Statistical Analysis Report

1. Method

In this project, we combine all the data from different categories and use them to interpret the profit relating to agriculture for each household in Ghana. Then we used linear regression as our model to see if there are any significant variables among all the data that we selected.

1.1 Data

In general, we selected the variables that we are interested in from household questionnaires and community service questionnaires and divided them into 2 types (numerical variables and categorical variables). Then we clean the data for numerical variables and categorical variables respectively and finally we merge all the variables with area code (clust) and household (nh) by the same EA number into one data frame. At last, we use our aggregated table containing profit per acre for each agricultural relating household and merge it with the previous data frame. Therefore, we create the data frame containing all the variables for the model.

1.1.1 Household Data

After reviewing all the household questionnaires, we find some interesting data and would like to see if there are any relations between the data and profit.

The first category is about educational attainment. We would like to see the influence on the profit for each household if the household ever attended school, or could read the local language and could do written calculations and anyway. Therefore, s2aq1, s2cq2, and s2cq5 are selected respectively. Our expectation is when holding other factors constant, the people who attended school and can do written calculations and could read local language will make a higher profit than those cannot, which means the better education the people received, the higher the profit they make.

The next category is about agriculture. We would like to know how the profit is influenced by the using of different types of agricultural equipment and crop are planted for each household. And also we would like to measure the performance of profit in different regions. The general expectation for the equipment is the modern and mechanical equipment should bring more profit than human power to each household. And also the bigger sale amount on honey should make more profit than the lower sale. At the same time, the different crop should also bring different profit for each household.

The following category is about employment and time use. We would like to consider the factors such as employment status, employers in the main job, whether the members of the household do the same work, whether the fathers of households do the same work, whether the family is working on agricultural activities and housekeeping activities.

At last, we would like to count in all the other factors, such as immigration status, house conditions, and incomes & expenditures.

The variables we have could be divided into 2 types: numeric, and categorical data. For the numeric data, we fill the NA data with 0. For the categorical data, because there existed some data that is categorical but not tidy(one household could have only one answers to one question), we have to transfer it as dummy variables and make each category under this data as 0 and 1, where 0 means no and 1 means yes. For the other tidy categorical data, we keep it as it is and use factor function directly on the data before we run the model. After this step, we combined all the data together on the same clust and nh number.

1.1.2 Community Service Data

In the very beginning of this part, we pick up some of the questions that we are interested in for further research from the community service questionnaires. Thus, we select totally 36 variables from the sections of demographic information(file:cs1), economy and infrastructure(file: cs2), education (file: cs3) as well as agriculture(file: cs5b) in community service questionnaires, including 28 categorical variables and 8 numerical variables.

Next, we convert all the outliers which are larger than 1.0e+100 to 0 in all the files we select. After that, we divide the variables into 2 parts (numerical variables and categorical variables) for data cleaning. As we found that in the community service data, there are some communities share the same EA number. However, for the further use, if we want to assign the household with community data, we need to aggregate those community with shared EA number into one big area, where those community would share the same community information. So, we aggregate those observations with the same EA number into 1 observation and obtain the average data for the area. As for the categorical variables, we use the round(mean) function to gain the nearest integer to the mean, which refers to the nearest answer in the questions. For numerical variables, we use average data directly.

Then, we merge all categorical variables into 1 data frame and create dummy variables for each of them. In this case, we gather all the categorical variables together into 2 columns('question ','answer') for each EA, after which we unite these two columns into one, in order to create names for every answer in the questionnaires. (e.g. s1q4_1: the 1st answer of question 4 in section 1). Then we create a new column with the value of 1 to assign those names with the value 1. Finally, we use spread function in R to set the 2 columns ('names' column and 'value_1' column) into multiple columns and replace all the NA by 0 to obtain dummy variables for each categorical variables.

Finally, to obtain our community service data frame, we will merge all the categorical and numeric data together into 1 data frame, and use left join function to assign each EA number with exact cluster number (clust) for further use when we merge the community data with the household information. [related file : cs.R, cs1.csv]

1.1.3 Profit

By understanding the whole dataset, we are trying to explore the explanatory factors that influence the household profit per acre. So for the first step, we define the output variable --- household profit per acre. Firstly, we can find the household agricultural profit data in "agg2.dta". In this table, we choose the column "agri1c" as the household agricultural profit data. This column contains the revenue from the sale of cash crops, roots/fruit/vegetables, other agricultural sources, transformed crop products, consumption of home-produced food, and deducts the agriculturally related expenditures. It conforms to the definition of profit which is the revenue deducts expenditure, so this is the household agricultural profit in our research.

For the next step, we need to find out the household land data. In “sec8a1”, we have the data related to household land. The column “s8aq4” stands for the land size owned by each household. In our research, we took this part as the calculation base. The column “s8aq3” is the unit of the land data in column “s8aq4”. So we exclude the data that have no unit information or have unreasonably large land data. Then we convert all the land size into the same unit “acre”. The last step is to calculate the profit per acre by using profit data divided by land size. Thus we have got the data of “profit per acre” as our output variable. After we visualized the distribution of the output variable, we noticed there are extreme values on both sides, especially on the right side. Thus we decided to exclude the first 3% percentiles and the last 20% percentiles to make sure the output variable is valid and less influenced by extreme values.

With this profit data, we use left join function to merge all the 3 data frame, where the left table is profit in the function. At last, we cleaned all the data and get one data frame containing all of the variables.

1.2 Model

In this project, we use a linear regression model and input all the variables we have as independents except profit as the only dependent. Because too many variables could result in a high value of R square but low value of adjusted R square, we select the AIC (Akaike information criterion) method to optimize our model and the selection of variables.

AIC is an estimator of the relative quality of statistical models for a given set of data and could measure the quality of the model. On the other hand, AIC could reward the goodness of fit and reduce the unnecessary variables from the model. Therefore we could use AIC to optimize it and get an effective and meaningful model.

In R, there is a package called MASS. In MASS, we could use the function called stepAIC to do AIC optimization on our model. And we use stepwise search to select the fitted variables on our model.

2. Hypothesis

In our model, we are aiming to find out what determines the agriculture profit of household. so we select the profit_per_acre variable as the dependent variable in this model and pick up the other variables as the independent variables. The independent variables could be divided into two categories: numerical variables and categorical variables. And the categorical variables could also be divided into two categories: dummy categorical variables, whose value is either 0 or 1, and factor categorical variables. We could find out their influence on unit agriculture profit of household.

In terms of hypothesis, we divide all the variables into 2 parts(dummy/ factor variables and numerical variables) to discuss.

For dummy/factor variables, there are only 2 values in dummy/factor variables, which is either 1 (or number n as a factor) or 0. The value of 1 represents the answer to the question (X0 = 1) is chosen, otherwise, the answer to the question (X0 = 0) isn't the fact for the household.

H0: on average, households who have the fact of X0 have the same profit as the one without the fact of X0.

H1: on average, households who have the fact of X0 have more (less) profit as the one without the fact of X0.

For numeric variables(X1):

H0: X1 does not have any effect on the household agriculture profit per acre

H1: X1 has an effect on the household agriculture profit per acre

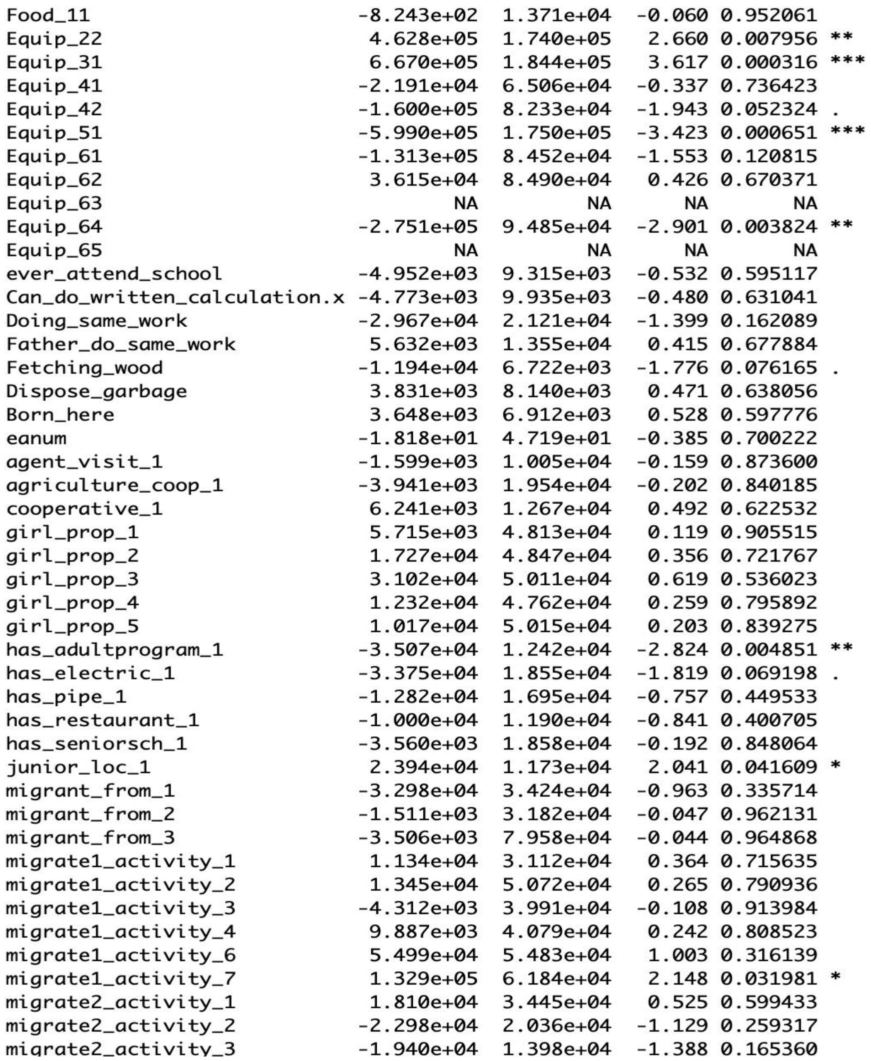
To test the model, we employ both linear regression and stepwise regression to automatically select the best model based on our full model(Zhang Z, 2016). In this model, we first use all the variables and lm() function to calculate our full model, then we applied stepAIC() function to perform selection procedure automatically by statistics packages and reject some unrelated variables, showing the best subset of the full model.

3. Result and Diagnostics

3.1 Result

Our full model yields the following results:

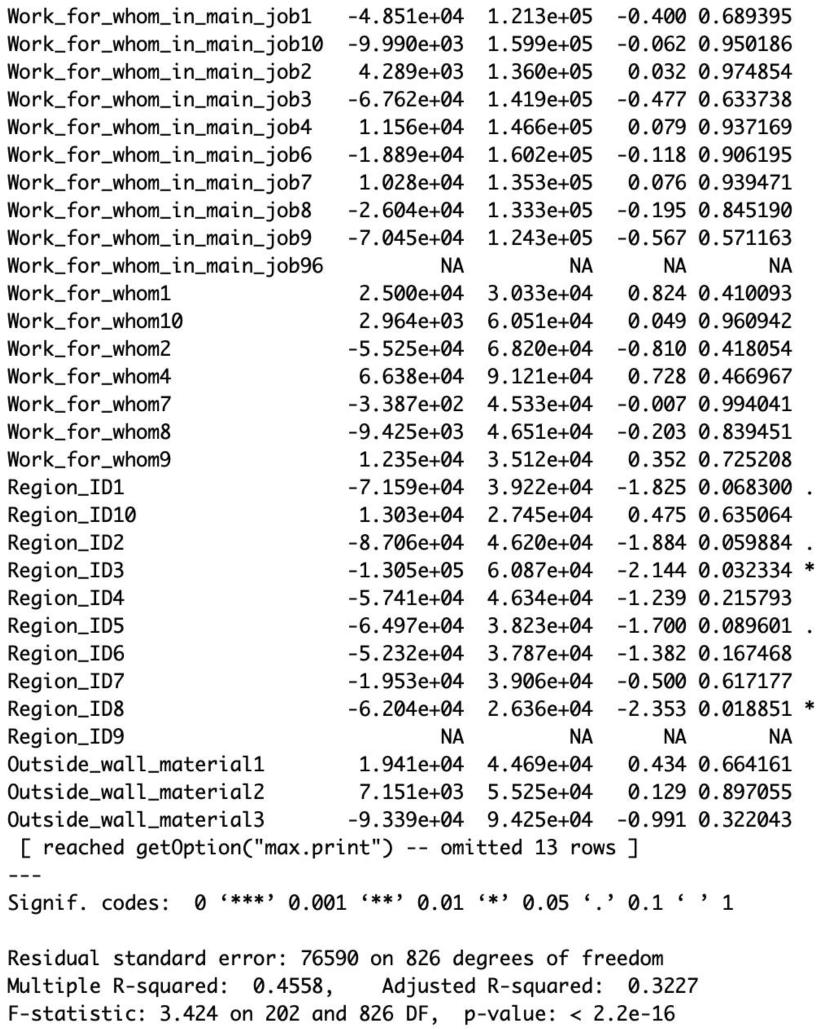
```
Call:
lm(formula = profit_per_acre ~ ., data = full)

Residuals:
    Min      1Q  Median      3Q     Max 
-238354  -45538  -10475   37213  260861 

Coefficients: (10 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  2.426e+05  2.658e+05   0.913 0.361569    
nh                          -2.645e+02  4.417e+02  -0.599 0.549465    
clust                       -6.985e+01  5.639e+01  -1.239 0.215800    
Crop_18                      4.376e+03  1.739e+04   0.252 0.801371    
Crop_19                      2.821e+04  1.374e+04   2.053 0.040398 *  
Crop_22                     -2.144e+03  1.375e+04  -0.156 0.876149    
Crop_23                      1.218e+04  1.484e+04   0.821 0.411959    
Crop_24                      6.879e+03  2.386e+04   0.288 0.773211    
Oil_Palm                     9.499e+03  9.000e+03   1.055 0.291529    
Plantain                     1.775e+04  1.092e+04   1.626 0.104299    
Banana                      -3.099e+03  9.818e+03  -0.316 0.752372    
Orange                       1.547e+04  9.892e+03   1.564 0.118214    
Cola_nut                     6.409e+04  1.779e+04   3.604 0.000333 ***
Pineapple                    4.619e+03  1.105e+04   0.418 0.676000    
Cassava                      2.778e+04  1.146e+04   2.424 0.015547 *  
Yam                         -5.736e+03  1.020e+04  -0.562 0.574162    
Cocoyam                     -2.850e+03  1.037e+04  -0.275 0.783498    
Potatoes                     6.717e+03  1.740e+04   0.386 0.699629    
Tomatoes                     2.563e+03  7.876e+03   0.325 0.744911    
Okro                        -9.116e+03  8.445e+03  -1.080 0.280672    
Garden_egg                   2.481e+02  9.557e+03   0.026 0.979292    
Pepper                       1.422e+04  8.343e+03   1.705 0.088604 .  
Leafy_vege                  -9.292e+03  9.203e+03  -1.010 0.312986    
Other_vege                   8.420e+03  1.357e+04   0.620 0.535238    
Onion                        4.843e+03  1.338e+04   0.362 0.717464    
Avacado_pear                -5.688e+02  1.033e+04  -0.055 0.956124    
Mango                       -2.953e+04  1.071e+04  -2.758 0.005936 ** 
Pawpaw                       7.352e+03  1.053e+04   0.698 0.485193    
Food_02                     -5.936e+03  2.509e+04  -0.237 0.813065    
Food_03                     -8.497e+03  1.674e+04  -0.508 0.611902    
Food_04                     -2.055e+04  1.618e+04  -1.271 0.204237    
Food_05                      4.478e+03  1.447e+04   0.310 0.756970    
Food_06                     -9.924e+03  1.867e+04  -0.532 0.595143    
Food_07                      2.821e+04  3.131e+04   0.901 0.367822    
Food_08                      5.777e+03  1.706e+04   0.339 0.734920    
Food_09                      2.892e+03  2.058e+04   0.141 0.888254    
Food_10                      1.596e+03  2.024e+04   0.079 0.937163    
```



```
Food_11                        -8.243e+02  1.371e+04  -0.060 0.952061    
Equip_22                        4.628e+05  1.740e+05   2.660 0.007956 ** 
Equip_31                        6.670e+05  1.844e+05   3.617 0.000316 ***
Equip_41                       -2.191e+04  6.506e+04  -0.337 0.736423    
Equip_42                       -1.600e+05  8.233e+04  -1.943 0.052324 .  
Equip_51                       -5.990e+05  1.750e+05  -3.423 0.000651 ***
Equip_61                       -1.313e+05  8.452e+04  -1.553 0.120815    
Equip_62                        3.615e+04  8.490e+04   0.426 0.670371    
Equip_63                               NA         NA      NA       NA    
Equip_64                       -2.751e+05  9.485e+04  -2.901 0.003824 ** 
Equip_65                               NA         NA      NA       NA    
ever_attend_school             -4.952e+03  9.315e+03  -0.532 0.595117    
Can_do_written_calculation.x -4.773e+03  9.935e+03  -0.480 0.631041    
Doing_same_work                -2.967e+04  2.121e+04  -1.399 0.162089    
Father_do_same_work             5.632e+03  1.355e+04   0.415 0.677884    
Fetching_wood                  -1.194e+04  6.722e+03  -1.776 0.076165 .  
Dispose_garbage                 3.831e+03  8.140e+03   0.471 0.638056    
Born_here                       3.648e+03  6.912e+03   0.528 0.597776    
eanum                          -1.818e+01  4.719e+01  -0.385 0.700222    
agent_visit_1                  -1.599e+03  1.005e+04  -0.159 0.873600    
agriculture_coop_1             -3.941e+03  1.954e+04  -0.202 0.840185    
cooperative_1                   6.241e+03  1.267e+04   0.492 0.622532    
girl_prop_1                     5.715e+03  4.813e+04   0.119 0.905515    
girl_prop_2                     1.727e+04  4.847e+04   0.356 0.721767    
girl_prop_3                     3.102e+04  5.011e+04   0.619 0.536023    
girl_prop_4                     1.232e+04  4.762e+04   0.259 0.795892    
girl_prop_5                     1.017e+04  5.015e+04   0.203 0.839275    
has_adultprogram_1             -3.507e+04  1.242e+04  -2.824 0.004851 ** 
has_electric_1                 -3.375e+04  1.855e+04  -1.819 0.069198 .  
has_pipe_1                     -1.282e+04  1.695e+04  -0.757 0.449533    
has_restaurant_1               -1.000e+04  1.190e+04  -0.841 0.400705    
has_seniorsch_1                -3.560e+03  1.858e+04  -0.192 0.848064    
junior_loc_1                    2.394e+04  1.173e+04   2.041 0.041609 *  
migrant_from_1                 -3.298e+04  3.424e+04  -0.963 0.335714    
migrant_from_2                 -1.511e+03  3.182e+04  -0.047 0.962131    
migrant_from_3                 -3.506e+03  7.958e+04  -0.044 0.964868    
migrate1_activity_1             1.134e+04  3.112e+04   0.364 0.715635    
migrate1_activity_2             1.345e+04  5.072e+04   0.265 0.790936    
migrate1_activity_3            -4.312e+03  3.991e+04  -0.108 0.913984    
migrate1_activity_4             9.887e+03  4.079e+04   0.242 0.808523    
migrate1_activity_6             5.499e+04  5.483e+04   1.003 0.316139    
migrate1_activity_7             1.329e+05  6.184e+04   2.148 0.031981 *  
migrate2_activity_1             1.810e+04  3.445e+04   0.525 0.599433    
migrate2_activity_2            -2.298e+04  2.036e+04  -1.129 0.259317    
migrate2_activity_3            -1.940e+04  1.398e+04  -1.388 0.165360    
```
]

```
migrate2_activity_7          -1.911e+05  1.023e+05  -1.868 0.062104 .
migrate2_activity_8                  NA         NA      NA       NA
move_1                        5.153e+04  2.590e+04   1.989 0.046990 *
move_2                        4.144e+04  2.626e+04   1.578 0.114999
move_3                        7.074e+04  5.103e+04   1.386 0.166034
rice_husking_1                3.430e+04  1.908e+04   1.798 0.072595 .
s2q17_1                       1.577e+03  1.935e+04   0.081 0.935065
s2q27_1                      -3.300e+03  1.496e+04  -0.221 0.825529
s3q1_1                        1.747e+04  1.539e+04   1.135 0.256819
s3q23_1                              NA         NA      NA       NA
s5bq15_1                     -3.975e+03  1.127e+04  -0.353 0.724476
s5bq17_1                      2.306e+04  1.955e+04   1.179 0.238548
s5bq23_1                     -4.633e+02  1.416e+04  -0.033 0.973896
s5bq5_1                       9.067e+03  1.571e+04   0.577 0.563887
sch_private_1                -2.134e+04  6.641e+04  -0.321 0.748063
sch_private_2                -7.462e+04  7.741e+04  -0.964 0.335360
school_gender_1                      NA         NA      NA       NA
school_gender_2              -2.807e+04  5.195e+04  -0.540 0.589122
school_gender_3                      NA         NA      NA       NA
sharecropper_1               -2.330e+04  1.594e+04  -1.462 0.144191
sharecropper_prop_1           1.678e+03  1.404e+04   0.119 0.904921
sharecropper_prop_2          -2.073e+04  1.636e+04  -1.267 0.205496
sharecropper_prop_3           3.238e+04  1.751e+04   1.849 0.064823 .
use_insecticides_1            2.133e+04  1.234e+04   1.729 0.084209 .
s1q1                          3.222e+00  4.645e+00   0.694 0.488024
s2q22                        -1.201e+02  1.478e+02  -0.812 0.416868
s2q25                        -4.608e+02  4.843e+02  -0.951 0.341631
s3q12                         6.205e+02  6.433e+02   0.965 0.335063
distance_seniorschool        -6.044e+01  6.295e+01  -0.960 0.337314
distance_agricenter           2.250e+02  1.230e+02   1.829 0.067692 .
tracer_num                   -3.202e+03  3.782e+03  -0.847 0.397460
Honey                        -2.853e-01  1.057e+00  -0.270 0.787421
s8aq6                         4.573e-97  1.392e-96   0.328 0.742655
s8aq10                        2.053e-96  3.324e-96   0.618 0.536949
s8aq14                        9.603e-97  7.324e-97   1.311 0.190178
s8aq15                               NA         NA      NA       NA
Harvest_Jan                  -3.480e-97  3.703e-97  -0.940 0.347561
Harvest_Feb                   1.314e-96  4.703e-97   2.795 0.005314 **
Harvest_Mar                   1.477e-97  4.484e-97   0.329 0.741905
Harvest_Apr                  -7.245e-97  4.577e-97  -1.583 0.113849
```

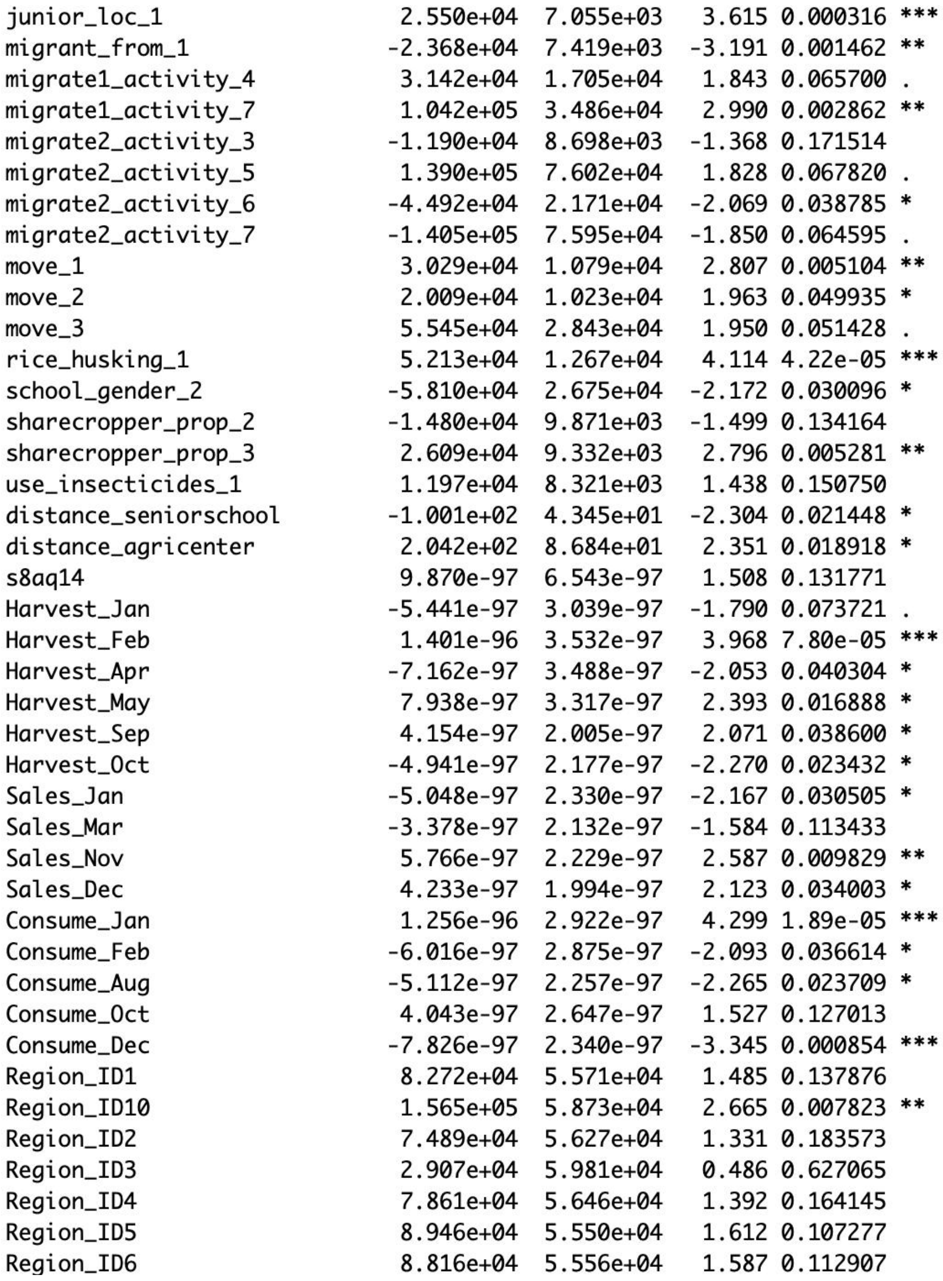
```
Harvest_May                   8.956e-97  4.560e-97   1.964 0.049864 *
Harvest_Jun                   2.636e-97  3.685e-97   0.715 0.474621
Harvest_Jul                   1.962e-97  3.236e-97   0.606 0.544522
Harvest_Aug                   1.657e-97  2.808e-97   0.590 0.555121
Harvest_Sep                   5.850e-97  2.931e-97   1.996 0.046303 *
Harvest_Oct                  -3.870e-97  2.843e-97  -1.361 0.173883
Harvest_Nov                  -2.060e-98  2.837e-97  -0.073 0.942120
Harvest_Dec                  -1.479e-99  2.869e-97  -0.005 0.995890
Sales_Jan                    -5.017e-97  3.118e-97  -1.609 0.107974
Sales_Feb                    -7.995e-99  3.448e-97  -0.023 0.981504
Sales_Mar                    -2.802e-97  3.374e-97  -0.831 0.406381
Sales_Apr                    -1.309e-97  3.424e-97  -0.382 0.702299
Sales_May                    -2.294e-97  3.752e-97  -0.612 0.540969
Sales_Jun                    -2.659e-97  3.875e-97  -0.686 0.492748
Sales_Jul                     7.543e-98  3.894e-97   0.194 0.846464
Sales_Aug                    -3.319e-97  3.440e-97  -0.965 0.334926
Sales_Sep                    -2.054e-97  3.402e-97  -0.604 0.546160
Sales_Oct                     7.074e-98  3.063e-97   0.231 0.817430
Sales_Nov                     3.877e-97  3.052e-97   1.270 0.204352
Sales_Dec                     4.951e-97  2.744e-97   1.804 0.071521 .
Consume_Jan                   1.047e-96  3.393e-97   3.085 0.002106 **
Consume_Feb                  -4.574e-97  3.674e-97  -1.245 0.213425
Consume_Mar                  -2.943e-97  2.819e-97  -1.044 0.296870
Consume_Apr                   1.578e-97  2.679e-97   0.589 0.555880
Consume_May                  -1.037e-97  2.478e-97  -0.418 0.675723
Consume_Jun                  -2.480e-97  3.032e-97  -0.818 0.413595
Consume_Jul                   1.122e-97  3.061e-97   0.366 0.714183
Consume_Aug                  -5.983e-97  3.329e-97  -1.798 0.072613 .
Consume_Sep                   3.668e-98  4.074e-97   0.090 0.928279
Consume_Oct                   4.502e-97  3.973e-97   1.133 0.257478
Consume_Nov                  -1.129e-97  3.440e-97  -0.328 0.742795
Consume_Dec                  -8.135e-97  3.014e-97  -2.699 0.007097 **
Employment_status1            1.076e+04  1.360e+05   0.079 0.936968
Employment_status2            8.820e+03  1.287e+05   0.069 0.945394
Employment_status3            7.290e+04  1.255e+05   0.581 0.561555
Employment_status4            1.167e+04  1.827e+05   0.064 0.949091
Employment_status5            6.445e+04  1.232e+05   0.523 0.601048
Employment_status6            4.390e+04  1.263e+05   0.347 0.728370
Employment_status7            3.715e+04  1.083e+05   0.343 0.731737
Read_Ghanaian_language.x1     1.055e+05  8.701e+04   1.213 0.225464
Read_Ghanaian_language.x2     1.070e+05  8.683e+04   1.233 0.218052
Read_Ghanaian_language.x3     8.473e+04  9.007e+04   0.941 0.347128
Read_Ghanaian_language.x4     1.112e+05  9.237e+04   1.204 0.228899
Read_Ghanaian_language.x5     1.252e+05  9.025e+04   1.387 0.165863
Read_Ghanaian_language.x8     1.232e+05  8.667e+04   1.421 0.155709
```



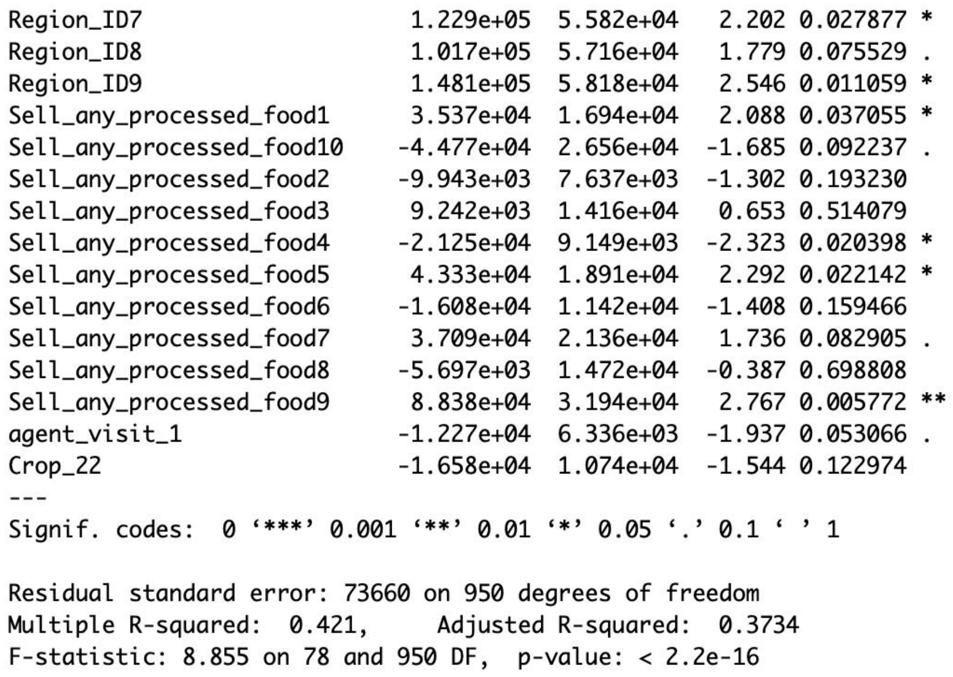
```
Work_for_whom_in_main_job1   -4.851e+04  1.213e+05  -0.400 0.689395
Work_for_whom_in_main_job10  -9.990e+03  1.599e+05  -0.062 0.950186
Work_for_whom_in_main_job2    4.289e+03  1.360e+05   0.032 0.974854
Work_for_whom_in_main_job3   -6.762e+04  1.419e+05  -0.477 0.633738
Work_for_whom_in_main_job4    1.156e+04  1.466e+05   0.079 0.937169
Work_for_whom_in_main_job6   -1.889e+04  1.602e+05  -0.118 0.906195
Work_for_whom_in_main_job7    1.028e+04  1.353e+05   0.076 0.939471
Work_for_whom_in_main_job8   -2.604e+04  1.333e+05  -0.195 0.845190
Work_for_whom_in_main_job9   -7.045e+04  1.243e+05  -0.567 0.571163
Work_for_whom_in_main_job96          NA         NA      NA       NA
Work_for_whom1                2.500e+04  3.033e+04   0.824 0.410093
Work_for_whom10               2.964e+03  6.051e+04   0.049 0.960942
Work_for_whom2               -5.525e+04  6.820e+04  -0.810 0.418054
Work_for_whom4                6.638e+04  9.121e+04   0.728 0.466967
Work_for_whom7               -3.387e+02  4.533e+04  -0.007 0.994041
Work_for_whom8               -9.425e+03  4.651e+04  -0.203 0.839451
Work_for_whom9                1.235e+04  3.512e+04   0.352 0.725208
Region_ID1                   -7.159e+04  3.922e+04  -1.825 0.068300 .
Region_ID10                   1.303e+04  2.745e+04   0.475 0.635064
Region_ID2                   -8.706e+04  4.620e+04  -1.884 0.059884 .
Region_ID3                   -1.305e+05  6.087e+04  -2.144 0.032334 *
Region_ID4                   -5.741e+04  4.634e+04  -1.239 0.215793
Region_ID5                   -6.497e+04  3.823e+04  -1.700 0.089601 .
Region_ID6                   -5.232e+04  3.787e+04  -1.382 0.167468
Region_ID7                   -1.953e+04  3.906e+04  -0.500 0.617177
Region_ID8                   -6.204e+04  2.636e+04  -2.353 0.018851 *
Region_ID9                           NA         NA      NA       NA
Outside_wall_material1        1.941e+04  4.469e+04   0.434 0.664161
Outside_wall_material2        7.151e+03  5.525e+04   0.129 0.897055
Outside_wall_material3       -9.339e+04  9.425e+04  -0.991 0.322043
 [ reached getOption("max.print") -- omitted 13 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76590 on 826 degrees of freedom
Multiple R-squared:  0.4558,    Adjusted R-squared:  0.3227
F-statistic: 3.424 on 202 and 826 DF,  p-value: < 2.2e-16
```

Because the difference between R-squared and the adjusted R-squared differs a lot, so we choose to use AIC model to improve our model. Below is the result run by AIC model.

```
Call:
lm(formula = profit_per_acre ~ clust + Crop_19 + Plantain + Orange +
    Cola_nut + Cassava + Okro + Pepper + Mango + Food_04 + Equip_22 +
    Equip_31 + Equip_42 + Equip_51 + Equip_61 + Equip_64 + Can_do_written_calculation.x +
    Doing_same_work + Father_do_same_work + Fetching_wood + has_adultprogram_1 +
    has_electric_1 + junior_loc_1 + migrant_from_1 + migrate1_activity_4 +
    migrate1_activity_7 + migrate2_activity_3 + migrate2_activity_5 +
    migrate2_activity_6 + migrate2_activity_7 + move_1 + move_2 +
    move_3 + rice_husking_1 + school_gender_2 + sharecropper_prop_2 +
    sharecropper_prop_3 + use_insecticides_1 + distance_seniorschool +
    distance_agricenter + s8aq14 + Harvest_Jan + Harvest_Feb +
    Harvest_Apr + Harvest_May + Harvest_Sep + Harvest_Oct + Sales_Jan +
    Sales_Mar + Sales_Nov + Sales_Dec + Consume_Jan + Consume_Feb +
    Consume_Aug + Consume_Oct + Consume_Dec + Region_ID + Sell_any_processed_food +
    agent_visit_1 + Crop_22, data = full)

Residuals:
    Min      1Q  Median      3Q     Max
-240915  -49083  -11620   35598  236858

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.825e+05  1.117e+05   2.529 0.011598 *
clust                       -6.945e+01  2.189e+01  -3.172 0.001560 **
Crop_19                      1.868e+04  8.058e+03   2.318 0.020639 *
Plantain                     1.179e+04  8.061e+03   1.462 0.143976
Orange                       1.683e+04  8.020e+03   2.099 0.036090 *
Cola_nut                     5.843e+04  1.490e+04   3.921 9.44e-05 ***
Cassava                      3.028e+04  8.137e+03   3.721 0.000210 ***
Okro                        -1.095e+04  6.900e+03  -1.587 0.112898
Pepper                       1.524e+04  6.924e+03   2.200 0.028023 *
Mango                       -2.303e+04  9.015e+03  -2.555 0.010778 *
Food_04                     -1.932e+04  1.144e+04  -1.688 0.091659 .
Equip_22                     5.046e+05  1.319e+05   3.826 0.000138 ***
Equip_31                     6.311e+05  1.529e+05   4.127 3.99e-05 ***
Equip_42                    -1.500e+05  7.562e+04  -1.984 0.047563 *
Equip_51                    -5.978e+05  1.532e+05  -3.902 0.000102 ***
Equip_61                    -1.322e+05  7.567e+04  -1.747 0.080945 .
Equip_64                    -2.735e+05  7.713e+04  -3.546 0.000410 ***
Can_do_written_calculation.x -1.180e+04  5.259e+03  -2.243 0.025132 *
Doing_same_work             -2.205e+04  1.078e+04  -2.046 0.041023 *
Father_do_same_work          9.868e+03  7.246e+03   1.362 0.173568
Fetching_wood               -1.112e+04  5.457e+03  -2.038 0.041843 *
has_adultprogram_1          -2.240e+04  7.974e+03  -2.809 0.005067 **
has_electric_1              -3.566e+04  8.668e+03  -4.114 4.23e-05 ***
```



```
junior_loc_1              2.550e+04  7.055e+03   3.615 0.000316 ***
migrant_from_1           -2.368e+04  7.419e+03  -3.191 0.001462 **
migrate1_activity_4       3.142e+04  1.705e+04   1.843 0.065700 .
migrate1_activity_7       1.042e+05  3.486e+04   2.990 0.002862 **
migrate2_activity_3      -1.190e+04  8.698e+03  -1.368 0.171514
migrate2_activity_5       1.390e+05  7.602e+04   1.828 0.067820 .
migrate2_activity_6      -4.492e+04  2.171e+04  -2.069 0.038785 *
migrate2_activity_7      -1.405e+05  7.595e+04  -1.850 0.064595 .
move_1                    3.029e+04  1.079e+04   2.807 0.005104 **
move_2                    2.009e+04  1.023e+04   1.963 0.049935 *
move_3                    5.545e+04  2.843e+04   1.950 0.051428 .
rice_husking_1            5.213e+04  1.267e+04   4.114 4.22e-05 ***
school_gender_2          -5.810e+04  2.675e+04  -2.172 0.030096 *
sharecropper_prop_2      -1.480e+04  9.871e+03  -1.499 0.134164
sharecropper_prop_3       2.609e+04  9.332e+03   2.796 0.005281 **
use_insecticides_1        1.197e+04  8.321e+03   1.438 0.150750
distance_seniorschool    -1.001e+02  4.345e+01  -2.304 0.021448 *
distance_agricenter       2.042e+02  8.684e+01   2.351 0.018918 *
s8aq14                    9.870e-97  6.543e-97   1.508 0.131771
Harvest_Jan              -5.441e-97  3.039e-97  -1.790 0.073721 .
Harvest_Feb               1.401e-96  3.532e-97   3.968 7.80e-05 ***
Harvest_Apr              -7.162e-97  3.488e-97  -2.053 0.040304 *
Harvest_May               7.938e-97  3.317e-97   2.393 0.016888 *
Harvest_Sep               4.154e-97  2.005e-97   2.071 0.038600 *
Harvest_Oct              -4.941e-97  2.177e-97  -2.270 0.023432 *
Sales_Jan                -5.048e-97  2.330e-97  -2.167 0.030505 *
Sales_Mar                -3.378e-97  2.132e-97  -1.584 0.113433
Sales_Nov                 5.766e-97  2.229e-97   2.587 0.009829 **
Sales_Dec                 4.233e-97  1.994e-97   2.123 0.034003 *
Consume_Jan               1.256e-96  2.922e-97   4.299 1.89e-05 ***
Consume_Feb              -6.016e-97  2.875e-97  -2.093 0.036614 *
Consume_Aug              -5.112e-97  2.257e-97  -2.265 0.023709 *
Consume_Oct               4.043e-97  2.647e-97   1.527 0.127013
Consume_Dec              -7.826e-97  2.340e-97  -3.345 0.000854 ***
Region_ID1                8.272e+04  5.571e+04   1.485 0.137876
Region_ID10               1.565e+05  5.873e+04   2.665 0.007823 **
Region_ID2                7.489e+04  5.627e+04   1.331 0.183573
Region_ID3                2.907e+04  5.981e+04   0.486 0.627065
Region_ID4                7.861e+04  5.646e+04   1.392 0.164145
Region_ID5                8.946e+04  5.550e+04   1.612 0.107277
Region_ID6                8.816e+04  5.556e+04   1.587 0.112907
```



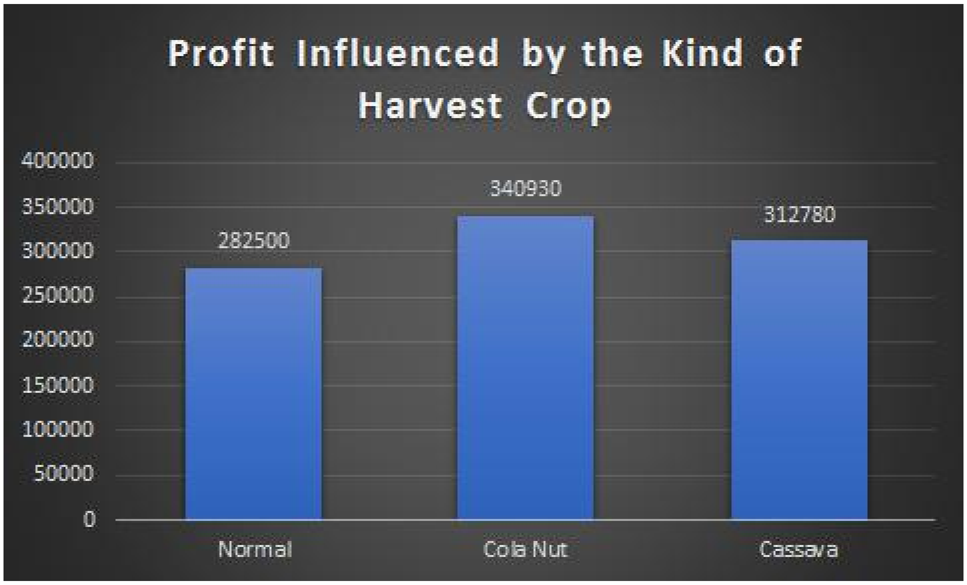
```
Region_ID7                  1.229e+05  5.582e+04   2.202 0.027877 *
Region_ID8                  1.017e+05  5.716e+04   1.779 0.075529 .
Region_ID9                  1.481e+05  5.818e+04   2.546 0.011059 *
Sell_any_processed_food1    3.537e+04  1.694e+04   2.088 0.037055 *
Sell_any_processed_food10  -4.477e+04  2.656e+04  -1.685 0.092237 .
Sell_any_processed_food2   -9.943e+03  7.637e+03  -1.302 0.193230
Sell_any_processed_food3    9.242e+03  1.416e+04   0.653 0.514079
Sell_any_processed_food4   -2.125e+04  9.149e+03  -2.323 0.020398 *
Sell_any_processed_food5    4.333e+04  1.891e+04   2.292 0.022142 *
Sell_any_processed_food6   -1.608e+04  1.142e+04  -1.408 0.159466
Sell_any_processed_food7    3.709e+04  2.136e+04   1.736 0.082905 .
Sell_any_processed_food8   -5.697e+03  1.472e+04  -0.387 0.698808
Sell_any_processed_food9    8.838e+04  3.194e+04   2.767 0.005772 **
agent_visit_1              -1.227e+04  6.336e+03  -1.937 0.053066 .
Crop_22                    -1.658e+04  1.074e+04  -1.544 0.122974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73660 on 950 degrees of freedom
Multiple R-squared:  0.421,     Adjusted R-squared:  0.3734
F-statistic: 8.855 on 78 and 950 DF,  p-value: < 2.2e-16
```

The F-statistic of 8.855 and its associated p-value that is less than 0.001 allow us to reject the null hypothesis that all of our explanatory variables have zero effect on the dependent variable. Therefore, we can conclude at the 0.1% significant level that our full model is valid in predicting agricultural profit. In addition, the coefficient of determination(R^2) tells that the variation in our explanatory variables explains 42.1%(37.3% after adjusting to the number of explanatory variables) of the variation in agricultural profit. Among the 73 explanatory variables in the model, the number of variables which we have enough evidence to infer that they are related to agricultural profit is 45 at the 10% level, 35 at the 5% level, 15 at the 1% level, and 9 at the 0.1% level.

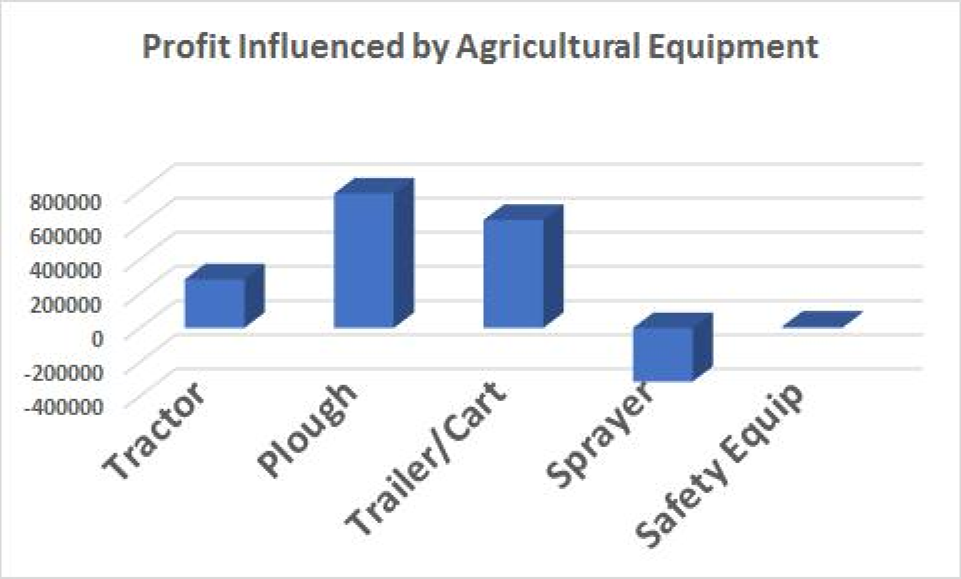
To present more informative interpretations, in this paper, we will only discuss details regarding the effects of explanatory variables that are statistically significant at the 1% level on the agricultural profit produced by households in different areas of Ghana. It turns out that all of those explanatory variables of a significance level of 1% are dummy variables as a result of our data cleaning. The information they contain falls into 7 categories: types of crops harvested, types of agriculture equipment owned, local education, local migration, local electricity availability, the proportion of local sharecroppers, and sale of processed crop/fish.

To prevent redundancy, note that all coming interpretations follow the standards:

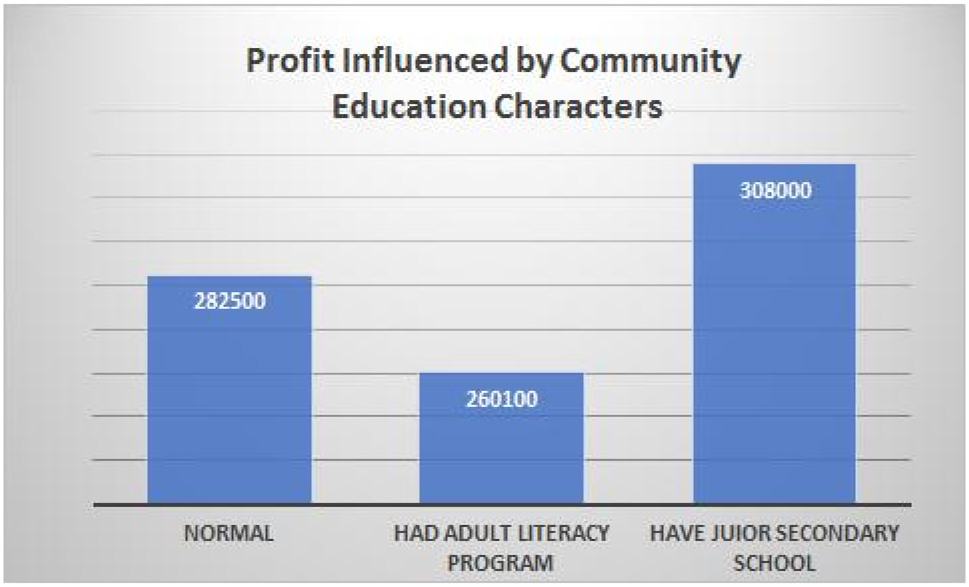
- The result is significant at a confidence level of 99%.
- The comparison is done by holding other variables constant.
- The unit of agriculture profit is Ghanaian cedi per acre.



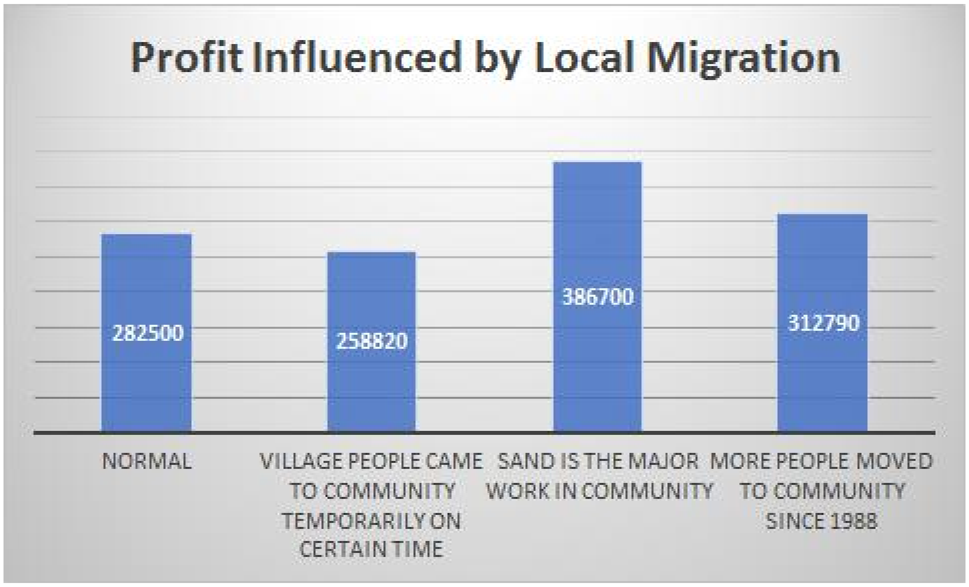
The harvest of cola nuts and cassava both seem to have a positive impact on the dependent variable. More specifically, households that harvested cola nut and cassava are estimated to be associated with an increase of 58340 and 30280 on agriculture profit respectively, in comparison to the ones did not harvest those two kinds of crops. This result is expected due to the fact that cola nut and cassava both belong to Ghana 's main agricultural crops types.



As for agriculture equipment, owning a plough and a trailer/cart are associated with a gain of 504600 and 631100 on profit than owning a tractor. The variable from the community survey, the existence of rice husking machine also has an association with a profit increase of 52130. In contrast, owning sprayer and safety equip is associated with profit decrease of 597800 and 273500. We believe that such results can be explained by the lack of agricultural mechanization in Ghana, giving households little experience with more advanced agricultural equipment, which might interfere with the increase of profit. In addition, mechanically advanced agricultural equipment sometimes needs a longer period of time to show its positive profit effect.



As to local educational condition, the existence(or former existence) of adult literacy program is associated with a profit decrease of 22400, whereas that of junior secondary school is associated with a profit increase of 25500. Being the main future labor force of agricultural activities, teenagers of age around 15-18 are likely to be positively related to agricultural profit. On the other hand, the negative relation between adult literacy program and agricultural profit is also reasonable, since adults usually choose to participate in literacy program usually when they want to acquire jobs other than farming.



Concerning local migration, there exist people who come to this community temporarily during certain times of the year to look for work. If those people are from villages, that's associated with 23680 less profit; but if they are here to do sand-winning, that's associated with 104200 more profit. Also, it is not a surprise that more arrival of immigrants is associated with an additional profit of 30290, as Ghana's main industry was agriculture in 1998-1999.

Notably, electricity availability in the community is negatively associated with the profit, with a decrease of -35660 versus not having electricity or a generator. This is probably because people who have electricity are more economically proficient and therefore tend to gain income from non-agricultural activities.

The proportion of sharecroppers being less than half, in contrast to more than half, is associated with 26090 more profit, which is reasonable considering sharecropping indicates sharing income with sharecroppers.

The last category is processed crops/fish selling. If households sold shea butter in the last 2 weeks, that's associated with a profit increase of 88380.

3.2 Diagnostics

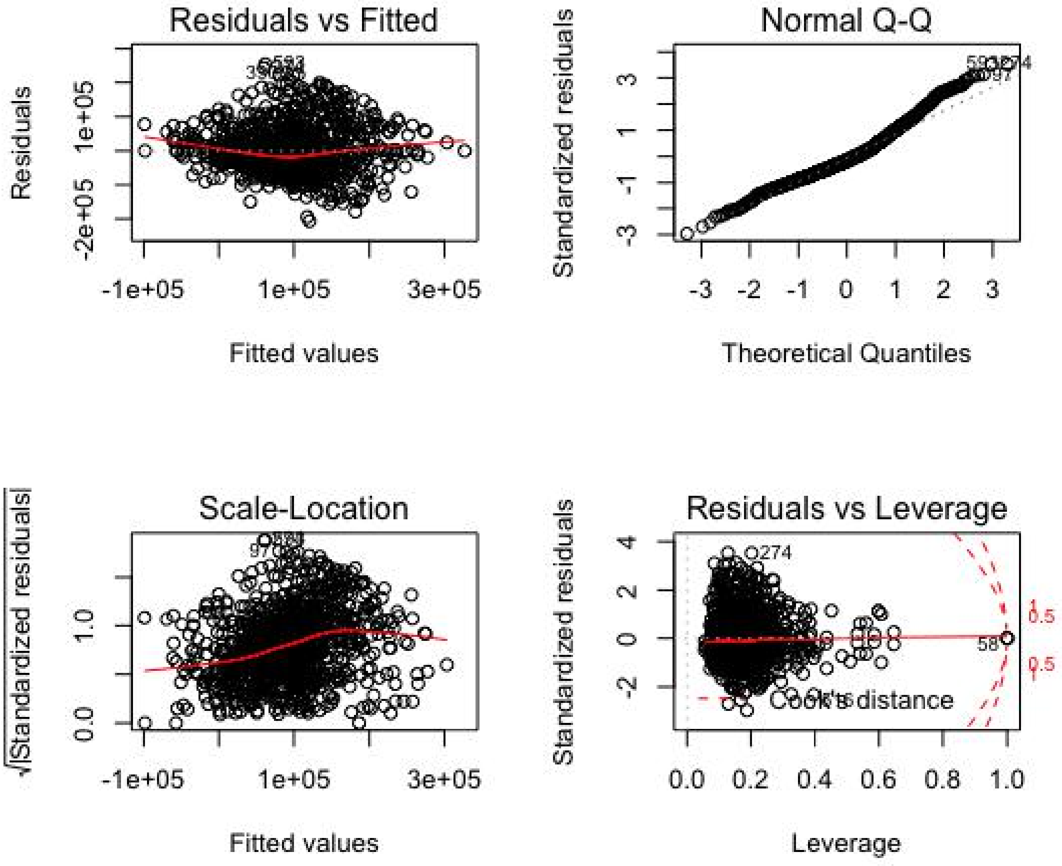


Figure 3.2.1

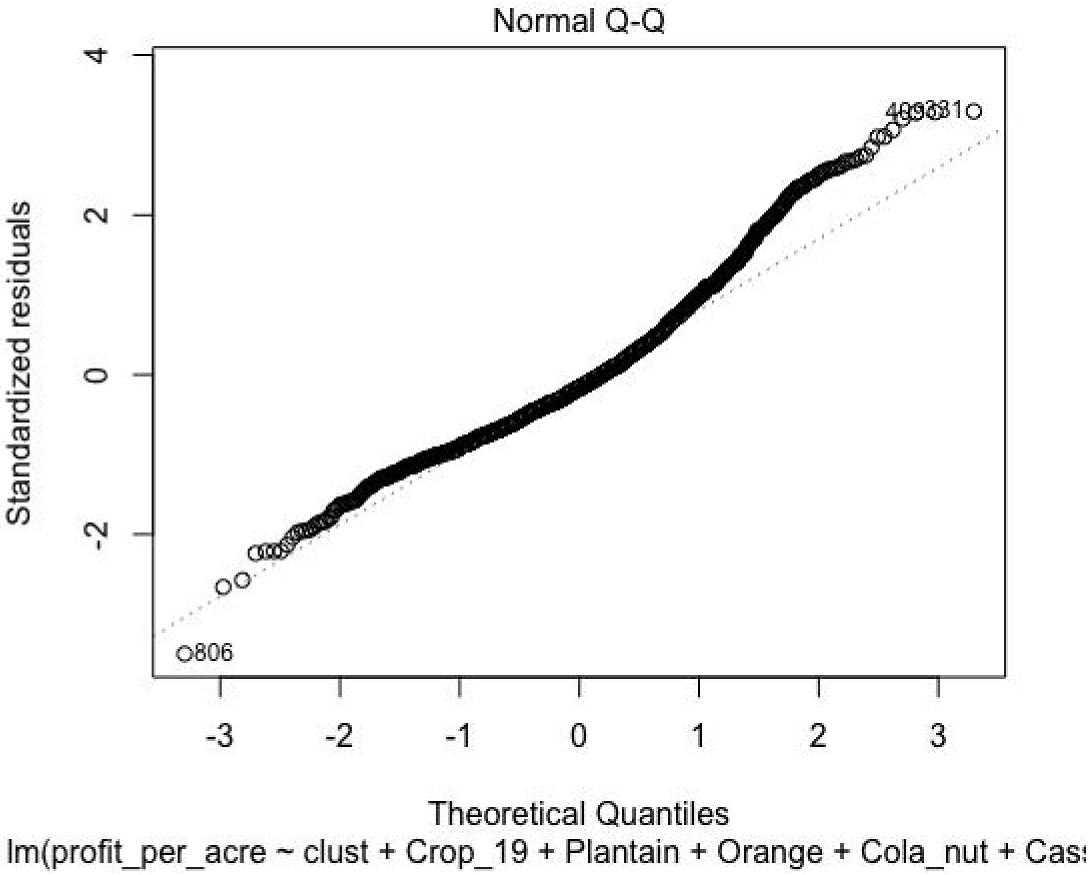


Figure 3.2.2

We tried to plot the residuals to diagnose the quality of the model, and we got the results shown as Figure 3.2.1. From Figure 3.2.2 of the standardized residuals versus the theoretical quantiles, our model seems to be quite satisfactory. The standardized residuals are close to the theoretical quantiles' line which means that the standardized residuals conform to a similar shape to normal distribution but not perfectly fitted somehow. We can also see this from the histogram in figure 3.2.3.

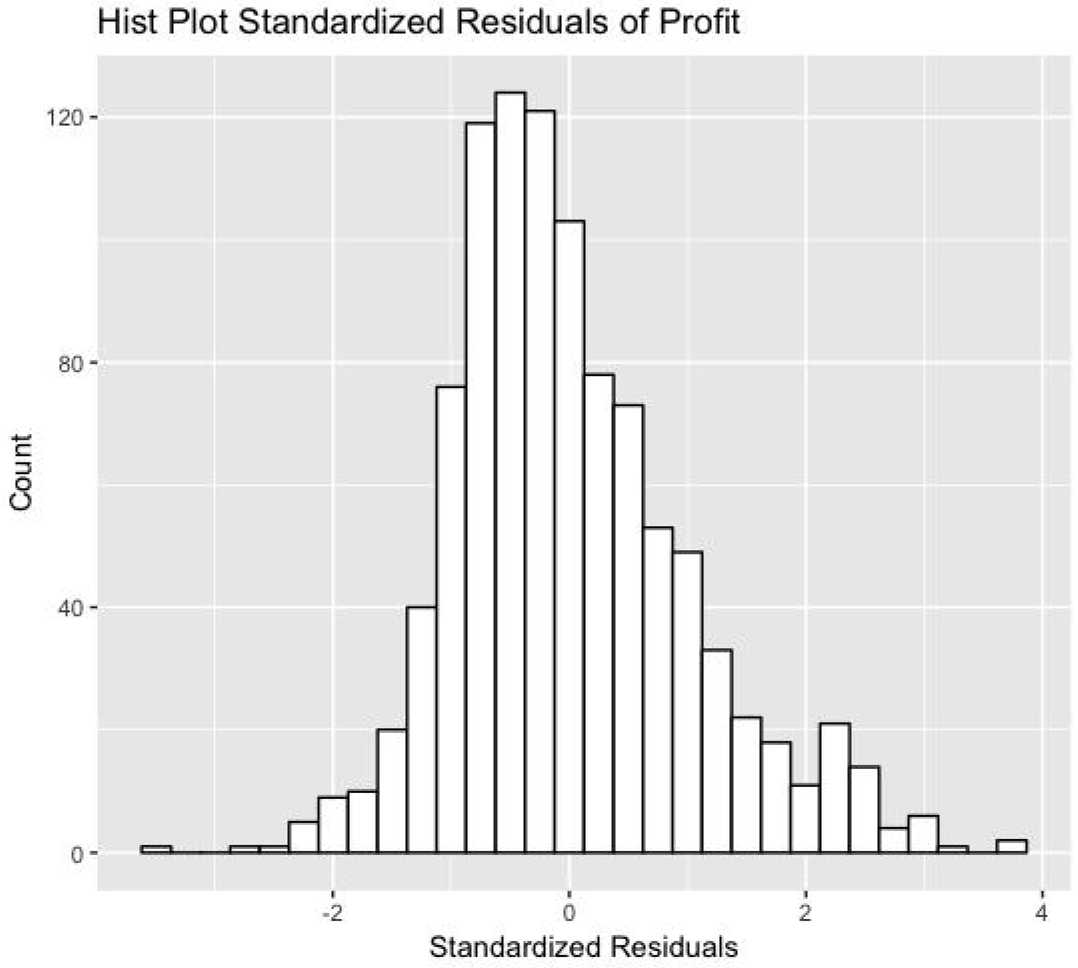


Figure 3.2.3

From this perspective, our model meets one of the requirements for the linear regression that the distribution of the error term is close to normally distributed. However it is a little bit left shifted, which means the standardized residuals of this model is acceptable but not really perfect fitted somehow.

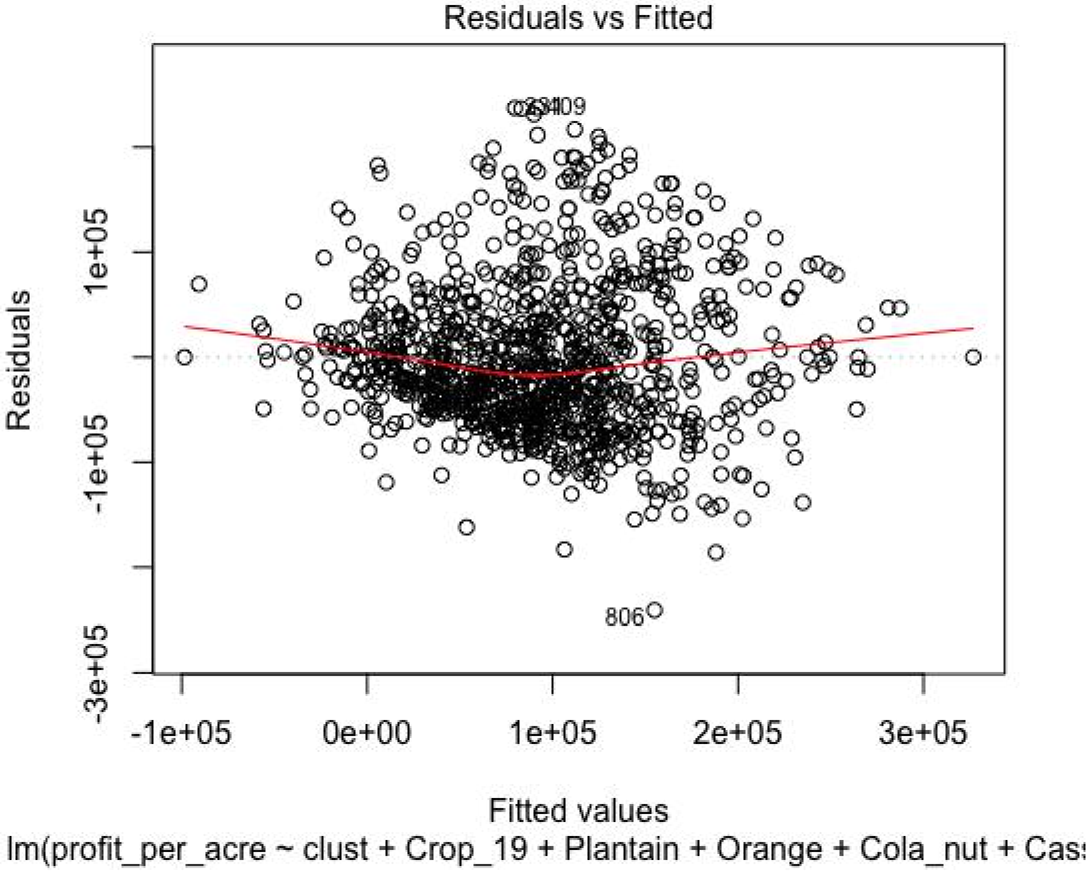


Figure 3.2.4

As for the graph of the residuals and the fitted values, shown as figure 3.2.4, it doesn’t turn out to be a good result. As the fitted value increases, the residuals’ variance is not stable and keeps changing. So the homoscedasticity assumption of the linear regression is not perfectly satisfied by our model.

Reference

Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med 2016;4(7):136. doi: 10.21037/atm.2016.03.35