March 2, 2025

Introduction

Many relationships in the real world are not linear. For example, the growth rate of a plant may vary with temperature in a nonlinear way. Polynomial regression is an extension of linear regression that allows for better capture of these complex relationships. This work aims to introduce this method, explore its fundamental properties, and discuss its practical applications

Imagine tracking the growth of a plant over time. Initially, it might shoot up rapidly, but as it matures, its growth rate slows down. A simple straight line (as in linear regression) wouldn't accurately represent this curved pattern. Polynomial regression allows us to fit a curve (a polynomial) that better follows this non-linear trend. It's like trying to draw a curve with a flexible pencil rather than a rigid ruler.

Polynomial regression

Linear Regression Problem

Linear regression is based on the assumption that a straight line can model the relationship between the data. However, in some cases, the data do not follow a linear relationship but rather a curvilinear one. In these situations, polynomial regression is more appropriate.

General Formula of the Polynomial Model

A polynomial model can be expressed as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon$$

where:

- Y is the dependent (or target) variable,

- X is the independent variable,
- $\beta_0, \beta_1, \ldots, \beta_d$ are the coefficients to be estimated,
- d is the degree of the polynomial,
- ϵ is a random error term.

Associated Concepts

- **Flexibility:** The flexibility of the model depends on the degree d. A higher degree allows the model to capture more complex trends in the data.
- **Risk of Overfitting:** When d is too large, the model may fit the training data perfectly, including the random noise, which harms its ability to generalize to new data.
- **Choice of Degree d:** A trade-off is needed between accuracy and simplicity. The choice of degree can be based on techniques like cross-validation to avoid both underfitting and overfitting.

General Equation of Polynomial Regression

The general equation of polynomial regression is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

where:

- $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients to be estimated.
- ε represents the random error or residual.

Coefficient Estimation

To estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_n$, the least squares method is typically used, which minimizes the sum of the squared differences between the observed values and the values predicted by the model. This method involves solving the following normal equation system:

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y} \tag{1}$$

where:

- $\mathbf{X}$ is the Vandermonde matrix constructed from the x values,
- $\boldsymbol{\beta}$ is the vector of coefficients $[\beta_0, \beta_1, \ldots, \beta_n]^T$,
- $\mathbf{y}$ is the vector of observations for y.

Coefficient Estimation β

Starting from the general form of the model:

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$$

The goal is to minimize the sum of squared errors:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \big(Y_i - f(x_i, \beta_0, \beta_1, \ldots, \beta_p)\big)^2$$

Fitting the Coefficients

The coefficients $\beta_0, \beta_1, \ldots, \beta_d$ are generally estimated using the least squares method, which minimizes the sum of squared errors between predicted and observed values:

$$\text{Minimize} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

where $\hat{Y}_i$ is the predicted value from the model.

Bias, Variance, MSE, and R^2 in Linear and Polynomial Regression

Bias

Bias measures the systematic error of a model. It can be mathematically expressed as:

$$\text{Bias}^2 = (\mathbb{E}[\hat{r}(X)] - r(X))^2$$

where $\hat{r}(X)$ is the model's average prediction, and $r(X)$ is the true function.

Variance

Variance measures the model's sensitivity to variations in the training data. It is given by:

$$\text{Variance} = \mathbb{E}\left[(\hat{r}(X) - \mathbb{E}[\hat{r}(X)])^2\right]$$

Mean Squared Error (MSE)

MSE measures the average error between the predicted and observed values:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

- **Error metrics:** Metrics such as the Mean Squared Error (MSE) or the coefficient of determination (R^2) are used to evaluate the quality of convergence.

Coefficient of Determination (R^2)

The coefficient R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2}$$

Bias-Variance Tradeoff in Regression

The bias-variance tradeoff is crucial to avoid both underfitting and overfitting. In linear and polynomial regression, this translates to:

- **Models with high bias (linear regression):** Simple but inaccurate for complex relationships.
- **Models with high variance (high-degree polynomial regression):** Flexible but sensitive to noise.

Illustrations

- **Linear fit:** The relationship is modeled by the equation: $y = ax + b$.
- **Polynomial regression:** A curve is fitted to better model the data.

Polynomial Regression and Matrices

For a polynomial regression of degree 2, the optimization problem becomes:

$$S = \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - y_i \right)^2$$

Partial derivatives give the normal equations:

$$\frac{\partial S}{\partial \beta_0} = 0 \implies n\beta_0 + \beta_1 \sum x_i + \beta_2 \sum x_i^2 = \sum y_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \implies \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 = \sum y_i x_i$$

$$\frac{\partial S}{\partial \beta_2} = 0 \implies \beta_0 \sum x_i^2 + \beta_1 \sum x_i^3 + \beta_2 \sum x_i^4 = \sum y_i x_i^2$$

This can be represented in matrix notation:

$$A \cdot X = Y$$

with:

$$A = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}, \quad X = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad Y = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{bmatrix}$$

To solve $X = A^{-1}Y$, we calculate A^{-1} as follows:

$$A^{-1} = \frac{1}{\det(A)} \cdot \text{Com}(A)^T$$

Successive Derivatives of the Polynomial

The successive derivatives of the polynomial are:

- First derivative (rate of change):

$$f'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + \cdots + n\beta_n x^{n-1} \tag{2}$$

- Second derivative (concavity):

$$f''(x) = 2\beta_2 + 6\beta_3 x + \cdots + n(n-1)\beta_n x^{n-2} \tag{3}$$

- The successive derivatives of a polynomial describe its local behavior (slope, concavity, etc.)

1 Convergence of Polynomial Regression

The convergence in polynomial regression refers to the model's ability to fit the training data properly while generalizing well to new data. Several factors influence this convergence:

1.1 Convergence Analysis

- **Cross-validation:** Essential to ensure that the model converges properly and generalizes well to new data.
- **Asymptotic convergence:** With sufficient data and a well-specified model, the estimated coefficients converge to their true values.

Asymptotic Convergence

Asymptotic convergence in polynomial regression describes the behavior of the coefficient estimators β as the sample size n tends to infinity. Specifically, it focuses on the probability convergence and the asymptotic distribution of these estimators.

Consistency

An estimator $\hat{\beta}$ is said to be *consistent* if its probability limit, as n tends to infinity, is equal to the true value β:

$$\text{plim}_{n\to\infty}(\hat{\beta}) = \beta$$

This means that as the sample size increases, the probability that $\hat{\beta}$ is close to β approaches 1.

Asymptotic Distribution

Under certain standard assumptions (such as homoscedasticity and the absence of autocorrelation in the errors), the least squares estimator $\hat{\beta}$ follows an asymptotic normal distribution:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(X^T X)^{-1}) \quad \text{as } n \to \infty$$

where:

- $\xrightarrow{d}$ denotes convergence in distribution.
- σ^2 is the variance of the errors ϵ.
- X is the Vandermonde matrix.

This property is crucial as it allows for the construction of confidence intervals and hypothesis testing for the coefficients β when n is sufficiently large.

Practical Implications

In practice, asymptotic convergence implies that:

- With a large sample, the coefficient estimates will be close to the true values.
- The asymptotic normal distribution can be used to evaluate the precision of the estimates and test hypotheses.

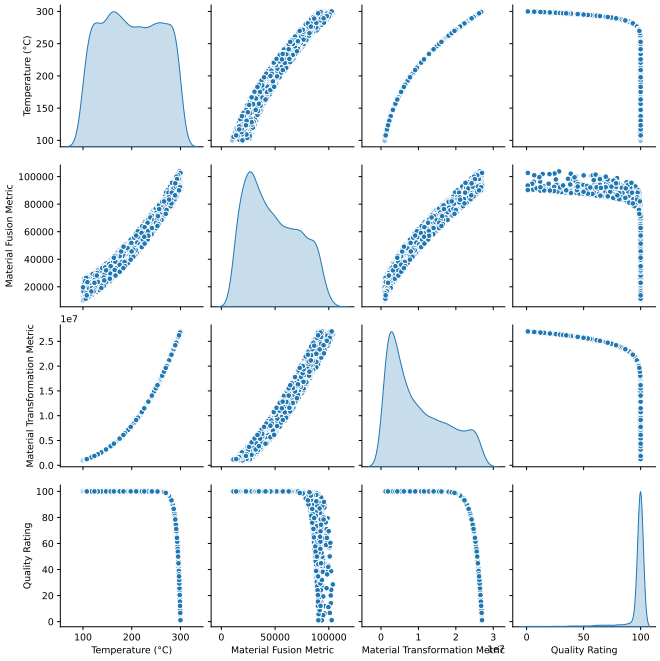
It is important to note that these results are asymptotic. For finite sample sizes, the properties of the estimators may differ.

This project explores the relationship between production variables (Temperature, Material Fusion Metric, and Material Transformation Metric) and the Quality Rating in an industrial context. The analysis implements polynomial regression to model the observed nonlinear relationships. The focus is on choosing the optimal polynomial degree, analyzing residuals, and performing cross-validation.

The data used in this study comes from a CSV file named manufacturing.csv. It contains temperature (°C) measurements, two material-related metrics (Material Fusion Metric and Material Transformation Metric), and a quality evaluation (Quality Rating).

The examination of the pairplot (Figure 1) reveals interesting relationships between the variables. The distribution of the temperature suggests the presence of two modes, which could indicate two distinct operating regimes in the manufacturing process. The material-related metrics show positively skewed distributions, which is common for this type of measurement. More importantly, nonlinear relationships are observed between temperature and the metrics, as well as between these metrics and the Quality Rating. Notably, the relationship between temperature and the Quality Rating shows a sharp drop in quality starting at approximately 275°C. These observations justify the use of a polynomial regression model to capture these nonlinearities.

The individual scatter plots (Figure 2) confirm the trends observed in the pairplot. The strong nonlinearity between temperature and the Quality Rating is clearly visible, as well as the less pronounced nonlinear relationships with the other metrics.



2 Results of Polynomial Regression

2.1 Choice of Polynomial Degree

Figure 3 compares the polynomial regression curves for degrees 1, 2, and 3. The linear model (degree 1) clearly fails to capture the curvature observed in the data. The degree 2 model offers an improvement, but the degree 3 model seems to follow the trend better, particularly the sharp drop in Quality Rating at high temperatures. The values of R^2 and MSE confirm this observation: the degree 3 model has the highest R^2 (0.69) and the lowest MSE (63.75).

The analysis of the bias-variance trade-off (Figure 4) reinforces this choice. The error on the test set (and cross-validation) reaches a minimum around degree 3, indicating a good balance between fit capacity and generalization.

2.2 Residual Analysis

The residual analysis (Figure 5) confirms the superiority of the polynomial model. The residuals of the linear model exhibit an asymmetric distribution, far from a normal distribution, indicating poor fit. In contrast, the residuals

of the polynomial model are more symmetrically distributed around zero, which is a positive sign.

Interpretation: This plot compares the error distribution for linear and polynomial models.

Errors for linear regression (blue) are widely spread, with significant negative peaks. Errors for polynomial regression (red) are more concentrated around zero, indicating a better overall fit.

Interpretation: This plot illustrates the relationship between temperature and quality rating for different polynomial regression degrees.

The degree 1 curve (linear regression) poorly captures the overall trend, as indicated by its low R^2 (0.21) and high MSE (160.61). The degree 2 curve improves the fit, with R^2 of 0.47 and MSE of 108.24. The degree 3 curve provides the best fit among the three models (R^2=0.69, MSE = 63.75), suggesting a nonlinear relationship is more appropriate.

Interpretation: This plot shows how errors evolve with the polynomial degree.

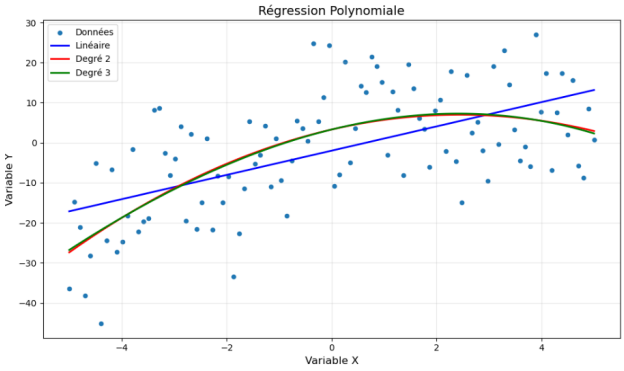
Errors decrease sharply between degrees 2 and 5, indicating that adding polynomial terms helps better fit the data. Beyond degree 5, errors stabilize, suggesting additional complexity offers no significant benefit.

Interpretation: This plot shows the relationship between each metric and quality rating:

Temperature vs Quality Rating: A nonlinear relationship is visible, where quality remains high up to 275°C and then drops sharply. Material Fusion Metric vs Quality Rating: A similar pattern is observed, with quality staying stable up to a certain value before a rapid decline. Material Transformation Metric vs Quality Rating: Quality remains constant, with a noticeable drop at high values.

Conclusion

Polynomial regression has proven effective in modeling the nonlinear relationships of production data, outperforming linear regression. The optimal choice of degree (d) is crucial to avoid underfitting and overfitting. Although effective, the method is sensitive to outliers. Future research could explore



[ ]:

Figure 1: Illustration de la régression polynomiale avec différents degrés.

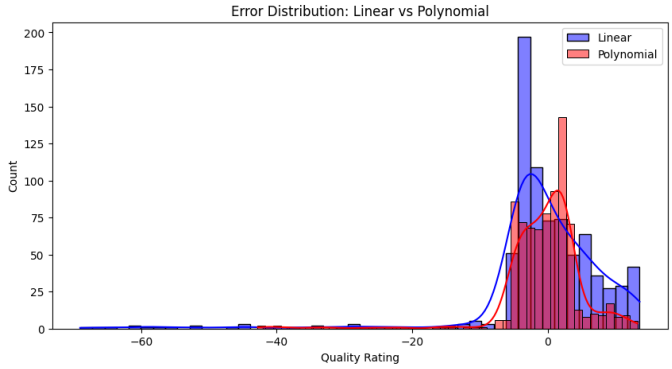


Figure 2: Residual error comparison

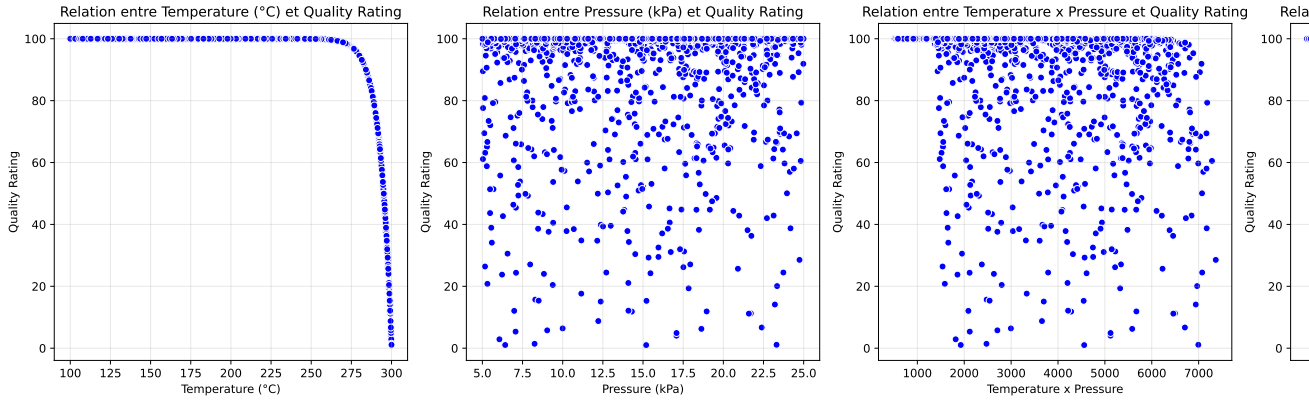


Figure 3: Scatter plot comparison

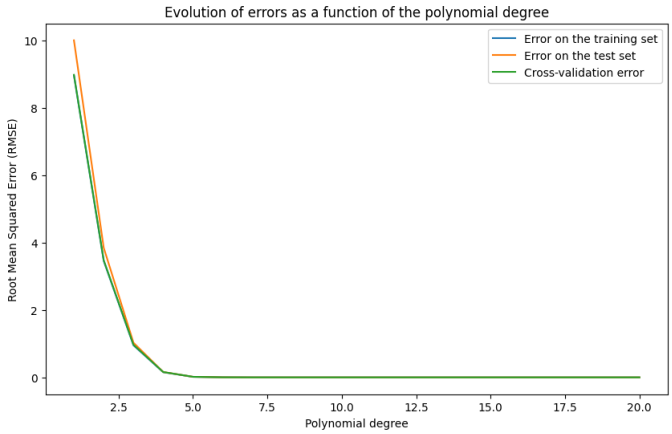


Figure 4: Error evolution with polynomial degree

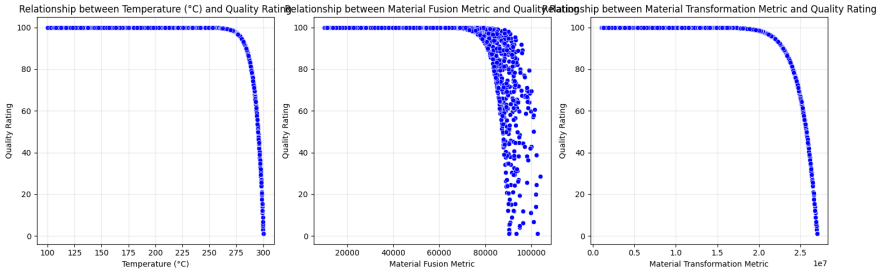


Figure 5: Quality rating vs metrics

regularization, asymptotic convergence, and application to other fields.