Glen Lewis, Sohrab Rajabi, Rosalie Sherry, Tamlyn Tamura

OMSBA-5112

Data Translation Challenge

Statistical Analysis

12/04/2020

Index of Variables Used in This Report

	Name	Alternate Name (or where data was taken from)	Type	Definition
Household Variables	region		Factor	Area of Ghana where household resides
	education_completed	s2aq2	Factor	Highest level of schooling completed
	av_hh_age	agey	Numeric	Average age in years for household. Determined by grouping individuals by household id then - agey = round(mean(agey), 1)
	male	sex	Numeric	Total number of males in household Determined by grouping individuals by household id then counting the number applicable factor levels
	female	sex	Numeric	Total number of females in household this also then gives us the total number of men
	male_help	s8cq17a	Numeric	Taken directly from sec8c data set, the total number of hired male hands
	female_help	s8cq17b	Numeric	Taken directly from sec8c data set, the total number of hired female hands
Land Variables	income10_final	income10, inc10	Numeric	Shows how much revenue was produced from each cash crop
	income11_final	income11, inc11	Numeric	Shows how much revenue was produced from each type of root/fruit/vegetable
	income12_final	income12, inc12	Numeric	Shows revenue from other types of agricultural income
	income13_final	income13, inc13	Numeric	Shows how much revenue is generated from transformed or processed foods
	expenditure3_final	expenditure3, exp3	Numeric	Shows the expenditure on renting farmland
	expenditure4_final	expenditure4, exp4	Numeric	Shows the expenditure on crop inputs
	expenditure5_final	expenditure5, exp5	Numeric	Shows the expenditure on livestock inputs
	expenditure6_final	expenditure6, exp6	Numeric	Shows labor costs on food processing
	expenditure7_final	expenditure7, exp7	Numeric	Shows how much of the output is consumed by the household
	agri_land_final	agri_land, sec8a1	Numeric	Shows whether a household currently owns any land
	agri_livestock_Fishing_final	agri_livestock_Fishing, sec8a2	Numeric	Shows how many of each type of livestock or fish is owned by each household
	agri_equipment_final	agri_equipment, sec8a3	Numeric	Shows how many of each type of agricultural equipment owned
	agri_plot_final	agri_plot, sec8b	Numeric	Shows farmland size in acres

Abstract

In determining what drives agricultural profit, three surveys were conducted in Ghana. Both at the individual and household level, participants were able to provide a first-hand look at their day-to-day life. While some of the data was focused on household, much of it was focused on land and farming. Analyzing the data from a birds-eye perspective generated several key insights related to profitability in Ghana and whether it is a strategic business move to invest here. From synthesizing and tidying the data, this paper aims to prove that investing in land in Ghana is not a sound business decision.

Background

The ACME Corporation has entrusted the DTC-4 Group with finding what factors, if any, drive agricultural profit per an acre in Ghana. For the DTC-4 group to draw an accurate conclusion, they have spent the last couple of weeks parsing through files and survey responses in order to get a full picture on what Ghanaian lifestyle. They also needed to provide a definition for profit so they could have a tool into which they could measure ACME's request. DTC-4 has determined that there are two main groups that all variables can be split into that determine profit— land and household. Each variable was grouped into one of these two groups, and each group then went through its own data wrangling process before being merged to create a final profit model. While looking at these variables the DTC-4 team determined that agricultural profit per an acre is how much money is generated for each acre after expenditures are considered.

Since Ghana is geographically, socioeconomically, and culturally different from the United States, DTC-4 members were not sure as to what to expect before diving into the data; they hypothesized that education and literacy were two key factors that would drive profitability as well as the number of hands working on the land. Since the question being answered is centralized around *agricultural* profitability, they hypothesized the land variables would have a much larger effect on profit than household variables. Since Ghana is made up of different ecological zones and districts, they also hypothesized that some areas of Ghana would be more profitable than others.

Variable Explanation

In order to split the data into their respective groupings, it first needed to be decided which variables from the data sets were relevant. The DTC-4 Group determined that 7 variables were relevant to the household portion and 13 were relevant to the land portion. The datasets for household that the variables were pulled from was The Household Roster (SEC1), Education (which consists of a combination of SEC2A: General Education, SEC2C: Education: Literacy/Apprenticeship), EXP3: LANDEXP = Expenditure on renting farmland, EXP4: CROPEXP = Expenditure on crop inputs, EXP6: which is made up of two subgroups: FDPREXP1: Labour[1] costs on food processing, FDREXP2: Other costs on food processing, and finally EXP7: HP = Consumption of home production. Each variable, in addition to being explained further in this section of the paper, can also be quickly referenced in the ***Index*** found at the beginning of this paper.

From these data sets, there were 7 variables that were decided upon for the household portion. The first variable was a factor we titled region as it explains the area of Ghana where the land resides. Since it was a factor DTC-4 compared each region to *Region 1* as it was the most profitable; all other regions paled in comparison. Education_completed is another factor variable used in the household model that combines literacy and level of schooling in order to get the education level of household members. Av_hh_age is a numeric variable that states the average age of each household member by calculating the sum of the ages of household members and dividing by the total number of household members. Next are variables male and female, both of which fall under the umbrella variable sex which explains the biological make up of each household. And the last two variables are male_help and female_help which explains the biological make up of those working on the land who do not belong to the household.

For household, only the variable region is a dummy variable. Meaning it is a categorical variable that is transcribed into being numeric. The data and corresponding documents defined

[1] The datasets and documentation were written in the Queen's English, when quoting said documentation this paper will use the language as it appears in the documentation, when writing it will continue to be written in American English style.

11 regions with varying environmental and economic factors that DTC-4 transformed into a variable that had values ranging from 1 – 11. This is important to take note of as it can be another source for a margin of error (further instances of such are explained in the section titled *Notice of Possible Errors* at the end of this paper).

The datasets used for the land variables provide data on revenue from sale of cash crops (INC10), revenue from sale of roots/fruits/vegetables (INC11), revenue from other agricultural income (INC12), and revenue from transformed (processed) crops (INC13). On the expense side, they chose to use the same expense variables listed above in household (EXP3 through EXP7). Other important variables used are land data (SEC8A1), livestock and fishing data (SEC8A2), agricultural equipment data (SEC8A3), and agricultural plot details (SEC8B).

INC10_FINAL was created by combining two variables within the dataset, CROPSV1 and CROPSV2 (which gave a total of total revenue from combined crops), and then mutating the data so that the crop code (variable name "cropcd") displayed a table with each unique name of the crop as its own column, instead of the numeric code associated with the crop in one column. After that, the crops with no values across the board were eliminated, and INC10_FINAL was created.

INC11_FINAL was created in a similar manner, by mutating the data so that the "root_name" code was transposed to the actual name of the root/fruit/vegetable having its own column. INC13_FINAL was also created in a similar manner, by mutating the data so that the "trans_crop_name" was transposed to displaying the actual name of the transformed crop assigned to its own column, instead of a series of numbers displayed under one variable "trans_crop_name". INC11_FINAL and INC13_FINAL both required additional data wrangling, taking a column of data and spreading it to create multiple other variables within the data sets. INC12_FINAL kept the same data but was renamed for convenience purposes.

DTC-4 mutated the data in EXP3 through EXP7 to display the respective expenses and codes, similarly to what was done in INC10 and INC11, to get to EXPENDITURE3_FINAL, EXPENDITURE4_FINAL, EXPENDITURE5_FINAL, EXPENDITURE6_FINAL AND EXPENDITURE7_FINAL.

AGRI_PLOT_FINAL was created to show how large each farmland area was in units of acres. Since some households chose to represent their farmland size in terms of ropes or poles, DTC-4 had to divide the number of poles by 210 and the number of ropes by 9 in order to transpose these units to acres across the board for a cleaner look at the data.

AGRI_LIVESTOCK_FISHING_FINAL and AGRI_EQUIPMENT_FINAL were created in a similar manner, by mutating the column to produce the actual name of the livestock animal or fish, or the actual name of the equipment, respectively.

AGRI_LAND_FINAL selected 3 columns from the original data, one of which consisted of a dummy variable; this variable shows a 1 if the household owns any land and 2 if the household does not own any land.

Implementation

Initially, the DTC-4 team looked at both household and land variables individually in order to have an indicator whether their initial hypothesizes were measuring up. This later turned out to be a poor indicator of what would be important once both data sets were combined. As an example, Table 1 displays the coefficients of each variable in the household model that relates to sex but *only* in the context of *the household*. The first two columns show the four variables relating to sex as they are initially. As the model progresses DTC-4 took out the variables that were not statistically significant (at the 10% level)[2]. As one can see with just the removal of female_help from the dataset (seen in the next three columns of the table) one can see an increase in the coefficients of the other variables and thereby their respective impact on profitability of the model.

After going through this process, DTC-4's working hypothesis was shifting towards believing that a major contributing factor to profitability per an acre was members of the household (regardless of sex) and the male help employed on the land. When looking at this early model DTC-4 saw that many of the households do not have hired help (the data provided a median of 0.000, mean of 2.081, and a max of 262.000), but, despite this, DTC-4 believed it had a major impact on the farms that used hired (male) help. Therefore, DTC-4 pivoted their

[2] In this section it is discussed getting rid of variables that are not significant at the 10% level while later the findings are presented with perspective from the 1% level. Though this may be confusing, this is not a mistake.

hypothesis to include the hired male help. Once the models were combined, it was then seen that none of the variables relating to sex ended up being statistically significant. Looking back on this portion of DTC-4's work process it was clear they were almost led astray by trying to draw conclusions too early with only half the data.

Next, DTC-4 combined the two branches, household and land, together for the purpose of comparing profit variables. Using the global key (labeled “key” in the database) that DTC-4 created the datasets were joined. This “key” was made by uniting the variables household ID and cluster number (“nh” and “clust”). The key is present in both the individual household and land data frames as well as the combined one. By doing this, DTC-4 was able to achieve a more specific explanation of where certain parts of the data were coming from.

Results

After reviewing our hypotheses there were a couple of surprises, the first being the influence education has on agricultural profit. Only two factor levels of completed education had statistically significant influence on profit at the 10% level: middle school completion and the koranic stage completion. One point of interest was while having completed middle school has a positive influence (coefficient: ¢ 193,867 GHS), the koranic stage training led to a significant negative influence on profit (coefficient: ¢ - 401,125 GHS). DTC-4 believes that a possible rationale for this is that the time spent pursuing this training takes away from a person’s ability to perform agricultural work. An interesting trend, though not significant, is that it appears as household members gained more education, the influence of education on profit turns increasingly negative, thus, disproving one of the initial hypothesizes made by DTC-4.

The second surprise was the minimal impact the average household age had on profit. The DTC-4 hypothesized that an older household could expect higher annual profits due to the buildup of knowledge and experience in farming over time. However, the average household age (variable: av_hh_age) while statistically insignificant (.0973) at the 10% level in influencing profit, the influence was only about ¢ 3,279 GHS per additional year of age. To put it further into perspective, this is about $560.99 in USD.

After merging the land and household data, DTC-4 took all the variables that were statistically significant at the 99% confidence level and generated the following final profit model:

```
agr_prod_profit_test3f <- lm(profit ~ factor(region) + male + I(male^2) +
        Husked_polished_rice + Gari + I(Gari^2) + Sheabutter + I(Sheabutter^2) +
        Pigs + Fish + Processed_fish + space_in_acres + I(space_in_acres^2),
        data = capstone_df)
```

DTC-4 chose this model over other models because it had the best fit. The adjusted R^2 was smaller in the other models they looked at, meaning, that the models accounted for a lower percentage of profit than this one. In some instances, the other models that DTC-4 created had long tails that presented a model with a non-normal distribution. These skewed models are due to the presence of outliers and because of these distributions DTC-4 did not feel confident in their ability to use them and be able to provide answers such as whether ACME should invest in Ghana. Also, this model appeared to be the most interesting to the team and allowed them, in their opinion, to best answer ACME's core question of whether or not to invest.

Looking at Figure 1, it makes sense to have a relatively symmetrical normal distribution with a mean of 0, since being at the 99% confidence level means 99% of the data falls within 3 standard deviations. When DTC-4 pulled up a summary of the statistics for the model, they were surprised to find that the adjusted R-squared only amounted to .2044--this means, that our model of 8 variables accounts for roughly 20% of what drives profit in Ghana, which is vastly different from our hypothesis that land alone should account for ~70% of profit. Nevertheless, to be able to represent a fifth of what factors drive profit within 8 variables out of thousands does hold a significant amount of influence that should not be overlooked.

Table 3 shows a summary of the coefficients from AGR_PROD_PROFIT_TEST3F. The table shows that all regions are statistically significant except for region 7. Region 1 (the intercept) has the highest coefficient estimate (7.918e+05), which means that each acre in Region 1 will produce about ¢ 791,800 GHS in profit, or $135,582 in USD. Since all the other regions have negative coefficient estimates, it shows that any other region will produce that amount less than Region 1 (since Region 1 is the represented as the intercept). Keeping the male variable in the final was essential to dampening out a couple of outliers that otherwise would have skewed the figures. The coefficient for males is positive, meaning that more males in a household will

lead to higher profit, most likely because there are more hands working on the farm, but it is not statistically significant, so it does a huge contributor in what determines profitability.

Husked polished rice does turn out to be profitable—it's cheap, easy to grow and can be stored for a long time. It also shows a linear pattern of profitability, meaning that the more rice grown, the more profit is returned. Gari also turns out to be a profitable crop, but only up until the 500,000 units mark, after which, profitability will start to decline. Shea butter also had interesting results—there is one outlier that produces a mass amount of shea butter, but the y-axis shows this farm lost money. Up until about 100,000 units, the profit starts to decline, so shea butter would not be a good area to invest in, since there's not a high enough profit on production to make it worthwhile. Pigs and processed fish are profitable, but regular fish is not profitable, as seen by looking at its negative coefficient estimate.

Looking at space_in_acres, it has a coefficient of 1.094e+05 meaning that for each acre increase, one can expect a ¢ 100,940 GHS increase in profit, or $18,732 USD. Thus, each additional acre is worth $18,732. As a caveat, Figure 3 shows that as farms approach roughly 103 acres, profitability begins to decrease.

Figure 2 shows the model's constant variance, where the plots of data are scattered around 0. Therefore, it is likely that the error terms' mean in DTC-4's profit model is close to 0. There is also a slight downward trend, which accounts for a few outliers, making the model appear to have homoskedasticity.

Conclusion

Based on our findings, the ACME Corporation should not invest in Ghana based on Agricultural land per an acre. DTC-4 only found one region of Ghana profitable (Region 1) and found that the land that did best across all regions were the ones that farmed rice. The point of profit maximization for land size is at 103.75 acres (Figure 3), after which point each additional acre is less profitable than the last. Thus, if ACME still chooses to invest and had to choose a target market, they should look at farms that are 103 acres large or less, for which each acre increase means $18,732 more than the last. If ACME is interested in pursuing this further, DTC-4 would suggest that they consider these factors, as the rest of the regions in Ghana do not appear to be a sound investment based off the data that has been collected and examined.

Notice of Possible Errors

It should be noted that this conclusion was drawn using data that did not have consistent methods of collection. One area to note is the AGRI_PLOT_FINAL, which shows how much land is owned in acres. In the survey, each household member entered a number followed by checking a ‘unit’ box (Acres, Ropes, Poles, Other), and while most of the observations were in acres, ropes and poles, there were 7 observations out of ~4100 that fell under the “Other” category. Even after extensively looking through the documentation, DTC-4 was unable to find what units these entries fell under, so they decided to leave these entries out.

Should the ACME Corporation want a more in-depth review or wishes to further pursue Ghana for various business interests the DTC-4 team would recommend finding a different surveying company as the previous one left many areas for error to be found in the DTC-4's analysis. Within the provided documentation it noted that many of the numbers were calculated to be on an annual scale but using data colleting from varying date ranges to get to make that estimation. Some of these data points were taken during regular intervals for it to be considered on an annual basis some of the data points were taken over a period as little over a two-week period and projected for an annual basis. For more information on the errors that may have occurred in this analysis please look through the Aggregate.pdf pages 13-16.

Figures & Tables

Table 1: Household Coefficients – Initial Findings

Variable	Coefficient	Variable	Coefficient	Percent Change
male	197340	male	202402	2.5%
female	96083	female	96713	.65%
male_help	22233	male_help	24537	9.4%
female_help	41004			

Table 2: Correlated Coefficients

	profit	Flour_from_other_grains	Husked_polished_rice	Home_brewed_drink	Processed_fish	Gari	Sheabutter	Other_nuts	Pigs	Chicken	Other_poultry	Fish
profit	1.00	0.01	0.08	0.03	0.06	0.17	-0.10	0.04	0.08	0.10	0.05	-0.07
Flour_from_other_grains	0.01	1.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Husked_polished_rice	0.08	0.00	1.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
Home_brewed_drink	0.03	0.00	0.00	1.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00
Processed_fish	0.06	0.01	0.00	0.00	1.00	0.00	0.00	0.00	0.01	0.00	-0.01	0.00
Gari	0.17	0.00	0.00	0.00	0.00	1.00	0.00	-0.01	0.00	0.00	-0.01	0.00
Sheabutter	-0.10	0.00	0.00	0.01	0.00	0.00	1.00	0.00	0.00	0.00	0.01	0.00
Other_nuts	0.04	0.00	0.00	0.00	0.00	-0.01	0.00	1.00	-0.01	0.00	-0.01	0.00
Pigs	0.08	0.00	-0.01	0.03	0.01	0.00	0.00	-0.01	1.00	0.05	0.09	0.00
Chicken	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	1.00	0.00	0.00
Other_poultry	0.05	0.01	0.00	0.00	-0.01	-0.01	0.01	-0.01	0.09	0.00	1.00	0.00
Fish	-0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 3: Summary list of Intercepts

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.918e+05	6.789e+04	11.663	< 2e-16	***
factor(regio)2	-5.160e+05	7.591e+04	-6.798	1.10e-11	***
factor(region)3	-6.712e+05	7.207e+04	-9.314	< 2e-16	***
factor(region)4	-8.864e+05	7.300e+04	-12.143	< 2e-16	***
factor(region)5	-3.934e+05	6.970e+04	-5.644	1.69e-08	***
factor(region)6	-5.615e+05	6.634e+04	-8.464	< 2e-16	***
factor(region)7	-7.361e+04	8.298e+04	-0.887	0.375	
factor(region)8	-1.211e+06	9.892e+04	-12.239	< 2e-16	***
factor(region)9	-1.253e+06	1.520e+05	-8.240	< 2e-16	***
factor(region)10	-1.103e+06	1.151e+05	-9.578	< 2e-16	***
male	3.659e+04	2.718e+04	1.347	0.178	
I(male^2)	1.589e+04	3.482e+03	4.562	5.10e-06	***
Husked_polished_rice	6.625e-01	5.776e-02	11.469	< 2e-16	***
Gari	2.326e+01	1.045e+00	22.251	< 2e-16	***
I(Gari^2)	-2.040e-05	1.420e-06	-14.364	< 2e-16	***
Sheabutter	3.457e+00	3.886e-01	8.898	< 2e-16	***
I(Sheabutter^2)	-1.879e-07	1.677e-08	-11.204	< 2e-16	***
Pigs	8.536e+04	8.568e+03	9.963	< 2e-16	***
Fish	-6.800e+02	6.765e+01	-10.052	< 2e-16	***
Processed_fish	3.057e-01	3.399e-02	8.993	< 2e-16	***
space_in_acres	1.094e+05	2.767e+03	39.547	< 2e-16	***

Figure 1: Distribution of Normality

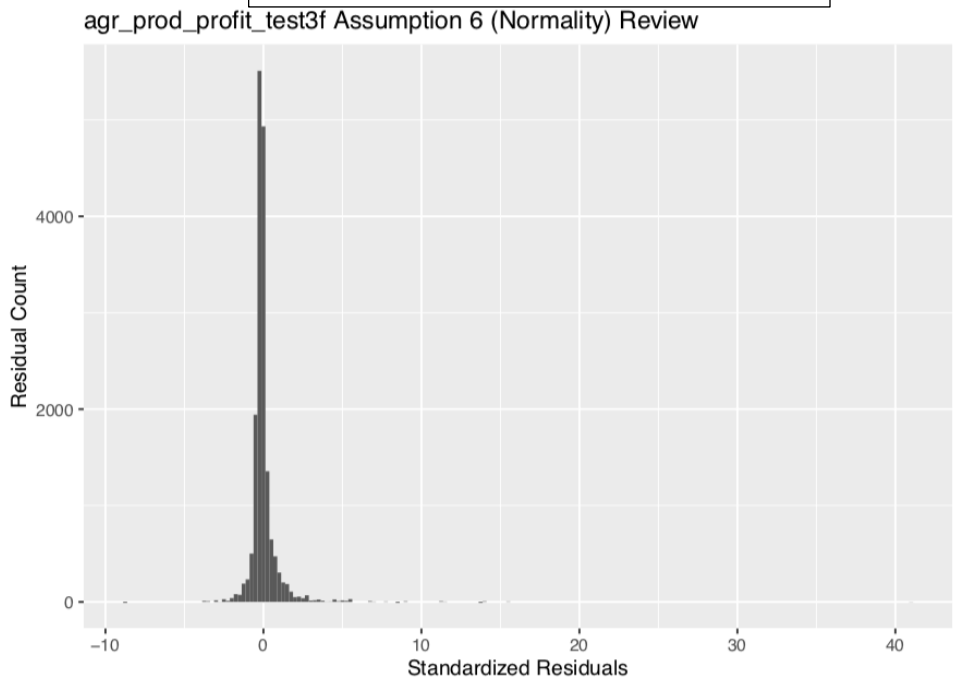

Figure 2: Variance using Fitted vs. Standard Residuals

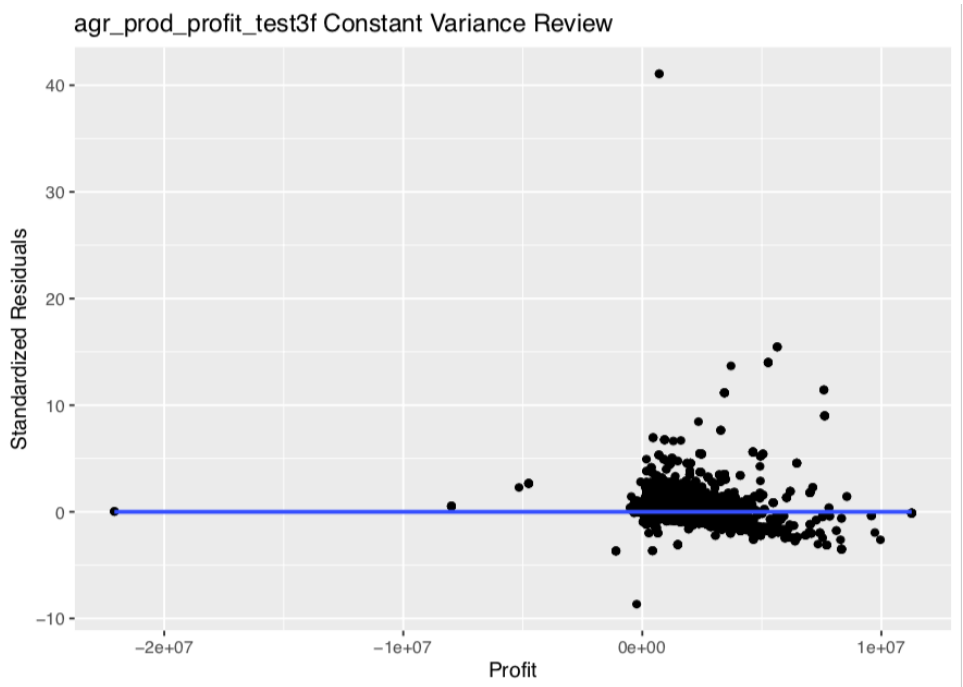

Figure 3: Profitability Model