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 techniques. 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.

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 is zero and the standard deviation is 1. Z-scores are computed by subtracting the mean, μ, and then dividing the standard deviation, σ, of the data to the observations.

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

It is important to make sure that the μ and σ used to standardize shall be the same for both the training and test set.

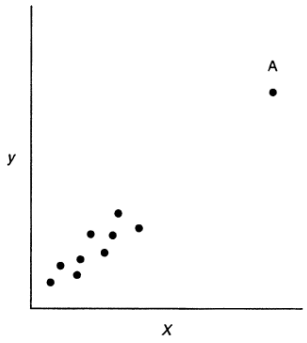
Outlier, Leverage and Measures of Influence

An *outlier* is a data point whose response, y, does not follow the general trend of the rest of the data. Conversely, a data point has *high leverage* if it has extreme predictor x values.

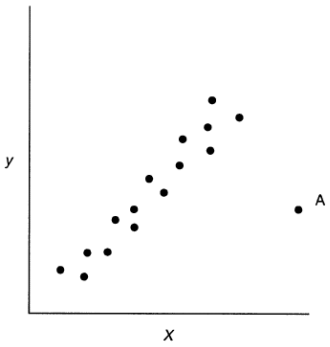
To discuss this further, note that leverage measures the distance of an observation from the center of the x-space. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values (e.g., with two predictors that are positively correlated, an unusual combination of predictor values might be a high value of one predictor paired with a low value of the other predictor)

A data point is *influential* if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. *Outliers and high leverage data points* have the potential to be influential. However, generally, an investigation shall be pursued to determine whether or not they are actually influential.

There are no general rules in choosing whether an influential data point or sets of influential data points is/are to be omitted. Hence, analysts often compare models with or without the influential data points in terms of coefficient of determination, R^2, changes in predictor coefficients and intercept and overall accuracy of prediction of the developed models. Shown below is an example of influential data point.



unusual x value
very little effect on regression coefficients



unusual x and y value
influential on regression coefficients

Sometimes,

Multicollinearity

Multicollinearity in linear regression 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 addition of interaction/s of involved predictor variables may fix the problem. Discussion on interaction evaluation will be specified in the next section.

Interactions

Inclusion of Interaction or Effect Modification in linear regression is done to take into account the multicollinearity of 2 or more predictor variables.

There are two questions to ask in deciding whether interaction terms are to be included in a linear model, namely: (1) Does the interaction make sense concepctually, or in other words, does it make sense that the effect of predictor variable 1 should depend on predictor variable 2?; and (2) Is the interaction term statistically significant, or whether or not the slopes of the regression lines significantly different.

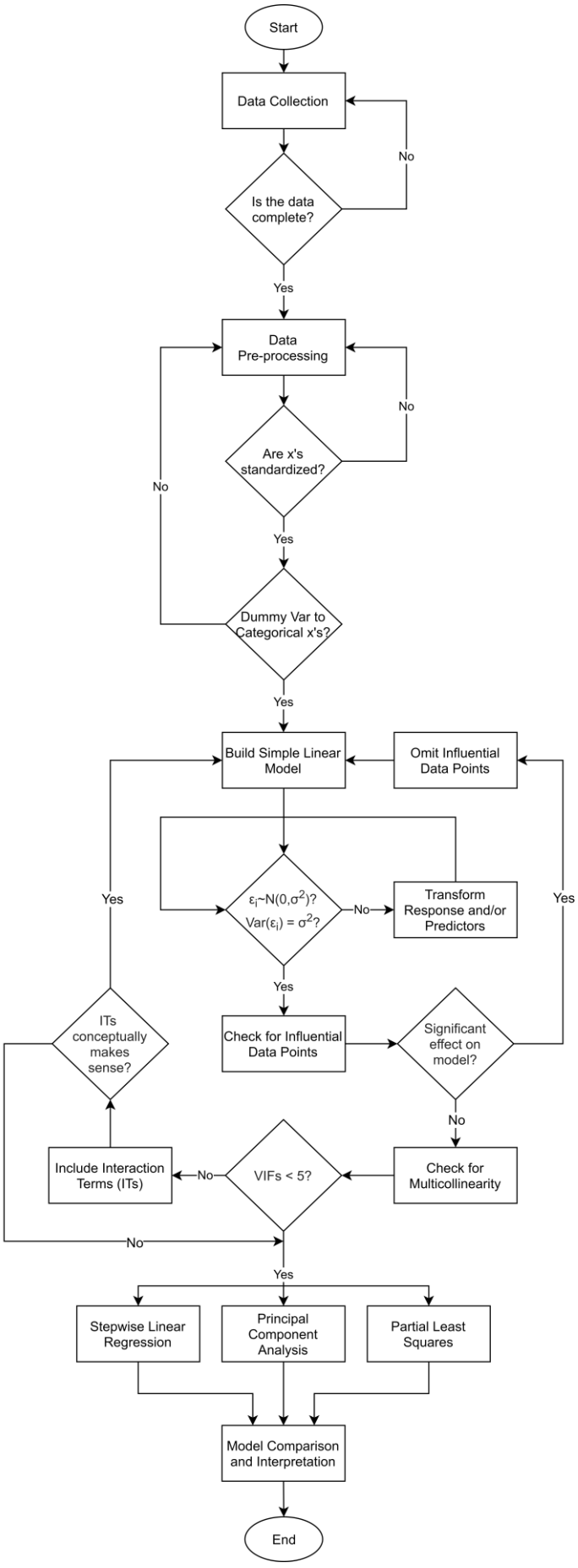
Stepwise Linear Regression

Principal Component Analysis

Partial Least Square Regression

III. Methodology

Linear Modeling Process Flowchart



The linear modeling process follows the flow chart as shown. Steps taken in this study for each section is discussed in detail below.

Data Collection

The data used in the study was sourced out from Honeywell Uniformance Process Studio, LabMate 5: Laboratory Information Mangement System and salt topping/changeout records from previous downstream engineers.

Relevant data were arranged in columns of MS Excel spreadsheet. Listed below are the data collected for this study.

Response Variable

D-4404 (Salt Filter) Salt Depletion Rate	-	mmpd	DepRate

Predictor Variables

Kerosene Feed	FI44005	MBSD	KerFeed
Kerosene Temperature	TI44002	°C	KerTemp
D-4403 (Water Wash) Level	LDIC44006	%	WaterLev
D-4406 (Drain Pot) Level	LDIC44008	%	BrineLev
D-4403 Water pH	-	-	WaterpH
D-4401 (Water Coalescer) Level	LDIC44002	%	WCBootLev
D-4402 (ECP) Inlet dP	PDIC44551	kPa(g)	ECPdP
D-4402 Level	LDIC44552	%	ECPLev
R-4401 (Reactor) Inlet Pressure	PI44005	kPa(g)	ReactorIP
R-4401 dP	PDI44004	kPa(a)	ReactordP
R-4401 Caustic Level	LDI44003	%	ReactorCauLev
D-4405 (Clay Filter) dP	PDI44006	kPa(a)	ClaydP
D-4405 Outlet Pressure	PIC44007	kPa(g)	ClayOP
Salt Density	-	kg/m^3	SaltDen
Salt Topping Timeframe	-	-	D_1 to D_8

Other Significant Data

Shipping Jet A1 Density	-	used in correcting kerosene feed flowrate recorded by FI44005
Historical Salt Topping Data and Densities	-	used to compute historical salt depletion rates of D-4404
D-4404 Dimensions	-	used to compute historical salt depletion rates of D-4404
D-4404 Moisture Differerential	-	used to compare depletion rates based on design/typical, based on moisture differential and predicted

Data Pre-processing

Data Pre-processing or data cleaning is done to either complete or omit sets with missing data, re-scale or standardize predictors with different units and ranges, or assign dummy variables to categorical predictors.

For this study, all predictors are standardized using Z-score standardization. Moreover, eight (8) dummy variables is assigned to nine (9) salt topping timeframes. Notice that salt depletion per timeframe is independent of each other, that is, measurement of total salt depleted resets to 0 after each salt topping.

	Salt Topping/Changeout Timeframes	*Actual DepRate*
t_1	December 2014 to October 1,2015	2780.00
t_2	October 2, 2015 to April 1, 2016	1190.00
t_3	April 2, 2016 to September 8, 2016	1945.00
t_4	September 9, 2016 to April 29, 2017	2447.00
t_5	April 30, 2017 to September 27, 2017	1240.00
t_6	September 28, 2017 to May 7, 2018	2061.00
t_7	May 8, 2018 to November 26, 2018	1987.80
t_8	November 27, 2018 to June 7, 2019	2260.00
t_9	June 8, 2019 to January 10, 2020	1201.00

$$Timeframe = \begin{cases} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{cases} \quad D_1 = \begin{cases} 1, & t_1 \\ 0, & otherwise \end{cases} \quad D_2 = \begin{cases} 1, & t_2 \\ 0, & otherwise \end{cases} \quad D_3 = \begin{cases} 1, & t_3 \\ 0, & otherwise \end{cases} \quad D_4 = \begin{cases} 1, & t_4 \\ 0, & otherwise \end{cases} \quad D_5 = \begin{cases} 1, & t_5 \\ 0, & otherwise \end{cases} \quad D_6 = \begin{cases} 1, & t_6 \\ 0, & otherwise \end{cases} \quad D_7 = \begin{cases} 1, & t_7 \\ 0, & otherwise \end{cases} \quad D_8 = \begin{cases} 1, & t_8 \\ 0, & otherwise \end{cases}$$

	Dummy Variable							
Category	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
t_1	1	0	0	0	0	0	0	0
t_2	0	1	0	0	0	0	0	0
t_3	0	0	1	0	0	0	0	0
t_4	0	0	0	1	0	0	0	0
t_5	0	0	0	0	1	0	0	0
t_6	0	0	0	0	0	1	0	0
t_7	0	0	0	0	0	0	1	0
t_8	0	0	0	0	0	0	0	1
t_9	0	0	0	0	0	0	0	0

Bases and assumptions for this study are summarized below:

1. Response variable, DepRate, is represented in millimeters of salt lost per day (mmpd).

2. Total actual salt depleted per timeframe was used to compute for actual salt depletion rate per day. The salt depletion was scaled against accumulated feed with the assumption that higher feed constitutes higher deleption rate. With this, shutdown periods are assumed to have minimal to zero depletion rates.

3. Shutdown periods in between timeframes are omitted from the population data set.

4. Kerosene feed flowrate measured by flow indicator FI44005 was corrected using averaged Jet-A1 shipping sample densities and kerosene temperature measured by TI44002.

5. Water wash vessel (D-4403) pH value ranges from 10-13 where values lowers per changeout. Since testing is not done everyday, previous values were copied until a new value comes up. For cases where pH is 0, values are replaced with 10 – signifying normal operation.

6. Moisture difference between salt filter's (D-4404) inlet and outlet is deemed unreliable in predicting salt depletion rate due to insufficient historical data.

7. Timeframe is treated as a categorical variable due to independence as discussed previously.

8. Predictor variables values are averaged per day. Moreover, standardization of the training set was done using the mean and standard deviation of the whole dataset from December 2014 to January 2020. Standardization for the test set is also done using the same mean and standard deviation.

`

[Simple Linear Modeling]

A
A
A
A
A
A
A

Aa
A

A
A
A
Aa

Aa
A
A
A
Aa
A
A
A
A
A
A
A

A
A
A
Aa

A

A
A
A
A
A
A
A
A
A
A

a

IV. Results and Discussion

Sets

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)