Application of Phillip's Model to Predict United States' Wage & Consumer Price Index using Two Step Least Square (2 SLS) Regression

Christopher Timothy Kwee (23101910029), Dennis Jonathan (23101910027), and Justin Jedidiah Sunarko (23101910024)

christopher.kwee@student.pmsbe.ac.id, dennis.jonathan@student.pmsbe.ac.id, justin.sunarko@student.pmsbe.ac.id

Business Mathematics 2019, School of Applied Science, Technology, Engineering and Mathematics, Universitas Prasetiya Mulya, BSD City Kavling Edutown I.1, Jl. BSD Raya Utama, BSD City, Kec. Pagedangan, Tangerang, Banten 15339, Indonesia

Abstract. Phillips' Curve and its implementation in the form of the wage and price model is one of the most important tools in macroeconomics. This research focused on the Phillips wage and price model found in Gujarati's *Basic Econometrics* book using 2 SLS regression, particularly examining whether this model can possibly be useful to determine wage and price with the given data. From the results, the model seems to perform well on data it has trained on but it has a very bad performance on new data. The model showed that wage and price were significantly correlated.

1. Introduction

The concept of one's wage and his or her ability to buy goods and services has been highly studied in the field of macroeconomics. Banton (2021) said that wage and price have a cause and effect relationship with one another. She added the increase in wage leads to an increase in disposable income, which in turn will also create an increase in price because when people have more money to spend, there is a tendency to demand for more goods and services, a concept which is commonly referred to as the wage and price spiral [1]. The ability to predict this relationship is especially important for policymakers when determining the trade off of macroeconomics instruments. In his book Principles of Macroeconomics, Mankiw (2018) explained that in Phillips' research of macroeconomics metrics in the United Kingdom, Phillips found that there is indeed a negative correlation between inflation and

unemployment rate, which was also backed up by a similar research done by Samuelson and Solow in the United States [2]. This cause and effect relationship is what in turn creates a circle which puts pressure on the economics and causes inflation and deflation when it is not controlled properly as described by Gordon (2021) [3]. Therefore, there is a need to be able to find out how macroeconomics metrics might affect wage and price which in turn will also affect the lives of many people.

In this study, the authors will delve more into how one can analyse the cause and effect of this phenomena using regression in trying to model wage and price, especially in the United States of America. The scope of the research will be limited from the year 2006 up to 2020, with data collected at monthly intervals. The model the authors used is based on the Phillip's Wage and Price model created by the economist William Phillips, specifically a simultaneous model taken from Gujarati's Basic Econometrics book with the objective of finding out whether this model can possibly be useful to determine wage and price with the given data. The research will be conducted using the programming language Python with the help of several libraries in order to model the data.

2. Methodology

This research is done to find out about the relationship of wage and price along with several other variables (unemployment, cost of capital, raw materials import). All of the data for the variables are acquired from the U.S. Bureau of Labor Statistics except for the cost of capital variable that is acquired from NYU Stern.

2.1 Simultaneous Equation Model

Simultaneous equations are a set of equations with a minimum of two equations where the response variables (y) are also explanatory variables (x) in the other equations. Therefore there is a back and forth causality between the x and y variables [4].

2.2 Two-Stage Least Squares (2 SLS)

Ordinary least squares model requires its errors to be uncorrelated with the independent variables (x) [5]. The problem with simultaneous equations is that the errors will be correlated with the endogenous variables. Therefore, two-stage least squares uses instrumental variables to create a proxy for the endogenous variables. The proxy is highly correlated with the endogenous variable but is not correlated with the error terms . To get the proxy, the endogenous variables will be regressed with all exogenous variables of the entire model as the exogenous variables. The prediction values from the regression will therefore be used as a proxy for the endogenous variables . That is the first stage, the second is to simply make a regression but replacing the endogenous variables with the proxies [6].

2.3 Phillips Curve and the Wage-Type Model

According to Mankiw (2018), the Phillips Curve is a curve which shows the relationship between inflation and unemployment based on William Phillips' article published in a British Journal called *Economica* [2]. The relationship between those two variables is indeed negative and valid only for the short-run. Kevin D. Hoover from the website econlib.org explained that Phillips found a consistent inverse relationship: when unemployment was high, wages increased slowly; when unemployment was low, wages rose rapidly [7].

The model used for this analysis was inspired from Phillips' founding and this system of equations is available on Gujarati's Basic Econometrics in chapter 18 [6], example 18.3. The system of equations is shown below:

$$\dot{\mathrm{W}}_t = \alpha_0 + \alpha_1 \mathrm{UN}_t + \alpha_2 \dot{\mathrm{P}}_t + u_{1t} \tag{1}$$

$$\dot{\mathrm{P}}_t = \beta_0 + \beta_1 \dot{\mathrm{W}}_t + \beta_2 \dot{\mathrm{R}}_t + \beta_3 \dot{\mathrm{M}}_t + u_{2t} \tag{2}$$

where the variables are defined as:

- $\dot{\mathrm{W}}_t$ is the money wages at time *t*
- UN_t is the unemployment rate percentage at time *t*
- $\dot{\mathrm{P}}_t$ is the consumer price index at time *t*
- $\dot{\mathrm{R}}_t$ is the rate of change of cost of capital at time *t*
- $\dot{\mathrm{M}}_t$ is prices of import raw material at time *t*
- u_{1t}, u_{2t} is the stochastic disturbance at time *t*

2.4 Identification problem (rank condition, order condition)

Each of the equations of a simultaneous equations model can be underidentified, identified, or overidentified. If an equation's parameters can not be obtained from its reduced form coefficients then the equation is said to be underidentified. On the other hand, if the parameters can be obtained from its reduced form coefficients then the equation is either identified or overidentified . If a unique value for the parameters can be obtained then the equation is identified but if there can be more than one value then the equation is overidentified.

For a model to be identified, there is a condition called order condition:

$$K - k \geq m - 1, \tag{3}$$

where:

- $K$is the number of exogenous variables in the model,
- $k$is the number of exogenous variables in the equation,
- m is the number of endogenous variables in the equation.

If $K - k = m - 1$ then the equation is said to be identified. Else, if $K - k > m - 1$then the equation is said to be overidentified. While the order condition is necessary, it is not enough. The other condition is the rank condition. The rank condition requires for all equations in a simultaneous equations model to have at least one nonzero determinant of order $(M - 1)(M - 1)$that is made from the coefficients of both endogenous and exogenous variables that are included in other equations but not is the particular equation where$M$is the number of endogenous variables in the model [6].

2.5. Root Mean Squared Error

According to Simon P. Neill, M Reza Hashemi (2018), on Fundamentals of Ocean Renewable Energy, The root mean squared error (RMSE) is the square root value obtained from the mean squared of all errors. RMSE serves as a general purpose error metric that is an excellent value for numerical prediction. The results of the RMSE measure are very good, but can only be used to compare the predictive error values of different models or model configurations for certain variables, but not between variables, because later it depends on the scale.[8].

$$RMSE = \sqrt{\frac{1}{n}\sum(Y_i - \hat{Y}_i)^2} \tag{4}$$

where:

- n is the number of observations
- Y_i is the actual value of Y at observation i
- $\hat{Y}_i$is the predicted value of Yat observation i

2.6. Hausmann's test for endogeneity

According to Glen (2007), The Hausman Test is also called the "Hausman specification test" that detects endogenous regressors or predictor variables in a regression model. Endogeneous variables have values that are determined by other variables in the system. The ordinary least squares estimators will fail if there are endogenous regressors in a model as one of the assumptions of OLS is that there is no correlation between the predictor variable and the error term. There are also alternatives to these cases, called the instrumental variables estimators. However, make sure that the predictor variables are endogenous. This is what the Hausman test can do [9].

A version of the Hausman specification error test that can be used for testing the simultaneity problem. The hypothesis for the Hausmann test can be written as follows:

$$H_0: \beta_{res} = 0$$

$$H_1: \beta_{res} \neq 0$$

Decision Rule:

H_0 is rejected if the p-value < α, which is equal to 0.05.

2.7. Cross Validation

As reported by Jason Brownlee (2018), cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. It is a popular method as the method is simple and easy to understand. It generally results in a less biased estimate of the model, such as a simple train test split [10].

2.8. Multicollinearity

According to Jim Frost (2017), multicollinearity will occur when the independent variables in a regression model are correlated. This can be said to be a problem because independent variables should be independent (uncorrelated). If the degree of correlation between the variables is high enough, it can cause problems when the model is fitted and the results are interpreted [11].

Multicollinearity problems will make the coefficient estimates swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model. It can also reduce the precision of the estimated coefficients, which weakens the statistical value and power of the regression model. The rule of thumb is as follow:

VIF	Interpretation
< 1	no multicollinearity
1-5	independent variable is moderately correlated
> 5	strong multicollinearity

Table 1. VIF Values Interpretation

2.9. Normality Q-Q plot and KDE D'Agostino's K-squared test

According to Sundaresh Chandran, Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting two different quantiles against each other. For example, if the variable tested is the distribution of wage of employees is normally distributed, the quantiles of the wage of employees and the quantile from a normally distributed curve are being compared. If the two quantiles are sampled from the same distribution, both of them should roughly fall in a straight line [12].

There is also another method to visualize the normality of the dataset, called KDE plot (Kernel Density Plot). KDE plot is a method for visualizing the distribution of observations in a dataset, analogous to a histogram. KDE can represent the data using a continuous probability density curve in one or more dimensions [13].

Based on Real Statistics, D'Agostino-Pearson Test can be used to determine whether a dataset is normally distributed using skewness and kurtosis analysis. The D'Agostino-Pearson Omnibus Test is based on the fact that when the data is normally distributed, the test statistic:

$$z_s^2 + z_k^2 \sim \chi^2(2) \tag{5}$$

The test should generally not be used for datasets with less than 20 observations [14]. The hypothesis test for normality using D'Agostino's K-Squared Test:

H_0: Normally distributed

H_1: Not normally distributed

Decision Rule:

H_0 is rejected if the p-value (D'Agostino's K Squared test) $< \alpha$, which is equal to 0.05.

2.10. Heteroscedasticity

Heteroscedasticity will happen when the standard deviations of the predicted variable, monitored over different values of an independent variable or as related to prior time periods, are not constant. The assumption is that we can make a good model if the error of the variance is constant over time [15].

One of the tools that can be used to test heteroskedasticity is using the Breusch-Pagan Test. The Breusch-Pagan Test statistic is a test for heteroskedasticity in the residuals from a statistical regression analysis. The hypothesis test for normality using Breusch-Pagan Test:

H_0:Homoscedastic

H_1: Heteroscedastic

Decision Rule:

H_0 is rejected if the p-value (Breusch-Pagan's K Squared test) $< \alpha$, which is equal to 0.05 [16].

3. Result and Discussion

3.1. Reduced Form

$$\dot{W}_t = \frac{\alpha_0+\alpha_2\beta_0}{1-\alpha_2\beta_1} + \frac{\alpha_1}{1-\alpha_2\beta_1}UN_t + \frac{\alpha_2\beta_2}{1-\alpha_2\beta_1}\dot{R}_t + \frac{\alpha_2\beta_3}{1-\alpha_2\beta_1}\dot{M}_t + u^* \tag{5}$$

$$\dot{P}_t = \frac{\beta_0+\beta_1\alpha_0}{1-\alpha_2\beta_1} + \frac{\alpha_1\beta_1}{1-\alpha_2\beta_1}UN_t + \frac{\beta 2}{1-\alpha_2\beta_1}\dot{R}_t + \frac{\beta_3}{1-\alpha_2\beta_1}\dot{M}_t + u^* \tag{6}$$

Equation (5) and **Equation (6)** above are the reduced form of the model. u^* is the linear combination of the errors divided by $1 - \alpha_2\beta_1$. One can observe that the coefficient for each variable is composed of a linear combination from the substituted equation.

3.2. Hausman's Test for Endogeneity

There are two ways of conducting Hausman's Test for Endogeneity, the first is doing it manually. The manual test starts by regressing an equation on every exogenous variable , for this step the authors chose the Wage equation and regressed it on 'unemployment', 'raw_mat', and 'cost_capital'. Regressing the Wage equation will give the predicted value for the variable 'wage' and also its residual. The next step involves regressing the other equation with this predicted value and its residual and checking whether the parameter for the residual is statistically significant. The result was that the parameter for the residuals of the Wage equation were indeed significant, signalling the variable 'Wage' was indeed endogenous.

Another way of conducting Hausman's Test for Endogeneity is using the method '.wu_hausman()'. This method uses the hypothesis in which all variables are exogenous, thus it can be rejected when p-value is less than α. The result from this method indeed confirmed that both variables ('CPI' and 'wage') were endogenous.

3.3. Model and Identification Problem

The first step in doing analysis of Simultaneous Equation Models is finding out whether it is possible to predict the value of the parameters for the model's equations or determining the identification for the particular equation. First is to check the Order condition.

Equation	K	k	K-k	m	m-1	Description
Wage	3	1	2	2	1	Overidentified
CPI	3	2	1	2	1	Exactly Identified

Table 2. Order Condition for Both Equations

For the Wage equation, it can be seen that $K - k$ was more than $m - 1$. This means that the Wage equation was overidentified. On the other hand, the CPI equation was exactly identified since $K - k$ was the same as $m - 1$. This indicates that one could solve and estimate the values of the parameters using a technique called two-stage least square.

Equation	W	UN	P	R	M	Rank≥ $M-1$	Description
Wage	x	x	x	0	0	$1 \geq 1$	Identified
CPI	x	0	x	x	x	$1 \geq 1$	Identified

Table 3. Rank Condition for Both Equations

The second step is to check the Rank condition. For each equation, a matrix is made by taking the coefficients from other equations whose coefficient from the original equation is zero. Since the values of the coefficients are unknown, 'x' is used to represent them. The next step is to check the rank of those matrices. If the rank is more than or equal to one ($M-1$), then the equation is either exactly identified or overidentified. Else, the equation is underidentified. Since the rank of all the matrices made for this model has a rank of one, then the equations of this model were all identified.

3.4. Modelling with Two Stage Least Squares

Before modelling, the data was split into two parts: training data and testing data. This step was done so the model could be evaluated not just on the data that it has trained on but also data that it has never seen before. The authors chose the training data to be 70% of the data and the testing data to be 30%. Since the dataset was observed by month, the division was not shuffled. The modelling process was done using the library linearmodels.IV using the function IV2SLS.

The result of the modelling process for both equations can be seen on the table below:

Equation	Exogenous Variables	Estimated Parameters	Adj. R-squared
Wage	Constant	$1.137 * 10^5$	0.9647
	Unemployment Rate (%)	-1881.5	
	CPI	158.55	
CPI	Constant	297.37	0.7565
	Money Wages	$-8.499 * 10^{-6}$	

Rate of change of cost of capital	−9.5966
Prices of Imported Raw Material	−0.7316

Table 4. Estimated Parameters for the Models with 2 SLS Method

The Adjusted R-squared for the Wage model is 96.47%, this value is really good. On the other hand, the CPI model has 75.65% Adjusted R-squared which is quite good though not optimal.

3.5. Residual Analysis

Equation	Train/Test	MSE	RMSE
Wage	Train	507,225	712
	Test	8,523,641	2,920
CPI	Train	35	6
	Test	2881	54

Table 5. Residual Metrics for Each Model

From the table above, it looks like the difference of the MSE and RMSE values for the wage model is really bad since the differences are huge. The same goes to the CPI model. This indicates that the models are not very helpful even though they have high Adjusted R-squared values.

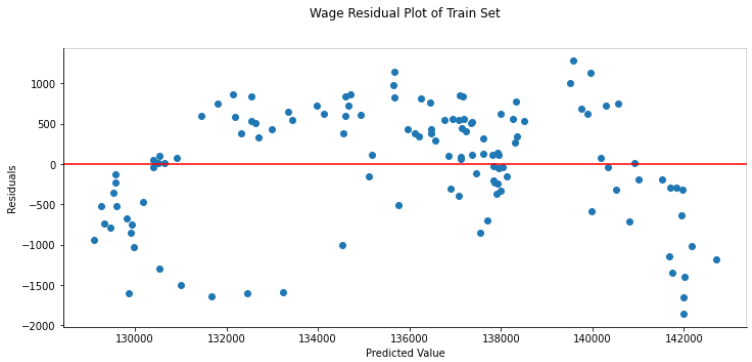

Figure 3.5.1. Wage Residual Plot of Train Test

The plot above looks pretty good except for the left side which looks quite strange. To be certain of the residuals, the result of statistical testing for normality of the residuals are shown in 3.5.2.

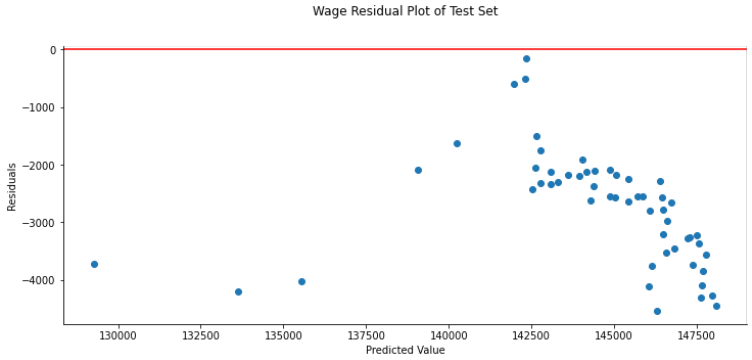

Figure 3.5.2. Wage Residual Plot of Test

From **Figure 3.5.2**, it can be seen that the residuals for the Wage model using the test set are bad since all the residuals are negative. This means that the predictions used from the model are not trustworthy even though the Adjusted R-squared and residuals from the training data are good.

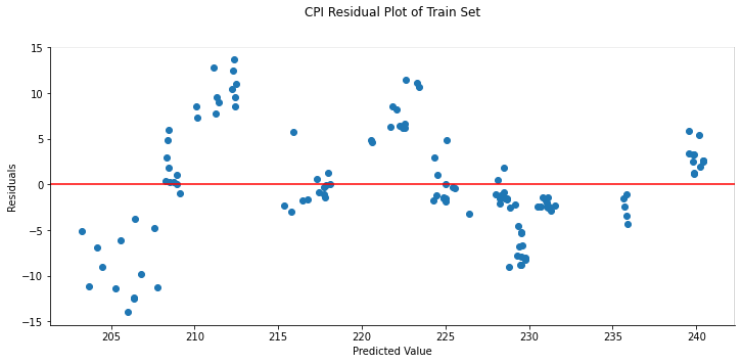

Figure 3.5.3. CPI Residual Plot of Train

Figure 3.5.3 shows that the residuals for the CPI model using training data look pretty good though not the best since clusters of values can be seen from the figure.

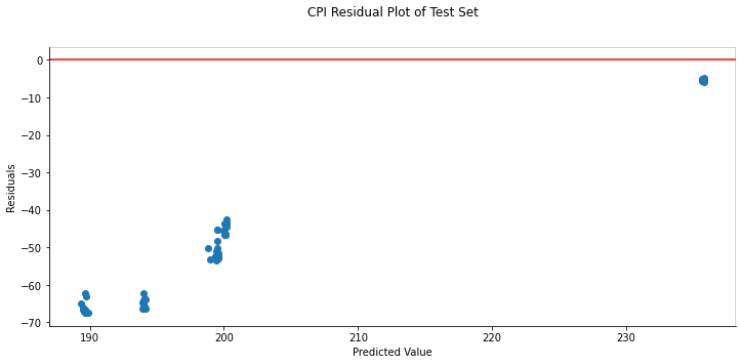

Figure 3.5.4. CPI Residual Plot of Test

As seen in **Figure 3.5.4**, the residuals are all negative and clusters can be seen. This means that the model is not trustworthy for unseen data.

3.6. Results for the Models

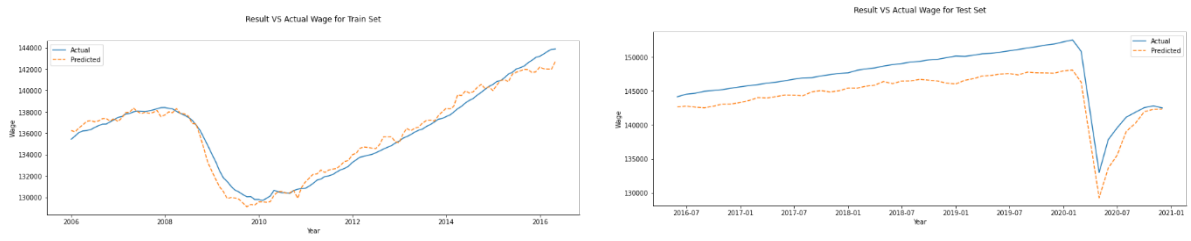

Figure 3.6.1. Train and Test Result of Wage Model

Figure 3.6.2. Train and Test Result of CPI Model

Figure 3.6.1 and **Figure 3.6.2** represents the result of the model when fitted with the train and test set. It can be seen that for the wage model, it can fit both datasets well and generally able to follow the trend. For the CPI model, it does not follow the actual data well for both datasets, it has a tendency to be quite jumpy and tend to be quite off from the actual values especially for the test set (the train set was still able to follow the actual data quite well even with its relatively jumpy characteristic).

3.7. Assumption Testing

When the authors used the Simultaneous Equation Model, there were 3 classical assumptions which must be fulfilled, namely:

- Multicollinearity
- Normality: the regression's error should be normally distributed.
- Heteroscedasticity: the variance of the error term should be constant.

3.7.1. Multicollinearity

Multicollinearity means that the regressors should not be highly linearly correlated to each other. First, the authors are going to test whether our features have a multicollinearity problem. One way of checking this is to calculate the Variance Inflation Factor or VIF in short. A VIF of 5 or higher can be considered to have a high probability of multicollinearity problems (see **Table 1**).

The hypothesis testing for multicollinearity is as follows:

H_0:There is no multicollinearity problem.

H_1:There is a multicollinearity problem.

Decision Rule:

H_0 is rejected if VIF > 5.

The authors will check the VIF for each model's regressor. The system of equations can be seen on **Equation (1)** and **Equation (2).**

When the authors were checking the VIF, the endogenous variables must be excluded. (since these variables will be highly correlated to each other). The Wage equation (W_t) only have one non-endogenous variable, thus it is impossible for this equation to have multicollinearity problems. We will now calculate the Price Equation (P_t).

Variable	VIF
Raw Material	1.032001611175605
Cost of Capital	1.032001611175605

Table 6. VIF Values of 2 SLS Regression

The VIF values for both columns are 1.032001611175605, which is less than 5, thus the authors can conclude that the regressors for the Price Equation do not have multicollinearity problems.

3.7.2. Normality

The assumption of normality means that the residuals are expected to be normally distributed. There are many statistical tests one can use to determine the normality, but for this case, the authors used D'Agostino's K-Squared Test. The following are the hypothesis and rejection region for the test:

H_0: Normally distributed

H_1: Not normally distributed

Decision Rule:

H_0 is rejected if the p-value (D'Agostino's K Squared test) $< \alpha$, which is equal to 0.05.

The result from the test can be seen from the table below:

Equation	Train/Test	P-Value
Wage	Train	0.0098
	Test	0.64
CPI	Train	0.85
	Test	0.0002

Table 7. P-Values from the D'Agostino's K-Squared Test

From the result on **Table 7**, it can be clearly seen that there were two datasets from two different models which yield residuals which are not normally distributed (which meant that the p-value was less than α). Those residuals came from the train set of the Wage model and the test set of the CPI model.

3.7.3. Heteroscedasticity

One of the tools that can be used to test heteroscedasticity is using the Breusch-Pagan Test. The tool is available on the 'statsmodels' package and using the function het_breuschpagan(). The p-values are as follows:

Equation	Train/Test	P-value
Wage	Train	0.32558276714959444
	Test	0.0008135307925442346
CPI	Train	$4.45645138287269 \cdot 10^{-26}$
	Test	$4.22683449823684 \cdot 10^{-11}$

Table 8. Statistic Value for Heteroscedasticity

From the result on **Table 8**, the authors can conclude that there is heteroscedasticity in the Wage (train) and CPI equation since the p-values are above 0.05.

4. Conclusion

From the simultaneous equations, the authors made two models, one to estimate wage and one to estimate CPI. From the Adjusted R-squared both models look good. However, the RMSE and MSE of the two models using training data and testing data are very different. The MSE and RMSE using the testing data are worse than using the training data. This means that the models are only good for estimating on the data they have trained on but not on data they have not seen before. Therefore, it can be concluded that the models have bad performance. The authors decided not to interfere by adding or removing variables to make the models better since the objective is to check whether the presented model is good or not.

5. References

[1] Banton, Caroline. "Explaining the Wage-Price Spiral and How It Relates to Inflation." Investopedia, Investopedia, 19 May 2021, www.investopedia.com/terms/w/wage-price-spiral.asp.

[2] Mankiw, N. Gregory, et al. *Principles of Macroeconomics*. Cengage, 2018.

[3] Gordon, Jason. "Wage-Price Spiral - Explained." *The Business Professor, LLC*, 1 July 2021, thebusinessprofessor.com/economic-analysis-monetary-policy/wage-price-spiral-definition.

[4] Glen, Stephanie. "Simultaneous Equations Model (SEM): Simple Definition." *Statistics How To*, 5 Dec. 2020, www.statisticshowto.com/simultaneous-equations-model/.

[5] IBM Corporation. *Two-Stage Least-Squares Regression*, www.ibm.com/docs/en/spss-statistics/version-missing?topic=regression-two-stage-least-squares.

[6] Gujarati, Damodar N., and Dawn C. Porter. *Basic Econometrics*. McGraw-Hill/Irwin, 2017.

[7] Hoover, Kevin D., et al. "Phillips Curve." *Econlib*, 5 Feb. 2018, www.econlib.org/library/Enc/PhillipsCurve.html.

[8] Simon, P. Neill, and Reza Hashemi. "Root-Mean-Squared Error." *Root-Mean-Squared Error - an Overview | ScienceDirect Topics*, 2018, www.sciencedirect.com/topics/engineering/root-mean-squared-error.

[9] Glen, Stephanie. "Hausman Test for Endogeneity (Hausman Specification Test)." *Statistics How To*, 16 Sept. 2020, www.statisticshowto.com/hausman-test/.

[10] Brownlee, Jason. "A Gentle Introduction to k-Fold Cross-Validation." *Machine Learning Mastery*, 2 Aug. 2020, machinelearningmastery.com/k-fold-cross-validation/.

[11] Frost, Jim, et al. "Multicollinearity in Regression Analysis: Problems, Detection, and Solutions." *Statistics By Jim*, 12 June 2021, statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.

[12] Chandran, Sundaresh. "Significance of Q-Q Plots." *Medium*, Towards Data Science, 6 June 2021, towardsdatascience.com/significance-of-q-q-plots-6f0c6e31c626.

[13] Waskom, Michael. “Seaborn.kdeplot.” *Seaborn.kdeplot - Seaborn 0.11.1 Documentation*, 2012, seaborn.pydata.org/generated/seaborn.kdeplot.html.

[14] Flórez, Jeffersson, et al. “Raghunath.” *Real Statistics Using Excel*, 2017, www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/dagostino-pearson-test/.

[15] Hayes, Adam. “Heteroskedasticity.” *Investopedia*, Investopedia, 7 July 2021, www.investopedia.com/terms/h/heteroskedasticity.asp.

[16] Breusch, Trevor, and Adrian Pagan. *Heteroscedasticity*, 2017, cran.r-project.org/web/packages/olsrr/vignettes/heteroskedasticity.html.

Acknowledgement

The authors would like to thank God for His divine grace and endless blessings during this tough period. Further appreciation for Gilbert Aurelio Sachio, Brandon Gabrielle Soetrisno, and last but not the least, Ms. Maria Zefanya Sampe, and fellow colleagues for the continuous support during the research.