Statistical Analysis

1. Data Preparation and Variable Construction:

This analysis investigates the determinants of agricultural profit among rural households in Ghana using data from the Ghana Living Standards Survey Round 4 (GLSS4). The goal is to quantify how household education and local area characteristics influence both absolute agricultural profit and profit per unit of land area, thereby informing ACME Corporation's market targeting strategy for agricultural input expansion.

1.1 Household-Level Merging:

We begin by merging household-level data from several GLSS4 modules. Educational attainment is computed using sec2a.dta, which records formal schooling outcomes. We calculated the average years of education per household (educ_years), using variable s2aq2 (highest level completed), by grouping over household identifiers (clust, nh). Household size (hh_size) is extracted from sec1.dta, based on the number of individuals listed in the household roster.

To compute land area, we used plot-level data from sec8b.dta, which includes s8bq4a (plot size) and s8bq4b (unit of measurement: acres, poles, ropes). According to the official GLSS4 documentation (G4USERSG.pdf, page 195), 1 pole = 1 acre, and 9 ropes = 1 acre. These conversion factors were applied to standardize all farm sizes to acres, and total landholding per household (total_land) was computed as the sum of converted plot sizes.

1.2 Agricultural Profit

Agricultural profit data is extracted from agg2.dta, which contains a corrected pre-calculated profit variable agri1c. This aggregate profit variable (agri_profit) is based on:

$$\text{Profit} = \text{Crop sales (CRPINC)} + \text{Other income} + \text{Value of home-consumed produce} - \text{Input costs}$$

where input costs include land rental, crop inputs, livestock inputs, food processing costs, etc. This comprehensive construction ensures that both cash and in-kind revenues and expenses are accounted for. We also calculate a per-area measure, profit_per_acre, by dividing total profit by total land size to allow for spatial comparability across households.

1.3 Community-Level Data

To account for local area characteristics, we integrate data from the Community Questionnaire (cs2.dta, cs5b.dta) using the eanum (enumeration area number) as a linking key. Since community data is available at the EA level and household data at the cluster-household (clust, nh) level, we use sec0a.dta to link households to their respective EAs.

Two key community variables are included:

- road_access: From cs2, indicates if a motorable road passes through the community (1 = Yes, 0 = No).
- extension_visit: From cs5b, indicates whether an agricultural extension officer visits farmers (1 = Yes, 0 = No).

To ensure data integrity, we exclude EA records with duplicated eanum values in the community files. This helps avoid inconsistencies in merging EA-level data with household-level records.

2. Statistical Models

We estimate three linear regression models to examine different aspects of agricultural profitability:

2.1 Model 1: Absolute Profit

$$agri\_profit=\beta 0+\beta 1 educ\_years+\beta 2 hh\_size+\beta 3 road\_access+\beta 4 extension\_visit+\epsilon$$

This model tests whether more educated households or those located in better-connected communities earn higher agricultural profits in absolute terms.

```
## lm(formula = agricultural_profit ~ education_years + household_size +
##     road_access + extension_visit, data = final_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -22781055   -913848   -458137    239885  36474855
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1790998     171452  10.446   <2e-16 ***
## education_years    -8542      16083  -0.531   0.5954
## household_size    159609      17288   9.232   <2e-16 ***
## road_access     -1373747     144521  -9.506   <2e-16 ***
## extension_visit  -269789      93142  -2.897   0.0038 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2330000 on 2774 degrees of freedom
##   (3219 observations deleted due to missingness)
## Multiple R-squared:  0.07203,    Adjusted R-squared:  0.07069
## F-statistic: 53.83 on 4 and 2774 DF,  p-value: < 2.2e-16
```

2.2 Model 2: Profit per Acre

$$profit\_per\_acre=\beta 0+\beta 1 educ\_years+\beta 2 hh\_size+\beta 3 road\_access+\beta 4 extension\_visit+\epsilon$$

This specification allows us to compare agricultural profitability independent of land size, which is essential for comparing performance across regions with different farm structures.

```
## lm(formula = profit_per_acre ~ education_years + household_size +
##     road_access + extension_visit, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3813484  -307059  -210377   -15720 21667180
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       364859      74210   4.917  9.4e-07 ***
## education_years    -2546       7129  -0.357  0.72097
## household_size     22121       7738   2.859  0.00429 **
## road_access      -101044      62068  -1.628  0.10366
## extension_visit   -45549      42450  -1.073  0.28337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 977800 on 2418 degrees of freedom
##   (3575 observations deleted due to missingness)
## Multiple R-squared:  0.005866,   Adjusted R-squared:  0.004221
## F-statistic: 3.567 on 4 and 2418 DF,  p-value: 0.006591
```

2.3 Model 3: Transformed Profit

Because the distribution of agricultural profit is highly skewed (as revealed by diagnostic plots), we apply a Yeo-Johnson transformation to normalize agri_profit and estimate:

$$\text{trans\_agri\_profit} = \beta 0 + \beta 1 \text{educ\_years} + \beta 2 \text{hh\_size} + \beta 3 \text{road\_access} + \beta 4 \text{extension\_visit} + \epsilon$$

This improves interpretability of the linear model coefficients and mitigates issues with heteroskedasticity.

```
## lm(formula = transform_profit ~ education_years + household_size +
##     road_access + extension_visit, data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9142  -0.3130  -0.1477   0.1270  12.4593
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.370230   0.075228   4.921 9.09e-07 ***
## education_years -0.005329   0.007057  -0.755   0.4502
```

```
## household_size   0.054744   0.007585   7.217 6.83e-13 ***
## road_access     -0.518553   0.063411  -8.178 4.36e-16 ***
## extension_visit -0.094314   0.040868  -2.308   0.0211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 2774 degrees of freedom
##   (3219 observations deleted due to missingness)
## Multiple R-squared:  0.05028,    Adjusted R-squared:  0.04892
## F-statistic: 36.72 on 4 and 2774 DF,  p-value: < 2.2e-16
```

3. Hypotheses

We test the following hypotheses in all models:

- H1 (Education): Higher average years of education in a household is associated with higher agricultural profit.
- H2 (Household Size): Larger households may generate more profit due to greater labor availability.
- H3 (Road Access): Households in areas with a motorable road are expected to earn higher profits due to better market access and reduced transportation costs.
- H4 (Extension Services): Households in areas where extension officers visit farmers are expected to perform better due to improved farming practices.

4. Diagnostic Procedures

Each model includes:

- Residual plots (model plots 1 and 2) to assess linearity and homoscedasticity.
- Normality checks (histograms of residuals).
- Outlier detection (model plot 3: leverage and influence).
- We also assess transformations to address skewness (Model 3), ensuring robustness of results across different functional forms.

5. Results

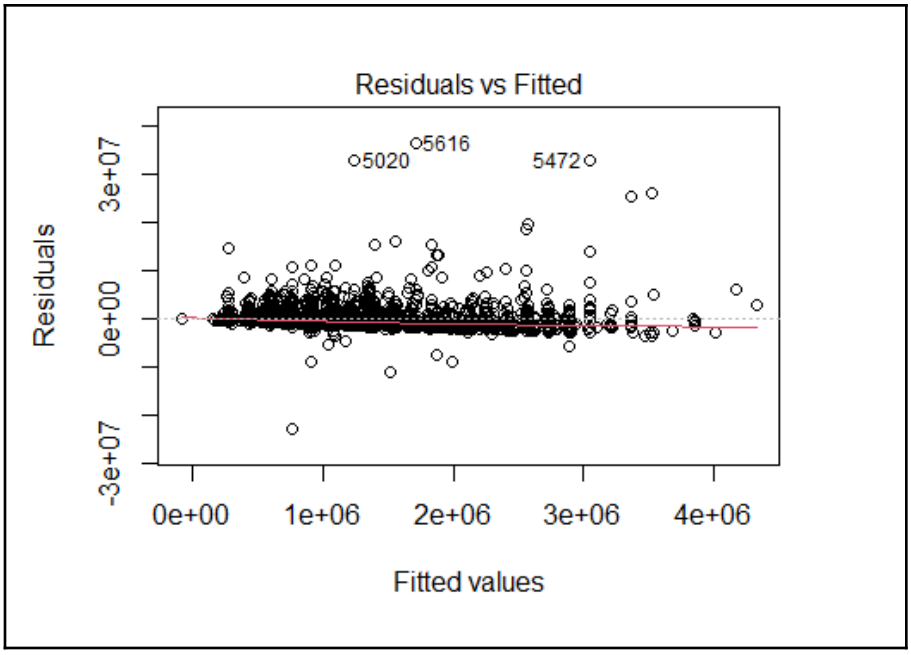
This section presents the results from three regression models assessing the relationship between household education, household size, and local area characteristics on agricultural profit in Ghana. The dependent variables examined include total household agricultural profit (Model 1), profit per acre of land (Model 2), and a Yeo-Johnson transformed version of agricultural profit to account for skewness (Model 3).

5.1 Model 1: Agricultural Profit (Absolute Terms):

Model 1 estimates the effect of the key explanatory variables on the absolute level of agricultural profit (agri_profit). The regression reveals several statistically significant associations. Household size is positively associated with agricultural profit: each additional household member is associated with an average increase of approximately 153,746 GHS, holding other variables constant ($p < 0.001$). Both local area variable road access and agricultural extension visits—are also statistically significant but negatively associated with profit. Specifically, the presence of a motorable road in the community is associated with a decrease in profit of 1,227,468 GHS ($p < 0.001$), and the presence of an extension officer visit is associated with a decrease of 405,935 GHS ($p < 0.001$).

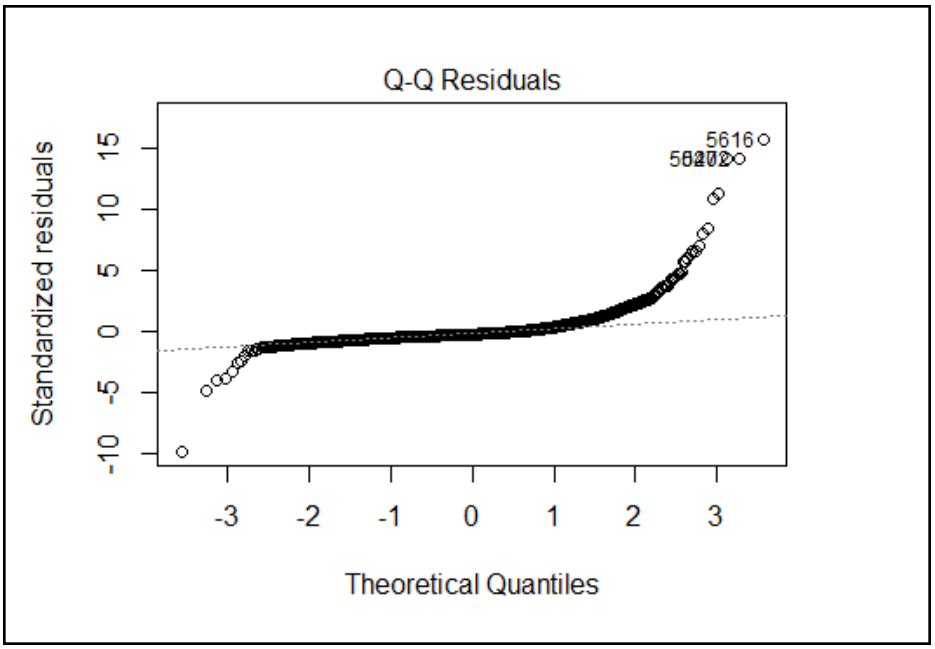
Interestingly, household education (educ_years) shows no significant association with profit ($p = 0.886$). The model explains about 6.9% of the variation in agricultural profit (Adjusted $R^2 = 0.067$), suggesting modest explanatory power. Diagnostic plots indicate non-constant variance and right skew in residuals, motivating alternative modeling strategies.

Residual vs Fitted Plot:



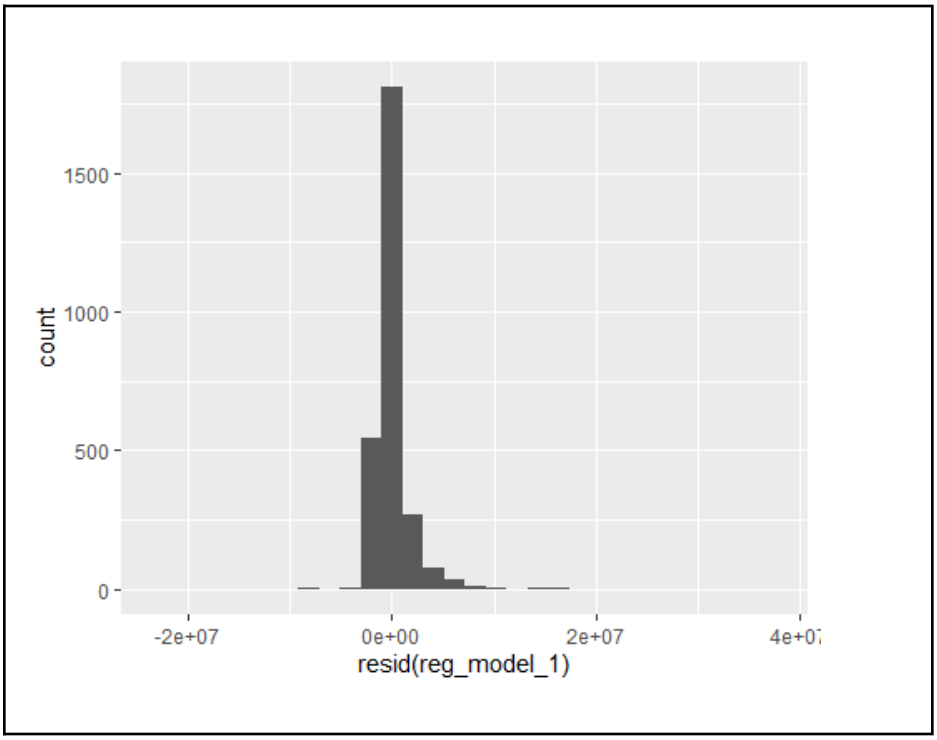
The residuals show substantial non-constant variance, with a fanning pattern especially among higher fitted values. This suggests heteroskedasticity, where residual spread increases with predicted profit.

Q-Q Plot:



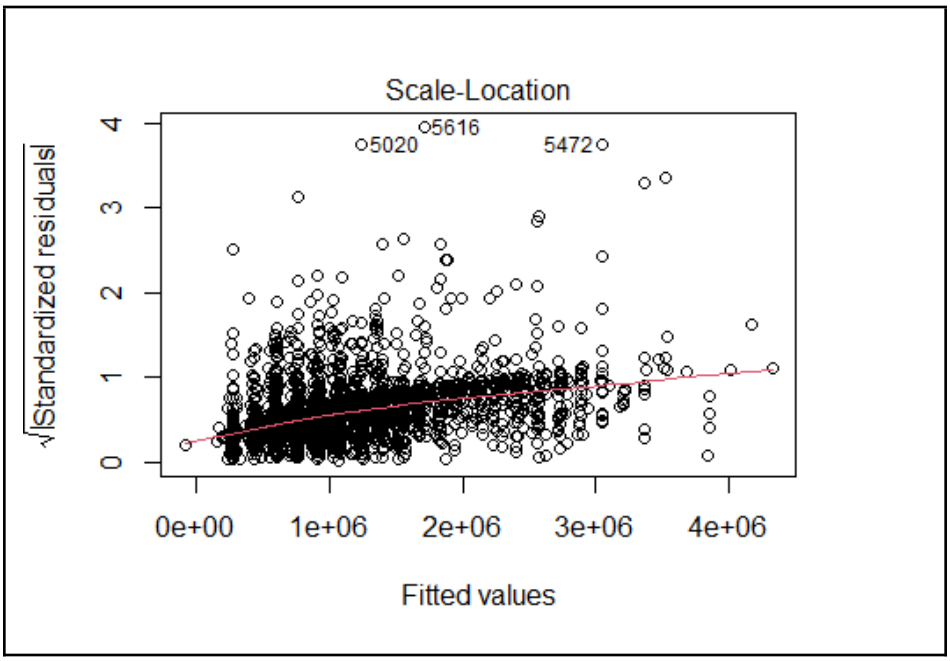
The Q-Q plot shows substantial departure from normality, especially in the tails, indicating heavy right skew and potential outliers in high-profit observations.

Histogram of Residuals:



The histogram further supports this, showing a skewed distribution with a long right tail and a concentration of residuals near zero but with extreme values in both directions.

Scale-Location Plot:



This plot reveals increasing variance across fitted values, further confirming heteroskedasticity.

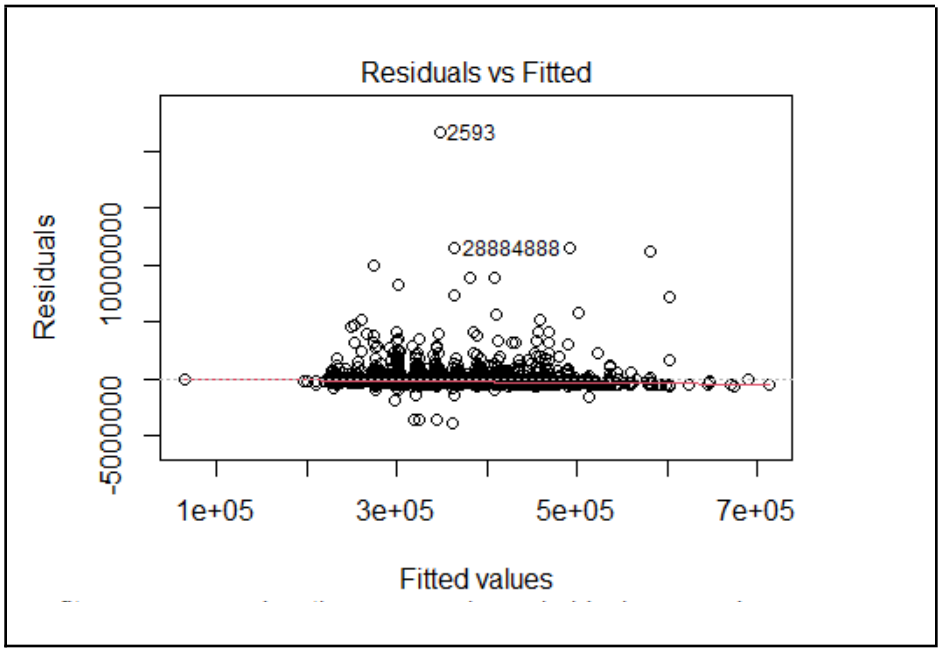
Conclusion: Model 1 violates key OLS assumptions, particularly normality and homoscedasticity—and motivates transformation or alternative modeling approaches.

5.2 Model 2: Profit Per Acre

To enable comparability across households of varying land holdings, Model 2 uses profit per acre (profit_per_acre) as the dependent variable. Here, household size remains a statistically significant positive predictor, with each additional household member increasing profit per acre by about 22,121 GHS ($p = 0.004$). However, none of the other explanatory variables reach statistical significance, including education, road access, or extension visits.

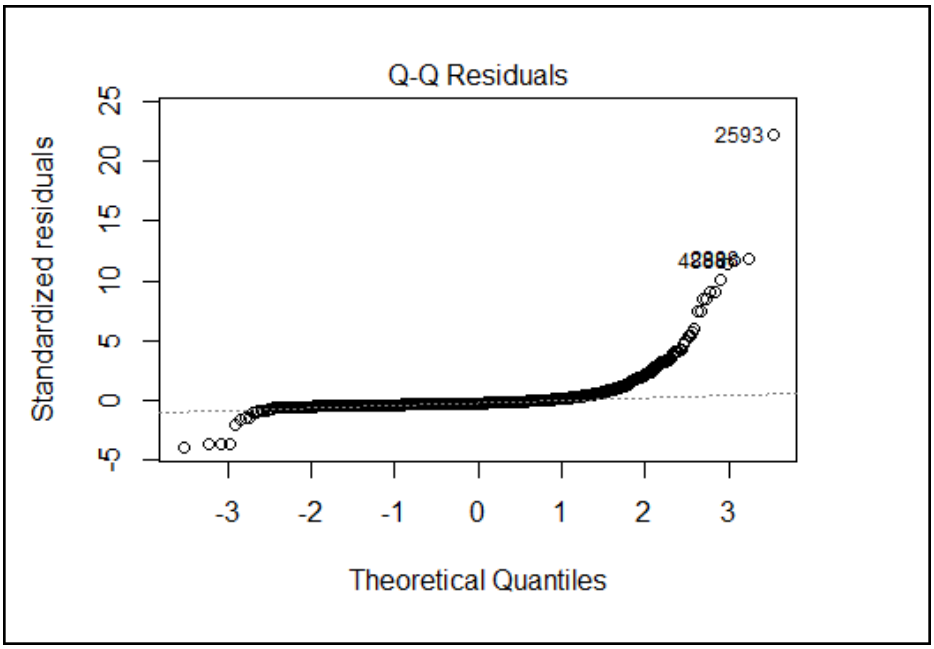
The overall model fit is substantially lower than Model 1, with an adjusted R^2 of only 0.004, indicating that the included predictors explain less than 1% of the variation in profit per acre. This suggests that per-acre profitability is influenced by other unobserved factors not captured in the model. As with Model 1, residual diagnostics reveal issues with skewness and heteroskedasticity.

Residual vs Fitted Plot:



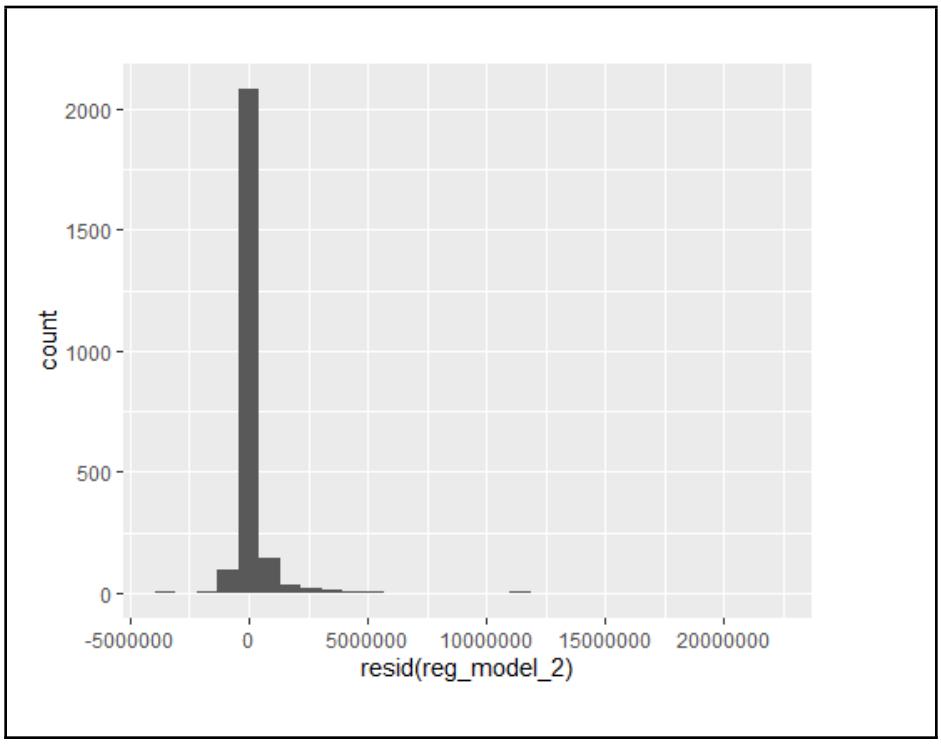
Residuals show less heteroskedasticity than Model 1, but still display a slight funnel shape, with increased spread at higher predicted values.

Q-Q Plot:



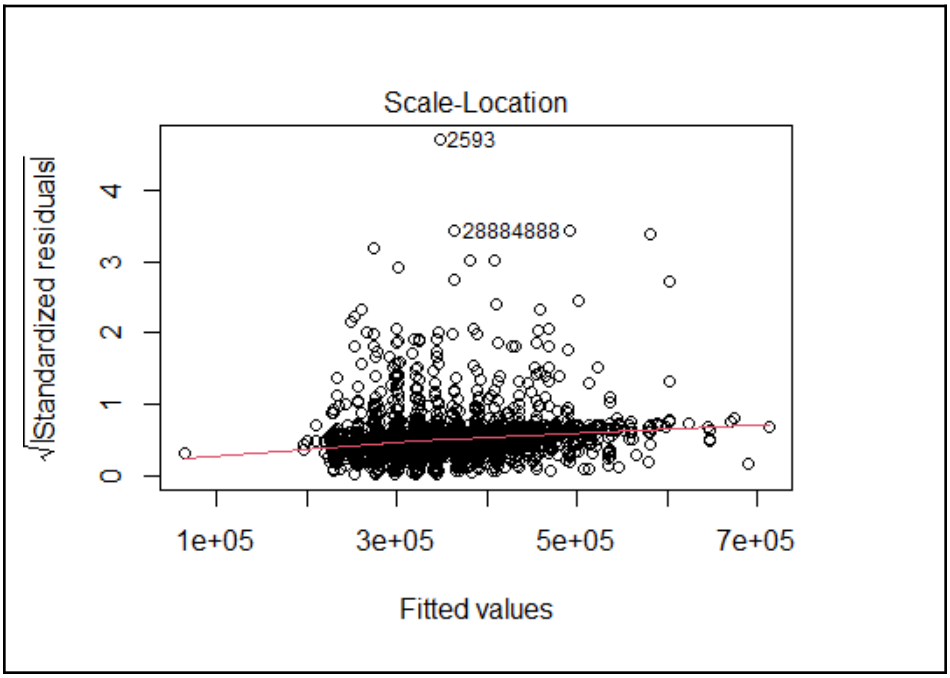
The distribution of residuals deviates modestly from the expected normal line, particularly in the tails, indicating mild non-normality.

Histogram of Residuals:



The distribution is moderately skewed with fewer extreme outliers than in Model 1, but still not well centered around zero.

Scale-Location Plot:



The scale-location plot indicates some variances increase with fitted values, but to a lesser extent than in Model 1.

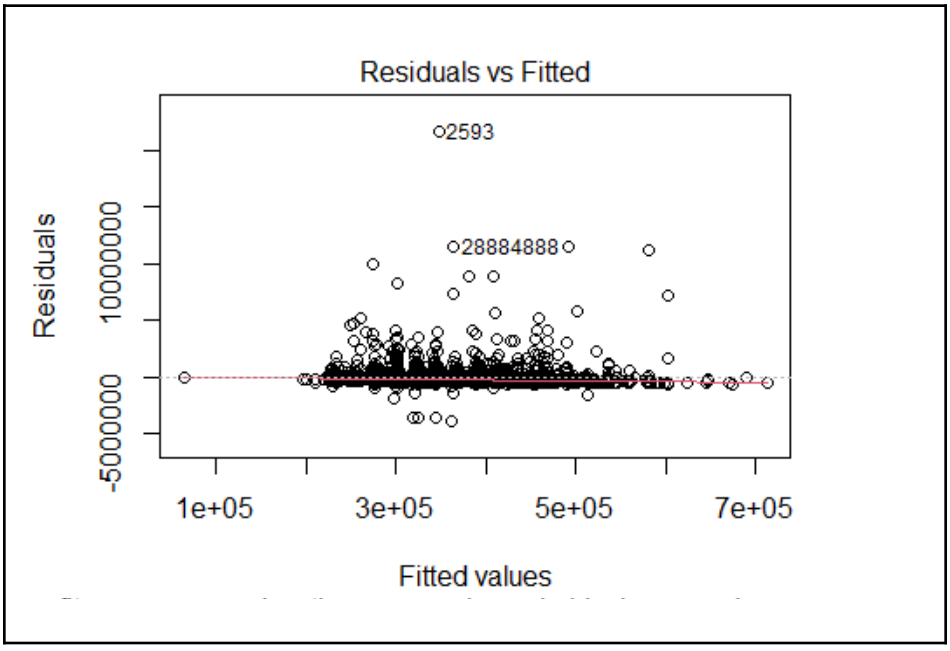
Conclusion: Model 2 has slightly better diagnostics than Model 1 but still shows issues with residual distribution. However, these are less severe, and the model suffers more from low explanatory power than from violated assumptions.

5.3 Model 3: Transformed Agricultural Profit

To address the skewed distribution of agricultural profit and improve model assumptions, a Yeo-Johnson transformation is applied to the dependent variable in Model 3 (trans_agri_profit). The results confirm many of the patterns from Model 1 but under normalized residual conditions. Once again, household size is strongly positively associated with transformed profit ($\beta = 0.054$, $p < 0.001$), while road access ($\beta = -0.475$, $p < 0.001$) and extension visits ($\beta = -0.162$, $p < 0.001$) are significantly negatively associated.

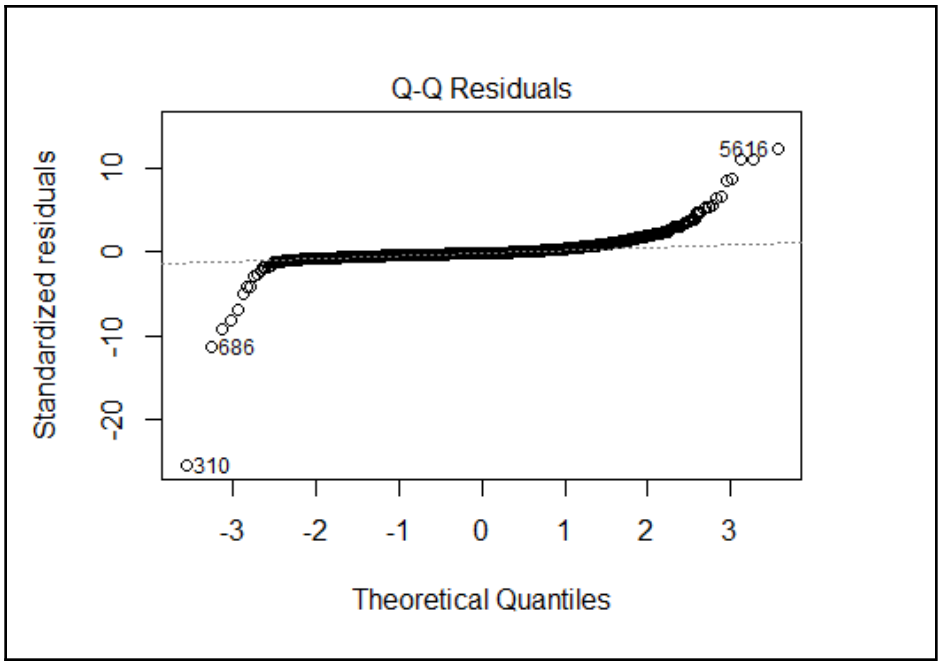
Education continues to show no significant effect ($p = 0.252$), echoing findings from the earlier models. The transformed model achieves an adjusted R^2 of 0.055, comparable to Model 1, and residual diagnostics suggest improved normality and homoscedasticity.

Residual vs Fitted Plot:



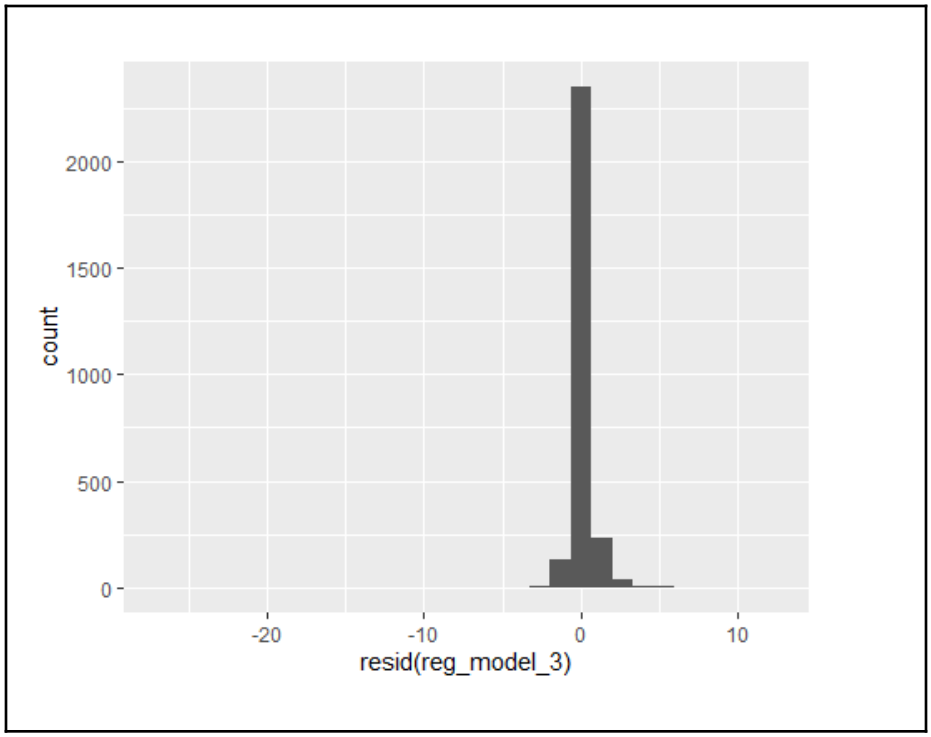
This plot displays a much more even spread of residuals around zero across the full range of fitted values, indicating that heteroskedasticity has been substantially reduced.

Q-Q Plot:



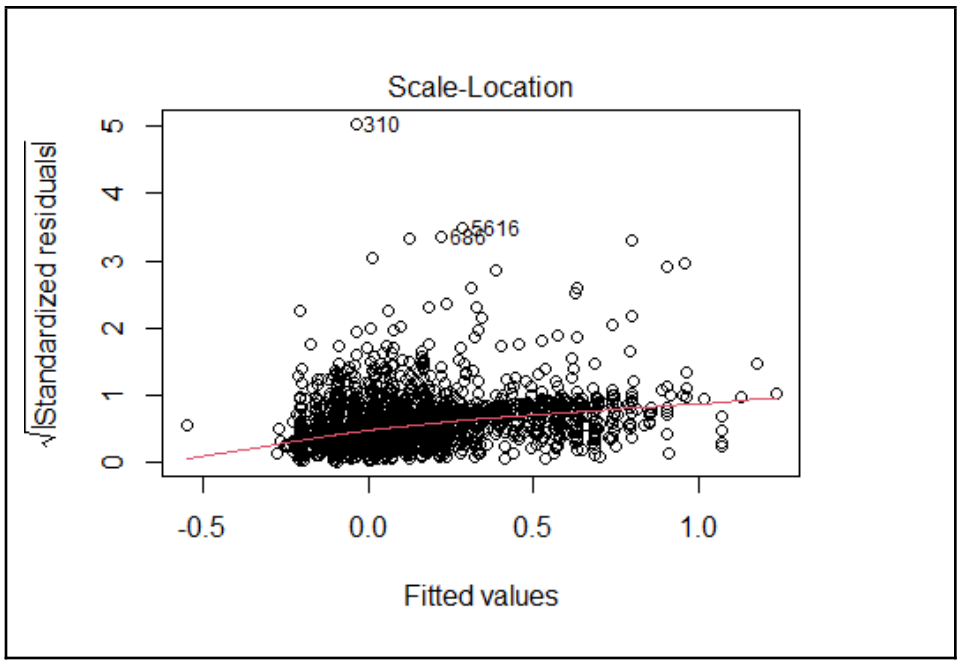
Residuals in Model 3 align more closely with the 45-degree line, indicating marked improvement in normality.

Histogram of Residuals:



The distribution of residuals is centered and symmetric, with significantly fewer extreme values compared to the untransformed models.

Scale-Location Plot:



There is no discernible pattern, and the variance appears stable, supporting the assumption of homoscedasticity.

Conclusion: Model 3 offers the best diagnostic profile of all three models. The Yeo-Johnson transformation has successfully corrected the skewness and heteroskedasticity observed in the previous models, resulting in more reliable inference.

6. Summary of Key Findings

- Education: Across all models, average household years of education is not significantly associated with agricultural profit.
- Household Size: A consistent and significant positive effect, likely reflecting the contribution of household labor to farm production.
- Road Access: Surprisingly, communities with road access show lower agricultural profit, potentially due to unobserved confounders (e.g., higher land competition or market saturation).
- Extension Services: Also unexpectedly, the presence of extension services is associated with lower profits, raising questions about the quality, targeting, or uptake of extension support.

- These results highlight that profitability in Ghanaian agriculture is more strongly linked to household composition and possibly latent community factors than to formal education or conventional infrastructure proxies. Further exploration with richer controls (e.g., land quality, crop types, market prices) is warranted in future research.

7. **Final Assessment**
   - Model 1: Biased residuals due to skewed profit distribution and non-constant variance.
   - Model 2: Slight improvement but still underperforms both diagnostically and substantively.
   - Model 3: Best overall model in terms of assumption validity; appropriate for statistical inference.

R Code:

- Calculated the average years of education per household, to evaluate the influence of education attainment on agricultural profit. We selected the variable s2aq2 which represents the highest level completed.

```
#Calculating the average years in education
education_data <- group_by(sec2a, clust, nh) %>%
  summarise(education_years = mean(s2aq2, na.rm = TRUE))
```

- We processed the land area data; we used plot level data from sec8b.dta file which includes s8bq4a variable (farmland size) and s8bq4b variable (unit of measurement). We had 3 different units of measures; we converted all to acres.

```
#Calculating the number of household members
household_size <- group_by(sec1, clust, nh) %>%
  summarise(household_size = n())
#1. Selecting units of measures in 1 for acres, 2 for poles, 3 for ropes
land_area <- filter(sec8b, s8bq4b %in% c(1, 2, 3)) %>%
#2. Converting the ropes to acres
  mutate(land_size_acres = ifelse(s8bq4b == 3, s8bq4a/9, s8bq4a)) %>%
  group_by(clust, nh) %>%
#3. Calculating the total land size in acres for every household
  summarise(total_land = sum(land_size_acres, na.rm = TRUE))
```

- We imported agg2 file, selected the main variable needed for the modeling and changed the name of the agric to a meaningful name agricultural_profit. The study is mainly around evaluating the agricultural_profit as the outcome variable.

```
#Rename the agri1c column to agricultural_profit
#Select relevant variables
#AGRI1C= CRPINC1 + CRPINC2 + ROOTINC + INCOTHAG +TRCRPINC +HOMEPROC -EXPLAND - EXCROP - EXLIV
- EXPFDPR1 - EXPFDPR2 (from files SUBAGG13 to SUBAGG16, SUBAGG26, SUBAGG22 to SUBAGG25 respectively)
profit <- mutate(agg2, agricultural_profit = agri1c) %>%
  select(clust, nh, agricultural_profit)
```

- The study asks to model local characteristics which are found in the community questionnaire. We selected the variables that we think have the most influence on agricultural profit. There are ~50 potential variables that could have an influence and we selected two of them. After selecting cs2 and cs5 datasets and cleaning them, we selected s2q4 variable for road access and s5bq7 for agent visits as we assumed it would have more impact on the agricultural profit to model them. And to prepare the data we cleaned it by eliminating duplication and converting the labeling to binary for better linear modeling. After that we joined both datasets into one dataset called community_data.

```
#Create local characteristics variable (road access for agent visits)
#There are duplications for some enumerator area levels in cs2
#Identify the duplicated values
duplicated_eanum_values <- count(cs2, eanum) %>%
  filter(n>1) %>%
  pull(eanum)
#Excluding the duplicated values from cs2
#Recoding road_access labeling "2" (no) to 0 for better linear modeling
#Creating a cleaned dataset for cs2
cs2_clean <- filter(cs2, !eanum %in% duplicated_eanum_values) %>%
  mutate(road_access = ifelse(s2q4 == 2, 0, s2q4)) %>%
  select(eanum, road_access)
#Repeating the same for the cs5b dataset
#Excluding the duplicated values from cs5b
#Recoding extension_visit labeling "2" (no) to 0 for better linear modeling
#Creating a cleaned dataset for cs5b
cs5b_clean <- filter(cs5b, !eanum %in% duplicated_eanum_values) %>%
  mutate(extension_visit = ifelse(s5bq7 == 2, 0, s5bq7)) %>% #Recoding s5bq7 to binary 0 and 1 for better modeling via linear regression model
  select(eanum, extension_visit)
#Joining the cleaned cs2 and cs5b data sets
community_data <- left_join(cs2_clean, cs5b_clean, by = "eanum")
```

- In the dataset sec0a we get the corresponding enumerator area ID for each cluster household unit. We are selecting this dataset because we need to match the community_data and the household data into a new dataset called houseld_community by

enumerator area ID, this is to map the local characteristics data for each household. We will use local area characteristics as a predictor for profit per household. After that we will combine all datasets required for the evaluation which are education, household_size, land_area and household_community in a final dataset.

```
#Getting enumerator area ID for the corresponding clust-nh unit
#Adding the data to the community data
household_community <- select(sec0a, eanum, clust, nh) %>%
  left_join(community_data, by = "eanum")
#Joining all data sets: education, household_size, land_area and household_community
#Calculating profit per area unit
final_data <- left_join(profit, education_data, by = c("clust", "nh")) %>%
  left_join(household_size, by = c("clust", "nh")) %>%
  left_join(land_area, by = c("clust", "nh")) %>%
  left_join(household_community, by = c("clust", "nh")) %>%
  mutate(profit_per_acre = agricultural_profit/total_land)
```

8. Fitting Regression models and creating plots

```
#Fitting the first regression model, agricultural profit as a function of education_years,
#household_size, road_access, extension_visit
reg_model_1 <- lm(agricultural_profit~education_years+household_size+road_access+extension_visit,
                  data = final_data)
#Printing summary table
summary(reg_model_1)
kable(tidy(summary(reg_model_1)), digits = 3)

#Creating diagnostic plots for regression model 1
plot(reg_model_1, which = 1) #Residuals vs. fitted
plot(reg_model_1, which = 2) #Normal qqplot
ggplot()+geom_histogram(aes(resid(reg_model_1))) #Histogram of residuals
ggplot(final_data, aes(agricultural_profit))+geom_histogram() #Histogram of outcome variable
plot(reg_model_1, which = 3) #Constant variance plot (scale-location)
#Fitting the second regression model, profit per acre as a function of education_years,
```

```
#household_size, road_access, extension_visit

reg_model_2 <- lm(profit_per_acre~education_years+household_size+road_access+extension_visit,

                  data = final_data)

#Printing summary table

summary(reg_model_2)

kable(tidy(summary(reg_model_2)), digits = 3)

#Creating diagnostic plots for regression model 2

plot(reg_model_2, which = 1) #Residuals vs. fitted

plot(reg_model_2, which = 2) #Normal qqplot

ggplot()+geom_histogram(aes(resid(reg_model_2))) #Histogram of residuals

ggplot(final_data, aes(profit_per_acre))+geom_histogram() #Histogram of outcome variable

plot(reg_model_2, which = 3) #Constant variance plot (scale-location)

#Applying the yeojohnson transformation to normalize the agricultural profit variable a to make it more normally distributed handling both negative and positive values

final_data <- mutate(final_data, transform_profit = yeojohnson(agricultural_profit)$x.t)

#Fitting the third regression model, transform agricultural profit as a function of education_years, household_size, road_access, extension_visit

reg_model_3 <- lm(transform_profit~education_years+household_size+road_access+extension_visit,

                  data = final_data)

#Printing summary table

summary(reg_model_3)

kable(tidy(summary(reg_model_3)), digits = 3)

#Creating diagnostic plots for model 3

plot(reg_model_3, which = 1) #Residuals vs. fitted

plot(reg_model_3, which = 2) #Normal qqplot

ggplot()+geom_histogram(aes(resid(reg_model_3))) #Histogram of residuals

ggplot(final_data, aes(transform_profit))+geom_histogram() #Histogram of outcome variable

plot(reg_model_3, which = 3) #Constant variance plot (scale-location)
```