The Sales Prediction of Abctronics' based on the Multi-factorial Linear Model in Terms of Variance Inflation Factor

Zheming Cao[1,†][a] and Sheng Luo[2,†][b]
[1]*Art & Science, University of Toronto, Ontario, Canada*

[2]*Rutgers Business School, Rutgers University, New Brunswick, U.S.A.*

[†]*These authors contributed equally*

Keywords: Multi-Factorial Linear Model, Inflation Factor, Linear Regression.

Abstract: To make the prediction of ABCtronics' sales, the multi-factorial linear model is used. In terms of variance inflation factor, a linear regression model can be set up by Minitab for us to analyze and predict the ABCtronics' sales. Based on the approach, the estimated multiple regression equation is derived for sales volume with respect to total market demand, price per chip, and economic situation. According to the analysis, the total market variable is not statistically significant at the 5% significance level. These results shed light on processing a more precise prediction by using multi-factorial linear model in terms of variance inflation factor.

1 INTRODUCTION

ABCtronics is looking for avenues to improve its manufacturing as well as quality control processes. In a crucial quarterly review meeting, Jim Morris, the chief operating officer (COO) of the Chip Manufacturing Unit, is meeting with the quality control, manufacturing, and marketing team to decide the way forward (Irwin, Mcclelland, 2001). The case provides the relevant data on the plant performance, detailed description of its operating procedures, quality control policies and issues currently faced at a client site. In order to offer a precise prediction of ABCtronics' sales, we will use the multifactorial linear regression and variance inflation factors to set up and analysis the model. Some scholars have discussed the multicollinearity illusion in moderated regression analysis (Disatnik, Sivan, 2016). Numerous papers in the fields of marketing and consumer behavior that utilize moderated multiple regression express concerns regarding the existence of a multicollinearity problem in their analyses. In most cases, however, as show in this paper, the perceived multicollinearity problem is merely an illusion that arises from misinterpreting high correlations between independent variables and interaction terms (Becker, Ringle, Sarstedt, Völckner, 2015) and a small sample performance of the Wald test in the sample selection model under the multicollinearity problem (Andrews, Brusco, Currim, Davis, 2010). This paper reviews and extends the literature on the finite sample behavior of tests for sample selection bias. According to previous analysis, the standard regression-based t-test and the asymptotically efficient Lagrange Multiplier test, are robust to nonnormality but have very little power (Kalnins, 2018). There are some miss leadings of multiple regression models were discussed. Moderated multiple regression models allow the simple relationship between the dependent variable and an independent variable to depend on the level of another independent variable. The moderated relationship, often referred to as the interaction, is modeled by including a product term as an additional independent variable. Multiple regression models not including a product term are widely used and well understood. It is argued that researchers have derived from this simpler type of multiple regression several data analysis heuristics that, when inappropriately generalized to moderated multiple regression, can result in faulty interpretations of model coefficients and incorrect statistical analyses.

[a] https://orcid.org/ 0000-0002-4671-0927
[b] https://orcid.org/ 0000-0002-6721-0200

Cao, Z. and Luo, S.
The Sales Prediction of ABCtronics' based on the Multi-factorial Linear Model in Terms of Variance Inflation Factor.
DOI: 10.5220/0011194700003440
In *Proceedings of the International Conference on Big Data Economy and Digital Management (BDEDM 2022)*, pages 587-591
ISBN: 978-989-758-593-7

Based on the theoretical arguments and constructed data sets, previous literature has discussed heuristics, discuss how they may easily be misapplied, and suggest some good practices for estimating, testing, and interpreting regression models that include moderated relationships (Yamagata, 2006). In order to figure out the collinearity impacts on mixture regression result, we collect some statement form Becker and Ringle. Mixture regression models are an important method for uncovering unobserved heterogeneity. A fundamental challenge in their application relates to the identification of the appropriate number of segments to retain from the data. Prior research has provided several simulation studies that compare the performance of different segment retention criteria.

Although collinearity be- tween the predictor variables is a common phenomenon in regression models, its effect on the performance of these criteria has not been analyzed thus far. We address this gap in research by examining the performance of segment retention criteria in mixture regression models characterized by systematically increased collinearity levels. The results have fundamental implications and provide guidance for using mixture regression models in empirical (marketing) studies (Yamagata, Orme, 2005). Based on some of their theories, we will use multi-factorial linear regression and variance inflation factors to predict ABCtroncs' sales. First, we will calculate date using Minitab and displayed in the original formula, e.g., multiple linear regression. Subsequently, a regression model is applied for ABCtronics' sales and carry out the analysis based on the software Minitab. The regression model test is widely available and has been used in many investigational studies. Traditionally, simple linear regression has been assessed by measuring. Subsequently, the reasons attributed to the prediction of ABCtronics are analyzed and the approach to predict it will also be demonstrated. Finally, the future usage of the opinions in the paper is demonstrated as well as the application.

2 DATA & METHOD

The research method is a simple linear regression model for ABCtronics' sales based on the analysis software Minitab. The regression model test is widely available and has been used in many investigational studies. Traditionally, a simple linear regression has been assessed by measuring.

$$Y = \beta^0 + \beta^1 X + \varepsilon, \varepsilon \sim Normal(0, \sigma^2) \quad (1)$$

We include the expected term with a normal distribution to improve the estimate. The simple linear regression model only explains the slight change in sales volume, which makes us doubt whether the model is enough to be a good forecast. Multiple regression models that do not include product terms are widely used and understood. Scholars believe that this simple multiple regression led to some data analysis heuristics. Moderately improper expansion of these heuristics in multiple regressions can lead to misinterpretation of model coefficients and inaccurate statistical analysis. In this case, it is necessary to consider multiple regression models. The main difference between simple linear regression and multiple linear regression is the number of independent variables. We decided first to use all three variables to run the multiple regression models X1, X2, X3 with the following regression equation:

$$Y = \beta^0 + \beta^1 X^1 + \beta^2 X^2 + \beta^3 X^3 + \varepsilon, \varepsilon \sim Normal(0, \sigma^2) (2)$$

Becker et al. claimed that it is usually assumed that the regression coefficients are constant (Becker, Ringle, Sarstedt, Völckner 2015). Besides, the same regression equation is sufficient to describe all members of the population. Under normal circumstances, this assumption of similar data structures does not hold. If not fully explained, non-uniformity can lead to inaccurate results and preliminary conclusions. Before drawing conclusions about the comparison between the two models, the final step is to test for multicollinearity. If there is multicollinearity in the latest multiple regression model, it fails. We test for multicollinearity by calculating the VIF variance inflation factor, which indicates the correlation between the two independent variables.

$$VIF(X_j) = \frac{1}{1 - R_j^2} \quad (3)$$

Numeric publications in the fields of marketing and consumer behavior that use mitigated multiple regression raise concerns about the existence of multicollinearity issues in analysis. However, in most cases, as shown in this paper, the perceived multicollinearity problem is merely an illusion resulting from a misunderstanding of the high correlation between the independent variable and the interaction term. In order to obtain multiple regression model and simple regression model, VIF, the Minitab software is utilized to perform calculations and analysis.

3 RESULTS

3.1 Simple Linear Regression

The regression result of simple linear regression is given as Eq. (4) and Tables. I~III:

$$Sales\ Volume = 0.753 + 0.00614\ Market\ Demand \quad (4)$$

Table 1: Coefficients- Simple Linear Regression.

Term	Coef	SE Coef	T-Value	P-value	VIF
Constant	0.753	0.779	0.97	0.362	
Overall Market Demand	0.00614	0.00344	1.79	0.112	1.00

Table 2: Model Summary- Simple Linear Regression.

S	R-sq	R-sq(adj)	R-sq(pred)
0.833884	28.50%	19.56%	0.00%

Table 3: Analysis of Variance- Simple Linear Regression.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2.217	2.2171	3.19	0.112
Overall Market Demand	1	2.217	2.2171	3.19	0.112
Error	8	5.563	0.6954		
Total	9	7.780			

In Table II, R^2 shows that only 28.5% of the variation in the sales volume is explained by this regression model. Then the estimated multiple linear regression equation of sales volume on overall market demand, price per chip, and economic condition are given.

3.2 Multiple Linear Regression

Table 4: Coefficients- Multiple Linear Regression.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.86	1.35	6.57	0.001	
Overall Market Demand	-0.00524	0.00258	-2.03	0.089	3.27
Price Per Chip	-5.505	0.881	-6.25	0.001	2.54
Economic Condition	1.130	0.342	3.30	0.016	2.44

Table 5: Model Summary- Multiple Linear Regression.

S	R-sq	R-sq(adj)	R-sq(pred)
0.346274	90.75%	86.13%	72.75%

Table 6: Analysis of Variance - Multiple Linear Regression.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	7.0606	2.3535	19.63	0.002
Overall Market Demand	1	0.4930	0.4930	4.11	0.089
Price Per Chip	1	4.6786	4.6786	39.02	0.001
Economic Condition	1	1.3089	1.3089	10.92	0.016
Error	6	0.7194	0.1199		
Total	9	7.7800			

$$Sales\ Volume = 8.86 - 0.00524 Market\ Demand - 5.505\ Price + 1.130\ Economic\ Condition \quad (5)$$

The regression result of mutiple linear regression is given in Eq. (7) and Tables. IV-VI. For Table IV, the p-value of overall market demand is only 0.089, which is not statistically significant, i.e., one needs to perform another regression.

3.3 Regression Analysis on Price and Condition

Table 7: Coefficients- Analysis on Price and Condition.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.452	0.766	8.42	0.000	
Price Per Chip	-4.136	0.680	-6.08	0.001	1.05
Economic Condition	0.606	0.269	2.25	0.059	1.05

Table 8: Model Summary- Analysis on Price and Condition.

S	R-sq	R-sq(adj)	R-sq(pred)
0.416172	84.42%	79.96%	64.42%

The regression analysis results of the price and condition are given in Eq. (6) and Table VII.

$$Sales\ Volume = 6.452 - 4.136\ Price + 0.606\ Economic\ Condition \quad (6)$$

For more accuracy, we use adjusted R^2 in Table VIII. The adjusted R^2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

3.4 Test Multicollinearity

Table 9: Coefficients- Test Multicollinearity.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.452	0.766	8.42	0.000	
Price Per Chip	-4.136	0.680	-6.08	0.001	1.05
Economic Condition	0.606	0.269	2.25	0.059	1.05

In Table XI, the VIF of Price Per Chip is 1.05 which is VIF (X2, X3) <5. Hence, there is no multicollinearity. Comparing the two results, it can be seen that Multiple Linear Regression is better than the simple linear regression model used by ABCtronics.

4 DISCUSSION

ABCtronics uses a simple linear regression model to forecast sales volume based on aggregate market demand. For example, according to a regression model, we take aggregate demand in the market as an independent variable. $Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim$ normal $(0, \boldsymbol{\sigma}^2)$. Here, we include the error term of the normal distribution for estimation accuracy. According to the evaluation from Minitab, the estimated parameters for $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ along with the estimated linear regression equations. The model summary of the two values shows that only 28.5% of sales fluctuations are the cause. This volume regression model explains why it makes us wonder if this model is sufficient for good predictions. Yao, & Li argued that linear regression varies from normal linear regression in that it models the conditional mean (Yao, Li, 2014). To address this issue, we need to consider the multiple regression model proposed by the SMT intern. The main difference between simple regression and multiple regression is the number of independent variables. Here, in Table III, we first run a multiple regression model using all three variables, and to simplify it, ABCtronics sales volume Y, total market demand X1, price per chip X2, and economic situation X3 are shown, respectively. The corresponding model used is $Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 X_2 + \boldsymbol{\beta}_3 X_3 + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim$ Normal $(0, \boldsymbol{\sigma}^2)$. Minitab is utilized to regress on X1 and X2, including error terms. Therefore, the estimated multiple regression equation for sales volume with respect to total market demand, price per chip, and economic situation.

Horssen et al. denotes that prediction ranges derived from logistic regression model predictions when uncertainties in explanatory variables and uncertainty in regression coefficients are considered (van Horssen, Pebesma, Schot 2002). To make a more rigorous estimate of the economic situation, we look further at the statistical significance of each independent variable, including the p-value, and level 5% of significance. It is verified again to ensure that the p-value for the market-wide concept is 0.089, which is greater than 5%. To sum up, we find that the total market variable is not statistically significant at the 5% significance level. Before proceeding to rebuild the multiple regression model, analysis claim why chose 5% as the standard. One can also find the p-value for total market demand from the previous simple linear regression model. This simple model gives a p-value of 0.112. A market-wide term is defined as not statistically significant, not to mention the 5% level, even if the significance level is 10%. This may work for higher levels of significance, but overall accuracy makes it less likely to select too high level of significance for each model. Therefore, if one omits the demand period, the multiple regression model to is able to estimate sales volume more accurately. The modified equation regresses only the price per chip and the economic situation, and the equation becomes Eq. (6).

This model uses adjusted R^2, and the summary is as follows. Approximately 79.96% of sales volume fluctuations are explained by this regression model. This is much higher than the data obtained from a simple regression model. The final step before concluding the contrast between the two models is to test multicollinearity. Yamagata and Orme indicate that when the “multicollinearity problem” is severe, the t-test based on the Heckman–Greene variance estimator can be unreliable, but the Likelihood Ratio test remains powerful, and nonnormality can be interpreted by Maximum Likelihood methods as severe sample selection bias, resulting in negative Wald statistics (Yamagata, 2006). Arturs Kalnins (Kalnins, 2018) states that even if the variables’ true effects are minor and of the same sign, calculated beta coefficients will trend to infinite magnitudes in opposite directions if they are associated via an unobservable common component. Diagnostics (e.g., VIF) can be used to falsely verify Type 1 mistakes as genuine results. If the latest multiple regression

model has multicollinearity, it will fail. To test multicollinearity, calculate the VIF variance factor. This gives a correlation between the two independent variables, which shows that it is 1.05, which is much smaller than the limit line of 5. To sum up, there exists no multicollinearity and we suggest using multiple linear regression models proposed by STM interns since it has stronger explanatory power. Takashi Yamagata shows that the Maximum Likelihood (ML) estimator is robust to such multicollinearity and can produce a more reliable estimator in the same circumstances (Yamagata, 2006).

The limitation of this sales forecasting study is that the data in the study is not enough. There are only 40 customers in the sample, and there is no in-depth investigation into the research data of a larger sa[ple size. Compared with the regional vegetation models presented in Ref. (van Horssen, Pebesma, Schot, 2002) are based on a database of 306 sample locations throughout the area. The abundance of 78 wetland plant species, as well as 21 environmental characteristics, were recorded at Each location. The sample of ABCtronics is insufficient. Additionally, the time span of historical data is that there is not enough new sales data for the period from 2004 to 2013, i.e., there will be slight differences in feedback on the recent situation. For the analysis of sales data, there is no specific analysis to the month, and the evaluation is also a parameter for one year. This may make people who need to know more specific to the month feel that there is not enough detail.

5 CONCLUSIONS

In summary, we construct a model to predict ABCtronics' sales based on the Multiple Linear Regression Model in terms of Variance Inflation Factor with the analysis software Minitab. According to the analysis, the model proposed by the SMT interns predicts the sales figure better than the model previously used by ABCtronics. ABCtronics use simple linear regression with low R^2-low accuracy. SIM interns propose multiple linear regression. Multi-linear regression gives a more precise prediction -- suggest using this one. The significance of the results is to help manufacturing companies make better sales forecasts. Therefore, it turns out that we are unable to solve for companies with large and complex data sources. During the research, the multicollinearity illusion in moderated regression analysis is discussed. The perceived multicollinearity problem is merely an illusion that arises from misinterpreting high correlations between independent variables and interaction terms. Moreover, based on Multi-linear regression and VIF, a more accurate sales forecast can be provided. It is also hoped that there will be more models of predictions that can help companies make decisions in the future. These results offer a guideline for companies to fast respond to market conditions promptly to adjust their sales strategies.

REFERENCES

Andrews, Brusco, M., Currim, I. S., & Davis, B. (2010). An Empirical Comparison of Methods for Clustering Problems: Are There Benefits from Having a Statistical Model? Review of Marketing Science, 8(1). https://doi.org/10.2202/1546-5616.1117

Arnab Adhikari, Indranil Biswas, Arnab Bisi(2016) Case—ABCtronics: Manufacturing, Quality Control, and Client Interfaces. INFORMS Transactions on Education 17(1):26-33. https://doi.org/10.1287/ited.2016.0158cs

Becker, Ringle, C. M., Sarstedt, M., & Völckner, F. (2015). How collinearity affects mixture regression results. Marketing Letters, 26(4), 643–659. https://doi.org/10.1007/s11002-014-9299-9

Disatnik, & Sivan, L. (2016). The multicollinearity illusion in moderated regression analysis. Marketing Letters, 27(2), 403–408. https://doi.org/10.1007/s11002-014-9339-5

Irwin, & Mcclelland, G. H. (2001). Misleading Heuristics and Moderated Multiple Regression Models. Journal of Marketing Research, 38(1), 100–109. https://doi.org/10.1509/jmkr.38.1.100.18835

Kalnins. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. Strategic Management Journal, 39(8), 2362–2385. https://doi.org/10.1002/smj.2783

P.W van Horssen, E.J Pebesma, P.P Schot, Uncertainties in spatially aggregated predictions from a logistic regression model, Ecological Modelling, Volume 154, Issues 1–2, 2002, Pages 93-101, ISSN 0304-3800, https://doi.org/10.1016/S0304-3800(02)00060-1

Yamagata, & Orme, C. D. (2005). On Testing Sample Selection Bias Under the Multicollinearity Problem. Econometric Reviews, 24(4), 467–481. https://doi.org/10.1080/02770900500406132

Yamagata. (2006). The small sample performance of the Wald test in the sample selection model under the multicollinearity problem. Economics Letters, 93(1), 75–81. https://doi.org/10.1016/j.econlet.2006.03.049

Yao, & Li, L. (2014). A New Regression Model: Modal Linear Regression. Scandinavian Journal of Statistics, 41(3), 656–671. https://doi.org/10.1111/sjos.12054.