The Effect of Years of Education on Total Annual Sales of Businesses in Somalia

Arnur Akhat Aldiyar Rustem
Zhalgas Kabidolda Diyana Orazayeva Assem Kenzhetayeva
7272

April 28, 2025

Abstract

This study investigates the effect of years of education of business owners on total annual sales in Somalia by using data from the 2019 World Bank survey, which includes 491 businesses in Mogadishu and Bosaso. We use multiple regression models, including log-transformed sales and various control variables such as owners' age, number of owners, experience, number of workers, special offers, and total assets. The final model involves squared terms and interaction effects and explains 39.7% of the variation in sales. The results show that each additional year of education increases sales by approximately 3.55%. These findings suggest that improving access to education can support business performance in developing economies as Somalia.

1 Introduction

This research project aims to investigate how years of education affect the annual sales of companies for all products and services in Somalia. We used the Somalia dataset, which was taken from the survey of business owners from the World Bank Open Data. Lack of infrastructure, furnishings, educational resources, and teachers prevented the advancement of education in rural regions of the country. According to UNESCO, about 3 million children in Somalia do not have the opportunity to attend school.[1] Therefore, we might assume that the Somali people have a low educational level. The existing literature supports the idea that economic progress depends on education. According to Taleb Da Costa (2021), there is a positive and statistically significant effect between literacy rate and human development index in low-income countries and a positive effect between years of education and GDP per capita and labor force employment. This shows that education is an essential factor in the development of the economy, which coincides with our hypothesis that the level of education affects the economy. According to Chiliya and Roberts-Lombard (2012), the age of the owner and the educational level showed a positive effect on the business performance. This study shows that it is reasonable to compare the effect of the same variables such as the age of the owner, the level of education, and the experience on business profitability in Somalia. According to Warsame (2023), fixed assets or technology is one of the important factors that can influence sales growth. This study revealed that the gender of top management does not have a significant effect on sales growth, that is why our model does not include the effect of gender, but does include assets on sales growth. This study demonstrates the importance of education for raising sales and overall expansion of the economy.

2 Data Description

The survey was conducted in 2019 by Global Indicators Department (2020). It was carried out in two cities in Somalia: Bosaso and Mogadishu. These cities are the biggest in Somalia. Mogadishu, the capital of Somalia, has a population of 2,3 million people, which is about 1/8 of the whole population of Somalia. Bosaso has a population of 600 thousand people. Mogadishu serves as an important port, connecting traders across the Indian Ocean. According to Janzen J.H. (2020), the economy of Somalia is dominated by small businesses that rely on livestock, agriculture, and trade. The author states that the country does not rely on manufacturing, and commodities are fulfilled by small businesses. Thus, analyzing the small businesses would allow us to analyze the structure and economy of Somalia. Therefore, the survey represents the environment and characteristics of businesses and their owners in an environment where the economy relies on small

[1] ascanbeseenonhttps://www.unicef.org/somalia/education

businesses. The data set contains information on the business environment and various aspects of business operations. It contains data on business ownership, management practices, sales performance, number of employees, number of assets, and other relevant operational indicators. Although the data set offers valuable information for analysis, it also includes certain variables that may not be relevant to the research focus. The survey has 7747 business responses, of which 491 responses were randomly selected. The data set has 491 observations and 231 variables. Our dependent variable is sales, we sorted all the variables that could affect the sales. As we have shown in the introduction, we assumed that education and age of the owner would affect the sales. Additionally, we picked the experience of the owner, the number of workers, and the fact that the owner provided special offers or not. These are our dependent variables, we analyzed them below.

We chose total sales as the dependent variable. This variable represents the total annual sales reported by each business. In some cases, respondents either answered “Don’t know” or refused to provide their sales. These type of responses were treated as missing values in the dataset. Totsale represents the success of the business and it is key a value to evaluating the connection between sales results and educational attainment.

Our main independent value is educ. It is the education of the owner of the business, and it is measured by years, including years from school and university. Initially, it was written as options: ”No School”, ”Primary School(1-4)”, ”Intermediate School(4-9)”, ”Secondary School(9-12)”, ”Undergraduate degree”, ”Vocational Training/Diploma”. We considered ”No School” as 0 years of education, ”Primary School(1-4)” as 4 years of education, ”Intermediate School(4-9)” as 9 years of education, ”Secondary School(9-12)” as 12 years of education, ”Undergraduate degree” as 16 years of education, ”Vocational Training/Diploma” as 14 years of education. Additional control variables include:

- **owner_age**: Represents the age of the business owner. As discussed in the introduction, we expect a positive relationship between age and sales, under the assumption that older individuals possess greater managerial experience and decision-making skills, which can enhance business performance.

- **owners_num**: Represents the total number of owners involved in the business. We hypothesize a positive effect on sales because multiple owners may contribute diverse skills and resources, thereby improving the overall efficiency and output of the business, holding other variables constant.

- **owner_experience**: Represents the number of years the owner has been engaged in business activities. Greater experience is assumed to enhance managerial effectiveness and operational efficiency, which should positively influence total sales. This assumption was outlined in the introduction.

- **workers_num**: Represents the total number of individuals working in the business, including both paid and unpaid workers as well as the owner. We expect that a larger workforce allows the business to carry out more activities and serve more customers, thus contributing positively to sales.
- **total_assets**: Shows the value of the business's physical assets, including tools and utensils, machinery and equipment, and transportation-related assets. Based on the literature review in the introduction, we expect a positive relationship between total assets and sales, as higher asset levels may improve production capacity and service quality.
- **special_offer**: A dummy variable that shows whether the business implemented any promotional offers. We assume that the presence of special offers increases consumer demand, thereby boosting sales.

While creating a new dataset with the necessary variables, we replaced invalid responses, such as negative values or empty spaces, to ensure data quality.

3 Methodology and Results

To analyze how a business owner's education level affects total annual sales, we developed a statistical model with the following hypotheses:

- Null hypothesis (H_0): Education has no effect on annual sales
- Alternative hypothesis (H_1): Education has a significant effect on annual sales

Initial Regression Model At first, we ran a basic linear regression using only education, as an independent variable and total annual sales as a dependent variable. We tested our hypothesis using a simple regression, where we examined how owner's education level relates to total annual sales. The results showed an R-squared of 0.0957, meaning that owner's education alone explains about 9.57% of the variation in annual sales. Despite being a basic model, the coefficient for educ was statistically significant, with a t-value equal to 6.73.

Then we began with a basic linear regression:

$$\begin{aligned}\widehat{\text{totsale}} = \beta_0 &+ \beta_1\text{educ} + \beta_2\text{workers\_num} + \beta_3\text{owners\_number}\\ &+ \beta_4\text{owner\_age} + \beta_5\text{owner\_experience}\\ &+ \beta_6\text{special\_offer} + \beta_7\text{total\_assets}\end{aligned}$$

We selected these variables, workers_num, owners_number, owner_age, owner_experience, special_offer, and total_assets, based on data from literature review and assumptions that we made during analysis of dataset.

The model has an R-squared of 0.2801, indicating that approximately 28.01% of the variation in annual sales is explained by the included variables. The model is statistically significant overall, with F(7,403) = 11.35. Therefore, this model explains better the dependent variable, totsale.

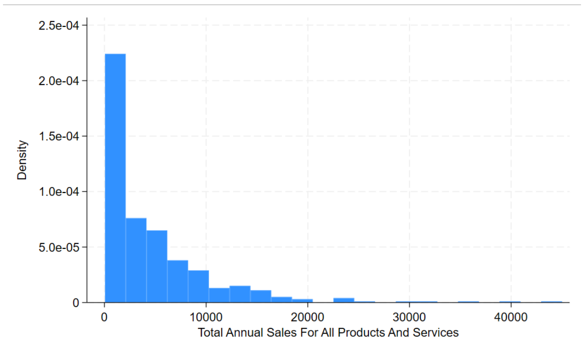

Figure 1: Histogram of *totsale*

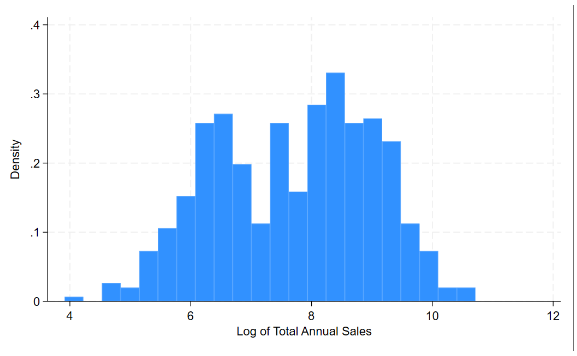

Figure 2: Histogram of *logtotSale*

Log Transformation Financial data such as sales is usually very skewed. Some businesses earn much more than others. It can be seen from Figure 1. Total annual sales for most of the businesses is about 0. That is why we used a log transformation on the sales variable to make the results more meaningful. Figure 2 shows how taking the logarithm of total annual sales compresses extreme values and produces a more symmetric distribution. This transformed variable, logtotSale, became our new dependent variable. Log transformation is efficient because it turns differences into percentage effects. We then re-ran the multivariable regression, this time with log-transformed annual sales as the dependent variable and educ, workers_num, owners_number, owner_age, owner_experience, special_offer, and total_assets as the independent variable, to see how the relationship changed with the new scale. It turned out that taking logarithm of sales has a positive effect on R-squared, which was equal to 0.2801, became 0.3204. After the log transformation the constant term and coefficient of education significantly changed. Constant term dropped from -2200.40 to 6.21. Coefficient for education was 216.59, meaning that each additional year of education was associated with an increase of about 217 units in annual sales. After the transformation, the coefficient dropped to 0.0429, which now implies a 4.29% increase in sales for each additional year of education. This shift reflects the change from absolute effects in the original model to relative (percentage-based) effects in the log-transformed model. These new values are much easier to interpret and makes the regression results more manageable. Overall, using the logged version of the dependent variable gave us a better-fitting model that is easier to interpret and has higher

R-squared.

Then we built a regression model that includes age squared and experience squared. Including squared terms for age and experience allows the model to show non-linear relationships, where the effect increases up to a point and then decreases. As a result, the model provides a more accurate and realistic representation of their impact on business performance. After including these controls, the model's explanatory power improved significantly. The R-squared increased from 0.3204 to 0.3886, showing that the agesq and expsq control variables helped better explain the variation in annual sales.

Finally, we introduced an interaction term between total assets and owner experience, `assetexp`. This interaction helps us to explore whether experienced owners benefit more from having higher assets, or whether the impact of assets depends on how experienced the owner is. The model explains 39.7% of the variation in logged sales and is statistically significant overall because $F(10, 400) = 27.77$, $p < 0.001$. Several variables remain significant. Education continues to positively impact sales, with a coefficient of 0.0355, suggesting that each additional year of education is associated with a 3.55% increase in sales. The interaction between assets and experience is significant. This means that the combination of having more assets and being experienced leads to even better sales performance.

Final linear regression:

$$\begin{aligned}\widehat{\log(\text{totsale})} = {} & \beta_0 + \beta_1 \text{educ} + \beta_2 \text{owner\_age} + \beta_3 \text{workers\_num} \\ & + \beta_4 \text{owner\_experience} + \beta_5 \text{special\_offer} \\ & + \beta_6 \text{total\_assets} + \beta_7 \text{agesq} + \beta_8 \text{expsq} + \beta_9 \text{assetexp}\end{aligned}$$

Diagnostic Tests

After finalizing our regression model, we decided to complete a diagnostic tests to check whether the model meets the key assumptions of linear regression. These checks are important because if the assumptions are not satisfied, the results can be misleading. This would lead to incorrect interpretations and conclusions.

1. Multicollinearity

We checked for multicollinearity to see whether any of the independent variables in our regression model were too closely related to each other. High multicollinearity can cause problems by inflating standard errors and making the coefficient estimates unreliable. To assess this, we used the Variance Inflation Factor (VIF), a common diagnostic tool that shows how much the variance of a regression coefficient is increased due to multicollinearity. Generally, a VIF value above 10 suggests a serious multicollinearity problem, while values below 5 are considered acceptable. In our model, all VIF values were above the threshold, with the highest being over 50, which confirms that multicollinearity is found in our model. This form of multicollinearity is considered standard and not problematic, as both terms are necessary to represent potential non-linear effects.

Variance Inflation Factors (VIF):

Variable	VIF
agesq	52.06
owner_age	51.08
expsq	5.48
owner_experience	5.39
assetexp	3.77
total_assets	2.79
workers_num	1.29
educ	1.27
special_offer	1.14
owners_number	1.11
Mean VIF	12.54

Table 1: Values of VIF for each regressor

2. T-test

The t-test is used to assess whether the coefficient of a specific explanatory variable is statistically different from zero. In the Model 6, we test the null hypothesis that the coefficient on the *education* variable is equal to zero against the alternative that it is not. The resulting t-statistic is approximately 3.5, and the associated p-value is less than 0.001. Since the p-value is far below the conventional significance levels, such as 0.05 or 0.01, we reject the null hypothesis. This provides strong evidence that the education variable has a statistically significant effect on *logtotSale*, and that its effect is not due to random chance.

3. F-test

The F-test for Model 6 yields a statistic of 34.70 with a p-value less than 0.001. This very small p-value allows us to confidently reject the null hypothesis that all slope coefficients are equal to zero. Therefore, we conclude that the model is statistically significant overall. This means that at least one of the explanatory variables in Model 6 has a meaningful effect on the dependent variable, *logtotSale*. The F-test confirms that the chosen set of predictors collectively explains a significant portion of the variation in sales.

4. Joint significance F-test

In addition to the typical F-test, which evaluates all slope coefficients, we also conducted a joint significance F-test on the *owner_age* and $owner\_age^2$ as they were insignificant at the T-test level across all models (see Table 2). Results show that jointly they are statistically insignificant, being p-value = 0.878. Consequently, it is better to drop these two variables as they don't serve any purpose in explaining the independent variable, `logtotSale`.

5. Heteroskedasticity

To assess whether our model suffers from heteroskedasticity, we conducted the Breusch-Pagan test. This test involves regressing the squared residuals on the independent variables in their original linear form to determine whether the variance of the residuals is constant. The results of the BP test yielded an F-statistic of 1.93 with 8 and 402 degrees of freedom, and a corresponding p-value of 0.0539. This p-value is slightly above the conventional 5% significance level but below the 10% level. Therefore, we interpret this as weak evidence of heteroskedasticity in the model. While we do not strongly reject the null hypothesis of homoskedasticity at the 5% level, the closeness of the p-value to the threshold suggests that the assumption of constant variance does not not hold strictly.

In addition to the Breusch–Pagan test, we performed a special case of the White test to further investigate the presence of heteroskedasticity. This involved regressing the squared residuals on the fitted values and their squares. The results confirmed the presence of heteroskedasticity, as both the fitted value and its square were statistically significant at the 1% level, with p-values of 0.000. The negative coefficient on the fitted value, –5.19, and the positive coefficient on its square, being equal to 0.314, suggest a U-shaped pattern in the variance of the residuals. This implies that the variance initially decreases and then increases as the fitted values rise, further supporting the presence of non-constant error variance in the model.

	totsale	*totsale*	*logtotSale*	*logtotSale*	*logtotSale*	*logtotSale*
Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
education	324.903***	216.593***	0.043***	0.038***	0.036***	0.035***
	(48.284)	(46.938)	(0.011)	(0.010)	(0.010)	(0.010)
owner's number		385.298	-0.014	0.028	0.015	0.014
		(769.227)	(0.121)	(0.113)	(0.114)	(0.114)
owner's age		-11.122	-0.001	0.037	0.023	—
		(38.148)	(0.008)	(0.054)	(0.052)	
owner's experience		355.066***	0.114***	0.284***	0.277***	0.280***
		(98.819)	(0.020)	(0.039)	(0.039)	(0.039)
special offer		1931.294***	0.442***	0.402***	0.439***	0.437***
		(467.257)	(0.124)	(0.120)	(0.121)	(0.121)
total assets		-0.178*	0.000	0.000	0.000**	0.000***
		(0.096)	(0.000)	(0.000)	(0.000)	(0.000)
workers number		1428.171***	0.271***	0.259***	0.268***	0.268***
		(529.129)	(0.078)	(0.073)	(0.072)	(0.072)
owner's age^2				0.000	0.000	—
				(0.001)	(0.001)	
owner's experience^2				-0.010***	-0.011***	-0.011***
				(0.003)	(0.003)	(0.003)
assets×*experience*					0.000***	0.000***
					(0.000)	(0.000)
Constant	1958.158***	-2208.397	6.206***	5.105***	5.447***	5.893***
	(326.854)	(2084.937)	(0.349)	(0.974)	(0.958)	(0.221)
R-squared	0.096	0.280	0.320	0.389	0.397	0.397
F-statistic	45.280	11.350	16.890	24.190	27.770	34.700
Prob > F	0.000	0.000	0.000	0.000	0.000	0.000
Observations	489	411	411	411	411	411

Table 2: Regression Results. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; robust standard errors in parentheses. Models 1 and 2 use total sales as an independent variable, while Models 3-6 use log-transformed version of the response variable as indicated in the first row of the table.

3.1 Discussion of Regression Results

Table 2 shows the results of six regression models examining the factors affecting business performance, measured in both total annual sales, *totsale*, and its logarithmic transformation, *logtotSale*.

In Models 1 and 2, where *totsale* is the dependent variable, we observe that education has a strong and statistically significant positive effect on sales, with a coefficient of 324.903 in Model 1 and 216.593 in Model 2. However, the fit of the model is relatively low in Model 1, $R^2 = 0.096$, and improves significantly in Model 2, being $R^2 = 0.280$, once additional variables are included. Despite the improvements, the non-log-transformed models are sensitive to extreme values because the sales data is skewed.

To solve this problem, Models 3 through 6 use the natural logarithm of total sales, *logtotSale*, as the dependent variable. This transformation reduces the skewness and allows us to interpret the coefficients as percentage changes. These models show significantly improved explanatory power, with R^2 increasing from 0.320 in Model 3 to 0.397 in Models 5 and 6.

Across Models 3 to 6, education remains a positive and statistically significant factor for business performance, with consistent coefficients around 0.035–0.043. This suggests that higher education among business owners is linked to about 3.5%–4.3% higher sales, holding other factors constant.

Initially, we assumed that the owner's age would contribute to the business's sales. However, neither the owner's age nor its quadratic term shows a significant impact on sales. This suggests that, in our sample, entrepreneurial success is not related to the owner's age, but rather to other factors.

Owner's experience is also a strong predictor, with a positive linear effect and a negative quadratic term. The significance of both the linear and squared terms shows that there are diminishing returns to experience: while more years of experience initially increase sales, the effect decreases after a certain point. This pattern is consistent in all models.

Special offers have a clear and large positive effect on sales, with coefficients ranging from 0.402 to 0.442, all significant at the 1% level. This shows that businesses that use promotional strategies perform much better in terms of revenue.

The number of workers employed also has a positive and statistically significant effect on log-transformed sales, potentially reflecting economies of scale or increased labor productivity.

Total assets show a small but statistically significant negative relationship with sales in Models 5 and 6, once interaction effects are added. Interestingly, the interaction term between total assets and experience is positive and highly significant, meaning that the negative effect of assets could be reduced by the owner's experience. This interaction effect highlights the importance of considering combined effects when analyzing business performance. For example, according to our model an owner with 10 years of experience and $100,000 in assets generates 0.81% higher sales compared to a less experienced peer with identical assets.

Among the six regression models estimated, Model 6 is our preferred model. It has the highest explanatory power, with an R^2 of 0.397, indicating that approximately 39.7% of the variation in logarithmic transformed sales is explained by the included variables. The log transformation corrects for the skewness in the sales variable. This model captures non-linear relationships and interaction effects, which provide more understanding of business dynamics. Using squared experience terms and interaction between assets and experience allows us to identify diminishing returns and conditional effects, which are empirically significant. Another reason we prefer this model is that it keeps only the statistically significant variables. This makes the results more reliable and reduces noise from variables that do not meaningfully affect sales. For these reasons, Model 6 offers the most accurate and informative way to analyze business performance in our dataset.

4 Conclusion

Our analysis demonstrates that level of education consistently has a positive and statistically significant effect on business total sales. The preferred model, Model 6, offers the highest explanatory power and includes only variables that are statistically significant. The model shows that each additional year of education is associated with a 3.55% increase in total annual sales. Other key factors of sales includes owner experience (with diminishing returns), number of workers, special offers, and an interaction between assets and experience. Interestingly, the age of the owner and its squared term does not have any significant effect, which can mean that it is not the age, but rather the accumulated experience and special offers in business, that have more effect on business sales.

Even though our analysis provides meaningful information, it is important to point certain limitations. First of all, the sample that we analyze is limited to 491 businesses located in Mogadishu and Bosaso, which may not fully show all businesses in Somalia, especially in rural or conflict-affected regions. This may lead to potential sample bias, as businesses in major cities may differ from those in less developed areas in terms of access to education, capital, and markets. Second, even though we include several key explanatory variables, our model may still suffer from omitted variable bias. Because factors such as access to credit, market competition, or the business owner's motivation and risk tolerance are not included in the dataset but they can also affect total sales of businesses.

From a policy perspective, our findings show that level of education and experience play a significant role in increasing business performance in Somalia. Encouraging access to education may significantly improve profitability of businesses. Moreover, training programs focused on effective asset management and the use of special offers as type of promotions could further improve small business performance.

Future study could further analyze gender differences, access to finance, and the role of digital tools in sales performance to build an expanded model for business policy in developing countries based on Somalia.

References

Chiliya, N., & Roberts-Lombard, M. (2012). Impact of level of education and experience on profitability of small grocery shops in south africa. *International Journal of Business Management and Economic Research*, *3*(1), 462–470.

Global Indicators Department, E. A. U. (B. (2020). Informal sector business survey 2019 [data set]. *World Bank, Development Data Group.*

Janzen J.H., L. I. (2020). Somalia. *Encyclopedia Britannica.*

Taleb Da Costa, M. (2021). An econometric study of the impact of education on the economic development of low-income countries.

Warsame, A. S. (2023). Factors influencing firm sales growth: An instrumental variable analysis. *International Journal of Marketing Studies*, *15*(2), 51.