Simple & Multiple Linear Regression

Understanding Regression

Regression

- What is regression?
- Types of regression analysis
- Purpose of regression analysis
- Understanding Simple Linear Regression
- Understanding Multiple Linear Regression

What is regression?

- A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

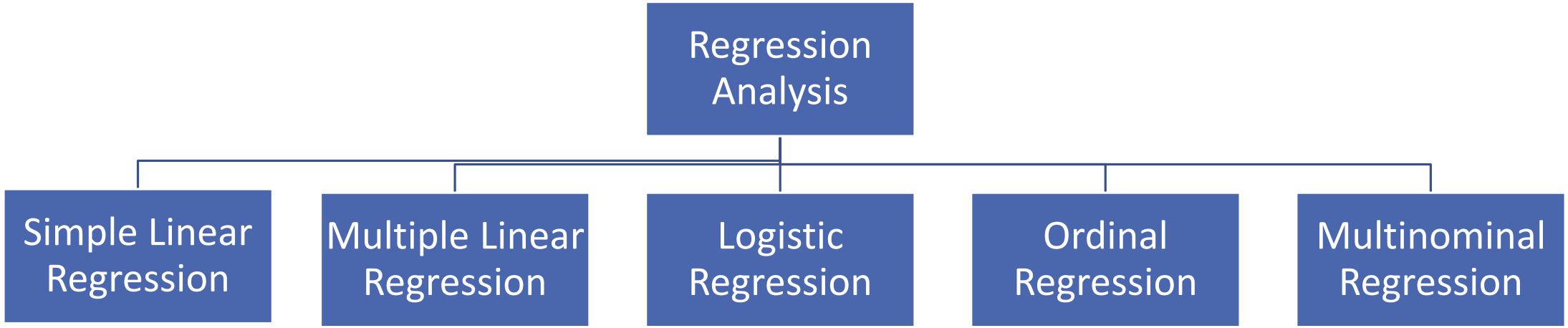
- It consists of 3 stages –

(1) analysing the correlation and directionality of the data, (correlation & covariance)

(2) estimating the model, i.e., fitting the line,

(3) evaluating the validity and usefulness of the model.

Types of regression analysis



1 DV - interval or ratio
1 IV - interval or ratio or dichotomous

1 DV - interval or ratio
1 IV - interval or ratio or dichotomous

1 DV - binary
2+ IV - interval or ratio or dichotomous

1 DV - ordinal
1 IV - nominal or dichotomous

1 DV – nominal
1+ IV - interval or ratio or dichotomous

Purpose of regression analysis

- The purpose of regression analysis is to analyse relationships among variables.
- Usually, the investigator seeks to ascertain the causal effect of one variable upon another. (causal analysis)
- The analysis is carried out through the estimation of a relationship

$$y = f(x1, x2,..., xk)$$

- The results serve the following two purposes:
  - Answer the question of how much y changes with changes in each of the Xs (x1, x2,...,xk). (forecasting an effect / impact of an effect)
  - Forecast or predict the value of y based on the values of the Xs. (trend forecasting)

Simple Linear Regression

- It is the simplest form with one dependent and one independent variable, is defined by the formula : -

$$y = a + b*x$$

Where,

y = estimated dependent,

a = intercept

b = regression coefficients

x = independent variable.

Multiple Linear Regression

- "Multiple regression" is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated.

- It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable.

Multiple Regression

Multiple Regression:

$$Y = a + b1X1 + b2X2 + B3X3 + \ldots + BtXt + u$$

Where:

Y= the variable that we are trying to predict(DV)

X= the variable that we are using to predict Y(IV)

a= the intercept

b= the slope (Coefficient of X1)

u= the regression residual (error term)

Typical Applications of Regression Analysis

Building of models to ascertain the pattern/behaviour of certain performance measures

- Asset Management : -
  - Predict performance
  - Maintenance cost
  - Asset at risk
  - Predict remnant life
  - Pump/compressor efficiency
  - Furnace efficiency
- Production operations : -
  - Proactive alerts
  - Heat exchanger – Heat duty prediction
  - Likelihood of Event (e.g. Trip, Failure,…)
  - Catalyst life prediction.

And many more….

What kind of data we needed?

- Dependent Variable(DV) needs to be measured on a continuous numerical scale
- Independent Variable(IV) can be continuous ,categorical or a mixture of two
- SAMPLE SIZE
  - 40 +K
  - 50+8K
  - 104+K

(Where K is the number of independent variables)

Regression :
Levels of measurement

- Dependent Variable(DV) = Continuous (Interval or Ratio)

- Independent Variable(IV) = Continuous or Dichotomous (may need to create dummy variables)

Assumptions in Multiple Regression Analysis

- Linearity of the phenomenon measured.
- Constant variance of the error terms.
- Independence of the error terms.
- Normality of the error term distribution.

Assumptions in Multiple Regression Analysis

- Errors (residuals) from the regression model:

$$e = (y - y\hat{})$$

- The model errors are independent and random
- The errors are normally distributed
- The mean of the errors is zero
- Errors have a constant variance

Covariance & Correlation

- **Covariance**: - It is a measure which helps to find out the direction of relationship between two variables.

  i.e. what happens to Y when X increases or decreases?

$$\text{cov}_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

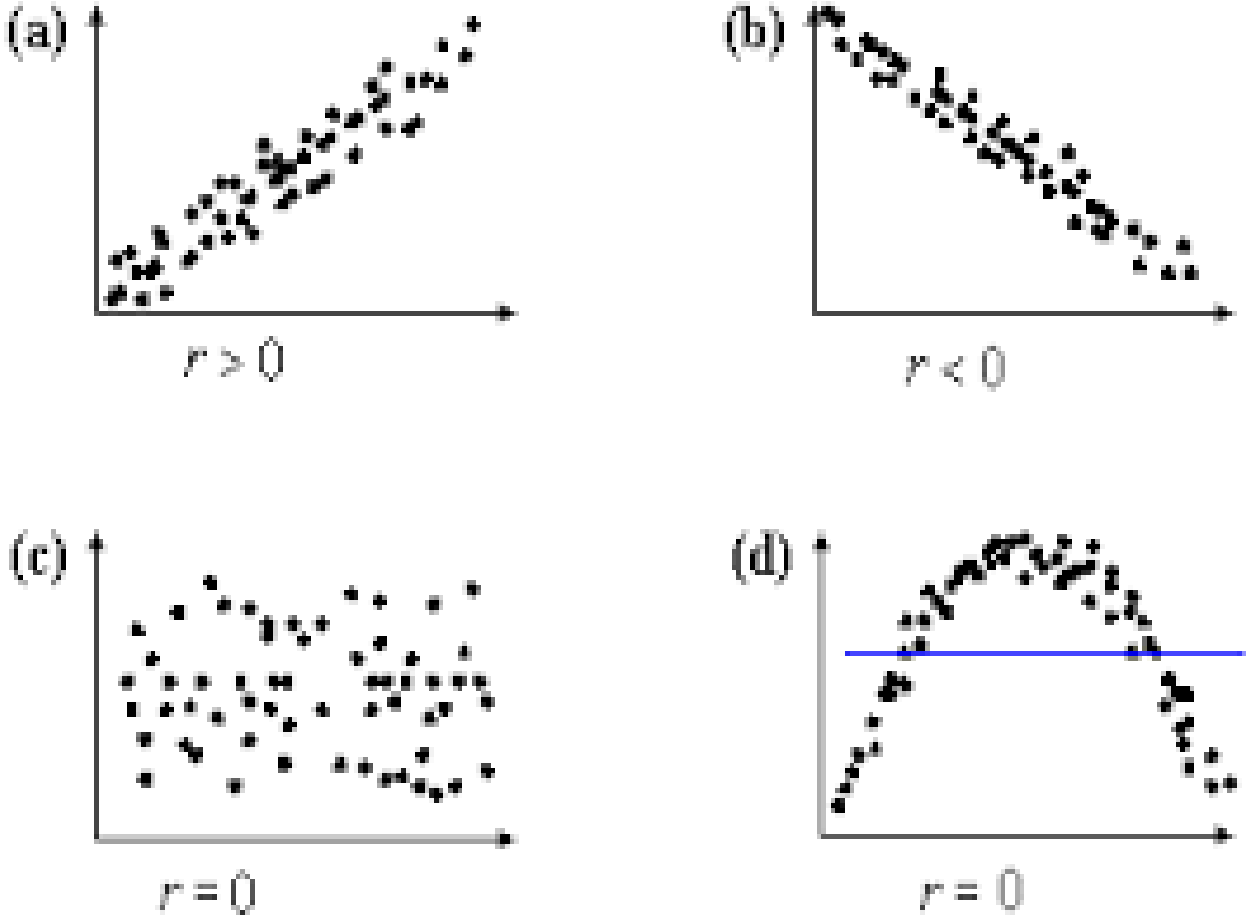
- **Correlation**: - Correlation signifies the strength of linear relationship. (It only captures linear relationship)

$$\rho_{xy} = \frac{\text{cov}_{xy}}{\sigma_x \sigma_y}$$

  - Population correlation is denoted by ρ. (rho)
  - Sample correlation is denoted by r.

Features of r (correlation)

- It is Unit free.
- Ranges between -1 and 1.
  a) The closer to 1, the stronger the positive linear relationship.
  b) The closer to -1, the stronger the negative linear relationship.
  c) The closer to 0, the weaker the linear relationship.



Testing the significance of the correlation coefficient

- Test whether the correlation between the population of two variables is equal to zero

  Null hypothesis, H_0: $r = 0$

- Assuming that the two populations are normally distributed, we can use a t-test to determine whether the null hypothesis should be rejected.
- The test statistic is computed using the sample correlation, r, with n – 2 degrees of freedom (df)

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

- Calculated test statistic is compared with the critical t-value for the appropriate degrees of freedom and level of significance
- Reject H_0 if $t > t_{critical}$ or $t < -t_{critical}$

R^2 (Coefficient of Determination)

defour analytics

❑ The **coefficient of determination** is the ratio of total variation in the dependent variable that is explained by variation in the independent variable.

$$R^2 = \frac{SSR}{SST} = \frac{sum\ of\ squares\ error\ explained\ by\ regression}{total\ sum\ of\ squares} \quad \text{where} \quad 0 \leq R^2 \leq 1$$

i.e. if R^2= 0.70 than 70% variance in Y is explained by X.

➢ For model to be good / acceptable value should be closer to 1.

Note: - In the single independent variable case, the coefficient of determination is

$$R^2 = r^2$$

Where:

R2 = Coefficient of determination
r = Simple correlation coefficient

R^2 (Coefficient of Determination)

Total variation is made up of two parts:

$$\mathbf{SST} = \mathbf{SSE} + \mathbf{RSS}$$

Total sum of Squares

Sum of Squared Errors

Regression Sum of Squares

Also known as

Square Sum of Regression SSR

$$SST = \sum (y - \bar{y})^2 \qquad SSE = \sum (y - \hat{y})^2 \qquad SSR = \sum (\hat{y} - \bar{y})^2$$

Where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

Term use in regression analysis

- **Explained variance** = R2 (coefficient of determination).
- **Unexplained variance** = residuals (error).
- **Adjusted R-Square** = reduces the R2 by taking into account the sample size and the number of independent variables in the regression model (It becomes smaller as we have fewer observations per independent variable).
- **Standard Error of the Estimate (SEE)** = a measure of the accuracy of the regression predictions. It estimates the variation of the dependent variable values around the regression line. It should get smaller as we

- **Total Sum of Squares (SST)** = total amount of variation that exists to be explained by the independent variables. TSS = the sum of SSE and SSR.
- **Sum of Squared Errors (SSE)** = the variance in the dependent variable not accounted for by the regression model = residual. The objective is to obtain the smallest possible sum of squared errors as a measure of prediction accuracy.
- **Sum of Squares Regression (SSR)** = the amount of improvement in explanation of the dependent variable attributable to the independent variables.

Coefficient of determination (R^2) and Adjusted R^2

- Coefficient of determination(R^2) can also be used to test the significance of the coefficients collectively apart from using F-test.

$$R^2 = \frac{SST - SSE}{SST} = \frac{RSS}{SST} = \frac{\text{Sum of Squares explained by regression}}{\text{Total Sum of Squares}}$$

- The drawback of using Coefficient of determination is that the value of the coefficient of determination always increases as the number of independent variables are increased even if the marginal contribution of the incoming variable is statistically insignificant.
- To take care of the above drawback, coefficient of determination is adjusted for the number of independent variables taken. This adjusted measure of coefficient of determination is called adjusted R^2
- Adjusted R^2 is given by the following formula:

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \times \left(1 - R^2\right) \right]$$

where

n = Number of Observations

k = Number of Independent Variables

R_a^2 = Adjusted R^2

Multicollinearity

❑ Significant problem faced in the Regression Analysis is when the independent variables or the linear combinations of the independent variables are correlated with each other.

❑ This correlation among the independent variables is called Multicollinearity which creates problems in conducting t-statistic for statistical significance.

❑High correlation among the independent variables suggests the presence of multicollinearity but lower values of correlations doesn't omit the chances of presence of multicollinearity.

❑ The most common method of correcting multicollinearity is by systematically removing the independent variable until multicollinearity is minimized.

- In R we use function called vif (variance inflation factor) to check if multicollinearity is present or not.
- Normally, if vif < 5 than no multicollinearity

  if vif > 5 than multicollinearity is present

CASE STUDY

Advance Analytics Case Study

HOW TO PREDICT FOULING FACTOR

IN

HEAT EXCHANGER

USING ADVANCE ANALYTICS ???

What is Heat Exchanger

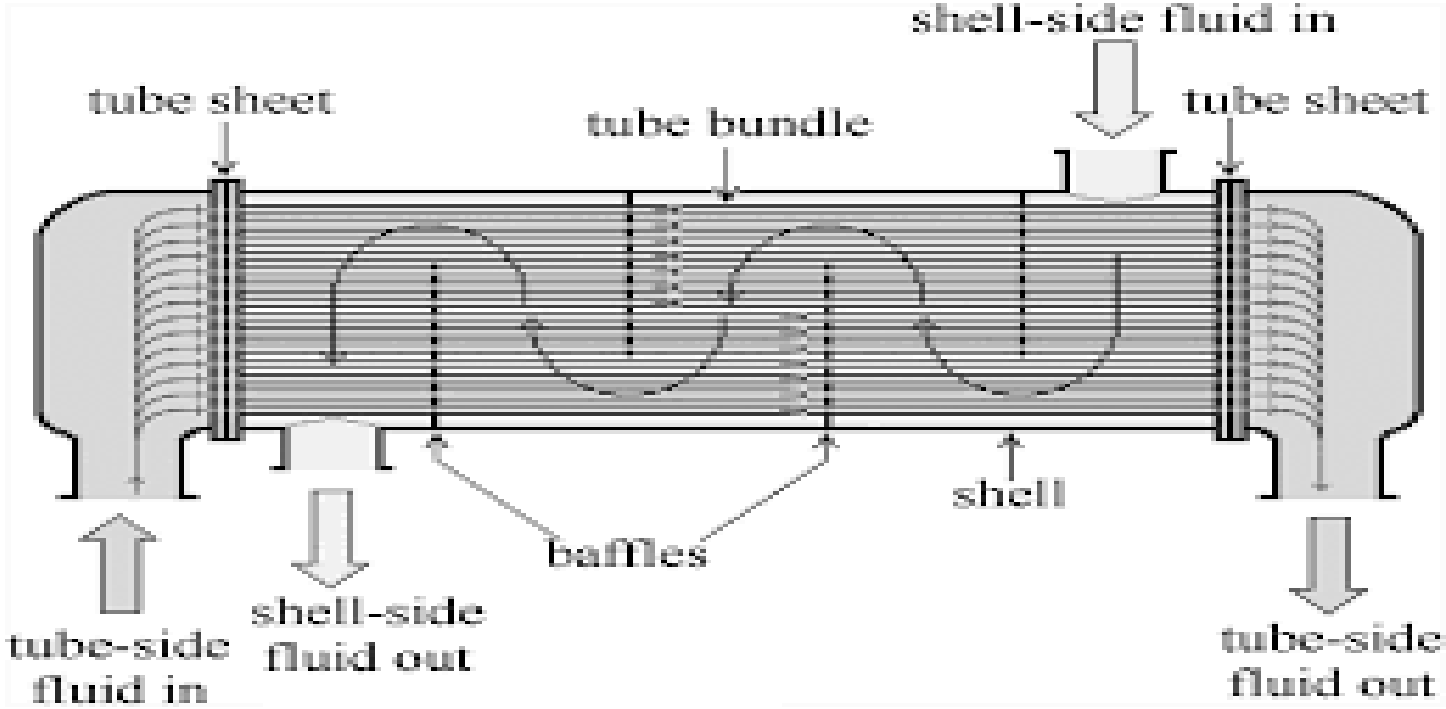
Definition of Heat Exchanger:-

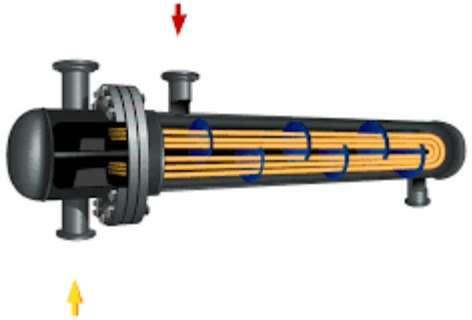
Heat exchanger are devices that facilitate the exchange of heat between two fluids that are at different temperatures while keeping them from mixing with each other.

Application:-

Heat exchangers are commonly used in practice in a wide range of applications, from heating and air conditioning systems in a household, to chemical processing, refinery and power production in large plants

Heat Exchanger- Images





Typical Operational Issues in Heat Exchanger

Exchanger Fouling

Fouling occurs when any type of particles both organic or inorganic plug or plate out on heat transfer surfaces creating a resistance to transfer energy

Corrosion:-

Severe corrosion can and does occur in tubing and very often with common fluids such as water.

Vibration:-

Vibration of the tubes as a result of the flow of the shell side past them is important phenomena specially when the H.X size and flow quantities of flow are increased

How to manage & ensure availability of Heat exchanger in case of fouling

- Predictive modelling of fouling factor by using advance analytics
- Proactive planning of maintenance to reduce its impact in productivity.
- Maintain a proper heat duty for efficient operation.

Advance analytics method to predict fouling factor

The crude preheat train (CPT) in a petroleum refinery consists of a set of large heat exchangers which recovers the waste heat from product streams to preheat the crude oil.

In these exchangers the overall heat transfer coefficient reduces significantly during operation due to fouling. The rate of fouling is highly dependent on the properties of the crude blends being processed as well as the operating temperature and flow conditions.

The objective is to develop a predictive modelling using advance analytics methods which can *a priori* predict the rate of the fouling and the decrease in heat transfer efficiency in a heat exchanger

Prediction of fouling factor

Using Multiple Linear Regression

Stepwise description of actions to be taken on data

1. When data is received prepare data dictionary or ask concerned person to provide it.
2. Understanding Data: -
   - Variable identification (DV, IV, Identifying data types)
   - DV exploration
     - Distribution analysis
     - Outlier treatment
   - IV exploration
     - Bivariate analysis
     - Variable transformation (if any) (Dummy variables creation, grouping of distinct values, mathematical transformation i.e. log, splines etc)

Stepwise description of actions to be taken on data

4. Fitting the regression
   - Variable selection
     - Check for the most suitable transformed variable
     - Select the transformation giving the best fit
     - Reject the statistically insignificant variables
5. Fitting the regression
   - Analysis of results
   - Model comparison
   - Model performance check
     - R^2
     - Actual vs Predicted comparison

R script

```
setwd("C://R")                                  #Set Working Directory
he=read.csv("htex.csv")                         #Read the data file
View(he)                                        #Check what is there in the file
head(he)                                        #Check if the data is populated/imported properly
tail(he)
summary(he)                                     #Check the summary of the file
plot(he$Fouling.Resistance)                     #Generate plot of Dependent variable (Fouling Resistance)
quantile(he$Fouling.Resistance, c(0,0.05,0.1,0.25,0.5,0.75,0.90,0.95,0.99,0.995,1)) #Check the quantile to find out the outlier limit
which(he$Fouling.Resistance>=0.010935163, arr.ind=TRUE)  #To find the row number for specific value
he1=he[-c(45:48,50:55,181),]                    #To delete rows
View(he1)                                       #View the dataset to check rows are been deleted
plot(he1$Fouling.Resistance)
library(ggplot2)
```

R script

```
h=ggplot(he1,aes(Crude.Temp.In,Fouling.Resistance))+geom_line(color="blue")

plot(h)

FitLinReg=lm(Fouling.Resistance~Crude.Temp.In+Crude.Temp.Out+CDU1.rate....KBPD.+Crude.Flow.rate..kg.hr.+Kero.Temp.In+
Kero.Temp.Out+Kero.Flow.m3.hr+Furnace.inlet.temp..deg.C.+Crude.Temp.Increase+Kero.Temp.decrease+Q..heat.exchanged.
.MW+Hot.in...Cold.out+Hot.out..Cold.in+LMTD+U.transfer.rate+Cummulative.Flow.Tones.per.day,data=he1)

summary(FitLinReg)

FitLinReg=lm(Fouling.Resistance~Crude.Temp.In+Crude.Temp.Out+Crude.Temp.Increase+Q..heat.exchanged..MW+LMTD+Cu
mmulative.Flow.Tones.per.day,data=he1)

summary(FitLinReg)

library(car)

vif(FitLinReg)

cti=lm(Fouling.Resistance~Crude.Temp.Increase,data=he1)

q=lm(Fouling.Resistance~Q..heat.exchanged..MW,data=he1)

summary(cti)

summary(q)
```

R script

```
FitLinReg=lm(Fouling.Resistance~Crude.Temp.In+Crude.Temp.Out+Q..heat.exchanged..MW+LMTD+Cummulative.Flow.Tones.per.day,data=he1)

summary(FitLinReg)

FitLinReg=lm(Fouling.Resistance~Q..heat.exchanged..MW+LMTD+Cummulative.Flow.Tones.per.day,data=he1)

summary(FitLinReg)

 vif(FitLinReg)

pred = predict(FitLinReg,type="response")

Pred

  # # write.csv(pred,"output_fitlinreg.csv", row.names = F)

   ## library("sandwich")

  ## vcovHC(FitLinReg,omega=NULL, type="HC4")

  ## install.packages("lmtest")

  ## library("lmtest")

  ## coeftest(FitLinReg,df=Inf,vcov=vcovHC(FitLinReg,type="HC4"))
```

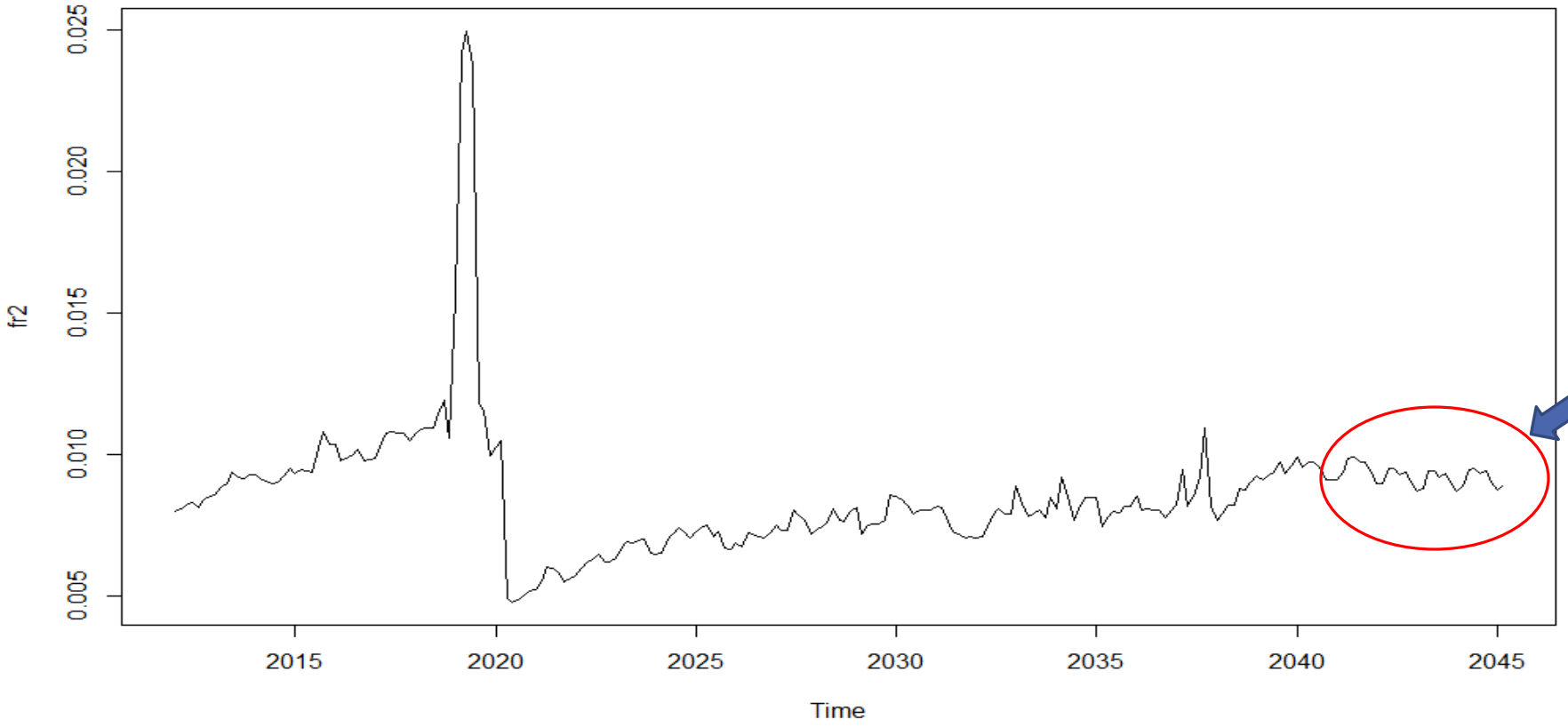
R script

```
he2=he1[,-c(2:11,13,14,16)]

View(he2)

he2$predicted = predict(FitLinReg,type="response")

write.csv(he2,"output_htex.csv", row.names = F)
```

Forecasting fouling factor

- Now we have linear regression model / equation.
- In order to forecast the trend of fouling factor we will have to forecast all the real time values (Machine readings) / variables using time series.
- Calculate necessary variable's value through formula.
- Substitute that values in the regression model / equation to get the predicted fouling factor trend.

GRAPH (Predicted fouling resistance)



Cyclic nature is predicted which is doubtful.

How model will be used daily ?

- Real time data through existing monitoring systems will get captured into the regression model which will display the forecasted values and graphs of the fouling factor and it's respective significant variables with respect to date selection basis.
- This will enable the end user to check the performance of the exchanger based on the predicted values of fouling factor as well as respective significant variables.
- Significant variables with respect to this case study are : -
  - Kero flow
  - Heat exchanged
  - LMTD
  - Cumulative flow tonnes per day

  For instance if we observe heat exchanged value is deviating from the standard norms we can take corrective measures in time than to wait till the scheduled maintainance / unscheduled breakdown.
- Also the prediction values keeps on changing everyday according to that day's circumstances / performance.

Design Thinking

Ask the participants what kind of problems related to their industry can be solved using this concept.

Question & Answers - students

defour analytics

THANK YOU