Ghana Agricultural Profit Analysis

By Jomaica Alfiler, Shruthi Dakur, Natasha Kacoroski, and Daizy Koech

Introduction:

ACME corporation is investigating the profitability of moving into agricultural inputs in Ghana and require additional information on what determines agricultural profit to support the decision-making process. The purpose of this report is to provide business insights on what determines agricultural profit in Ghana by looking at educational attainment and local area characteristics.

Methods:

Data was sourced from Ghana Statistical Service's fourth Living Standard Survey and associated aggregate tables prepared for a World Bank report. All analysis was completed in RStudio.

To allow for regional comparison, agricultural profit was calculated per acre. Data used to calculate agricultural profit was loaded from the "agg2.dta" and sec8b.dta" files. The first file was used to determine the agricultural profit per household. The corrected agricultural income variable was used for agricultural profit because it is standardized. Agricultural profit was merged with agricultural land and unit per household using the second file. Observations where the land unit was missing were removed and all areas were converted acreage. The agricultural profit was divided by the acreage to get the agricultural profit per acre for each household observation.

Educational attainment was prepared from the "sec1.dta" and "sec2.dta" files. Similar to the documentation provided with the data, educational attainment was bucketed into the following categories: no education, less than a basic education, basic education (MSLC/BECE qualification), and secondary or higher educational qualification. Educational attainment was filtered to individuals at least 15 years of age, because that is the minimum age needed to have enough time to complete a basic education qualification. The highest educational attainment was determined for each household.

Regional information was prepared from the "sec0a.dta" file. Due to missing values in the region and district variables that indicated a potential region mismatch, these regional variables were dropped as the data is suspect. Regional variables were provided for ecological zone, urban versus rural, a combination of the two, and urban versus rural versus capital. As the community variables are only present for rural areas, ecological zone will be the primary regional variable.

Community variables were prepared from the community data files provided. The community size was taken from the "cs1.dta" file and the distance to primary school from the "cs3.dta" file. As there are multiple communities in some enumeration areas, the community size was summed for each enumeration area. Primary school distances where the school was in the community

were updated from missing to zero. Remaining missing values for both variables were replaced with the median as both distributions are skewed right. Duplicate rows were dropped.

The base data for developing a model was prepared by using a left join from agricultural profit per acre (explanatory variable) to all other prepared data tables. Observations that did not contain any community data were dropped.

Descriptive statistics were completed for the base model data, including a correlation matrix of the continuous variables. Models were developed using various combinations of the variables and variable manipulations. Results were interpreted for each model and models were validated using F-statistics, adjusted R-squared values, and if normality and constant variance assumptions were met.

Data Summary:

Variables in the base model data include agricultural profit per acre in cedis (profit), community size (size), distance to primary school in kilometers (school distance), educational attainment (education), and ecological zone. The response variable is profit, the remaining variables are explanatory. There are 3,286 observations. All observations are at the household level.

From the summary statistics for continuous variables (Table 1), size, school distance, and profit are all skewed to the right. There also appear to be outliers present. Agricultural profit per acre may be negative or zero. This makes sense as it is possible that farming expenses are greater than the income for certain households.

Table 1. Continuous Variable Summary Statistics

Variable	Size	School Distance	Profit
Minimum	25.0	0	-3,451,796
1st Quartile	341.0	0	42,835
Median	649.0	0	133,690
Mean	952.4	2.731	427,864
3rd Quartile	1051.0	0	334,029
Maximum	7620.0	360	51,470,781

From the ecological zone percentage breakdown (Table 2), over half of the observations are on forested areas, a little over a quarter on savannah areas, and less than a quarter on coastal areas. As the observations were collected randomly, there is not a concern with oversampling from a particular zone.

Table 2. Ecological Zone Percentage Breakdown

Ecological Zone	Households (%)
Forest	52.8
Savannah	27.8
Coastal	19.5

From the educational attainment percentage breakdown (Table 3), almost half of households have someone with a basic education, for over a quarter of households all members have never attended school, under a quarter of households have less than a basic education, and only around 10% of households have someone with a secondary or higher educational qualification.

Table 3. Education Percentage Breakdown

Educational Attainment	**Households (%)**
Never attended school	27.9
Less than basic education	21.0
Basic education qualification	40.8
Secondary or higher qualification	10.3

From the continuous variable correlation matrix (Table 4), all relationships between variables are weak. As the community size increases, the school distance and profit slightly decrease. As the school distance increases, the profit slightly increases.

Table 4. Continuous Variable Correlation Matrix

Variable	**Size**	**School Distance**	**Profit**
Size	1	-0.032	-0.018
School Distance	-0.032	1	0.005
Profit	-0.018	0.005	1

From the average agricultural profit per acre by highest household educational attainment (Table 5), a basic education qualification is associated with the highest median agricultural profit per acre, followed by less than basic education, secondary or higher education, and never attended school. Using mean agricultural profit per acre, the highest amount is associated with less than a basic education, followed by basic education qualification, never attended school, and secondary or higher education qualification.

Table 5. Average agricultural profit per acre by highest household educational attainment.

Highest Household Educational Attainment	**Median Profit**	**Mean Profit**
Never attended school	101,665	375,964
Less than basic education	137,675	579,625
Basic education qualification	157,563	401,418
Secondary or higher education qualification	121,002	363,149

From the average agricultural profit per acre by ecological zone (Table 6), households in a forest ecological zone are associated with the highest median agricultural profit per acre, followed by savannah, and coastal. For mean agricultural profit per acre, coastal is associated with the highest value, followed by forest and savannah.

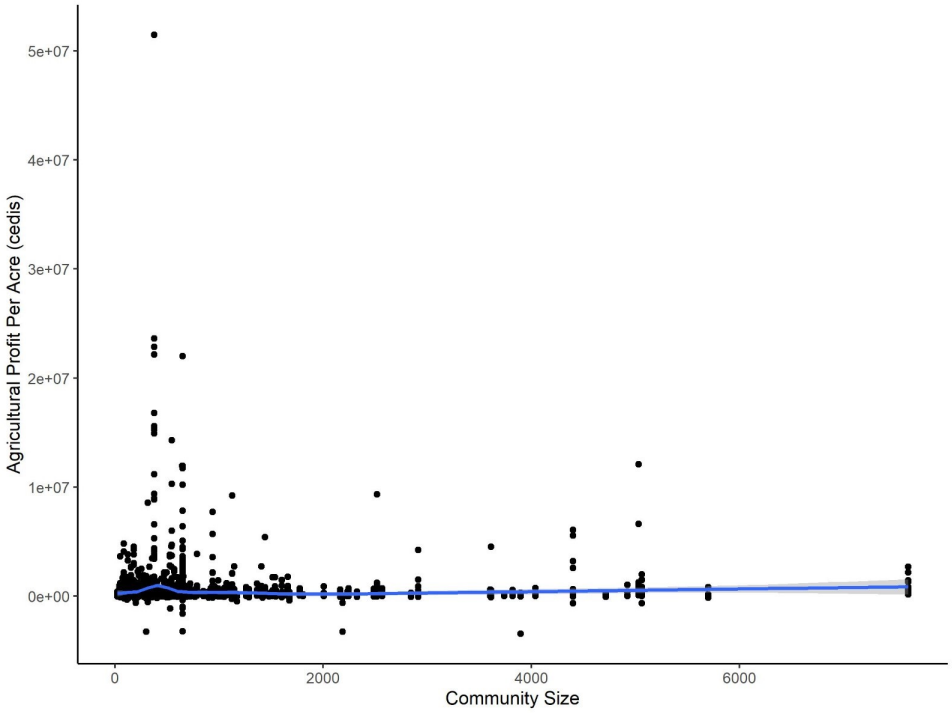
Table 6. Average agricultural profit per acre by ecological zone.

Ecological Zone	**Median Profit**	**Mean Profit**

Forest	171,092	412,054
Savannah	107,758	321,531
Coastal	62,388	622,223

From the agricultural profit per acre by community size (Figure 1), there does not seem to be a relationship between community size and agricultural profit per acre.

Figure 1. Agricultural profit per acre (cedis) by community size



Based on this information and general knowledge, our hypotheses are as follows:

1. Community size does not influence agricultural profit per acre, or a small negative effect. As a community gets larger, development pressures could make farm land more expensive.
2. Distance to primary school has a small positive effect on agricultural profit per acre. As the distance increases, the agricultural profit increases. It seems like distance to primary school may be less of a proxy of education level and more of an indicator whether there is development or open space for farming.
3. Ecological zone has a strong effect on agricultural profit per acre, with forest having a positive impact. Savannah areas are likely to not have enough rain for supporting crops,

coastal areas may have rocky or sandy soil. Areas where forests can grow may have good rainfall and soil quality for agriculture.

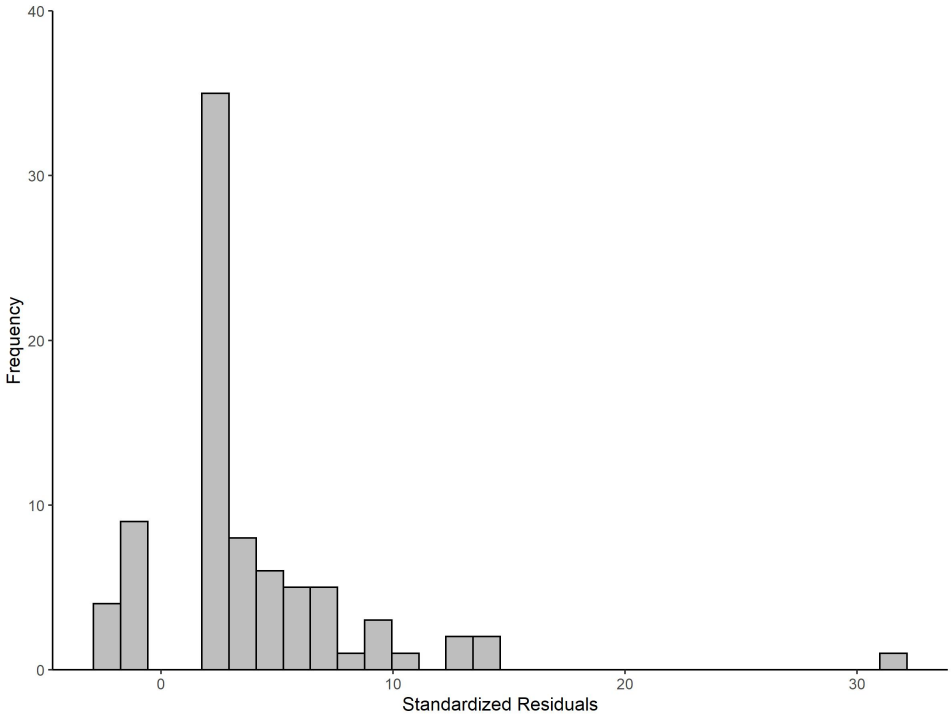
4. Educational attainment has a small positive effect on agricultural profit per acre. It may be sufficient as a proxy variable for innate ability of the household to work with technology and understand how to maximize income, but it is likely one of many factors impacting agricultural profit.

Results:

The first model included all variables and was the effect of community size, highest educational attainment per household, ecological zone, and distance to primary school on agricultural profit per acre. The coefficient for community size was -31.63 and is not statistically significant. For every additional person in the community, the agricultural profit per acre decreases by 31.63 cedis, holding all other variables constant. The coefficient for an educational level of MSLC/BECE (basic education) relative to an educational level less than a basic education was -166,232.19. It means that compared to a household where the highest educational level attained is less than basic, households with a basic education level have an agricultural profit per acre that is lower by 166,232.19. cedis on average. It was statistically significant at the 5% level. The coefficient for an educational level of none, relative to an educational level less than a basic education is -176,017.49 cedis. It means that compared to a household where the highest educational level attained is less than basic, households with no education have an agricultural profit per acre that is lower by 176,017,49. cedis on average. It was statistically significant at the 5% level. The coefficient for an educational level of secondary or higher relative to an educational level less than a basic education is -189,097.60 cedis. It means that compared to a household where the highest educational level attained is less than basic, households where the highest education level is secondary or higher have an agricultural profit per acre that is lower by -189,097.60. cedis on average. It was statistically significant at the 10% level. The coefficient for a forested ecological zone relative to a coastal ecological zone is -205,146.87 and is statistically significant at the 1% level. It means that in forested areas the agricultural profit per acre is 205,146.87 less than coastal zones on average, holding all other variables constant. The coefficient for a savannah ecological zone relative to a coastal ecological zone is -299,568.65 and was statistically significant at the 0.01% level. It means that in savannah areas the agricultural profit per acre is 299,568.65 less than coastal zones on average, holding all other variables constant. The coefficient for distance to primary school was 468.88 and was not statistically significant. It means that for every additional kilometer that the primary school is further away, the agricultural profit by acre increases by 468.88 cedis on average holding all other variables constant.

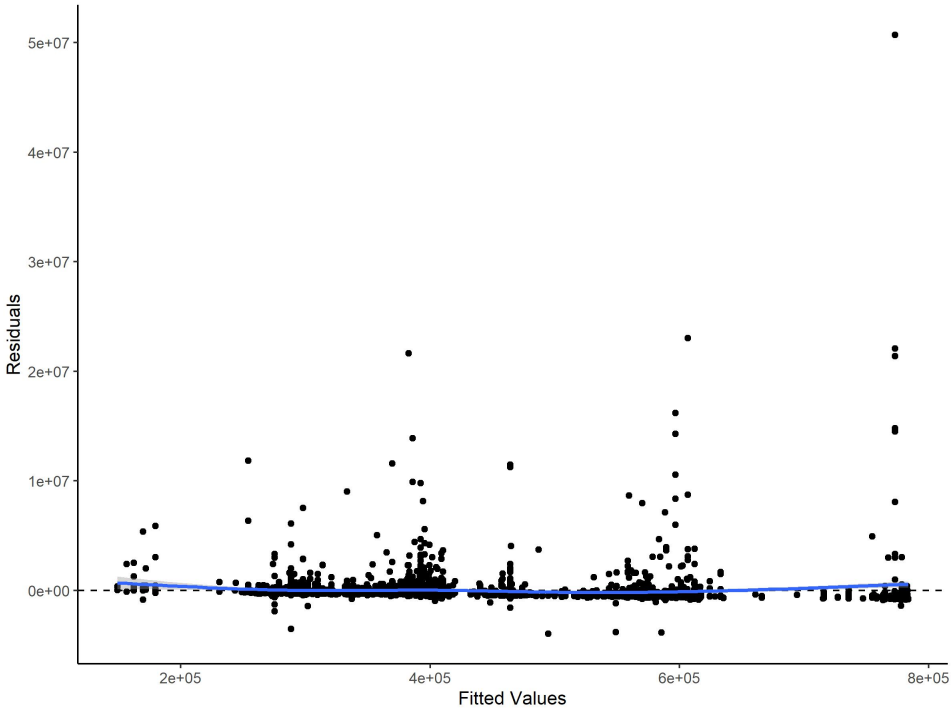
The adjusted R-squared value was 0.004467, so the model explains very little of the variation seen in the response variable. The F-statistic is 3.106 and statistically significant at the 1% level. The distribution of residuals is skewed right (Figure 2) so the normality assumption fails, meaning that the errors are biased, but it does not impact getting unbiased estimates.

Figure 2. Model 1 Standardized Residual Distribution



Graphing the residuals against the fitted values (Figure 3) shows that the mean seems to remain constant and the spread is relatively balanced, indicating homoskedascity.

Figure 3. Model 1 Residuals vs. Fitted Values



As distance to primary school was the least statistically significant, the variable was dropped from Model 2. The adjusted R-squared value increased to 0.004706 and F-statistic increased to 3.589, still significant at the 1% level.

For Model 3, the interaction effect between educational attainment and community size was tested. Note of the interaction coefficients were statistically significant, size was still insignificant, and secondary education or higher attainment relative to less than a basic education was no longer significant. The adjusted R-squared value decreased to 0.00437 and the F-statistic decreased to 2.605, still significant at the 1% level.

Model 4 included community size squared, which improved the model. The adjusted R-squared value increased to 0.006278. The F-statistic increased to 3.965 and was significant at the 0.1% level. The coefficient for community size was –183.5 and significant at the 1% level. The coefficient for community size squared was 0.03029 and significant at the 5% level. The turning point is approximately 3,030. As the coefficient for community size is negative and the coefficient for community size squared is positive, community size follows a convex curve. Holding all other variables constant, as the community size increases, the agricultural profit per acre decreases with diminishing returns. At about a community size of 3,030 the direction turns and as community size increases, the agricultural profit per acre increases. 94.6% of the observations occur before the turning point where agricultural profit per acre decreases with community size. As most of the data is before the turning point, it is likely that the relationship is negative and the observations after the turning point are outliers. It is possible that there are some large farming operations where the community is big and the agricultural profit per acre is high, however overall, as more individuals move into an area the pressure for development increases the costs for agriculture.

Model 5 removed community size and community size squared to test if removing the variable created a better model than keeping in as a polynomial. The adjusted R-squared value decreased to 0.004563. The F-statistic was higher at 4.012 but back to the 1% significance level.

Model 6 tested if adding the log of community size created a better model. The adjusted R-squared value decreased to 0.004372 and the F-statistic decreased to 3.404, significant at the 1% level.

Overall, the best model was Model 4 with explanatory variables: community size, community size squared, highest educational attainment per household, and ecological zone. To check if the model is valid, the residual distribution (Figure 4) and residuals versus fitted (Figure 5) graphs were examined. They do not appear to have changed significantly. The distribution of the residuals is still skewed right. For the residuals versus fitted values, the mean still appears to be stable and spread mostly balanced even though there are outliers. The model is still valid though the errors are biased. It would be helpful to have more data, or it might make sense to try removing some outliers.

Figure 4. Model 4 Standardized Residual Distribution

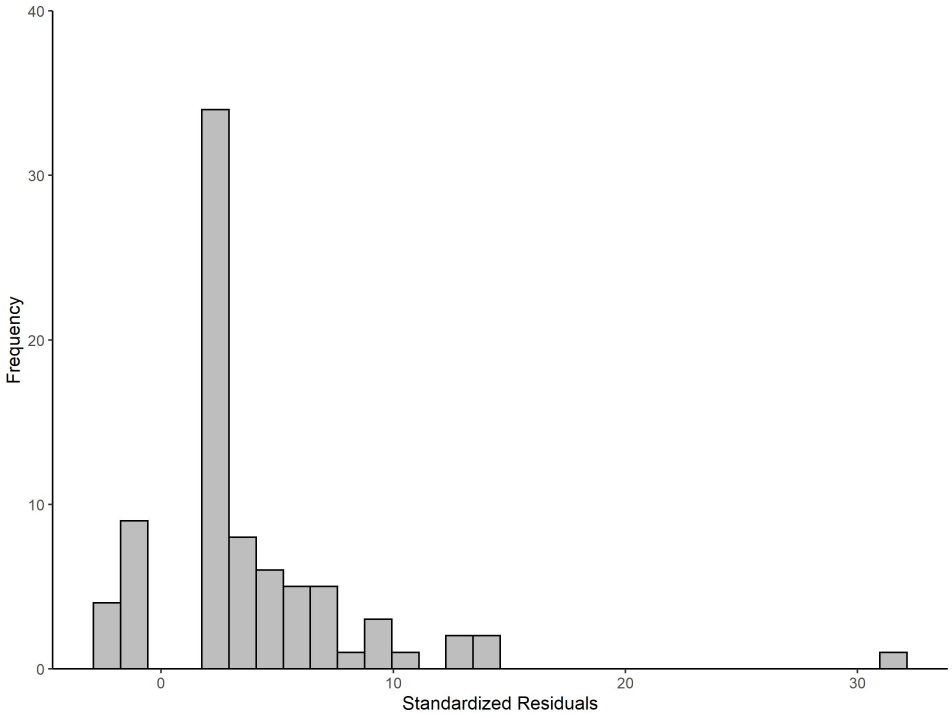
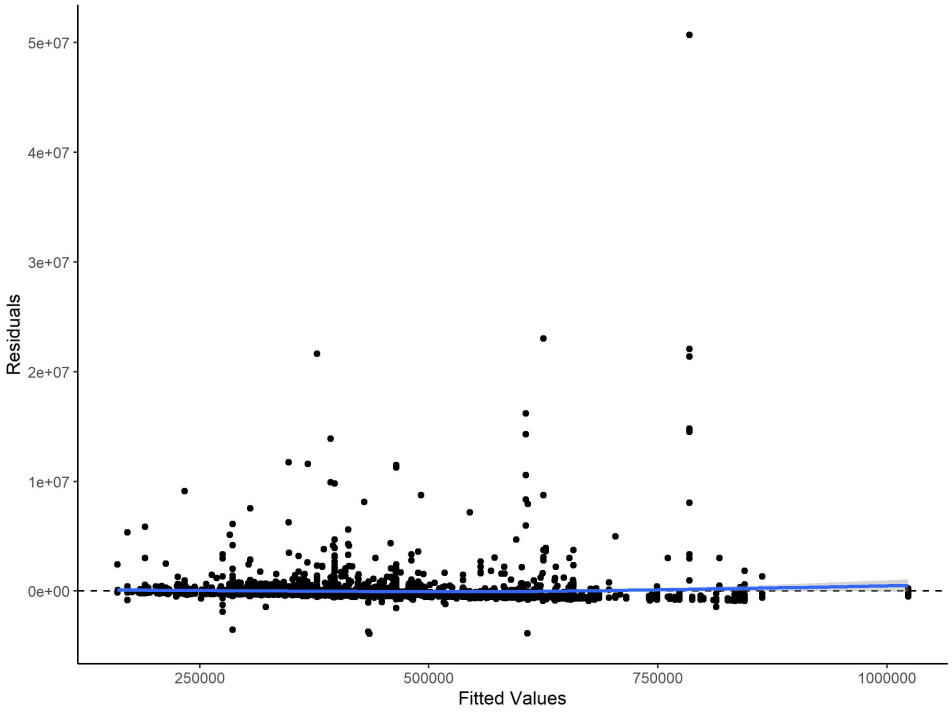


Figure 5. Model 4 Residuals vs. Fitted Values



Conclusion:

Based on our initial model results, a coastal ecological zone in a small community where the highest educational attainment per household is between no education and a basic education is associated with the highest agricultural profit per acre. These results conflict with the descriptive statistics of the data, most likely because of the non-normal distribution. Reviewing descriptive statistics in tandem with our model results, it seems like a forest ecological zone in a small community where the highest educational attainment per household is between no education and a basic education is associated with the highest agricultural profit per acre.

More research is needed, however, as very little of the variation in agricultural profit per acre is explained by our model. There are additional variables within the data provided that have not been evaluated yet. Some of these variables are associated with agriculture and may provide additional insights, such as presence of a local market or presence of irrigation. It is also important to note that there is a wide range of climatic factors not captured in the data that might be important too, such as rainfall and temperature.

Appendix:

Analysis Scripts in Project Repository

- Educational Attainment Data Preparation – edu_attainment.R
- Agricultural Profit by Acre Data Preparation – ag_profit_by_area.R
- Community Size Data Preparation – community_factors.R
- Primary School Distance Data Preparation – community.R
- Ecological Zone Data Preparation – hh_regions.R
- Base Model Data Preparation – prepare_model_data.R
- Model Development – develop_model.R