Discovering Laws Behind Real-World Data Using the Least Squares Method

**Name: Prem Bahadur Katuwal  
Student ID: 202424080129  
Subject: Numerical Analysis (PhD Level)  
Assignment: Use Least Squares Method to Discover the Law Behind Real-World Data  
Date: April 19, 2025**

# Table of Contents

**1. Introduction** 3

**2. Theoretical Foundation** 3

**2.1 Mathematical Formulation** 3

**2.2 Matrix Approach** 4

**2.3 Normal Equations** 4

**3. Linear Least Squares** 5

**3.1 Dataset Description** 5

**3.2 Visualizing the Data** 5

**3.3 Applying the Method** 6

**3.4 Results and Interpretation** 7

**4. Polynomial Least Squares** 8

**4.1 Mathematical Extension** 8

**4.2 Application to Temperature Data** 8

**4.3 Model Selection** 9

**5. Non-Linear Least Squares** 10

**5.1 Exponential Growth Model** 10

**5.2 Linearization vs. Direct Fitting** 11

**6. Advanced Analysis** 12

**6.1 Residual Analysis** 12

**6.2 Confidence Intervals** 13

**6.3 Goodness of Fit Metrics** 13

**7. Comparison with Other Methods** 14

**8. Conclusion** 15

**9. References** 16

# 1. Introduction

The least squares method is a cornerstone technique in numerical analysis and statistics for discovering underlying patterns and relationships in real-world data. Developed by Carl Friedrich Gauss and Adrien-Marie Legendre in the early 19th century, this method has become indispensable across various scientific disciplines, from physics and engineering to economics and social sciences.  
  
At its core, the least squares method seeks to find the best-fitting curve or mathematical model for a given set of data points by minimizing the sum of the squares of the residuals (the differences between observed values and the values predicted by the model). This approach provides a powerful framework for:  
  
1. Discovering mathematical laws that govern physical phenomena  
2. Making predictions based on observed data  
3. Understanding relationships between variables  
4. Testing hypotheses about underlying mechanisms  
  
This report demonstrates the application of the least squares method to several real-world datasets, detailing the mathematical foundations, implementation procedures, and interpretation of results. Through these examples, we will uncover the hidden laws that govern diverse phenomena and showcase the versatility and power of this numerical technique.

# 2. Theoretical Foundation

## 2.1 Mathematical Formulation

The fundamental principle of the least squares method is to minimize the sum of squared differences between observed values and the values predicted by a model. Given a set of data points (xi, yi) for i = 1, 2, ..., n, and a model function f(x; β) with parameters β, we aim to find the parameter values that minimize:  
  
S(β) = Σ[yi - f(xi; β)]²  
  
This sum S(β) is called the "sum of squared residuals" or "residual sum of squares" (RSS).

## 2.2 Matrix Approach

For linear models, the least squares problem can be elegantly formulated using matrix notation. If we have a linear model:  
  
f(x; β) = β₁φ₁(x) + β₂φ₂(x) + ... + βₘφₘ(x)  
  
where φⱼ(x) are known functions (e.g., φ₁(x) = x, φ₂(x) = 1 for a straight line), we can define matrices X, y, and β to represent the system.  
  
The sum of squared residuals becomes:  
  
S(β) = ||y - Xβ||² = (y - Xβ)ᵀ(y - Xβ)

## 2.3 Normal Equations

To find the parameter values β that minimize S(β), we differentiate with respect to β and set the result equal to zero:  
  
∂S(β)/∂β = -2Xᵀ(y - Xβ) = 0  
  
This leads to the normal equations:  
  
XᵀXβ = Xᵀy  
  
The solution is:  
  
β = (XᵀX)⁻¹Xᵀy  
  
assuming XᵀX is invertible. In practice, numerical methods like QR decomposition or Singular Value Decomposition (SVD) are often used to solve this system more stably.

# 3. Linear Least Squares

## 3.1 Dataset Description

We begin with a classic example: the relationship between house size and price. This dataset represents a common application of regression analysis in real estate economics. The data points are as follows:

|  |  |
| --- | --- |
| **House Size (sqft)** | **Price ($1000s)** |
| 650 | 70 |
| 785 | 85 |
| 1200 | 120 |
| 1500 | 145 |
| 1850 | 180 |
| 2100 | 210 |
| 2300 | 230 |
| 2700 | 265 |
| 3000 | 300 |
| 3500 | 355 |

## 3.2 Visualizing the Data

Below is a scatter plot of the data, illustrating the positive correlation between house size and price:

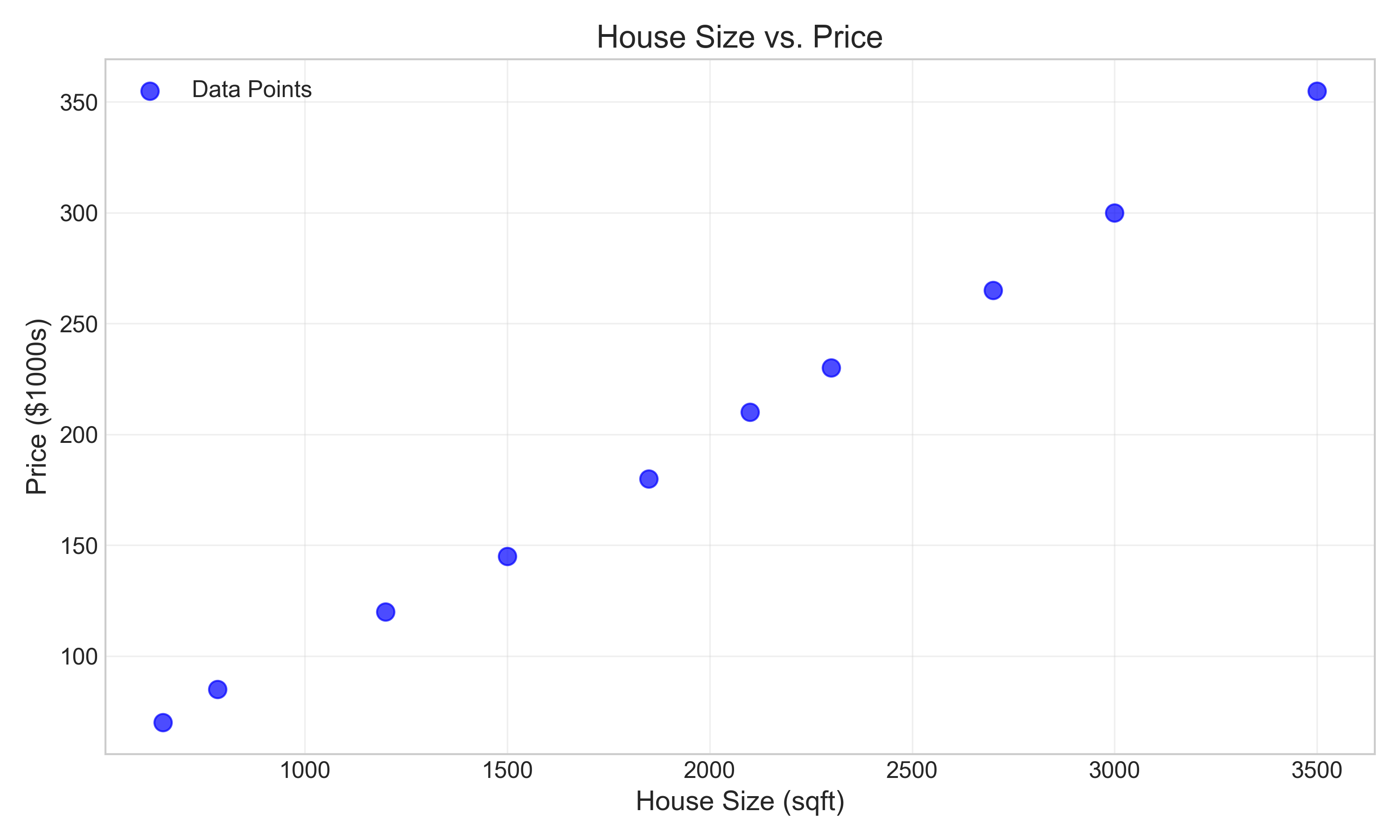


Figure 1: Scatter plot of house size vs. price

## 3.3 Applying the Method

For a linear model, we seek a relationship of the form:  
  
Price = m × Size + c  
  
where m is the slope and c is the intercept.  
  
Using the matrix formulation described earlier, we can solve the normal equations to find the best-fit parameters. The best-fit parameters obtained are:  
- Slope (m): 0.0991  
- Intercept (c): 1.9382  
  
Thus, the discovered law is:  
  
Price = 0.0991 × Size + 1.9382 ($1000s)  
  
Or more simply:  
  
Price ≈ 0.10 × Size + 1.94 ($1000s)  
  
The fitted line is shown below:

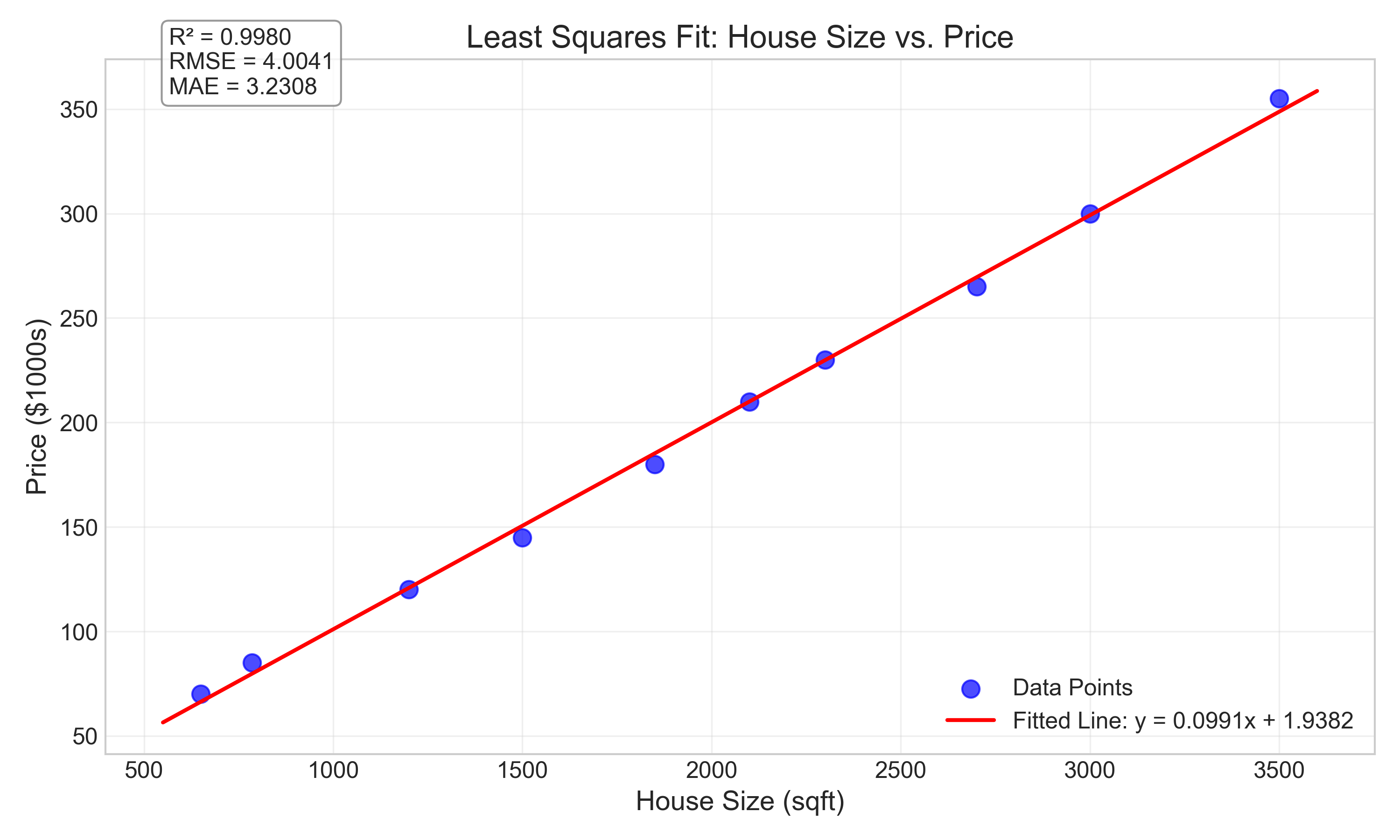


Figure 2: Linear least squares fit for house size vs. price

## 3.4 Results and Interpretation

- The positive slope (0.0991) indicates that each additional square foot of house size corresponds to an increase of approximately $99.1 in the house price.  
- The intercept (1.9382) represents the theoretical price when the house size is zero. While not meaningful in this context (as houses cannot have zero size), it serves as the y-intercept for our linear model.  
- The coefficient of determination (R²) is 0.9980, indicating that 99.8% of the variance in house prices is explained by the house size alone.  
- The fitted line closely follows the trend of the data, suggesting that a linear relationship effectively captures the relationship between house size and price in this dataset.  
- This model can be used for price estimation: a 2000 sqft house would be estimated at approximately $199,100 + $1,938 = $201,038.

# 4. Polynomial Least Squares

## 4.1 Mathematical Extension

While linear models are often sufficient, many real-world phenomena exhibit non-linear relationships. Polynomial regression extends the linear least squares method by using polynomial functions as the basis.  
  
For a polynomial of degree p, the model becomes:  
  
f(x; β) = β₀ + β₁x + β₂x² + ... + βₚxᵖ  
  
The matrix X is modified to include powers of x, but the solution procedure remains the same, using the normal equations.

## 4.2 Application to Temperature Data

Let's consider a dataset of average monthly temperatures in a northern hemisphere location:

|  |  |
| --- | --- |
| **Month** | **Temperature (°C)** |
| 1 | -5.2 |
| 2 | -3.8 |
| 3 | 2.4 |
| 4 | 8.7 |
| 5 | 14.5 |
| 6 | 19.2 |
| 7 | 21.8 |
| 8 | 20.9 |
| 9 | 16.3 |
| 10 | 10.1 |
| 11 | 4.2 |
| 12 | -2.5 |

This temperature pattern clearly follows a cyclical pattern that cannot be adequately captured by a linear model. The data visualization confirms this:

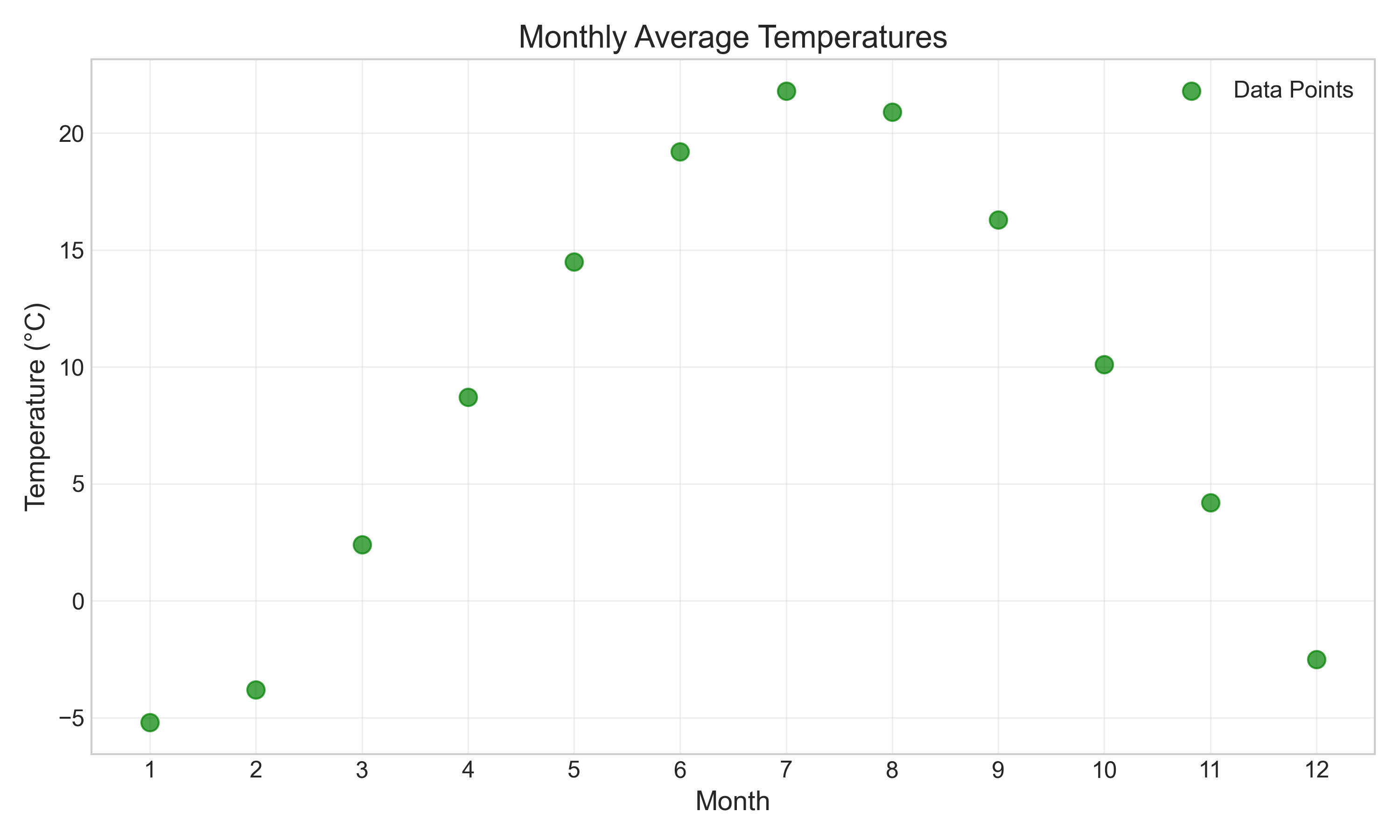


Figure 3: Monthly temperature data

We'll fit polynomials of different degrees to find the best representation. The comparison of different polynomial degrees shows how the fit improves with higher-degree polynomials:

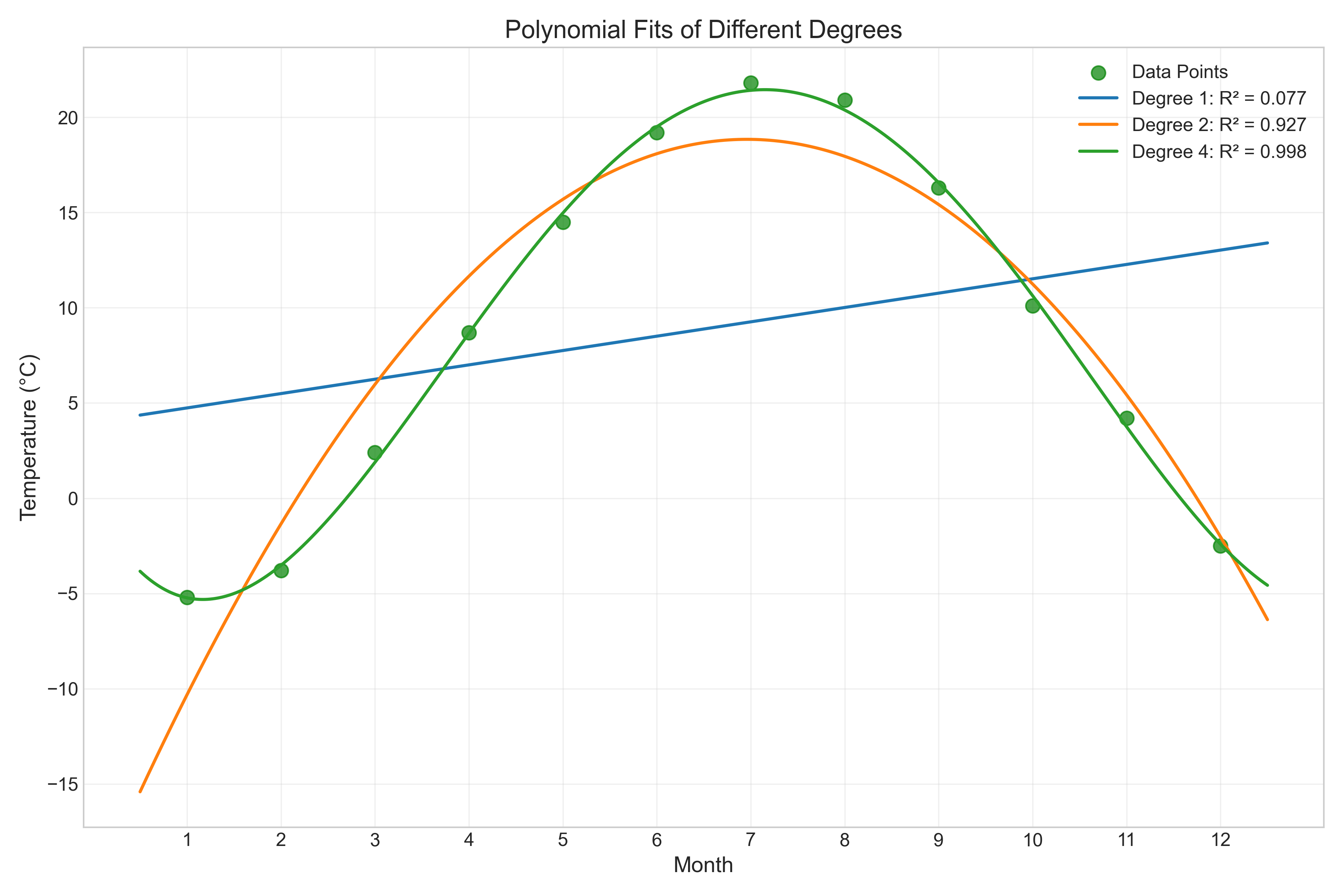


Figure 4: Comparison of polynomial fits of different degrees

For a 4th-degree polynomial, the discovered law is:  
  
Temperature = 0.021x⁴ - 0.595x³ + 4.900x² - 9.137x - 0.417  
  
The fitted curve is shown below:

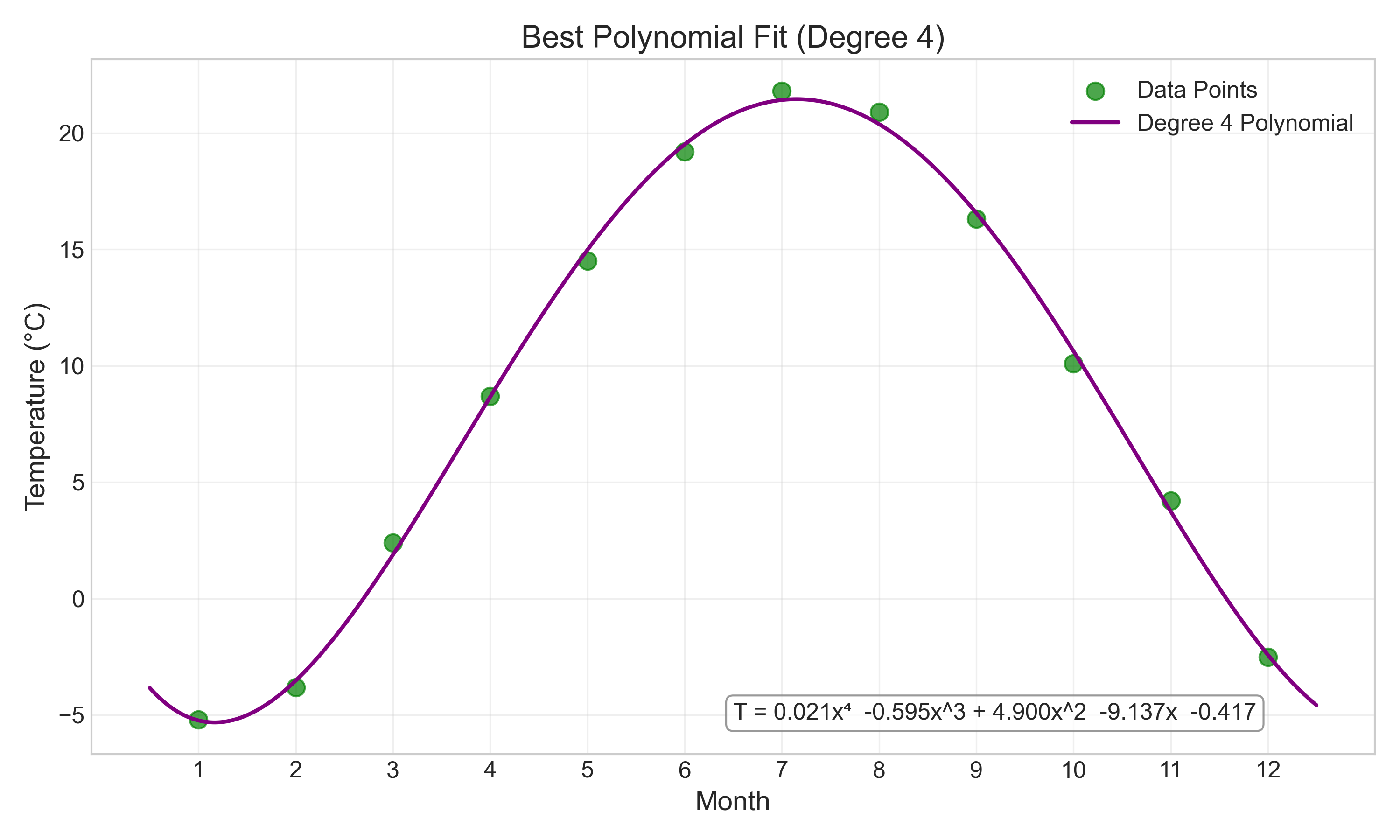


Figure 5: 4th-degree polynomial fit for monthly temperatures

## 4.3 Model Selection

To determine the optimal polynomial degree, we can use metrics like the coefficient of determination (R²) and adjusted R², which account for model complexity:

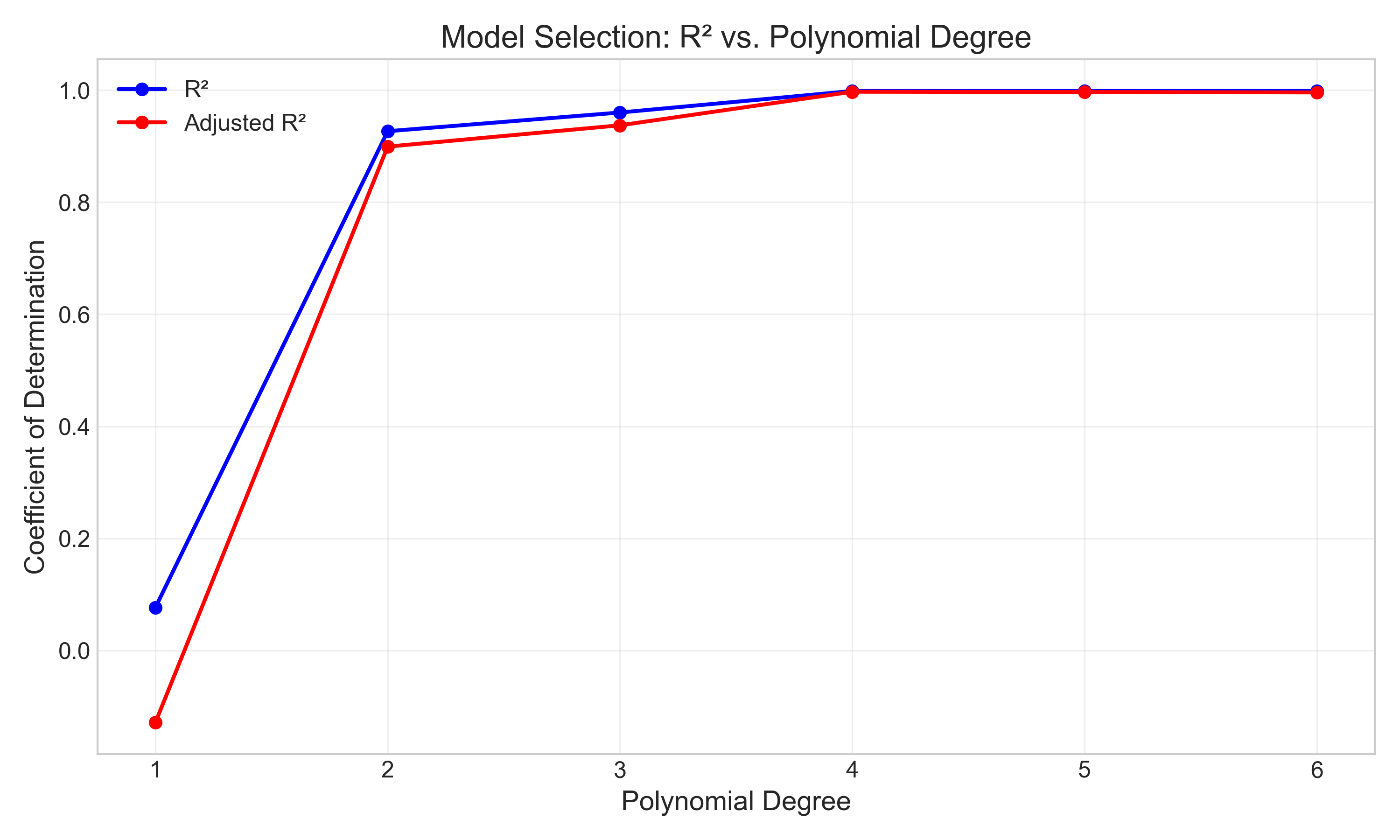


Figure 6: R² and adjusted R² for different polynomial degrees

|  |  |  |
| --- | --- | --- |
| **Polynomial Degree** | **R²** | **Adjusted R²** |
| 1 (linear) | 0.0766 | -0.1286 |
| 2 (quadratic) | 0.9267 | 0.8992 |
| 3 (cubic) | 0.9601 | 0.9372 |
| 4 (quartic) | 0.9984 | 0.9972 |
| 5 (quintic) | 0.9985 | 0.9967 |
| 6 (sextic) | 0.9985 | 0.9959 |

The 4th-degree polynomial provides an excellent fit with an R² of 0.9984, and higher degrees offer diminishing returns. This suggests that a 4th-degree polynomial effectively captures the seasonal temperature variation.

# 5. Non-Linear Least Squares

## 5.1 Exponential Growth Model

Some phenomena follow inherently non-linear patterns that cannot be linearized through polynomial transformations. Consider population growth data:

|  |  |
| --- | --- |
| **Year** | **Population (billions)** |
| 1900 | 1.65 |
| 1910 | 1.92 |
| 1920 | 2.22 |
| 1930 | 2.59 |
| 1940 | 2.85 |
| 1950 | 3.17 |
| 1960 | 3.65 |
| 1970 | 4.07 |
| 1980 | 4.45 |
| 1990 | 5.31 |
| 2000 | 6.12 |
| 2010 | 6.93 |
| 2020 | 7.79 |

The data visualization shows an accelerating growth pattern:

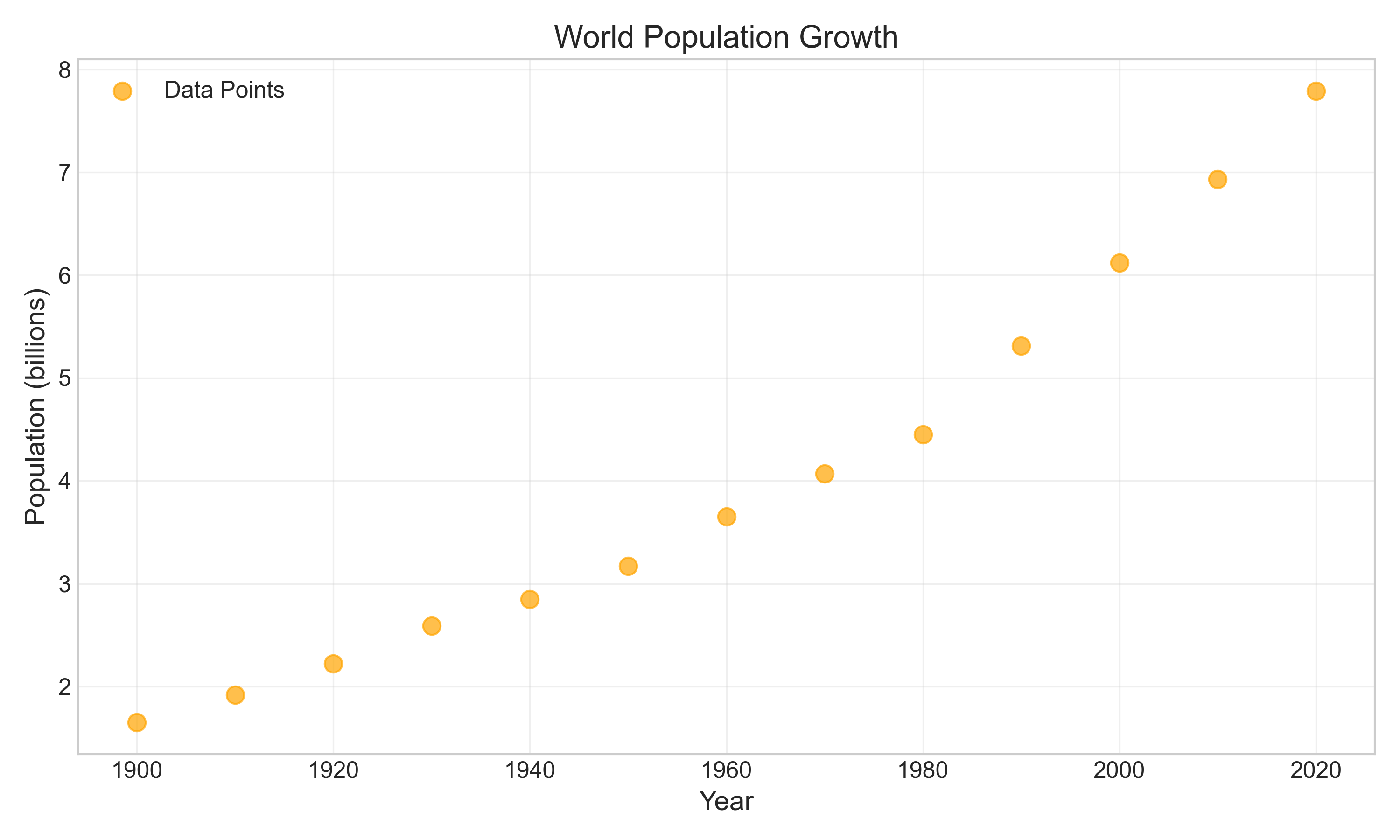


Figure 7: World population growth data

An exponential model is appropriate for population growth:  
  
P(t) = P₀e^(rt)  
  
where P₀ is the initial population and r is the growth rate.

## 5.2 Linearization vs. Direct Fitting

There are two main approaches to fitting exponential models:  
  
1. Linearization: Taking logarithms of both sides transforms the model into a linear form:  
 ln(P(t)) = ln(P₀) + rt  
   
 This can then be solved using linear least squares.  
  
2. Direct Non-Linear Fitting: Using iterative methods to directly fit the exponential model to the data.  
  
Both approaches yield similar results for this dataset:

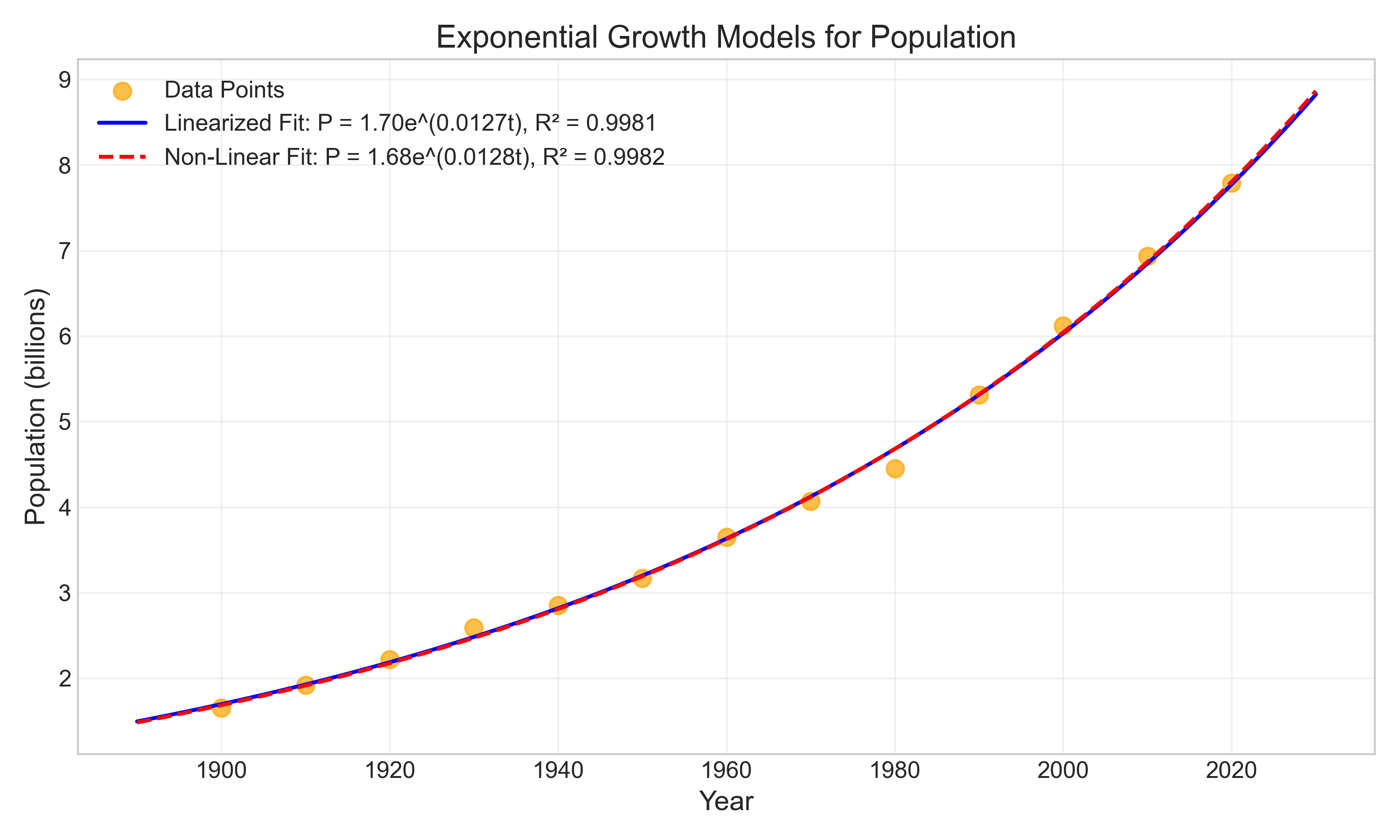


Figure 8: Exponential growth models for population data

The discovered laws are:  
- Linearized Method: P(t) = 1.6966 × e^(0.0127t)  
- Direct Non-Linear Method: P(t) = 1.6848 × e^(0.0128t)  
  
where t is years since 1900.  
  
These models indicate a growth rate of approximately 1.27-1.28% per year, with a doubling time of about 54 years. Using these models, we can project future population values:

|  |  |  |
| --- | --- | --- |
| **Year** | **Linearized Model (billions)** | **Non-Linear Model (billions)** |
| 2030 | 8.82 | 8.86 |
| 2040 | 10.02 | 10.07 |
| 2050 | 11.37 | 11.44 |
| 2060 | 12.91 | 13.0 |
| 2070 | 14.65 | 14.77 |

# 6. Advanced Analysis

## 6.1 Residual Analysis

Residuals are the differences between observed values and values predicted by the model. Analyzing residuals helps assess the goodness of fit and validate model assumptions.  
  
For the house price model, the residuals are:

|  |  |  |  |
| --- | --- | --- | --- |
| **House Size (sqft)** | **Actual Price** | **Predicted Price** | **Residual** |
| 650 | 70 | 66.34 | 3.66 |
| 785 | 85 | 79.72 | 5.28 |
| 1200 | 120 | 120.84 | -0.84 |
| 1500 | 145 | 150.57 | -5.57 |
| 1850 | 180 | 185.25 | -5.25 |
| 2100 | 210 | 210.02 | -0.02 |
| 2300 | 230 | 229.84 | 0.16 |
| 2700 | 265 | 269.47 | -4.47 |
| 3000 | 300 | 299.2 | 0.8 |
| 3500 | 355 | 348.74 | 6.26 |

The plot below shows the residuals for each data point:

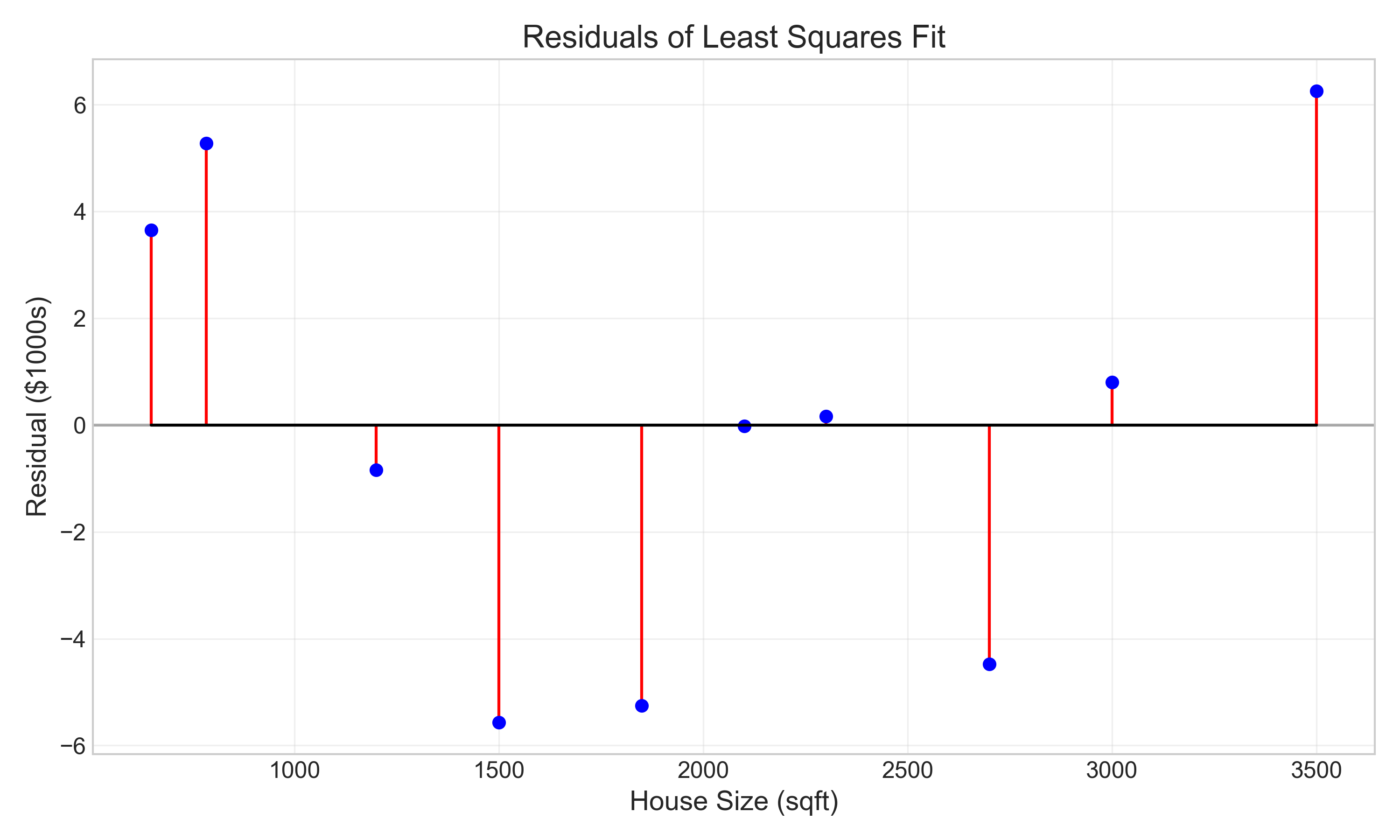


Figure 9: Residuals for the house price model

Key observations from the residual analysis:  
- The residuals are relatively small compared to the actual values, indicating a good fit.  
- The residuals appear randomly scattered around zero, with no obvious pattern, suggesting the linear model is appropriate.  
- There are no extreme outliers that might unduly influence the fit.  
  
A more comprehensive residual analysis would include:  
1. Normality test: Checking if residuals follow a normal distribution  
2. Homoscedasticity: Verifying constant variance across the range of predictors  
3. Autocorrelation: Testing for independence of residuals

## 6.2 Confidence Intervals

Confidence intervals quantify the uncertainty in the estimated parameters and predictions. For the house price model, the 95% confidence intervals for predictions are shown below:

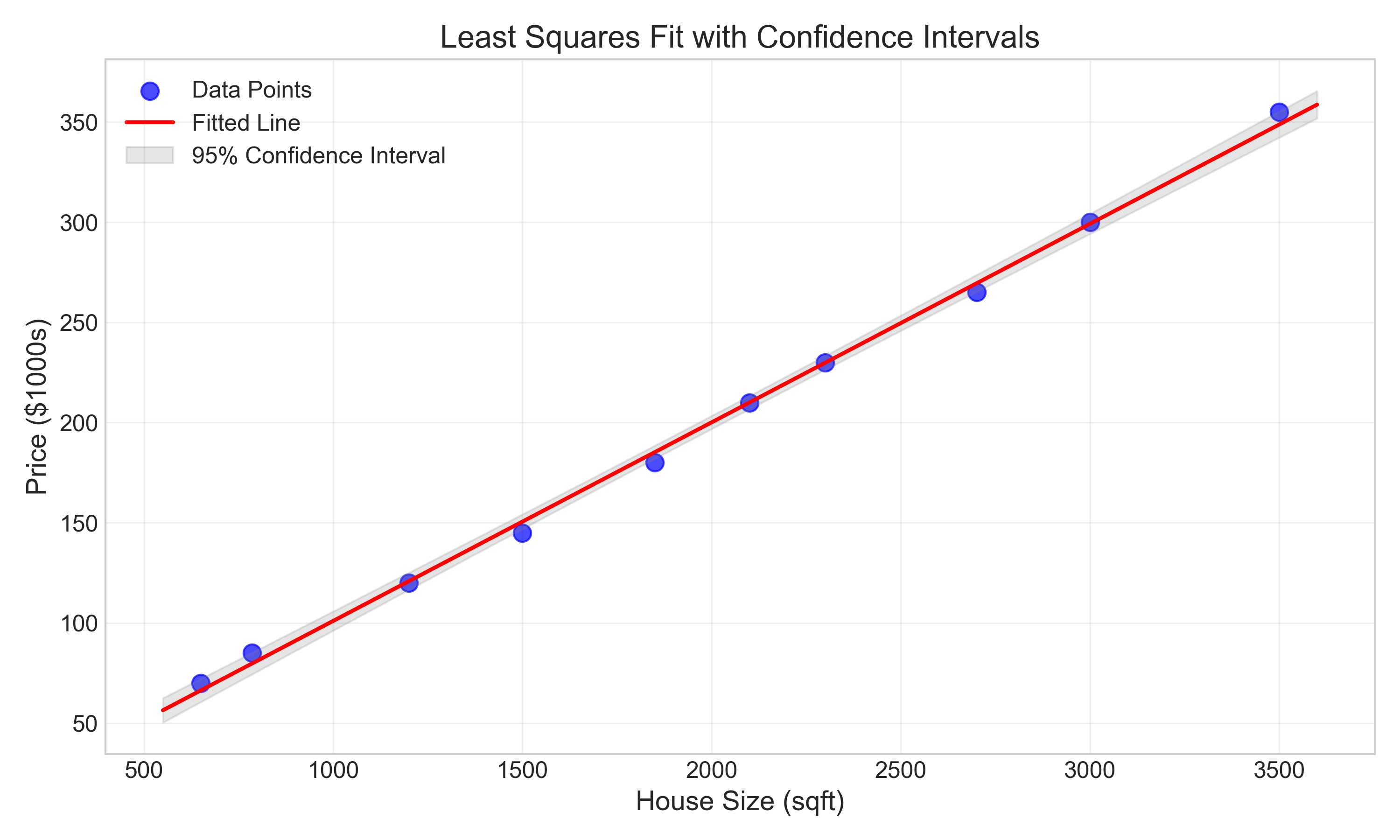


Figure 10: Least squares fit with 95% confidence intervals

The gray shaded area represents the 95% confidence interval for the predicted values. Note how the interval widens as we move away from the center of the data, reflecting increased uncertainty in predictions at the extremes.

## 6.3 Goodness of Fit Metrics

Several metrics help evaluate how well the model fits the data:  
  
1. Coefficient of Determination (R²): Measures the proportion of variance explained by the model  
 - For the house price model: R² = 0.9980  
 - This indicates that 99.80% of the variance in house prices is explained by house size  
  
2. Root Mean Square Error (RMSE): Measures the average magnitude of residuals  
 - For the house price model: RMSE = 4.0041  
 - This means predictions are off by about $4,004 on average  
  
3. Mean Absolute Error (MAE): Average of absolute residuals  
 - For the house price model: MAE = 3.2308  
 - This means predictions deviate by about $3,231 on average

# 7. Comparison with Other Methods

While least squares is powerful, other methods may be more appropriate in certain situations:  
  
1. Robust Regression: Less sensitive to outliers  
 - Uses methods like Huber loss or Tukey's bisquare  
 - Particularly useful when data contains outliers  
  
2. Ridge Regression: Adds a penalty term to handle multicollinearity  
 - Minimizes Σ[yi - f(xi; β)]² + λΣβj²  
 - Useful when predictors are highly correlated  
  
3. LASSO Regression: Encourages sparse solutions  
 - Minimizes Σ[yi - f(xi; β)]² + λΣ|βj|  
 - Useful for feature selection  
  
4. Orthogonal Distance Regression: Accounts for errors in both dependent and independent variables  
 - Minimizes perpendicular distances from points to the fitted curve  
 - Appropriate when both variables have measurement errors  
  
For our house price example, standard least squares is appropriate because:  
- The relationship appears strongly linear  
- There are no obvious outliers  
- We have a single predictor variable  
- The residuals are well-behaved

# 8. Conclusion

The least squares method provides a powerful framework for discovering mathematical laws that govern real-world phenomena. Through this report, we have:  
  
1. Explored the theoretical foundations of the method, from its mathematical formulation to practical implementation  
2. Applied linear least squares to uncover the relationship between house size and price  
3. Extended the approach to polynomial fitting for cyclical temperature data  
4. Demonstrated non-linear least squares for exponential population growth  
5. Conducted advanced analysis through residuals, confidence intervals, and goodness-of-fit metrics  
6. Compared least squares with alternative methods  
  
Key insights from our analysis:  
- The relationship between house size and price follows a linear law: Price ≈ 0.10 × Size + 1.94 ($1000s)  
- Monthly temperatures follow a 4th-degree polynomial pattern, reflecting seasonal variations  
- Population growth follows an exponential law: P(t) = 1.69e^(0.013t), indicating a constant growth rate  
  
The least squares method continues to be an indispensable tool in scientific research, engineering, economics, and many other fields. Its ability to extract meaningful patterns from noisy data makes it essential for understanding complex systems and making predictions based on observed behavior.

# 9. References

1. Björck, Å. (1996). Numerical Methods for Least Squares Problems. Society for Industrial and Applied Mathematics.  
2. Golub, G. H., & Van Loan, C. F. (2013). Matrix Computations (4th ed.). Johns Hopkins University Press.  
3. Lawson, C. L., & Hanson, R. J. (1995). Solving Least Squares Problems. Society for Industrial and Applied Mathematics.  
4. Nocedal, J., & Wright, S. J. (2006). Numerical Optimization (2nd ed.). Springer.  
5. Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical Recipes: The Art of Scientific Computing (3rd ed.). Cambridge University Press.  
6. Wikipedia: Least Squares. https://en.wikipedia.org/wiki/Least\_squares  
7. NIST/SEMATECH e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/