Code ▾

Education Relationship with Aggricultural Profits

By: Rommel Artola

To analyze the effect of education on agricultural profits, we’re going to, firstly, choose just the head of household as the “target” individual to represent that specific household (or, profit point). Since the data set did not lend itself to a scalar variable of education very easily, we’ll proceed to use the binary (dummy) variables of whether the head of household had any level of education or not, where 1 is educated and 0 is not educated.

In the head() function below we can see a brief overview of the data set. Where House_ID is the unique identifier of the property. Attended_School is the binary value to depict whether the head of household of that specific property had any level of school. Lastly, the remaining three columns are tied to the property overall where:

1. Savings show the amount of savings participants disclosed on the survey as a numeric value,
2. Max_Farm_Land_Size_Norm is the normalized farm land size in acres, and
3. Profit_Per_Acre is the amount of profit in Ghanaian Cedis divided by the acres of the farm. We will use profit per acres (instead of total profits) to normalize inequality in farm land sizes.

Hide

```
Df_Subset <-  Main_Dataset_Household %>%
  select(House_ID, Attended_School_TF_Head, Tot_Savings, Max_Farm_Land_Size_Norm, Profit_Per_Acr
e) %>%
  rename(Attended_School = Attended_School_TF_Head)

head(Df_Subset)
```

	House_ID <chr>	Attended_School <dbl>	Tot_Savings <dbl>	Max_Farm_Land_Size_Norm <dbl>	Profit_Per_Acre <dbl>
1	4002-1	1	0	4	122600.00
2	4002-11	0	0	11	3385.00
3	4002-12	1	11000	4	109871.25
4	4002-13	0	0	3	84333.33
5	4002-14	0	0	1	839500.00
6	4002-16	0	0	21	6476.19

6 rows

The null hypothesis on this is that attending any amount of schooling will not have an impact on the agricultural profits. Let’s first examine a linear model on the variables, and then change the relationship to better fit the model if, and, as needed. This subset data frame has 4,107 observations across 5 variables.

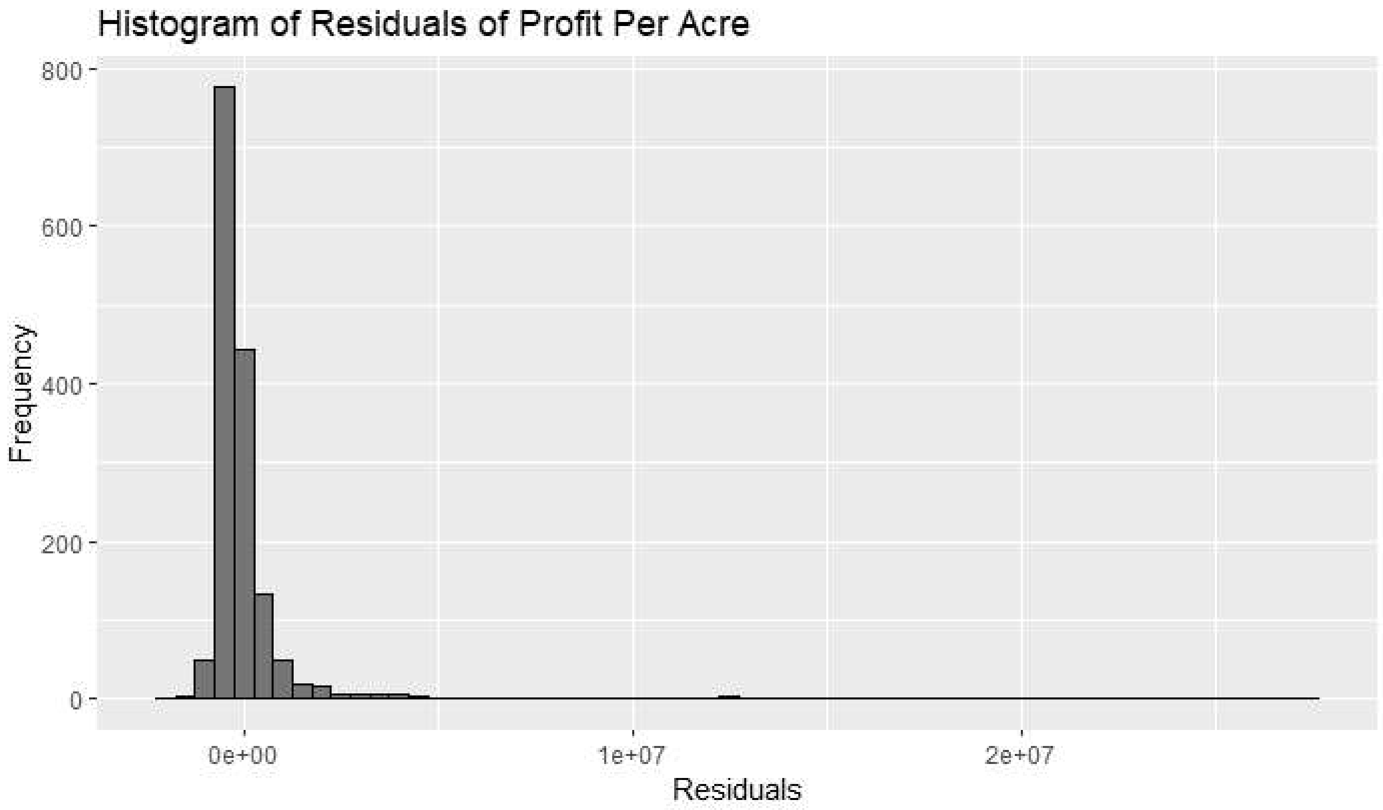
Hide

```
model_profit_on_educ_linear <- lm(formula = Profit_Per_Acre ~ Attended_School +
                                                            Tot_Savings +
                                                            Max_Farm_Land_Size_Norm,
                           data = Df_Subset)

summary(model_profit_on_educ_linear)
```

```

Call:
lm(formula = Profit_Per_Acre ~ Attended_School + Tot_Savings + 
    Max_Farm_Land_Size_Norm, data = Df_Subset)

Residuals:
     Min       1Q   Median       3Q      Max 
-2064932  -472732  -287277    33251 27364828 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              5.073e+05  6.255e+04   8.109 1.03e-15 ***
Attended_School          1.621e+05  7.526e+04   2.154   0.0314 *  
Tot_Savings              3.547e-02  3.897e-02   0.910   0.3629    
Max_Farm_Land_Size_Norm -1.845e+04  4.342e+03  -4.249 2.28e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1423000 on 1531 degrees of freedom
  (2572 observations deleted due to missingness)
Multiple R-squared:  0.01415,   Adjusted R-squared:  0.01222 
F-statistic: 7.326 on 3 and 1531 DF,  p-value: 7.077e-05
```

Hide

```
ggplot(model_profit_on_educ_linear, aes(x = model_profit_on_educ_linear$residuals)) +
  geom_histogram(bins = 60, fill = "steelblue", color = "black") +
  labs(title = "Histogram of Residuals of Profit Per Acre",
       x = "Residuals",
       y = "Frequency",
       caption = "Dependant variables were schooling, savings, and farm land size.")
```


Histogram of Residuals of Profit Per Acre

Dependant variables were schooling, savings, and farm land size.

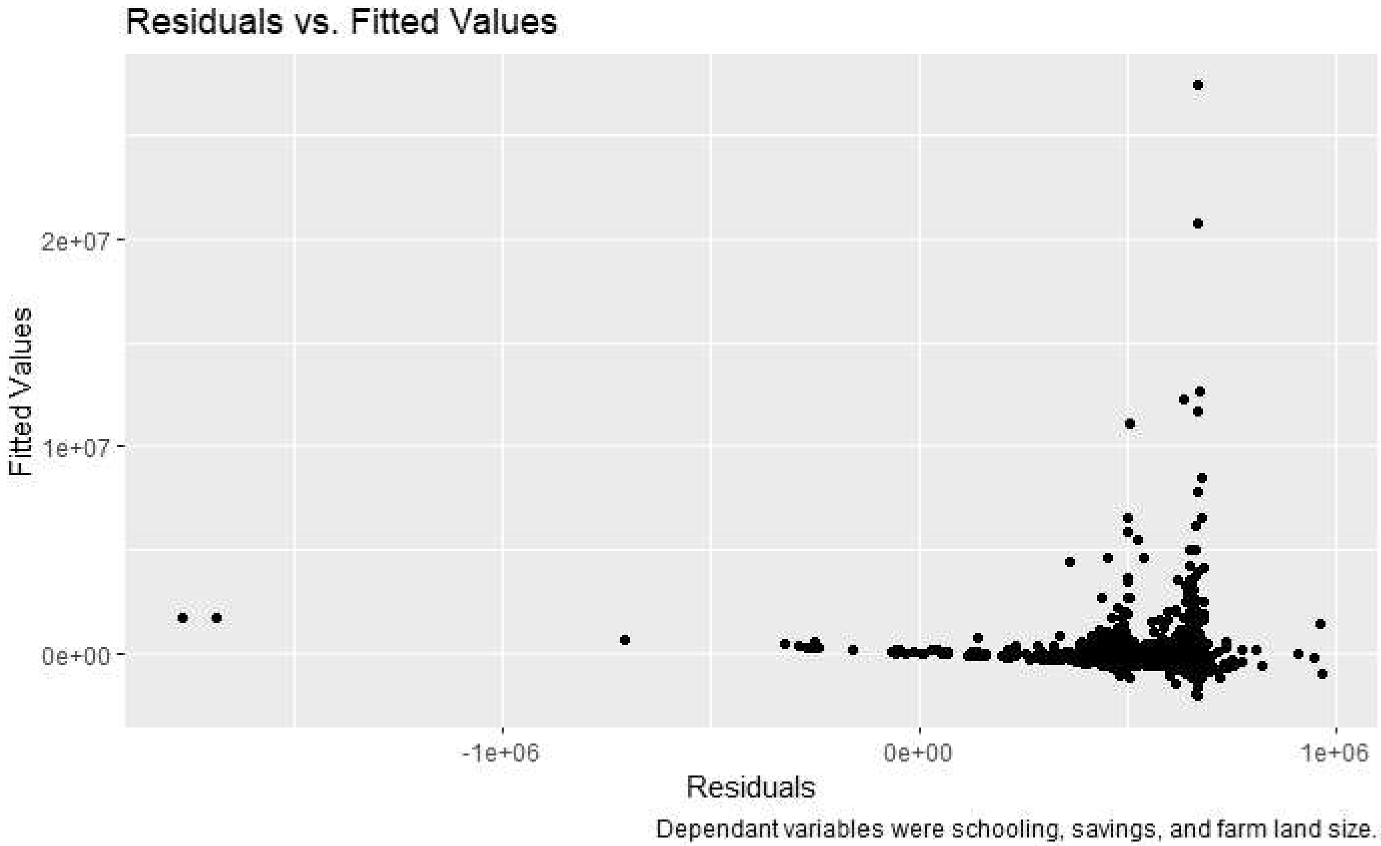
The basic linear regression model shows that attending school, at any level, leads to an estimated increase in agricultural profits of 162,100 Ghanaian Cedis (GHS). Additionally, this is statistically significant at the 5% level, and since our Pr(>|t|) value is less than our significance level of 0.05, we fail to reject the null hypothesis in favor of the alternative hypothesis. However, we do so with caution, as with a multiple r-squared of 0.01415, this indicates that this model is only representative for 1.415% of that variance, which is quite poor.

As an additional analysis, total savings is not significant in this model, but as one might expect, the size of the farm is very statistically significant – interestingly enough, farm size is related inversely. In other words, this model suggests that having too great of a farm size leads to lower agricultural profits. This could be to a potential polynomial relationship when we're only analyzing a linear model, or because having too many acres becomes unsustainable at a certain point for some individuals which leads to not being able to benefit from scales of economies.

Lastly, the fitted residuals histogram is relatively normally distributed, when discounting the outliers in the residuals. We can see them in the scatter below.

Hide

```
ggplot(model_profit_on_educ_linear, aes(x = model_profit_on_educ_linear$fitted.values,
                                        y = model_profit_on_educ_linear$residuals)) +
         geom_point() +
         labs(title = "Residuals vs. Fitted Values",
              x = "Residuals",
              y = "Fitted Values",
              caption = "Dependant variables were schooling, savings, and farm land size.")
```



Residuals vs. Fitted Values

Dependant variables were schooling, savings, and farm land size.

In this model, we can see some heteroscedasticity as the randomness of the fitted vs residuals become less constant on the low and high values of the fitted values.

To try and improve the model, we will examine a semi-log relationship on the dependent variable and a polynomial interaction on farm size independent variable. For this model, it is critical to note that since we'll be taking the log() of the dependent variable, profit per acre, we will have to filter for only values that are greater than 0. Due to this exclusion, the observations have been reduced to 3,552, down from our original 4,107 observations.

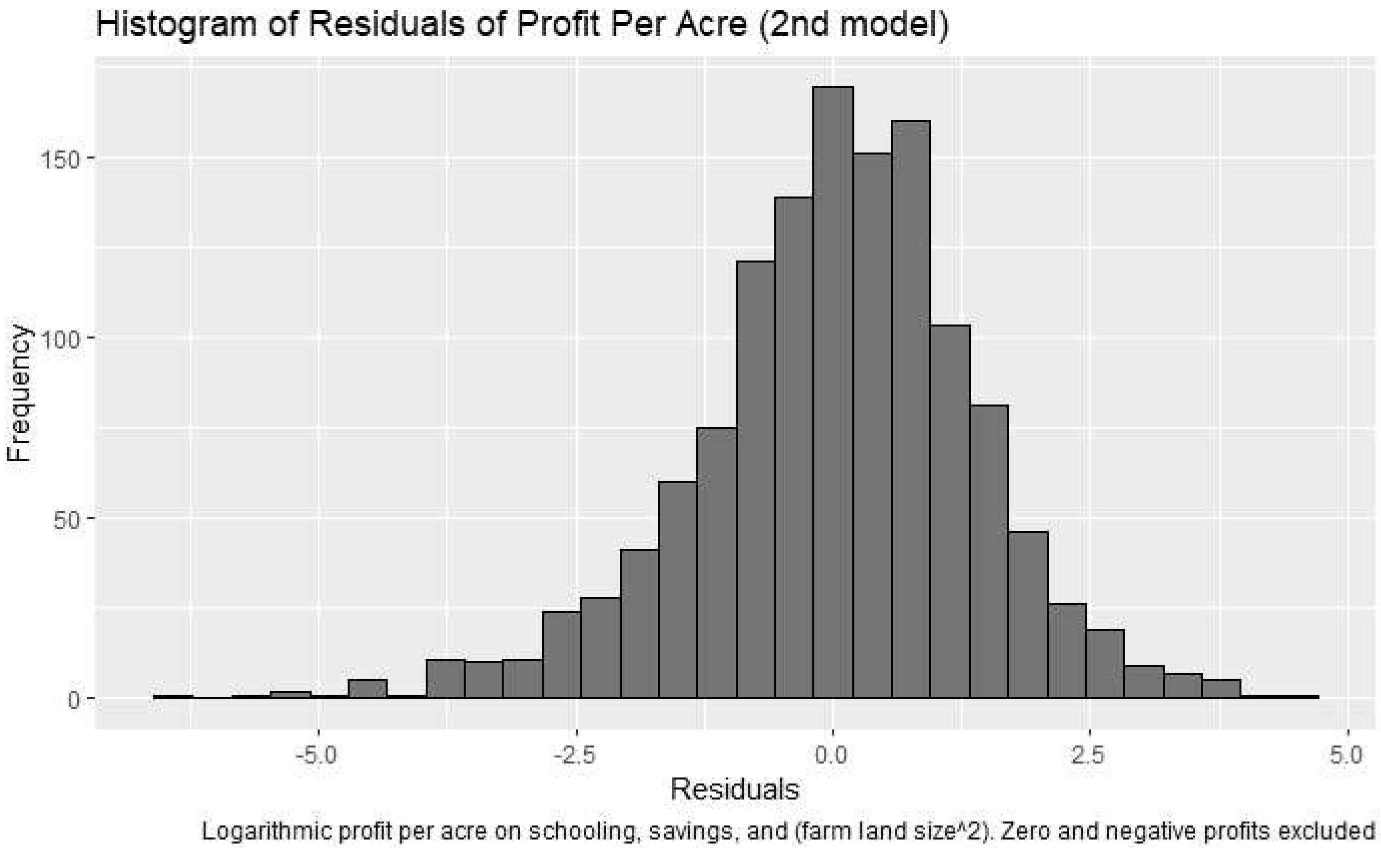
Hide

```
Df_Subset_Positive <- Df_Subset %>%
  filter(Profit_Per_Acre > 0)

model_profit_on_educ_2 <- lm(formula = log(Profit_Per_Acre) ~ Attended_School +
                                                            Tot_Savings +
                                                            I(Max_Farm_Land_Size_Norm*2),
                             data = Df_Subset_Positive)

summary(model_profit_on_educ_2)
```

```

Call:
lm(formula = log(Profit_Per_Acre) ~ Attended_School + Tot_Savings + 
    I(Max_Farm_Land_Size_Norm * 2), data = Df_Subset_Positive)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4710 -0.7675  0.0831  0.8900  4.4659 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.240e+01  6.548e-02 189.337  < 2e-16 ***
Attended_School                 2.894e-01  7.936e-02   3.646 0.000277 ***
Tot_Savings                     1.506e-07  4.121e-08   3.655 0.000267 ***
I(Max_Farm_Land_Size_Norm * 2) -2.176e-02  2.204e-03  -9.872  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.393 on 1305 degrees of freedom
  (2243 observations deleted due to missingness)
Multiple R-squared:  0.08022,	Adjusted R-squared:  0.07811 
F-statistic: 37.94 on 3 and 1305 DF,  p-value: < 2.2e-16
```

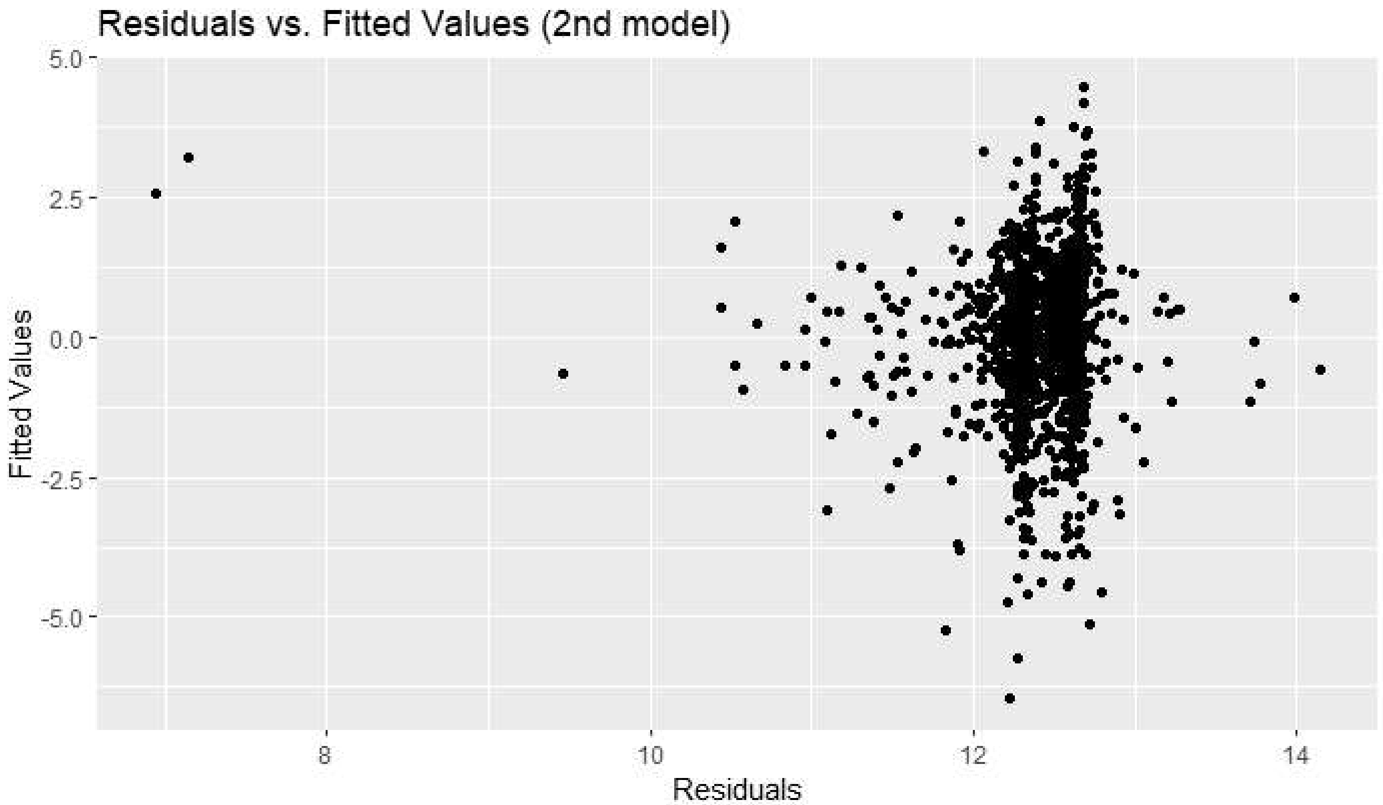
Hide

```
ggplot(model_profit_on_educ_2, aes(x = model_profit_on_educ_2$residuals)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black") +
  labs(title = "Histogram of Residuals of Profit Per Acre (2nd model)",
       x = "Residuals",
       y = "Frequency",
       caption = "Logarithmic profit per acre on schooling, savings, and (farm land size^2). Zer
o and negative profits excluded")
```



Hide

```
ggplot(model_profit_on_educ_2, aes(x = model_profit_on_educ_2$fitted.values,
                                        y = model_profit_on_educ_2$residuals)) +
         geom_point() +
         labs(title = "Residuals vs. Fitted Values (2nd model)",
              x = "Residuals",
              y = "Fitted Values",
              caption = "Logarithmic profit per acre on schooling, savings, and (farm land size^
2). Zero and negative profits excluded")
```



Logarithmic profit per acre on schooling, savings, and (farm land size^2). Zero and negative profits excluded

After having removed the negative and 0 values from the agricultural profits due to performing a log, as well as making farm_size a polynomial interaction, we see a vast improvement in the statistical significance of our independent variables. In addition to that, our t-values are now all much closer to 0 across the board and our histogram of residuals also looks much more normal straight out of the box. Lastly, our multiple R-squared increased several folds, to now this model explains a hair over 8% of the variance.

Looking at our visuals now we see that our fitted versus residual values above seems to have more randomness than before, and although there are still a few values on the extremes ends, there are not as many as before.

For the last test, we're going to see if a third degree polynomial of farm size improves our estimate, t-value, or multiple R-squared of the farm size variable noting that our values were -2.176e-02, < 2e-16, and 0.08022 respectively.

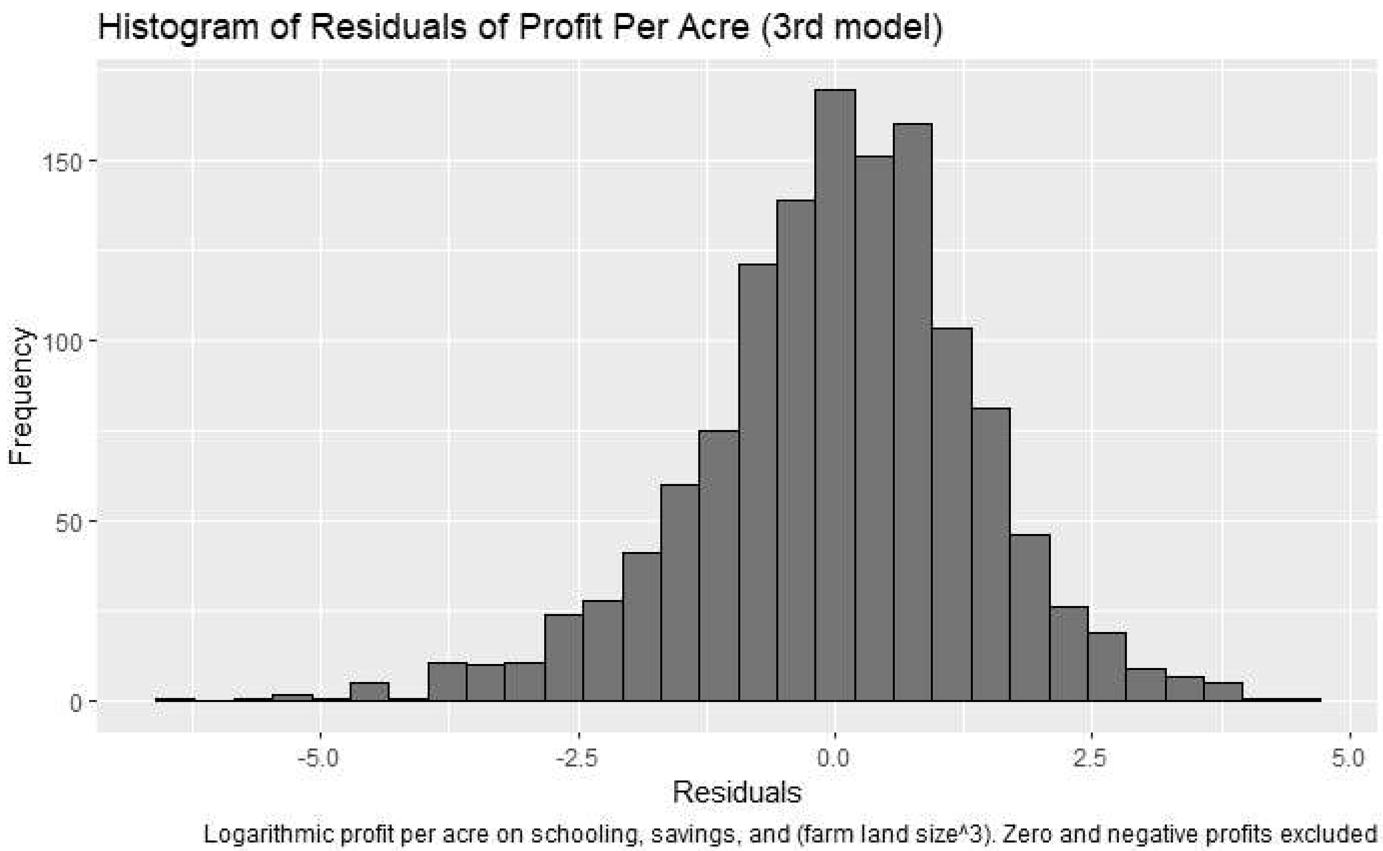
Hide

```
model_profit_on_educ_3 <- lm(formula = log(Profit_Per_Acre) ~ Attended_School +
                                                             Tot_Savings +
                                                             I(Max_Farm_Land_Size_Norm*3),
                             data = Df_Subset_Positive)

summary(model_profit_on_educ_3)
```

```

Call:
lm(formula = log(Profit_Per_Acre) ~ Attended_School + Tot_Savings + 
    I(Max_Farm_Land_Size_Norm * 3), data = Df_Subset_Positive)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4710 -0.7675  0.0831  0.8900  4.4659 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.240e+01  6.548e-02 189.337  < 2e-16 ***
Attended_School                 2.894e-01  7.936e-02   3.646 0.000277 ***
Tot_Savings                     1.506e-07  4.121e-08   3.655 0.000267 ***
I(Max_Farm_Land_Size_Norm * 3) -1.451e-02  1.470e-03  -9.872  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.393 on 1305 degrees of freedom
  (2243 observations deleted due to missingness)
Multiple R-squared:  0.08022,	Adjusted R-squared:  0.07811 
F-statistic: 37.94 on 3 and 1305 DF,  p-value: < 2.2e-16
```

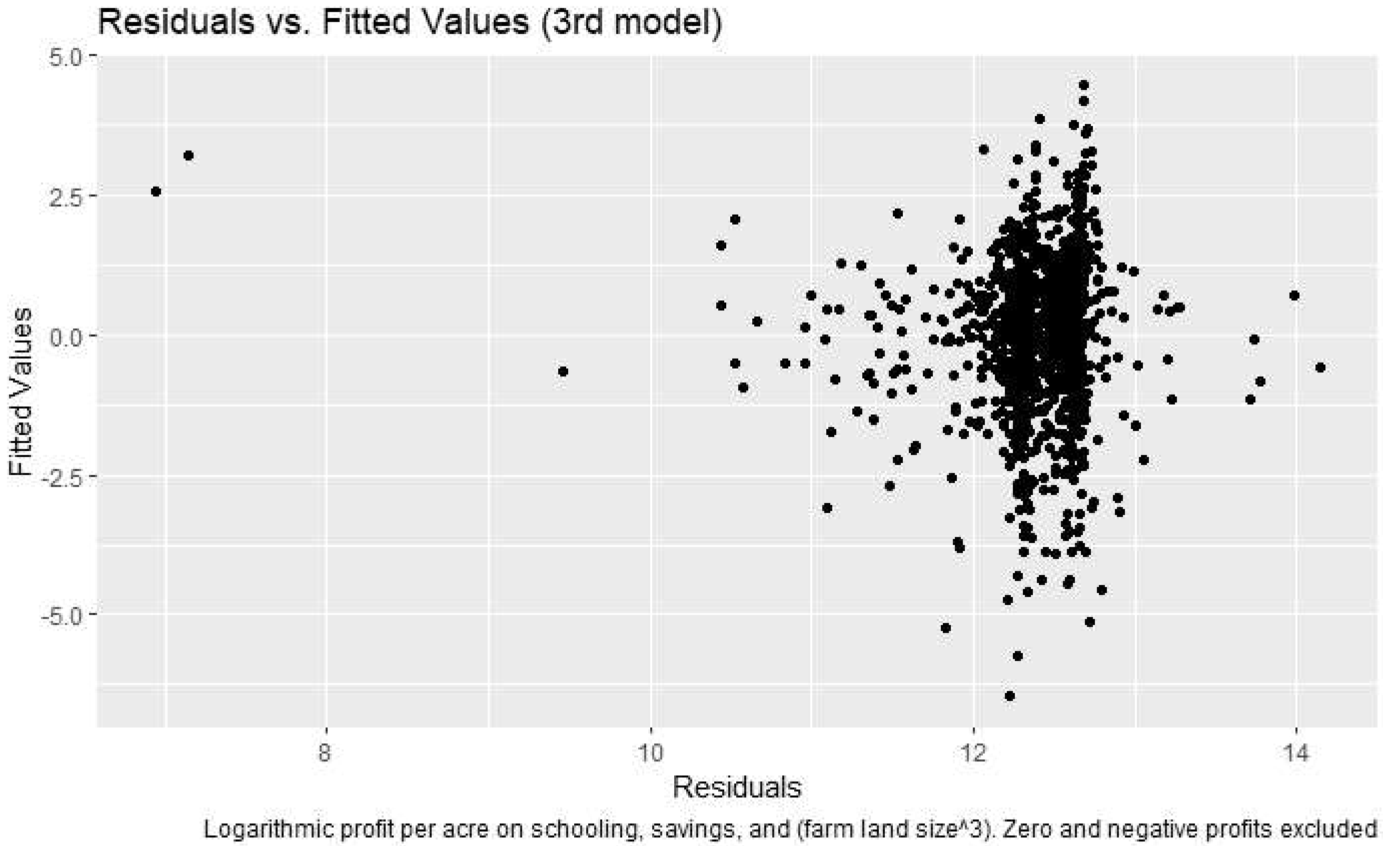
Hide

```
#Histogram
ggplot(model_profit_on_educ_2, aes(x = model_profit_on_educ_3$residuals)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black") +
  labs(title = "Histogram of Residuals of Profit Per Acre (3rd model)",
       x = "Residuals",
       y = "Frequency",
       caption = "Logarithmic profit per acre on schooling, savings, and (farm land size^3). Zer
o and negative profits excluded")
```



Hide

```
#Fitted vs Residuals
ggplot(model_profit_on_educ_3, aes(x = model_profit_on_educ_3$fitted.values,
                                        y = model_profit_on_educ_3$residuals)) +
         geom_point() +
         labs(title = "Residuals vs. Fitted Values (3rd model)",
              x = "Residuals",
              y = "Fitted Values",
              caption = "Logarithmic profit per acre on schooling, savings, and (farm land size^3). Zero and negative profits excluded")
```


Residuals vs. Fitted Values (3rd model)

Logarithmic profit per acre on schooling, savings, and (farm land size^3). Zero and negative profits excluded

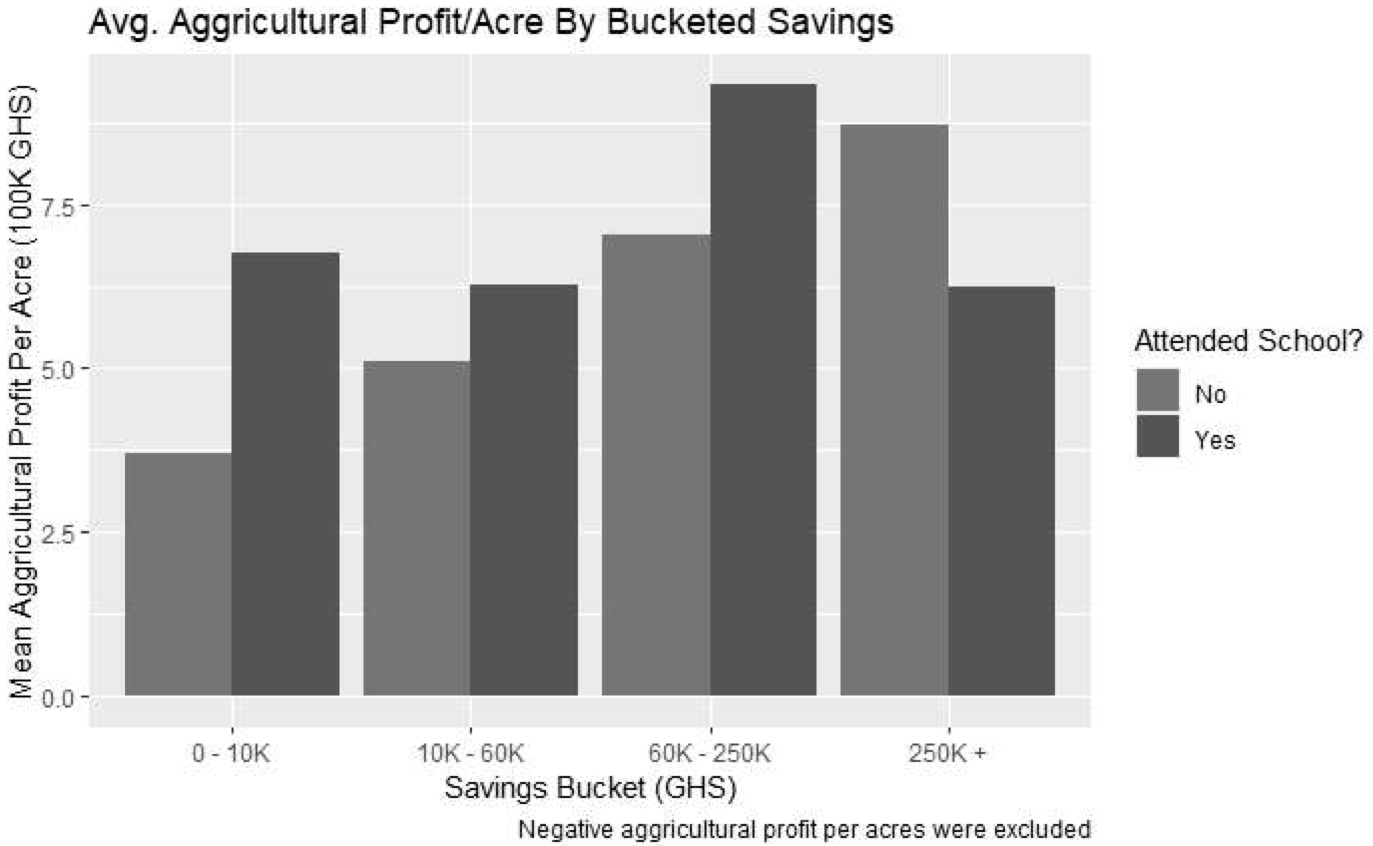
The new estimates is now -1.451e-02, with a t-value of < 2e-16, and a new multiple R-squared of 0.08022. Additionally, the histogram and scatter of fitted vs residuals remains largely unchanged. This leads to a small change in the context of the big picture, which leads me to believe that a two-degree polynomial is better-fitted than a 3rd degree polynomial. However, when we looked back, a two-degree polynomial is better than a non-polynomial interaction on farm size. Additionally, not much more variance was able to be explained with this model using the multiple R-squared. This model does not seem to improve the fit over the model with 2nd degree polynomial.

Now that we've identified the best model to answer the original hypothesis, we will proceed to answer that question with the semi-logarithmic model with a 2nd degree polynomial. So, due to the p-value being lower than the significance level, we proceed with rejecting the null hypothesis in favor of the alternative hypothesis that education has an effect on agricultural profits. In fact, the model suggests that those individuals with agricultural profits greater than 0 GHS, and that had any level of education, are likely to enjoy a 28.94% higher level of profits per acre. However, as an additional warning, this model is only able to explain roughly 8% of the variance on our dependent variable. So, although significant, it does not tell the entire story, not even close.

As a last graphical representation of the data, we will bucket the ranges of savings into 4 buckets of similarly sized n and show average agricultural profits per acre split into educated individuals vs non-educated individuals. As a warning for this graph, though, do note that, like our log model, negative profits have been excluded from the analysis, otherwise we would not have been able to determine a reliable means of agricultural profit per acre.

Hide

```
#Used to manually format x-axis of ggplot.
level_order <-  c("0 - 10K","10K - 60K","60K - 250K","250K +")

Agg_Profit_On_Education_DF_Bucket <- Df_Subset_Positive %>%
  #Adding a custom bucketing column based on savings amounts
  mutate(Savings_Bucket = case_when(Tot_Savings >= 250000 ~ "250K +",
                                    Tot_Savings >= 60000 ~ "60K - 250K",
                                    Tot_Savings >= 10000 ~ "10K - 60K",
                                    Tot_Savings >= 0 ~ "0 - 10K",
                                    TRUE ~ "DNI")) %>%
  #Filter out the ones that don't match the range and the profits less than 0 due to errors othe
rwise
  filter(Savings_Bucket != "DNI",
        Profit_Per_Acre >= 0) %>%
 #Grouping by two columns to be able to show difference with schooling and wrap graphs
  group_by(Savings_Bucket, Attended_School) %>%
              #tally() #just a counter middle step to get the groups into somewhat similar size
s, manual step.
  #summarizing mean of profit per acre based on the group statement above
  summarize(Avg_Agg_Profit_Per_Acre = mean(Profit_Per_Acre)/100000) %>%
  #Renaming the Attended_School variable to be categorical
  mutate(Attended_School = case_when(Attended_School == 0 ~ "No",
                                        Attended_School == 1 ~ "Yes")) %>%
  drop_na()



Bucketed_Savings_ggplot <- ggplot(Agg_Profit_On_Education_DF_Bucket, aes(x = factor(Savings_Buck
et, level = level_order),
                                                                         y = Avg_Agg_Profit_Per_
Acre,
                                                                         fill = Attended_Schoo
l)) +
  geom_bar(stat = "identity", #summarise the data of the bars
           position = "dodge") + #position is to stack bars next to each other.
  labs(title = "Avg. Aggricultural Profit/Acre By Bucketed Savings",
       x = "Savings Bucket (GHS)",
       y = "Mean Aggricultural Profit Per Acre (100K GHS)",
       caption = "Negative aggricultural profit per acres were excluded",
       fill = "Attended School?") +
  #manually filling colors to match the rest
  scale_fill_manual(values = c("steelblue", "saddlebrown"))


Bucketed_Savings_ggplot
```



Interestingly, individuals that had some form of education outperformed those with no schooling in average profits per acre in every bucket except for those that had a profit of 250K+ GHS. Although no conclusions or further testing was done on this, it is a point that brings up many questions and the possibility of even deeper levels of analysis.