Factors Affecting the Current Capital of MSEs: A Multiple Linear Regression Approach

Authors: Natnael G. Mengistu*, Yoseph H. Tilahun*, Yonas D. Melese*, Abenezer Y. Bekele*, Fitsum A. Zewude*, Rahel D. Shikur*

Context: Micro and Small Enterprises (MSEs) play a vital role in the economic development of Ethiopia. They contribute significantly to employment creation, poverty reduction, and overall economic growth, particularly in urban and semi-urban areas. Understanding the factors that influence the performance and growth of MSEs is therefore crucial for both policymakers and business practitioners.

Core project task: Regressing current capital on initial capital, years since establishment, years of schooling of owner(s), number of paid workers, gender, having a business license, access to a loan from micro finance institution, and access to land/working premises.

1. Introduction

Micro and Small Enterprises (MSEs) play an important role in Ethiopia's economic system. Gebreeyesus, M. (2011) puts MSEs as the heart of Ethiopia's industrial sector. Understanding what factors contribute to the success and growth of MSEs is really important. One way to measure success of such firms is by understanding factors that affect the current capital MSEs.

In this project, we explore the effect of several characteristics such as initial capital, years of establishment, education level of the owner, access to loans, number of paid workers, ownership of business licenses, and gender of owner on the current capital of MSEs and how they relate to it.

2. Objectives

The main objective of this project is to investigate factors affecting the current capital of Micro and Small Enterprises (MSEs). This is expected to contribute to our understanding of MSE success and potentially help owners see where they are and where they can be with few improvements in the right direction.

Specific objectives include the following:

- Build a multiple linear regression able to explain current capital by using other independent factors
- See if the owner's educational background and years of schooling are strong determinants of current capital.
- Analyze current capital as opposed to the gender of the owner.
- Explore sub-sectors of MSEs

3. Methods

3.1 Data Source

We are using the data that was provided to us by our instructor. It includes 15 columns in addition to the column

*Undergraduate Student, Department of Statistics, College of Natural and Computational Sciences, Addis Ababa University, Ethiopia.

"current_capital", each with 860 non-empty entries. See Appendix A for a detailed data dictionary.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 860 entries, 0 to 859
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   business_registered  860 non-null    category
 1   TIN                  860 non-null    category
 2   business_license     860 non-null    category
 3   subsector            860 non-null    category
 4   MSE_type             860 non-null    category
 5   loan                 860 non-null    category
 6   land_premise         860 non-null    category
 7   one_stop_service     860 non-null    category
 8   util_rate            860 non-null    int64
 9   current_capital      860 non-null    int64
 10  initial_capital      860 non-null    float64
 11  gender               860 non-null    category
 12  years_establish      860 non-null    int64
 13  years_schooling      860 non-null    int64
 14  no_paid_workers      860 non-null    int64
 15  highest_education    860 non-null    category
dtypes: category(10), float64(1), int64(5)
memory usage: 50.3 KB
```

Figure 3.1: Data information

3.2 Data Exploration and Hypothesis Testing

We used Python and R languages with libraries like "seaborn", "matplotlib", "plotly" (for python) and "ggplot2", "dplyr" (for R) to explore data with plots and summaries.

Since our data was made up of many categorical variables, we used chi-square tests to see if two variables are associated or not. When needed, we converted the "current_capital" column to categorical using appropriate data recoding. These chi-square tests were followed by post-hoc standardized residuals analysis whenever we found potential associations to see exactly how these associations played out.

3.3 Multiple Linear Regression

As suggested by our instructor, we focused on regressing current capital on initial capital, years since establishment, years of schooling of owner(s), number of paid workers, gender, having business license, access to loan from micro finance institution, and access to land/working premises.

We explored three models in total.

1. Initial model (first model): fitted using ordinal least squares (OLS) on the variables mentioned above. We used this to check assumptions and guide us in the next direction.
2. Log transformed model (second model): fitted with log transformed dependent variable and other transformed independent variables. We also used OLS here and checked assumptions again.
3. Weighted least squares model with log transformation (third model): fitted using weighted least squares instead of OLS as suggested in lecture notes and text (Zelalem T., 2025, and Kutner et al., 2005) during the presence of heteroscedasticity. We used the same functional form and transformations from our second model.

4. Analysis Results

4.1 Exploratory Data Analysis and Variables Association Exploration

The very first thing we noticed when we started exploring the data was how "current_capital" was extremely skewed (figure 4.1). This signaled to us from the start a possible transformation on our dependent variable was a good idea.

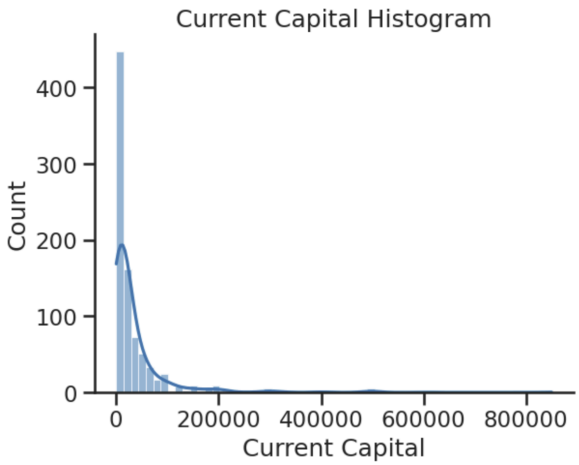


Figure 4.1: Current capital distribution

This type of high skewness was also observed in many of the other numerical variables including initial capital and years since establishment.

We were also interested in exploring the highest level of education vs gender. We plotted the multi-bar chart seen in figure 4.2 just for this purpose.

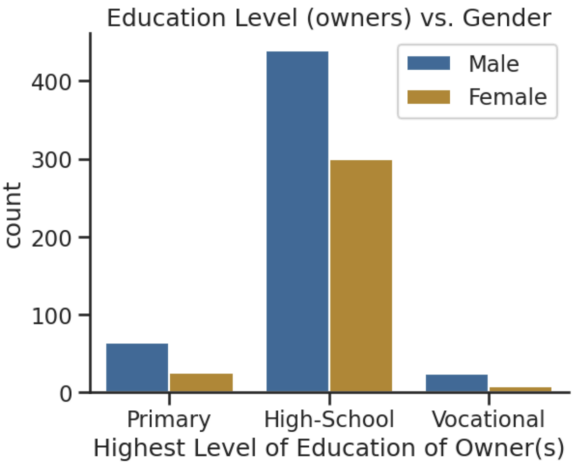


Figure 4.2: Education level of owner(s) vs gender

It can be seen that at each education level the number of females in that category is lower than that of males.

Next, we explored the composition of MSEs by sub-sector (figure 4.3).

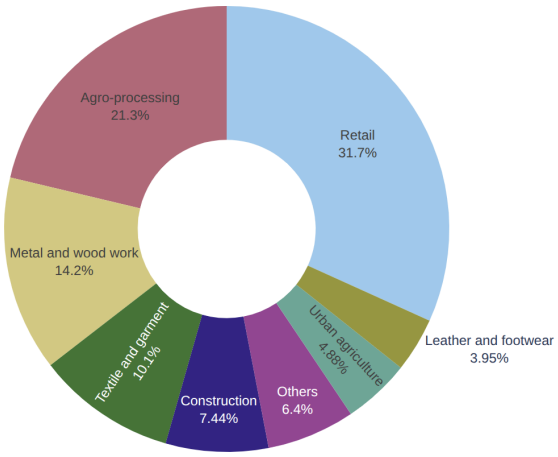


Figure 4.3: MSEs by sub-sector

Leading the way is the retail sector with 31.7% of the MSEs being in its category, followed by agro-processing at 21.3%. Leather and footwear and urban agriculture are at the bottom with 3.95% and 4.88% of the MSEs belonging in these sub-sectors, respectively.

To explore association between current capital, gender, and levels of education we decoded current capital to a categorical variable using its four quartiles, ranging from low to high. We then proceeded with chi-square tests as discussed in the methods section of this paper. We found the results displayed in table 4.1.

Current Capital	**Chi-square statistic**	**p-value**
vs. Gender	56.90504500890765	2.692663875235754e-12
vs. Highest Education of Owner(s)	17.94179115735739	0.006379320255157455

Table 4.1: Results of Chi-square

In both cases, the p-values are lower than 0.05. We followed post-hoc analysis by looking at the standardized residuals from observed and expected crosstabulations used in the chi-square tests. Figure 4.4 shows the results for the post-hoc analysis as heatmaps.

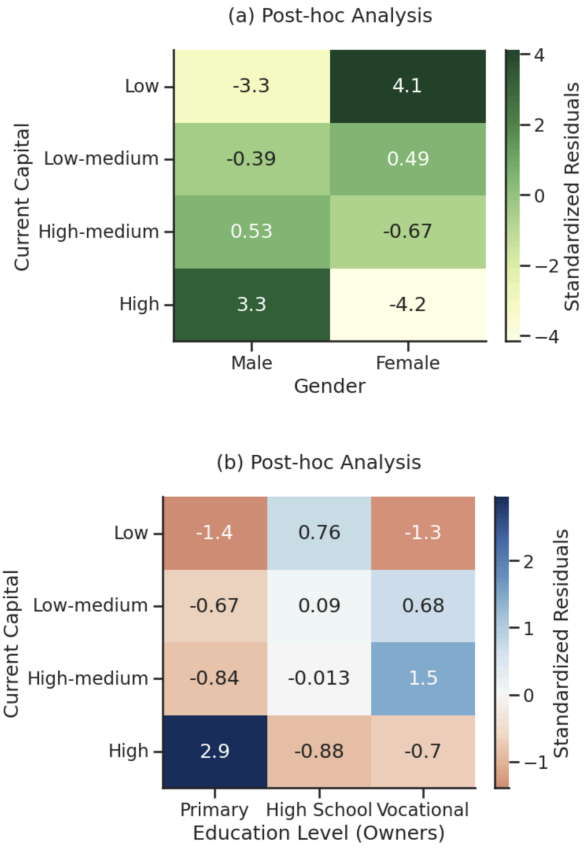


Figure 4.4: (a) (Above) post-hoc analysis of the association between current capital and gender. (b) (Below) post-hoc analysis of the association between current capital and highest education level of owner(s)

Figure 4.4 (a) shows there is more than the expected number of males in the high current capital category. In the case of females, it is exactly the opposite, with more than expected number of females found in the lowest current capital category and lower than expected found in the high current capital section. Figure 4.4 (b) shows an interesting result where there are more than expected owners in the high current capital category with primary level of education.

4.2 Multiple Linear Regression

4.2.1 Initial Full Model (1st model)

Model formula:

Current Capital$_i$ = β_0 + β_1 **Initial Capital**$_i$ + β_2 **Years Established**$_i$ + β_3 **Years of Schooling**$_i$ + β_4 **Number of Paid Workers**$_i$ + β_5 **Business License**$_i$ + β_6 **Loan**$_i$ + β_7 **Land Premise**$_i$ + β_8 **Gender**$_i$ + ε_i

This model turned out to have an r-square and adjusted r-square values of 38.5% and 37.9% respectively. (See Appendix B for further detail.) We primarily used this model to check assumptions and point us in the right direction. While we found assumptions of no autocorrelation and no multicollinearity in the model as seen by the results in table 4.2.1, q-q plots of residuals and residuals vs fitted values plots of figure 4.2.1 tell a different story when it comes to normality of errors and homoscedasticity.

Auto-correlation test	Result
Durbin-Watson	1.820738953540482
Breusch-Godfrey	p-value = 0.0202
Multi-collinearity test	
VIF	All independent variables < 1.14
Hetroscedasticity test	

Breusch-pagan	p-value = 0.0005

Table 4.2.1: Results of assumption tests

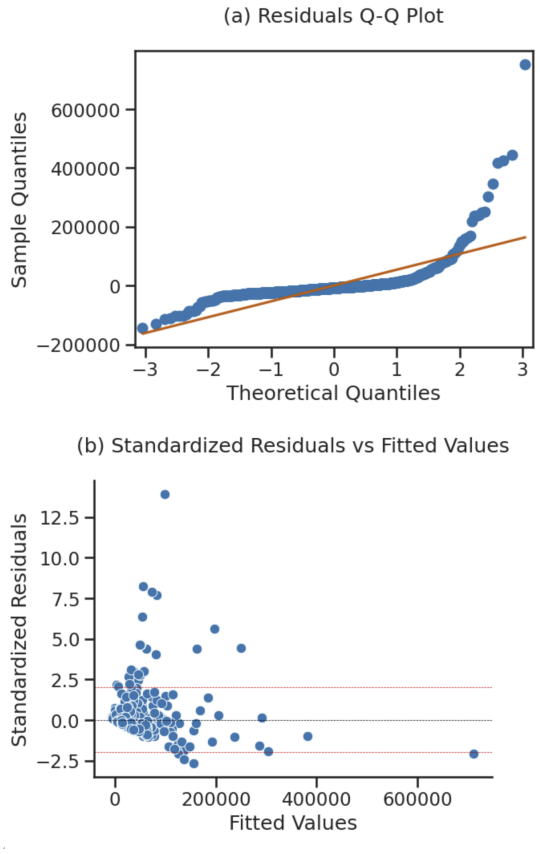


Figure 4.2.1: (a) (Above) Q-Q plot of residuals. (b) (Below) Standardized residuals vs fitted values

Following results from our tests of assumption on the initial model and our observation of the skewness exhibited in many of the variables, we fitted a second model with appropriate transformations applied.

4.2.2 Log Transformed Model (2nd model)

Model formula:

log(Current Capital$_i$) = β_0 + β_1 **log(Initial Capital**$_i$) + β_2 **log(Years Established**$_i$) + β_3 **Years of Schooling**$_i^2$ + β_4 **log(Number of Paid Workers**$_i$) + β_5 **Business License**$_i$ + β_6 **Loan**$_i$ + β_7 **Land Premise**$_i$ + β_8 **Gender**$_i$ + ε_i

See Appendix C for before and after transformation effects of the dependent and some independent variables.

The r-square and adjusted r-square of the transformed model improved to 59.6% and 59.2%, respectively. Durbin-Watson statistic stood at 1.82 and the condition number was found to be 759. Go to Appendix D for detailed results on the summary of this model. Variance Inflation Factor (VIF) for all independent variables is below 1.23.

The Q-Q plots of the residuals and the standardized residuals vs fitted values plots (figure 4.2.2) this time tell a new story.

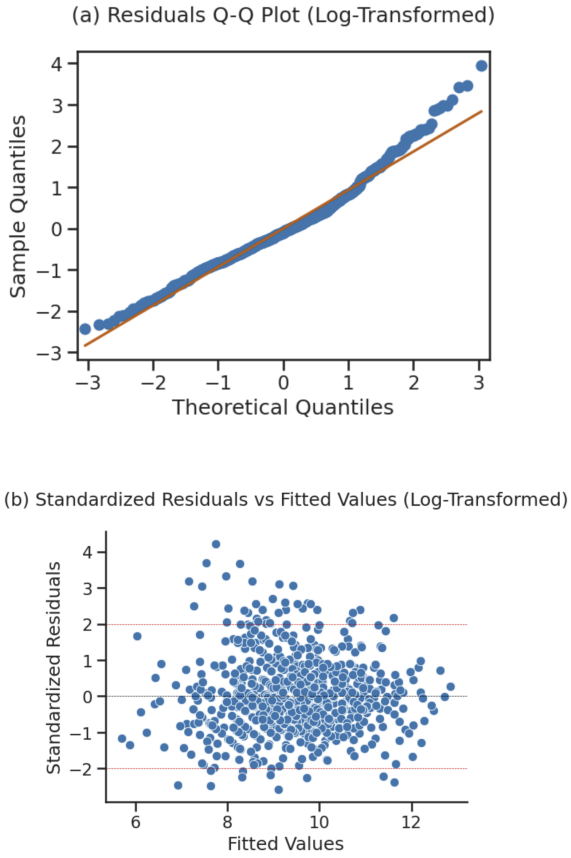


Figure 4.2.2: (a) (Above) Q-Q plot of residuals from our second model. (b) (Below) Standardized residuals vs fitted values from our second model

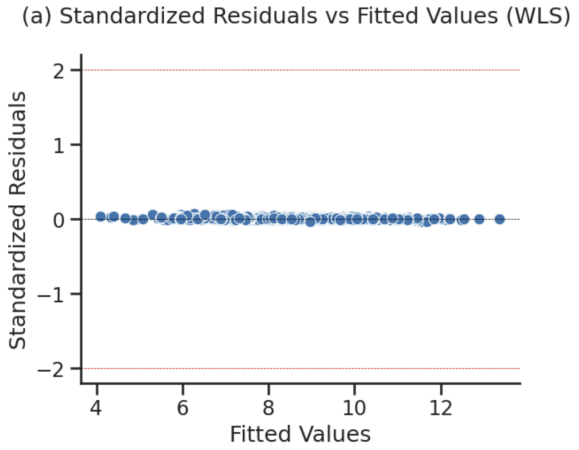
Despite improvements on the residuals vs fitted values, we still can see points outside the (+2, -2) line and Breusch-Pagan test gave a p-value of 0.0000. These results pushed us into fitting a weighted least square model with the same model formula.

4.2.3 Weighted Least Squares Model with Log Transformation (3rd model)

The model formula is the same with our second formula. The weights were found by using the steps suggested in our lectures (Zelalem T., 2025). We fitted the squared residuals on our predictors and then used the reciprocal of the fitted values from this model as our weights.

R-square and adjusted r-square jumped to 81.6% and 81.4%, respectively. Durbin-Watson statistic was found to be 1.966, with all VIF values for the independent variables below 1.23. (See Appendix E for further details.)

Figure 4.2.3 shows the new standardized residuals from our WLS model vs fitted values and the partial autocorrelation plot of the same residuals.



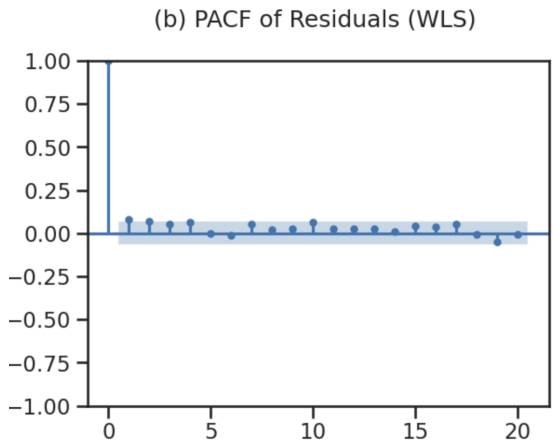


Figure 4.2.2: (a) (Above) Standardized residuals vs fitted values from our third model. (b) (Below) Partial autocorrelation plot of residuals from our third model.

5. Discussions

There is a potential gap in the education level between female and male MSE owners. For each education level, the number of females in that level was lower than males. While this might be the result of education level differences between the two sexes, it could also have come as a result of there being fewer female MSE owners in this particular data than males.

Retail and agro-processing are the most dominant sub-sectors in the MSEs we explored. We can also see there is a clear distinction between the two genders and the current capital of MSEs. There are more than the expected number of males with high current capital, while there are more than the expected number of females with low current capital. This is not due to there being fewer female owners than males, as we are comparing observed counts with expected ones.

There is also an association between the highest level of education of the owner(s) and current capital. More than the expected number of owners with primary

education have a high current capital. We can't, however, say that current capital becomes less as education level increases as there are more education levels missing in the data such as university graduates.

The initial model we fitted violated the assumptions of normality of errors as well as homoscedasticity. Both show improvements when we use log transformations in our second model. Hetroscedasticity was still present. Our third and final model solved for this problem by using weighted least squares. It also managed to explain about 81.6% of the variability in log current capital. All three models didn't violate assumptions of no autocorrelation and no multicollinearity. The interpretations following next are based on our third model.

Initial capital is the strongest predictor of current capital, with a 1% increase in log initial capital being associated with a 0.91% increase in log current capital keeping all other factors constant. Female owned MSEs have about 14.6% lower current capital than male owned MSEs, holding other factors constant. The positive coefficients for the square years of schooling point to additional schooling having a larger positive effect on current capital. Having business licenses and loan access from microfinance institutions actually affect current capital negatively, while the number of paid workers is positively related.

The logarithm of years since establishment and the receiving of land/premises are found to be statistically not significant.

6 Conclusion

Initial capital is the strongest predictor of current capital, while female owners tend to have lower current capital than male ones. We would like to explore why this is so in future studies.

Years since establishment and the receiving of land/premises were not found to be statistically significant.

Owners with a primary level of education as their highest level of education appear to have higher current capital than expected. When it comes to years of schooling, it is actually positively related with current capital.

Most MSEs are in the retail and agro-processing sub-sectors.

References

Gebreeyesus, M. (2011). Industrial policy and development in Ethiopia: Evolution and present state. *Africa Growth Initiative Working Paper*, 6. Washington, DC: Brookings Institution.

Kutner, M. H., Nachtshein, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (5th ed.). McGraw-Hill.

Zelalem, T. (2023). *Lecture note: Econometrics (Stat4061)*. Addis Ababa University.

Appendices

Appendix A: Data Dictionary

Variable	description	Coding
business_registered	Is the business registered?	0 = No, 1 = Yes
TIN	Do you have TIN?	0 = No, 1 = Yes
business_license	Do you have business license or certificate?	0 = No, 1 = Yes
MSE_type	Type of MSE	1 = Micro, 2 = Small
loan	Have you obtained loan from any micro finance institution?	0 = No, 1 = Yes
land_premise	Have you ever received land or working premises for your enterprise?	0 = No, 1 = Yes
one_stop_service	Did you receive one-stop service at the nearest administration?	0 = No, 1 = Yes
util_rate	Enterprise's current capacity utilization rate	Actual data
current_capital	Current capital (in birr)	Actual data
initial_capital	Initial capital (in birr)	Actual data
years_establish	Years since establishment	Actual data
years_schooling	Years of schooling of owner(s)	Actual data
no_paid_workers	Number of paid workers	Actual data
gender	Gender of owner	0 = Male, 1 = Female
highest_education	Highest level of education of owner(s)	1 = Illiterate, 2 = Read and 3 = Primary, 4 = High scho 5 = Vocational training, 6 = University degree
subsector	Sub-sector of MSE	1 = Metal and wood work 2 = Construction 3 = Agro-processing 4 = Textile and garment 5 = Leather and footwear 6 = Retail 7 = Urban agriculture subsector Sub-sector of MSE 8 = Others

Figure A.1: Data dictionary

Appendix B: Initial Model Results

OLS Regression Results

Dep. Variable:	current_capital	R-squared:	0.385
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	66.58
Date:	Fri, 23 Jan 2026	Prob (F-statistic):	1.12e-84
Time:	18:47:45	Log-Likelihood:	-10589.
No. Observations:	860	AIC:	2.120e+04
Df Residuals:	851	BIC:	2.124e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5221.3189	1.03e+04	0.506	0.613	-1.51e+04	2.55e+04
C(business_license)[T.1]	8061.7434	4838.307	1.666	0.096	-1434.670	1.76e+04
C(loan)[T.1]	2717.9583	4587.289	0.592	0.554	-6285.768	1.17e+04
C(land_premise)[T.1]	-1704.9100	4217.519	-0.404	0.686	-9982.869	6573.049
C(gender)[T.1]	-1.67e+04	3903.874	-4.278	0.000	-2.44e+04	-9038.310
initial_capital	1.3882	0.067	20.576	0.000	1.256	1.521
years_establish	1057.6029	321.640	3.288	0.001	426.303	1688.903
years_schooling	583.8037	906.773	0.644	0.520	-1195.969	2363.577
no_paid_workers	1497.7008	493.886	3.032	0.002	528.323	2467.079

Omnibus:	1068.625	Durbin-Watson:	1.821
Prob(Omnibus):	0.000	Jarque-Bera (JB):	145408.082
Skew:	6.282	Prob(JB):	0.00
Kurtosis:	65.450	Cond. No.	1.72e+05

Figure B.1: Initial OLS model summary

Appendix C: Effects of Transformation

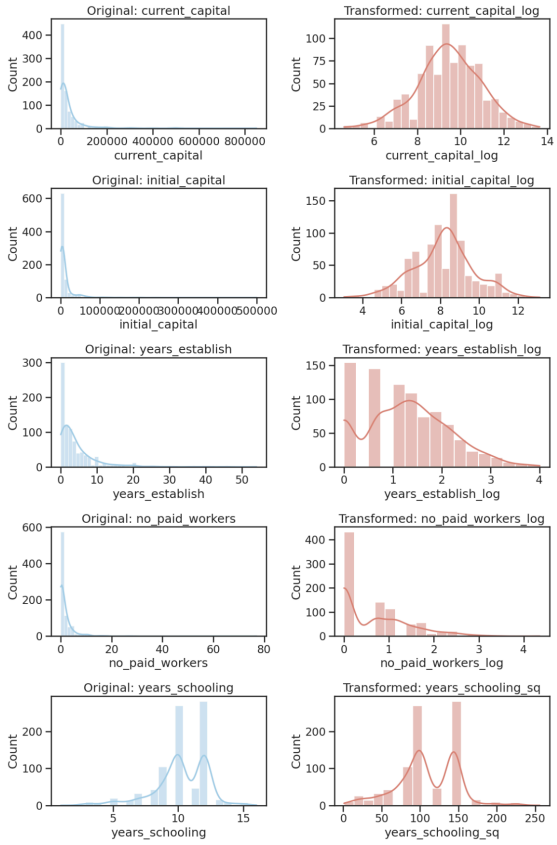


Figure C.1: Before and after transformation distributions of the dependent and some independent varaibles

Appendix D: Log Transformed Model results

OLS Regression Results

Dep. Variable:	I(np.log1p(current_capital))	R-squared:	0.596
Model:	OLS	Adj. R-squared:	0.592
Method:	Least Squares	F-statistic:	157.0
Date:	Fri, 23 Jan 2026	Prob (F-statistic):	8.28e-162
Time:	18:47:54	Log-Likelihood:	-1159.3
No. Observations:	860	AIC:	2337.
Df Residuals:	851	BIC:	2379.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.7557	0.208	18.058	0.000	3.347	4.164
C(gender)[T.1]	-0.3578	0.069	-5.196	0.000	-0.493	-0.223
C(business_license)[T.1]	0.1466	0.086	1.701	0.089	-0.023	0.316
C(loan)[T.1]	-0.1091	0.080	-1.370	0.171	-0.265	0.047
C(land_premise)[T.1]	-0.0009	0.073	-0.012	0.990	-0.144	0.143
I(np.log1p(initial_capital))	0.6291	0.022	28.584	0.000	0.586	0.672
I(np.log1p(years_establish))	0.3795	0.038	9.917	0.000	0.304	0.455
I(years_schooling ** 2)	-0.0004	0.001	-0.478	0.633	-0.002	0.001
I(np.log1p(no_paid_workers))	0.2285	0.046	4.940	0.000	0.138	0.319

Omnibus:	70.507	Durbin-Watson:	1.820
Prob(Omnibus):	0.000	Jarque-Bera (JB):	97.714
Skew:	0.647	Prob(JB):	6.05e-22
Kurtosis:	4.025	Cond. No.	759.

Figure D.1: Log transformed model summary

Appendix E: Weighted Least Squares with Log Transformation Results

WLS Regression Results

Dep. Variable:	I(np.log1p(current_capital))	R-squared:	0.816
Model:	WLS	Adj. R-squared:	0.814
Method:	Least Squares	F-statistic:	472.2
Date:	Fri, 23 Jan 2026	Prob (F-statistic):	7.69e-307
Time:	20:02:24	Log-Likelihood:	-4590.2
No. Observations:	860	AIC:	9198.
Df Residuals:	851	BIC:	9241.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3678	0.232	5.891	0.000	0.912	1.824
C(gender)[T.1]	-0.1455	0.032	-4.487	0.000	-0.209	-0.082
C(business_license)[T.1]	-0.1191	0.036	-3.267	0.001	-0.191	-0.048
C(loan)[T.1]	-0.1100	0.064	-1.731	0.084	-0.235	0.015
C(land_premise)[T.1]	-0.0121	0.038	-0.318	0.750	-0.086	0.062
I(np.log1p(initial_capital))	0.9074	0.020	45.430	0.000	0.868	0.947
I(np.log1p(years_establish))	-0.0426	0.029	-1.458	0.145	-0.100	0.015
I(years_schooling ** 2)	0.0012	0.000	2.580	0.010	0.000	0.002
I(np.log1p(no_paid_workers))	0.0957	0.033	2.916	0.004	0.031	0.160

Omnibus:	1186.158	Durbin-Watson:	1.966
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1574579.814
Skew:	-6.638	Prob(JB):	0.00
Kurtosis:	212.202	Cond. No.	2.05e+03

Figure E.1: Weighted Least Squares with log transformation model summary