From Operations to Outcomes: The ROA Story of Taiwanese Enterprises

Krish Sanjay Bhalala

20 November, 2024

Introduction:

This project investigates the financial performance of Taiwanese companies listed on the Taiwan Stock Exchange from 1999 to 2009. Our focus is on understanding how specific financial indicators influence a company's profitability, particularly its Return on Assets (ROA) before interest and after taxes. By examining this relationship, we aim to provide valuable insights for investors, managers, and financial analysts interested in Taiwanese corporate performance for their corporate investments (i.e ROA).

The response variable we seek to model is **Return on Assets (ROA) before interest and after tax**. ROA is a crucial metric that indicates how efficiently a company utilizes its assets to generate profits. It provides a clear picture of overall financial health and management effectiveness, making it an ideal measure of a company's performance.

We propose two explanatory variables for our model: Operating Profit Rate and Non-industry income and expenditure revenue.

The Operating Profit Rate reflects core profitability, with a higher rate indicating more efficient asset utilization. We expect a positive relationship between this variable and ROA, as companies with efficient operations are likely to achieve higher returns.

The second variable, **Non-industry income and expenditure revenue**, captures performance outside core business activities. While significant non-industry income might boost overall profitability and ROA, excessive reliance on it could indicate potential issues with the main business model. By combining these two variables, we aim to provide a comprehensive view of factors influencing ROA. The Operating Profit Rate offers insight into operational efficiency, while Non-industry income accounts for additional profitability sources. Together, they should yield a nuanced prediction of a company's ability to generate returns from its assets.

Data Set:

```r
data <- read.csv("data.csv")

set.seed(92)
# Select the first 50 rows
data <- data[sample(1:nrow(data), 50), ]

# Select the specific columns
selected_columns <- c("ROA.A..before.interest.and...after.tax",
                      "Operating.Profit.Rate",
                      "Non.industry.income.and.expenditure.revenue")
```

```r
data <- data[selected_columns]

# Rename columns for easier access
colnames(data) <- c("ROA_Before_Interest_After_Tax", "Operating_Profit_Rate",
                    "Non_Industry_Income_Expenditure")

data[,2] <- data[,2] + runif(42, -0.0005, 0.0005)

# Display the first few rows of the dataset

# Check if packages are installed, if not, install them
if (!requireNamespace("knitr", quietly = TRUE)) install.packages("knitr")
if (!requireNamespace("kableExtra", quietly = TRUE)) install.packages("kableExtra")

library(knitr)
library(kableExtra)

kable(data,
      col.names = c("ROA Before Interest & After Tax",
                    "Operating Profit Rate",
                    "Non-Industry Income & Expenditure"),
      align = c('r', 'r', 'r'),
      caption = "Financial Metrics (First 50 Rows)",
      digits = 7) |>
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = FALSE) |>
  column_spec(1, border_right = TRUE) |>
  column_spec(2, border_right = TRUE) |>
  column_spec(3, border_right = TRUE)
```

Table 1: Financial Metrics (First 50 Rows)

	ROA Before Interest & After Tax	Operating Profit Rate	Non-Industry Income & Expenditure
4562	0.6046119	0.9992759	0.3038578
1218	0.5860772	0.9988207	0.3035347
3895	0.4917139	0.9986146	0.3034627
1290	0.5836241	0.9985814	0.3036534
4844	0.6037942	0.9991142	0.3039151
5500	0.5165722	0.9992885	0.3033790
2652	0.5686328	0.9991915	0.3035228
4127	0.5547863	0.9989025	0.3034991
2501	0.6844200	0.9989166	0.3107813
4752	0.5245857	0.9986382	0.3034172
3078	0.5613280	0.9987504	0.3035151
4504	0.6032490	0.9993038	0.3035098
5812	0.4844636	0.9985496	0.3033628
507	0.5445377	0.9990867	0.3034602
982	0.5415940	0.9986752	0.3034175
5236	0.5835696	0.9995645	0.3035390
1376	0.6669211	0.9986907	0.3062369
14	0.5333079	0.9990668	0.3034917

3238	0.5503162	0.9990000	0.3035262
5078	0.5892935	0.9995572	0.3035400
2906	0.6164413	0.9988764	0.3034709
4820	0.4788487	0.9987451	0.3034327
4228	0.4536633	0.9990393	0.3035071
1530	0.5773005	0.9992316	0.3036809
5922	0.5589294	0.9991996	0.3035685
4739	0.6102268	0.9991540	0.3036437
351	0.6203663	0.9992450	0.3040420
2017	0.5008177	0.9990906	0.3035078
2526	0.5372874	0.9986928	0.3033679
2970	0.5629634	0.9991843	0.3039521
3762	0.5506978	0.9986807	0.3034356
4700	0.6053205	0.9987833	0.3036891
2213	0.5367423	0.9981974	0.3050245
2714	0.6202028	0.9997414	0.3035480
2472	0.6301788	0.9996684	0.3093391
2786	0.5714130	0.9993705	0.3034037
6627	0.4655473	0.9989408	0.3032183
328	0.5542412	0.9986714	0.3035388
4907	0.5938181	0.9988170	0.3035850
3361	0.5877671	0.9992337	0.3035008
3120	0.5783908	0.9995830	0.3035414
4123	0.6073921	0.9992563	0.3036082
2646	0.5542957	0.9993100	0.3034846
2975	0.5998147	0.9987055	0.3035163
4704	0.4979830	0.9983747	0.3031571
2247	0.5528238	0.9986247	0.3035427
5834	0.4843000	0.9989637	0.3034914
4721	0.6674117	0.9997407	0.3041032
3435	0.5454645	0.9990555	0.3035340
2444	0.5493349	0.9989494	0.3034849

The Taiwanese Bankruptcy Prediction dataset, donated to UC Irvine Machine Learning Repository on 6/27/2020, contains financial data from 6,819 companies listed on the Taiwan Stock Exchange between 1999 and 2009. The dataset includes 95 financial variables and does not have any missing values. But for this project we will use 3 financial variables only in order to predict one from other 2

MLA Citation:

Taiwanese Bankruptcy Prediction [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D.

Response Variable

- **ROA (Y): Return on Assets before interest and after tax**, *expressed as a percentage* (%). Specifically, ROA is calculated by dividing a company's net income (before interest and after tax) by its total assets. It represents the percentage of profit a company generates relative to its total assets, indicating how effectively the company uses its assets to produce earnings.

Explainatory variable

- **Operating Profit Rate** (X_1)**:** This is the ratio of operating income to net sales, *expressed as a percentage* (%). It measures the profitability of a company's core business operations, excluding the effects of investments, taxes, and interest expenses.In other words, how much profit a company makes on each dollar of sales from its main business activities, before accounting for non-operating expenses. A higher operating profit rate suggests more efficient operations and stronger pricing power.
- **Non-industry income and expenditure revenue** (X_2)**:** This represents the monetary value of revenue or expenses not directly related to the company's main business activities, *measured in New Taiwan Dollars (NT$)* like Interest income from investments, Gains or losses from foreign exchange transactions, etc

Now lets first fit the simple linear model for ROA (Y) Return on Assets before interest and after tax and Operating Profit Rate (X_1)

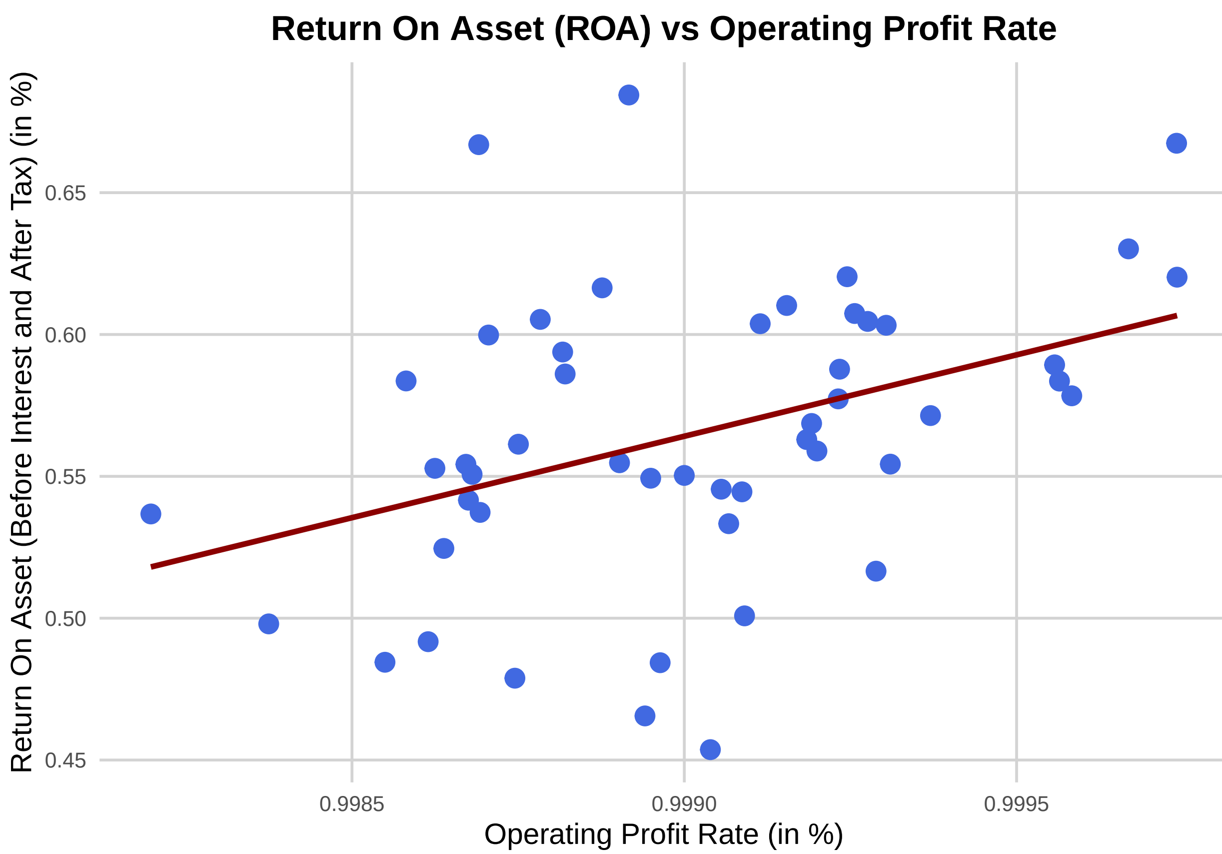
```r
# Load necessary library
library(ggplot2)

# Fit linear model Return on Asset ~ Operating profit rate
ROA_OPR.lm <- lm(ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate,
               data = data)

# Create a custom theme for consistent styling
custom_theme <- theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    axis.title.x = element_text(size = 11),
    axis.title.y = element_text(size = 11),
    axis.text = element_text(size = 8),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank()
  )

# Create scatterplot for ROA vs Operating Profit Rate
plot1 <- ggplot(data, aes(x = Operating_Profit_Rate,
                          y = ROA_Before_Interest_After_Tax)) +
  geom_point(color = "royalblue", size = 3, shape = 19) +
  geom_smooth(method = "lm", color = "darkred", se = FALSE) +
  labs(title = "Return On Asset (ROA) vs Operating Profit Rate",
       x = "Operating Profit Rate (in %)",
       y = "Return On Asset (Before Interest and After Tax) (in %)") +
  custom_theme

# Display the first model
plot1
```




```r
# Calculate r-squared values
summary(ROA_OPR.lm)$r.squared
```

```
## [1] 0.1562472
```

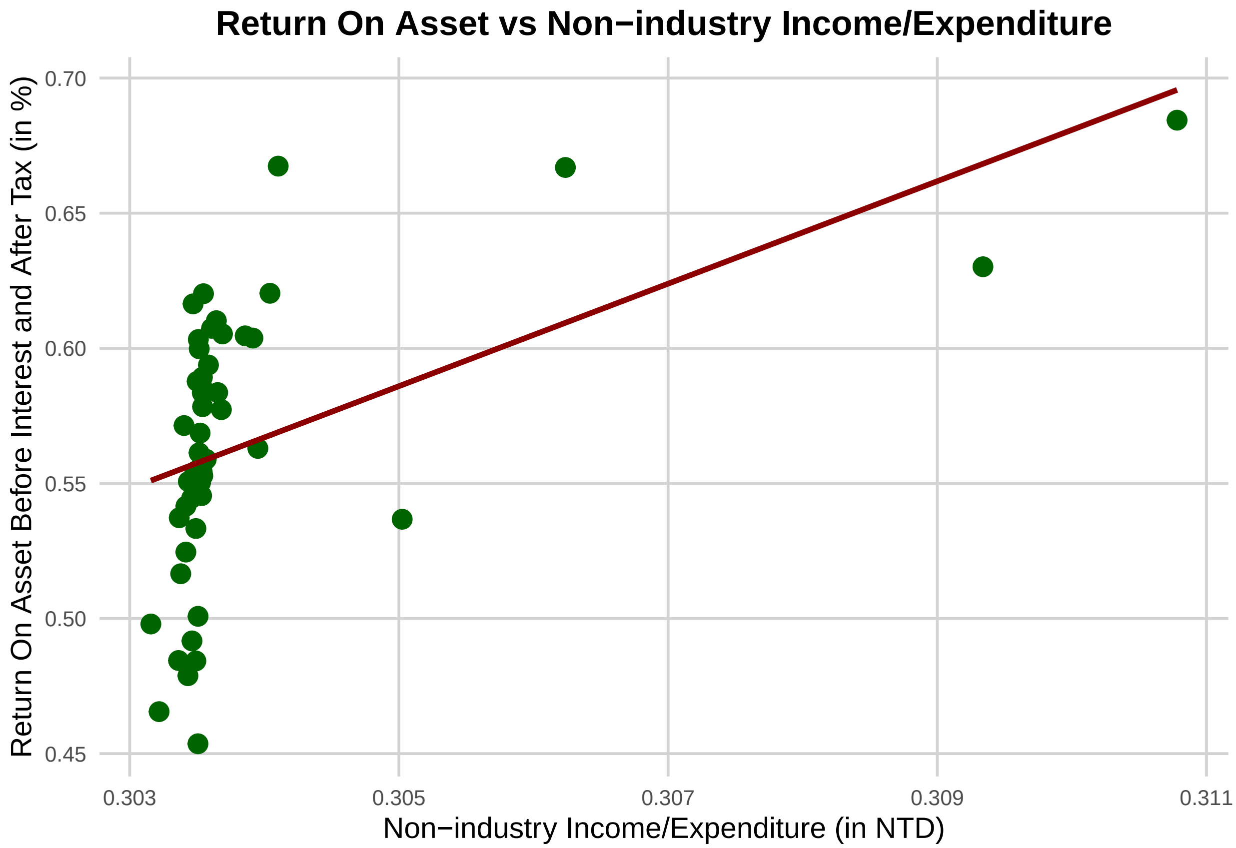
The r^2 value for ROA vs Operating Profit Rate is 0.16, indicating a weaker relationship.

Now we will fit the simple linear model for ROA (Y) Return on Assets before interest and after tax and Non Industry Income Expenditure (X_2)

```r
# Fit linear model Return on Asset ~ Non Industry Income Expenditure
ROA_NIIE.lm <- lm(ROA_Before_Interest_After_Tax ~ Non_Industry_Income_Expenditure,
               data = data)

# Create scatterplot for ROA vs Non-industry Income/Expenditure
plot2 <- ggplot(data, aes(x = Non_Industry_Income_Expenditure,
                          y = ROA_Before_Interest_After_Tax)) +
  geom_point(color = "darkgreen", size = 3, shape = 19) +
  geom_smooth(method = "lm", color = "darkred", se = FALSE) +
  labs(title = "Return On Asset vs Non-industry Income/Expenditure",
       x = "Non-industry Income/Expenditure (in NTD)",
       y = "Return On Asset Before Interest and After Tax (in %)") +
  custom_theme

# Display the second plot
plot2
```




```r
summary(ROA_NIIE.lm)$r.squared
```

```
## [1] 0.2541733
```

The r^2 value for ROA vs Non-industry income and expenditure revenue is 0.25, suggesting a weaker but still slightly better relationship than previous model.

Preliminary Model:

Model for ROA and Operating Profit Rate

We will first fit a simple linear regression model with Return on Assets (ROA) as the response variable and Operating Profit Rate as the predictor variable.

```r
summary(ROA_OPR.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.112716 -0.021185 -0.003818  0.028337  0.125084
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -56.78      19.23  -2.952  0.00487 **
## Operating_Profit_Rate    57.40      19.25   2.981  0.00450 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04755 on 48 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1387
## F-statistic: 8.889 on 1 and 48 DF,  p-value: 0.004496
```

The adjusted R-squared (R^2_{adj}) for this Operating Profit Rate model is 0.14

The resulting regression equation for our simple regression model Return on Asset before Tax and Interest $\sim$ Operating Profit Rate is:

$$\hat{\text{ROA}} = -56.78 + 57.40 \cdot \text{Operating Profit Rate}$$

Model for ROA and Non-industry Income and Expenditure

Next, we will fit a simple linear regression model with ROA as the response variable and Non-industry Income and Expenditure as the predictor variable.

```r
summary(ROA_NIIE.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Non_Industry_Income_Expenditure,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10401 -0.02249 -0.00284  0.03375  0.09844
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -5.198      1.425  -3.648 0.000650 ***
## Non_Industry_Income_Expenditure   18.965      4.689   4.045 0.000189 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0447 on 48 degrees of freedom
## Multiple R-squared:  0.2542, Adjusted R-squared:  0.2386
## F-statistic: 16.36 on 1 and 48 DF,  p-value: 0.0001893
```

The adjusted R-squared (R^2_{adj}) for this Non-industry Income and Expenditure model is 0.24

The resulting regression equation for our simple regression model Return on Asset before Tax and Interest $\sim$ Non-industry Income and Expenditure is:

$$\hat{\text{ROA}} = -5.20 + 18.97 \cdot \text{Non-industry Income and Expenditure}$$

Combined Model for ROA

We will fit a multiple linear regression model incorporating both Operating Profit Rate and Non-industry Income and Expenditure as predictors for ROA.

```
ROA_combined.lm <- lm(ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
                      Non_Industry_Income_Expenditure, data = data)
summary(ROA_combined.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.105750 -0.016165  0.001691  0.029435  0.076875
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -55.84      16.71  -3.343 0.001636 **
## Operating_Profit_Rate              51.06      16.79   3.041 0.003851 **
## Non_Industry_Income_Expenditure    17.74       4.35   4.078 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0413 on 47 degrees of freedom
## Multiple R-squared:  0.3768, Adjusted R-squared:  0.3503
## F-statistic: 14.21 on 2 and 47 DF,  p-value: 1.493e-05
```

The resulting regression equation for our simple regression model Return on Asset before Tax and Interest $\sim$ Operating Profit Rate + Non-industry Income and Expenditure is:

$$\hat{\text{ROA}} = -55.84 + 51.06 \cdot \text{Operating Profit Rate} + 17.74 \cdot \text{Non-industry Income and Expenditure}$$

The adjusted R-squared (R^2_{adj}) for this combined model is 0.35, which is higher than both the ROA $\sim$ Operating Profit Rate model (0.14) and the ROA $\sim$ Non-industry Income and Expenditure model (0.24) by at least 0.1 units. This increase in R^2_{adj} indicates that the combined model provides a better fit to the data compared to the individual models, explaining a larger proportion of the variance in ROA.

Full Second-Order Model

Finally, we will fit a full second-order model, including quadratic terms and an interaction term to see if it is better model as compared to previous models or not

```
ROA_full_model.lm <- lm(ROA_Before_Interest_After_Tax ~
                   Operating_Profit_Rate +
                   Non_Industry_Income_Expenditure +
                   I(Operating_Profit_Rate^2) +
                   I(Non_Industry_Income_Expenditure^2) +
                   Operating_Profit_Rate:Non_Industry_Income_Expenditure,
                 data = data)
```

```r
summary(ROA_full_model.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure + I(Operating_Profit_Rate^2) +
##     I(Non_Industry_Income_Expenditure^2) + Operating_Profit_Rate:Non_Industry_Income_Expenditure,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.104188 -0.017968  0.000951  0.020771  0.071256
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                              -13695      38209
## Operating_Profit_Rate                                     20120      76530
## Non_Industry_Income_Expenditure                           23755      10581
## I(Operating_Profit_Rate^2)                                -6982      38468
## I(Non_Industry_Income_Expenditure^2)                      -5934       2578
## Operating_Profit_Rate:Non_Industry_Income_Expenditure   -20116      10592
##                                                        t value Pr(>|t|)
## (Intercept)                                             -0.358   0.7217
## Operating_Profit_Rate                                    0.263   0.7939
## Non_Industry_Income_Expenditure                          2.245   0.0298 *
## I(Operating_Profit_Rate^2)                              -0.182   0.8568
## I(Non_Industry_Income_Expenditure^2)                    -2.302   0.0261 *
## Operating_Profit_Rate:Non_Industry_Income_Expenditure  -1.899   0.0641 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03861 on 44 degrees of freedom
## Multiple R-squared:   0.49,  Adjusted R-squared:  0.4321
## F-statistic: 8.456 on 5 and 44 DF,  p-value: 1.138e-05
```

The regression equation for this model is:

$$
\begin{aligned}
\widehat{\mathrm{ROA}} = -13695 &+ 20120 \cdot (\text{Operating Profit Rate}) + 23755 \cdot (\text{Non-industry Income and Expenditure}) \\
&- 6982 \cdot (\text{Operating Profit Rate})^2 - 5934 \cdot (\text{Non-industry Income and Expenditure})^2 \\
&- 20116 \cdot (\text{Operating Profit Rate} \times \text{Non-industry Income and Expenditure})
\end{aligned}
$$

The adjusted R-squared (R^2_{adj}) for this Full Second Order model mentioned above is 0.43, which is slightly improved from the Combined Model (with $R^2_{adj} = 0.35$) but not too much. So we will need to refine our full model by removing less significant terms.

ANOVA Test

But lets first performed an Analysis of Variance (ANOVA) test to determine if at least one of the model terms in the full second-order model is statistically significant.

```r
summary(ROA_full_model.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure + I(Operating_Profit_Rate^2) +
##     I(Non_Industry_Income_Expenditure^2) + Operating_Profit_Rate:Non_Industry_Income_Expenditure,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.104188 -0.017968  0.000951  0.020771  0.071256
##
## Coefficients:
##                                                       Estimate Std. Error
## (Intercept)                                             -13695      38209
## Operating_Profit_Rate                                    20120      76530
## Non_Industry_Income_Expenditure                          23755      10581
## I(Operating_Profit_Rate^2)                               -6982      38468
## I(Non_Industry_Income_Expenditure^2)                     -5934       2578
## Operating_Profit_Rate:Non_Industry_Income_Expenditure   -20116      10592
##                                                       t value Pr(>|t|)
## (Intercept)                                            -0.358   0.7217
## Operating_Profit_Rate                                   0.263   0.7939
## Non_Industry_Income_Expenditure                         2.245   0.0298 *
## I(Operating_Profit_Rate^2)                             -0.182   0.8568
## I(Non_Industry_Income_Expenditure^2)                   -2.302   0.0261 *
## Operating_Profit_Rate:Non_Industry_Income_Expenditure  -1.899   0.0641 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03861 on 44 degrees of freedom
## Multiple R-squared:   0.49,  Adjusted R-squared:  0.4321
## F-statistic: 8.456 on 5 and 44 DF,  p-value: 1.138e-05
```

1. LEVEL OF SIGNIFICANCE: $\alpha = 0.05$
2. HYPOTHESES:

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
- H_a : At least one $\beta_i \neq 0$ for $i = 1, 2, 3, 4, 5$

3. DECISION RULE: Reject H_0 if p-value $\leq \alpha$.
4. TEST STATISTIC: $F = 8.456$
5. P-VALUE: $1.138 \times 10^5 \approx 0$
6. CONCLUSION: Reject H_0. Conclude there is strong evidence that at least one of the coefficients of (Operating Profit Rate, Non-Industry Income and Expenditure, (Operating Profit Rate)^2, (Non-Industry Income and Expenditure)^2, Operating Profit Rate $\times$ Non-Industry Income and Expenditure) in the full second-order model is non-zero and at least of those model terms mentioned before has a significant effect on Return on Assets before interest and after tax.

Model Refinement:

Now lets refine our complete second order model. We will do that by finding which terms of complete model are significant using summary()

```r
summary(ROA_full_model.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure + I(Operating_Profit_Rate^2) +
##     I(Non_Industry_Income_Expenditure^2) + Operating_Profit_Rate:Non_Industry_Income_Expenditure,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.104188 -0.017968  0.000951  0.020771  0.071256
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                              -13695      38209
## Operating_Profit_Rate                                     20120      76530
## Non_Industry_Income_Expenditure                           23755      10581
## I(Operating_Profit_Rate^2)                                -6982      38468
## I(Non_Industry_Income_Expenditure^2)                      -5934       2578
## Operating_Profit_Rate:Non_Industry_Income_Expenditure    -20116      10592
##                                                        t value Pr(>|t|)
## (Intercept)                                             -0.358   0.7217
## Operating_Profit_Rate                                    0.263   0.7939
## Non_Industry_Income_Expenditure                          2.245   0.0298 *
## I(Operating_Profit_Rate^2)                              -0.182   0.8568
## I(Non_Industry_Income_Expenditure^2)                    -2.302   0.0261 *
## Operating_Profit_Rate:Non_Industry_Income_Expenditure  -1.899   0.0641 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03861 on 44 degrees of freedom
## Multiple R-squared:   0.49,  Adjusted R-squared:  0.4321
## F-statistic: 8.456 on 5 and 44 DF,  p-value: 1.138e-05
```

Based on the summary output, the following coefficients appear to be significant:

- Non Industry Income Expenditure ($p - value = 0.03 < 0.05$)
- Non Industry Income Expenditure^2(($p - value = 0.03 < 0.05$))

Based on the significance of terms in the refined model, our proposed reduced model is:

ROA Before Interest After Tax $\sim$ Non Industry Income Expenditure + Non Industry Income Expenditure^2

```r
ROA_reduced_model.lm <- lm(ROA_Before_Interest_After_Tax ~
                          Non_Industry_Income_Expenditure +
                          I(Non_Industry_Income_Expenditure^2) ,
                        data = data)
```

Now lets perform a nested F-test to compare the full model to the above proposed reduced model to see whether it does a good job at predicting ROA before interest and after tax:

```
anova(ROA_reduced_model.lm, ROA_full_model.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ROA_Before_Interest_After_Tax ~ Non_Industry_Income_Expenditure +
##     I(Non_Industry_Income_Expenditure^2)
## Model 2: ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate + Non_Industry_Income_Expenditure +
##     I(Operating_Profit_Rate^2) + I(Non_Industry_Income_Expenditure^2) +
##     Operating_Profit_Rate:Non_Industry_Income_Expenditure
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1     47 0.086643
## 2     44 0.065588  3  0.021056 4.7085 0.006174 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value $= 0.006 < \alpha = 0.05$. There is sufficient evidence to conclude that at least one of the coefficients for Operating Profit Rate, (Operating Profit Rate)^2, or Operating_Profit_Rate $\times$ Non Industry Income Expenditure are non-zero. The reduced model, which excludes these terms, is not appropriate for predicting the Return on Assets before interest and after tax.

This means we have removed too many terms, so we need to analyze further to add back the important term.

Now, we will use the Variance Inflation Factor (VIF) function to check for multicollinearity and remove unnecessary terms from our full model. We will take out the term with the highest VIF and any terms with a VIF greater than 5 until all remaining terms have a VIF of less than 5, in other words untill all the model terms become significant.

```
library(car)
vif(ROA_full_model.lm)
```

```
##                                 Operating_Profit_Rate
##                                            23965156.9
##                       Non_Industry_Income_Expenditure
##                                             6826699.3
##                            I(Operating_Profit_Rate^2)
##                                            24173005.2
##                  I(Non_Industry_Income_Expenditure^2)
##                                              152378.1
## Operating_Profit_Rate:Non_Industry_Income_Expenditure
##                                             6971174.6
```

Since the VIF for Operating Profit Rate^2 $= 24,173,005.2$, which is greater than 5 and the highest among all terms, we will remove it from our model and then recalculate the VIF for the remaining variables.

```
# Removing Operating_Profit_Rate^2
ROA_reduced_model.lm <- lm(ROA_Before_Interest_After_Tax ~
                   Operating_Profit_Rate +
                   Non_Industry_Income_Expenditure +
                   I(Non_Industry_Income_Expenditure^2) +
```

```r
                   Operating_Profit_Rate:Non_Industry_Income_Expenditure,
                 data = data)
summary(ROA_reduced_model.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure + I(Non_Industry_Income_Expenditure^2) +
##     Operating_Profit_Rate:Non_Industry_Income_Expenditure, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.103453 -0.017806  0.001976  0.020813  0.071801
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                               -6784       3159
## Operating_Profit_Rate                                      6241       3162
## Non_Industry_Income_Expenditure                           23899      10438
## I(Non_Industry_Income_Expenditure^2)                      -5782       2412
## Operating_Profit_Rate:Non_Industry_Income_Expenditure   -20353      10398
##                                                        t value Pr(>|t|)
## (Intercept)                                             -2.147   0.0372 *
## Operating_Profit_Rate                                    1.973   0.0546 .
## Non_Industry_Income_Expenditure                          2.290   0.0268 *
## I(Non_Industry_Income_Expenditure^2)                    -2.398   0.0207 *
## Operating_Profit_Rate:Non_Industry_Income_Expenditure  -1.957   0.0565 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03819 on 45 degrees of freedom
## Multiple R-squared:  0.4896, Adjusted R-squared:  0.4443
## F-statistic: 10.79 on 4 and 45 DF,  p-value: 3.212e-06
```

Still our model terms Operating Profit Rate and Operating Profit Rate × Non Industry Income Expenditure are statistically insignificant (since pval $\approx 0 \leq \alpha = 0.05$). Consequently, we must recalculate the Variance Inflation Factor (VIF) for all model terms to evaluate multicollinearity and refine our model's predictive accuracy.

```r
vif(ROA_reduced_model.lm)
```

```
##                                 Operating_Profit_Rate
##                                              41820.87
##                       Non_Industry_Income_Expenditure
##                                            6788451.61
##                  I(Non_Industry_Income_Expenditure^2)
##                                             136300.16
## Operating_Profit_Rate:Non_Industry_Income_Expenditure
##                                            6864975.62
```

Since the VIF for Operating Profit Rate × Non Industry Income Expenditure = 6864975.62, which is greater than 5 and the highest among all terms, we will remove it from our model and then recalculate the VIF for the remaining variables.

```r
# Removing Operating_Profit_Rate:Non_Industry_Income_Expenditure
ROA_reduced_model.lm <- lm(ROA_Before_Interest_After_Tax ~
                   Operating_Profit_Rate +
                   Non_Industry_Income_Expenditure +
                   I(Non_Industry_Income_Expenditure^2),
                 data = data)
summary(ROA_reduced_model.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure + I(Non_Industry_Income_Expenditure^2),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10310 -0.01457  0.00413  0.02450  0.06991
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -615.8      233.8  -2.635  0.01144 *
## Operating_Profit_Rate                     50.5       16.0   3.156  0.00282 **
## Non_Industry_Income_Expenditure         3674.2     1522.7   2.413  0.01987 *
## I(Non_Industry_Income_Expenditure^2)  -5962.0     2482.9  -2.401  0.02044 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03935 on 46 degrees of freedom
## Multiple R-squared:  0.4462, Adjusted R-squared:  0.4101
## F-statistic: 12.35 on 3 and 46 DF,  p-value: 4.71e-06
```

All terms in our reduced model are now significant since their p-value > 0, indicating a refined model. Our final reduced model is:

ROA Before Interest After Tax $\sim$ Operating Profit Rate + Non Industry Income Expenditure + Non Industry Income Expenditure^2

Now, we will perform a nested F-test to compare the full model with this new reduced model to see how well it predicts ROA before interest and after tax.

```r
# Adding back the Non_Industry_Income_Expenditure
ROA_reduced_model.lm <- lm(ROA_Before_Interest_After_Tax ~
                   Operating_Profit_Rate +
                   Non_Industry_Income_Expenditure +
                   I(Non_Industry_Income_Expenditure^2),
                 data = data)

anova(ROA_reduced_model.lm, ROA_full_model.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate + Non_Industry_Income_Expenditure +
##     I(Non_Industry_Income_Expenditure^2)
```

```
## Model 2: ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate + Non_Industry_Income_Expenditure +
##     I(Operating_Profit_Rate^2) + I(Non_Industry_Income_Expenditure^2) +
##     Operating_Profit_Rate:Non_Industry_Income_Expenditure
##   Res.Df      RSS Df Sum of Sq     F Pr(>F)
## 1     46 0.071225
## 2     44 0.065588  2 0.0056376 1.891  0.163
```

1. LEVEL OF SIGNIFICANCE: $\alpha = 0.05$
2. HYPOTHESES:

- $H_0 : \beta_3 = \beta_5 = 0$
- H_a : At least one $\beta_i \neq 0$ for $i = 3, 5$

3. DECISION RULE: Reject H_0 if p-value $\leq \alpha$
4. TEST STATISTIC: $F = 1.89$
5. P-VALUE: p-value $= 0.16 > \alpha = 0.05$
6. CONCLUSION: Fail to Reject H_0 since p-value $> \alpha$. There is insufficient evidence to conclude that at least one of the coefficients for (Operating Profit Rate)^2, or Operating_Profit_Rate $\times$ Non Industry Income Expenditure are non-zero. The reduced model, which excludes these terms, is appropriate for predicting the Return on Assets before interest and after tax.

So our final reduced model is:

ROA Before Interest After Tax $\sim$ Operating Profit Rate + Non Industry Income Expenditure + Non Industry Income Expenditure ^{2}

Final Model and Assessment:

ANOVA Test on Reduced Model

let's perform an ANOVA test on our reduced model to determine whether it effectively predicts ROA before interest and after tax.

```r
summary(ROA_reduced_model.lm)
```

```
##
## Call:
## lm(formula = ROA_Before_Interest_After_Tax ~ Operating_Profit_Rate +
##     Non_Industry_Income_Expenditure + I(Non_Industry_Income_Expenditure^2),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10310 -0.01457  0.00413  0.02450  0.06991
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -615.8      233.8  -2.635  0.01144 *
## Operating_Profit_Rate                     50.5       16.0   3.156  0.00282 **
```

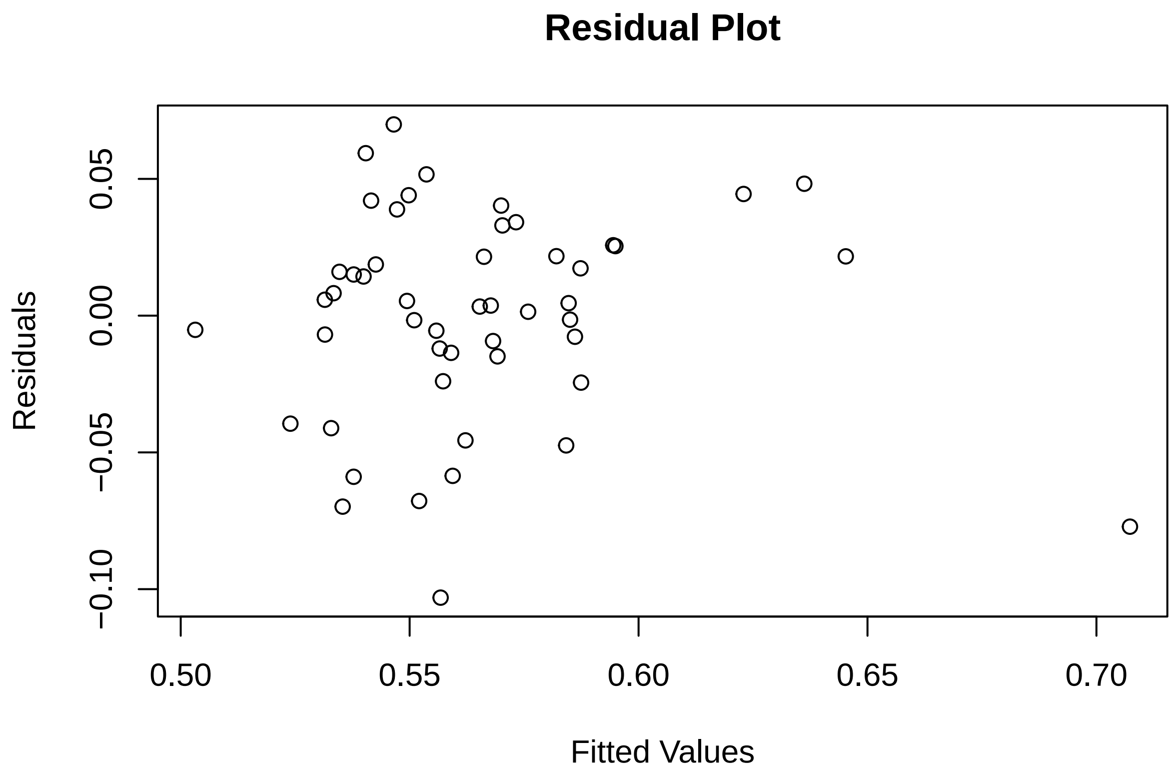
```
## Non_Industry_Income_Expenditure        3674.2     1522.7   2.413  0.01987 *
## I(Non_Industry_Income_Expenditure^2)  -5962.0     2482.9  -2.401  0.02044 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03935 on 46 degrees of freedom
## Multiple R-squared:  0.4462, Adjusted R-squared:  0.4101
## F-statistic: 12.35 on 3 and 46 DF,  p-value: 4.71e-06
```

1. LEVEL OF SIGNIFICANCE: $\alpha = 0.05$
2. HYPOTHESES:
   - $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
   - H_a : At least one $\beta_i \neq 0$ for $i = 1, 2, 3$
3. DECISION RULE: Reject H_0 if p-value $< \alpha$
4. TEST STATISTIC: F = 12.35
5. P-VALUE: p-value $= 4.71 \times 10^{-6}$
6. CONCLUSION: Reject H_0 since p-value $< \alpha$, we conclude that the model terms (Operating Profit Rate , Non Industry Income Expenditure , Non Industry Income Expenditure^2) significantly explains the variation in ROA before interest and after tax.

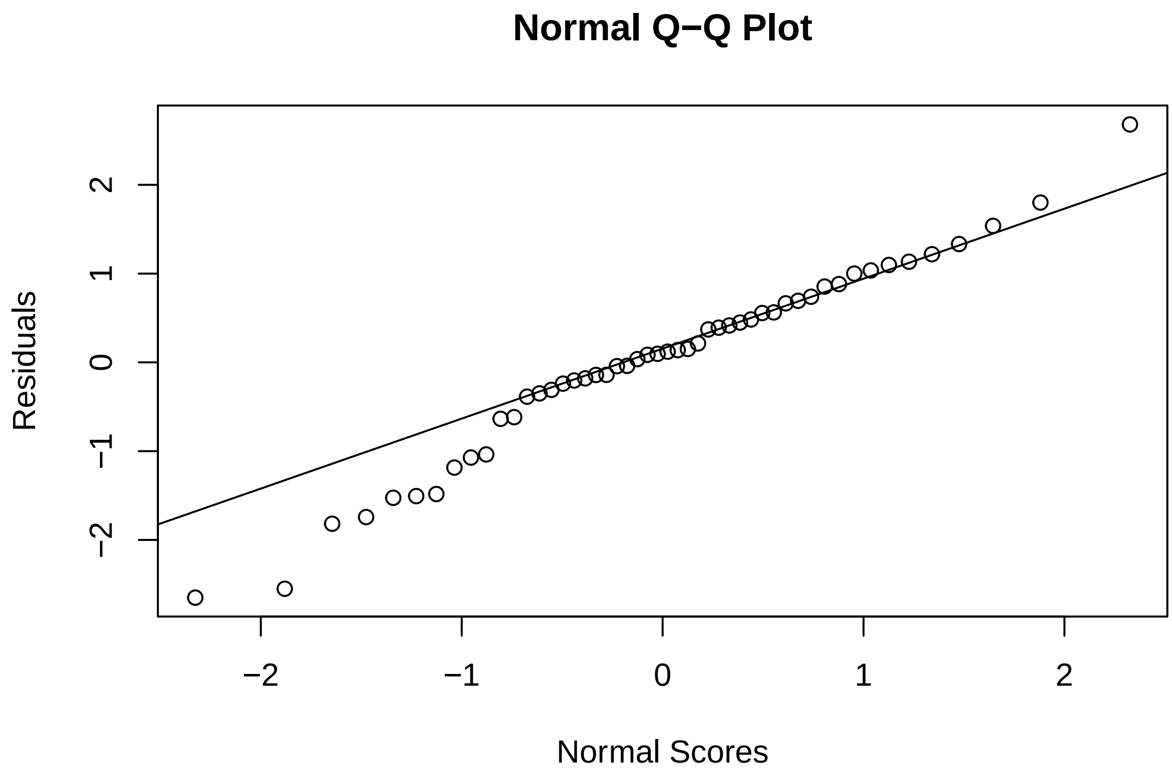
Residual Plot and Normal Quantile Plot

Now lets check whether our constant variance and normality assumptions are met appropriately or not. We will do it using the residual plot and Normality Quantile plots

```r
# Residual Plot
ROA.residual <- resid(ROA_reduced_model.lm)
ROA.fitted <- fitted.values(ROA_reduced_model.lm)
plot(ROA.fitted, ROA.residual, ylab="Residuals", xlab="Fitted Values", main="Residual Plot")
```




```r
# Normal Quantile Plot
ROA.stdresidual <- rstandard(ROA_reduced_model.lm)
qqnorm(ROA.stdresidual, ylab="Residuals", xlab="Normal Scores", main="Normal Q-Q Plot")
qqline(ROA.stdresidual)
```



Based on the residual plot, the points appear to be randomly scattered around the horizontal line, suggesting that the **linearity** and **constant variance** assumptions are reasonably met. There are no clear patterns or funneling effects that would indicate serious violations of these assumptions.

The normal Q-Q plot shows that most points follow the diagonal line closely, particularly in the middle range. There are deviations at the extreme ends, which is not uncommon in real-world data. Overall, the **normality** assumption seems to be adequately met.

While there are no extreme deviations from our assumptions, the slight departures at the tails of the normal Q-Q plot suggest we should use the model with some caution, particularly when making predictions for extreme values of ROA. However, for most practical purposes within the observed range of our data, the model appears to satisfy the key assumptions of linear regression.

The **randomness** assumption is satisfied because we randomly selected 50 companies from a total of 6,819 listed on the Taiwanese Stock Exchange between 1999 and 2009, ensuring that every company had an equal chance of being included. The **independence** assumption is also met, as each company operates independently with its own unique assets, expenditures, and liabilities, meaning that the performance of one company does not affect another. Hence, these assumptions support the validity of our multiple linear regression analysis.

Conclusion:

Our analysis of Taiwanese companies listed on the Taiwan Stock Exchange from 1999 to 2009 demonstrates that Operating Profit Rate is a significant predictor of Return on Assets (ROA), meaning that as a company's operating efficiency improves, so does its profitability relative to its assets. Additionally, Non-industry Income and Expenditure initially boosts ROA but starts to negatively impact it at higher levels. Overall, this model highlights the importance of managing core business operations effectively while being cautious about the effects of non-operating income. Overall the final model explains approximately 44.62% of the variation in ROA, indicating good predictive power as compared to indivual contributing parameters.

Hence our best estimate of the relationship between ROA (Y), Operating Profit Rate (X_1), and Non-industry income and expenditure revenue (X_2) is:

$$\begin{aligned}\widehat{\text{ROA before interest and after tax}} = &- 615.8 + 50.5 \cdot \text{Operating Profit Rate} \\ &+ 3674.2 \cdot \text{Non-industry Income and Expenditure} \\ &- 5962 \cdot \text{Non-industry Income and Expenditure}^2\end{aligned}$$

This model provides valuable insights for investors and managers in assessing and predicting the financial performance of Taiwanese companies, it suggests that a company's main business operations and some non-core activities can help boost Return on Assets before interset and after taxes (i.e profits on investments), but there's a sweet spot for non-core expenditures - too much can be harmful, emphasizing the critical importance of efficient core operations in driving overall profitability, while also considering the significant effect of non-core financial activities.