ACME CORPORATION EXPLORATORY MARKET ANALYSIS

An analysis of agricultural profit and market conditions in Ghana.

Group 07:
Siqi Zhang
Huibo Jia
Qiyu Ye
Mark Russeff

Abstract

This paper attempts to explore what determines agricultural profits in Ghana. The data for the study was sourced from the fourth round of the Ghana Living Standard Survey (GLSS 4). Six multiple linear regression models were estimated to determine the effect of household member information (sex, education, age), local area characteristics, livestock, equipment, crops, market proximity and transportation availability on household agricultural profits. First, an unrestricted model consisting of all agricultural related variables was estimated in order to get a broad picture of the determinates of agricultural profit in Ghana. Then the restricted models were estimated with carefully specified variables to explore a more focused statistical analysis. Our analysis reveals that educational attainment has no statistically significant effect on household agricultural profit; while the forest ecological zone has a significant positive effect on agricultural profit compared to the coastal ecological zone, proximity to a public market has a significant negative effect, and varying types of root crops could significantly influence agricultural profit in either direction.

1. Introduction

The agricultural sector is a significant part of the Ghanaian economy, approximately 2.7 million households in Ghana either own or operate a farm or keep livestock and more than half of those households hire additional labor for their operations. As the agricultural sector in Ghana continues to grow, many multinational agricultural companies are seeking to gain entry to this thriving market. The main purpose of this project is to provide ACME Corporation with a statistical analysis determining the factors that most significantly contribute to household agricultural profits in Ghana so that they can more effectively target their initial sales efforts upon market entry. The board of ACME Corp has requested an analysis of the effect of educational attainment and local characteristics on household profit. However, this project will attempt to identify and quantify any statistically significant variable that could affect household agricultural profit.

2. Data Review and Variable Selection

2.1 Basic information on the GLSS 4

Data for the study was sourced from the Ghana Living Standard Survey (GLSS 4) collected between April 1998 and March 1999. It took household as a key social and economic unit, provides valuable insights into living conditions in Ghana. The GLSS 4 is a nation-wide survey which collected detailed information on a variety of topics, including demographic characteristics of the population, education, health, employment and time use, migration, housing conditions, household agriculture and non-farm businesses. A representative nationwide sample of more than 5,998 households, containing over 25,000 persons, was covered in GLSS 4. Detailed information was collected on all aspects of living conditions, including health, education, employment, housing, agricultural activities, the operation of non-farm establishments, remittances, savings, and credit and assets. It also contains information on farm levels, household level characteristics and socio- demographic characteristics.

Therefore, the GLSS 4 is the most complete and comprehensive resource available to access data pertaining to household agricultural profits and inputs in Ghana.

2.2 Datasets and variable selection

Upon reviewing all documentation related to GLSS 4; The Estimation of Components of Households Incomes and Expenditures (Aggregate), Ghana Living Standards Survey Report of The Fourth Round (GLSS 4), and the Data User's Guide (G4USERSG) were the most useful components in the selection of variables. From the aggregate data, AGRI1 was chosen as the dependent variable for agricultural profits as it incorporated all relevant agricultural income variables minus all expenses. Farm equipment depreciation was then subtracted from AGRI1 to get a more accurate household agriculture profit level. Finally, AGRI1 was divided by land size in acres in order to mitigate the effects of very large or small farms on the profit data, yielding a household agricultural profit per acre figure that was used exclusively moving forward.

Then variables were extracted for use as determinants of agricultural profits from household basic information, household member information, household member education information, household livestock count and types, household agricultural equipment type, household harvested crop count and types, household harvested root count and types, as well as characteristics of the local area, including community, region, transport, and market information. The variables were all tidyed and aggregated to the household lever prior to regression analysis. A large quantity of explanatory variables was initially chosen in order to construct an unrestricted model that would yield a broader picture of the agricultural market in Ghana at the household level. Furthermore, the results of the unrestricted multiple linear regression model played an important role in variable selection for the restricted models later in the analysis.

3. Methodology

3.1 Analysis Steps and Explanations

After reading through the documentation, the data files containing variables related to education, local area characteristics, household livestock/crop/equipment, and community market were identified as the required data sets for analyzing the factors that influence agricultural profits in Ghana. Those data sets were then tidied and aggregated onto the household level and fit to several different multiple linear regression models containing varying combinations of carefully selected variables.

Here is a detailed analysis of steps and explanations with code, output and graphs written in R notebook: Analysis_Steps_and_Explanations

3.2 Models, Hypothesis and Regression Analysis

In order to achieve the objectives of this paper, we used multiple regression method to build models. Six different models were estimated. The first is an unrestricted model for all features we found. The other five models are restricted models for different combinations of variables separately.

Model 1: unrestricted model (UR)

First, we fit an unrestricted model. From the summary of the linear model we can see that there are some variables that's significant on 10% level like ezSavannah and cropcd5, while some others are more significant on 1% level like livstcd6 and cropcd8. Besides these, variables like livstcd5 and rootcd18 are extremely significant on 0.1% level. The R squared is 0.13 which means roughly 13% of variation in profit could be explained by this model, which is not the best but is a good start point. And the F-statistic value is 5.4 on 102 and 3510 DF, which is significant on 1% level (since $F_{0.01,102,3510} = 1.36$). Although during hypothesis test, we found that there are some variables that are highly correlated with others like loc5 and cropcd0. So we decided to pick the significant variables, remove the highly correlated variables and fit a restricted model and see if we could get a even better model.

profit ~ reslan + ez + loc2 + loc5 + loc3 + female + age + avgAge +
maxAge + minAge + educ + market + transport + livstcdTypeCount +
livstcd1 + livstcd2 + livstcd3 + livstcd4 + livstcd5 + livstcd6 +
livstcd7 + livstcd8 + livstcd9 + livstcd10 + livstcd11 +
livstcd12 + equipTypeCount + eqcdown21 + eqcdown22 + eqcdown31 +
eqcdown51 + eqcdown61 + eqcdown62 + eqcdown63 + eqcdown64 +
eqcdown65 + cropTypeCount + cropcd0 + cropcd1 + cropcd2 +
cropcd3 + cropcd4 + cropcd5 + cropcd6 + cropcd8 + cropcd9 +
cropcd10 + cropcd11 + cropcd12 + cropcd13 + cropcd14 + cropcd15 +
cropcd16 + cropcd17 + cropcd18 + cropcd19 + cropcd20 + cropcd21 +
cropcd22 + cropcd23 + cropcd24 + cropcd25 + cropcd26 + cropcd27 +
cropcd28 + cropcd29 + cropcd31 + cropcd32 + cropcd33 + cropcd34 +
cropcd35 + rootTypeCount + rootcd0 + rootcd5 + rootcd6 +
rootcd7 + rootcd8 + rootcd9 + rootcd11 + rootcd14 + rootcd16 +
rootcd18 + rootcd19 + rootcd20 + rootcd21 + rootcd22 + rootcd25 +
rootcd26 + rootcd27 + rootcd29 + rootcd30 + rootcd31 + rootcd33 +
rootcd34 + rootcd35 + rootcd36

Hypothesis:

reslanAkan =reslanEwe=reslanGaAdangbe =reslanDagbani =reslanHausa
reslanOther=reslanUnknown=ezForest=ezSavannah =loc2Rural=
loc5RuralCoastal =loc5RuralForest=loc5RuralSavannah =loc3Rural
=femaleTRUE =age=avgAge =maxAge=minAge= educBasicEducation

=educSecondaryEducation =educTertiaryEducation =educOther =marketTRUE =transportTRUE =livstcdTypeCount =livstcd1=livstcd2=livstcd3=livstcd4 =livstcd5=livstcd6=livstcd7=livstcd8=livstcd9=livstcd10=livstcd11=livstcd12 =equipTypeCount=eqcdown21=eqcdown22 =eqcdown31=eqcdown51=eqcdown61 =eqcdown62=eqcdown63=eqcdown64=eqcdown65=cropTypeCount=cropcd0 =cropcd1=cropcd2=cropcd3=cropcd4=cropcd5=cropcd6=cropcd8=cropcd9 =cropcd10=cropcd11=cropcd12=cropcd13=cropcd14=cropcd15=cropcd16= cropcd17=cropcd18=cropcd19=cropcd20=cropcd21=cropcd22=cropcd23= cropcd24=cropcd25=cropcd26=cropcd27=cropcd28=cropcd29=cropcd31= cropcd32=cropcd33=cropcd34=cropcd35=rootTypeCount =rootcd0=rootcd5= rootcd6=rootcd7=rootcd8=rootcd9=rootcd11=rootcd14=rootcd16= rootcd18= rootcd19=rootcd20=rootcd21=rootcd22=rootcd25=rootcd26=rootcd27=rootcd29 =rootcd30=rootcd31=rootcd33=rootcd34=rootcd35 = rootcd36 = 0

Model 2: restricted model with significant variables (R1)

We picked significant variables and fit a restricted model (shown as below). Looking at the variable coefficients, most of them are significant. For example, rootcd20 (Cocoyam) is significant on 0.1% level with an estimate of 4.45e+04, which means one unit of increase in Cocoyam would lead to 44500 Cedi increase in profit. And rootcd27 (Eggplant) is also significant on 0.1% level but with an negative estimate of -1.61e+04, which means one unit of increase in Eggplant could result in 16100 Cedi decrease in profit. The R squared of this model dropped slightly to 0.12, but the adjusted R squared didn't change much. This is because even though we have less variables in restricted model, most of them are significant and contribute to explanation of variation in profit. In other words, the non-significant variables in unrestricted model does not help much in explaining variation in profit. The F-statistic is 18.44 on 28 and 3584 DF, which is significant on 1% level ($F_{0.01,28,3584} = 1.72$). This is way better than the F-statistic of the unrestricted model and indicates this is a better fit to the data, which is expected because we used only significant variables.

profit ~ reslan + ez + age + market + livstcd5 + livstcd6 + livstcd7 +

livstcd10 + equipTypeCount + eqcdown61 + cropcd5 + cropcd8 +

cropcd11 + cropcd25 + cropcd29 + rootcd7 + rootcd18 + rootcd20 +

rootcd27 + rootcd33 + rootcd36

Hypothesis:

reslanAkan = reslanEwe = reslanGaAdangbe = reslanDagbani = reslanHausa = reslanOther = reslanUnknown = ezForest = ezSavannah = educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = marketTRUE = livstcd5 = livstcd6 = equipTypeCount = cropTypeCount = cropcd8 = cropcd11 = cropcd25 = rootcd8 = rootcd18 = rootcd20 = rootcd27 = rootcd33 = 0

Model 3: restricted model with top features from agricultural characteristics information (TOP)

We are also interested to see how the model will perform if fitted only with top 15 features from agricultural characteristics data, including livestock, crop, root and equipment. We first calculate the correlation of each variable with profit and pick the top 15 variables with higher absolute correlation. We used absolute correlation value because we want both positive and negative correlated variables. The model is shown as below. The model summary shows that there are half a dozen variables that are significant. Although both R squared and adjusted R squared is lower than that of model R1. This is because some of the agricultural characteristics variables does not have enough correlation comparing to the local area and community variables in model R1.

profit ~ cropcd25 + rootcd18 + rootcd20 + rootTypeCount + eqcdown61 +

rootcd36 + livstcd5 + equipTypeCount + rootcd6 + rootcd7 +

livstcd10 + rootcd30 + rootcd5 + rootcd8 + livstcd2

Hypothesis:

cropcd25 = rootcd18 = rootcd20 = rootTypeCount = eqcdown61 = rootcd36 = livstcd5 = equipTypeCount = rootcd6 = rootcd7 = livstcd10 = rootcd30 = rootcd5 = rootcd8 = livstcd2 = 0

Model 4: restricted model with only education and local characteristic variables (R2)

Our curiosity leads us to fitting a model with only education and local characteristic variables. We would like to see without agricultural characteristic information, how does this model perform and if any variable will stand out. The model summary shows that only the marketTRUE is significant on 1% level and R squared is only 0.006 which is very low. Although we notice that loc5 and loc3 are correlated with other variables. So we decided to remove them and fit the model again.

profit ~ educ + ez + loc2 + loc5 + loc3 + market + transport

Hypothesis:

educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = ezForest = ezSavannah = loc2Rural = loc5RuralCoastal = loc5RuralForest = loc5RuralSavannah = loc3Rural = marketTRUE = transportTRUE = 0

Model 5: restricted model removing loc5 and loc3 from R2 (R3)

After removing loc5 and loc3, ezForest becomes significant on 5% level. marketTRUE remains significant on 1% level. Both of their coefficients estimate didn’t change much from model R2. Although the R squared is still very low, and F-statistic is 2.5 with 9 and 3603 DF and p-value is 0.006 which is not as good as model R1. What’s out of expectation is that none of the education variables are significant. We think maybe education combined with other variables might be significant.

profit ~ educ + ez + loc2 + market + transport

Hypothesis:

educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = ezForest = ezSavannah = loc2Rural = marketTRUE = transportTRUE = 0

Model 6: restricted model with education * age (R4)

We decided to combine education with age. Our idea is that age is a representation of experience, and people with both education and experience might have some influence on profit. Our model is as below. However, the model summary doesn't show as we expected. The education together with age is still not significant. This means based on this data set, education is not an influencing factor on Ghana agricultural profit. Although this does not necessarily mean that education does not have effect at all, only that further analysis maybe required.

*profit ~ educ * age + female + ez + loc2 + market + transport*

Hypothesis:

educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = age = femaleTRUE = ezForest = ezSavannah = loc2Rural = marketTRUE=transportTRUE=educBasicEducation:age=educSecondaryEducation:age = educTertiaryEducation:age = educOther:age = 0

4. Results and discussion

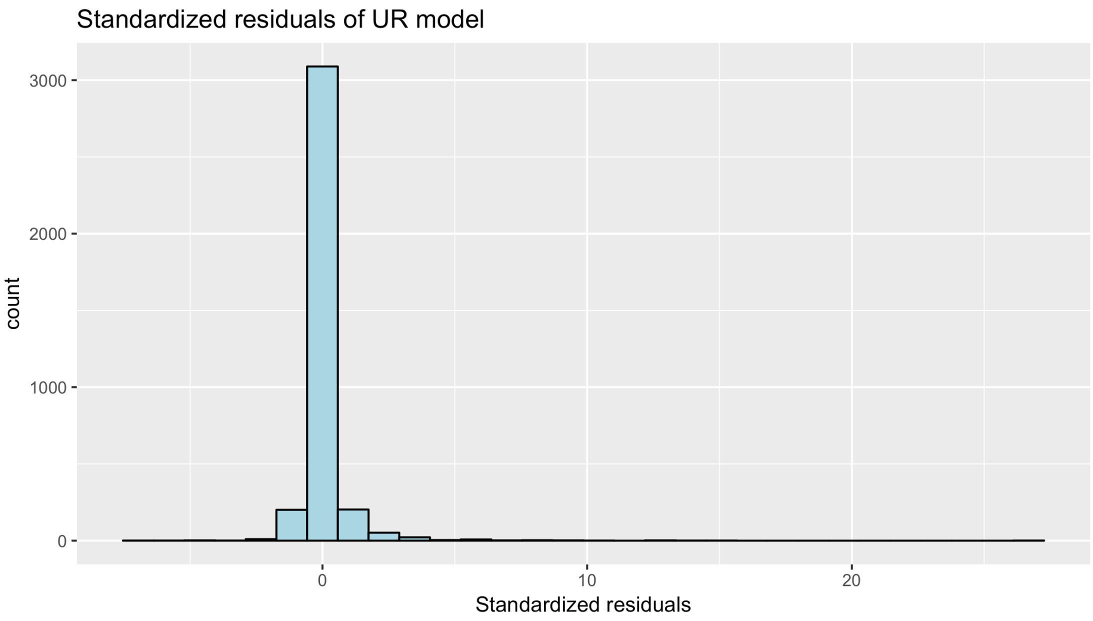
4.1 Descriptive Statistics

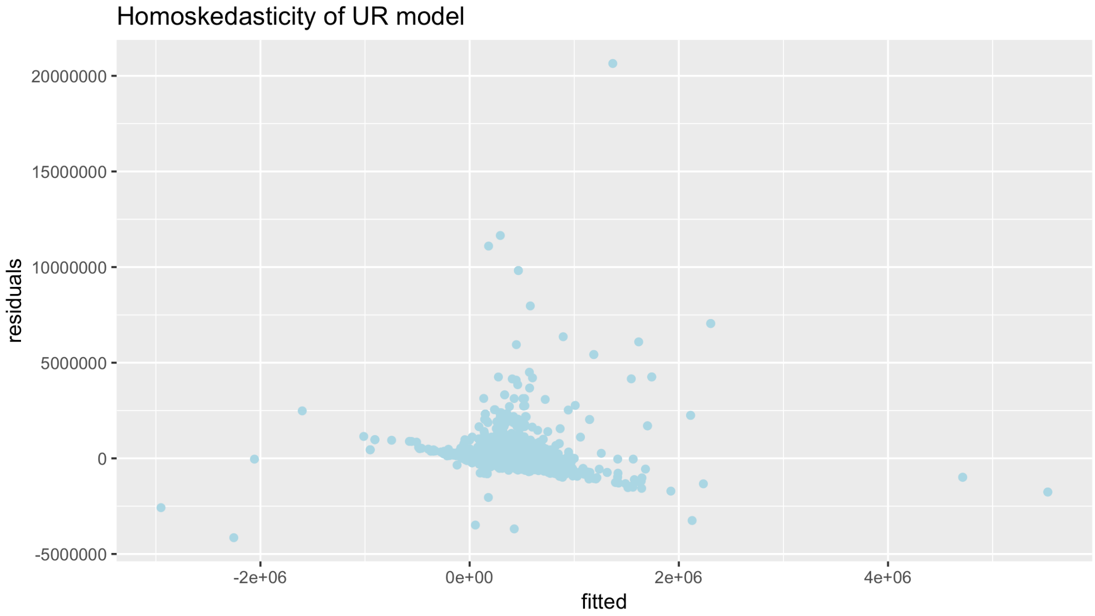
In general, the average household agriculture profit in Ghana is about 314881.5 cedi per acre, when disaggregated, the averages are 314,737.2 and 324,043 per acre for rural and urban households respectively. Also, about 84% of household agriculture profit per acre is generated by household living in rural area. These values have very high standard deviation indicating the presence of outliers.

From our unrestricted model i.e. hh_profit_model_ur, we can see that livstcd5, eqcdown61, rootcd18, rootcd20, rootcd27, rootce36 are statistically significant, which means households that own pigs as livestock, own outboard motor as agriculture equipment, or grow root crops such as cassava, cocoyam, eggplant and pawpaw can affect agriculture profit significantly. It should be noted that eqcdown61 and rootcde27 both have negative effect on our training target, which means household that own outboard motor or grow eggplant as root crops can generate decrease in household profit per acre. When disaggregated, the unrestricted model for rural area shows that same factors as in unrestricted model that affect rural household agriculture profit significantly, however, in unrestricted model for urban area, there are no variables showing strong statistically significance.

4.1 Regression diagnostics:

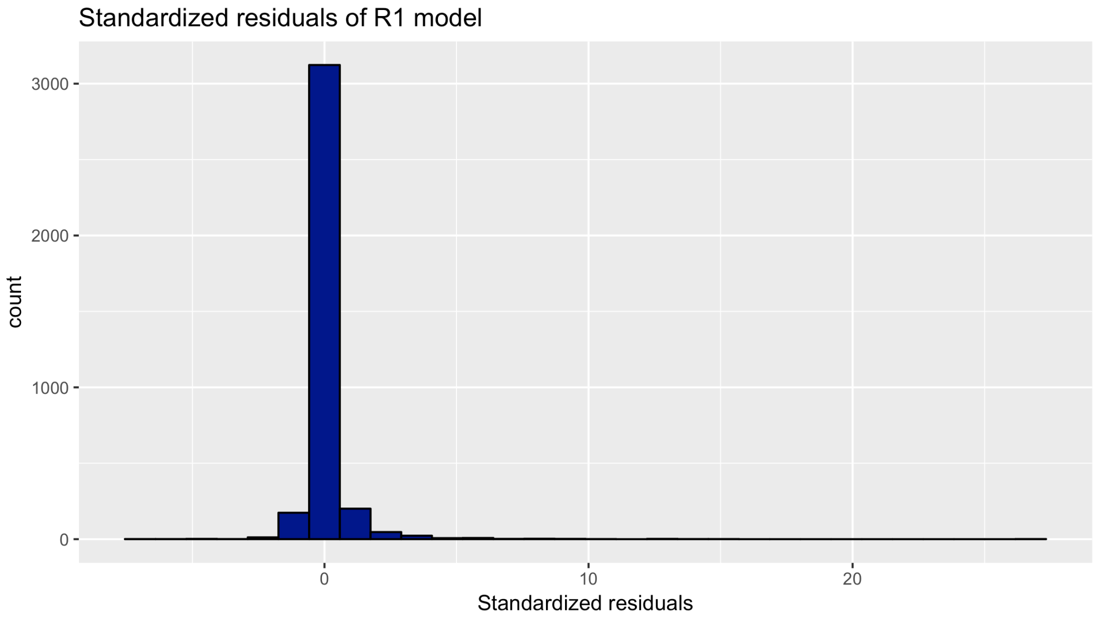
Model 1 UR:

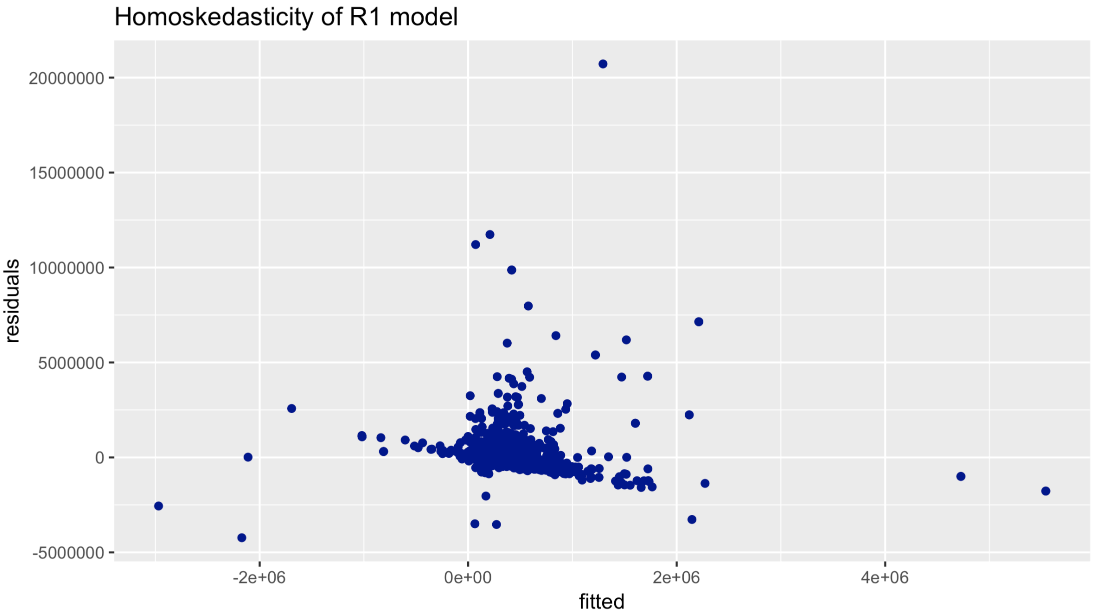




The histogram plot illustrates an approximately normal distribution of standardized residuals produced by Model 1, which is symmetric. In the residuals vs fitted plot, besides the outliers, the majority shows a decreasing trend in residuals as fitted value goes larger, and the variance becomes smaller. This could be because there are other important variables we didn't include in our model.

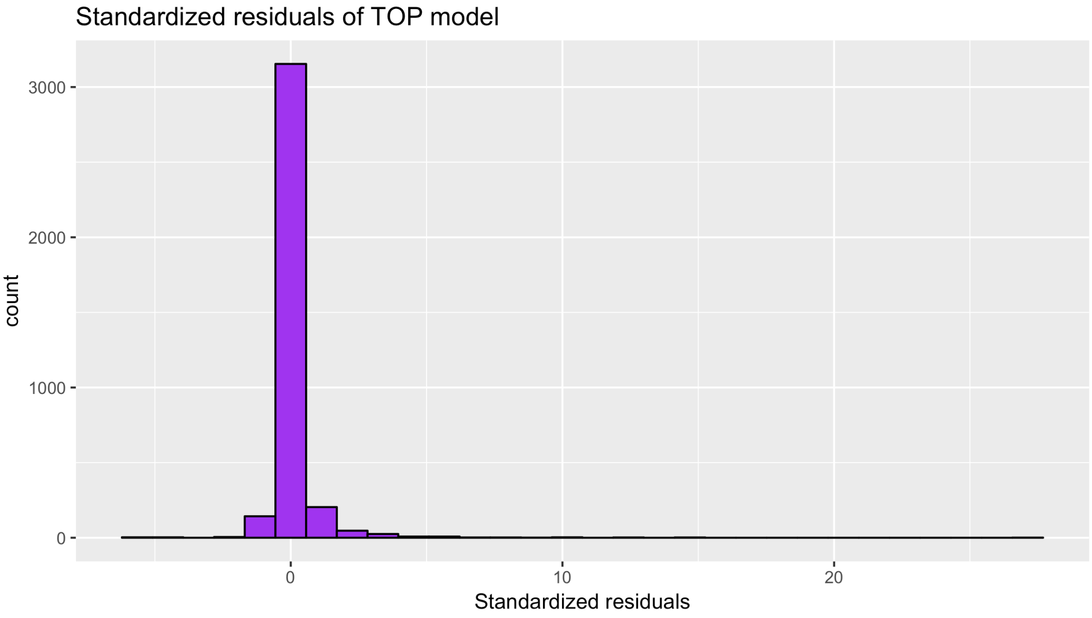
Model 2 R1:

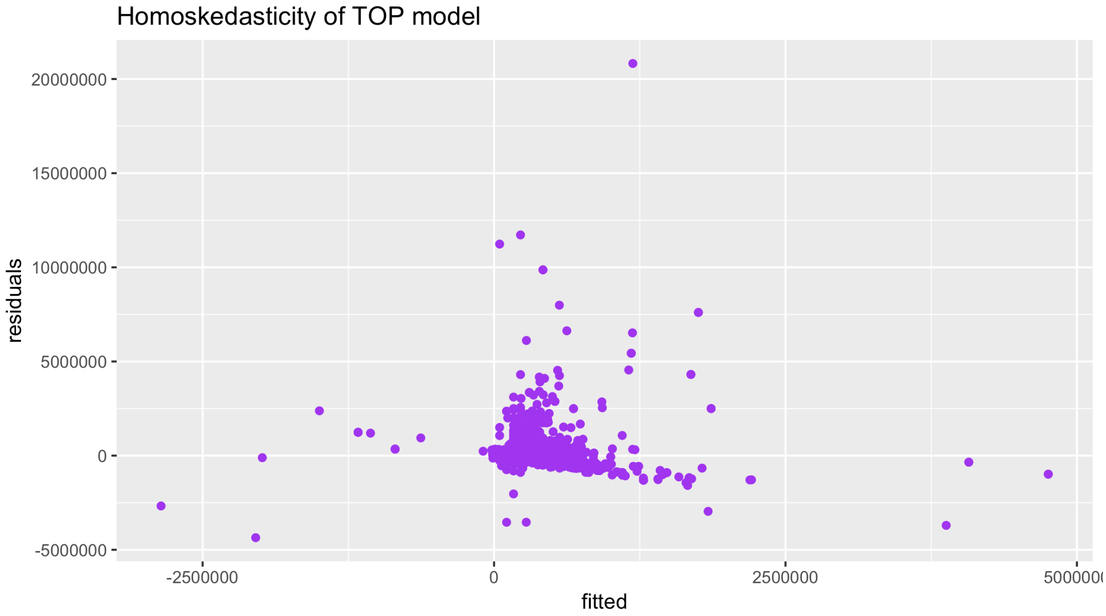




Similar to Model 1, the standardized residuals histogram is normal, and residuals vs fitted shows a decreasing trend and variance is not constant.

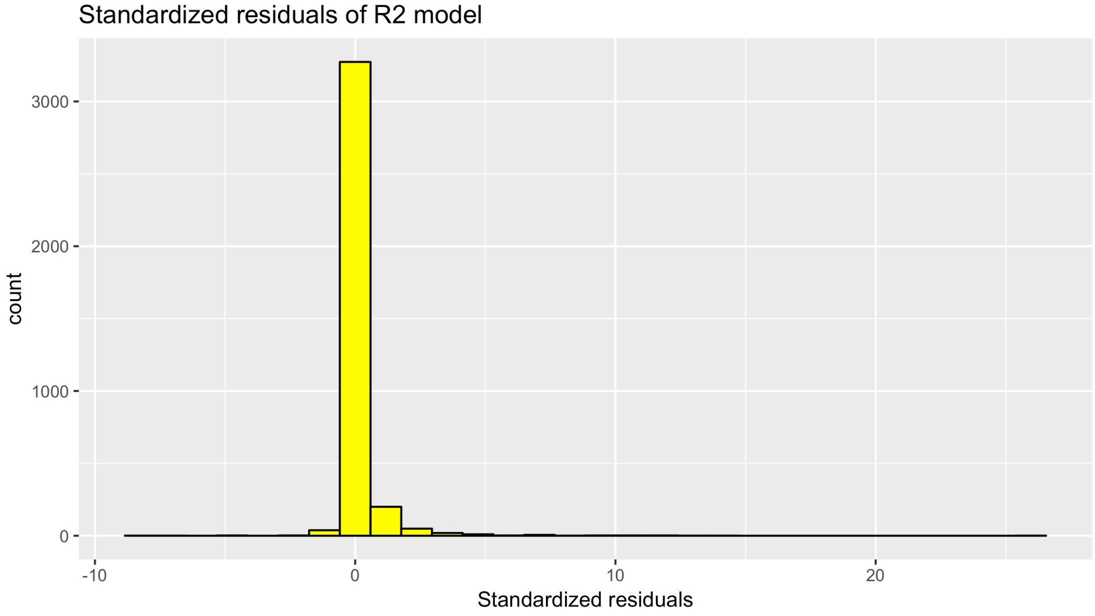
Model 3 TOP:

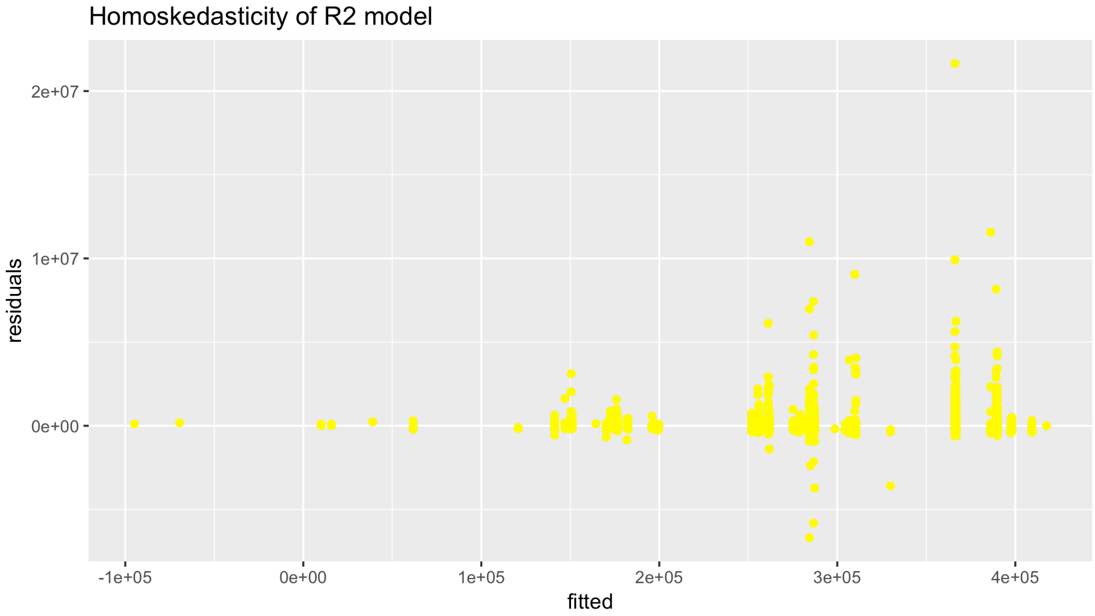




Model 3 has the similar distribution as Model 1, which is also fit but can be improved as well.

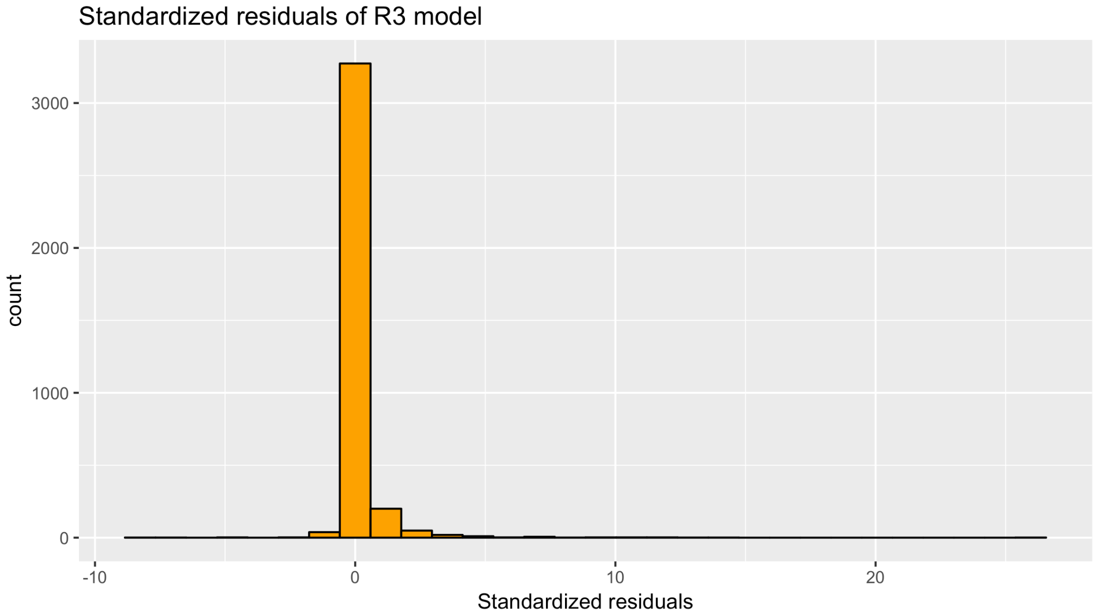
Model 4 R2:

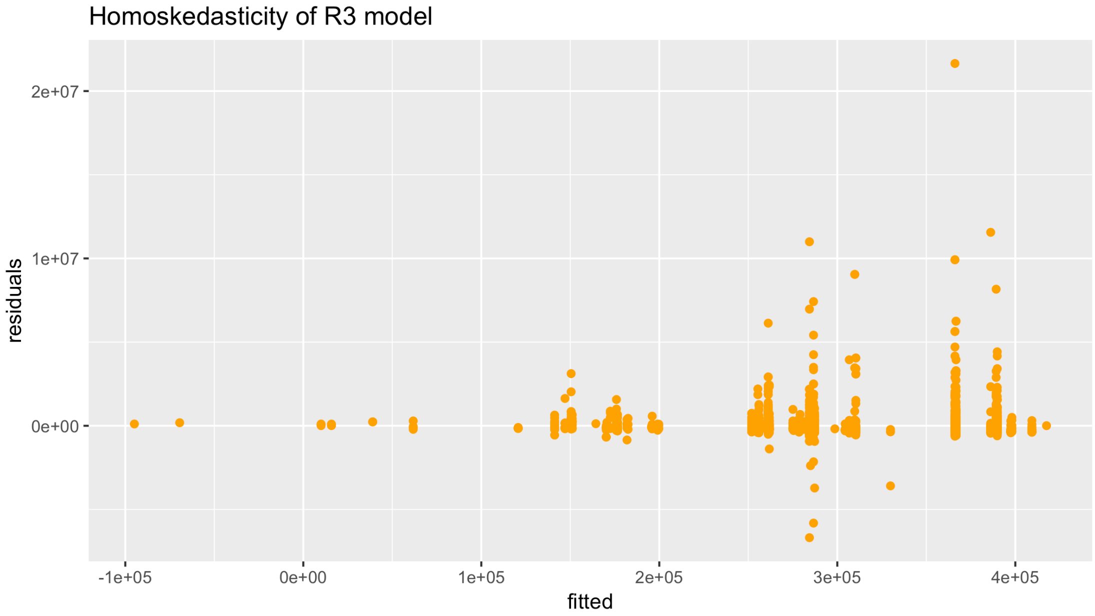




The plot shows a y-axis unbalanced residual , which means this model can be made significantly more accurate.

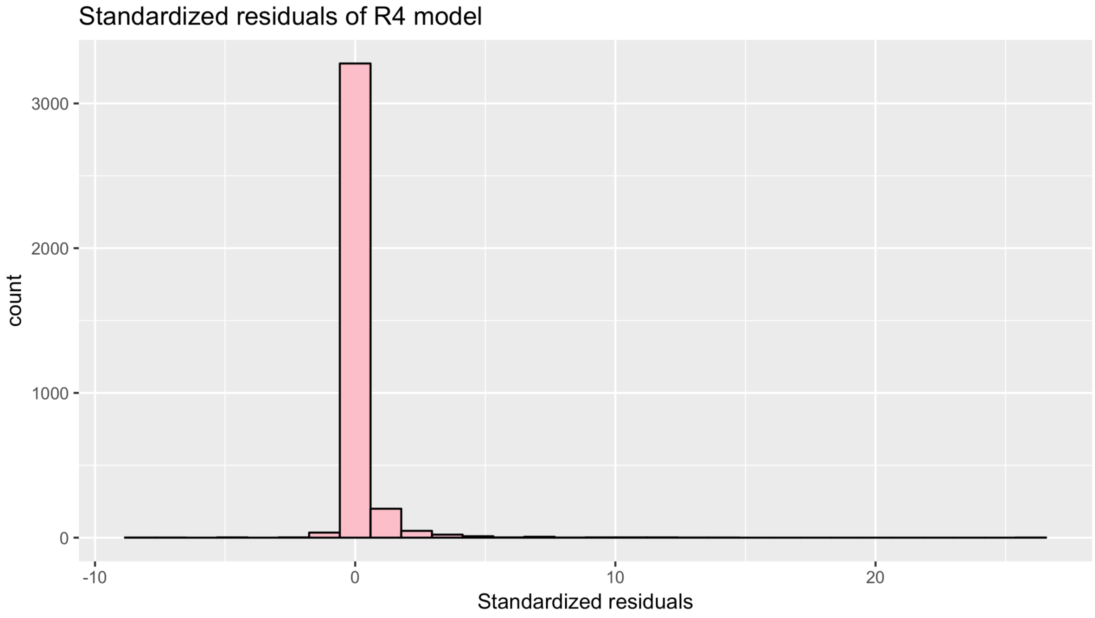
Model 5 R3:

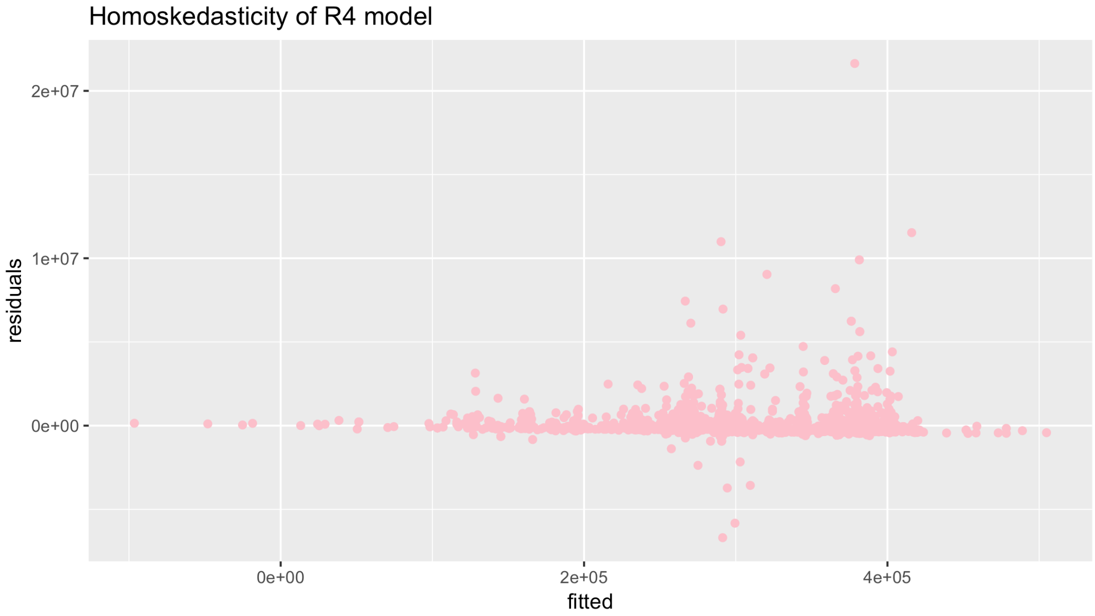




Same problem as Model 4, hence, this model can be made significantly more accurate.

Model 6 R4:





Model 6 plots show most of the predictions are a bit too high, and then some would be way too low and aren't evenly distributed vertically. This almost always means the model can be made significantly more accurate. Most of the time we'll find that the model was directionally correct but pretty inaccurate relative to an improved version.

5. Conclusions:

Based on the analysis above, we conclude that the restricted model R1 is the best fit among all models. It could roughly explain 12% variation in profit per acre. Education does not have much effect on profit, and ecological zone, market and crops have significant influence on profit.

Appendix

Find correlations between features and training target (HH profit per acre)

For features already grouped by (clust, nh)

To calculate the correlation between feature and training target, i.e. HH profit per acre, firstly we added features from aggregates data in batch mode. For features already grouped by (clust, nh), we added feature directly using the function below:

```
AddFeature <- function(data, feature){
  all_features <- data %>%
    left_join(feature, by=c("clust", "nh"))    %>%
    replace(., is.na(.), 0)
  return (all_features)
}
```

To add this kind of feature more efficiently, we developed a function so that we could add features from files in batch mode.

```
AddAggFeaturesFromFiles <- function(data, path, namePattern) {
  files <- list.files(path, full.names = TRUE, pattern = namePattern)
  for (i in 1:length(files)) {
    file <- files[i]
    data <- AddFeature(data, read_dta(file))
  }
  return (data)
}
```

For features not grouped by (clust, nh)

In this situation, we aggregated features by (clust, nh), generate (mean, sum) of one feature, using the function shown below:

```
GetAggFeature <- function(data, dataCol){
  aggData <- data %>%
     select(c("clust", "nh", dataCol)) %>%
     group_by(clust, nh) %>%
     summarise_at(c(3), funs(mean, sum))

  names(aggData)[3] <- paste(dataCol, colnames(aggData)[3], sep="_")
  names(aggData)[4] <- paste(dataCol, colnames(aggData)[4], sep="_")
```

```
    return (aggData)
}
```

After adding training target to all features list, 1 training target and 157 features were generated.

Calculate the correlations

In this part, we calculated the correlations in batch mode, and generated a new data frame *all_correlations* with three columns. Column names are *index, colName*, and *correlation,* ordered by correlation in descending order. Below is the function we used.

```
findCorrelation <- function(a) {
    df <- data.frame(index = (NA), colName=(NA), correlation = (NA))
    for (i in 1:ncol(a)) {
        correlationP <- cor(a[i], a[1])
        row <- c(i, colnames(a[i]),correlationP)
        df<- rbind(df, row)
    }
    df <- df %>%
        filter(!is.na(colName))
    df <- df[order(df$correlation, decreasing = T),]
    return (df)
}
```

After applying above function to calculate the correlations, we selected top n features, and passed them to *lm()* parameter. Through the comparison, it is not hard to find that these features overlap the features that we used to calculate our training target. Thus, we later dropped these features.

Same method has been applied to raw data. We found that features such as livstcd5, livstcd6, cropcd8, cropcd11, rootcd8, rootcd18, etc. are statistic significant compared to other features, thus we trained model based on these features.