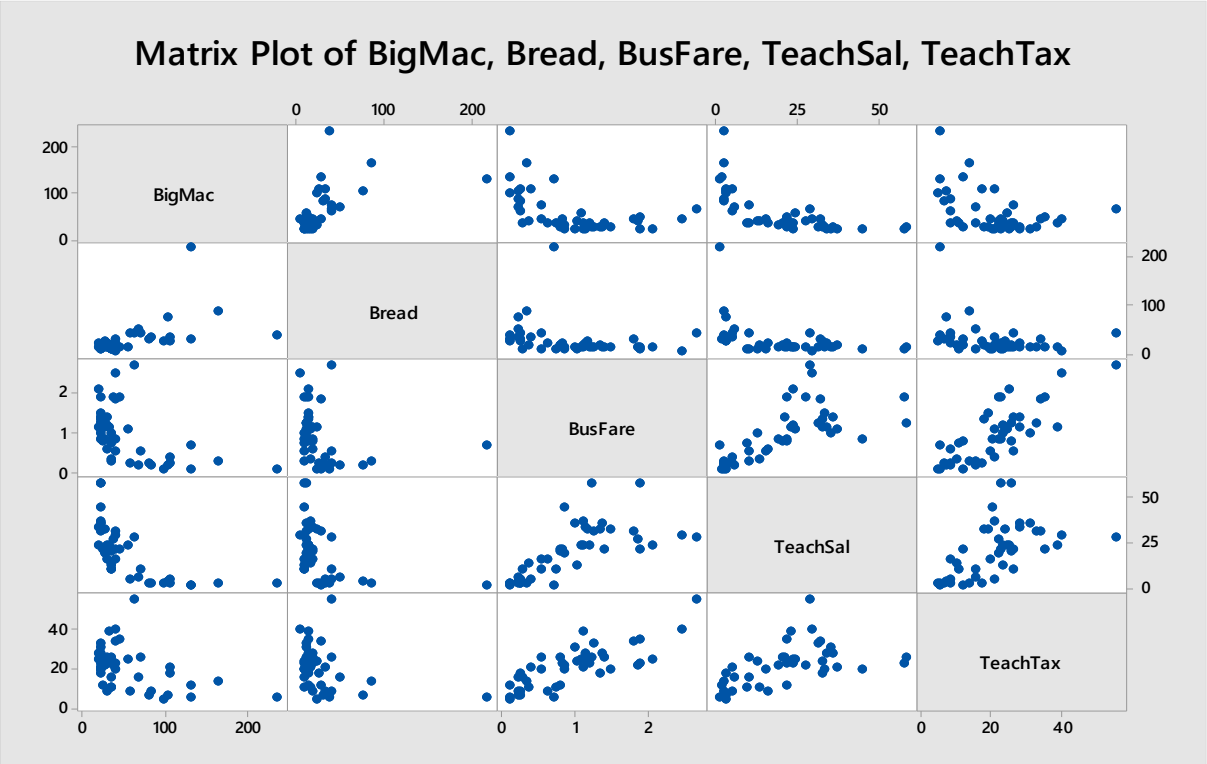
Applied Regression Modelling Assignment 2

Introduction: *Student Id: 17230755*

The Big Mac hamburger is a simple commodity which is identical throughout the world. Hence, we might expect that, the price of Big Mac will be same all around the world. However, the price of Big Mac is not the same and changes over various places. This may be due to the inefficiency in currency exchange.

The main aim of this study is to find how the cost of Big Mac hamburger varies over the places with economic indicators that describe each city.

Q1: Matrix Scatter plot showing relationship between Bread, BusFare, TeachSal, TeachTax.



Below are the variables which shows approximately **linear** relationship (Y-axis vs X-axis):

- TeachSal vs BusFare (Teacher salary increase with increase in bus fare)
- TeachSal vs TeachTax (Teacher salary increase with increase in tax)
- BusFare vs TeachTax (Bus Fare increases with increase in tax)

Below are the variables which shows **non - linear** relationship (Y-axis vs X-axis):

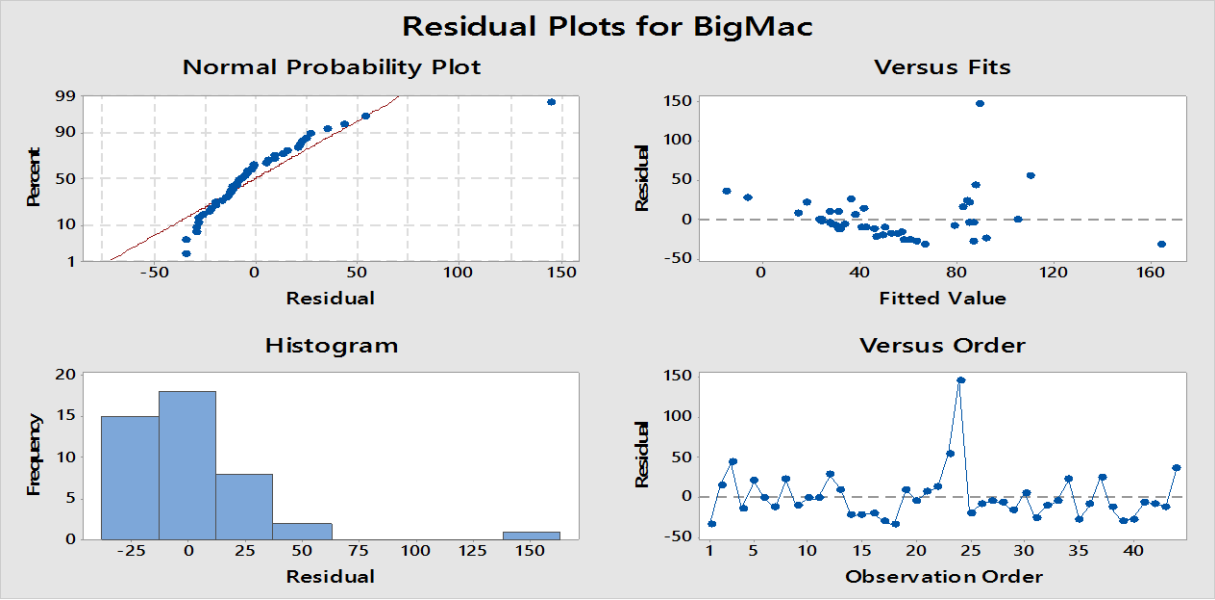
- Bread vs (BusFare or TeachTax or TeachSal) – Looks like Bread price remains same (small variations) irrespective any changes of the above-mentioned variables.

Cost of Big Mac and Tax Paid: Looks like Big Mac Cost decreases with increase in tax paid i.e. tax paid by teacher and engineer.

Price of bread and the primary teacher salary: From above matrix plot we can say that there is no linear relationship between the price of bread and salary. In addition, looks like price of bread remains constant irrespective of the salary of the teacher.

Salary earned, and the tax paid by a primary school teacher: From above matrix scatter plot, as tax increases the salary of teacher also increases. This may be as tax increases, teacher may demand more salary.

Q2: Regression Model using Bread price, BusFare, TeachSal, TeachTax as predictor(X):



Regression Analysis: BigMac versus Bread, BusFare, TeachSal, TeachTax

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	48440.0	12110.0	11.67	0.000
Bread	1	7653.2	7653.2	7.37	0.010
BusFare	1	732.2	732.2	0.71	0.406
TeachSal	1	7702.3	7702.3	7.42	0.010
TeachTax	1	74.8	74.8	0.07	0.790
Error	39	40475.1	1037.8		
Total	43	88915.2			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
32.2153	54.48%	49.81%	10.76%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	75.6	13.8	5.46	0.000	
Bread	0.449	0.165	2.72	0.010	1.30
BusFare	-11.9	14.2	-0.84	0.406	3.61
TeachSal	-1.329	0.488	-2.72	0.010	2.20
TeachTax	0.208	0.775	0.27	0.790	2.83

Regression Equation

BigMac = 75.6 + 0.449 Bread - 11.9 BusFare - 1.329 TeachSal + 0.208 TeachTax

Fits and Diagnostics for Unusual Observations

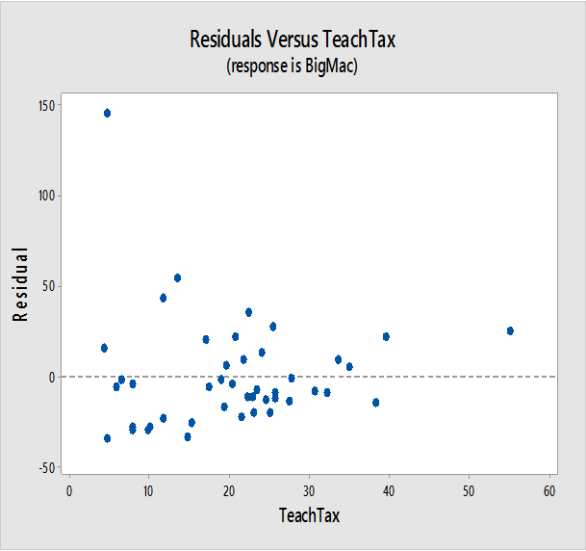
Obs	BigMac	Fit	Resid	Std Resid		
18	130.0	164.3	-34.3	-2.41	R	X
24	235.0	89.4	145.6	4.71	R	
37	61.0	35.9	25.1	0.98		X

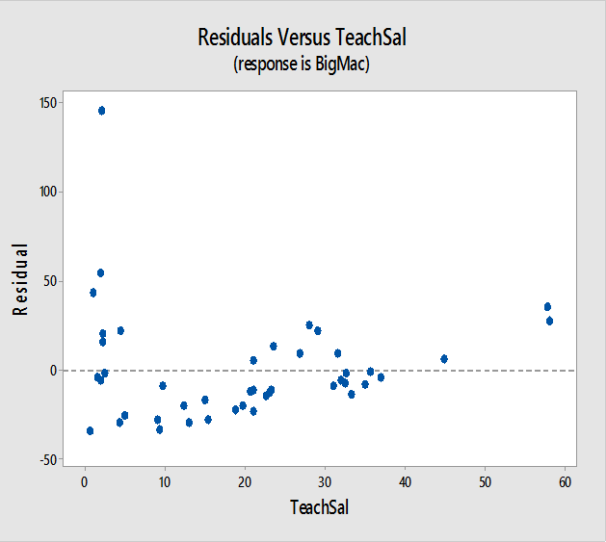
R Large residual
X Unusual X

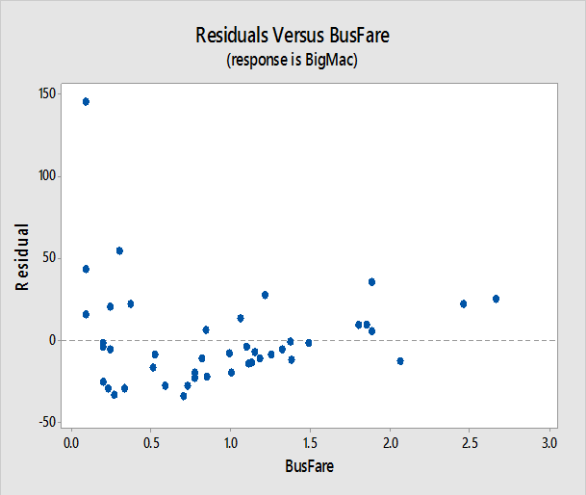
Observations:

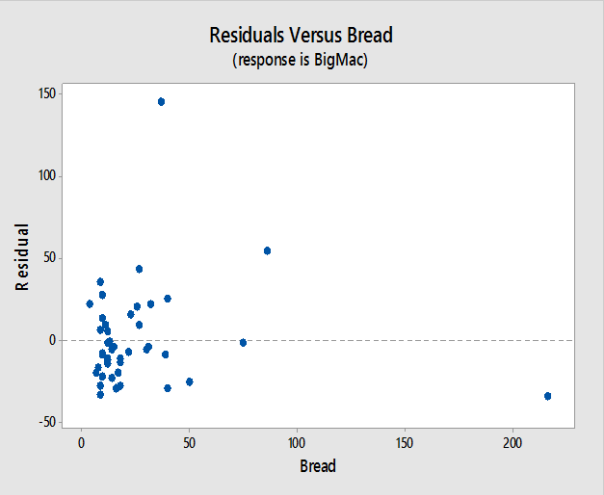
Normal Probability Plot: From the normality graph, we can see that few points do not lie in a straight line. This, indicates this is taken from normally distributed population with few outliers. If we trans form the data for normality we may achieve better linearity in response.

Residual Vs Fitted Value Plot: No pattern is observed in residual. Hence, we can say the error is random variable. Moreover, point with residual = 150, which is possible outlier.









The regression model is designed based on the assumption of the constant variance i.e. the error should be randomly distributed. There should be no pattern in the residual plot. If there are any observed pattern, then this crosschecked. Several transformations also used to achieve linearity, normality etc. By observing the residual plot of each predictor, we need to ensure that residuals are random and equally spread around the fitted line.

Observations:

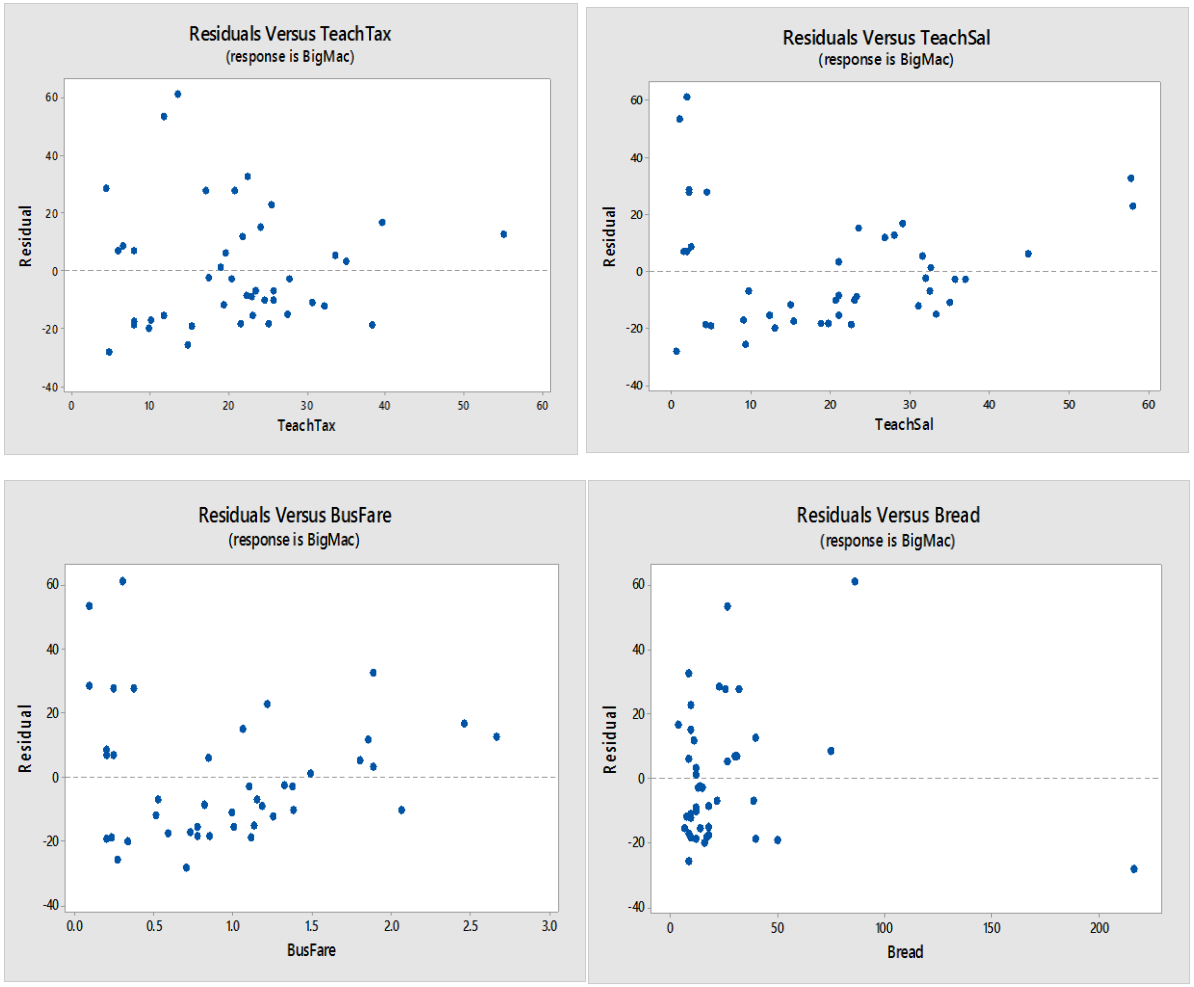
- From Graph 1: Possible outlier i.e. for residual = 150. Moreover, looks like rest all points are randomly scattered, hence there is no need of transformation for the variable TeachTax.
- From Graph 2 and 3 and 4: We can the observe approximate linear pattern in residual, hence this will affect the linearity in response. Hence, to make residual random as spread around the

fitted line, transformation is needed. TeachSal, BusFare needs transformation. Graph 4 is densely packed on one position, to make these points spread, better transform the variable **Bread** too.

Q3: <u>Presence of the outlier.</u>

Looks like **Mexico City** which has a response of 235 and the large residual of **149** from fitted line, this can be observed from the scatter plot and previously run regression analysis. Let us remover this entry and observe the effect of removing the outlier.

Examining the residual plot after removing outlier.



We can observe that; large residual is removed from the plot and **R squared** value also improved to **68.51%.**

Regression Analysis: BigMac versus Bread, BusFare, TeachSal, TeachTax

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	37894.8	9473.7	20.66	0.000
Bread	1	9095.1	9095.1	19.84	0.000
BusFare	1	822.1	822.1	1.79	0.188
TeachSal	1	5774.4	5774.4	12.60	0.001
TeachTax	1	730.5	730.5	1.59	0.215
Error	38	17421.6	458.5		
Total	42	55316.5			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
21.4118	68.51%	65.19%	23.10%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	59.07	9.50	6.22	0.000	
Bread	0.490	0.110	4.45	0.000	1.30
BusFare	-12.64	9.44	-1.34	0.188	3.47
TeachSal	-1.154	0.325	-3.55	0.001	2.13
TeachTax	0.655	0.519	1.26	0.215	2.73

Regression Equation

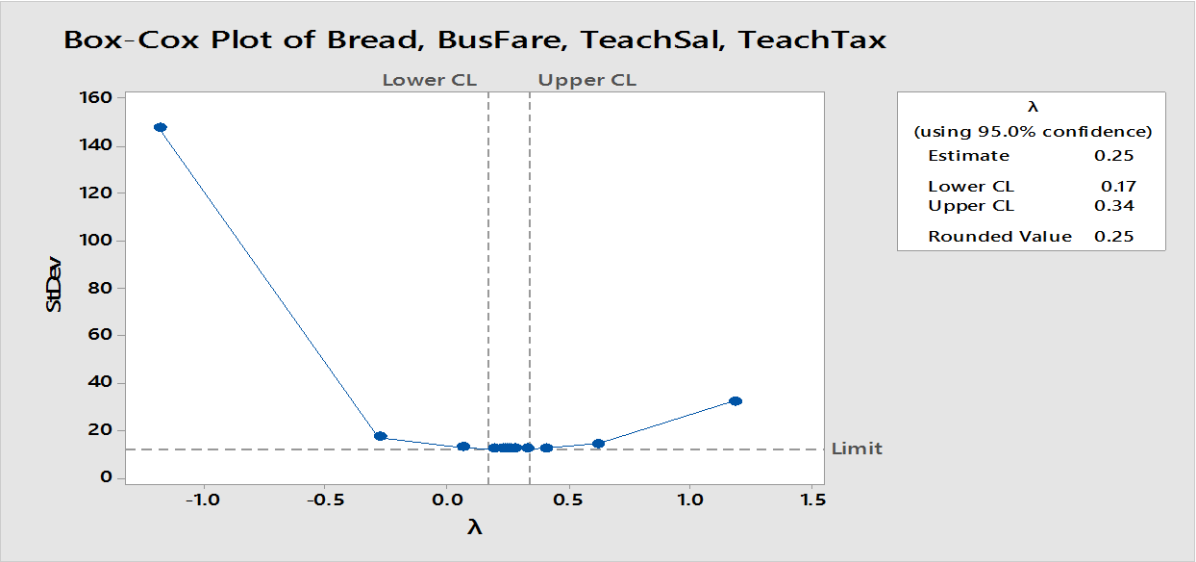
BigMac = 59.07 + 0.490 Bread - 12.64 BusFare - 1.154 TeachSal + 0.655 TeachTax

Fits and Diagnostics for Unusual Observations

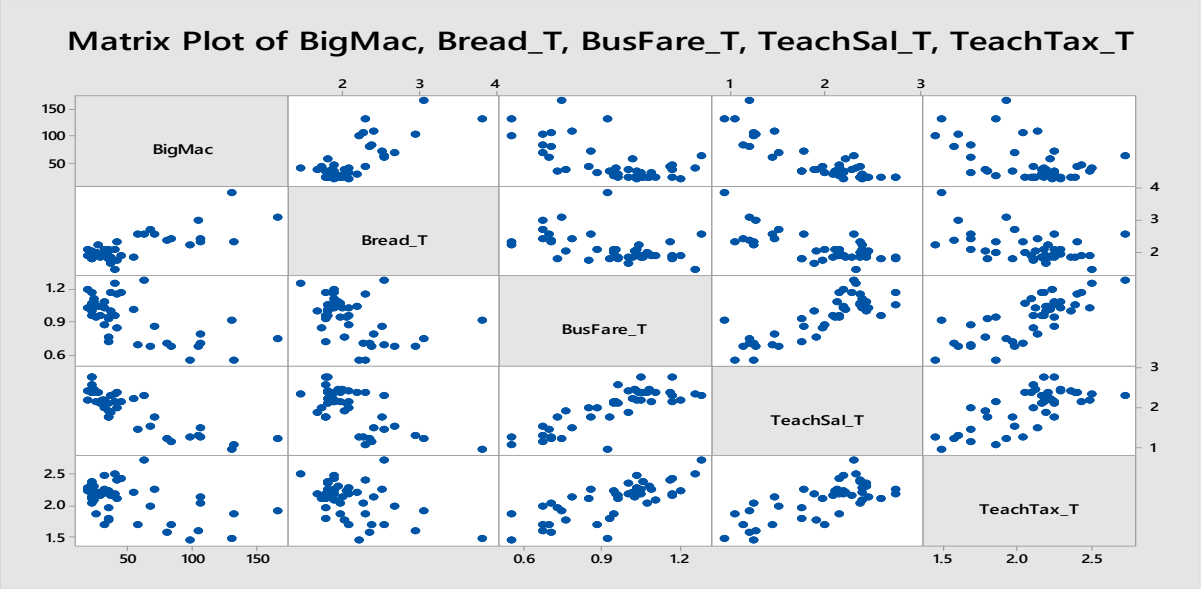
Obs	BigMac	Fit	Resid	Std Resid		
3	131.0	77.5	53.5	2.60	R	
18	130.0	158.3	-28.3	-3.00	R	X
23	165.0	103.9	61.1	3.03	R	
36	61.0	48.6	12.4	0.73		X

R Large residual
X Unusual X

Q4: log transformation:

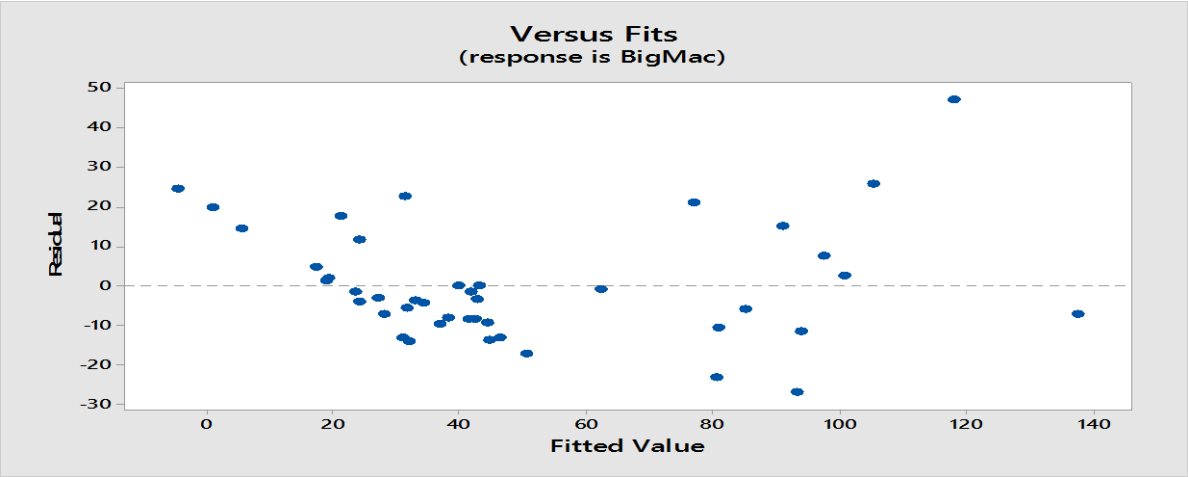


Scatter Plot after transformation:



From the above graph, linearity between Big Mac and transformed variables are improved after transformation.

Constant Variance:



From these we can say that error is scattered randomly around fitted line. Hence, possibility of variance being constant is more and better than before.

Regression Analysis: BigMac versus Bread_T, BusFare_T , TeachTax_T

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	46300.9	11575.2	48.79	0.000
Bread_T	1	3426.0	3426.0	14.44	0.001
BusFare_T	1	83.5	83.5	0.35	0.556
TeachSal_T	1	6417.0	6417.0	27.05	0.000
TeachTax_T	1	1857.9	1857.9	7.83	0.008
Error	38	9015.6	237.3		
Total	42	55316.5			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
15.4030	83.70%	81.99%	77.94%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	36.0	30.3	1.19	0.241	
Bread_T	29.27	7.70	3.80	0.001	2.04
BusFare_T	-15.1	25.4	-0.59	0.556	4.14
TeachSal_T	-55.6	10.7	-5.20	0.000	5.03
TeachTax_T	36.6	13.1	2.80	0.008	2.58

Regression Equation

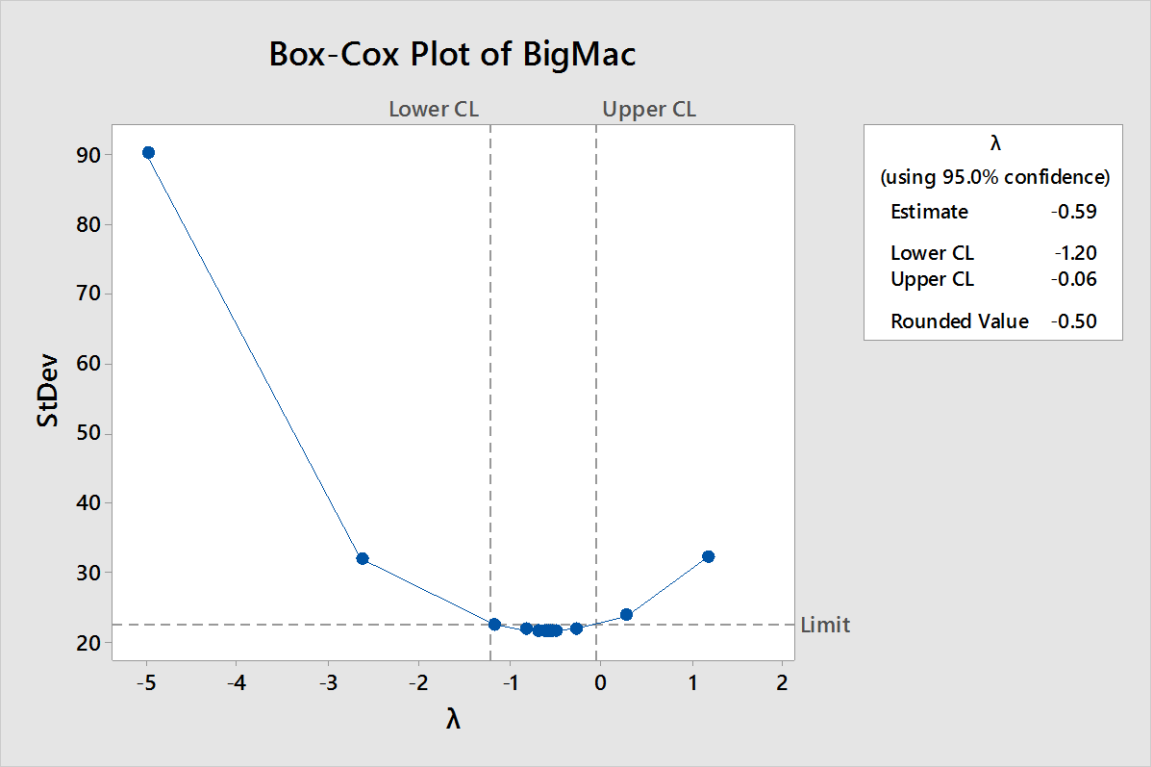
BigMac = 36.0 + 29.27 Bread_T - 15.1 BusFare_T - 55.6 TeachSal_T + 36.6 TeachTax_T

Fits and Diagnostics for Unusual Observations

Obs	BigMac	Fit	Resid	Std Resid	
18	130.00	137.26	-7.26	-0.73	X
23	165.00	117.91	47.09	3.32	R

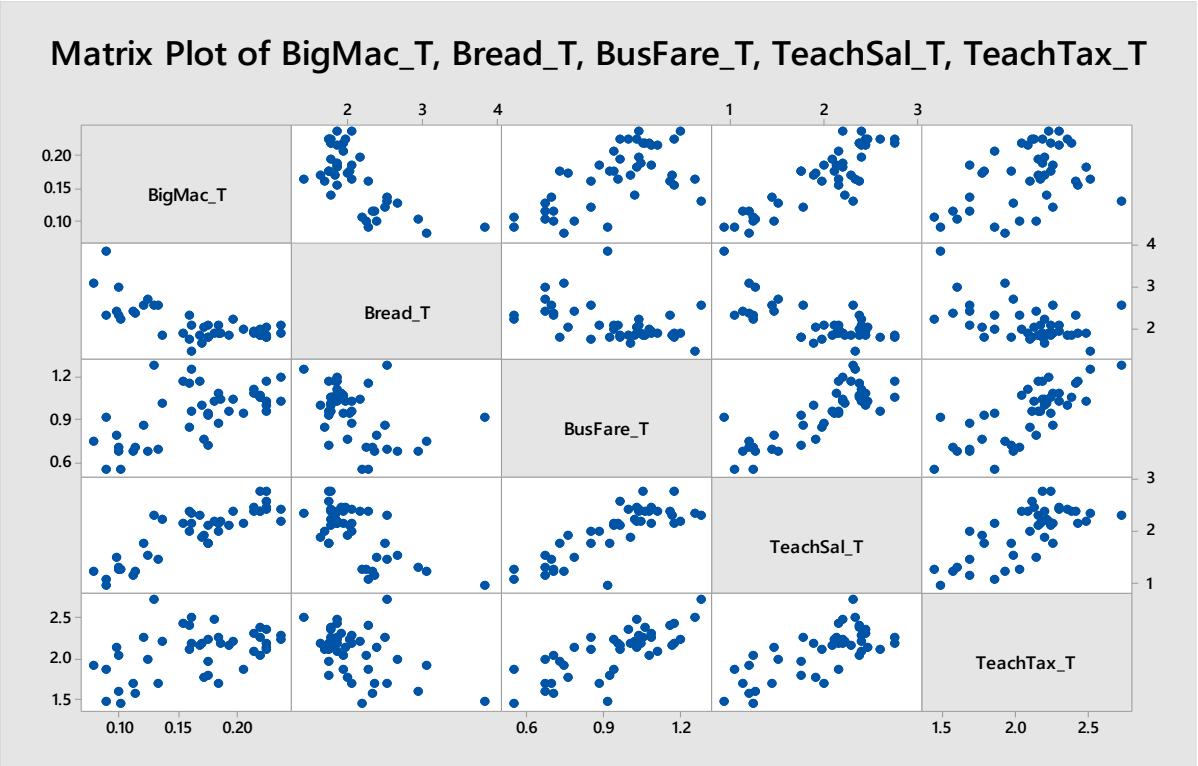
R Large residual
X Unusual X

Q5: Transformation of response variable:



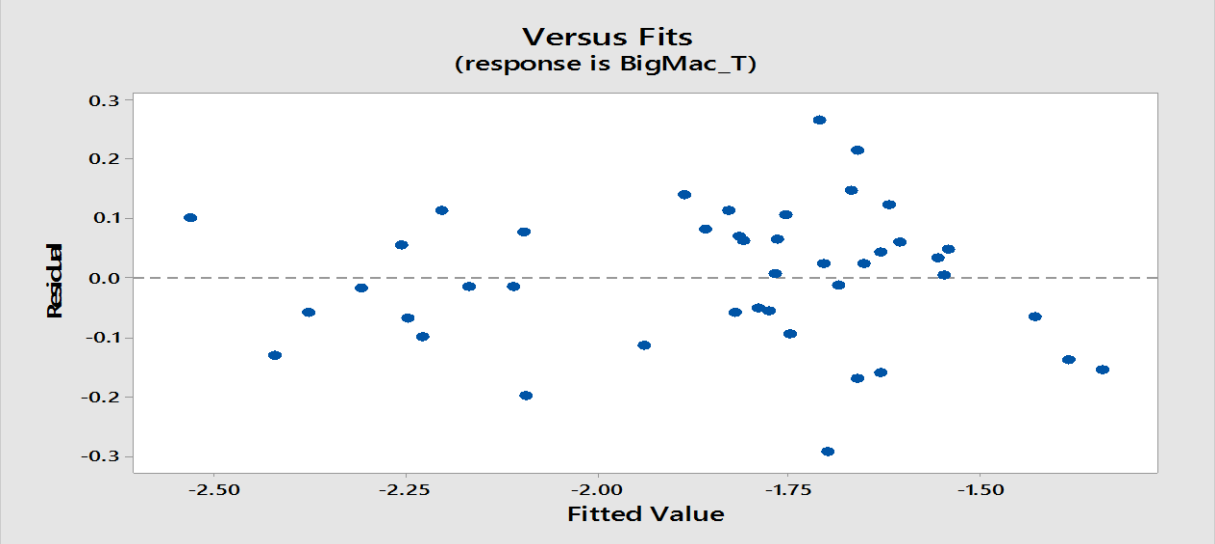
Lambda value is -0.59. However, since it is negative and for easy to interpret we take lambda = 0 and run the regression analysis on transformed response.

Q6: <u>Scatter plot matrix for transformed variables:</u>



The relationship between transformed response and transformed predictor looks slightly better linear than before.

Constant variance: This can be explained better in terms of residual vs fitted value plot which is obtained after running the regression analysis for lambda = 0.



From the above graph, and matrix scatter plot we can say that error is randomly distributed, and relationship is more linear. For constant variance error should be randomly distributed. $E[e_i] = 0$ and $\{e_i\}$ is independent an identically distributed.

Q7: Regression Model with transformed response and transformed predictor:

Box-Cox transformation λ = 0

Analysis of Variance for Transformed Response

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	3.59244	0.898110	61.65	0.000
Bread_T	1	0.11734	0.117341	8.05	0.007
BusFare_T	1	0.00266	0.002660	0.18	0.672
TeachSal_T	1	0.71724	0.717238	49.23	0.000
TeachTax_T	1	0.20208	0.202075	13.87	0.001
Error	38	0.55358	0.014568		
Total	42	4.14601			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.120697	86.65%	85.24%	81.92%

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.932	0.237	-8.15	0.000	
Bread_T	-0.1713	0.0604	-2.84	0.007	2.04
BusFare_T	0.085	0.199	0.43	0.672	4.14
TeachSal_T	0.5878	0.0838	7.02	0.000	5.03
TeachTax_T	-0.382	0.103	-3.72	0.001	2.58

Regression Equation

ln(BigMac_T) = -1.932 - 0.1713 Bread_T + 0.085 BusFare_T + 0.5878 TeachSal_T - 0.382 TeachTax_T

Fits and Diagnostics for Unusual Observations

Original Response

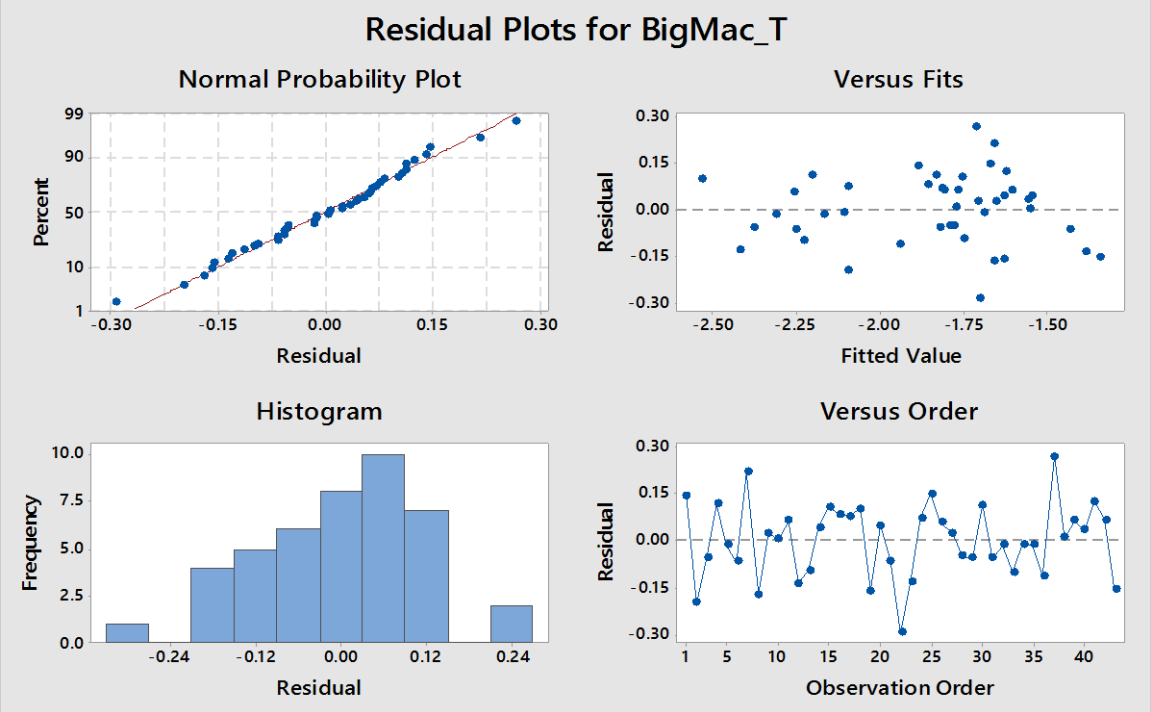
Obs	BigMac_T	Fit
18	0.08771	0.07933
22	0.13608	0.18249
37	0.23570	0.18070

Fits and Diagnostics for Unusual Observations

Transformed Response

Obs	BigMac_T'	Fit	Resid	Std Resid	
18	-2.4338	-2.5342	0.1004	1.29	X
22	-1.9945	-1.7011	-0.2934	-2.48	R
37	-1.4452	-1.7109	0.2657	2.35	R

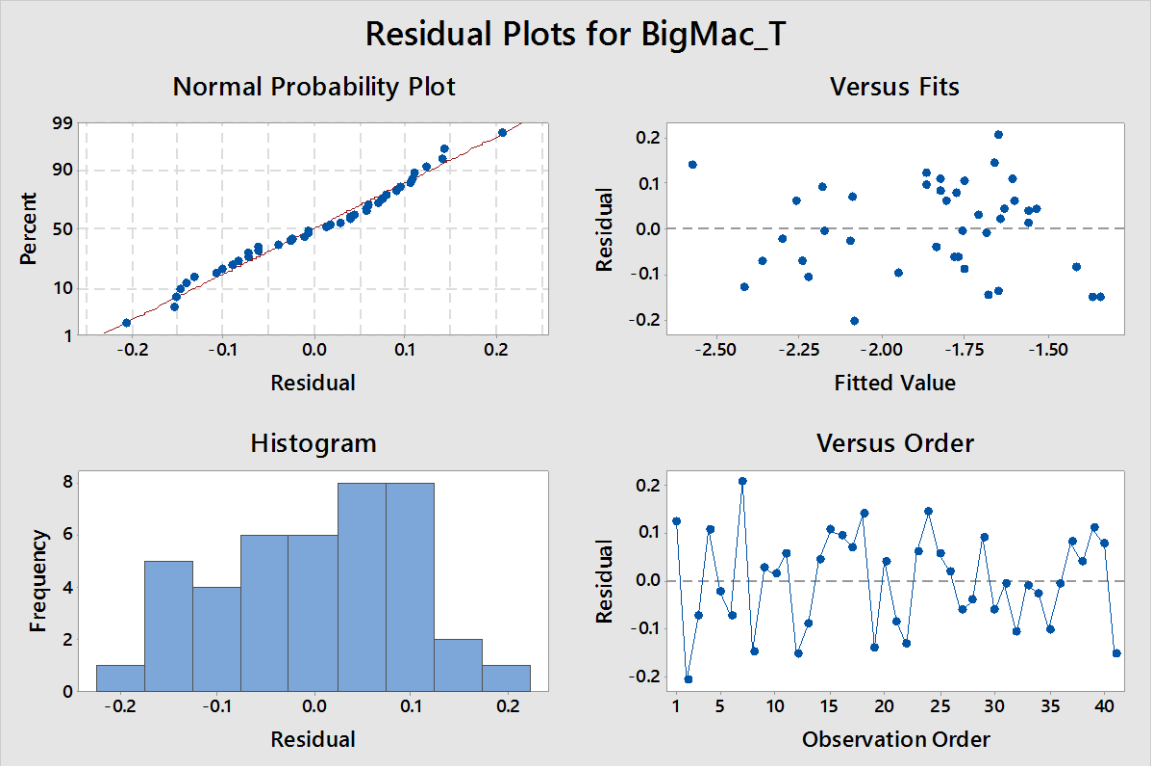
BigMac_T' = transformed response
R Large residual
X Unusual X



<u>Cities with largest residual:</u> Madrid, Sydney

Madrid:	22	-1.9945	-1.7011	-0.2934	-2.48	R
Sydney:	37	-1.4452	-1.7109	0.2657	2.35	R

Removing cities Madrid and Sydney and fitting regression model.



Box-Cox transformation λ = 0

Analysis of Variance for Transformed Response

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	3.57138	0.892846	82.37	0.000
Bread_T	1	0.10416	0.104165	9.61	0.004
BusFare_T	1	0.00098	0.000984	0.09	0.765
TeachSal_T	1	0.77089	0.770889	71.12	0.000
TeachTax_T	1	0.16990	0.169902	15.67	0.000
Error	36	0.39023	0.010840		
Total	40	3.96162			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.104115	90.15%	89.06%	84.59%

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.950	0.205	-9.49	0.000	
Bread_T	-0.1637	0.0528	-3.10	0.004	2.06
BusFare_T	-0.055	0.181	-0.30	0.765	4.36
TeachSal_T	0.6240	0.0740	8.43	0.000	5.22
TeachTax_T	-0.3525	0.0890	-3.96	0.000	2.58

Regression Equation

ln(BigMac_T) = -1.950 - 0.1637 Bread_T - 0.055 BusFare_T + 0.6240 TeachSal_T - 0.3525 TeachTax_T

Fits and Diagnostics for Unusual Observations

Original Response

Obs	BigMac_T	Fit
2	0.10102	0.12428
7	0.23570	0.19166
18	0.08771	0.07627

Fits and Diagnostics for Unusual Observations

Transformed Response

Obs	BigMac_T'	Fit	Resid	Std Resid		
2	-2.2925	-2.0852	-0.2073	-2.19	R	
7	-1.4452	-1.6520	0.2068	2.06	R	
18	-2.4338	-2.5735	0.1398	2.12	R	X

Observations:

- R squared an adjusted r squared value has been improved.
- R squared adjusted before: 85.24%
- R squared adjusted new: 89%
- Moreover, frequency vs residual plot is also near to normal
- The price of big mac of Sydney and Madrid have a considerable effect on fitted response. However, it does not have strong influence on fitted response. The reason for this is improved adjusted r square is only 4%.

Q8: Statistical test interpretation about how the response Big Mac depends on predictor variable.

Bread_T	1	0.10416	0.104165	9.61	0.004
BusFare_T	1	0.00098	0.000984	0.09	0.765
TeachSal_T	1	0.77089	0.770889	71.12	0.000
TeachTax_T	1	0.16990	0.169902	15.67	0.000

For 95% CI alpha value is 0.05.

If p value is less than alpha reject null hypothesis. Here p value for BusFare_T > alpha, which indicates there is no enough evidence that BusFare_T is significant. However, rest all other predictor variable Breard, Bus, TeachSal, TeachTax are all significant i.e. its p value is less than 0.05. The order of significant predictor is as below:

- TeachTax and TeachSal
- Bread
- BusFare – no evidence of significance to the response variable.

Discussion and Conclusion:

We can conclude the price of Big Mac in various cities depends on Tax Paid, Salary, and Bread cost in different cities. In this case I have removed outliers considering it as bad data. However, we need to investigate thoroughly before making any decision of removing outlier. It may happen that sometime the outlier indicates some hidden feature about the data. Hence careful evaluation outlier or large residuals is necessary and find out the reason behind it.

Future scope: We need to evaluate the response of price on another predictor for e.g. EngSal, EngTax etc. and find the best predictors which describe variation of response better if there is any better predictor present.

Reference:

[1] Interpreting p value: https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/basics/example-of-getting-and-interpreting-a-p-value/

[2] Data Transformation: http://blog.minitab.com/blog/statistics-and-quality-improvement/see-how-easily-you-can-do-a-box-cox-transformation-in-regression

[3] Residual: http://blog.minitab.com/blog/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses