Kerosene Merox Treater No. 2 (KMX-2) Salt Filter (D-4404) Salt Depletion Modeling via Stepwise Linear Regression (SLR), Principal Component Analysis (PCA) and Partial Least Squares (PLS)

Pascual, Ronald Patrick D.[1,2], Arellano, Roman Christopher T.[1,3], Pumatong, Jurrel D.[1,4]
[1]Process Engineering Department, Technical Services Division, Production A, Petron Bataan Refinery, Limay, Bataan, Philippines
[2]rdpascual@petron.com, [3]rtarellano@petron.com, [4]jdpumatong@petron.com

ABSTRACT

I. Introduction

Kerosene Merox Treater No.2 (KMX-2) aims to chemically treat kerosene to convert sulfur present as mercaptans to a less objectionable sulfur form which are disulfides. Part of the purification process is passing the treated kerosene to a fixed bed of salt to remove entrained water.

The level of salt in the salt filter (D-4404) shall be maintained at a critical level of 40% . Hence, salt depletion rate prediction is vital not only in stable operation but also in planning shutdown durations and salt topping activity requirements such as salt inventory and manpower.

Historically, salt depletion is based on typical or averaged values only, resulting to large deviations from the actual value. Thus, this study aims to explore linear regression techniques (SLR, PCA and PLS) in predicting salt depletion to eliminate this operational uncertainty.

The study is delimited to evenly distributed depletion rate per timeframe. Moreover, the study only focuses on linear regression. Non-linear and machine learning techniques are not covered.

II. Theoretical Framework

Multiple Linear Regression Model

The basic linear model assumes that there exists a linear relationship between two variables X and Y. This relationship is not perfect and distributed by some random error. For each value of X, say x, the corresponding Y-value is of the form:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where β_0 and β_1 are called as the intercept parameter and the slope parameter, respectively. If we have observed n values of x (i.e., $x_i, i = 1,2,\cdots,n$) with errors ε_i, then the resulting random variables $Y_1, Y_2, \cdots Y_n$ will be defined as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2,\cdots,n.$$

Similarly, the multiple linear regression model relates a response variable to p predictor variables. The model can be written as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where,
$y_i = i^{th}$ value of the response
$x_{ij} = i^{th}$ value of the j^{th} predictor variable
β_j = regression coefficient of the j^{th} predictor
ε_i = random error term, $i = 1,2,\cdots,n;\ j = 1,2,\cdots,p$.

Simple Linear Regression models, as discussed above, are prevalently misused due to failure of testing for model adequacy. That is, simple linear regression models can only be used when mathematical assumptions are met. Assumptions are enumerated and discussed briefly below:

1. *Linearity of the model*: The regression model is linear in the parameters. It is not necessary that Y and X are linearly related.

2. *Exact Measurement of the Covariates*: The covariate/s (predictor variable/s) must be recorded without measurement error. Otherwise, the interpretation of the error term will include not only the effect of unspecified predictor variables but also the systematic errors incurred in measuring X.

3. *Correct Specification of the Regression Model*: The regression model is correctly specified. This consists of developing the appropriate functional form of the model and selecting which variables to include.

4. *Zero Mean*: This critical assumption states that the mean of the error terms, ε_i, is equal to zero. This means that the influence of other factors not included in the model is essentially random.
$$E(\varepsilon_i) = 0$$

5. *Homoscedasticity or Constant Variance*: The variance of the error terms, ε_i, must be constant in relation to zero mean assumption.
$$Var(\varepsilon_i) = \sigma^2$$

6. *No Autocorrelation*: The error terms must be not correlated.
$$Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$$

7. *No Multicollinearity*: Multicollinearity between two or more covariates (predictor variables) means that the one should not be a linear function of another.

8. *Normality of Error Terms*: It is assumed that the error terms follow a normal distribution with zero mean and constant variance.
$$\varepsilon_i \sim N(0, \sigma^2)$$

Quantitative Data Transformations

Real data (complex with multiple predictors) often fail to meet normality and homoscedasticity criteria of simple linear regression model. To augment this, data analysts transform either the response variable, the predictor variable/s or both. Note that there is nothing illicit in transforming variables, however, the analysis of the results with transformed variables must be done cautiously.

There are several ways to conduct variance stabilizing transformations (to conform to $\varepsilon_i \sim N(0, \sigma^2)$ and consequently $Var(\varepsilon_i) = \sigma^2$) as summarized below.

Generally, for *right-skewed data* (tail is on the right or positively skewed), the appropriate transformations include $\sqrt{y}$, $\sqrt[3]{y}$, and $\ln(y)$. Conversely, for *left-skewed data* (tail is on the left or negatively skewed), the appropriate transformations include, y^2, $\sqrt{a-y}$, $\sqrt[3]{a-y}$, and $\ln(a-y)$ where a is an arbitrary constant.

1. Logarithmic Transformation

Holding all other factors constant, summarized below are the changes in interpretation once log transformation is employed.

Case	Regression Specification	Interpretation of β_1
linear-log	$Y = \beta_0 + \beta_1 \ln x + \varepsilon$	a percent change in x corresponds to $\ln(1.01) * \beta_1$ or $0.01\beta_1$ unit change in Y
log-linear	$\ln Y = \beta_0 + \beta_1 x + \varepsilon$	a unit change in x corresponds to $100 * (e^{\beta_1} - 1)$ or $100\beta_1$ percent change in Y
log-log	$\ln Y = \beta_0 + \beta_1 \ln x + \varepsilon$	a percent change in x corresponds to $100*(1.01^{\beta_1} - 1)$ percentage increase in Y

Note that because ln(0) is undefined, as is ln of any negative number, when using a log transformation, a constant should be added to all values to make them all positive before transformation.

2. Box-Cox Transformation

Box-Cox method considers a family of transformations on strictly positive response variables where the parameter, λ, is chosen by numerically maximizing the log-likelihood.

$$g_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Note that when λ approaches 1, there is essential no transformation. This does not change the variance nor make the model fit better. Otherwise, when λ approaches 0, the appropriate transformation is logarithmic.

3. Tukey's Ladder of Power Transformation

Unlike Box-Cox where the response variable is transformed to $\frac{y^\lambda - 1}{\lambda}$, Tukey's Ladder of Power, with the form y^λ, iteratively finds a λ that maximizes the W-statistic. In essence, this finds the power transformation that makes the data fit the normal distribution as closely as possible.

Qualitative Data Transformations

Dummy variables are used to allow the use of categorical variables as predictor variables in regression models. There are many instances where categorical variables are used as predictors, here are some:

a. The variable has no intrinsic quantitative characteristic (e.g. pass/fail, sex etc.)
b. The categorical responses are not measured numerically but instead by category (e.g. education attainment: elementary, undergraduate etc.)
c. The dummy variables are used as time identifiers for group data

A dummy variable, denoted by D is a dichotomous variable defined as:

$$D = \begin{cases} 1, & belongs\ to \\ 0, & otherwise \end{cases}$$

In general, if the variable has m categories, there should be $m - 1$ dummy variables created. An example is shown below.

$$Variable = \begin{cases} a \\ b \\ c \end{cases}$$

$$D_1 = \begin{cases} 1, & a \\ 0, & otherwise \end{cases} \qquad D_2 = \begin{cases} 1, & b \\ 0, & otherwise \end{cases}$$

	Dummy Variable	
category	D_1	D_2
a	1	0
b	0	1
c	0	0

Note that setting a D_3 where c is assigned with 1 will result to a multicollinearity issue ($D_3 = 1 - D_1 - D_2$). Hence it should be omitted.

(insert discussion on dummy variable and interaction interpretation)

Standardization

Data Standardization is the process of putting different variables on the same scale. This is primarily rooted in the problematic effects of variation in magnitude of predictors. For instance, if two variables with ranges of 0 to 10 and $1e10^5$ to $1e10^6$ are used in linear modeling, slight changes in

the latter will cause significant changes in the predicted response.

One way to scale data is via Z-score standardization where data are re-scaled so that the mean zero and the standard deviation is 1. Z-scores are computed by subtracting the mean the data and then dividing the standard deviation of the data to the observations.

$$Z_i = \frac{x_i - \mu}{\sigma}, \qquad i = 1,2,\cdots,n$$

Multicollinearity

Multicollinearity means that some of the predictors are linear factors of other fellow predictors. This phenomenon disrupts the precision of estimate coefficients, that is, even if the p-values are less than $\propto$ (which means that the predictor is statistically significant), the model produced will not be accurate. This poses difficulty in deciding which variables to include in the model.

Variance Inflation Factor (VIF) helps identify correlation between independent variables. Generally, VIFs larger than 5 are considered significant.

$$VIF_j = (1 - R_j^2)^{-1}$$

R_j^2 is the coefficient of determination of the regression model when the predictor j is predicted from all other predictors.

Structural multicollinearity (caused by the model and interaction of chosen predictors) is often fixed through standardization. However, if VIF is still > 5 after standardization, the multicollinearity is therefore caused by the data itself and omission of predictor variable/s with highest VIF/s may fix the problem.

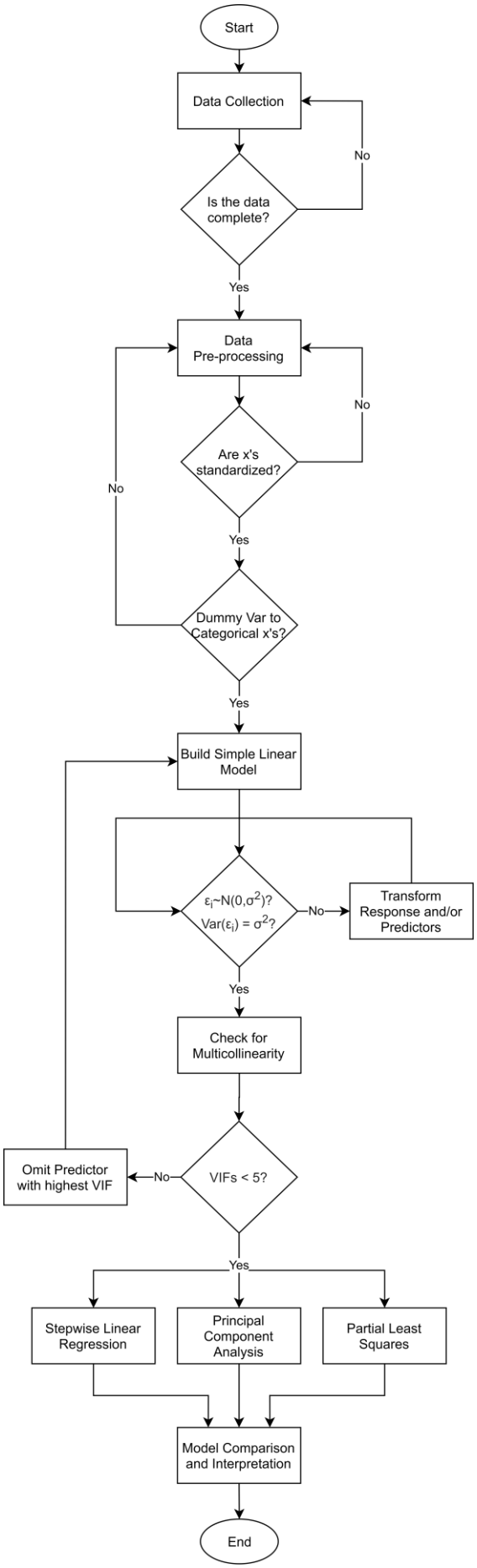
Stepwise Linear Regression

Principal Component Analysis

Partial Least Square Regression

III. Methodology

Modeling Process Flowchart



Data Collection

chika

Data Pre-processing

sdad

Zzzzzzzzz

Zz

Zzzzzzzzzzzzzzzzzzzzzzzzz

IV. Results and Discussion

- discuss the results
- show relevant plots
- show accuracy of prediction

V. Conclusions and Recommendations

- discuss conclusions
- recommend improvements to future engineers

VI. References

-list of all references in bibliography format

VII. Appendices

-add codes (discuss codes)