Statistical Data Science

Application example

Stuti Malik

Table of Contents

- Project overview and objective
- Assumptions and data exploration
- Process of selecting a multivariate statistical model
- Model selection
- Model testing for an adequate fit
- Hypothesis testing
- Applying the model
- Optimisation model
- Further discussions

Project overview and objective

A file containing data from an experiment on a chemical reactor contains the following variables:

- $X1$: the reaction temperature, in degrees Celsius
- $X2$: the pH of the reaction
- $X3$: the pressure under which the reaction is run, measured in kPa
- Y: the yield of the reaction as a percentage

Task

The aim of this project is to recommend settings of temperature, pH and pressure which will maximise the yield of the reaction

Assumptions and data exploration

Testing assumptions

- Distribution of the error terms - check if higher order term is needed or a transformation need to be applied - does its variance stay constant or have a trend? (residual plot) - does the error terms come from normal distribution for each response variable (QQ plot), with mean 0 and constant variance
- Independence of response variables, no autocorrelation
- Box-Cox transformation parameter, to see if a transformation of response is needed

Data exploration

- Consider the range of data: extrapolation/predictions should be made within the range
- Check for influential outlier effect (Cook's distance)
- Check for multicolinearity and interactions (Variance inflation factor)

Process of selecting a multivariate statistical model

Model selection process

A process involved in creating a multivariate regression model: To test if a model is a good-fit, consider the statistic values and plots from the model which illustrate the adequacy of the model

- Global F-test (is any of the term related to response)
- Partial F-test (for testing higher order terms, terms with non-significant t-test, variables not believed to be linked)
- Analysis of variance (ANOVA)
- Analysis of covariance, for example, at a set of different pH levels

Initial process for model selection

- Number of observations = n, number of parameters = p
- Number of subsets of linear regression models = 2^p
- Maximum number of polynomial terms, for a full polynomial of order d, with m features = $1 + md + mC2$ (in this example: d=2, m=3)

Model selection

Several linear multivariate regression models were fitted, and analysed in R, using the MASS library, to estimate the expected yield at different settings of temperature, pH and pressure.

Regressors, and their corresponding residual sum of squares considered.

All possible linear regressors/ explanatory variable terms used in the linear predictor:

- Constant model: None
- Univariate: X_1
- Univariate: X_2
- Univariate: X_3
- Multivariate: X_1, X_2
- Multivariate: X_1, X_3
- Multivariate: X_2, X_3
- Multivariate, full model: X_1, X_2, X_3

Considerring the 2nd order polynomial terms:

- Full model: $X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_1X_2, X_1X_3, X_2X_3$

Model testing for an adequate fit

All subset regression, or a forward selection, backward elimination, or step-wise selection can be tried

Example of test statistics ($\bar{Y}$: mean response (scalar), $\hat{Y}$: model estimate)

- Error sum of squares, $SS_E = \Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)$
- Regression/model sum of squares, $SS_R = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})$
- Total sum of squares, $SS_T = SS_R + SS_E = \Sigma_{i=1}^{n}(Y_i - \bar{Y})$
- Mean squard error, $MS_E = \frac{SS_E}{n-p}$ (minimise)
- R^2 / adjusted R^2 (consider the ratio of model sum of squares, SS_R, and total sum of squares, SS_T) (close to 1?, close to the full model?)
- (Adjusted) C_p (close to p, simplest model, minimise for minimising MSE)
- Generalisation of C_p statistic, for any modal- AIC (minimise)

Hypothesis testing

Chi squared statistics- for applications in sample standard deviation estimates (student t-distribution) and F-tests

- Hypothesis testing: T-statistic (for each coefficient, modelled as a random variable)
  $\hat{\beta} = (X^T X)^{-1}(X^T Y)$
  $\hat{\beta}_j \sim N(\beta_j, \sigma^2[(X^T X)^{-1}]_{jj})$ - σ estimated with same variance, giving rise to t-test
  β_j lies between $[\ \hat{\beta}_j - t_{n-p,1-/\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\ ,\ \hat{\beta}_j + t_{n-p,1-/\frac{\alpha}{2}} s.e.(\hat{\beta}_j)\ ]$
- Hypothesis testing: F-statistic (two types: global/partial)
  Variance ration, $\frac{MS_R}{MS_E} = F_{(p-1,n-p)}$, where $MR_R = \frac{SS_R}{p-1}$, $MS_E = \frac{SS_E}{n-p}$, have Chi-squared distribution with (p-1) and (n-p) degrees of freedom, respectively

Applying the model

Answering the question

At what settings would you recommend the reaction be run in order to maximise the yield?

- Summarise the steps involved in using the selected model to solve the optimisation problem (finding stationary points)

Model used to make predictions (considering functional marginality):
$lm(Y \sim X1 + X2 + X3 + I(X1^2) + I(X2^2) + I(X3^2) + I(X1 * X3) + I(X2 * X3))$

- R^2 and adjusted R^2 close to 1, and to that of a full model, AIC better than that of a full model, which includes the X1*X2 term
- The Normal QQ-plot shows that the standardised/studentised residuals ($r_i = \frac{e_i}{\sqrt{(1-h_{ii})s^2}}$, σ estimated from the sample:$=s$) come from a normal distribution
- The standardised residuals, r_i against $\hat{y}_i$ are uniformly scattered

Optimisation model

We want to find the values of explanatory variables which maximise the expected response, the percentage yield of reaction. Hence, selecting a model with a stationary point, like a second order polynomial regression model, is preferred.

Second order polynomial (linear in parameter) model

For q explanatory variables:

$E(Y_i) = \beta_0 + \Sigma_{j=1}^{q} \beta_j x_{ji} + \Sigma_{j=1}^{q} \beta_{jj} x_{ji}^2 + \Sigma_{j=1}^{q-1} \Sigma_{k=j+1}^{q} \beta_{jk} x_{ji} x_{ki}$

$E(Y_i) = \hat{\beta}_0 + \underline{x_i}^T \underline{\hat{b}} + \underline{x_i}^T \hat{\beta} \underline{x_i}$ - differentiate (w.r.t $\underline{x}$) to get $\underline{\hat{b}} + 2\hat{\beta}\underline{x}$, and equate to 0, to find $\underline{\hat{x_s}}$

The stationary point, $\underline{\hat{x_s}} = -\frac{1}{2}\hat{\beta}^{-1}\underline{\hat{b}}$

If all eigenvalues of $\hat{\beta}$ are negative, the stationary point is a maximum

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_{11} & \frac{1}{2}\hat{\beta}_{12} & . & . & \frac{1}{2}\hat{\beta}_{1q} \\ \frac{1}{2}\hat{\beta}_{12} & \hat{\beta}_{22} & . & . & \\ . & & & & \\ \frac{1}{2}\hat{\beta}_{1q} & . & . & . & \hat{\beta}_{qq} \end{pmatrix}, \text{ and } \underline{\hat{b}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ \hat{\beta}_q \end{pmatrix}$$

Further discussions

- Statistical modelling and data science techniques, applied in agricultural and environmental research
- Data integration and data fusion techniques from various data sets, meta-analysis, considering the existing research, and selecting the right tools for analysis
- Challenges for agriculture

Additional slide: Thinking of a modelling approach

The data is coded so that:

- $X1$ = (Temperature - 200)/50
- $X2$ = pH - 5
- $X3$ = (Pressure - 175)/25
- Y: the yield of the reaction as a percentage

The standard operating conditions for the reaction are Temperature = 200, pH = 5 and Pressure = 175.

Models to consider

- Simple linear regression
- Multiple linear regression
- Polynomial regression (with or without interaction terms)
- Qualitative explanatory variables (observations per category)

Additional slide: Regression model design

Regression analysis is used to model the relationship between a dependent variable and one or more independent variables.

Linear in the parameters model

$\hat{y}=\hat{f}(x)=\beta_0+\beta_1 f_1(x)+\beta_2 f_2(x)+\dots\beta_p f_p(x)$, where f_i are the basis functions or feature mappings that we choose, and β_i are the model parameters that we choose.

The Prediction error or residual, $r_i = y_i$ - $\hat{y}_i$

Least squares

Common method for choosing model parameters, $\underline{\hat{\beta}}$, is to minimise the root mean square (RMS) prediction error, $\epsilon = \sqrt{\frac{(r_i^2+\dots r_n^2)}{n}}$ on the given data set, which is same as minimising the sum of squares of the prediction error.

Example of a multivariate linear model, in a matrix form

The predicted or expected value of the response variable, $\hat{Y} = \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ . \\ . \\ \hat{y_n} \end{pmatrix} =$ $X\hat{B}$, where X is the data matrix and $\hat{\beta}$ is the vector of optimal parameters selected.

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2p} \\ . & & & & & & \\ . & & & & & & \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix} \text{ and } \hat{B} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ . \\ \hat{\beta}_p \end{pmatrix}$$

A normal linear model

$Y \sim N_n(X\beta, \sigma^2 I)$, where Y is a $n \times 1$ vector, X is a $n \times p$ matrix, and β is a $p \times 1$ vector