

**Multiple Regression**

**Data Analysis & Research**

CSX2006/ITX2006

Mathematics and Statistics for Data Science

FULL REPORT

Pakin Charoenchanachai ID 6210195

Submitted in Partial Fulfillment of the Requirements for the Course Final Assessment

Semester 1/2020

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, the Benefits of using statistics in business and daily 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.

Contents

[Chapter 1: State of Problem 4](#_Toc88599918)

[Chapter 2: Objectives 5](#_Toc88599919)

[Chapter 3: Benefits 5](#_Toc88599920)

[Chapter 4: Contributions 6](#_Toc88599921)

[Chapter 5: Theorems 7](#_Toc88599922)

[Chapter 6: Datasets and Stories 12](#_Toc88599923)

[Chapter 7: Results of Analysis 14](#_Toc88599924)

[Chapter 8: Conclusion 23](#_Toc88599925)

[Chapter 9: Appendix 24](#_Toc88599926)

# 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, many technologies 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 others more. A lot of prediction solution has been used at 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, many people 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 in measuring the healthiness of the people. So, we came up with an idea to predict 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 use 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.

# 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.

# Chapter 3: Benefits

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

1. Performance management

* To analyze where were the gaps in performance management.
* Know your business strong points.
* Know your business weak points.

1. Product development

* Develop product according to the condition of the market.
* Product improvement.
* Analyze feedback from customers.

1. Cost Analysis

* Manage business finances.
* Know what part you need to spend money on.
* Manage the cost of ensuring the benefit.

4. Risk/Return on Investments

* Optimize the return
* Minimize the risk
* evaluate the project under different economic environments

The benefit of using regression analysis to your business. You can predict some pattern from data or feedback from customers. Also, can predict efficiencies, a strong point of your business. And there are some of the benefits from using regression analysis:

1. Predictive Analytics

* predicts the number of items that a consumer will probably purchase.
* Financial management

1. Operation Efficiency

* Analyze service performance
* Maximize the impact on operational efficiency and revenues.

1. 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 help 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 there are some field examples of which statistics taking place:

1. Business

* Statistics play a very big role in business. Many successful businessmen must have very quick and accurate in solving problems and decision making. In this case, statistics help them to analyze the data and feedback, for easier and quick decision making

1. 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.

1. 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.

1. Banking

* Statistics plays an important role in banking. Banks make use of statistics for several 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 them to others on interest. Bankers use statistical approaches based on probability to estimate the number of deposits and their claims for a certain day.

1. 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 were done in different fields in our life. After what I have mentioned before we can know that statistics is improving our life quality. Almost every part of the world has used statistics to improve in many parts all over.

# Chapter 5: Theorems

5.1 Regression Analysis

One of the processes in Statistical Modeling is Regression Analysis. Regression analysis looks at how dependent variables relate to invariable variables. The basic 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

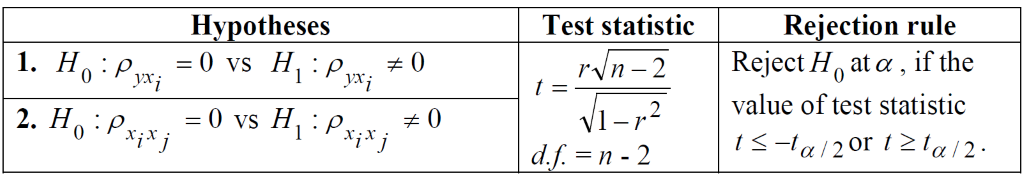
5.2 Multiple Regression

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

Step 1 Construct the scatter plot for each *y* and *xi*.

Step 2 Compute the correlation coefficients *ryxi*, *rxix j*, and then conduct the hypothesis

        testing of those population correlation coefficients. The hypotheses and test statistics are

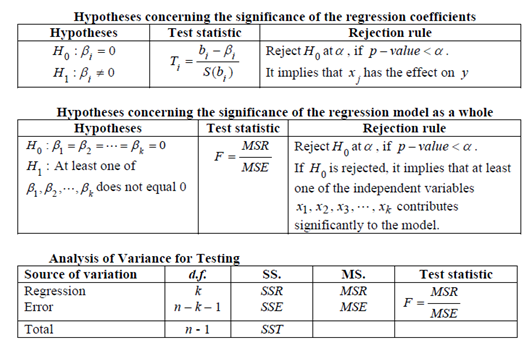


 Step 3 Diagnosing the effects of multicollinearity and correcting them.

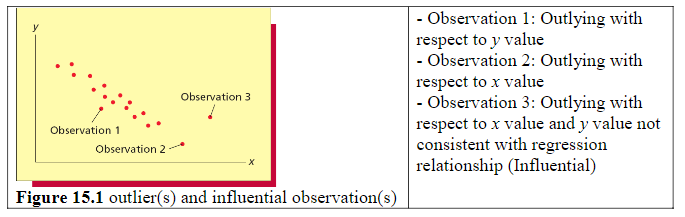
* Multicollinearity is the condition where the independent variables are correlated with each other.
* Detection
* Correlation matrix
* Variance inflation factors for the independent variable *x j*
* *VIFj* = 1 implies *x j* not related to other predictors
* The largest *VIFj* is greater than ten suggest severe multicollinearity
* Average *VIF* substantially greater than one suggests severe multicollinearity

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

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



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



* Cook’s Distance Measure can use to identify influential observations.
* An observation will be the influential observations if the value of *Di* >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.
* Large means are greater than twice the average of all the leverage values.
* 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.
* 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 is 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*(ε) = 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 the time period(t) could see by the mean of k value. An average value has been assigned to each observation. The moving average is not working well with trend and seasonal types. The analyst needs to pick a value of periods *k*, and a moving average. The big or large number is likable when widely in use and the value is not undulation in the series value. Many times, the moving series is used 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

Single Exponential Smoothing is a prediction method used 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 Double exponential smoothing is well fitted to trended data of the form.

Double Exponential Smoothing equation:

* = = forecast for p periods into the future
* = the exponentially smoothed series or current level estimate

regularly, we could set the first smoothed level equal to the first observation

* + *α* = smoothing constant for the level (0 < *α* < 1)
  + *β* = smoothing constant for trend estimate (0 < *β* < 1)
  + *p* = periods to be forecast into the future

5.3.4 Time Series Regression (Seasonal Without Trend)

Time series regression is a technique used for predicting a future response based on the recorded history and the transfer of dynamics from related predictors. Time series regression could give help so you could understand and predict the behavior of dynamic systems from experience or the data which has been recorded. Time series regression is 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 include the heteroscedasticity or autocorrelation effects.

5.3.5 ARIMA

ARIMA models are one of the techniques used for time series prediction. Illustrated smoothing and ARIMA models are the two most popular techniques used to predict the time series and add complementary approaches to the problem. During the exponential smoothing models are related to the description of the trend and seasonality data, The ARIMA models focus to explain 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 say 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.

# 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(x1)

 From the scatter plot of Systolic(y) and Age(x1) above, we can see that there are positively related in a linear sense.



Figure 7.2 Scatterplot of Systolic blood pressure(y) vs Weight(x2)

From the scatter plot of Systolic(y) and Weight(x2) 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: 𝜌𝑦𝑥1=0 (There is no linear relationship between the Age (x1) and Systolic Blood Pressure (y))

H1: 𝜌𝑦𝑥1≠0 (There is a linear relationship between 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 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: 𝜌𝑦𝑥2=0 (There is no linear relationship between the Weight (x2) and Systolic Blood Pressure (y))

H1: 𝜌𝑦𝑥2≠0 (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 a linear relationship between weight (x2) and Systolic blood pressure (y).

(3) H0: 𝜌𝑥1𝑥2=0 (There is no linear relationship between the Age (x1) and Weight (x2))

H1: 𝜌𝑥1𝑥2≠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 Age (x1) and Weight (x2). The p-value 0, which is lower than 𝑎 = 0.05 so we reject H0 and conclude that there is a 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  = 0.05, this means that the y-intercept is significant and this model can be used to predict.

Since this model can be able to predict let’s proceed to the next step.

Durbin-Watson Statistic

From Table of the Critical Values for the Durbin-Watson Statistic at ,

*n* = 15, and *k* = 2. We can get and .543

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

0 = 0.946 = 1.543 = 2.457  = 3.054

Since the first model has the Durbin-Watson of 1.82287 which is in the part of Accepting H0 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 α = 0.05, so H0 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

In the case of 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.

# Chapter 8: Conclusion

8.1) Multiple Regression: systolic blood pressure prediction

|  |  |  |  |
| --- | --- | --- | --- |
| Age x1 | Weight x2 | 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 variables x1 and x2 to predict systolic blood pressure (Actual Value).

# 

# 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 ,  It has a positive autocorrelation. | The test is inconclusive. | Accept: There is no autocorrelation. | The test is inconclusive. | Reject ,  It has a negative autocorrelation. |

0 = 0.946 = 1.543 = 2.457  = 3.054

This model is invalid due to the amount of Durbin-Watson value which is 0.895180 so it rejects the H0 and has a positive autocorrelation since the dL value is 0.946.

Model 4: Regression analysis: Systolic blood pressure Y versus AgeX1 (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

Model 4 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 X2

**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 = 0.05.

Model 6: Regression analysis: Systolic blood pressure Y versus Weight X­2 (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 ,  It has a positive autocorrelation. | The test is inconclusive. | Accept: There is no autocorrelation. | The test is inconclusive. | Reject ,  It has a negative autocorrelation. |

0 = 0.946 = 1.543 = 2.457  = 3.054

This model is invalid due to the amount of Durbin-Watson value which is 2.08396 so it rejects the H0 and has a positive autocorrelation since the dL value is 0.946.

References

1. <https://www.surveygizmo.com/resources/blog/regression-analysis/>
2. <https://medium.com/@JackieMolloy22/the-advantages-of-statistics-in-business-d7c7eda73333>
3. <https://www.quora.com/What-is-the-use-of-statistics-in-business>
4. <https://bizfluent.com/about-6360783-importance-statistics-industry-business.html>
5. <https://www.newgenapps.com/blog/business-applications-uses-regression-analysis-advantages/>
6. <https://www.emathzone.com/tutorials/basic-statistics/importance-of-statistics-in-different-fields.html>
7. <https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr02.html>