# MAT314 Regression Analysis

## Project 2: Polynomial Regression Models

Ang Jin Wei
Wang Yihan
Lai Yupeng
Lew Wai Hong
Wang Fangyu

June 2026

# Theory(1): Polynomial Regression Model

**Polynomial regression model:**

- A regression model that uses polynomial terms of a predictor, such as $x^2$ and $x^3$, to describe a nonlinear relationship.
- Although the model is nonlinear in x, it is still linear in the coefficients.

**One-variable model:**

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

**Two-variable model:**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

# Theory(2): $k^{th}$ Order Polynomial Regression Model

**$k^{th}$ Order Polynomial Regression Model:**

- A regression model that uses polynomial terms of predictor up to degree k, to describe a nonlinear relationship.
- The highest power of the predictor can go only up to k.

**For example, given k = 3,**

Model 1: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

Model 2: $y = \beta_0 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

Model 3: $y = \beta_0 + \beta_3 x^3 + \varepsilon$

# Theory(2): $k^{th}$ Order Polynomial Regression Model

**Restrictions:**

1. No term can have a degree greater than k.
2. For the model to be truly of $k^{th}$-order, the coefficient of $x^k$ must not be equal to 0.
3. The number of observations must be greater than the number of estimated parameters, so that the error's degrees of freedom is positive.
4. The model is nonlinear in x, but linear in the parameters (power of $\beta$s are always 1).

# Theory(3): Is $k$ the higher the better?

**Answer: No**

- Higher k usually means more coefficients ($\beta$) included in the model.
- This causes the degrees of freedom of residuals ($df_{RES}$) to decrease.

$$df_{RES} = n - (k + 1)$$

- If the predictor added does not contribute much in prediction, sum of squares of residuals ($SS_{RES}$) only decrease a little bit.

$$MeanSquaredError(MSE) = SS_{RES}/df_{RES}$$

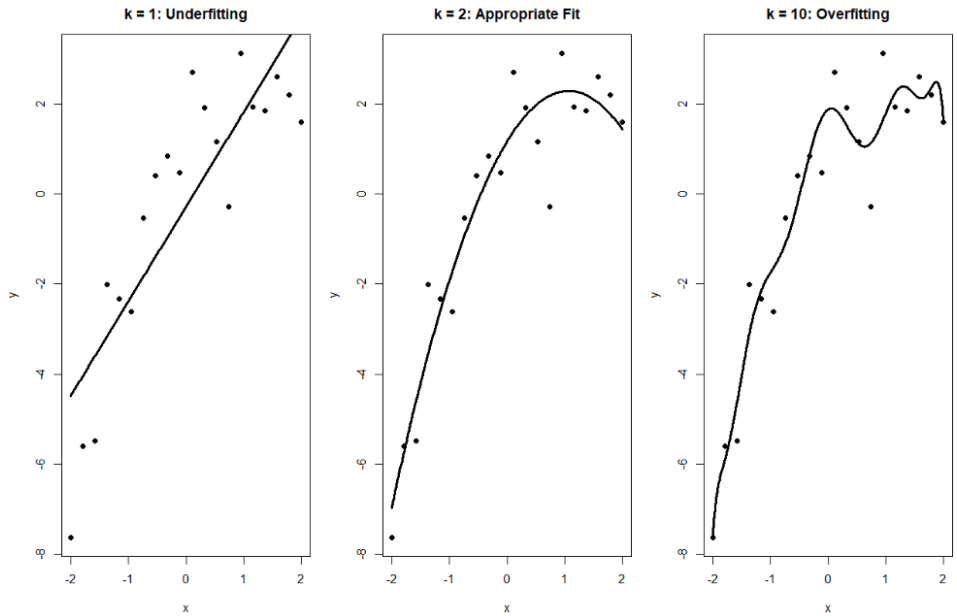
- MSE will increase if the reduction in $df_{RES}$ is much larger than $SS_{RES}$.

# Theory(4): Overfitting in Polynomial Regression Model

**Overfitting:**

- The model is too complex and fits the random noise in the data instead of only capturing the underlying relationship.
- Overfitting model has smaller training error but perform poorly on new data.
- An appropriate k should be chosen to prevent the model from overfitting and underfitting.

# Theory(4): Overfitting in Polynomial Regression Model




- When k is very small, the model cannot capture the underlying relationship within the data. (Underfitting)
- When k is appropriate, the model can capture the underlying relationship without being affected too much by each data points.
- When k is too large, the model tends to fit all data points too closely. (Overfitting)

# Theory(5): Forward Selection Procedure

**Steps:**

1. Start with the simplest linear model (order 1):

$$\text{Fit } k = 1: \qquad y = \beta_0 + \beta_1 x + \varepsilon$$

2. Add the first polynomial term (order 2):

$$\text{Fit } k = 2: \qquad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

After fitting the second-order model, compare it with the first-order model to see whether adding $x^2$ improves the model.

# Theory(5): Forward Selection Procedure

3. Use a partial $F$-test to make this comparison.
   - If the $p$-value for adding $x^2$ is $< 0.05$, keep $x^2$ and continue by testing $x^3$.
   - If the $p$-value is $\geq 0.05$, do not include $x^2$ and choose the first-order model.
4. Continue this process until the newly added polynomial term is not significant.

   **Note:** Lower-order terms should be retained when higher-order terms are included.

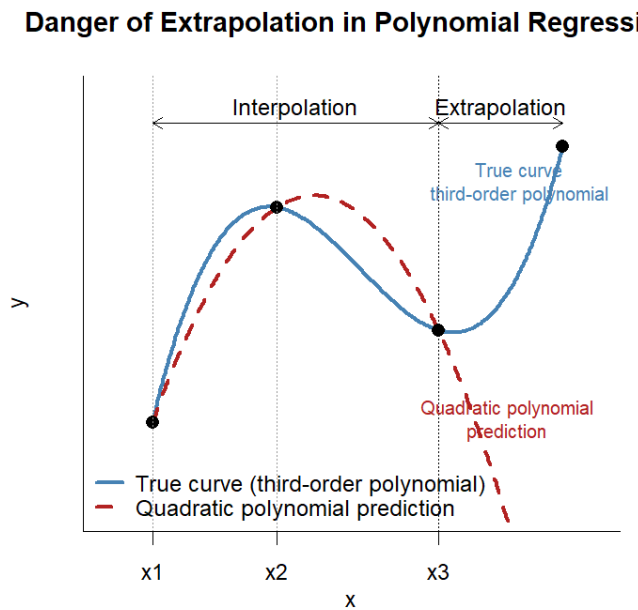
# Theory(6): Extrapolation in Polynomial Regression

**Definition:** Extrapolation uses a fitted regression model to predict $y$ for $x$-values outside the observed data range.

**Why extrapolation is dangerous:**

- Prediction accuracy generally decreases outside the observed range, where no data are available to verify the results.
- A polynomial curve may change direction or increase rapidly beyond the fitted data.
- High-degree terms such as $x^k$ may dominate for large $|x|$, leading to large prediction errors.

# Theory(6): Extrapolation in Polynomial Regression




**Graph interpretation:**

- Within the observed range $x_1$ to $x_3$, the fitted quadratic curve follows the true curve reasonably well.
- Beyond $x_3$, the fitted curve moves downward while the true curve turns upward, showing a clear deviation.
- Polynomial extrapolation is risky because curves may behave unpredictably outside the observed range, especially in higher-order models.

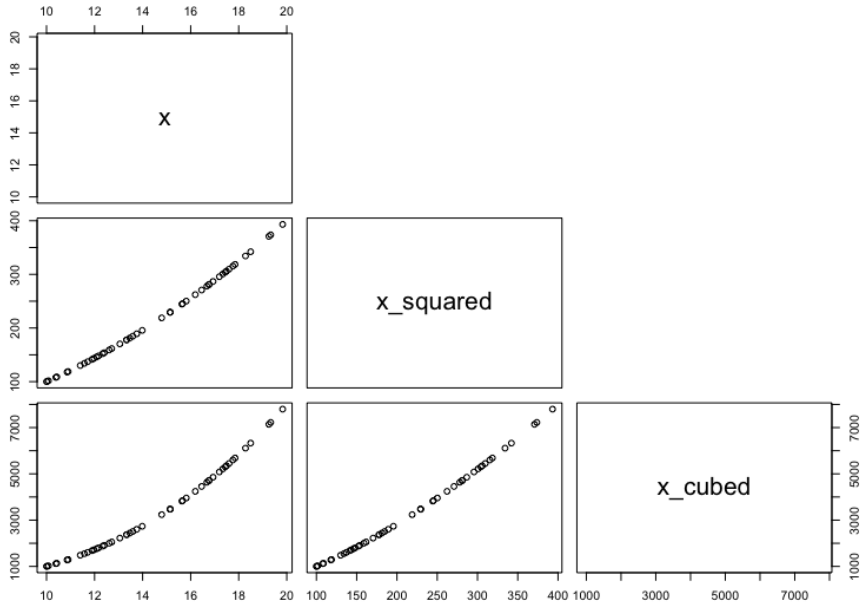
# Theory(7):Multicollinearity in Polynomial Regression

**What is multicollinearity?**
When two or more predictor variables are highly correlated, coefficient estimates may become unstable and unreliable.

**Why it occurs in polynomial regression:**

- Polynomial predictors such as $x$, $x^2$, and $x^3$ are all derived from the same original variable.
- These terms are not independent and may be highly correlated with each other.
- This can lead to large standard errors and unstable coefficient estimates.

# Theory(7):Multicollinearity before Mean Centering



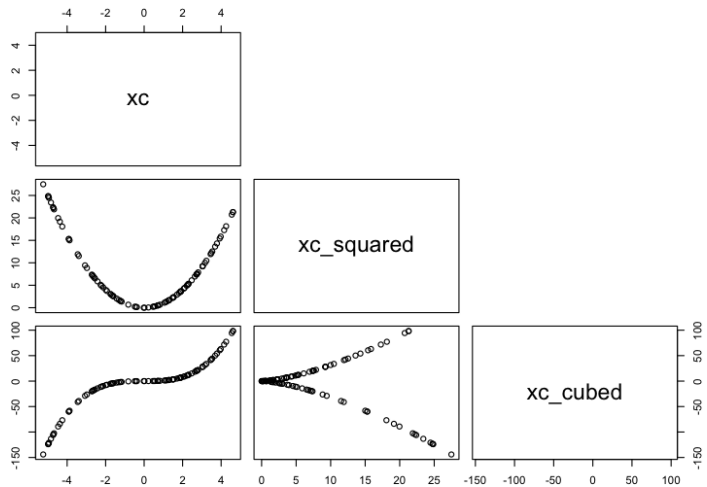

**Graph interpretation:**

- A strong linear trend is shown between $x$ and $x^2$.
- A near-linear trend is shown between $x$ and $x^3$.
- This indicates high correlation among polynomial terms because they are derived from the same original variable.

# Theory(7): Mean Centering as the Remedial Measure

**Mean centering:** subtract the mean of the predictor variable from each observation.

$$x_c = x - \bar{x}$$




**After Mean Centering:**

- The mean of $x_c$ becomes zero.
- The relationship between $x_c$, $x_c^2$ and $x_c^3$ becomes more symmetric and less linearly correlated.
- Mean centering reduces correlation among polynomial terms, but may not completely remove multicollinearity.

# Theory(7):Multicollinearity by Variance Inflation Factor (VIF)

**Uncentered Polynomial Terms**

| Term | VIF |
|---|---|
| $x$ | 9339.627 |
| $x^2$ | 39266.049 |
| $x^3$ | 10524.208 |

**Centered Polynomial Terms**

| Term | VIF |
|---|---|
| $xc$ | 6.461981 |
| $xc^2$ | 1.115907 |
| $xc^3$ | 6.749814 |

- Before mean centering, the VIF value of all the polynomial terms are absolutely larger than 5, which means that they are severely linearly correlated to each other.
- After mean centering, the VIF values of all polynomial terms have been decreased, and $xc^2$ have VIF value less than 5, showing that multicollinearity is reduced successfully.

# Theory(8): Hierarchical Polynomial Models

**Definition:** A hierarchical polynomial model includes all lower-order terms whenever a higher-order term is included.

- Example:
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$
- This model is hierarchical because $x$ and $x^2$ are retained when $x^3$ is included.

**Advantages:**

- Provides a complete and logically consistent polynomial model.
- Remains stable when $x$ is shifted or mean-centered.
- Allows clear comparisons between nested polynomial models.

# Theory(8): Non-hierarchical Polynomial Models

**Definition:** A non-hierarchical polynomial model includes a higher-order term without including all required lower-order terms.

- Example:
$$y = \beta_0 + \beta_1 x + \beta_3 x^3 + \varepsilon$$
- This model is non-hierarchical because $x^2$ is omitted even though $x^3$ is included.

**Advantages:**

- Uses fewer parameters and preserves more degrees of freedom.
- May be useful when theory suggests some lower-order terms should be zero.
- Can be simpler, but is usually harder to justify.

# Data analysis: Introduction

- The purpose of this project is to study polynomial regression models and apply them to the given dataset, `WineData.csv`.
- In this dataset, the response variable is pH, and the regressor is fixed acidity.
- Fixed acidity measures the total amount of acids in the wine, while pH measures the strength of acidity.
- Since the relationship between fixed acidity and pH may not be adequately described by a straight line, polynomial regression models are considered.
- The analysis mainly includes data visualisation, model selection by forward selection, multicollinearity checking, residual analysis, model validation, and the final fitted model.

# Data Analysis: Data Splitting

The full dataset is randomly divided into an estimation dataset and a prediction dataset. The estimation dataset is used to build and compare the candidate polynomial regression models.

```
WineData <- read.csv("WineData.csv")

set.seed(1)
index_e <- sort(sample(nrow(WineData), nrow(WineData) * .8))

WineData_e <- WineData[index_e, ]
WineData_p <- WineData[-index_e, ]

nrow(WineData_e)
nrow(WineData_p)
```

# Data Analysis: Data Splitting Result

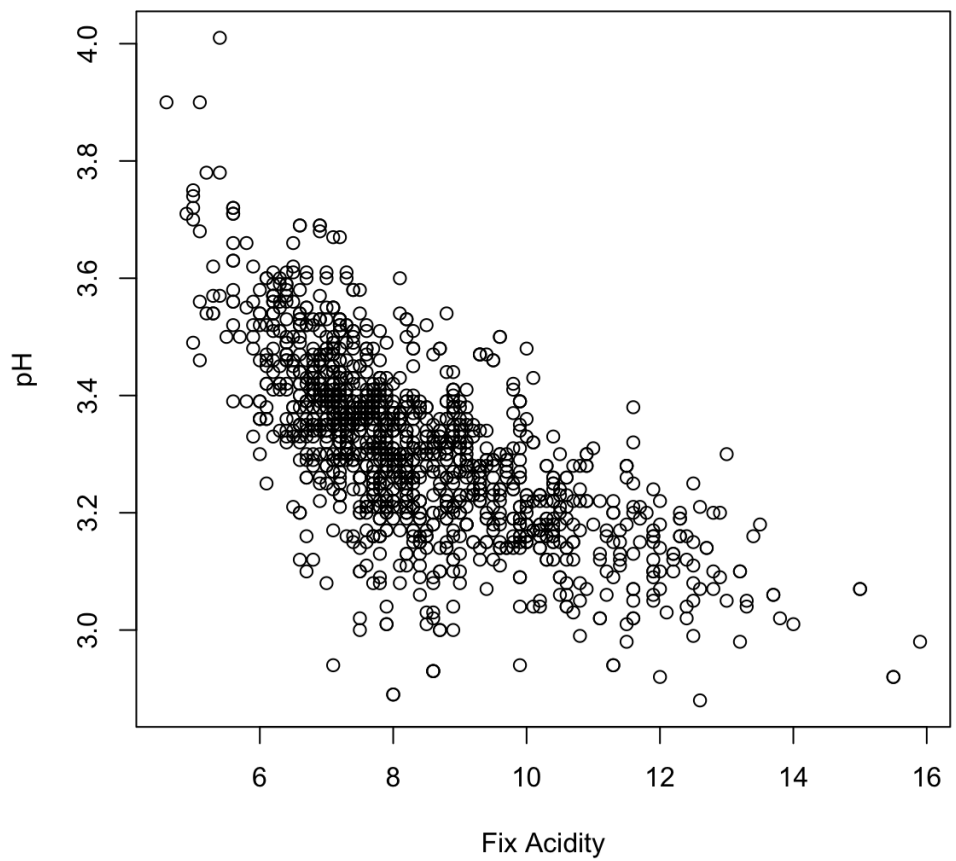
| Dataset | Number of Observations |
|---|---|
| Estimation dataset | 1279 |
| Prediction dataset | 320 |

The estimation dataset contains 80% of the observations and is used for model building. The remaining 20% is kept for later prediction checking.

# Data Analysis: Visualisation

The scatter diagram is used to inspect the relationship between fixed acidity and pH before fitting the polynomial models.

```
plot(WineData_e$fixed.acidity, WineData_e$pH,
     xlab = "Fixed Acidity", ylab = "pH")
```



# Data Analysis: Visualisation

- The response variable is pH and the regressor is fixed acidity.
- Inspection of the scatter diagram indicates that the relationship between pH and fixed acidity may not be linear.
- The rate of decrease becomes smaller as fixed acidity increases. Therefore, restricting the relationship to a straight line may be too simple, and a polynomial regression model is considered appropriate.

# Data Analysis: Polynomial Terms

The regressor is denoted by $x$, where

$$x = \text{fixed acidity}.$$

Polynomial terms up to the fifth order are created.

```
WineData_e$x <- WineData_e$fixed.acidity
WineData_e$x2 <- WineData_e$x^2
WineData_e$x3 <- WineData_e$x^3
WineData_e$x4 <- WineData_e$x^4
WineData_e$x5 <- WineData_e$x^5
```

# Data Analysis: Forward Selection

Candidate models are fitted by adding one higher-order polynomial term at a time.

```
model1 <- lm(pH ~ x, data = WineData_e)
model2 <- lm(pH ~ x + x2, data = WineData_e)
model3 <- lm(pH ~ x + x2 + x3, data = WineData_e)
model4 <- lm(pH ~ x + x2 + x3 + x4, data = WineData_e)
model5 <- lm(pH ~ x + x2 + x3 + x4 + x5,
             data = WineData_e)
```

# Data Analysis: Partial F-tests

Consecutive nested models are compared using partial F-tests.

```
anova(model1, model2)
anova(model2, model3)
anova(model3, model4)
anova(model4, model5)
```

The above code using R shows the test statistic $F$ and its corresponding $p$-value. At the significance level $\alpha = 0.05$, the polynomial term is considered statistically significant if the $p$-value is less than $\alpha$.

# Data Analysis: Forward Selection Result

- The quadratic, cubic, and fourth-order terms are significant at the 5% significance level.
- The fifth-order term is not significant because its $P$-value is 0.8842, which is larger than 0.05.
- Therefore, the selected polynomial order was $k = 4$.

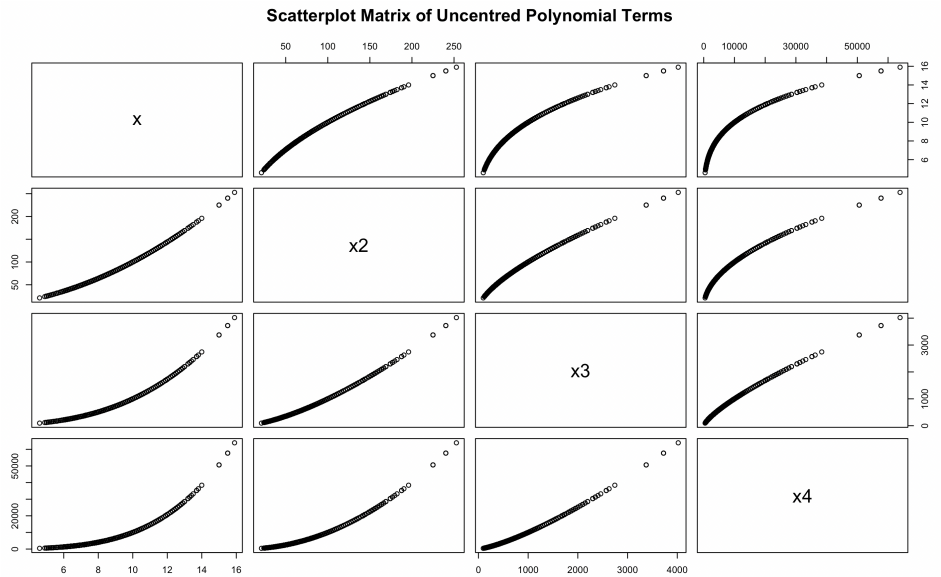
| **Comparison** | **Added term** | **F** | **P-value** | **Decision** |
|---|---|---|---|---|
| Model 1 vs Model 2 | $x^2$ | 95.597 | $< 2.2 \times 10^{-16}$ | Retain |
| Model 2 vs Model 3 | $x^3$ | 41.988 | $1.309 \times 10^{-10}$ | Retain |
| Model 3 vs Model 4 | $x^4$ | 4.6474 | 0.03129 | Retain |
| Model 4 vs Model 5 | $x^5$ | 0.0212 | 0.8842 | Stop |

# Data Analysis: Connection with Theory

- The scatter diagram motivates a polynomial regression model rather than a simple straight-line model.
- The candidate models follow the form of a $k^{\text{th}}$ order polynomial model in one variable.
- Forward selection uses partial F-tests to decide whether a higher-order term gives a significant improvement.
- Since the selected model contains powers of the same regressor, multicollinearity checking is the next step.

# Data Analysis : Checking for Multicollinearity

## Scatterplot Matrix of Uncentred Polynomial Terms




- Follow narrow, smooth patterns instead of random cloud
- The term show strong systematic dependence visually.

# Data Analysis : Checking for Multicollinearity

**Correlation Matrix**

| | $x$ | $x^2$ | $x^3$ | $x^4$ |
|---|---|---|---|---|
| $x$ | 1.000 | **0.991** | **0.964** | **0.920** |
| $x^2$ | **0.991** | 1.000 | **0.991** | **0.964** |
| $x^3$ | **0.964** | **0.991** | 1.000 | **0.991** |
| $x^4$ | **0.920** | **0.964** | **0.991** | 1.000 |

- Correlation matrix which $|r| > 0.8$ shows high pairwise linear correlations.

# Data Analysis : Checking for Multicollinearity

**VIF of uncentred polynomial terms**

| Term | VIF |
|---|---|
| $x$ | $22,752.64$ |
| $x^2$ | $200,324.21$ |
| $x^3$ | $207,149.52$ |
| $x^4$ | $25,121.24$ |

- All VIF values are larger than 5, this indicates multicollinearity is present in the uncentred fourth-order polynomial model.
- Therefore, mean centring is required before proceeding to the final model-selection stage.

# Data Analysis : Mean-Centring Diagnosis

**Mean-centring procedure**

After applying mean-centring,

$$x_c = x - \bar{x}$$

From estimation dataset,

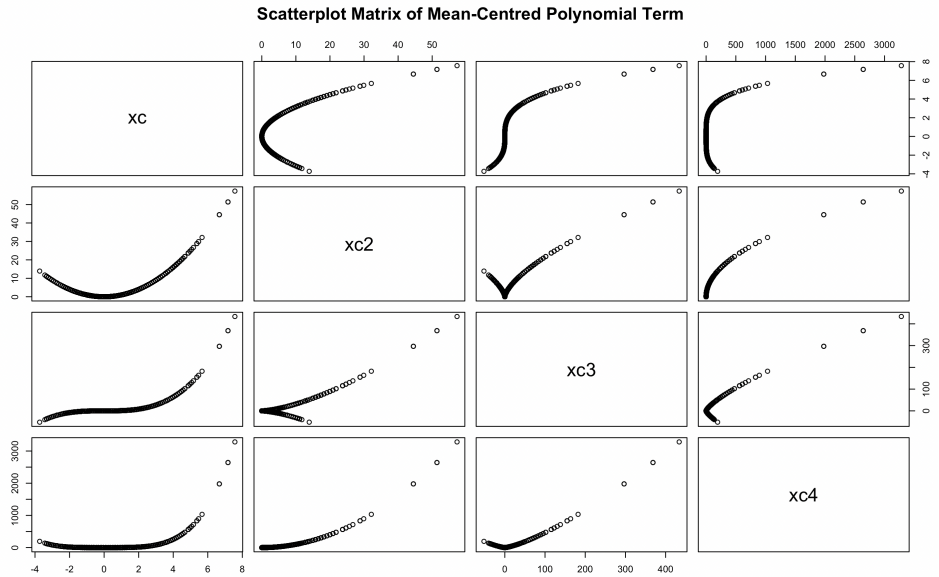
$$\bar{x} = 8.330414$$

Therefore, mean-centred variable is,

$$x_c = x - 8.330414$$

# Data Analysis : Checking for Multicollinearity

## Scatterplot Matrix of Mean-Centred Polynomial Terms




- U-shaped or S-shaped patterns appear.
- The regressors show reduced linear dependence compared with the uncentred polynomial terms

# Data Analysis : Checking for Multicollinearity

**Correlation Matrix of Mean-Centred Polynomial Terms**

| | $x_c$ | $x_c^2$ | $x_c^3$ | $x_c^4$ |
|---|---|---|---|---|
| $x_c$ | 1.000 | 0.551 | 0.699 | 0.467 |
| $x_c^2$ | 0.551 | 1.000 | **0.847** | **0.856** |
| $x_c^3$ | 0.699 | **0.847** | 1.000 | **0.934** |
| $x_c^4$ | 0.467 | **0.856** | **0.934** | 1.000 |

- Before centring, all six pairwise correlations exceeded 0.8. However, after centring, only three higher-order pairs remained above 0.8

# Data Analysis : Checking for Multicollinearity

**VIF of Mean-Centred Polynomial Term**

| Term | Uncentred VIF | Mean-Centred VIF |
|---|---|---|
| Linear term | 22,752.64 | 4.32 |
| Quadratic term | 200,324.21 | 4.19 |
| Cubic term | 207,149.52 | **25.27** |
| Fourth-order term | 25,121.24 | **19.33** |

- For VIF of $x_c$ and $x_c^2$ are below than 5. However, for the $x_c^3$ and $x_c^4$ still remain above 5, which are 25.27 and 19.33 but whole VIFs is already decrease dramatically
- Multicollinearity was reduced but not completely eliminated

# Data Analysis : Foward Selection with Mean-Centred Terms

Mean centring reduced multicollinearity, but did not completely eliminate it among higher-order terms. Therefore, forward selection is repeated using the mean-centred terms.

$$
\begin{aligned}
M_{1c}: &\quad pH = \beta_0 + \beta_1 x_c + \varepsilon, \\
M_{2c}: &\quad pH = \beta_0 + \beta_1 x_c + \beta_2 x_c^2 + \varepsilon, \\
M_{3c}: &\quad pH = \beta_0 + \beta_1 x_c + \beta_2 x_c^2 + \beta_3 x_c^3 + \varepsilon. \\
M_{4c}: &\quad pH = \beta_0 + \beta_1 x_c + \beta_2 x_c^2 + \beta_3 x_c^3 + \beta_4 x_c^4 + \varepsilon. \\
M_{5c}: &\quad pH = \beta_0 + \beta_1 x_c + \beta_2 x_c^2 + \beta_3 x_c^3 + \beta_4 x_c^4 + \beta_5 x_c^5 + \varepsilon.
\end{aligned}
$$

**Hypothesis :**

$$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0.$$

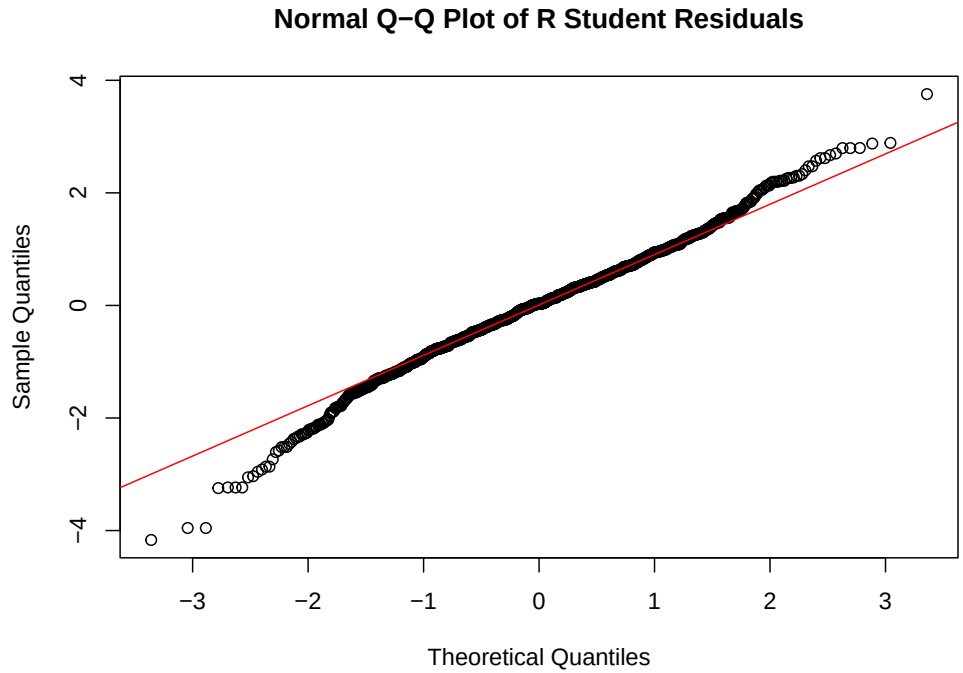
# Data Analysis : Forward Selection with Mean-Centred Terms

| **Comparison** | **Added term** | $F$ | $p$-**value** | **Decision** |
|---|---|---|---|---|
| $M_{1c}$ vs. $M_{2c}$ | $x_c^2$ | 95.597 | $< 2.2 \times 10^{-16}$ | **Retain** |
| $M_{2c}$ vs. $M_{3c}$ | $x_c^3$ | 41.988 | $1.309 \times 10^{-10}$ | **Retain** |
| $M_{3c}$ vs. $M_{4c}$ | $x_c^4$ | 4.6474 | 0.03129 | **Retain** |
| $M_{4c}$ vs. $M_{5c}$ | $x_c^5$ | 0.0212 | 0.8842 | **Stop** |

**Conclusion :** Since the $x_c^5$ term was not significant, the forward selection procedure stopped at the fifth-order comparison. Therefore, the centred fourth-order hierarchical polynomial model was selected, with $k = 4$.

# Data Analysis : Normality and Independence Assumptions

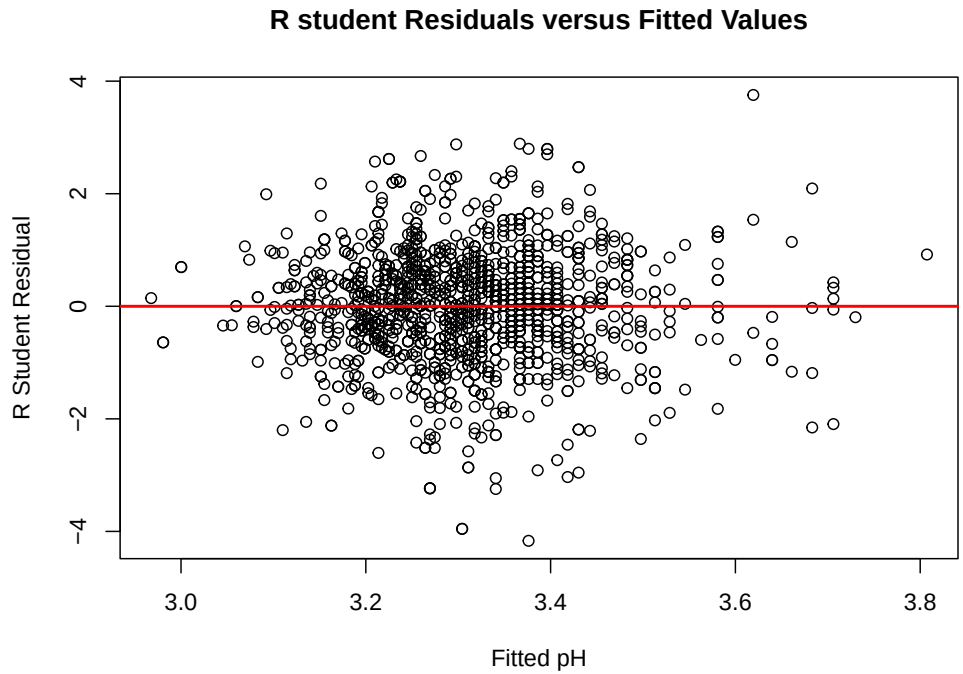
## Plot of Normal Q-Q plot of R Student Residuals




- The normality assumption is plausible since points appear to fall close to the reference line.
- We omit the time-sequence plot because the exact time of data collection was not recorded.
- We assume the independence assumption holds.

# Data Analysis : Linearity and Homoscedasticity Assumptions

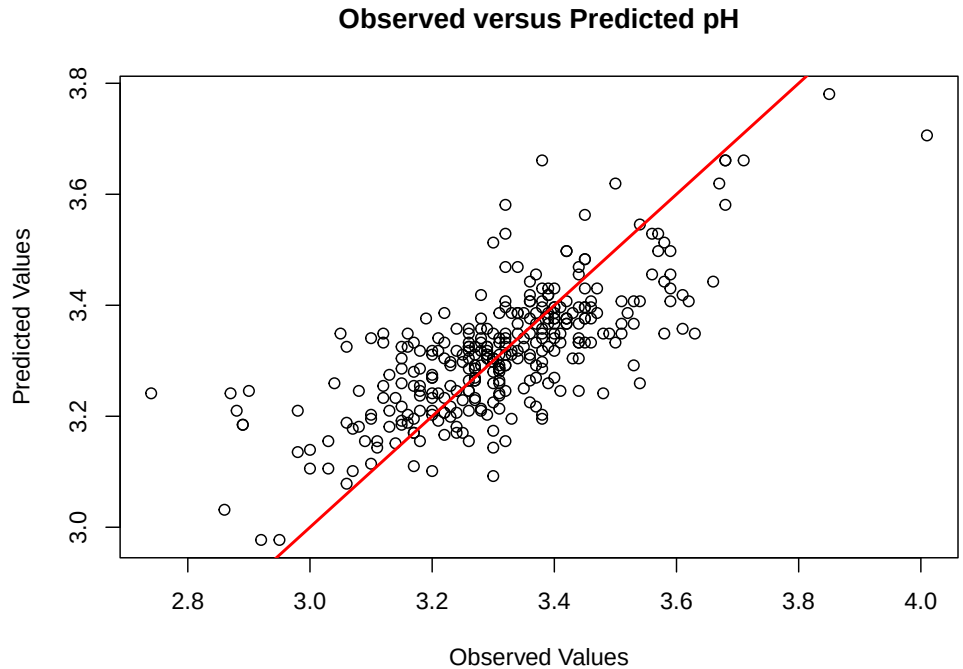
## Plot of R student Residuals versus Fitted Values




- The linearity assumption is not violated since plot distribute quite symmetric above and below the line: residual $= 0$.
- The homoscedasticity assumption is only a minor violation.

# Data Analysis : Model validation

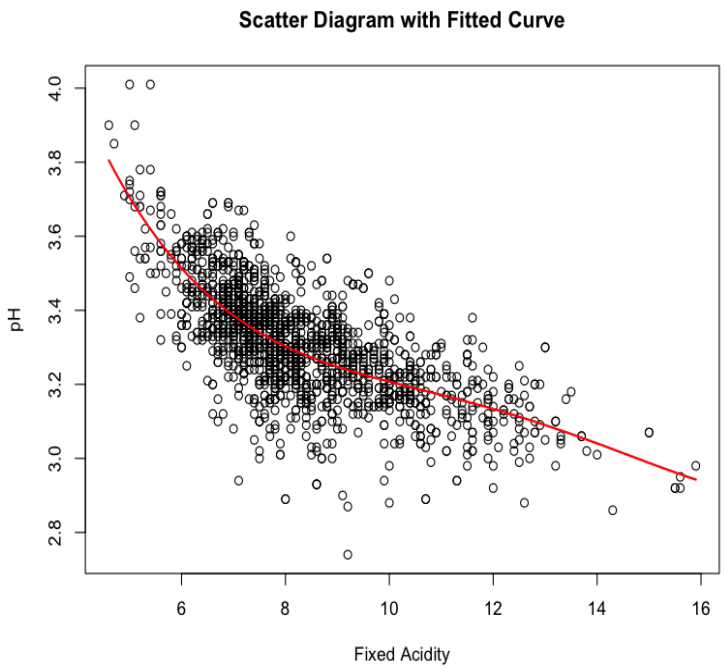
## Plot of Observed values versus Predicted values




- The plot is generated based on the prediction dataset.
- The red reference line is y=x.
- The data points align closely along this reference line with only few outliers.

# Data Analysis : Scatter diagram with the fitted curve

## Plot of Scatter Diagram with Fitted Curve




- The adopted model:

$$pH = \beta_0 + \beta_1 x_c + \beta_2 x_c^2 + \beta_3 x_c^3 + \beta_4 x_c^4.$$

- The fitted function:

$$\widehat{pH} = 3.2830 - 0.0586 x_c + 0.0116 x_c^2 - 0.0023 x_c^3 + 0.0001 x_c^4,$$

where $x_c = x - 8.330414$.

# Strengths and Weaknesses

**Strengths**

- The model successfully captures the nonlinear characteristics.
- The model is not overfitting.
- The model's structure and fitted values remain consistent after mean-centred.

**Weaknesses**

- Limitations of the model's applicability(Danger of Extrapolation).
- Multicollinearity of higher-order terms reduced but not completely eliminated.
- Tendency to over-predict and under-predict at the extreme ends of the observed data range.