSO	0.3172	0.093	0.3196	42.0988	43.54706	55.1451

Table 6: Results of Multiple Linear Regression model tests

The results of these tests for the selected models are detailed in Table 6. For the Pricing models (pmlr1, pmlr2, pmlr3), the models fitted with variables chosen by best subset selection and Lasso produce the strongest results, with the lowest AIC, AICc and BIC. However when conducting tests for multicollinearity with the models (constructing correlograms for variable correlation values, computing condition numbers[30] and variance inflation factors (VIF)[31]), the best subset model showed strong collinearity in the differenced Retail Sales and differenced Capacity Utilisation variables, leading to the Lasso model being determined as the most appropriate model for this technique. We can see that all Pricing models produce low adjusted R-squared scores, and perform poorly in the out-of-sample tests (comparing relative Fit and OoS MSEs), suggesting this may be a difficult series to model. With the Volume models (vmlr1, vmlr2, vmlr3), the model fitted with best subset selection variables is clearly the best performing, with the strongest scores for all tests.

Independent Variables	Coefficient	S.E.	p-value
Constant	-71.3837	6.9589187	1.09E-13
Q2	111.0939	9.8859534	4.85E-15
Q3	79.0702	10.2594435	6.12E-10
Q4	111.7578	10.1005996	8.33E-15
RealGDPGrowth	-5.8316	2.2196499	0.011521
JapanXRate.YentoUSD (Differenced)	2.8327	0.6499202	6.88E-05
HYOAS (Differenced)	-10.3476	2.8178289	6.04E-04
PersConsumpDurable (Differenced, 1 Lag)	3.5068	0.9985588	9.80E-04

Table 7: Coefficient estimates for Volume MLR model

Independent Variables	Coefficient	S.E.	p-value
Constant	0.1046541	0.0480093	0.0332
RetailSales (Differenced)	-0.0026112	0.0009158	0.00596
NomDispIncomeGrowth (Differenced, 1 Lag)	-0.0089631	0.0062301	0.15544
CorporateProfits (Differenced, 2 Lag)	-0.0013475	0.0004982	0.00888
JapanRealGDPgrowth (Differenced, 3 Lag)	-0.0123617	0.0071068	0.08709

Table 8: Coefficient estimates for Pricing MLR model

Tables 7 and 8 provide the coefficient estimates, their standard errors (the estimate for the standard deviation of the coefficients) and their respective p-values for the selected models. Examination of the p-values for the regressors in the models suggests that all of the variables in the Volume model are significant, while the Nominal Disposable Income Growth variable in the Pricing model could be deemed insignificant by its high p-value. This is likely due to the variable selection technique used to pick the variables in the Pricing model using an alternative method to p-values in the selection process (the Lasso method). Predictions for the forecast scenarios (explained in Section 2) were generated for the models, and the plots of the models actual versus fitted and out-of-sample values, along with their forecasts, are displayed in Figures 10 and 11. Note the dependent variable data was provided up to Q1 2016, while the forecast data are provided from Q1 2017, so predictions were made for the remaining quarters of 2016 using the independent variable data (provided until Q4 2016).

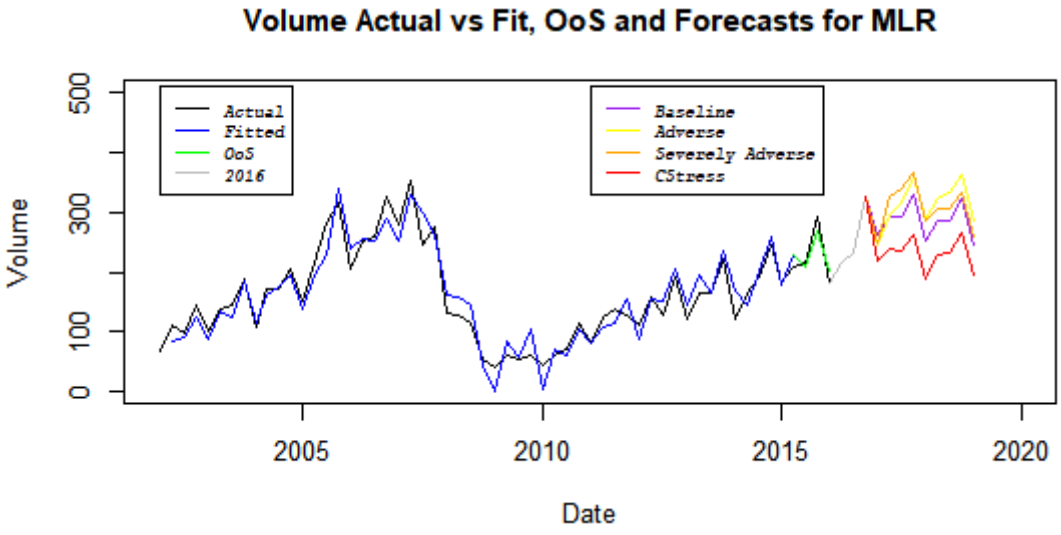

Figure 10: Plot of Actual versus Fitted and Out-of-Sample, and Forecasts for Volume MLR model

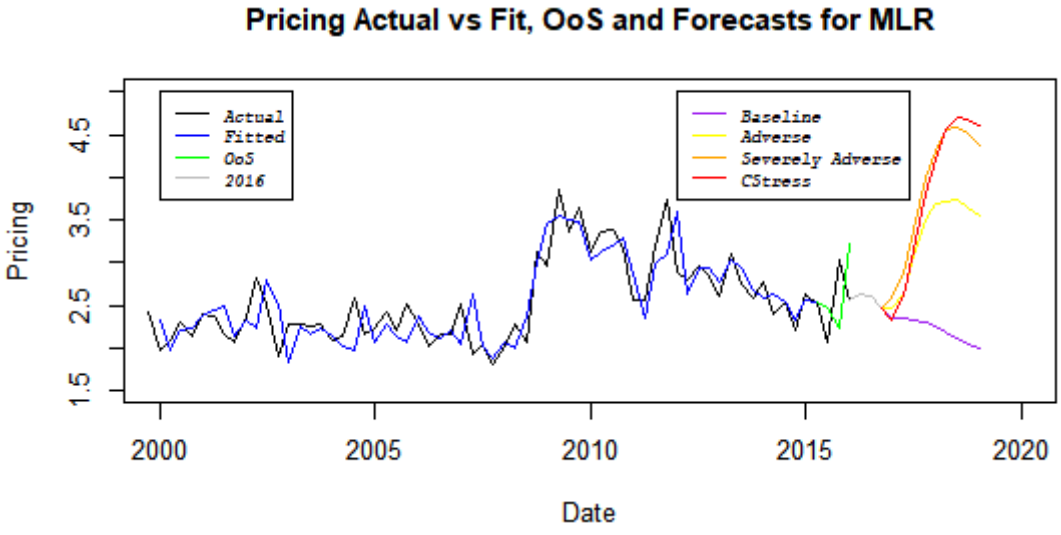

Figure 11: Plot of Actual versus Fitted and Out-of-Sample, and Forecasts for Pricing MLR model

From Figure 10 we can see that the Volume model does not perform as desired for the forecasting scenarios, with a limited impact felt by the time series in even the most severe stress scenarios. The forecasts largely replicate the seasonal pattern in the historical time series, with a slight downward trend in each scenario. Examining the data for the variables used in the model could perhaps provide some rationale for this, with some of the variables suffering a less severe impact in the forecast scenarios than during the credit crisis in 2008. A larger model with more variables being modelled could potentially have picked up the broader impact of the stress scenar-

ios. Figure 11 however shows a more desirable forecasting outcome, with the Pricing time series experiencing a sharp jump in the stressed scenarios, with the more pronounced jumps seen in times of highest stress, reaching or even surpassing levels seen during the financial crisis. Inspecting the data confirms that levels for the variables used in the model are even further stressed than during the financial crisis, helping explain the higher prices forecasted by the model.

The chosen models (vmlr1 and pmlr3) were tested against the assumptions of the multiple linear regression model outlined in Section 3.2. Normality of residuals was examined via QQPlots, where the quantiles of the residuals data set are plotted against those of a normal distribution. Heteroskedasticity of residuals was analysed through the Breusch-Pagan test[32], which tests whether the variance of the residuals is dependent on the values of the independent variables used in the regression (indicating heteroskedasticity) via a chi-squared test. Multicollinearity was tested as described in the previous passage. Both models resulted in an insignificant p-value for the BP test, meaning no reason to reject the null hypothesis of homoskedasticity, and small values for condition numbers and VIFs, indicating no presence of multicollinearity in either model. From Figure 12, the residuals for the Volume model appear to fit a normal distribution fairly well (left hand plot), but the residuals from the Pricing model look to have fatter tails than the standard normal distribution (left hand plot).

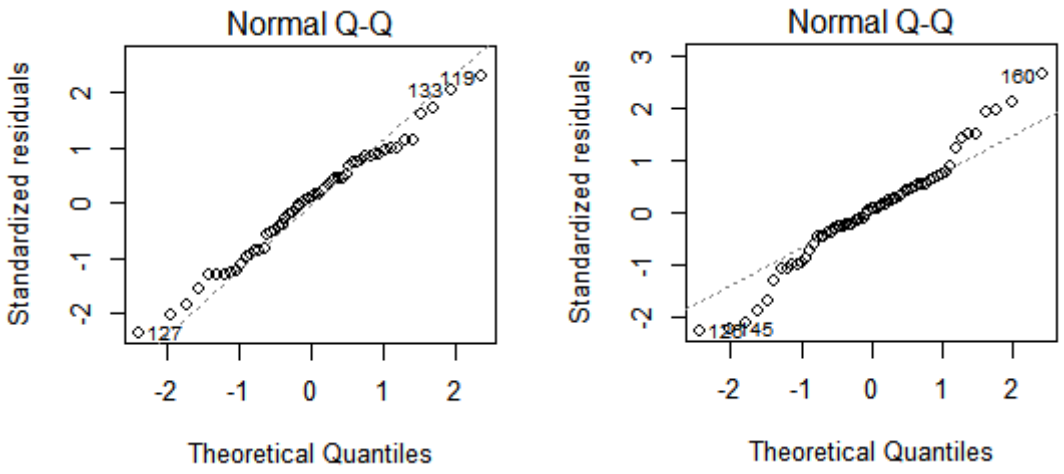

Figure 12: QQPlots for vmlr1 (Left) and pmlr3 (Right)

Further analysis of the residuals of the models was conducted, by examining residuals versus fitted values plots, and checking for any serial correlations in ACF and PACF plots. These plots are displayed below, and while neither set of residuals shows a noticeable trend in 13, the ACF and PACF plots of the Pricing model in 14 and 15 show an unwanted first-order serial correlation,

which would need correcting in further model building processes if this was to be used in practice.

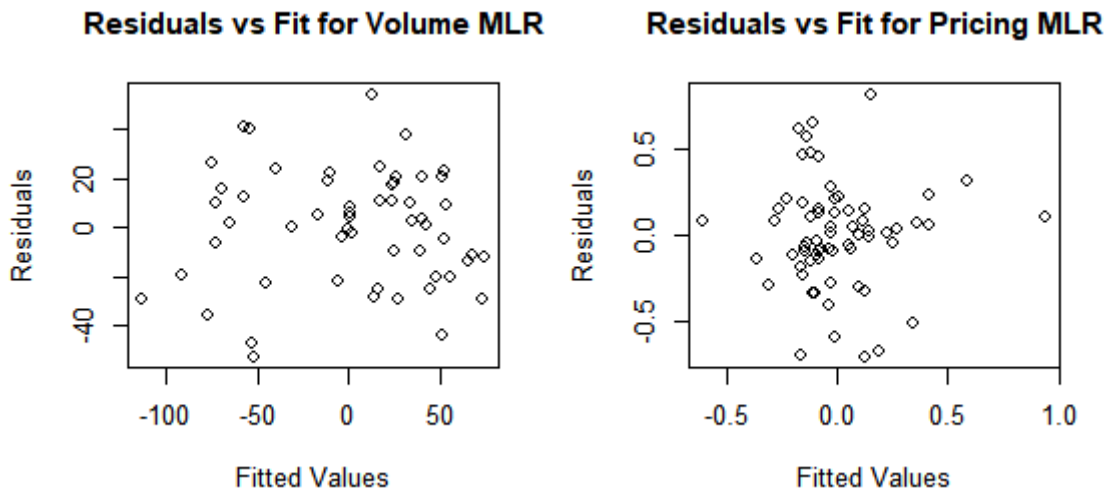

Figure 13: Residuals versus Fitted Values for vmlr1 (Left) and pmlr3 (Right)

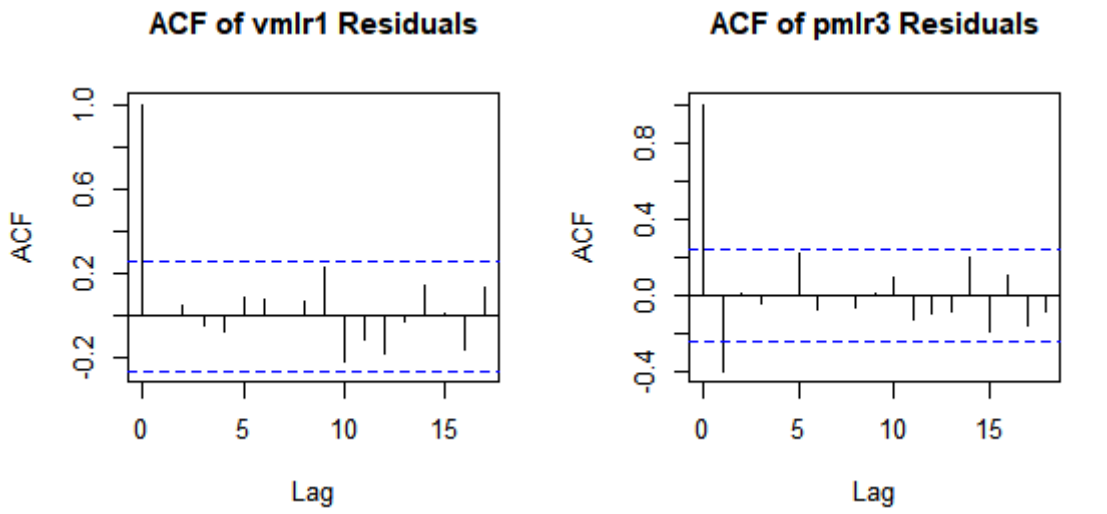

Figure 14: ACF of the residuals for vmlr1 (Left) and pmlr3 (Right)

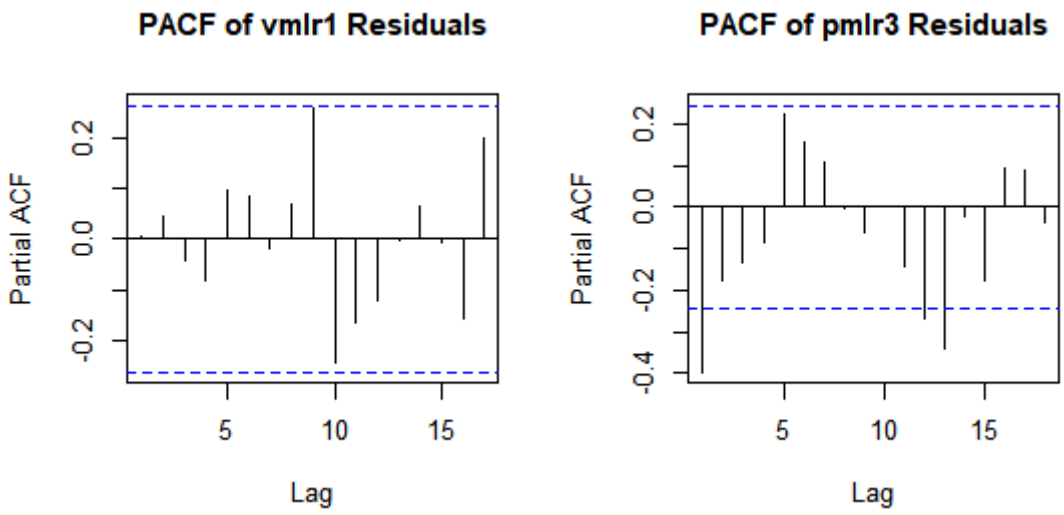

Figure 15: PACF of the residuals for vmlr1 (Left) and pmlr3 (Right)

4.3 ARIMAX

ARIMA models were fitted for the two dependent variables. The auto.arima() function in the forecast package in R conducts a search over all possible models, based on a set of order constraints, and determines the best model using a specified information criteria (AIC, AICc or BIC). From the analysis of stationarity in Section 4.1, both variables require a first difference ($d = 1$), so this can be entered as a constraint in the function. The models selected with the lowest AIC, AICc or BIC are an ARIMA$(0,1,0)(2,1,0)_4$ for Volume, and an ARIMA$(0,1,1)$ for Pricing. For the Volume model, this translates to:

$$(1 - \Phi_1\mathcal{B}^4 - \Phi_2\mathcal{B}^8)(1 - \mathcal{B})(1 - \mathcal{B}^4)y_t = \epsilon_t$$

and with coefficients:

$$(1 + 0.5148\mathcal{B}^4 + 0.2636\mathcal{B}^8)(1 - \mathcal{B})(1 - \mathcal{B}^4)y_t = \epsilon_t$$

For Pricing, the model translates to:

$$(1 - \mathcal{B})y_t = (1 + \theta_1\mathcal{B})\epsilon_t$$

and with coefficients:

$$(1 - \mathcal{B})y_t = (1 - 0.4360\mathcal{B})\epsilon_t$$

Due to the simplistic nature of the technique, and inflexible predictions it produces (largely quickly converging towards the series mean, with some seasonal effect if applicable), ARIMA models alone are not deemed as a sophisticated enough technique to provide an enhancement to the multiple

linear regression models.

The ARIMAX models were fit as linear regressions with ARIMA errors, as per the rationale provided in Section 3.3.

The model building process was carried out as follows:

- Using a set of variables most highly correlated with the stationary dependent variables, compute the coefficient and orders of a one variable ARIMAX model.
- Run the same model tests as with multiple linear regression (AIC, AICc, BIC, MSE of fitted values, MSE of out-of-sample values) for each of the models, and select the best model.
- Include the next best fitting additional variables, and re-compute the model. Re-run model tests to select the best model.
- Repeat process until a four variable model is fitted.
- Additionally, compute the models using the variables chosen in the variable selection techniques in Section 4.2.

Tables 9 and 10 display the model test results for the models built with the process described above. The models with the best scoring information criteria and MSEs were then selected for the two dependent variables.

Model	**ARIMA**	**Fit MSE**	**OOS MSE**	**AIC**	**AICc**	**BIC**
ARIMA	(0,0,0)(2,1,0)[4]	1548.442	481.2706	499.92	500.47	505.54
ARIMAX Manual	(0,0,1)(1,1,0)[4]	876.9115	4610.0083	478.2395	480.2883	489.4667
ARIMAX Lasso	(1,0,0)(1,1,0)[4]	686.7222	7060.2077	469.1782	471.9782	482.2766
ARIMAX Best Sub	(0,0,0)(1,1,0)[4]	762.3472	3757.9711	472.1584	474.2071	483.3856
ARIMAX Best Back	(1,0,0)(1,1,0)[4]	857.6598	4958.6904	475.7895	477.7283	486.6967

Table 9: Test results for ARIMA and ARIMAX Error models for Volume

Model	ARIMA	Fit MSE	OOS MSE	AIC	AICc	BIC
ARIMA	(0,0,1)	0.1113	0.1182	44.4307	44.6376	48.6524
ARIMAX Manual	(0,0,1)	0.0701	0.2894	23.5272	25.0827	36.1924
ARIMAX Lasso	(0,0,1)	0.0595	0.3802	15.4304	17.5436	30.2065
ARIMAX Best Sub	(0,0,1)	0.0593	0.3901	15.2227	17.3359	29.9988
ARIMAX Best Back	(0,0,1)	0.071	0.3175	26.4211	28.5343	41.1973

Table 10: Test results for ARIMA and ARIMAX Error models for Pricing

The tables show that the ARIMA models actually provided very accurate relative out-of-sample predictions for both models. This can likely be put down to the Volume out-of-sample values following the general upwards seasonal trend of the model, which the ARIMA model will predict, and the Pricing out-of-sample values fluctuating around a value of around the historical average of the series, which again is what an ARIMA model would predict. Looking at the information criteria scores and general MSE values for the fitted series, the ARIMA model performs poorly compared to the ARIMAX models. The selected model for both series was deemed to be the ARIMAX model built with variables chosen from the best subset selection method. These selected model details are found in Table 11, and the model parameters are listed in Tables 12 and 13 for the Volume and Pricing models respectively.

Series	Model Name	Variable Selection	ARIMA Error Order
Volume	v.arima.X4	Best Sub	(0,1,0)(1,1,0)[4]
Pricing	p.arima.X4	Best Sub	(0,1,1)

Table 11: Table of model dynamic for ARIMAX models

Independent Variable	Coefficient	S.E.	Coef/S.E.
Seasonal Autoregression(1) Component	-0.5669	0.1205	-4.7045
RealGDPGrowth	-5.198	2.1624	-2.4038
JapanXRate.YentoUSD (Differenced)	3.3727	0.6757	4.9914
HYOAS (Differenced)	-9.6173	2.4966	-3.8521
PersConsumpDurable (Differenced, 1 Lag)	3.7954	0.989	3.8376

Table 12: Table of model parameters for Volume ARIMAX model

Independent Variable	Coefficient	S.E.	Coef/S.E.
Moving Average (1) component	-0.5532	0.117	-4.7282
Constant	0.1168	0.0346	3.3757
RetailSales (Differenced)	-0.0028	0.0008	-3.5
CorporateProfits (Differenced, 2 Lag)	-0.0012	0.0006	-2
JapanRealGDPgrowth (Differenced, 3 Lag)	-0.0215	0.0069	-3.1159
CapacityUtilization (Differenced)	0.0449	0.0325	1.3815

Table 13: Table of model parameters for Pricing ARIMAX model

Inspection of the standard errors of the coefficient estimates for the models suggests that all variables look to be significant, with the only slight question mark being over the Capacity Utilisation variable in the Pricing model. Figures 16 and 17 display the performance of the model against the dependent time series along with the forecasts for the scenarios. The Volume model shows almost no impact felt by the stress testing scenarios, which would not be a desirable trait. As with the model used in multiple linear regression, this is likely due to the variables included experiencing minimal effects from the adverse scenarios, meaning this model may not lend itself well to forecasting stress scenarios. The Pricing model demonstrates behaviour more in line with expectations for the various scenarios, with sharp rises in price during times of economic recession.

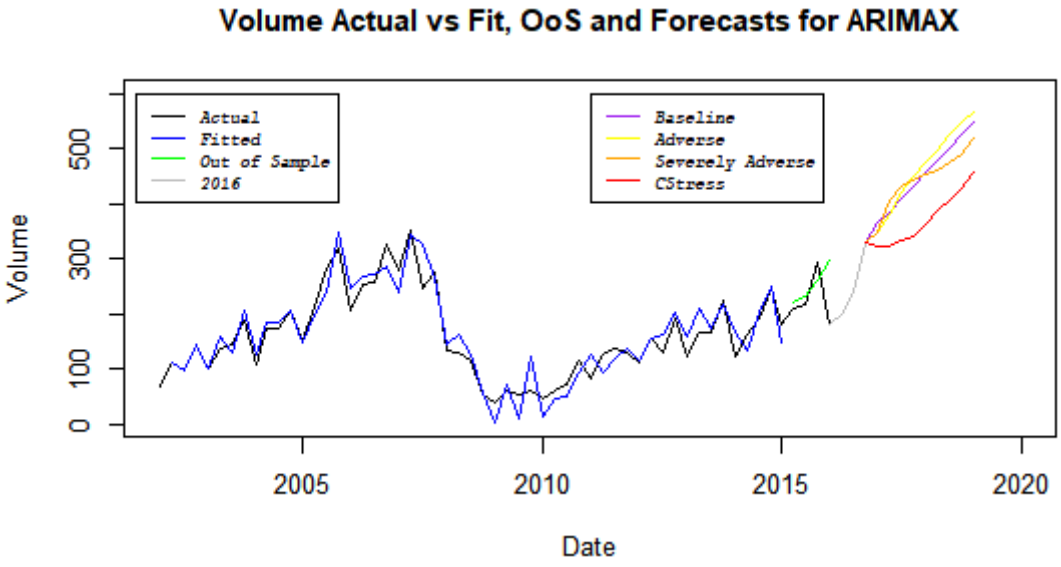

Figure 16: Plot of Actual versus Fitted and Out-of-Sample, and Forecasts for Volume ARIMAX model

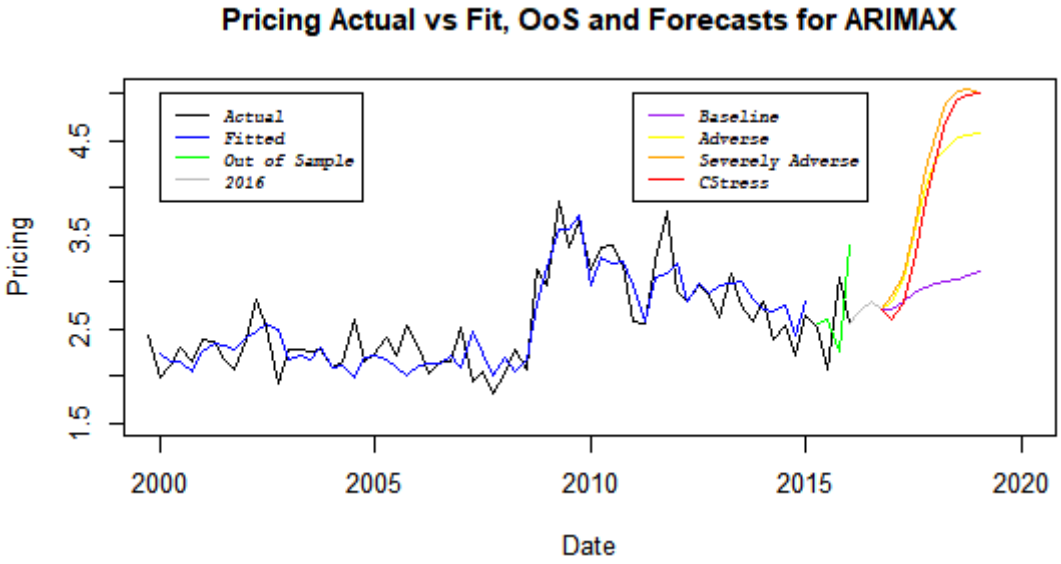

Figure 17: Plot of Actual versus Fitted and Out-of-Sample, and Forecasts for Pricing ARIMAX model

The models were then tested against the ARIMAX modelling assumptions listed in Section 3.3. Residual normality was checked in the same manner as with multiple linear regression, and homoskedasticity via a visual inspection of the time series. A useful, commonly used technique to check for no serial autocorrelation is to compute the Ljung-Box statistics for the residual time series. The Ljung-Box statistic[33] is a function of the accumulated sample autocorrelation, r_j, up to any specified time lag m. This is calculated as

$$Q(m) = n(n+2)\sum_{j=1}^{m}\frac{r_j^2}{n-j},$$

where n is the number of data points after any differencing. The value for $Q(m)$ should be around 0 for any lag, and the test is run with the null hypothesis of the data being independently distribution, and the plotted p-values for each lag should all demonstrate insignificance for this to hold.

In Figures 18 and 19 below, these tests are presented for the two ARIMAX models, and both demonstrate that the residuals are homoskedastic, normally distributed, and without serial autocorrelation.

A final further analysis of the residuals is performed by checking the residuals versus fitted plots, and the absence of patterns in Figure 20 confirms that the models have no serial correlation between residual and fitted values.

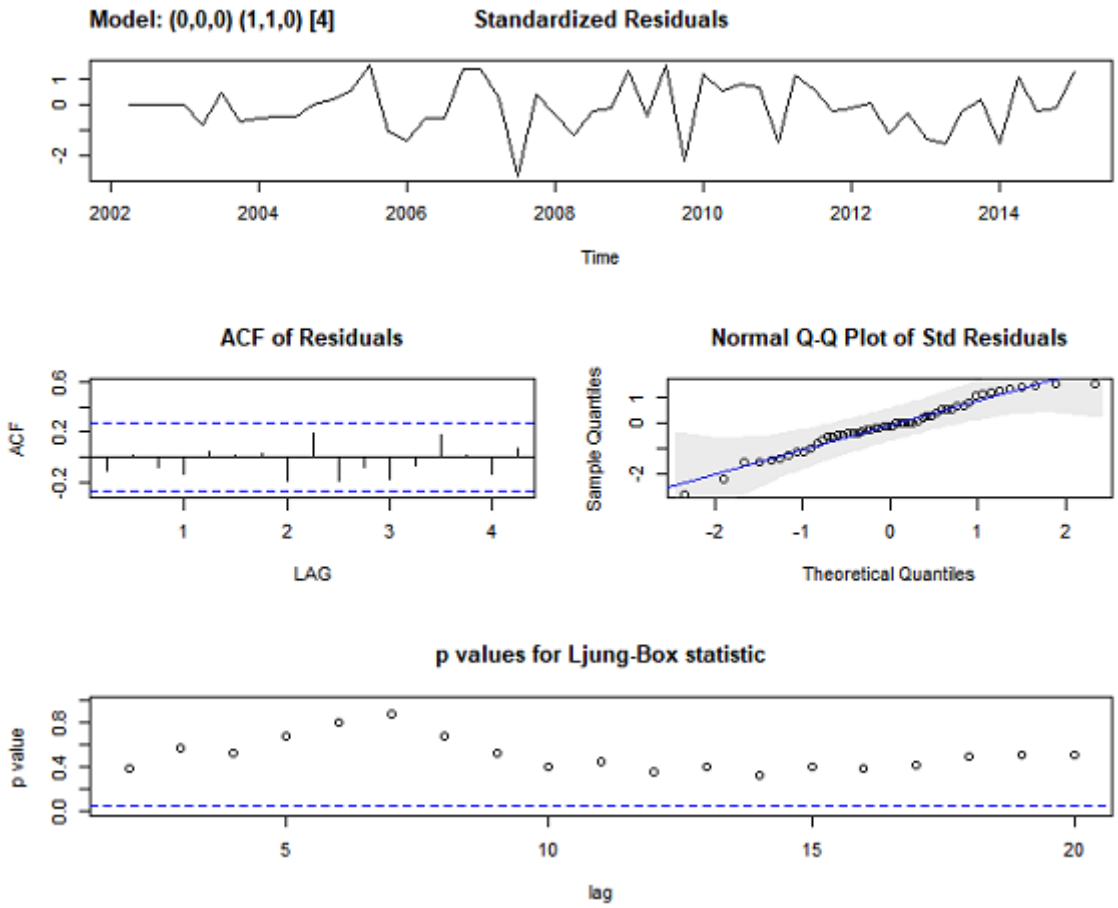

Figure 18: Plots of residual time series (Top), residuals ACF (Left Middle), residuals QQplot (Right Middle) and Ljung-Box p-values (Bottom) for residuals for v.arima.X4 model

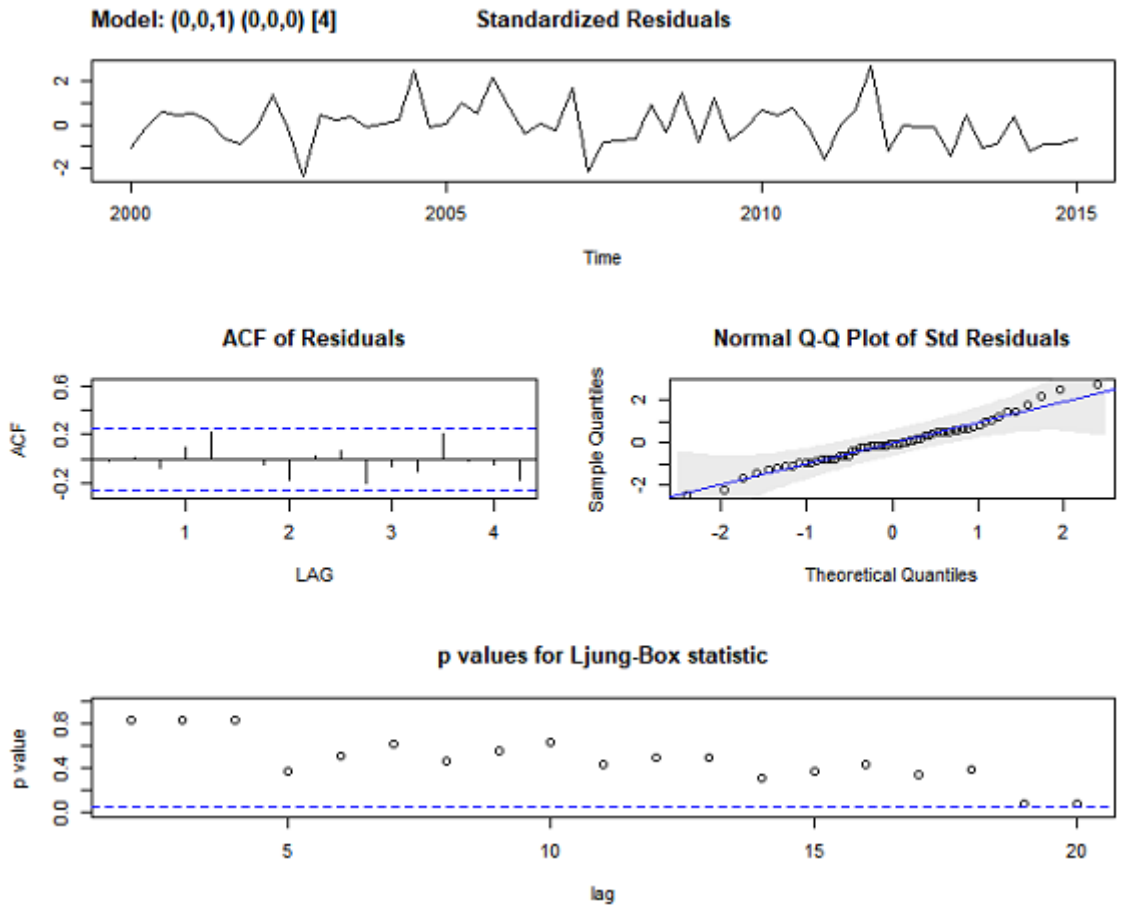

Figure 19: Plots of residual time series (Top), residuals ACF (Left Middle), residuals QQplot (Right Middle) and Ljung-Box p-values (Bottom) for residuals for p.arima.X4 model

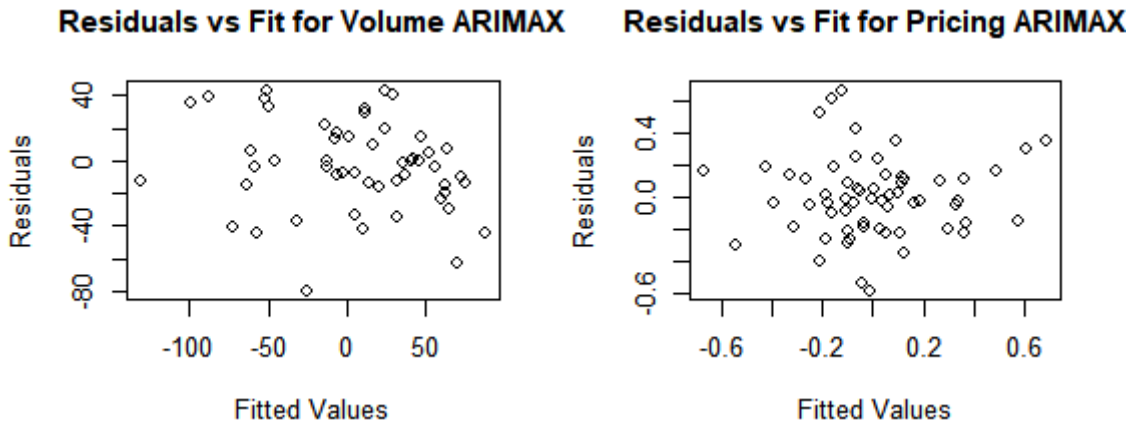

Figure 20: Plots of Residuals vs Fitted values for ARIMAX Volume (Left) and Pricing (Right) models

4.4 PCR

As PCA is not scale invariant, it is suggested that data is standardised before PCR is carried out. By applying the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}},$$

to each variable enables each to have a standard deviation of one, which will prevent the algorithm being skewed towards variables with higher absolute values, rather than those with the highest relative variance.

Models were built for both the untransformed dependent variables against the original independent variables, and also the differenced dependent variables against the differenced independent variables. The analysis of the models fitted to the untransformed variables is detailed below.

Remark 4.1. The differenced model proved to be a poor fit for both dependent variables, with a very high number of principal components required to explain enough of the variance in the response variable.

The results of the computation of the MSEP (which corresponds to MSE) calculated by cross-validation, and adjusted cross-validation (which tries to adjust for overestimation in cross-validation)[34] for each number of principal components is presented in Figure 21. For Volume (left hand plot), the lowest value looks to be obtained for under 5 components, and for Pricing (right hand plot) the minimal value is somewhere between the 10 and 20 component mark, though it is noted that the MSE only reduces by a small amount when the number of components is increased above 7.

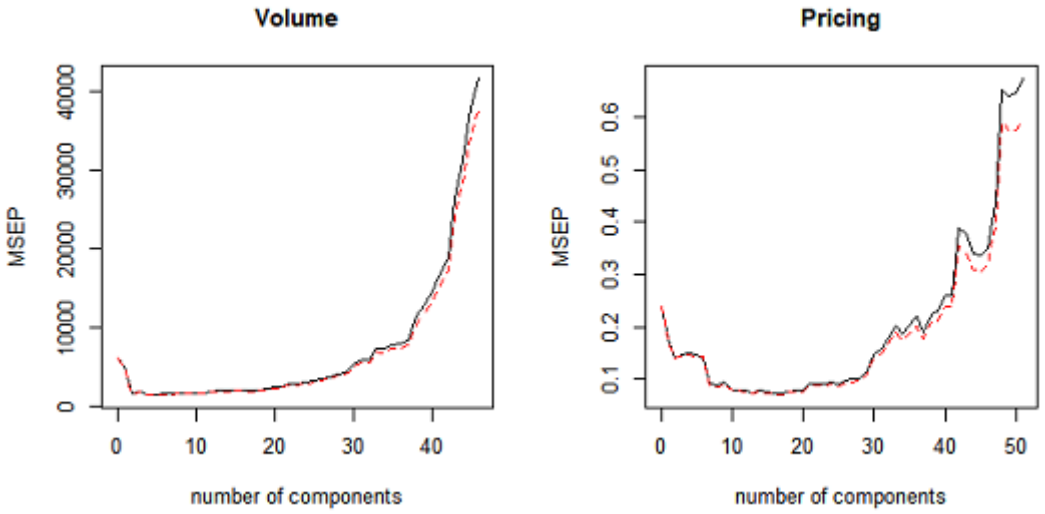

Figure 21: Number of components versus Mean Squared Error Prediction (MSEP) computed by cross-validation for Volume (Left) and Pricing (Right)

Analysis of the amount of total variance explained in the independent and dependent variables by models of varying number of components can be seen in Tables 14 and 15. From Table 14 we can see that up to 4 principal components, the assumption of a consistent direction of variance for the dependent and independent variables holds, with both having around 80% of their variance explained by the 4 components. The smaller patterns in the dependent variable become harder to ascertain with the principal components built for the low variance vectors however. The assumption proves to be less valid for the Pricing model, with a substantially smaller amount of the variance for the dependent variable explained by the components containing the variance in the independent variables. The model for Pricing was fitted using 7 components, with around 75% of the total variance explained for the dependent variable, and around 90% of the total variance of the independent variables explained, and a relatively low MSEP. The Volume model was fitted using 4 components, due to around 80% of the variance in both the dependent and independent variables explained, and a low MSEP.

No. of Comps	1	2	3	4	5	6	7	8	9	10
% var X	37.99	60.23	73.15	80.65	84.62	87.77	90.53	92.2	93.58	94.71
% var Volume	20.67	75.29	75.5	80.45	80.49	80.75	82.27	82.64	82.72	83.07

Table 14: Percentage of variance explained in predictors and response variables for Volume PCR

No. of Comps	1	2	3	4	5	6	7	8	9	10
% var X	40.58	61.05	72.32	79.28	83.12	86.52	89.37	91.56	93.09	94.23
% var Pricing	31.11	47.65	48.29	50.93	54.74	54.85	75.7	76.35	76.7	78.12

Table 15: Percentage of variance explained in predictors and response variables for Pricing PCR

The results of the computed MSEs for the fitted models on the training data and the out-of-sample MSEs are shown in Table 16, and the plot of the fitted model and forecasts are displayed in Figures 22 and 23. The forecasts for both scenarios in general show the expected behaviour, though the curves (in particular those for the Volume PCR model) are smoother than desired. The Pricing model also predicts a sharp decline in price in the cstress scenario, followed by an eventual steep increase, and the forecasts in general for Pricing do not show the extreme nature we might expect for the stress scenarios. This may be due to noise from the original variable set influencing the model in an unanticipated way.

Model	**Dependent Variable**	**Comps**	**MSE (Fitted)**	**MSE (Out-of-Sample)**
v.pcr	Volume	4	1162.48	1700.84
p.pcr	Pricing	7	0.0626	0.1467

Table 16: Model results for PCR

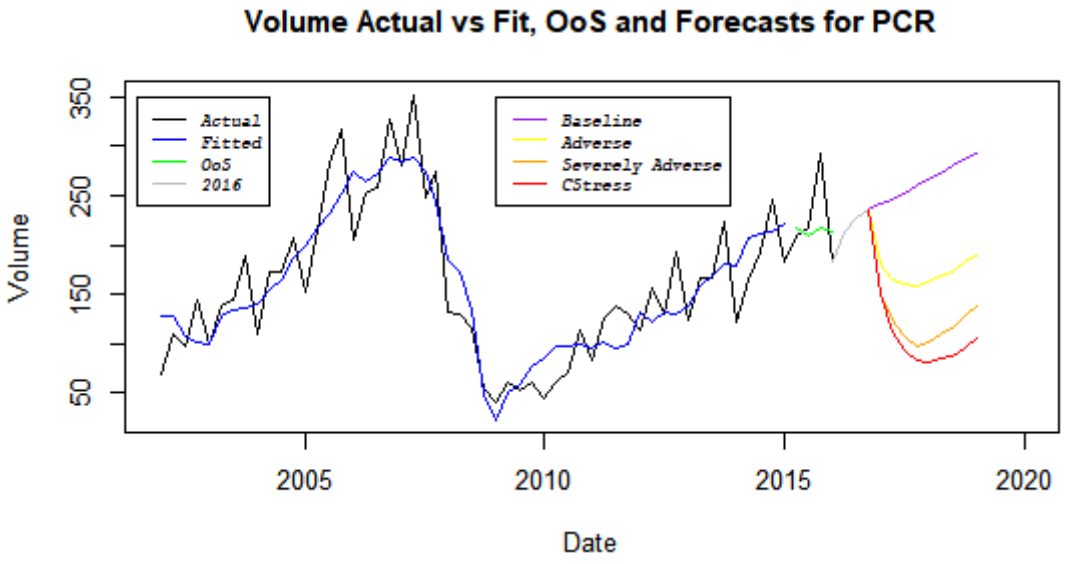

Figure 22: Plot of actual versus fitted and out-of-sample, and forecasts for Volume PCR model

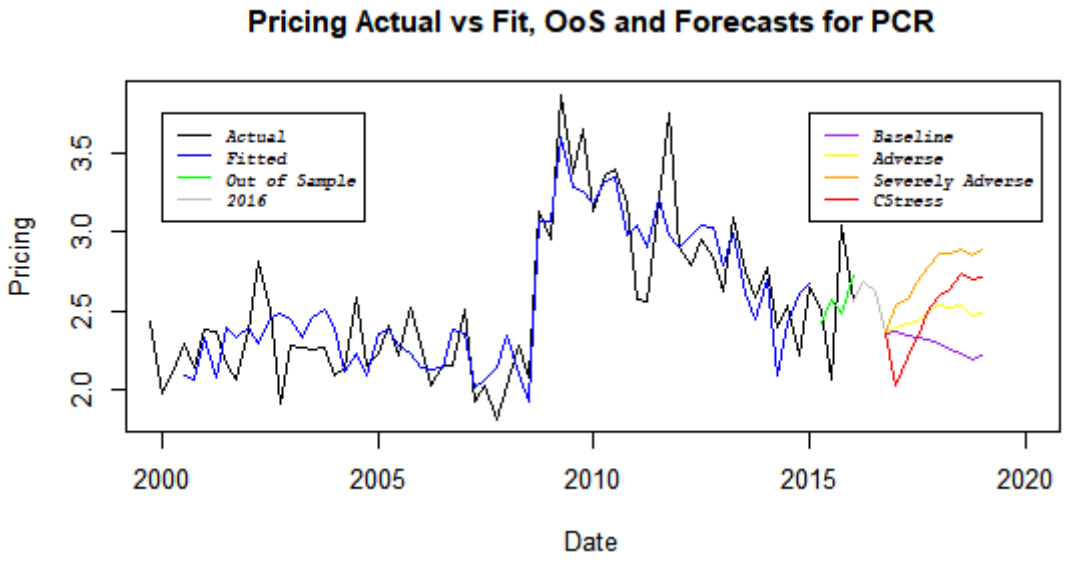

Figure 23: Plot of actual versus fitted and out-of-sample, and forecasts for Pricing PCR model

The plots for the residuals versus fitted values for the models are depicted in Figure 24. The residuals for the Volume model show potential unwanted behaviour of increasing with size of the fitted value, while the residuals from the Pricing variable display no obvious pattern. Analysing the Volume model's residuals further by looking at ACFs and PACFs indicates that it fails to model the seasonality of the series, shown by the serial autocorrelations around the 4 quarter (1 year) lags.

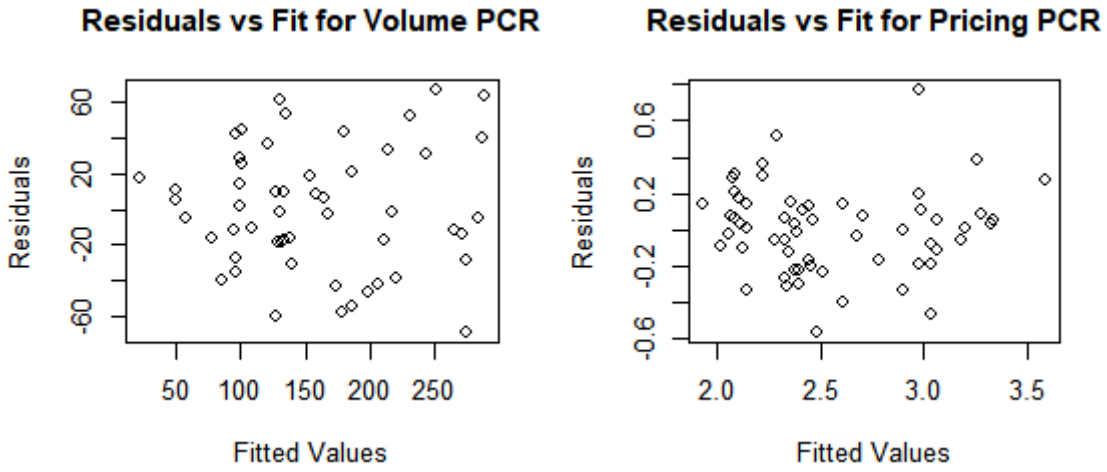

Figure 24: Plots of residuals versus fitted values for PCR Volume (Left) and Pricing (Right) models

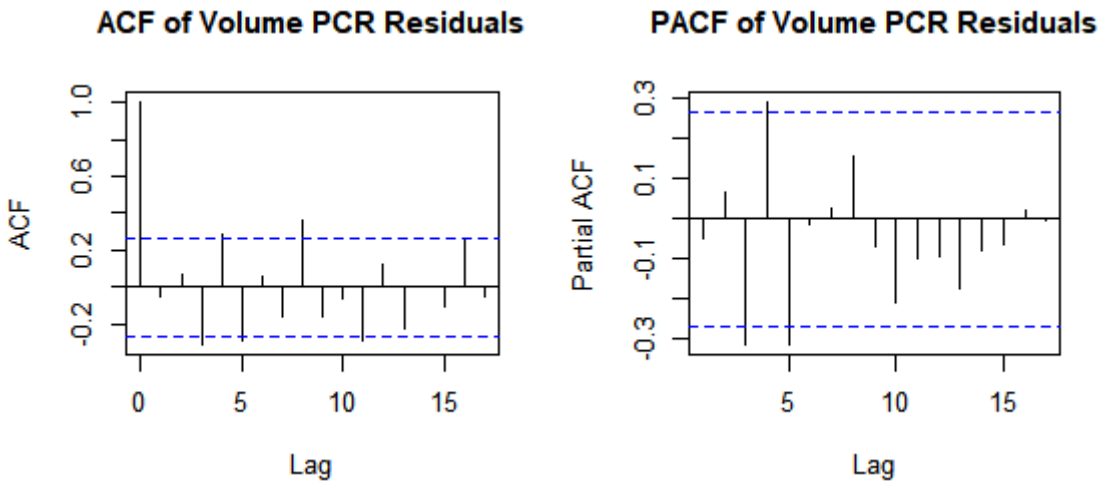

Figure 25: Plots of residuals ACF and PACF for the PCR Volume model

4.5 KNN Regression

After normalising the variables as specified in Section 3.5, models fit by KNN regression were built for:

- Original dependent variables using original independent variables,
- Original dependent variables using the independent variables most correlated with these, and
- Stationary dependent variables using the independent variables most correlated with these.
- Stationary dependent variables using the independent variables chosen by the variable selection techniques in Section 4.2.

Values for K were determined using random k-folds cross-validation, and models were constructed to compute MSEs for the fitted and out-of-sample values. The models for both Volume and Pricing that produced the best statistics was fitted using the original untransformed variables, and the results for these models are presented below in Table 17. The plots of the fitted values and forecasts for the models are displayed in Figures 26 and 27.

Model	**Dependent Variable**	**K**	**MSE (Fitted)**	**MSE (Out-of-Sample)**
v.knn	Volume	10	1,642.27	3,176.28
p.knn	Pricing	5	0.0683	0.1195

Table 17: Pricing model results for KNN

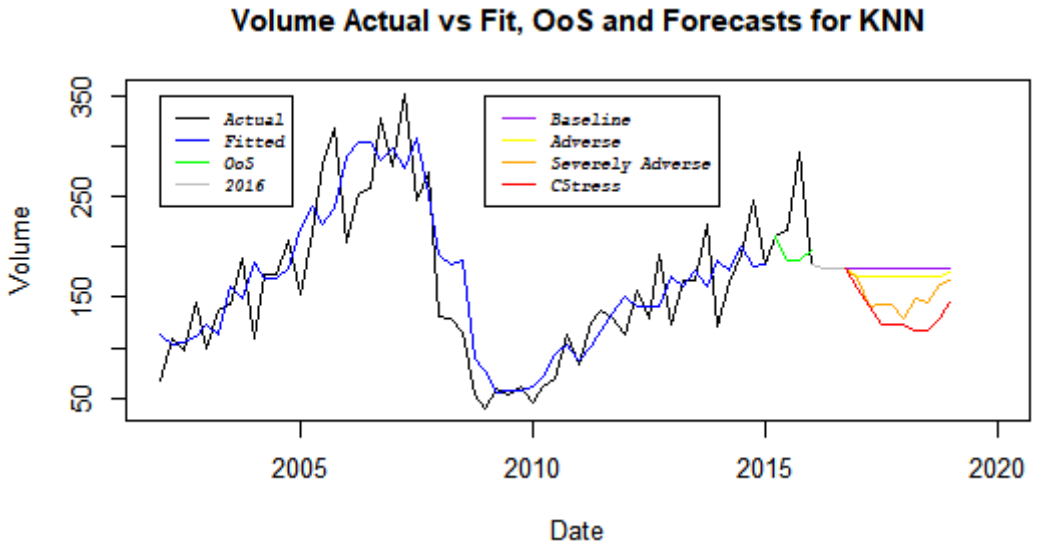

Figure 26: Plot of actual versus fitted and out-of-sample, and forecasts for Volume KNN model

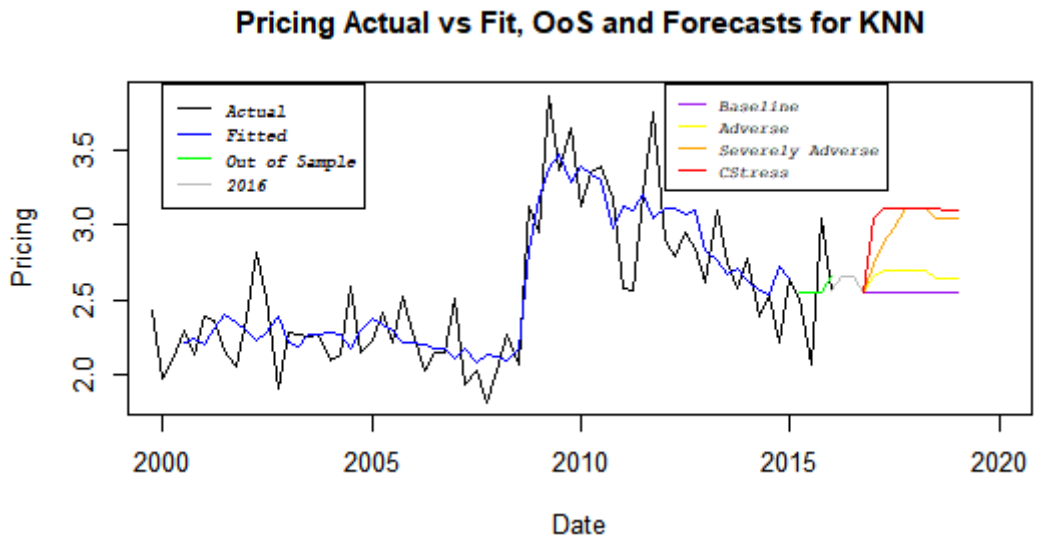

Figure 27: Plot of actual versus fitted and out-of-sample, and forecasts for Pricing KNN model

We can see from Figures 26 and 27 that the model forecasts troughs and peaks in line with scenario expectations, though the extent of the impact of the scenarios is less than that of the financial crisis in 2008. Due to the model taking an average of a set of 10 and 5 values respectively (the values for K in the models), the curves show little in the way of fluctuations and produce largely smooth lines. A potential improvement to the process would be to implement a weighted

voting system, with the inclusion in the algorithm of a weighting that is proportional to the inverse of the distances to the test point, ensuring that the closer neighbours contribute more to the final value. The residuals versus fitted values are plotted in Figure 28, and for Pricing demonstrate no pattern in the residuals as desired for a model. For the Volume model however, there appears to be a pattern of a increasing residuals with fitted value size. Closer inspection of the ACF and PACF for the residuals from the Volume model in Figure 29 again indicate (as was the case in PCR) that the model fails to produce the seasonal pattern of the Volume series.

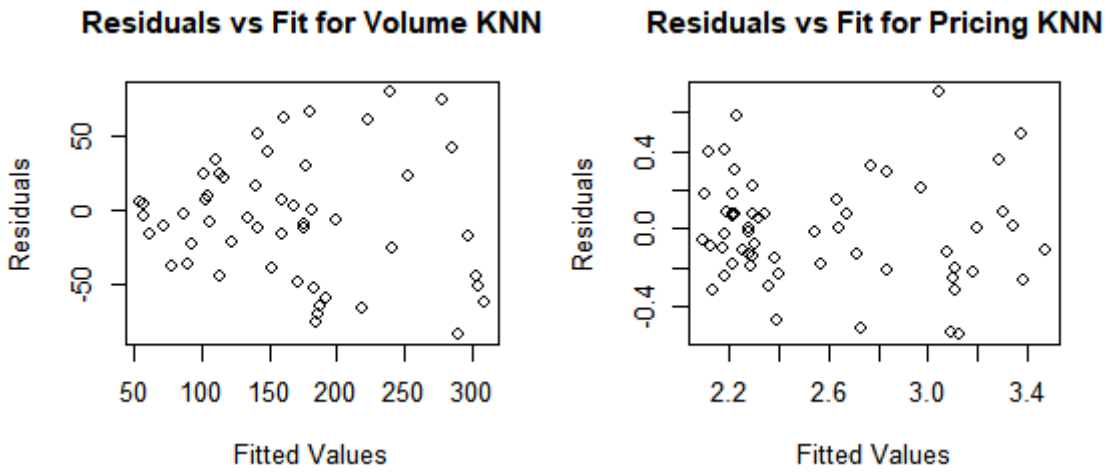

Figure 28: Plot of residuals versus fitted values for KNN Pricing model

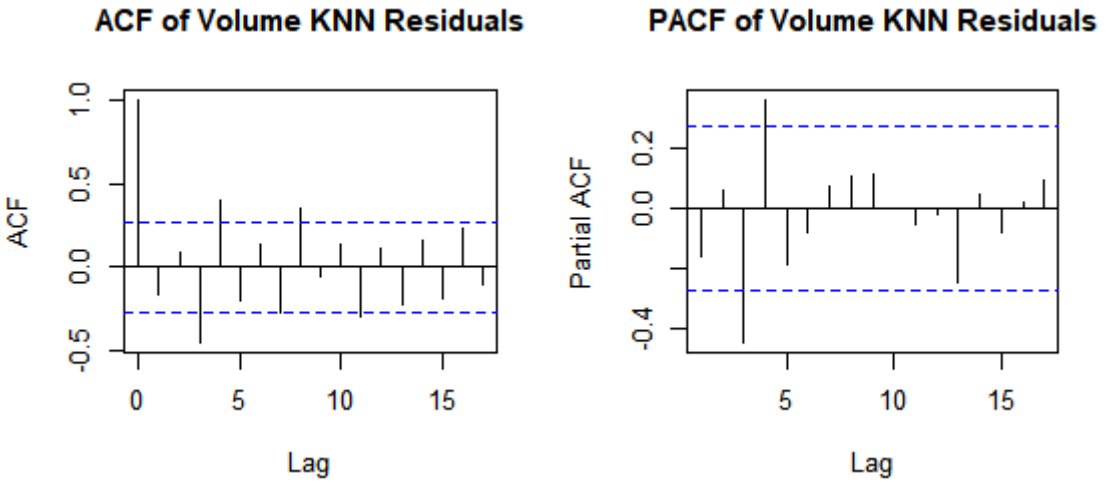

Figure 29: Plots of residuals ACF and PACF for the KNN Volume model

4.6 Discussion of Results

In the previous sections, for each modelling technique a model has been selected as the optimum for the implemented technique. A direct comparison of the models can be carried out by comparing the MSEs computed for the fitted values, and the out-of-sample values. Tables 6, 9, 16 and 17 indicate the multiple linear regression provided the closest fit to the data set for the Volume variable, while 6, 10, 16 and 17 suggest that the ARIMAX model was the closest fit for Pricing, though the PCR and KNN models provided better predictions for the out-of-sample data set, which are favourable to stress testing modelling.

When forecasting values for the scenarios provided, we would expect the Volume time series to fall as the adverse conditions play out, with a more pronounced descent, and potentially a sustained period of low values, the more severe the scenario. The Pricing time series would be expected to rise sharply in times of economic stress, again with a sharper rise the more troubling the scenario. These behaviours would mimic those that occurred during the 2008/2009 credit crisis. Forecasting for the Volume time series proved to be difficult, with some models not displaying a significant impact from the given scenario, raising questions about whether the model exhibits sufficient macroeconomic sensitivity. As mentioned earlier, the choice of variables selected in the models for multiple linear regression and ARIMAX errors may contribute to this, which would mean the models would need revisiting with a more business focussed outlook in the variable selection process. The Volume models constructed by PCR and KNN regression do however produce a forecast that follows anticipated behaviour to some extent, with a more severe and prolonged drop the more stressed the scenario becomes. This may be due to the techniques retaining the majority of the variability in all the variables, as opposed to only a select few. The forecasts for Pricing models largely generate the expected behaviour, with the Pricing time series experiencing greater spikes as the severity of the scenario increases. The impact of the scenarios was somewhat muted for the PCR and KNN models, potentially suppressed due to noisy, unhelpful variables used in the construction of the components in PCR, and in the computations of neighbour values in KNN.

Both PCR and KNN struggled to model the seasonality in the Volume time series, with multiple linear regression and ARIMAX both having clear methods of including the seasonality aspect in their techniques. A further area of interest could be to investigate if a quarterly indicator variable can be included as a regressor in the PCR models, to help model the seasonality effect.

5 Conclusion

While the Federal Reserve Bank are compelling banks to move their PPNR systems forward into more sophisticated modelling techniques than the classical linear regression, a definitive superior technique is yet to be widely employed. In this paper, the potential benefits and possible downsides to a set of modelling techniques were discussed and analysed, with a view for comparison to the multiple linear regression model. These techniques were: an extension to the linear regression with the inclusion of ARIMA errors; a supervised learning application of principal components in a regression model; and finally a utilisation of the K-nearest neighbours algorithm in an instance based learning regression. Models were fitted using the empirical data provided for simulation of the stress testing scenarios devised by the FRB, and tests of the performance of the models, and the validity of their assumptions, were conducted.

Techniques that rely on assumptions on the underlying data were shown to fall foul of these assumptions in some cases, such as the supposed normality of the residuals in a multiple linear regression. Extending the commonplace regression framework with the residuals modelled using an ARIMA process offers a potentially simple upgrade to the current models. The dimension reduction technique applied in principal components regression could also reap benefits in combination with other methods, to achieve fewer coefficient estimates. As a further improvement to PCR for the purposes of predicting the outcome, in [35] Jolliffe notes that a more careful analysis of the principal components with low variances may also be important in the prediction forecasts. Machine learning techniques may yet prove to be of use in these scenarios as well, but improved data storage may be required for these to become effectual. Any technique that is applied will need further scrutiny with regards to business application, with the potential for both PCR and KNN to be significantly improved by utilising a business analysis of the variables prior to model fitting, to remove any variables deemed uninformative for predictions of values in times of economic stress.

References

[1] Federal Reserve Bank (2017). "Stress tests capital planning". White paper report. https://www.federalreserve.gov/

[2] PWC (2014). "Passing the stress test PwC survey on regulatory stress testing in banks". White paper report. https://www.pwc.com/gx/en/financial-services/publications/

[3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel (2008). "Time Series Analysis: Forecasting and Control". John Wiley and Sons Inc, New York.

[4] K. Pearson (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine. 2 (11): 559572.

[5] A. J. McNeil, R. Frey, P. Embrechts (2015). "Quantitative Risk Management: Concepts, Techniques and Tools", Revised Edition. Princeton University Press, Princeton.

[6] H. Akaike (1974), "A new look at the statistical model identification". IEEE Transactions on Automatic Control, 19 (6): 716723.

[7] G. Claeskens, N. L. Hjort (2008). "Model Selection and Model Averaging". Cambridge University Press.

[8] C. Giraud (2015). "Introduction to High-Dimensional Statistics". CRC Press.

[9] J. E. Cavanaugh (1997). "Unifying the derivations of the Akaike and corrected Akaike information criteria". Statistics & Probability Letters, 31: 201208.

[10] K. P. Burnham, D. R. Anderson (2004), "Multimodel inference: understanding AIC and BIC in Model Selection". Sociological Methods and Research, 33: 261304.

[11] G. E. Schwarz (1978). "Estimating the dimension of a model". Annals of Statistics, 6 (2): 461464.

[12] K. N. Berk (1978). "Comparing subset regression procedures". Technometrics, 20, 1-6.

[13] E. M. L. Beale (1970). "Note on procedures for variable selection in multiple regression". Technometrics, 12, 909-914.

[14] F. E. Harrell (2001). "Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis". Springer-Verlag, New York.

[15] R. Tibshirani (1996). Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological) 58 (1). Wiley: 26788.

[16] H. Zou, T. Hastie (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (statistical Methodology) 67 (2). Wiley: 30120.

[17] M. Yuan, Y. Lin (2006). Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society. Series B (statistical Methodology) 68 (1). Wiley: 4967.

[18] A. Pankratz (1991). "Forecasting with Dynamic Regression Models". Wiley-Interscience.

[19] H.J. Bierens (1987). "ARMAX Model Specification Testing, With an Application to Unemployment to The Netherlands". Journal of Econometrics, 35 (February), 161-90.

[20] G. Athanasopoulos, R. Hyndman (2013). "Forecasting: Principles and Practice". OTexts.

[21] R. Hyndman. "The ARIMAX model muddle". Blog. https://robjhyndman.com/hyndsight/arimax/

[22] D. A. Dickey, W. A. Fuller (1979). "Distributions of the Estimators for Autoregressive Time Series with a Unit Root". Journal of the American Statistical Association, 74(366), 427-431.

[23] D. A. Dickey, W. A. Fuller (1981). "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root". Econometrica, 49(4), 1057-1072.

[24] S. E. Said, D. A. Dickey (1984). "Test for Unit Roots in Autoregressive-Moving Average Models of Unknown Order." Biometrika, 71(3), 599-607.

[25] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, Y. Shin (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root". Journal of Econometrics. 54 (13): 159178.

[26] G. James, D. Witten, T. Hastie, R. Tibshirani (2013). "An Introduction to Statistical Learning". Springer.

[27] J. Shlens (2003). "A Tutorial on Principal Component Analysis". Princeton. 7-11.

[28] D. G. Stork, P. E. Hart, R. O. Duda (1973). "Pattern Classification". John Wiley and Sons, New York.

[29] A. Banerjee, J. J. Dolado, J. W. Galbraith, and D. F. Hendry (1993). "Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data". Oxford University Press, Oxford.

[30] D. A. Belsley, E. Kuh, R. E. Welsch (1980). "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity". John Wiley and Sons Inc, New York. 100104.

[31] M. H. Kutner, C. J. Nachtsheim, J. Neter (2004). "Applied Linear Regression Models" (4th ed.). McGraw-Hill Irwin.

[32] T. S. Breusch, A. R. Pagan (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". Econometrica. 47 (5): 12871294.

[33] G. M. Ljung, G. E. P. Box (1978). "On a Measure of a Lack of Fit in Time Series Models". Biometrika. 65 (2): 297303.

[34] B.-H. Mevik, H. R. Cederkvist (2004). "Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR)". Journal of Chemometrics 18 422429.

[35] I. T. Jolliffe (1982). "A note on the Use of Principal Components in Regression". Journal of the Royal Statistical Society, Series C. 31 (3): 300303.