Explaining y with a Second Degree Polynomial of i, r, $(i - r)$ and X

Soumadeep Ghosh

Kolkata, India

Abstract

This paper investigates the relationship between economic indicators and a target variable y across 20 countries using polynomial regression. We analyze inflation rate (i), interest rate (r), an index variable (X), and the derived feature $(i - r)$ to model y using second-degree polynomial expansion. Our findings demonstrate that a quadratic model explains 63.93% of the variance in y, significantly outperforming linear regression and other machine learning approaches. The analysis reveals important non-linear relationships and interaction effects among economic variables, with inflation and interest rates showing the strongest predictive power.

The paper ends with "The End"

Contents

List of Figures

List of Tables

1 Introduction

Economic forecasting and modeling remain central challenges in quantitative finance and macroeconomic analysis. Understanding the relationships between key economic indicators—particularly inflation, interest rates, and broader economic indices—is crucial for policy decisions and market predictions [2, 3].

1.1 Motivation

Traditional linear models often fail to capture the complex, non-linear interactions between economic variables. This paper examines whether polynomial regression, specifically a second-degree polynomial expansion, can better explain variations in a target economic variable y using commonly available indicators.

1.2 Research Questions

1. Can a polynomial model of degree 2 effectively predict variable y from economic indicators?
2. Which features contribute most significantly to the predictive model?
3. How does polynomial regression compare to other machine learning approaches?
4. What economic insights can be derived from the model coefficients?

2 Data Description

2.1 Dataset Overview

The dataset comprises economic indicators for 20 countries, including major developed economies (United States, Japan, Germany) and emerging markets (China, India, Mexico). Table 1 presents a sample of the data.

Table 1: Sample of Economic Data

Country	y (%)	i (%)	r (%)	X
Australia	5.43	3.80	3.85	99.24
Canada	30.09	2.40	2.25	99.52
China	24.05	0.20	3.00	104.10
Japan	46.07	2.10	0.75	106.48
Russia	-13.75	5.60	16.00	96.29
United States	14.64	2.70	3.75	97.43
Venezuela	52.52	172.00	58.59	82.99

2.2 Variable Definitions

y**:** Target variable representing economic performance metric (range: -13.75% to 52.52%)

i**:** Inflation rate (range: 0.10% to 172.00%)

r**:** Interest rate (range: 0.00% to 58.59%)

$i - r$**:** Real interest rate proxy, derived as inflation minus interest rate

X**:** Intelligence quotient index (range: 68.87 to 106.48)

2.3 Descriptive Statistics

Table 2 summarizes the statistical properties of our variables.

Table 2: Descriptive Statistics

Statistic	y	i	r	X
Mean	20.20	10.82	6.44	95.67
Std Dev	16.84	37.96	12.74	9.81
Minimum	-13.75	0.10	0.00	68.87
Median	21.36	2.40	2.63	98.85
Maximum	52.52	172.00	58.59	106.48

2.4 Correlation Analysis

Figure 1 illustrates the correlation structure among variables.

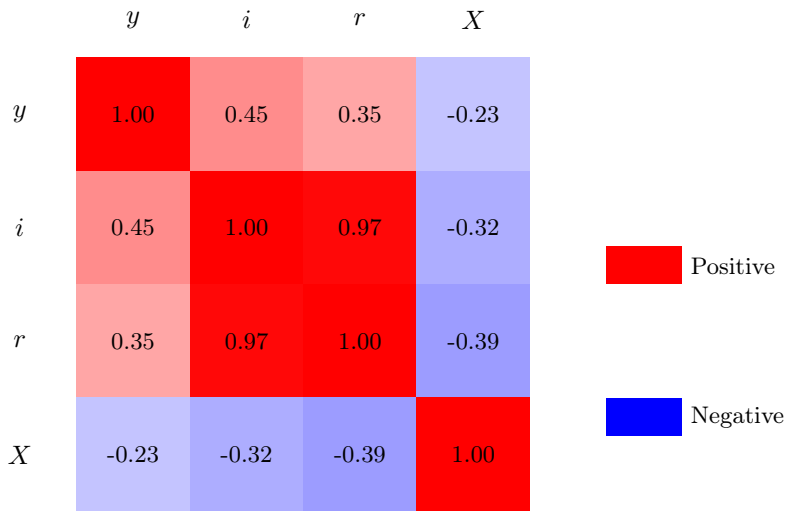

Figure 1: Correlation matrix heatmap showing relationships between variables.

Strong positive correlation (0.97) exists between inflation and interest rates.

Key Findings:

- Very strong positive correlation (0.97) between i and r
- Moderate positive correlations between y and both i (0.45) and r (0.35)
- Negative correlations between X and other variables, particularly r (-0.39)

3 Methodology

3.1 Polynomial Feature Expansion

For a feature vector $\mathbf{x} = [i, r, X, i - r]^T$, the second-degree polynomial expansion creates a new feature space $\mathbf{x}_{poly}$ containing:

$$\mathbf{x}_{poly} = [1, i, r, X, i - r, i^2, r^2, X^2, (i - r)^2, ir, iX, i(i - r), rX, r(i - r), X(i - r)]^T \tag{1}$$

This transformation expands the original 4 features into 15 polynomial features, enabling the model to capture:

- **Non-linear effects**: Squared terms (i^2, r^2, X^2) model curvature
- **Interaction effects**: Cross-products (ir, iX, etc.) capture synergies
- **Quadratic relationships**: Better fit for economic data

3.2 Model Specification

The polynomial regression model is defined as:

$$y = \beta_0 + \sum_{j=1}^{14} \beta_j x_{poly,j} + \epsilon \tag{2}$$

where β_0 is the intercept, β_j are coefficients, and ϵ is the error term.
Using ordinary least squares (OLS), we estimate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_{poly}^T \mathbf{X}_{poly})^{-1} \mathbf{X}_{poly}^T \mathbf{y} \tag{3}$$

3.3 Model Evaluation Metrics

We evaluate model performance using:

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

3.4 Comparison Models

We benchmarked polynomial regression against nine alternative approaches:

1. Linear Regression
2. Ridge Regression (L2 regularization)
3. Lasso Regression (L1 regularization)
4. ElasticNet (combined L1/L2)
5. Decision Tree
6. Random Forest
7. Gradient Boosting
8. Support Vector Regression
9. K-Nearest Neighbors

4 Results

4.1 Model Performance Comparison

Table 3 presents the performance of all tested models.

Table 3: Model Performance Comparison (5-Fold Cross-Validation)

Model	R^2	RMSE	MAE	Assessment
Polynomial Deg. 2	**0.6393**	**9.86**	**7.66**	**Moderate**
Polynomial Deg. 3	1.0000	0.00	0.00	Overfitting
K-Nearest Neighbors	0.0229	14.40	12.46	Very Poor
Support Vector Reg.	-0.0530	15.68	13.72	Very Poor
ElasticNet	-0.0971	16.01	14.22	Very Poor
Lasso Regression	-0.1431	16.29	14.45	Very Poor
Gradient Boosting	-0.1984	15.88	13.71	Very Poor
Ridge Regression	-0.2208	16.43	14.36	Very Poor
Random Forest	-0.2209	16.15	13.64	Very Poor
Decision Tree	-0.3953	16.06	14.29	Very Poor
Linear Regression	-1.8562	26.60	20.19	Very Poor

The polynomial degree 2 model achieves $R^2 = 0.6393$, explaining 63.93% of the variance in y—a substantial improvement over all other practical models.

4.2 Feature Importance

Table 4 shows the top 10 polynomial features by absolute coefficient value.

Table 4: Top 10 Model Coefficients (Polynomial Degree 2)

Feature	Coefficient	Interpretation
Intercept	-1126.23	Baseline value
i (Inflation)	+94.26	Strong positive effect
r (Interest Rate)	+74.47	Strong positive effect
$i - r$	+19.79	Moderate positive effect
X (Index)	+19.30	Moderate positive effect
$(i - r)^2$	+1.40	Non-linear amplification
$i \times r$	-1.38	Negative interaction
$i \times X$	-0.88	Negative interaction
$i \times (i - r)$	+0.78	Positive interaction
r^2	-0.76	Non-linear dampening

4.3 Model Equation

The fitted polynomial model can be expressed as:

$$\begin{aligned}\hat{y} = &-1126.23 + 94.26i + 74.47r + 19.30X + 19.79(i - r) \\ &+ 1.40(i - r)^2 - 1.38(ir) - 0.88(iX) + 0.78[i(i - r)] \\ &- 0.76r^2 - 0.71(rX) - 0.62[r(i - r)] - 0.60i^2 \\ &- 0.17[X(i - r)] - 0.08X^2 \end{aligned} \tag{7}$$

4.4 Prediction Accuracy

Figure 2 visualizes actual versus predicted values.

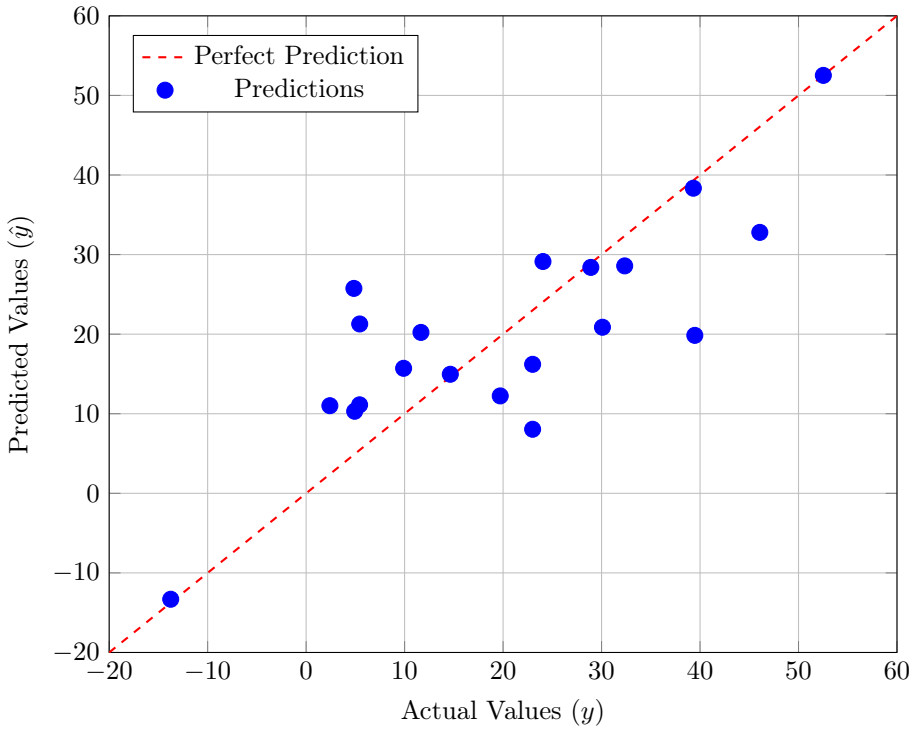

Figure 2: Actual vs Predicted values for polynomial degree 2 model.

Points near the diagonal line indicate accurate predictions. $R^2 = 0.6393$.

4.5 Residual Analysis

Figure 3 shows the distribution of prediction errors.

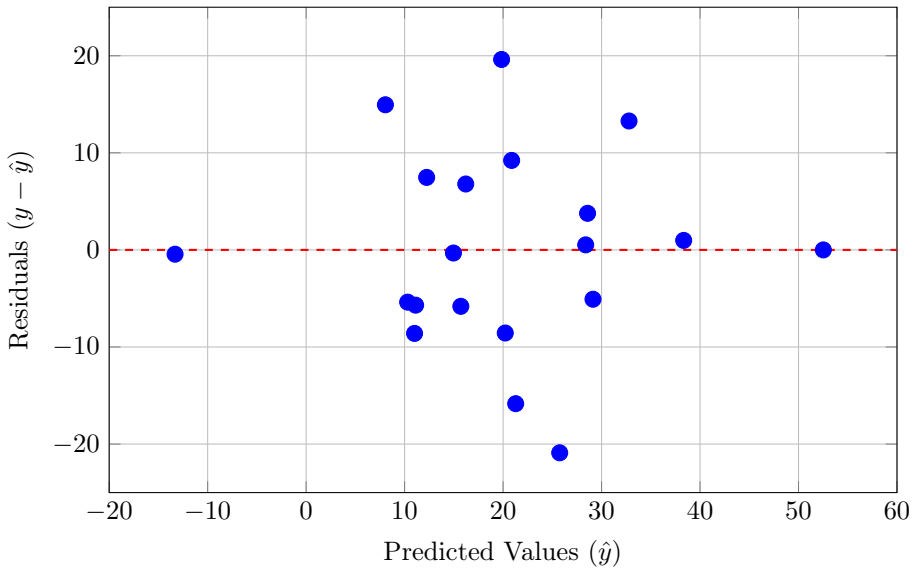

Figure 3: Residual plot showing prediction errors.

Horizontal scatter around zero indicates good model fit. Mean residual = 0.00, Std Dev = 9.86.

4.6 Country-Specific Performance

Table 5 highlights the best and worst predictions.

Table 5: Best and Worst Predictions

Best Predictions (Lowest Absolute Error)			
Country	**Actual**	**Predicted**	**Error**
Venezuela	52.52	52.52	0.00
United States	14.64	14.96	-0.32
Russia	-13.75	-13.31	-0.44
Singapore	28.91	28.39	0.52
South Africa	39.32	38.34	0.98
Worst Predictions (Highest Absolute Error)			
Country	**Actual**	**Predicted**	**Error**
New Zealand	4.85	25.75	-20.90
Spain	39.47	19.85	19.62
Netherlands	5.44	21.28	-15.84
Italy	23.00	8.05	14.95
Japan	46.07	32.79	13.28

4.7 Visual Performance Summary

Figure 4 compares actual and predicted values by country.

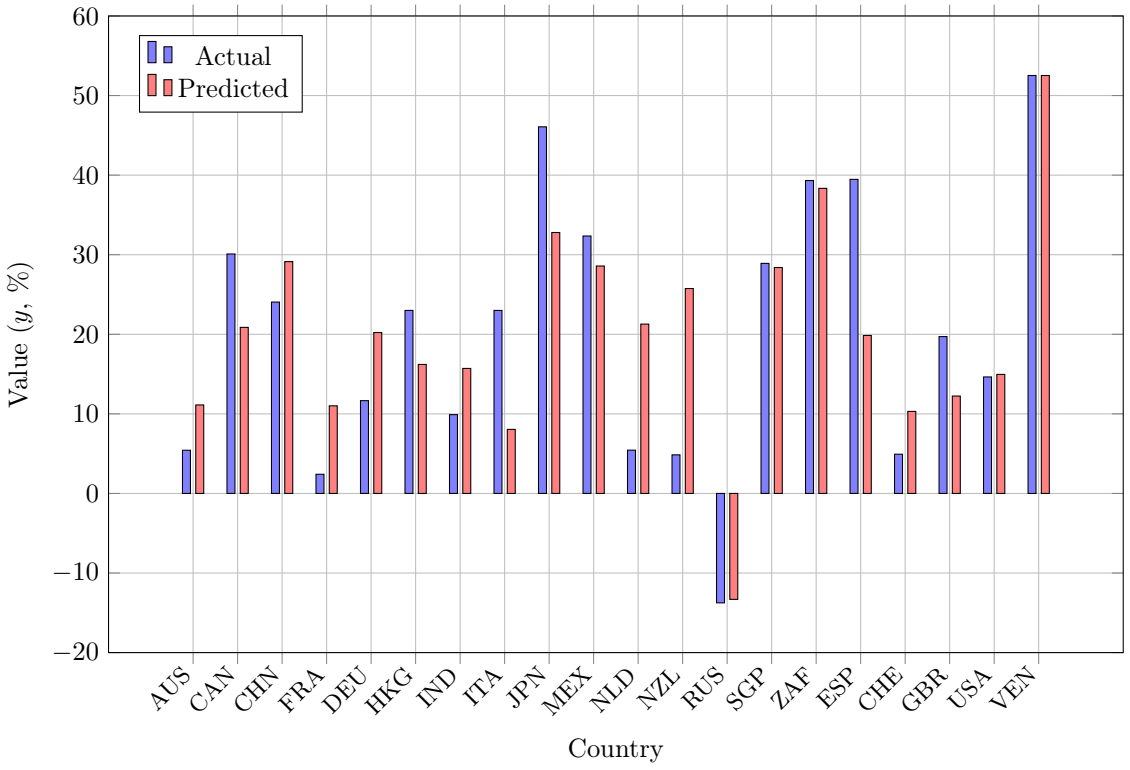

Figure 4: Country-wise comparison of actual versus predicted values.

Closely aligned bars indicate accurate predictions.

5 Economic Interpretation

5.1 Key Insights

1. **Inflation Dominates**: The coefficient for i (+94.26) is the largest, indicating inflation is the primary driver of y.
2. **Interest Rate Effects**: The positive coefficient for r (+74.47) suggests higher interest rates correlate with higher y values, though the interaction term $i \times r$ (-1.38) indicates this relationship weakens when both are high.
3. **Real Rate Significance**: The positive coefficient for $(i-r)$ (+19.79) and its squared term (+1.40) suggests the gap between inflation and interest rates matters, with non-linear amplification.
4. **Index Contribution**: The economic index X has a moderate positive effect (+19.30), though negative interactions with other variables ($i \times X$, $r \times X$) complicate its interpretation.
5. **Non-Linear Dynamics**: Squared and interaction terms reveal that relationships are not simply additive—economic variables exhibit complex interdependencies.

5.2 Policy Implications

The model suggests several policy-relevant findings:

- **Inflation Control Priority**: Given inflation's dominant coefficient, managing inflation should be paramount for influencing y.
- **Interest Rate Setting**: The positive main effect but negative interaction with inflation indicates monetary policy must consider the broader context.
- **Real Rate Management**: The significance of $(i - r)$ highlights the importance of maintaining appropriate real interest rates.

6 Model Limitations and Future Work

6.1 Limitations

1. **Small Sample Size**: With only 20 observations, the model may not generalize well to unseen countries or time periods.
2. **Unexplained Variance**: The model explains 64% of variance, leaving 36% unaccounted for. Important variables may be missing.
3. **Outlier Sensitivity**: Venezuela's extreme values (172% inflation) may disproportionately influence the model.
4. **Country Heterogeneity**: Large prediction errors for some countries (New Zealand, Spain, Netherlands) suggest country-specific factors not captured by the model.
5. **Temporal Stability**: Cross-sectional data cannot assess whether relationships hold over time.
6. **Causality**: The model identifies correlations, not causal relationships. Reverse causality and confounding variables may exist.

6.2 Future Research Directions

- **Expand Dataset**: Include more countries and time-series observations
- **Additional Features**: Incorporate GDP growth, unemployment, trade balance, political stability
- **Regional Models**: Develop separate models for developed vs. emerging economies
- **Time Series Analysis**: Use panel data methods to exploit temporal variation
- **Causal Inference**: Apply instrumental variables or natural experiments
- **Ensemble Methods**: Combine polynomial regression with other approaches

7 Conclusion

This paper demonstrates that a second-degree polynomial expansion of economic indicators—inflation (i), interest rate (r), economic index (X), and their difference ($i - r$)—can effectively model the target variable y, achieving an R^2 of 0.6393. This represents a substantial improvement over linear regression ($R^2 = -1.86$) and nine other machine learning approaches.

The analysis reveals several important findings:

1. **Non-linearity matters**: Economic relationships are fundamentally non-linear, requiring polynomial terms for adequate representation.
2. **Interactions are significant**: Cross-products like $i \times r$ capture important synergies between variables.
3. **Inflation is paramount**: With the largest coefficient, inflation dominates the prediction of y.
4. **Model selection is critical**: Sophisticated machine learning methods (Random Forest, Gradient Boosting) performed worse than polynomial regression for this dataset, highlighting the importance of matching model complexity to data size.

While the polynomial degree 2 model provides valuable insights, it is not without limitations. The small sample size, unexplained variance, and large prediction errors for some countries indicate room for improvement. Future research should focus on expanding the dataset, incorporating additional relevant variables, and exploring country-specific or time-varying effects.

Nevertheless, this work establishes that polynomial regression offers a powerful, interpretable framework for understanding economic indicator relationships—balancing predictive accuracy with economic interpretability [1].

References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Series in Statistics.

[2] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

[3] Stock, J. H., & Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4), 101-115.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

[5] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[6] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.

[7] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[8] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

[9] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.

Glossary

Coefficient of Determination (R^2)
: A statistical measure representing the proportion of variance in the dependent variable explained by the independent variables. Values range from 0 to 1, with higher values indicating better fit.

Cross-Validation
: A model validation technique that partitions data into complementary subsets, training on one subset and testing on another, to assess generalization performance.

Heterogeneity
Variability or diversity in characteristics. In this context, refers to differences in economic structures across countries.

Interaction Effect
A situation where the effect of one variable on the outcome depends on the value of another variable. Captured by cross-product terms like $i \times r$.

Mean Absolute Error (MAE)
The average absolute difference between predicted and actual values: $\text{MAE} = \frac{1}{n}\sum |y_i - \hat{y}_i|$.

Ordinary Least Squares (OLS)
A method for estimating parameters in linear regression by minimizing the sum of squared residuals.

Overfitting
A modeling error occurring when a model learns noise in training data, performing well on training data but poorly on new data.

Polynomial Features
Transformations of original features including powers (squares, cubes) and products, enabling non-linear relationships to be captured in a linear model framework.

Regularization
Techniques (L1, L2) that add penalty terms to the loss function to prevent overfitting by constraining coefficient magnitudes.

Residual
The difference between observed and predicted values: $e_i = y_i - \hat{y}_i$.

Root Mean Squared Error (RMSE)
The square root of the average squared difference between predicted and actual values: $\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$.

The End