Case Study Report 2 – Exercise 3

Attribution (5 marks)

The data for this report is sourced from the Australian Bureau of Statistics (ABS) for Building Approvals, Australia – 8731.0, Table 30, Reference Period: January 2024. The data is collected monthly, using both the original (Series ID: A422168F) and seasonally adjusted (Series ID: A422168F) series from February 2014 to January 2024, comprising 108 observations each.

Scope (5 marks)

This report aims to evaluate the accuracy of a multiple linear regression model with time as an intercept and the months February to December as dummy variables. The objective is to predict building approvals for the period from February 2023 to January 2024.

The analysis is conducted using Excel's Regression feature, with model evaluation based on R^2 (ranging from 0 to 1 for explanatory power) and Mean Squared Error (MSE), which quantifies the closeness of predictions to actual data points. The statistical significance of each variable is assessed using P-Value, considering values below 0.05 as significant to the model. Auto Correlation Function (ACF) and Residual Error Plot will be used to evaluate model adequacy.

Application (5 marks)

In this regression, we estimate a seasonal time series model where the target variable, total value of building jobs ($ '000) is explained by the quasi-explanatory variables, Time (t) and 11 monthly dummy variables, February to December (a_1 to a_{11}) respectively. The number of dummy variables created is $n - 1$. January is set as baseline reference to avoid multicollinearity. The model is simplified to equation below.

$$Y_t = \beta_0 + \beta_1 * t + (a_1 * D_1) + (a_2 * D_2) + (a_3 * D_3) + \cdots + (a_{11} * D_{11}) + \varepsilon t$$

Where:

Y_t: The target dependent variable

β_0: the intercept value

β_1: coefficient of indepedent time variable

a_i: coefficient of independent monthly dummy variables

D_i: independent monthly dummy variables

ε_t: random error that accounts for unexplained variation in Y_t

Dummy variables are binary, where $D_i = 1$ if observation is month i, otherwise $D_i = 0$.

E.g. For month December: $D_1 \ to \ D_{10} = 0 \ and \ D_{11} = 1$

The model can be simplified to:

$$Y_t = \beta_0 + \beta_1 * t + (a_{11} * D_{11}) + \varepsilon t$$

Analysis (10 marks)

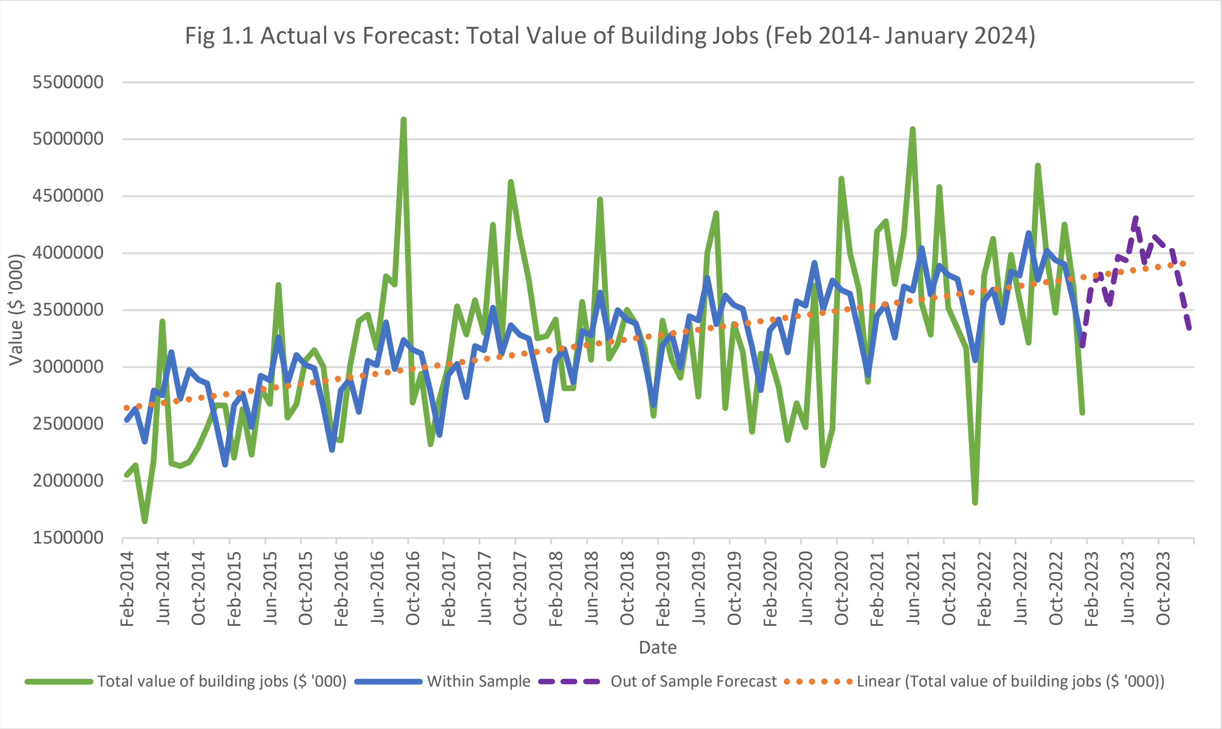


Fig 1.1 Actual vs Forecast: Total Value of Building Jobs (Feb 2014- January 2024)

The model captures the increasing trend but fails to adequately model seasonal fluctuations and trend reversals around 2020. Building approvals generally spike in June and decline in December, attributed to industry norms and holiday season influences. A large underprediction in 2020 is linked to the COVID-19 pandemic and subsequent economic measures, including reduced interest rates to 0.10% to boost the construction industry. The sharp decline at the beginning of 2022 correlates with the Reserve Bank of Australia's interest rate hikes.

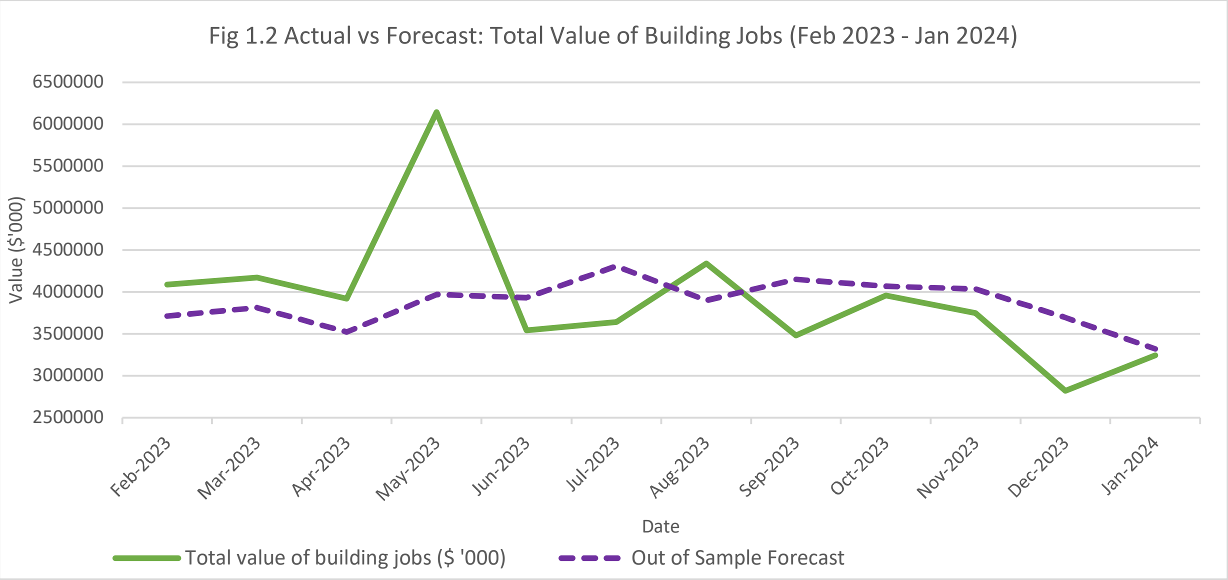
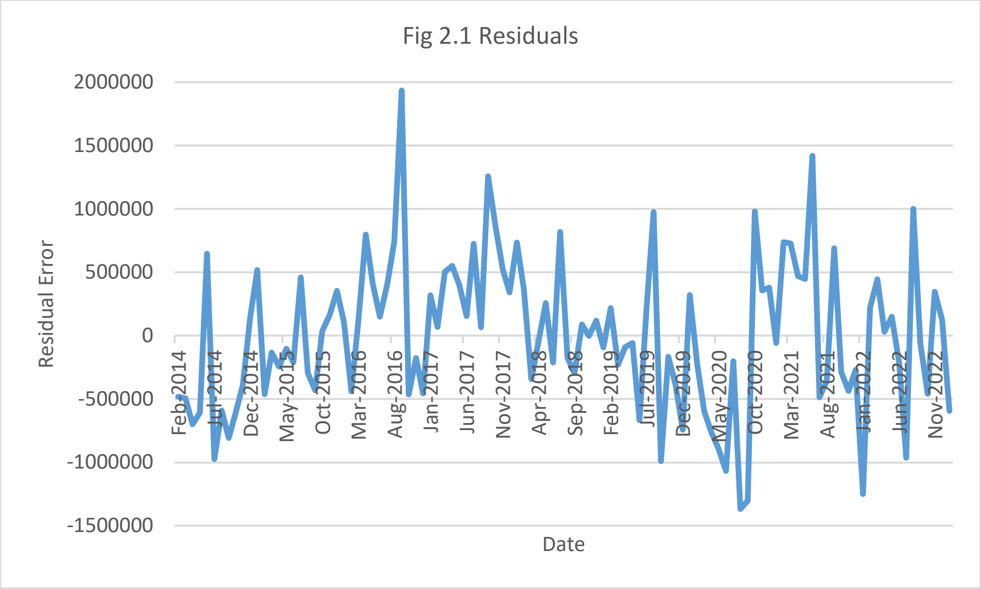


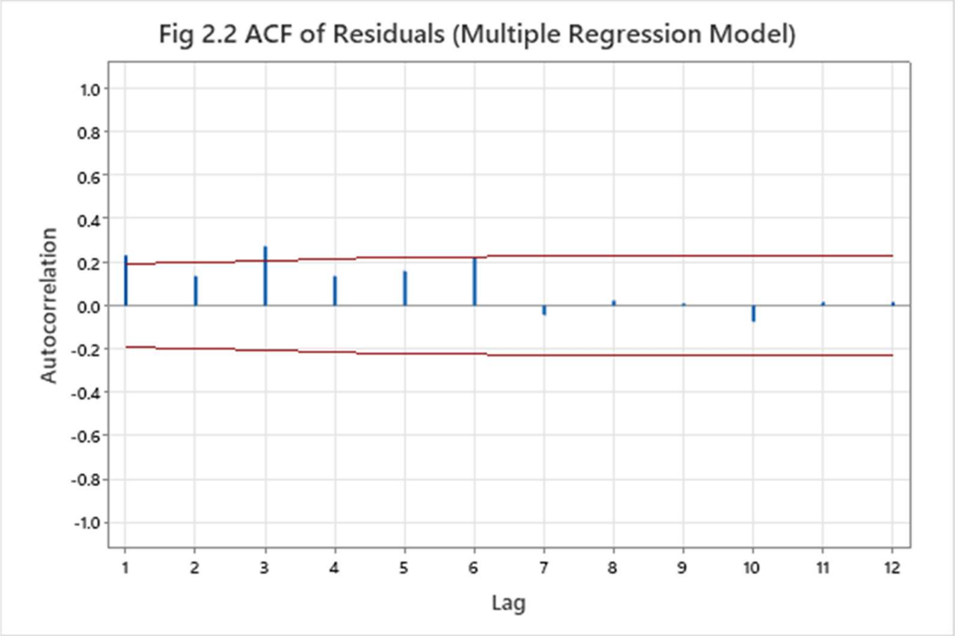
Fig 1.2 Actual vs Forecast: Total Value of Building Jobs (Feb 2023 - Jan 2024)

The out-of-sample forecast performance indicates an inability to generalize data patterns adequately, likely due to training on unstable economic periods and volatile interest rates. Additionally, supply shortages of building materials due to Eastern European conflicts exacerbate this issue.

Articulation of Issues (10 marks)



Referring to Fig 2.2, the residuals are not homoskedastic as variance does not consistently increase over time. Variance is largest at the extremes around 2016 and 2020, and smallest from 2018 to 2019, suggesting model misspecification.



The ACF test (Fig 2.2) shows several significant lag values at (1, 3, 6) that exceeds the confidence interval, indicating non-random components. The low R^2(0.35) and adjusted R^2 (0.25) supports the idea of model misspecification.

Critique (15 marks)

The overall model is significant (Sig F stat: 3.06E-05 < 0.05) in predicting building value. However, the low R^2 and adjusted R^2 value shows that 65% of variation has not been captured by the model. This suggests potential overfitting and unidentified significant predictors. Insignificant predictors such as February, April, and December (P-values > 0.05) contribute minimally to the model's explanatory power.

Fig 3.1

R^2	0.35
Adjusted R^2	0.26

Predictors	*P-value*
Intercept	4.72675E-13
Time	2.31026E-07
Feb	0.09
Mar	0.05
Apr	0.32
May	0.01
Jun	0.02
Jul	0.00
Aug	0.04
Sep	0.00
Oct	0.01
Nov	0.01
Dec	0.20

Improvement could be achieved by removing these predictors and including additional factors such as interest rates or raw material costs. Segmenting data with its own regression model could also better capture trends.

Fig 3.2 Error Comparisons	MAE	MSE	RMSE	MAPE
Multiple Regression	470293.4	3.96E+11	629063.4	15.49118
WES Multiplicative (Optimised)	598427.4	5.78E+11	760577.3	18.84453

Multiple Regression performs better than WES Multiplicative across all error metrics. Indicating that Multiple Regression is generally a better forecasting model. From a business forecasting perspective, both models have merits. WES Multiplicative captures trends and seasonality effectively but struggles with external factors such as the COVID-19 crisis. In contrast, Multiple Regression can model external factors as predictors, improving performance but is limited to additive seasonality and linear trends.

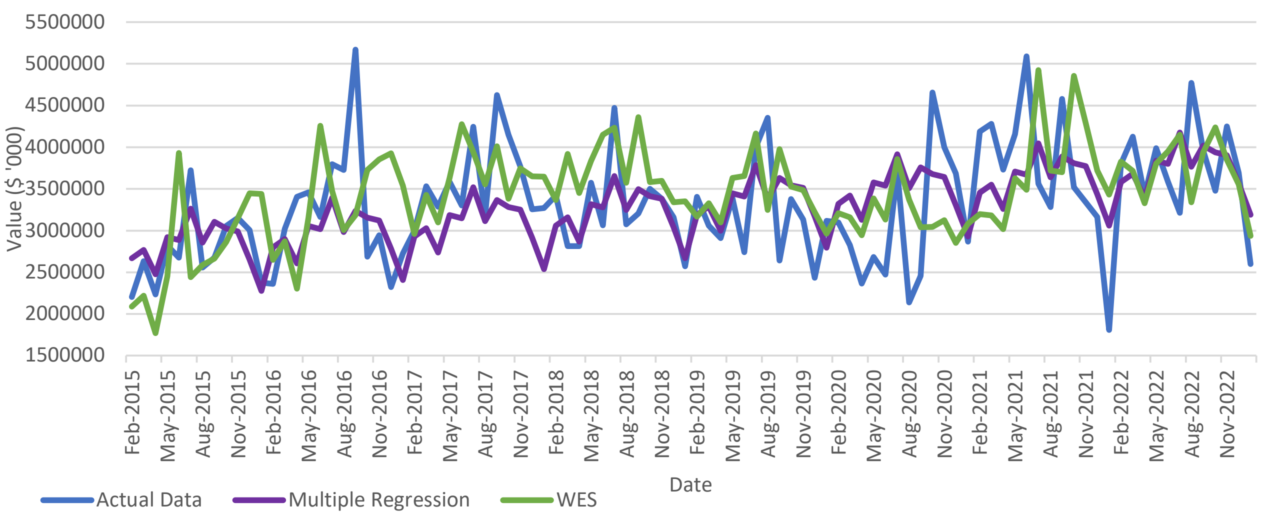


Fig 3.3 Comparison of WES Multiplicative, Multiple Regression and Actual Data

Position (10 marks)

The Multiple Regression model has a low R^2 value and high residual volatility, indicating room for improvement in accuracy and adequacy. Removing insignificant variables and identifying new external factors could enhance forecasting accuracy. A hybrid model, using WES Multiplicative for seasonal trends and Multiple Regression on residuals to capture external factors, may provide unbiased and robust forecasts. This combined approach leverages the strengths of both models, ensuring more reliable and actionable predictions for building approvals in New South Wales, Australia.