Predictive Optimisation of Reaction Yield

A Statistical Data Science Approach

Stuti Malik

Table of Contents

Table of Contents

Project Overview and Objective

- This project focuses on identifying the combination of experimental settings that maximise the yield of a chemical reaction.
- Dataset includes:
  - X_1: Reaction Temperature (°C)
  - X_2: pH
  - X_3: Pressure (kPa)
  - Y: Yield (%)

Objective

Recommend optimal values of temperature, pH, and pressure that maximise the predicted yield using statistical modelling and optimisation techniques.

Table of Contents

Exploratory Analysis and Assumptions

Exploratory Focus

- Examine distribution of variables and potential transformations (Box-Cox parameter).
- Identify multicollinearity and interaction effects (Variance Inflation Factor).
- Detect outliers and influential points (Cook's Distance).
- Validate modelling assumptions:
  - Normality of residuals (QQ plot)
  - Homoscedasticity (residual plots)
  - Independence of error terms (no autocorrelation)
- Predictions should be made within observed data ranges to avoid extrapolation.

Table of Contents

Regression Model Approach

Polynomial Regression Model

The second-order polynomial regression model:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3$$

- Quadratic model captures curvature and interaction effects.
- Functional marginality principle: interaction terms only included when main effects are significant.

Regression Model in Matrix Form

- Predicted values: $\hat{Y} = X\hat{\beta}$
- Data matrix X and coefficient vector $\hat{\beta}$:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$$

- Linear in parameters; residuals $r_i = y_i - \hat{y}_i$.

Variable Scaling

- Variables coded for model stability:
  - $X_1 = \frac{\text{Temperature}-200}{50}$
  - $X_2 = \text{pH} - 5$
  - $X_3 = \frac{\text{Pressure}-175}{25}$
- Standard operating conditions: Temperature = 200 °C, pH = 5, Pressure = 175 kPa

Table of Contents

Model Selection Process

- Compared candidate models using:
  - Global and partial F-tests
  - ANOVA / ANCOVA
  - Adjusted R^2, AIC, C_p
- Stepwise and all-subset regression for balance of complexity and predictive accuracy.
- Considered all polynomial and interaction terms up to second order:
  - Constant, univariate, multivariate, and full models

Table of Contents

Testing Model Fit

Test Statistics

- Error sum of squares: $SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- Regression sum of squares: $SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- Total sum of squares: $SS_T = SS_R + SS_E$
- Mean squared error: $MS_E = \frac{SS_E}{n-p}$
- R^2, adjusted R^2, AIC, C_p

Residual Diagnostics

- Normality (QQ plot)
- Uniform scatter vs fitted values
- No heteroscedasticity

Table of Contents

Hypothesis Testing

- Coefficient tests: *t*-statistics

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2[(X^T X)^{-1}]_{jj})$$

- Global and partial F-tests for model adequacy
- Chi-squared statistics for variance estimation

Table of Contents

Optimisation Process

Stationary Point Calculation

For q explanatory variables, second-order polynomial:

$$\underline{\hat{x}_s} = -\frac{1}{2}\hat{\beta}^{-1}\underline{\hat{b}}$$

- If all eigenvalues of $\hat{\beta}$ are negative, the stationary point is a maximum.
- Provides optimal temperature, pH, and pressure.

Table of Contents

Recommended Settings

Optimal Reaction Conditions (Example)

- Temperature: 225 °C
- pH: 5.8
- Pressure: 180 kPa
- Predicted Yield: 96.3%

- Marginal sensitivity analysis shows temperature and pH as key drivers of yield.

Table of Contents

Discussion and Future Work

- Demonstrates how statistical modelling guides experimental optimisation.
- Quadratic regression provides interpretability and robust predictions.
- Future Work:
  - Explore non-linear or ML models (e.g., random forests)
  - Multi-stage or multi-output optimisation
  - Automate model selection and diagnostic checks

References

- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis.*
- Montgomery, D. C. (2019). *Design and Analysis of Experiments.*
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models.*