International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

Analysis of Industrial Production Data-CVPP Regression Model Approach

K. Sivakumar
Professor
Department of Mathematics,
Sathyabama Institute of Science and Technology,
Chennai, Tamilnadu, India.

M. Nirmala
Associate Professor,
Department of Mathematics,
Sathyabama Institute of Science and Technology,
Chennai,Tamilnadu, India.

Abstract **- The voluminous measure of information being gathered by different national and state associations those are uninhibitedly accessible with the end goal of factual examination. Be that as it may, the information has not been fundamentally investigated and inspected for its effect on national economy. It is an inevitable the information must be completely inspected utilizing all around laid factual devices and strategies to have further understanding on the information like location of anomalies, significance of factors and regular varieties and so on. In this article the factual examination for registering different modern generation exhibitions dependent on their noteworthy distinction and assessing through Cross Validity Predictive Power (CVPP) model of relapse approach is dissected to get development and assembling process. Number of exceptions is recognized and potential purposes behind the equivalent are talked about.**

Keywords-Industrial Production data, linear regression, Outliers, Cross Validity Predictive Power Model

I. INTRODUCTION

Indian Industry at display advancing at a fast pace yet observed from recorded viewpoint it is recapturing a situation. The adverse impact of colonial rule has been seen as uniform and linear, so that at least three distinct phases to this impact have been observed. Industry analysts predict that with this momentum of growth the country will be the third largest economy by 2040. The purpose of regression modeling is to find out how the conditional distribution of the response depends on the explanatory variables. In usual approach, the conditional distribution is derived from the distribution assumption and predicted mean. An usual method is to model the independent variables for the data and to approximate the conditional distributions. A large amount of data has been analyzed from industrial production process. The data contain information that is useful in controlling the production process. The problem is to extract information from the data. This has commonly been done by developing statistical prediction models which are then utilized in process control, process planning and product planning (Khattree and Rao, 2003).

A model can be defined as a statistical description of the data that describes the variables their relationships, parameters, assumptions, constraints and the underlying distributions. In a review of industrialization in developing economics, Pack (1988) has drawn attention to the contrasting experience of the developing economies in this respect. Statistical analysis helps to understand the basis and differences in various industrial productions while prediction is of use for the constructed model to quantify the contribution of the underlying factors (Cooper and Richard 1999). An approximate formula for obtaining the levels of significant for the statistical data discussed (Childs et al 2006). Draper and Smith (1998) explained for a general survey regression analysis which was the statistical methodology for predicting values of one or more response variables from a collection of predictor variable values.

A measure of the degree of linearity and its square is obtained by Pearson's correlation then the coefficient of determination R^2 measures the amount of variability in Y explained by the predictor variable. The remaining variability is measured by the residuals. Thus any systematic pattern revealed in a plot of the residuals indicates departure from linearity. A virtue of the normal probability plot is that comparing the cumulative distribution of actual data values with the cumulative distribution of a normal distribution. By checking linearity the normal distribution forms a straight diagonal line. Errors are the residuals or difference between the actual score for a case and the score estimated by the regression equation. For the observation Y_i corresponding to predictors X_{i1}, X_{i2},……….X_{in} that can be calculated through $Y_i = b_0 + b_1X_{i1} + b_2X_{i2} +\ldots\ldots\ldots\ldots.+b_kX_{in}$. The difference $e_i = y_i - \hat{y}$ are called residuals for i = 1, 2, ……n. These residuals were used to diagnose the validity of the model. The diagnostic plot for multiple regressions is a scatter plot of the residuals e_i against the predicted values Y_i. Such plot may be used to see if the predictions can be improved by identifying outliers.

II. OUTLIER DETECTION METHODS

An outlier is authentic as 'an ascertainment whose amount is not in the arrangement of ethics produced by the blow of the data' (Daniel 1960). Han and Kamber (2000) authentic as outliers are those credibility which are altered from or inconsistent with the blow of the data. Literature is abounding with procedures for the apprehension and testing for outliers in sample data. Balasooriya et al(1987) did an experimental examination to distinguish the best of seven usually utilized techniques for recognizing anomalies and compelling perceptions in straight relapse models dependent on a few informational indexes. Based on study, it was seen that the techniques don't generally concur and recommended a prudent blend of systems. Their exact examinations likewise uncovered that the outcomes will in general firmly differ on account of numerous exceptions. Outlier is a case with such an extreme value on one variable (a univariate outlier) or such a strange combination of scores (a multivariate outlier) that they distort statistics. Outliers can affect the mean and the variance of a univariate distribution. Outlier detection is useful in various industrial productions for finding unusual responses to various productions. It is useful to the investigator for anomalies to be assembled with the goal that comparative exceptions can be dissected together. Another crucial scientific categorization of exception discovery strategies is between parametric (measurable) techniques and non parametric strategies that are sans show (Williams et al 2002). Different categorizations of anomaly location techniques can be found in (Barnett and Lewis 1994, Papadimitriou et al 2002, Acuna and Rodriguez 2004).

III. METHODOLOGY

Cross Validity Prediction Power (CVPP), p^2_{cv} was applied to test out the adequacy or soundness of the fitted model. Cross Validity Predictive Power (CVPP), ρ^2_{cv} is a well known model validation technique (Stevens, 1996 and Ali, 2000). The value $\rho^2 = 0$ ($\rho^2 = 1$) indicates that the fitted model has no validity (100% validity). Thus, the expected value of ρ^2 should lie within the interval (0,1). It is also a function of the coefficient of multiple determinations R for $R^2 = 0$ implies CVPP is negative. It was suggested that Cross Validity Predictive Model was also employed as model validation technique (Islam 2003).

The CVPP p^2_{cv} is defined as Stevens (1996) $p^2_{cv} = 1–W (1–R^2)$, where W= $\dfrac{(n-1)(n-2)(n+1)}{n(n-k-1)(n-k-2)} \; 0 \le R^2 \le 1$, n is the number of observations, k is the number of regressors, R^2 is the coefficient of multiple determination and the suffix cv of r indicates cross validation. The stability (n) of a fitted model examines the ability of fitting performance of the model over the industrial production data considered in the research study and is computed as $\eta = R^2 - p^2_{cv}$. Additionally, 1-shrinkage is the stability of R^2 of the model.

IV. DATA ANALYSIS

The source of the data for the analysis is the Central Statistical Organization (CSO) of Monthly Abstract of Industrial Statistics Database. This dataset provides analysis for four industrial productions of Cosmetics and Toiletries, Beverages and Tobacco, Leather and Cotton Textiles for the period over 2006 to 2015 on monthly basis. Linear regression technique was applied for fitting the linear model. The first step is to employ a number of variables to be identified that could explain a 'significant' portion for each industrial production, which are used for analysis and validated using Cross Validity Predictive Power Model (CVPP). The data was quite extensive with each of the industrial product has numerous auxiliary variables. However, the research study was restricted to two variables identified based on its economic importance and other related model aspects. It was observed that R^2 value for the remaining set of auxiliary variables was too small for consideration.

The approach are dealt in three ways

(i) A linear regression model was fitted for the years 2006 to 2015 based on production.

(ii) Similar analysis has been performed for annual total production of each industrial product.

(iii) Univariate Mahalonobis outlier detection technique was applied for the analysis that discriminated the statistical variables across the years. As described under, it is a useful way of determining similarity of an unknown sample set to a known one. As to exception location, the outlier detection, a standardized residual was plotted and the results were analyzed based on the range of the residuals lying within ±3.0 in absolute value.

Tables I and II provide the least square regression estimation that are resulted from the impact of technological knowledge through various factors on the productivity of the industrial data. Bold numbers in the tables indicate the p value significant at 5% level.

Table I. Estimates of parameters in linear regression model for Cosmetics and Toiletries, Beverages and Tobacco

	Cosmetics and Toiletries								Beverages and Tobacco							
	Soap all kinds(SSI)				Agarbathi				Country Liquor				Indian made foreign liquor			
Year	**R^2**	**P Value**	**Intercept A**	**Slope b**	**R^2**	**P Value**	**Intercept a**	**Slope b**	**R^2**	**P Value**	**Intercept a**	**Slope b**	**R^2**	**P Value**	**Intercept a**	**Slope b**
2006	0.21	0.13	93.88	0.32	0.29	0.16	98.67	1.58	0.67	**0.05**	-256983	3.20	0.68	**0.03**	-111588	3.20
			0	0.13			0	0.16			-0.05	-0.03			-0.04	-0.03
2007	0.85	**0**	-4861.15	0.14	0.31	0.82	101.17	-0.16	0.08	0.65	25654.56	-0.45	0.04	0.55	19406.74	-0.45
			0	0			0	0.82			-0.65	-0.55			-0.48	-0.55
2008	0.81	**0**	179.56	9.35	0.29	0.35	88.50	-1.54	0.58	**0.01**	254698	5.27	0.54	**0.02**	-186761	5.27
			0	0			0	0.35			-0.01	-0.01			-0.01	-0.01
2009	0.16	0.20	256.17	0.71	0.25	0.71	82	0.16	0.65	**0.01**	-254698	4.78	0.55	**0.03**	-170178	4.78
			0	0.20			0	0.71			-0.05	-0.01			-0.01	-0.01
2010	0.25	0.10	265.55	1.18	0.78	**0**	257.56	28.64	0.89	**0**	-480667	12.32	0.77	**0**	-480667	12.32
			0	0.10			0	0			0	0			0	0
2011	0.62	**0**	292.29	1.45	0.88	**0**	400.73	16.88	0.54	0.07	-448796	5.97	0.31	0.06	-306254	7.96
			0	0			0	0			-0.05	-0.04			-0.07	-0.06
2012	0.64	**0.02**	299.33	1.20	0.47	0.08	386.10	-3.23	0.25	0.44	-55243.90	6.70	0.06	0.44	254596.90	1.70
			0	0.02			0	0.08			-0.52	-0.44			-0.52	0.56
2013	0.10	0.31	303.92	0.08	0.81	**0**	409.08	7.00	0.53	**0.01**	562541.80	-6.93	0.53	**0.01**	286349.80	-6.93
			0	0.31			0	0			-0.01	-0.01			-0.01	-0.01
2014	0.95	**0**	304.19	0.35	0.31	0.73	417.25	0.50	0.54	0.09	-254873	5.86	0.63	**0**	-179754	4.83
			0	0			0	0.73			0	0			0	0
2015	0.97	**0**	308.06	0.31	0.12	0	1.14	1.14	0	0.52	0.02	-0.07	0	0.89	16063.89	0.08
			0	0			0.27	0.27			-0.25	-0.79			-0.51	-0.89
Total	0.81	**0**	-588764	295.61	0.83	**0**	-113081	566.58	0.94	**0**	-310000	15486.60	0.88	**0**	-500000	25185.38
			0	0			0	0			0	0			0	0

Table II. Estimates of parameters in linear regression model for Leather, Cotton Textiles

Leather									Cotton Textiles							
	Indian foot wear				Shoe Upper				Cotton Yarn				Cotton Hoisery			
Year	**R^2**	**P Value**	**Intercept a**	**Slope b**	**R^2**	**P Value**	**Intercept a**	**Slope b**	**R^2**	**P Value**	**Intercept a**	**Slope b**	**R^2**	**P Value**	**Intercept a**	**Slope b**
2006	0.84	**0**	10889.42	247.28	0.07	0.41	393.33	3.15	0.90	**0**	45987	7.37	0.46	0.99	56659	9.25
			0	0			0	0.41			-0.05	-0.01			-0.25	0
2007	0.53	**0.01**	14213.42	14213.42	0.15	0.21	441.67	4.62	0.03	0.79	54978.23	5.32	0.79	**0**	65879	7.33
			0	0.01			0	0.21			-0.42	-0.03			0	0
2008	0.87	**0**	12117.75	-694.42	0.68	**0.03**	559.75	17.98	0.80	**0**	54879	-6.21	0.49	0.79	32564	5.32
			0	0			0	0.03			-0.05	-0.25			0	-0.03
2009	0.31	0.75	11834	-19.37	0.56	**0.02**	551.83	10.83	0.79	**0**	-47895	8.37	0.89	**0**	25647	5.32
			0	0.75			0	0.02			0	-0.05			-0.12	-0.01
2010	0.43	7.59	12679.38	200.58	0	0.95	666	0.65	0.49	0.65	58794	-26.33	0.88	**0**	-45623	33.37
			0	0.02			0	0.95			-0.05	0			-0.01	0
2011	0.11	0.30	13504.64	17.83	0.03	0.60	544.50	5.10	0.46	0.08	25487	10.24	0.89	**0**	35.32	23.32
			0	0.30			0	0.60			-0.07	0			0	-0.01
2012	0.73	**0**	13932.28	0	0.11	0.29	571.67	6.76	0.80	**0.01**	-58794	1.55	0.26	0.99	-12665	7.33
			0	0			0	0.29			-0.05	-0.03			0	0
2013	0.80	**0**	14557.26	56.10	0.60	**0**	574.33	19.80	0.69	**0**	256487	4.37	0.78	**0**	33541	10.02
			0	0			0	0			-0.01	-0.03			-0.03	0
2014	0.51	0.21	15088.92	28.42	0.02	0.65	681.75	0	0.85	**0**	-78952	5.36	0.34	0.87	-78952	23.33
			0	0.01			0	0.65			0	0			0	0
2015	0.04	0.52	15688.99	17.84	0.22	0.12	751.25	12.03	0.24	0.90	254863	2.37	0.33	1	32564	5.36
			0	0.52			0	0.12			-0.05	0			-0.05	0
Total	0.67	**0**	830321	4992.12	0.71	**0**	-708234	357.29	0.77	**0**	5987.32	325464	0.90	**0**	65987	283654
			0	0			0	0			0	0			0	0

V. RESULTS AND DISCUSSION

On account of Soap of numerous types (SSI), it was seen that the proposed variable quantifiably immense for the years 2007, 2008, 2011, 2012, 2014 and 2015 with positive slopes and no huge for the years 2006, 2009, 2010 and 2011. Further, year wise production was compared with overall years production which showed that significant of the variation in Soap all kinds (SSI) explained by the total production. The coefficient value of b 295.61 which indicates positive slope revealed that one unit positive tones of production in one year, will increase 295.61 tones of Soap all Kinds (SSI) for the overall years of production. On the study of Agarbathi, the variable has statistically significant in the years 2010, 2011, 2013 with positive slopes however not in the years 2006, 2007, 2008, 2009, 2012, 2014, 2015. For over all data, coefficient value of b 566.58 which projected positive slope with one unit positive tones of Agarbathi production in one year, will expand 566.58 tones of Agarbathi for the overall years of production. The item Country Liquor, the variable has measurably huge for the years 1997, 1999, 2000, 2001 with positive inclines and the year 2013 with negative inclination. The variable does not demonstrate any measurably noteworthy for the years 2007, 2011, 2012, 2014 and 2015. A significant variation was observed for the production Country liquor (R^2=0.94) for overall data with coefficient value of b was 15486.60. The variable does not demonstrate any measurably noteworthy for the years 2007, 2011, 2012, 2014 and 2015. On account of Indian made foreign liquor, the variable was statistically significant for the years 2006, 2008, 2009, 2010, 2014 with positive slopes and negative slope for the year 2013. The variable does not show any statistically significant for the year 2007, 2011, 2012 and 2015. Further, it was compared with sum of annual production of overall years. A significant variation was observed in the overall for the Indian made foreign liquor (R^2=0.88). The coefficient value of b was 25185.38 which indicates positive slope with one unit increase 25,185.38 of Indian made foreign liquor (kls.) in one year, will increase 25,185.38 kilolitres of Indian made foreign liquor for the overall years of production. At the point when Indian footwear was considered, the variable has factual huge in the years 2006, 2007, 2012 and 2013 with positive slants and 2008 with negative slant. The variable does not demonstrate any measurably critical for the year 2009, 2010, 2011, 2014 and 2015. It was additionally contrasted and total of in general year generation which demonstrated that critical of the variety in Indian footwear clarified by the complete creation. Likewise coefficient estimation of b indicated as 4992.12 which shown positive slope with one unit positive tones of production, in one year will increase 4992.12 tones of Indian footwear for the overall years of production. In the case of Shoe upper, the years 2008, 2009 and 2013 have shown significant but not for the years 2006, 2007, 2010, 2011, 2012, 2014 and 2015. Also, sum of overall year production which showed that significant of the variation with coefficient value of b was 357.29 which indicated positive slope. In managing Cotton yarn, the variable has factually huge with positive slants in the years 1997, 2000, 2003, 2004 and 2005 and measurably noteworthy with negative incline for the period 2008 yet not for the years 2007, 2010, 2011, and 2015. Additionally coefficient estimation of b for by and large was 325464 which shown positive slant with one unit positive tones of Cotton yarn creation in one year will expand 325464 tones for the general long periods of generation. The variable has measurable noteworthy for Cotton hosiery, with positive inclines in the years 2007, 2009, 2010, 2011 and 2013 however not for the rest of the years with all out year savvy coefficient estimation of b was 283654 which shown positive slant with one unit positive tones of cotton hosiery creation in one year will build 283654 tones for the general long stretches of generation. Then, it is necessary to perform diagnostic checking using detection of Mahalonobis outliers. Further, model validation technique (Stevens 1996 and Khan and Ali 2003) applied over the study and thereby standardized residual was plotted and checked for the auxiliary variables considered for the research study whose range of standardized residuals were checked for ±3.0 in absolute value. Table III provides the summary of these techniques so that a comparative study on outlier detection could be made.

Table III. Results obtained from outlier detection techniques and the residual from normal probability plots

S.No	Products	Possibilities of Outliers	Residuals
1	Soap all kinds(SSI)	Nil	Normal
2	Agarbathi	Nil	Normal
3	Country Liquor	Nil	Normal
4	Indian made foreign liquor	Nil	Normal
5	Indian foot wear	September 2007	4.64
6	Shoe Upper	September 2010	-5.32
7	Cotton Yarn	July 2008	Normal
8	Cotton Hoisery	NIL	Normal

In the case of Indian foot wear, it was found that one of the standardized residual was lying out of the range ±3.0 in absolute value and hence there was an outlier found in data which lies in the month of September 2007 which could be due to heavy demand for the Indian footwear. For the Shoe upper data there was an outlier found in data which lies in the month of September 2010 which could be due to increase of Shoe upper price and similar observation has been found in Cotton yarn data which lies in the month of July 2008 which could be due to exorbitant tax levied on the export in supplying the Cotton yarn. There was an outlier in Nylon tyre cord product which lies in the month of July 2011 which could be due to shortage of production and outlier in the Middling data which lies in the month of March 2007 that could be limiting of raw material and heavy tax imposed over raw materials supplied to the industry. Similarly, an outlier was observed through the analysis of bicycle tubes data in the month of November 2008 which could be due to heavy tax levied. It was found that two outliers, one in the month of August 2006 which could be due to devaluation and another in the month of July 2009 was observed in reactive dyes data that could be due to monetary policy. Further, to check the validity of the model, Cross Validity Predictive Power (CVPP); the shrinkage coefficient of the model $\eta = (R^2 - \rho^2_{cv})$ where ρ^2_{cv} is CVPP and R^2 the coefficient of determination. The results obtained for all the industrial products were presented in Table IV. From the results, it could be observed that the fitted model for all the products has been validated and there is no strong evidence for revising the model for all the products considered in the research.

Table IV. Showed model validation techniques Cross Validity Predictive Power (CVPP), coefficient of determination (R^2), and shrinkage factor (η).

Products	ρ^2_{cv}	R^2	η
Soap all kinds(SSI)	0.7637	0.8090	0.0453
Agarbathi	0.7917	0.8314	0.0397
Country Liquor	0.9200	0.9350	0.0150
Indian made foreign liquor	0.8475	0.8769	0.0299
Indian foot wear	0.5921	0.6710	0.0788
Shoe Upper	0.6396	0.7086	0.0068
Cotton Yarn	0.7139	0.7725	0.0586
Cotton Hosiery	0.8298	0.8954	0.0656

VI. CONCLUSION

The eight industrial products dataset given for the period over 2006 to 2015 on monthly basis considered for linear regression analysis for fitting the linear model. The investigation incorporates dataset and it was analyzed for the relapse presumptions of homocedasticity, linearity, typicality, autonomy of residuals and multicollinearities inspected for CVPP model. The examination study was limited to two factors distinguished dependent on its monetary significance and other related model aspects. The linear model was fitted for the appropriate variables listed in each category of products over the periods and sum of year wise production was compared with over all years production. Univariate Mahalonobis anomaly location method was applied for the evaluation that separated the factual factors over the years. Concerning the outlier detection a standardized residual was plotted and the results were analyzed based on the range of the residuals lying within ±3.0 in absolute value. Further, to check the legitimacy of the model, the shrinkage coefficient of the model η has been computed.

The regression coefficients have been estimated and results are obtained for the coefficients of determination. The significant for the coefficients are tested at 5% level for all auxiliary variables and the overall estimated values are also obtained. The outlier test indicated that all models pertain to the considered products do not have much outliers except for the products; Indian foot wear, Shoe Upper and Cotton Yarn. On the observed results, it could be seen that the fitted model for all the products has been validated and there is no strong evidence for revising the model for all the products considered in the research. Several univariate and multivariate correlation measurements could be used to determine relationship among variables using additive models. However, in some multivariate data sets additive models may fail to determine certain complex relationships. Therefore, multiplicative models could be appropriate for assessing the relationship between the variables.

REFERENCES

[1] Pack H. (1988), ‘Industrialization and trade’, in Chenery H.B. and Srinivasan T.N. (eds.), Handbook of Development Economics, North Holland, Amsterdam, Vol. I, pp. 334-380.

[2] Cooper N. and Richard (1999), ‘Key currencies after the Euro’, The world Economy, (22), CSIRO Technical Report CMIS-02/102.

[3] Daniel C. (1960), ‘Locating outliers in factorial experiments’, Technometrics, Vol. 2, pp. 149-156.

[4] Han and Kamber (2000), ‘Data Mining: Concepts and Techniques’, 2nd Edition, Morgan Kaufmann Publishers.

[5] Balasooriya U. and Tse Y.K. (1987), 'Outlier detection in linear models: a comparative study in simple linear regression', Journal of the American Statistical Association, Vol. 68, pp. 941-943.

[6] (Stevens, 1996 and Ali, 2000) 'Applied multivariate statistics for the social sciences', Lawrence Erlbaum Associates, New Jersey and USA.

[7] Islam R. (2003), 'Modeling of Demographic parameters of Bangladesh-An Empirical Forecasting', Unpublished Ph.D. Thesis, Rajshahi University.

[8] Childs A., Balakrishnan N. and Srinivasan M.R. (2006), 'Unified scheme for testing for outlier in linear models', Journal of Statistical Computation and Simulation, Vol. 96, pp. 21-39.

[9] Williams G.J., Baxter R.A., He H.X., Hawkins S. and Gu L. (2002), 'A Comparative Study of RNN for Outlier Detection in Data Mining', IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan.

[10] Draper N.R. and Smith H. (1998), 'Applied Regression Analysis', Third edition, John Wiley & Sons Inc., New York.

[11] Khattree R. and Rao C.R. (2003), 'Time Series in Industry Handbook of Statistics', (22), Statistics in Industry, Elsevier Science B.V. (eds.).

[12] Barnett V. and Lewis T. (1994), 'Outliers in Statistical Data', 3rd Edition, John Wiley and Sons, Chichester, England.

[13] Acuna E. and Rodriguez C. (2004), 'A Meta-analysis study of outlier detection methods in classification', Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrieved from academic.uprm.edu/eacuna/paperout.pdf, in Proceedings IPSI 2004, Vemice.

[14] Papadimitriou S., Kitawaga H., Gibbons P.G., Faloutsos C. (2002), LOCI: Fast Outlier Detection using the Local Correlation Integral', Intel Research Laboratory Technical report no.IRP-TR-02-09.