Multiple Regression
Times Series Analysis

CSX2006/ITX2006
Mathematics and Statistics for Data Science

Kittiphop Jitlerdwilai ID 6216916
Pakin Charoenchanachai ID 6210195
Rattanapol Wattanasup ID 6211515

FULL REPORT

ABSTRACT

This report has been created for use in the Final report for CSX2006 / ITX2006 Mathematics and Statistics for Data Science. This report contains information about how statistics impact our daily life and business, Benefits of using statistics in business and dairy life, and some examples of working fields where they use statistics to help their job This report also includes an example dataset for Multiple regression and Time series analysis and explanation.

TABLE OF CONTENTS

Chapter 1: State of Problem

Past ten to twenty years, our world has changed a lot. There are many new technologies which are well-developed better than what we have in the past. Indeed, there are many technologies that have been improved to use in business. And have created competition against business. Now every single kind of business has to rely on statistics. Even in sports, it has statistics to predict the overall. Statistics also come to take place in our diary life example for Disease predicting, Political campaign, Weather forecast, and many other more. A lot of prediction solution has been using in present. We have chosen a few solutions from which are multiple regression and times series regression to explain further in this report.

In our current project, we decided to predict the value of systolic blood pressure with the multi regression technique. In these modern days, there are many people who are unhealthy or came up with the disease. So, we would like to say that this systolic blood pressure is also counted as a factor of measuring the healthiness of the people. So, we came up with an idea to predict the systolic blood pressure by using ages and weight. Why age? We are using age because at different ages we have a different growth rate of blood pressure and also why using weight? We are using weight because it tells about people's body structure whether they are fat or thin. We are going to study and make sure that this predicted systolic blood pressure can be used as a standard at a specific age and current weight of themselves.

In the second project, we would like to predict the temperature change of Anhui province within the Republic of China. As you can see, we can say that those seasonal changes are affecting the people within that province since the Anhui province is doing agriculture, so the temperature or the weather is affecting the way of their living and work. So, we are using temperature data that we have to predict the trends.

Chapter 2: Objectives

- To study or predict the blood pressure growth with the age and weight of the patient whether it is high or low and take that predicted value as a standard then we compared it to the actual systolic blood pressure.

- To study and test the assumption that the GDP growth rate from the period of 2012 to 2014 of Australia.

Chapter 3: Benefits

In the time when every company has been using statistics to predict and analysis. This is why understand what regression analysis and linear regression and regression method is that much impacted to our business will gain your business an opportunity to get succeed in the future. And this is some benefit example of using them:

1. Performance management
   - To analyze where were the gaps in the performance management.
   - Know your business strong points.
   - Know your business weak points.
2. Product development
   - Develop product according to condition of the market.
   - Product improvement.
   - Analyze feedback from customer.
3. Cost Analysis
   - Manage business financial.
   - Know what part you need to spend money.
   - Manage cost of ensuring the benefit.
4. Risk/Return on Investments
   - Optimize the return
   - Minimize the risk
   - evaluate the project under different economic environments

Benefit of using regression analysis to your business. You can predict some pattern from data or feedback from customer. Also, can predict the efficiency, a strong point of your business. And this is some of the benefit from using regression analysis:

1. Predictive Analytics
   - predicts the number of items which a consumer will probably purchase.
   - Financial management
2. Operation Efficiency
   - Analyze service performance
   - Maximize the impact on the operational efficiency and revenues.
3. Supporting Decisions
   - smarter and more accurate decisions.
   - test a hypothesis before diving into execution.
4. Correcting Errors
   - identifying errors in judgment.
   - quantitative support for decisions and prevent mistakes due to manager’s intuitions.
5. New Insights
   - potential to yield valuable insights.
   - analysis of data from point of sales systems and purchase accounts.

Chapter 4: Contributions

Statistics plays a vital role in every field of human activity. Statistics helps in determining the existing position of per capita income, unemployment, population growth rates, housing, schooling medical facilities, etc., in a country.

Now statistics holds a central position in almost every field, including industry, commerce, trade, physics, chemistry, economics, mathematics, biology, botany, psychology, astronomy, etc., so the application of statistics is very wide. Now we shall discuss some important fields in which statistics is commonly applied. And this is some field example which statistic taking place:

1. Business
   - Statistics play a very big role in business. Many successful businessmen must have a very quick and accurate in solving problem and decision making. In this case statistic help them to analysis the data and feedback, for easier and quickly decision making
2. Economics
   - Economics largely depends upon statistics. National income accounts are multipurpose indicators for economists and administrators, and statistical methods are used to prepare these accounts. In economics research, statistical methods are used to collect and analyze the data and test hypotheses.
3. Mathematics
   - Statistics helps in describing these measurements more precisely. Statistics is a branch of applied mathematics. A large number of statistical methods like probability averages, dispersions, estimation, etc., is used in mathematics, and different techniques of pure mathematics like integration, differentiation and algebra is used in statistics.
4. Banking
   - Statistics plays an important role in banking. Banks make use of statistics for a number of purposes. They work on the principle that everyone who deposits their money with the banks does not withdraw it at the same time. The bank earns profits out of these deposits by lending it to others on interest. Bankers use statistical approaches based on probability to estimate the number of deposits and their claims for a certain day.
5. Astronomy
   - Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, and sizes, masses and densities of heavenly bodies by means of observations. During these measurements' errors are unavoidable, so the most probable measurements are found by using statistical methods.

As we have explained about what statistics done in difference fields in our life. After what I have mentioned before we can know that statistics is improving our life quality. Almost every part in the world have use statistics to improving in many parts all over.

Chapter 5: Theorems

5.1 Regression Analysis

The one of the process in Statistical Modeling is Regression Analysis. Regression analysis is look how dependent variable relate to invariable variables. The basis way of prediction is linear regression. But in this report, we have chosen multiple regression analysis to predict the values from the given dataset. But for multiple regression there will be other 2 more predictor needs

5.2 Multiple Regression

5.2.1Equations

- The multiple linear regression model uses the equation:

- The estimated multiple linear regression equations:

5.2.2Building a Multiple Regression Model

5.2.3 Steps in developing a regression model

The steps to developing a multiple regression model are as follows:

Step 1 Construct the scatter plot for each y and x_i .

Step 2 Compute the correlation coefficients r_{yxi} , $r_{xi}x_j$, and then conduct the hypothesis testing of those population correlation coefficients. The hypotheses and test statistic are

Hypotheses	Test statistic	Rejection rule
1. $H_0 : \rho_{yx_i} = 0$ vs $H_1 : \rho_{yx_i} \neq 0$	$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ $d.f. = n - 2$	Reject H_0 at α , if the value of test statistic $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$.
2. $H_0 : \rho_{x_i x_j} = 0$ vs $H_1 : \rho_{x_i x_j} \neq 0$		

Step 3 Diagnosing the **effects of multicollinearity** and correct them.

- **Multicollinearity** is the condition where among the independent variables are correlated with each other.
- **Detection**
  - Correlation matrix

  - Variance inflation factors for the independent variable x_j
    - VIF_j = 1 implies x_j not related to other predictors
    - Largest VIF_j is greater than ten suggest severe multicollinearity
    - Average VIF substantially greater than one suggests severe multicollinearity

- **Impact of Multicollinearity**
  - *Makes the value of R^2 to increase. This means that it is not the real value of R^2 of our model.*

Step 4 Compute the regression coefficients b_i and then conduct the hypothesis testing of the population regression coefficients and the regression model.

Hypotheses concerning the significance of the regression coefficients

Hypotheses	Test statistic	Rejection rule
$H_0 : \beta_i = 0$ $H_1 : \beta_i \neq 0$	$T_i = \dfrac{b_i - \beta_i}{S(b_i)}$	Reject H_0 at α , if $p-value < \alpha$. It implies that x_j has the effect on y

Hypotheses concerning the significance of the regression model as a whole

Hypotheses	Test statistic	Rejection rule
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ H_1 : At least one of $\beta_1, \beta_2, \cdots, \beta_k$ does not equal 0	$F = \dfrac{MSR}{MSE}$	Reject H_0 at α , if $p-value < \alpha$. If H_0 is rejected, it implies that at least one of the independent variables $x_1, x_2, x_3, \cdots, x_k$ contributes significantly to the model.

Analysis of Variance for Testing

Source of variation	d.f.	SS.	MS.	Test statistic
Regression Error	k $n-k-1$	SSR SSE	MSR MSE	$F = \dfrac{MSR}{MSE}$
Total	n - 1	SST		

Step 5 Diagnosing the outlier(s) and influential observation(s) and correct them.

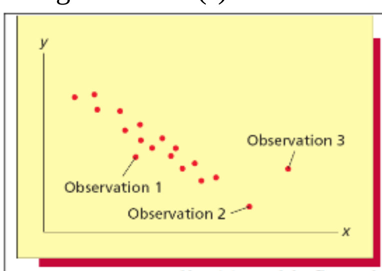


Figure 15.1 outlier(s) and influential observation(s)

- Observation 1: Outlying with respect to y value
- Observation 2: Outlying with respect to x value
- Observation 3: Outlying with respect to x value and y value not consistent with regression relationship (Influential)

- **Cook's Distance Measure** can use to identify influential observations
    1. An observation will be the influential observations, if the value of D_i >1
- **Leverage values** can help us identify outliers

  The leverage value for an observation is the distance value. This value is a measure of the distance between the x value and the center of the experimental region
- If the leverage value for an observation is large, it is an outlier with respect to its x value

2. Large means greater than twice the average of all the leverage values
3. An observation will be the outliers, if the leverage value is greater than

$2(k+1)$ / n

What to do About Outliers?

- First, check to see if the data was recorded correctly – If not correct, discard the observation and rerun.

4. If correct, search for a reason for the observation. It might be caused by a situation we do not wish to model. If so, drop the observation.

- If no reason found, consider that there might be an important independent variable not currently included in the model.

Step 6 Developing the best estimated regression model.

Step 7 Conduct the residual analysis of the model is obtained from step 6 with the following assumptions:

  - Constant Variance Assumptions, by examining residual plots against the predicted *y* values.
  - Normality Assumption and $E(\varepsilon) = 0$, by using the Anderson Darling test statistic.
  - Independence Assumption, by using the Durbin-Watson test statistic.

Transforming the Dependent and Independent Variables

- A possible remedy for violations of the constant variance, correct functional form and normality assumptions is to transform the dependent variable
- Possible transformations include: Square root, Quartic root, Logarithmic
- The appropriate transformation will depend on the specific problem with the original data set

Step 8 Indicate the best estimated regression model.

5.3 Time Series Technique

5.3.1 Moving Average

An average value for time period(t) could see by the mean of k value. In an average value has been assigned to each observation. The moving average is not working well with trend and seasonal type. The analyst needs to pick a value of periods *k*, and a moving average. The big or large number is likeable when widely in use and the value is not undulation in the series value. Many times, the moving series are using with quarterly type and monthly type to help to clear the time series. For quarterly data, the four-quarter moving average value, MA(4), yields an average of the four quarters, and for monthly data, a 12-month moving average, MA(12), delete or averages out the effects from seasonal. The order of moving average and the smoothing effect will be increasing together in order.

5.3.2 Single Exponential Smoothing

For Single Exponential Smoothing is a prediction method use to predict the unequal value to the time series. This unequal value is maneuverable by using a smoothing constant that determines how much value is attached to each value separately. Simple exponential smoothing is not created to use with trend or seasonality data, but for Double exponential smoothing are well fitted to trended data of the form.

Double Exponential Smoothing equation:

- $\hat{y}_{t+p} = L_t + pT_t$ = forecast for p periods into the future
- $L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1})$ = the exponentially smoothed series or current level estimate

  regularly, we could set the first smoothed level equal to the first observation
- $L_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$
- α = smoothing constant for the level ($0 < \alpha < 1$)
- β = smoothing constant for trend estimate ($0 < \beta < 1$)
- p = periods to be forecast into the future

5.3.4 Time Series Regression (Seasonal Without Trend)

Time series regression is a technique using for predicting a future response based on the recorded history and the transfer of dynamics from related predictors. Time series regression could give a help so you could understand and predict the behavior of dynamic systems from experience or the data which has been recorded. Time series regression are regularly used with the modeling and forecasting of economic, financial, and biological systems. We can start a time series analysis by First, we need to build a design matrix (), which can include the observations data ordered by time.

Second, apply ordinary least squares to the multiple linear regression model. Then, to get a prediction of a linear relationship of the response and the design matrix. β is representing the linear parameter estimates to be computed and () represents the innovation terms. The terms can be extended in the MLR model to including the heteroscedasticity or autocorrelation effects.

5.3.5 ARIMA

ARIMA models one of the techniques use for time series prediction. Illustrated the smoothing and ARIMA models are the two most popular technique use to predict the time series and add complementary approaches to the problem. During the exponential smoothing models are relate to the description of the trend and seasonality data, The ARIMA models focusing to explain about the autocorrelations in the data.

Chapter 6: Datasets and Stories

In this part we will present a dataset which we will use in this report.

SystolicBP_Y	Age_X1	Weight_X2
132	52	78
143	59	83
153	67	87
162	73	95
154	64	89
168	74	100
137	54	85
149	61	85
159	65	93
128	46	75
166	72	98
132	52	78
143	59	83
153	67	87
162	73	95

Table 6.1: Systolic Blood Pressure by Age and Weight

In this data set, some says that it was collected manually and there is no specific source. I found this data set from [https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr02.html] and I think that I can make use of it and develop it to be used.

Time	GDP growth rate
Q1_2008	1.1
Q2_2008	0.2
Q3_2008	0.8

Q4_2008	-0.5
Q1-2009	0.1
Q2-2009	0.6
Q3-2009	0.3
Q4-2009	0.7
Q1-2010	0.5
Q2-2010	0.5
Q3-2010	0.7
Q4-2010	1.0
Q1-2011	-0.3
Q2-2011	1.2
Q3-2011	1.4
Q4-2011	1.1
Q1_2012	0.9
Q2_2012	0.8
Q3_2012	0.6
Q4_2012	0.6
Q1_2013	0.3
Q2_2013	0.5
Q3_2013	0.8
Q4_2013	0.8
Q1_2014	0.7
Q2_2014	0.5
Q3_2014	0.5
Q4_2014	0.4
Q1_2015	0.8
Q2_2015	0.1
Q3_2015	1.1
Q4_2015	0.6
Q1_2016	0.9
Q2_2016	0.7
Q3_2016	0.1
Q4_2016	1.0
Q1_2017	0.3
Q2_2017	0.7
Q3_2017	1.0
Q4_2017	0.5
Q1_2018	0.9
Q2_2018	0.8
Q3_2018	0.3
Q4_2018	0.2
Q1_2019	0.5

Q2_2019	0.6
Q3_2019	0.6
Q4_2019	0.5
Q1_2020	-0.3

Table 6.2: Monthly growth rate of the GDP from the year of 2008 to 2019 of the Australia country.

Chapter 7: Results of Analysis

7.1 Multiple Regression Model: Systolic blood pressure prediction

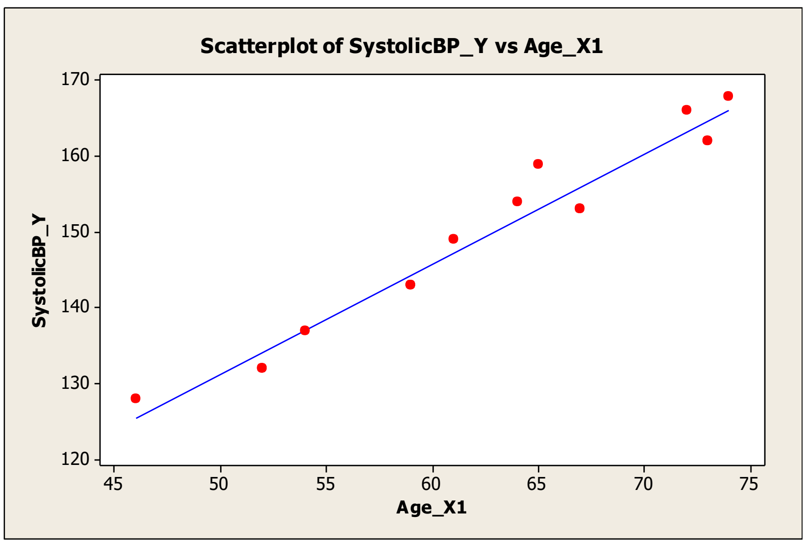


Figure 7.1 Scatterplot of Systolic blood pressure(y) vs Age(x_1)

From the scatter plot of Systolic(y) and Age(x_1) above, we can see that there are positively related in a linear sense.

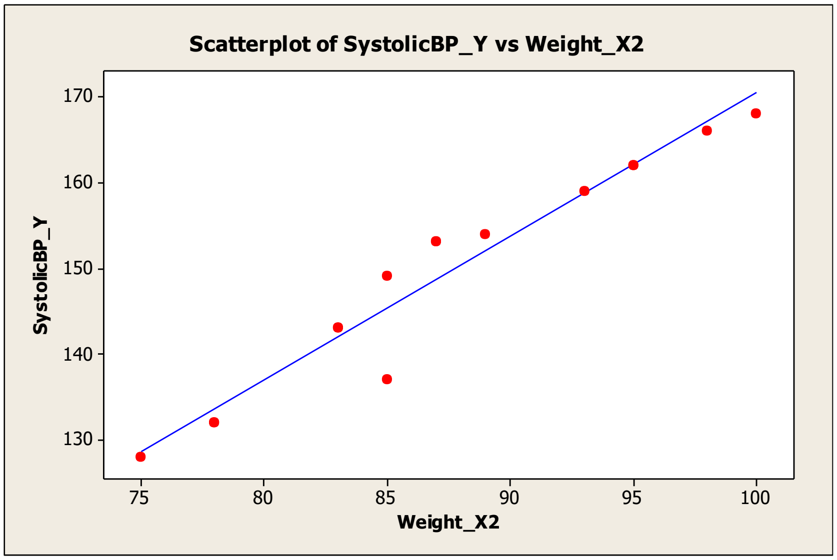


Figure 7.2 Scatterplot of Systolic blood pressure(y) vs Weight(x_2)

From the scatter plot of Systolic(y) and Weight(x_2) above, we can see that those plots are related in a linear sense.

Correlations: SystolicBP_Y, Age_X1, Weight_X2

```
             SystolicBP_Y      Age_X1
Age_X1              0.977
                    0.000

Weight_X2           0.970       0.939
                    0.000       0.000


Cell Contents: Pearson correlation
               P-Value
```

(1)H0: $\rho yx1=0$ (There is no linear relationship between the Age (x1) and Systolic Blood Pressure (y))

H1: $\rho yx1\neq0$ (There is a linear relationship between the Age (x1) and Systolic Blood Pressure (y))

We see that the correlation coefficient of 0.977 is close to one and indicates a high correlation between the Age (x1) and Systolic blood pressure (y). However, the p-value 0 is smaller than a = 0.05 so we reject the null hypothesis H0 and conclude that there is a linear relationship between Age (x1) and Systolic blood pressure (y).

(2) H0: $\rho yx2=0$ (There is no linear relationship between the Weight (x2) and Systolic Blood Pressure (y))

H1: $\rho yx2\neq0$ (There is a linear relationship between the Weight (x2) and Systolic Blood Pressure (y))
We see that the correlation coefficient of 0.970 is very close to one and indicates a high correlation between the Weight (x2) and Systolic Blood Pressure (y); they have the p-value 0 is lower than a = 0.05

so we reject the null hypothesis H0 and conclude that there is linear relationship between weight (x2) and Systolic blood pressure (y).

(3) H0: $\rho x1x2=0$ (There is no linear relationship between the Age (x1) and Weight (x2))

H1: $\rho x1x2\neq 0$ (There is a linear relationship between the Age (x1) and Weight (x2))

We see that the correlation coefficient of 0.939 is very close to 1 and indicates a high correlation between the Age (x1) and Weight (x2). The p-value 0 which is lower than α = 0.05 so we reject H0 and conclude that there is linear relationship between Age (x1) and Weight (x2).

1)Fit-Regression Analysis: Systolic blood pressure Y versus Age X1 and Weight X2

Regression Analysis: SystolicBP_Y versus Age_X1, Weight_X2

```
The regression equation is
SystolicBP_Y = 30.5 + 0.828 Age_X1 + 0.767 Weight_X2


Predictor    Coef  SE Coef     T      P
Constant   30.544    8.686  3.52  0.004
Age_X1     0.8281   0.1804  4.59  0.001
Weight_X2  0.7674   0.2098  3.66  0.003


S = 2.04964   R-Sq = 97.9%   R-Sq(adj) = 97.5%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  2327.2  1163.6  276.98  0.000
Residual Error  12    50.4     4.2
Total           14  2377.6


Source     DF  Seq SS
Age_X1      1  2271.0
Weight_X2   1    56.2


Unusual Observations

Obs  Age_X1  SystolicBP_Y      Fit  SE Fit  Residual  St Resid
  7    54.0       137.000  140.491   1.204    -3.491     -2.10R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.82287
```

Conclusion: With this model since the p-value of y-intercept is 0.004 which is less than α = 0.05, this mean that y-intercept is significant and this model can be used to predicted.

Since this model can be able to predicted let's proceed to the next step.

Durbin-Watson Statistic

From Table of the Critical Values for the Durbin-Watson Statistic at $\alpha = 0.05$,
n = 15, and k = 2. We can get $d_L = 0.946$ and $d_U = 1.543$

Regions of Acceptance and Rejection of the Null Hypothesis				
Reject H_0, It has the positive autocorrelation.	The test is inconclusive.	Accept H_0: There is no autocorrelation.	The test is inconclusive.	**Reject H_0, It has the negative autocorrelation.**

0   $d_L = 0.946$ $d_U = 1.543$ $4 - d_U = 2.457$ $4 - d_L = 3.054$

Since the first model has the Durbin-Watson of 1.82287 which is in the part of Accepting H_0 so there's no autocorrelation.

Normal Distribution Assumption

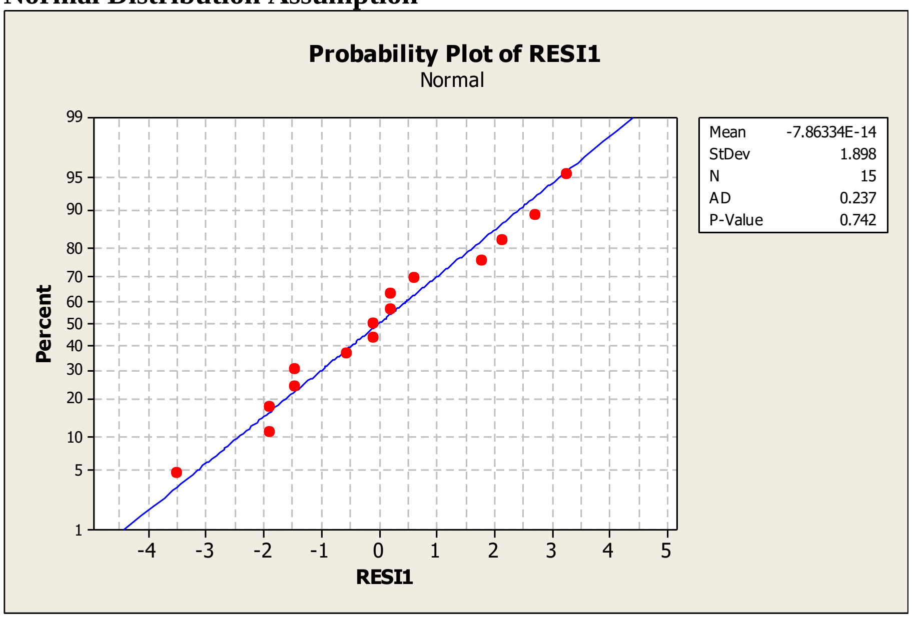


Figure 7.3: Probability plot of model 1

The mean of the residual = -7.86334E-14 and it means that's E(ε) = 0. And the p-value = 0.742 which is greater than $\alpha = 0.05$, so H_0 is accepted and conclude that the residual distribution are normally distributed.

Constant Variance Assumption

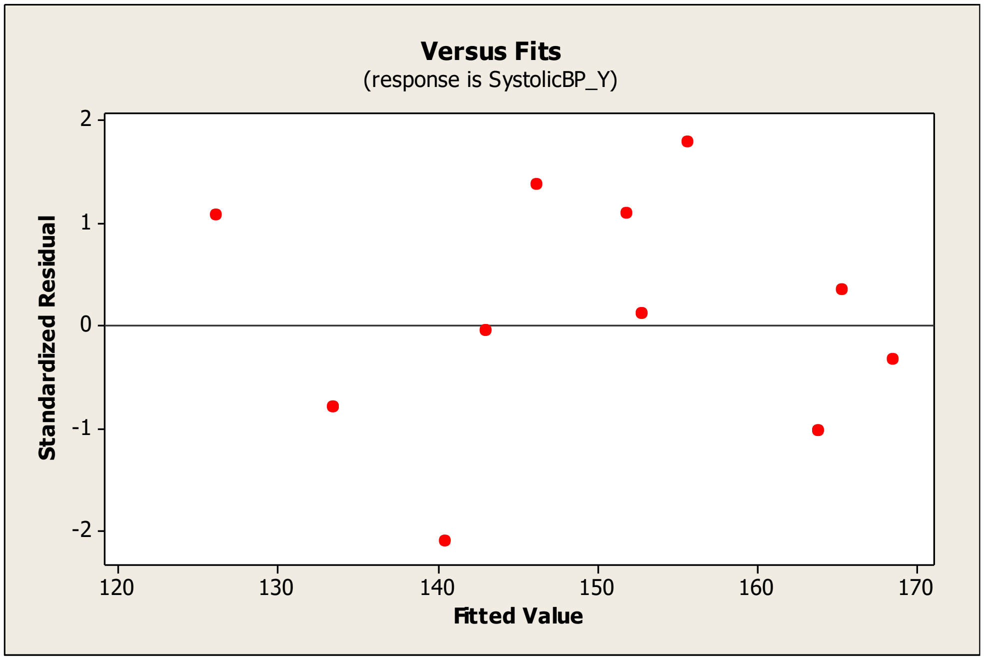


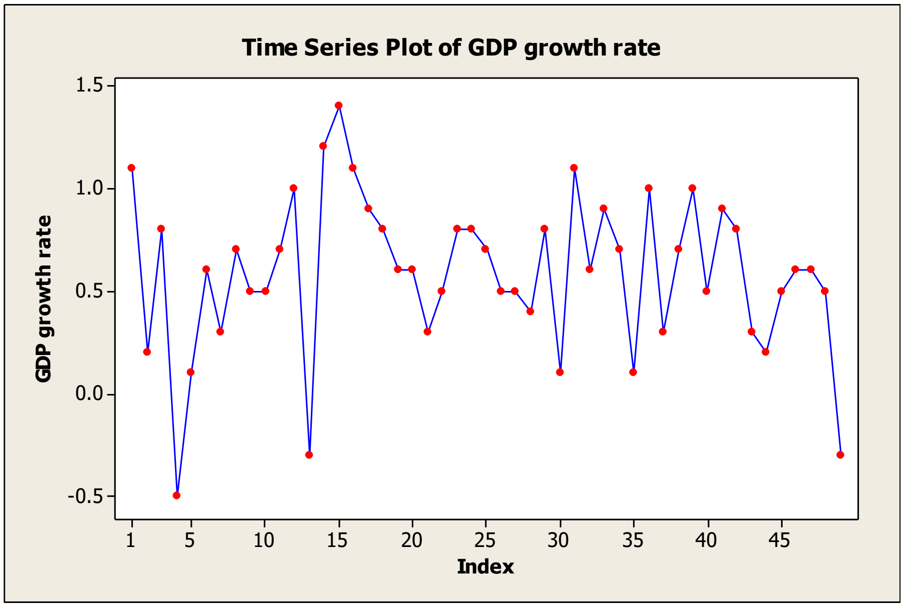
Figure 7.4: Versus Fits of Systolic Blood Pressure (y) of model 1

The standardized residual plot shows that the residual does fluctuate around the mean of zero, so the constant variance assumption is valid.

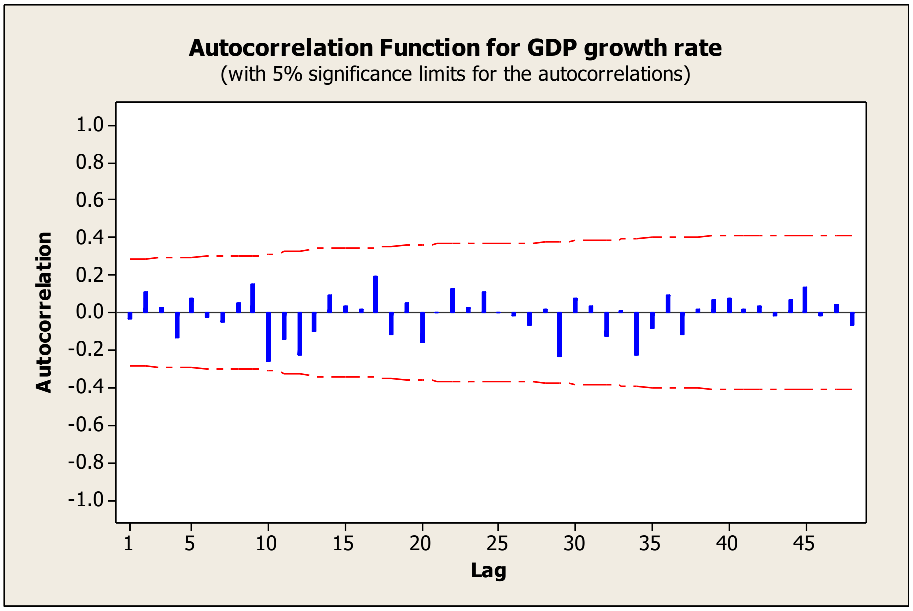
Conclusion for this model 1)

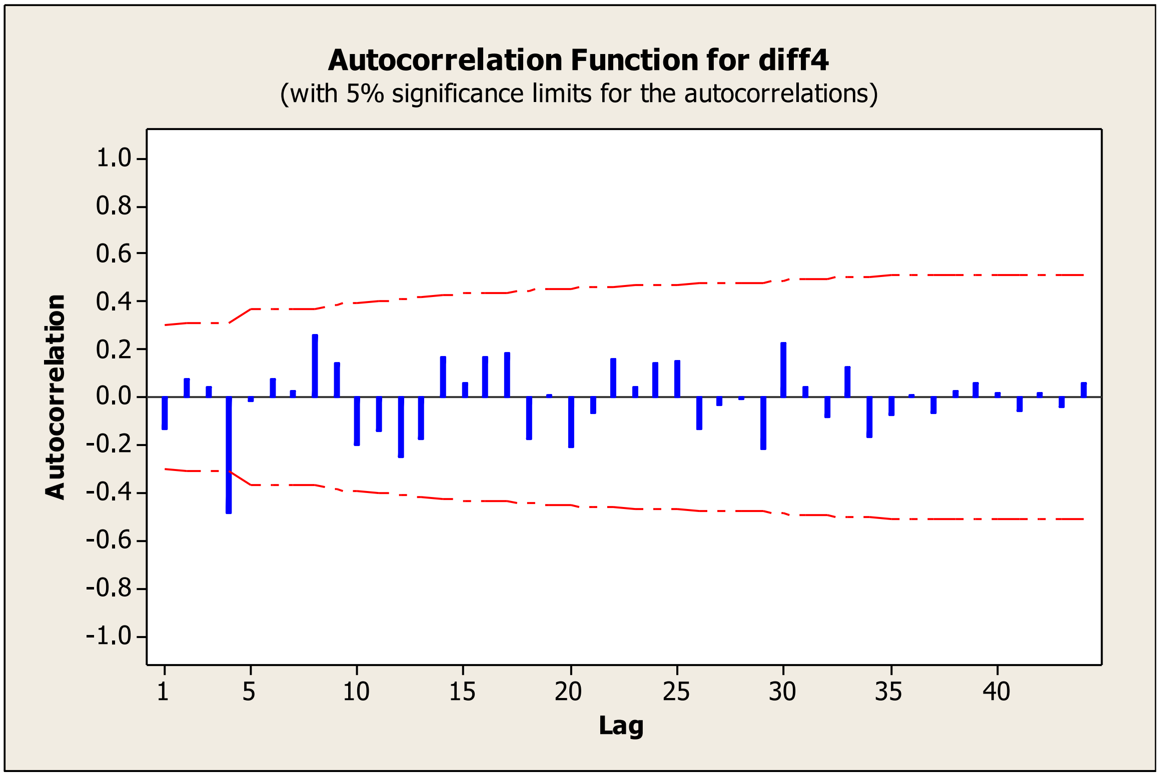
The residuals are normally distributed with the mean of zero, the constant variance assumption is valid and also there is no autocorrelation. The equation `SystolicBP_Y = 30.5 + 0.828 Age_X1 + 0.767 Weight_X2` should be used to predict.

7.2 Time Series ARIMA model: Quarterly GDP growth rate from the year of 2008 to 2019



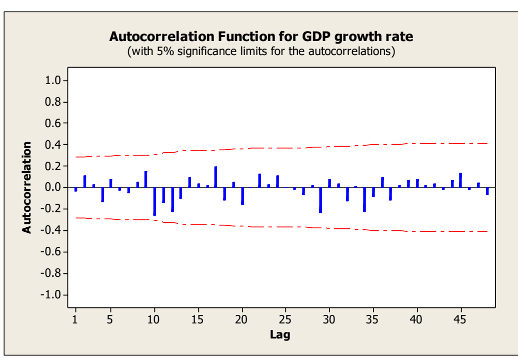
Checking for trends within the data set for regular and seasonal parts.

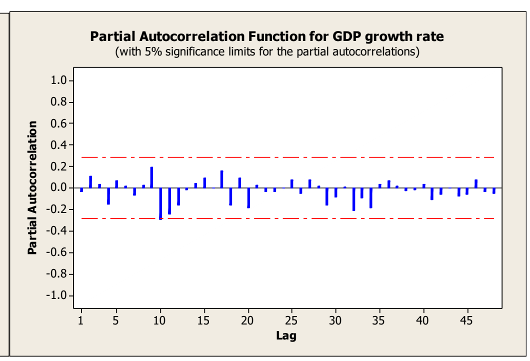




After analyzing those lags, we can see that there are no trends at all so the difference function are not require for this data set.

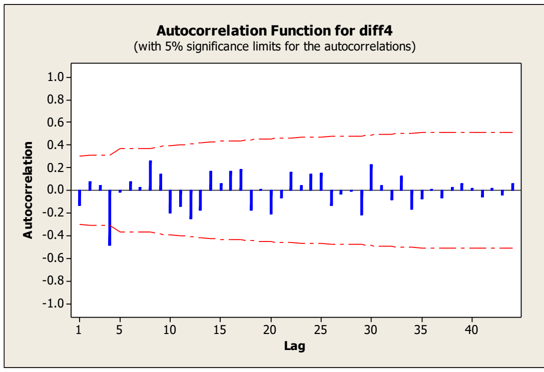
Developing Models

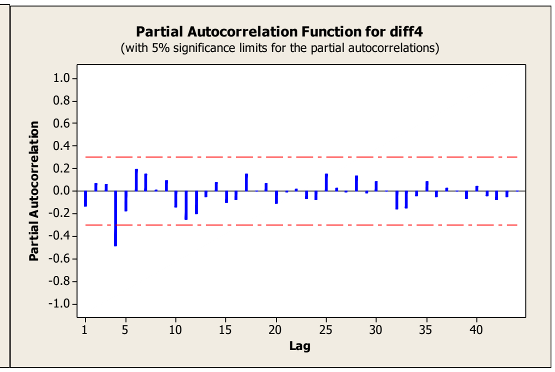




Judging from the pattern of the graph, I decide to use ARIMA $(2,0,1)(0,0,0)_4$ due to the pattern of itself if we compare it to comparison diagram.

Developing model for seasonal parts





Judging from the pattern we can say that in term of seasonal part, we should use model ARIMA(2,0,1)(1,0,1)$_4$ due to the pattern of itself if we compare it to comparison diagram.

Finding the best model

Model 1: ARIMA(2,0,1)(0,0,0)

ARIMA Model: GDP growth rate

```
Final Estimates of Parameters

Type          Coef   SE Coef     T      P
AR    1     0.0394    1.1537  0.03  0.973
AR    2     0.1378    0.1607  0.86  0.395
MA    1     0.0736    1.1647  0.06  0.950
Constant   0.47986   0.05262  9.12  0.000
Mean       0.58324   0.06395


Number of observations:  49
Residuals:    SS =  7.03421  (backforecasts excluded)
              MS =  0.15632  DF = 45


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   13.8   24.9   46.8   79.7
DF              8     20     32     44
P-Value     0.086  0.207  0.044  0.001
```

From all the constant above except the Y value, they are all not significant due to the value which is greater than α=0.05. Also the value of error, MS = 0.15632 .

Model 2: ARIMA(2,0,1)(1,0,1)$_4$

ARIMA Model: GDP growth rate

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	0.3592	0.4646	0.77	0.444
AR	2	0.3081	0.1638	1.88	0.067
SAR	4	-0.9945	0.0336	-29.59	0.000
MA	1	0.4874	0.4818	1.01	0.317
SMA	4	-0.8834	0.1392	-6.35	0.000
Constant		0.37639	0.05454	6.90	0.000
Mean		0.56735	0.08221		

```
Number of observations:  49
Residuals:     SS =  6.01730 (backforecasts excluded)
               MS =  0.13994  DF = 43
```

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	13.1	19.4	40.0	81.6
DF	6	18	30	42
P-Value	0.042	0.371	0.104	0.000

From the p-value of SAR and SMA which is equal to 0 so we can conclude that they are all significant so this model is valid, the MS value is equal to 0.13994 .

Conclusion

The best model that is going to be use to forecast this data set is from ARIMA(2,0,1)(1,0,1)$_4$ because all constant are significant, meanwhile the other model some of the constant are not significant so it can't be used to predict.

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \cdots - \omega_q \varepsilon_{t-q}$$

Chapter 8: Conclusion

8.1) Multiple Regression: systolic blood pressure prediction

Age x_1	**Weight x_2**	**Actual Value**	**PredictedValue**
52	78	132	**133**
59	83	143	**143**
67	87	153	**153**
73	95	162	**164**
64	89	154	**152**
74	100	168	**168**
54	85	137	**140**
61	85	149	**146**
65	93	159	**156**
46	75	128	**126**
72	98	166	**165**
52	78	132	**133**
59	83	143	**143**
67	87	153	**153**

The model uses the variable x_1 and x_2 to predict the systolic blood pressure (Actual Value).

8.2)Times Series: GDP growth rate prediction

Year	Quarter	t	T = 0.6825 - 0.006765t	S	Forecast $\widehat{y_t} = \widehat{T_t} * \widehat{S_t}$
2016	Q1	17	0.567495	1.0639	0.567495 * 1.0639 = 0.603741
	Q2	18	0.560730	0.8755	0.560730 * 0.8755 = 0.490947
	Q3	19	0.553965	0.9366	0.553965 * 0.9366 = 0.518860
	Q4	20	0.547200	1.12396	0.547200 * 1.12396 = 0.615031

By using decomposition method me can forecast these value repectively.

Chapter 9: Appendix

In this chapter we are going to show the possible model for the multi regression data set.
Model 2: Regression Analysis: Systolic blood pressure Y versus Age X1 and Weight X2(Not fit)

Regression Analysis: SystolicBP_Y versus Age_X1, Weight_X2

```
The regression equation is
SystolicBP_Y = 0.395 Age_X1 + 1.43 Weight_X2


Predictor     Coef  SE Coef      T      P      VIF
Noconstant
Age_X1      0.3952   0.1806   2.19  0.047  247.382
Weight_X2   1.4253   0.1299  10.97  0.000  247.382


S = 2.80610


Analysis of Variance

Source          DF      SS      MS         F      P
Regression       2  337081  168540  21404.14  0.000
Residual Error  13     102       8
Total           15  337183


Source     DF  Seq SS
Age_X1      1  336133
Weight_X2   1     948


Unusual Observations

Obs  Age_X1  SystolicBP_Y      Fit  SE Fit  Residual  St Resid
  7    54.0       137.000  142.493   1.452    -5.493     -2.29R

R denotes an observation with a large standardized residual.


Durbin-Watson statistic = 2.27014
```

In this model, due to the extremely high amount of VIF value so this model is going to have a multicollinearity problem and so it is invalid for this model.
Model 3: Fit-Regression analysis: Systolic blood pressure Y versus AgeX1.

```
Regression Analysis: SystolicBP_Y versus Age_X1

The regression equation is
SystolicBP_Y = 58.9 + 1.45 Age_X1


Predictor      Coef  SE Coef      T      P    VIF
Constant     58.877    5.490  10.72  0.000
Age_X1      1.44759  0.08699  16.64  0.000  1.000


S = 2.86387   R-Sq = 95.5%   R-Sq(adj) = 95.2%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  2271.0  2271.0  276.89  0.000
Residual Error  13   106.6     8.2
Total           14  2377.6


Unusual Observations

Obs  Age_X1  SystolicBP_Y      Fit  SE Fit  Residual  St Resid
  9    65.0       159.000  152.971   0.770     6.029     2.19R

R denotes an observation with a large standardized residual.


Durbin-Watson statistic = 0.895180
```

Regions of Acceptance and Rejection of the Null Hypothesis				
Reject H_0, It has the positive autocorrelation.	The test is inconclusive.	Accept H_0: There is no autocorrelation.	The test is inconclusive.	**Reject H_0, It has the negative autocorrelation**.

0 d_L = 0.946 d_U = 1.543 $4 - d_U$ = 2.457 $4 - d_L$ = 3.054

This model is invalid due to the amount of Durbin-Watson value which is 0.895180 so it is reject the H_0 and have a positive autocorrelation since the d_L value is 0.946.

Model 4: Regression analysis: Systolic blood pressure Y versus AgeX_1 (Not fit)

Regression Analysis: SystolicBP_Y versus Age_X1

```
The regression equation is
SystolicBP_Y = 2.37 Age_X1


Predictor       Coef  SE Coef      T      P    VIF
Noconstant
Age_X1       2.37205  0.03543  66.95  0.000  1.000


S = 8.65991


Analysis of Variance

Source          DF      SS      MS        F      P
Regression       1  336133  336133  4482.13  0.000
Residual Error  14    1050      75
Total           15  337183


Unusual Observations

Obs  Age_X1  SystolicBP_Y     Fit  SE Fit  Residual  St Resid
 10    46.0        128.00  109.11    1.63     18.89     2.22R

R denotes an observation with a large standardized residual.


Durbin-Watson statistic = 1.69743
```

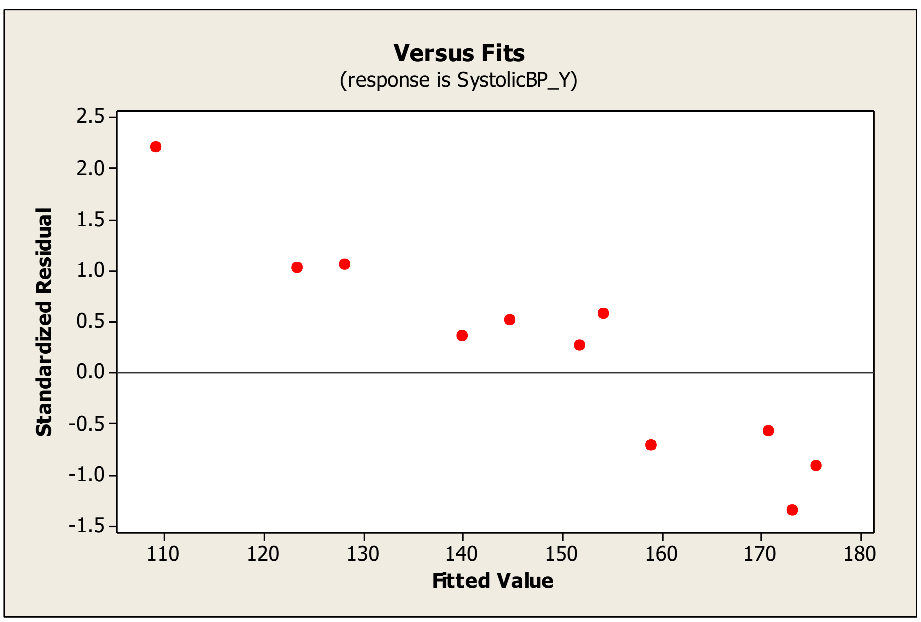


Figure 7.5: Versus Fits of Systolic Blood Pressure (y) of model 4

This model can’t be used because from the constant variance assumption analysis it is invalid due to the non-fluctuate around the 0 mean of the graph so it is invalid.
Model 5: Fit-Regression analysis: Systolic blood pressure Y versus Weight X_2

Regression Analysis: SystolicBP_Y versus Weight_X2

```
The regression equation is
SystolicBP_Y = 3.3 + 1.67 Weight_X2


Predictor    Coef  SE Coef      T      P    VIF
Constant     3.34    10.13   0.33  0.747
Weight_X2  1.6712   0.1155  14.47  0.000  1.000


S = 3.26893   R-Sq = 94.2%   R-Sq(adj) = 93.7%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  2238.7  2238.7  209.50  0.000
Residual Error  13   138.9    10.7
Total           14  2377.6


Unusual Observations

Obs  Weight_X2  SystolicBP_Y      Fit  SE Fit  Residual  St Resid
  7         85       137.000  145.389   0.888    -8.389     -2.67R

R denotes an observation with a large standardized residual.


Durbin-Watson statistic = 2.08312
```

This model is invalid since the p-value of the constant is greater than the $\alpha = 0.05$.

Model 6: Regression analysis: Systolic blood pressure Y versus Weight X_2 (Not fit)

Regression Analysis: SystolicBP_Y versus Weight_X2

```
The regression equation is
SystolicBP_Y = 1.71 Weight_X2


Predictor      Coef  SE Coef       T      P    VIF
Noconstant
Weight_X2   1.70912  0.00931  183.53  0.000  1.000


S = 3.16318


Analysis of Variance

Source          DF      SS      MS         F      P
Regression       1  337043  337043  33684.98  0.000
Residual Error  14     140      10
Total           15  337183


Unusual Observations

Obs  Weight_X2  SystolicBP_Y      Fit  SE Fit  Residual  St Resid
  7         85       137.000  145.275   0.792    -8.275     -2.70R

R denotes an observation with a large standardized residual.


Durbin-Watson statistic = 2.08396
```

This analyst has less VIF so this model is not going to have a multicollinearity problem.

Regions of Acceptance and Rejection of the Null Hypothesis				
Reject H_0, It has the positive autocorrelation.	The test is inconclusive.	Accept H_0: There is no autocorrelation.	The test is inconclusive.	**Reject H_0, It has the negative autocorrelation**.

0 $d_L = 0.946$ $d_U = 1.543$ $4 - d_U = 2.457$ $4 - d_L = 3.054$

This model is invalid due to the amount of Durbin-Watson value which is 2.08396 so it is reject the H_0 and have a positive autocorrelation since the d_L value is 0.946.

References

i. https://www.surveygizmo.com/resources/blog/regression-analysis/
ii. https://medium.com/@JackieMolloy22/the-advantages-of-statistics-in-business-d7c7eda73333
iii. https://www.quora.com/What-is-the-use-of-statistics-in-business
iv. https://bizfluent.com/about-6360783-importance-statistics-industry-business.html
v. https://www.newgenapps.com/blog/business-applications-uses-regression-analysis-advantages/
vi. https://www.emathzone.com/tutorials/basic-statistics/importance-of-statistics-in-different-fields.html