Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/45/20230302

Case Study of Setting Up Linear Regression Model for Turnover of a High-tech Company

Keren Yang[1,a,*]

[1]*University of Toronto Scarborough, 1265 Military Trail, Scarborough, ON, M1C 1A4, Canada*
a. karen.kr.yang@gmail.com
**corresponding author*

Abstract: This research paper wished to discover whether a linear regression model is enough to summarize the relationship between different factors and the turnover of a high-tech company. To learn it, we first read through a lot of references to find out the mysterious origin of the linear regression model. Moreover, we must pay attention to the similarities and differences between scientists' discoveries and our goal. Methodology in the whole process was also impressive since the data was obtained from employees working for the company, which means the data was brand new and confidential in certain respects. Therefore, we tried our best to protect all the involvers. The following parts are the results statement and discussion. We used the multiple linear regression model in Excel to obtain tables and figures, which helped to directly understand the data and analyze the correlation between the independent variables (i.e. what we always used to estimate turnover), including dummy variables and the dependent variable (i.e. turnover). Although we could improve, we still criticize ourselves in the discussion. Finally, we made conclusions in our research with the help of official websites, references mentioned in the last part, participants offering experimental data, writing classes provided, etc. The conclusion might be one-sided, but the arduous research process we have gone through made the conclusion credible, valuable, and precious.

Keywords: linear regression model, a high-tech company's turnover, Excel

1. Introduction

The linear regression model is a fundamental tool in our statistics study and daily life. There are many applied aspects, for instance, design for artificial intelligence, calculation and prediction of economic and financial products and projects, and estimation of turnover for a high-tech company monthly and yearly. When looking through hundreds of thousands of academic and literal works and websites online, there were few data sets for companies' profit or turnover. Therefore, we decided to invite some employees who are responsible for predicting the future and calculating the actual budget in the same company to be the participants and asked them to offer me real data for further analysis. The reason why we choose this topic to be the subject is not only because of the shortage of data, but also because we do think it is meaningful because it can help us understand the structure of a company's turnover, and the correlation among factors as well. Our goal of the research is to check the availability of a linear regression model in transforming data to be a line expression which we can directly understand. In the meantime, the way employees carry out to predict turnover successfully

DOI: 10.54254/2754-1169/45/20230302

based on the human resources needed, labor epiboly needed, actual applying fees, and actual epiboly fees is the important harvest we got from the research as well. This section lacks support from the literature, please add citations.

2. Literature Review

Linear regression is a widely used statistical technique for quantifying the relationship between two or more variables according to Kleinbaum et al. [1]. It is a fundamental tool in statistical modelling and analysis and has a long history dating back to the 19th century. To start our research, we need to be familiar with the history of linear regression, its first application, and its impact on statistical and scientific research.

Karl Pearson is always considered the inventor of linear regression based on the name: Pearson Product Moment Correlation Coefficient (PPMCC). However, according to Paul [2], Sir Francis Galton was the first to figure out the structure of correlation and regression. As one of the greatest scientists in the 19th century, Galton was not famous among the public since he was a cousin of Charles Darwin. Galton made a great distribution to statistics, biology, etc., which also led to the original imagination towards regression and PPMCC [3].

Galton was interested in exploring the relationship between traits in parents and offspring and developed the concept of regression to the mean [4][5]. Regression to the mean refers to the phenomenon where extreme values in a dataset tend to move closer to the mean in subsequent measurements. Galton developed methods for estimating this regression effect, which involved fitting a straight line to the data.

The first systematic use of linear regression as a statistical method was by the American mathematician Francis Anscombe in the mid-20th century. Anscombe's work was motivated by the problem of estimating the relationship between two variables, such as height and weight, in a way that minimized the error between the observed data and the fitted model. He developed a method for estimating the slope and intercept of a straight line that provided the best fit to a given dataset, which became known as simple linear regression [6].

The Framingham Heart Study is a notable example of linear regression applied to a real-world problem [7]. Began in 1948, this study used linear regression to model the relationship between a variety of risk factors and the likelihood of developing cardiovascular disease. Using this model, researchers were able to identify key risk factors, such as high blood pressure and smoking, and develop interventions to reduce cardiovascular disease risk. Similarly, we could use modern technology and software to help us to stimulate methods of discovering the effectiveness of regression models from sales behavior and related aspects of a firm.

In conclusion, linear regression has a long and fascinating history, dating back to the pioneering work of Sir Francis Galton. It has become an essential tool in many areas of science, enabling researchers to model complex relationships and make accurate predictions. As new techniques for regression analysis are developed, linear regression will continue to play a vital role in shaping our understanding of the natural world and commercial world.

3. Methodology

To directly analyze a high-tech company’s annual turnover of products based on quantity needed by clients, their different requirements on the products, for instance, human resources needed, labor epiboly needed, actual applying fees, and actual epiboly fees spent on the development and sales of various software products. Moreover, the average correlation between turnover and total cost is also wanted in the study.

DOI: 10.54254/2754-1169/45/20230302

The dataset, which will be used in the essay, would be the first-hand data provided by the company's employees. The content is formed by a lot of projects that the company concluded contracts with other corporations last year. All the numbers and values were mainly collected through interviewing and negotiating with some employees who participated in the relative work in the past few years, and it is organized by them and me from a great number of mixed reports. It is noteworthy that due to the privacy problem, only part of the data can be shown with no exact company's name mentioned in the whole passage. All the participants and records in the study will be entirely anonymous.

The quantitative method is the essence of the research, which will be analyzed using Excel in the following paragraph. There are 469 records for all completed projects in 2022 for this company. With net order value representing turnover, the four and actual costs are considered. To be more specific, we first used a linear regression model to figure out the linear relationship between those cost factors and total cost. After getting the expression of the line, use the correlation tool to further develop the inner relations. For the aim of investigating the relationship between developing software, clients' needs towards the products, costs of production and the turnover of the company. Then, the correlations among variables will also be analyzed using PPMCC (r). Moreover, from the ANOVA table, we can learn more about the coefficients of each variable, T Statistic, p-value, etc.

We used Excel for the category of data collected from involvers and for the convenience of data presenting to the reader. The easy way to handle it and the complete information included are also reasons I used Excel. The following parts will explain the results and messages got during the progress, as well as the weaknesses that need to be improved.

4. Results

Table 1: ANOVA table for how expected turnover related to labor epiboly needed, actual applying fees and actual epiboly fees for the company in 2022.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.977353				
R Square	0.955219				
Adjusted R Square	0.954833				
Standard Error	555250.1557				
Observations	469				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	3.0514E+15	7.6286E+14	2474.3762	0
Residual	464	1.4305E+14	3.0830E+11		
Total	468	3.1945E+15			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	164751.1126	31744.0996	5.1900	3.1509E-07	

DOI: 10.54254/2754-1169/45/20230302

Table 1: (continued).

human resources needed	1114.2864	77.1016	14.4522	2.3619E-39	
labor epiboly needed	878.3343	226.3062	3.8812	1.1901E-04	
actual applying fees	6.9200	0.3142	22.0244	4.0921E-74	
actual epiboly fees	1.4740	0.1354	10.8886	9.6229E-25	

Firstly, as Table.1 shows, $\bar{Y} = 164751.1126 + 1114.2864X_1 + 878.3343X_2 + 6.9200X_3 + 1.4740X_4$. ($\bar{Y}$, X_1, X_2, X_3, X_4 representing expected turnover, human resources needed, labor epiboly needed, actual applying fees, and actual epiboly fees in the following paragraphs.) Coefficients in front of X1 and X2 represent fees per person or labor epiboly involved. The R-squared of the model is 0.955219, and the adjusted R-squared is 0.954833. From the short review above, the key findings emerge the turnover of the company is highly related to four factors included and are all positively related.

Table 2: Correlation table for among human resources needed, labor epiboly needed, actual applying fees and actual epiboly fees for the company in 2022.

	human resources needed	labor epiboly needed	actual applying fees	actual epiboly fees
human resources needed	1			
labor epiboly needed	0.1623	1		
actual applying fees	0.9240	0.1938	1	
actual epiboly fees	0.3247	0.0730	0.2943	1

Secondly, Table.2 describes the result of the inner correlation of X_1, X_2, X_3, X_4. A novel finding is that correlation between X_1 and X3 is 0.9240, which is extremely higher than the expected normal value should be between two independent variables. Therefore, we tried to delete X_1 and X_3 respectively and check the ANOVA Table again (Table.3 and Table.4).

Table 3: ANOVA Table for how expected turnover related to labor epiboly needed, actual applying fees and actual epiboly fees for the company in 2022.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.966986				
R Square	0.935061				
Adjusted R Square	0.934642				
Standard Error	667922.8644				
Observations	469				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2.9870E+15	9.9568E+14	2231.8566	1.3469E-275
Residual	465	2.0745E+14	4.4612E+11		
Total	468	3.1945E+15			

DOI: 10.54254/2754-1169/45/20230302

Table 3: (continued).

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	311876.6240	36168.8217	8.6228	1.0351E-16	
labor epiboly needed	722.8318	271.9209	2.6582	8.1256E-03	
actual applying fees	11.0739	0.1527	72.5271	1.2110E-255	
actual epiboly fees	1.7581	0.1611	10.9121	7.7435E-25	

Table 4: ANOVA Table for how expected turnover related to human resources needed, labor epiboly needed and actual epiboly fees for the company in 2022.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.953102				
R Square	0.908404				
Adjusted R Square	0.907813				
Standard Error	793253.6438				
Observations	469				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2.9019E+15	9.6729E+14	1537.2121	7.1118E-241
Residual	465	2.9260E+14	6.2925E+11		
Total	468	3.1945E+15			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	61272.0078	44851.4614	1.3661	1.7256E-01	
human resources needed	2667.6698	44.4983	59.9499	6.6071E-221	
labor epiboly needed	1458.6717	321.1115	4.5426	7.0908E-06	
actual epiboly fees	1.4192	0.1934	7.3396	9.6358E-13	

For Table.3, after deleting the variable X_1, $\bar{Y} = 311876.624 + 722.8318X_2 + 11.0739X_3 + 11.7581X_4$, with newly R-squared equals 0.935061, and new adjusted R-squared is 0.934642. Similar to Table.4, with no X_3 included, $\bar{Y} = 61272.0078 + 2667.6698X_1 + 1458.6717X_2 + 1.4192X_4$, with R-squared is 0.908404, and adjusted R-squared is 0.907813. Our results demonstrated that both R-squared and adjusted R-squared decrease after dropping highly correlated variables.

DOI: 10.54254/2754-1169/45/20230302

Table 5: ANOVA table for how expected turnover related to human resources needed, labor epiboly needed, actual applying fees and actual epiboly fees with no constant term in the expression for the company in 2022.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.983352				
R Square	0.966982				
Adjusted R Square	0.964618				
Standard Error	570524.8772				
Observations	469				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	4.4327E+15	1.1082E+15	3404.5270	0
Residual	465	1.5136E+14	3.2550E+11		
Total	469	4.5840E+15			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	0	#N/A	#N/A	#N/A	
human resources needed	1242.6141	75.0383	16.5597	9.4203E-49	
labor epiboly needed	1151.7511	226.1435	5.0930	5.1305E-07	
actual applying fees	6.6787	0.3193	20.9175	5.7656E-69	
actual epiboly fees	1.4757	0.1391	10.6095	1.0730E-23	

Table.5 ANOVA Table for how expected turnover related to human resources needed, labor epiboly needed, actual applying fees and actual epiboly fees with no constant term in the expression for the company in 2022.

Based on our involvers' saying, the company always asked them to use a specific coefficient for estimating the turnover without adding a constant. Therefore, Table.5 indicates the ANOVA Table with no intercept (i.e. constant term). The new line expression for the same data becomes $\bar{Y} = 1242.6141X_1 + 1151.7511X_2 + 6.6787X_3 + 1.4757X_4$. After this change, R-squared increases to 0.966982, and adjusted R-squared increases to 0.964618. This result highlights that our participants gave out correct information and helped us to improve the model's quality and fitting degree.

5. Discussion

The results indicate that a company's turnover can be predicted with human resources needed, labor epiboly needed, actual applying fees and actual epiboly fees. According to Chicco et al. [8], R-squared, the coefficient of determination, is significant not only for the number itself but for representing the goodness of fit of a linear regression model. In our case study, R-squared is 0.955219 for the one with intercept and 0.966982 for the one without intercept. More than 95% means we at least set an

DOI: 10.54254/2754-1169/45/20230302

appropriate model for the existing data and we can get a reliable linear expression. Furthermore, adjusted R-squared being 0.954833 and 0.964618 means the availability of our applying multiple linear regression model instead of others. The results for deleting the constant term are with higher fitness to the model based on increased R-squared, which agrees with our involvers' statement of dropping the constant term. These results help to build on existing evidence of the correctness of establishing a linear relationship between turnover and different cost factors.

However, it is mentioned above that the correlation between human resources needed and actual application fees is too high. The existence of the multicollinearity implies these two independent variables are highly correlated and we should remove one of them. Therefore, we have new results appearing in Table.3 and Table.4, depicting results without one of two variables. Unfortunately, the lower number for R-squared and adjusted R-squared contradicts the claims of Anderson et al. that dropping one of the highly correlated variables would improve the model [9].

Exploring more dummy variables is also beyond this study's scope. In fact, labor epiboly needed and actual epiboly fees in the data set include a lot of zeros, which means some projects done by the company in 2022 are free of these potential fees. Moreover, the generalizability is limited by intercept or slope dummy variables consideration as well.

Therefore, further research is needed to check the existence of contradiction with the definition of multicollinearity and whether there are more independent variables made up of existing ones. Similarly, if the R-squared and adjusted R-squared increase, we will get a better-fitting linear regression.

6. Conclusions

This research aimed to set up a specific linear regression model for estimating a company's turnover. Based on a quantitative analysis of a company's turnover in response to the effect of human resources needed, labor epiboly needed, actual applying fees, and actual epiboly fees, it can be concluded that these are significant factors to determine the amount of turnover. The results indicate that potential variables can help to establish the linear relationship, but it also raises the question of failing to eliminate multicollinearity and difficulty of increment parameters such as intercept and slope dummy variables.

Based on these conclusions, practitioners should consider how to improve their daily work of estimating total cost and net order volumes. To better understand the implications of these results, future studies could address how the change of coefficient will affect turnover. Moreover, whether there are other factors, for instance, the existence of COVID-19, quantities of projects, and the percentage of completeness. Further research is needed to determine whether multicollinearity influences the linear expression and whether there will be improvement in having more independent dummy variables.

The linear regression model is now being applied in various developing fields. Therefore, case studies, as this research paper did, are necessary in order to be able to explore more of its nature and principles.

References

[1] *Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). Applied regression analysis and other multivariable methods: student solutions manual. Cengage Learning.*

[2] *Paul, D. B. (1995). Controlling human heredity: 1865 to the present. Humanities Press.*

[3] *Fitzpatrick, P. J. (1960). Leading British Statisticians of the Nineteenth Century. Journal of the American Statistical Association, 55(289), 38–70. https://doi.org/10.1080/01621459.1960.10482048.*

[4] *Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246–263. https://doi.org/10.2307/2841583.*

[5] *Pearson, K. (1924). The Life, Letters and Labours of Francis Galton. Cambridge University Press.*

DOI: 10.54254/2754-1169/45/20230302

[6] *Anscombe, F. J. (1948). The Validity of Comparative Experiments. Journal of the Royal Statistical Society. Series a (General), 111(3), 181. https://doi.org/10.2307/2984159*.

[7] *Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. The Lancet, 383(9921), 999–1008. https://doi.org/10.1016/s0140-6736(13)61752-3.*

[8] *Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7(5), e623. https://doi.org/10.7717/peerj-cs.623.*

[9] *Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran, J. J. (2014). Essentials of Statistics for Business and Economics + Lms Integrated for Cengagenow, 1-term Access for Business and Economics. South-Western Pub.*