DATA ANALYSIS PROJECT
PRESIDENCY UNIVERSITY

ANKAN CHAKRABORTY

January 16, 2020

1 INTRODUCTION

In brief, extrusion is a process in which dough-like raw material is pushed through a machine and the machine puts out product in desired form, followed by some finishing touches. One important characteristic of an item produced is its weight. If weight is too low, product may be weak. If weight is too high, it may mean wastage of raw material. Hence a manufacturer is keen to know the relation between product weight and various parameters of manufacturing process. In a particular factory 3 parameters likely to affect weight were monitored. These were Extruder RPM, current and conveyer speed. My project is to analyse the given data and find whether there is any significant dependendence of product weight on the other 3 parameters in the light of given data and make a justifiable inference with the help of statistical tools, techniques and notions.

2 DATA ANALYSIS

2.1 EXPLORATORY DATA ANALYSIS :

At first let us read the data and observe the content.

R Chunk - 2.1.1

```
  WEIGHT EXTRUDER.RPM CURRENT ConveyerSpeed
1   4.46           60      69            70
2   4.46           59      70            69
```

So, we can observe there are 4 variable columns such as :-

1. **"WEIGHT"** - which is **weight of product** (contains numeric elements)
2. **"EXTRUDER.RPM"** - which is **extruder speed [RPM- revolutions per minute]** (contains integer elements)
3. **"CURRENT**" - (contains integer elements)
4. **"ConveyerSpeed"** - (contains integer elements)

Now let us draw each scatterplot of weight and 3 other parameters separately and find out the individual correlations.

R Chunk - 2.1.2

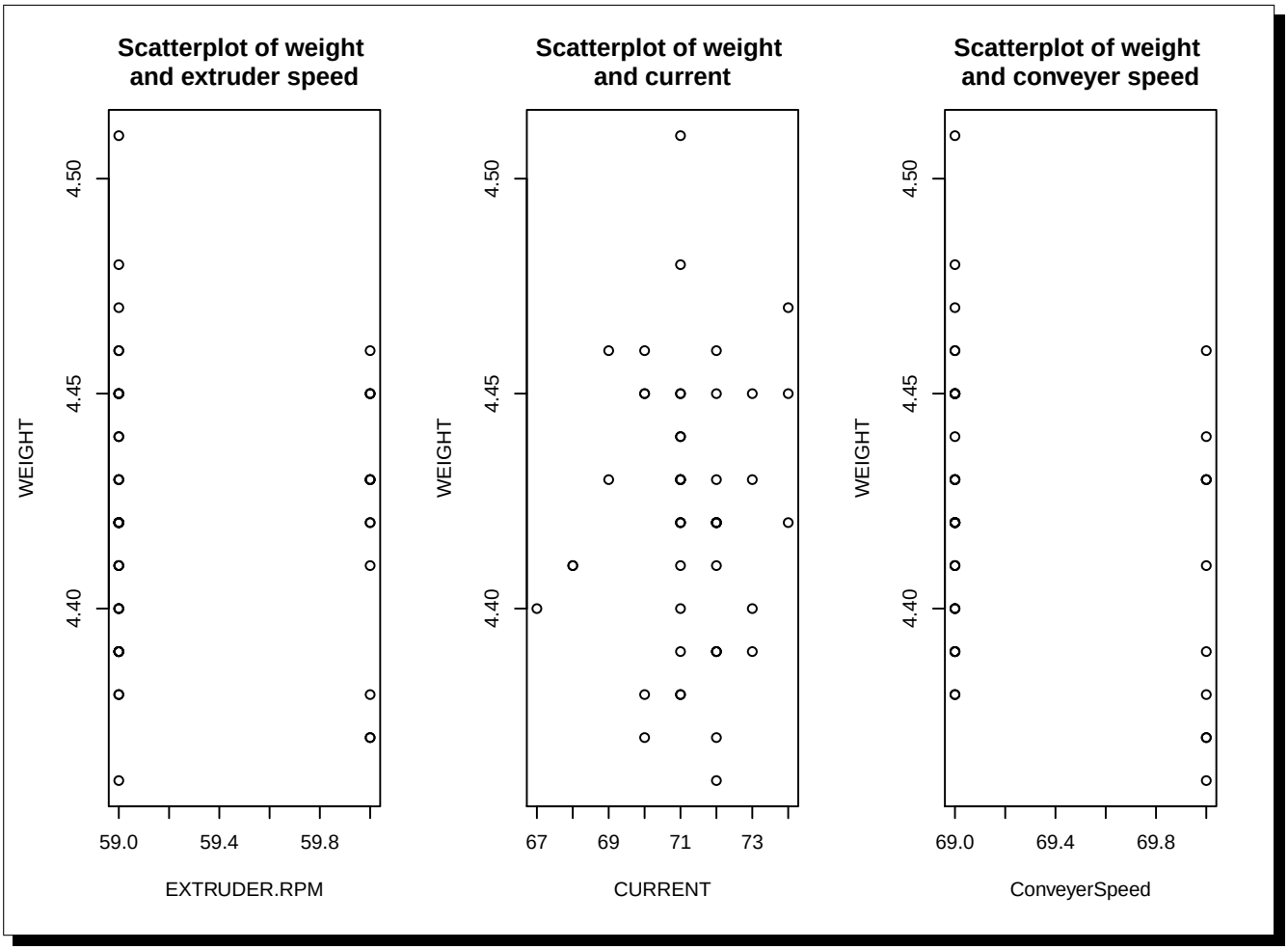

Fig:-2.1.1

From the above scatterplots we can observe that there is probably no such linear dependence between the variables for each plot. Let's see and verify by finding correlation.

R Chunk - 2.1.3

```
The correlation between weight and extruder speed is -
[1] -0.0168408
The correlation between weight and current is -
[1] 0.04313673
The correlation between weight and conveyer speed is -
[1] -0.2787137
```

Here we see that for each case the correlation between the variables are quite small or nearly about 0 i.e. there doesn't exist any linear relationship between the variables. So now one can think off going for **multiple regression** whether there is simultaneous dependence of product weight on the other 3 parameters.

- ## Multiple Regression :-

Now, let us construct the multiple regression model. Let y, x_1, x_2 and x_3 denote weight of product, extruder speed, current flown and conveyer speed respectively. So, our **multiple regression model** is -

$$y = a + bx_1 + cx_2 + dx_3 + \epsilon$$

,where ϵ is the error term in multiple regression model and it is assumed that $E(\epsilon) = 0$, thus our **multiple regression equation** becomes as following -

$$E(y) = a + bx_1 + cx_2 + dx_3$$

and **estimated multiple regression equation** is -

$$\hat{y} = a_1 + b_1x_1 + c_1x_2 + d_1x_3$$

,where a_1, b_1, c_1, d_1 are estimates of a, b, c, d respectively, $\hat{y}$ =predicted value of weight of product (dependent variable).

R Chunk - 2.1.4

```
Call:
lm(formula = WEIGHT ~ ., data = data)

Coefficients:
  (Intercept)   EXTRUDER.RPM        CURRENT  ConveyerSpeed
    5.5749207      0.0083687      0.0002617     -0.0240877
```

$$\hat{y} = 5.5749207 + 0.0083687x_1 + 0.0002617x_2 - 0.0240877x_3$$

At this point we have evaluated estimated multiple regression equation. But we don't know whether multiple regression model fits our data or not. Thus we do summary measure on our model.

R Chunk - 2.1.5

```
Call:
lm(formula = WEIGHT ~ ., data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.045197 -0.024151 -0.005197  0.020522  0.084803

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.5749207  0.8945421   6.232 1.41e-07 ***
EXTRUDER.RPM   0.0083687  0.0108561   0.771   0.4448
CURRENT        0.0002617  0.0031629   0.083   0.9344
ConveyerSpeed -0.0240877  0.0115526  -2.085   0.0428 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03052 on 45 degrees of freedom
Multiple R-squared:  0.08982,Adjusted R-squared:  0.02915
F-statistic:  1.48 on 3 and 45 DF,  p-value: 0.2326
```

Now, **multiple R^2in regression** is the **coefficient of determination**, which can be interpreted as **the percentage of variance in the dependent variable that can be explained by the predictors.** And the value of multiple R^2 in regression lies between 0 and 1. For our regression model, the value of **multiple R^2 is = 0.08982**, which is very close to 0 rather than 1. Thus from here we can say that our multiple regression model is unable to explain the data in a better way. In addition to this, we will perform more analysis such as residual analysis, identification of outliers to justify the above claim.

- **Residual Analysis and identification of outliers :-**

Residual analysis consists of two primary things. One is it tells us how good the model we have produced fits the data we are looking at or in another words how is are error; is are error large or small ? And the number two which is most importantly, whether or not the model we are using is appropriate to the data we are looking at. As we know there are many many ways to model a data set and certain models are more appropriate than the others. And residuals can help us to side that.

The **residuals** are defined as **the difference of predicted value from the observe value of a certain variable under interest** i.e. -

$$residuals = observed\ value - predicted\ value$$

$$i.e.\ residuals = y_i - \hat{y}_i\ for\ all\ i = 1(1)n$$

, n =number of observed data under the variable of interest.

Let us perform this in R code and plot the residuals and standard residuals separately against the predicted values $\hat{y}_i$.

R Chunk - 2.1.6

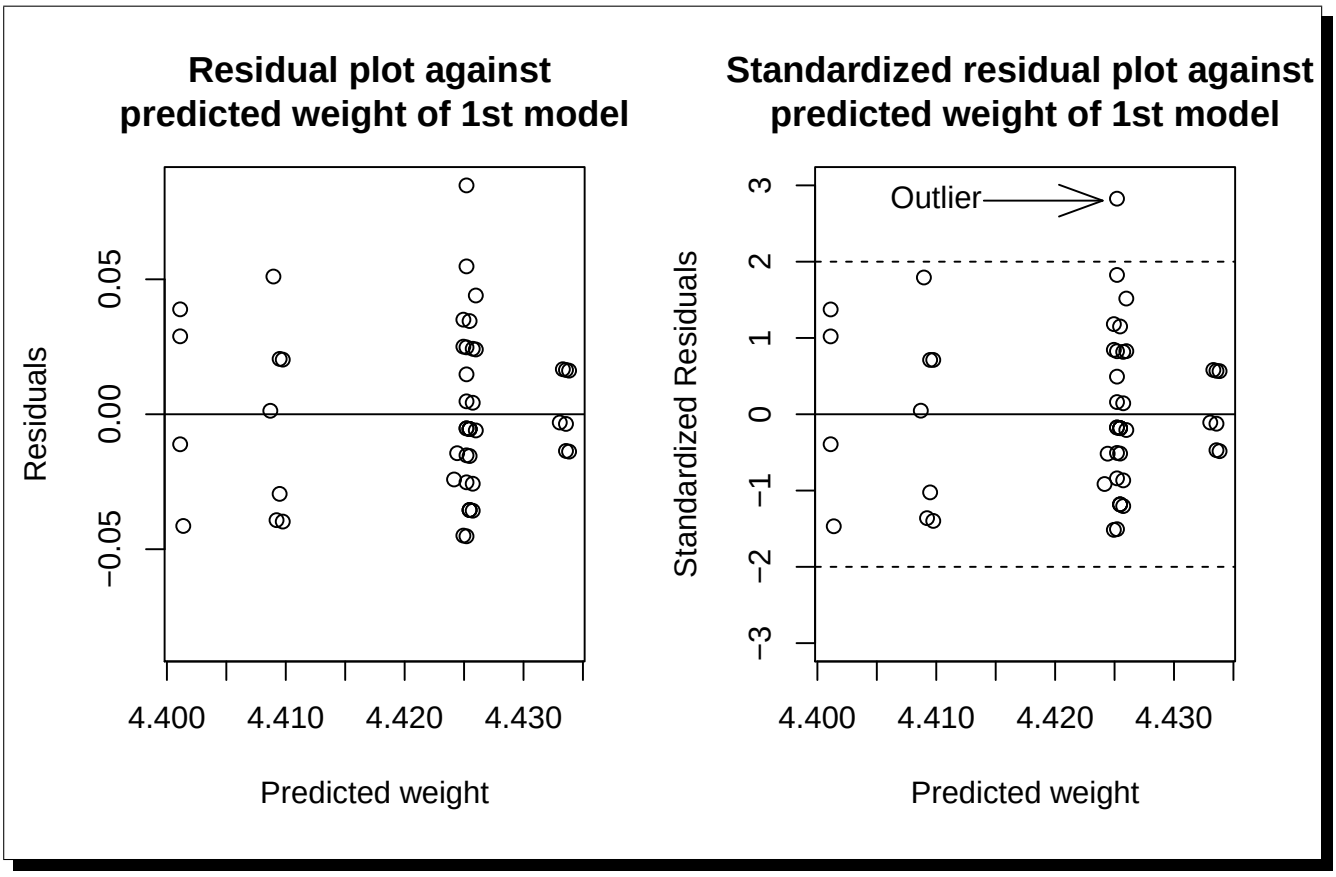

Fig:-2.1.2

Rule of Thumb for Interpreting Standardized Residuals :-

- If the standard residual is less than -2, then the observed value is less than the expected value.
- If the standard residual is greater than $+2$, then the observed value is greater than the expected value.
- So, combining above two situations, if for any observation **the absolute value of standard residual is greater than $+2$ then those observations are considered to be outliers**, which may seriously affect any statistical model.

Here, we can see from the scatterplot that a residual value is greater than $+2$, so we can consider the observation corresponding to that residual value is an outlier. We can explicitly find out that outlying observation by running a simple chunk of R code.

R Chunk - 2.1.6

```
(1) The absolute value of residual(s), which is(are) greater than 2 is(are) -
[1] 2.826174

(2) The outlying observation(s) is(are) -
  WEIGHT EXTRUDER.RPM CURRENT ConveyerSpeed
8   4.51           59      71            69
```

So, removing that outlier from 1st regression model we will again fit another multiple regression model(say, 2nd model) and repeat our analysis.

R Chunk - 2.1.7

```
Call:
lm(formula = WEIGHT ~ ., data = data_1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.042268 -0.019998 -0.002985  0.018980  0.057732

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.2905645  0.8256195   6.408 8.47e-08 ***
EXTRUDER.RPM   0.0106772  0.0099850   1.069   0.2908
CURRENT        0.0007567  0.0029054   0.260   0.7957
ConveyerSpeed -0.0224924  0.0106083  -2.120   0.0397 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02799 on 44 degrees of freedom
Multiple R-squared:  0.09728,Adjusted R-squared:  0.03573
F-statistic:  1.58 on 3 and 44 DF,  p-value: 0.2076
```

Thus our **estimated multiple regression equation of 2nd model** is -

$$\hat{y} = 5.2905645 + 0.0106772x_1 + 0.0007567x_2 - 0.0224924x_3$$

R Chunk - 2.1.8

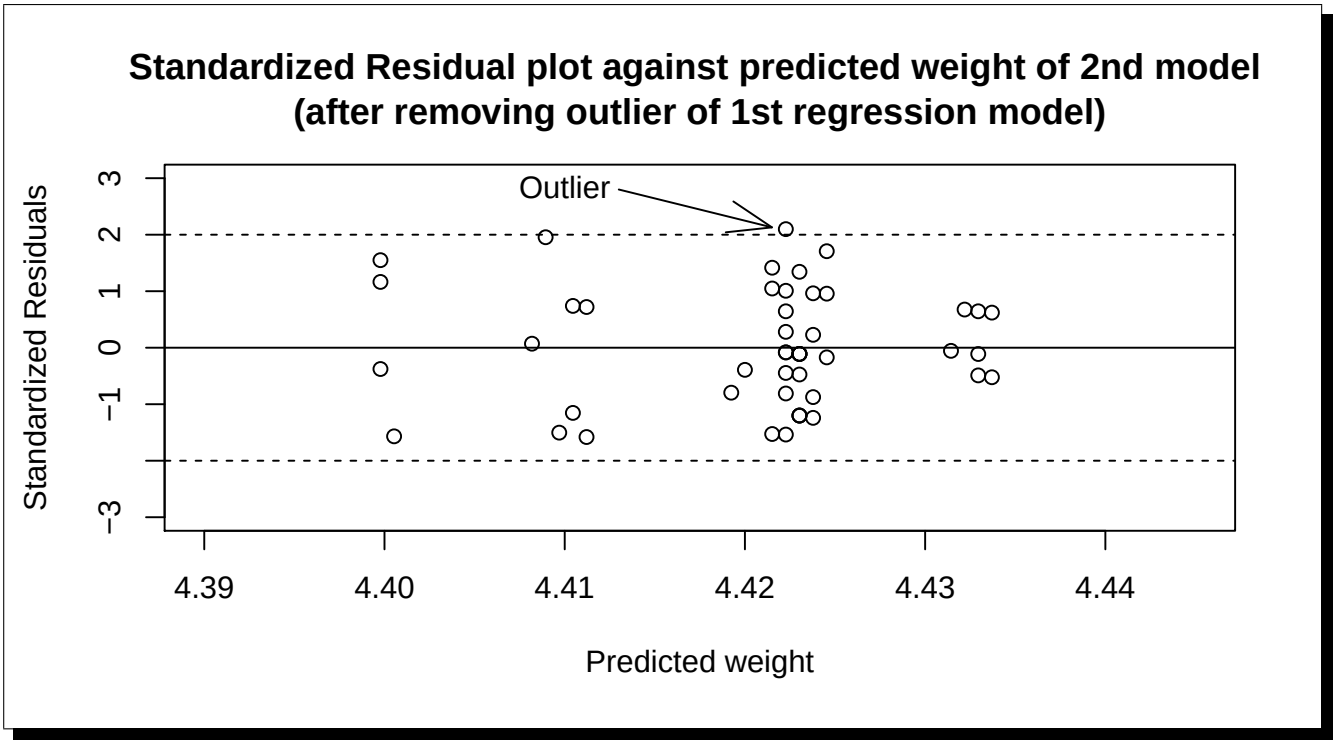

Fig:-2.1.3

And moreover our new fitted 2nd regression model has again outlier. So, removing outlier we fit another regresion model(3rd model).

R Chunk - 2.1.9

```
Call:
lm(formula = WEIGHT ~ ., data = data_2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041476 -0.019249 -0.001535  0.019293  0.049727

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.292e+00  8.564e-01   6.180 2.19e-07 ***
EXTRUDER.RPM   1.103e-02  9.852e-03   1.120   0.2692
CURRENT        5.907e-05  3.207e-03   0.018   0.9854
ConveyerSpeed -2.212e-02  1.032e-02  -2.144   0.0379 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02703 on 42 degrees of freedom
Multiple R-squared:  0.1024,Adjusted R-squared:  0.03832
F-statistic: 1.598 on 3 and 42 DF,  p-value: 0.2042
```

Thus our **estimated multiple regression equation of 3rd model** is -

$$\hat{y} = 5.090054 + 0.012305x_1 + 0.001106x_2 - 0.021367x_3$$

And now let us concentrate whether we can observe any pattern or shape of the standardized residual scatterplot of 3rd regression model.

R Chunk - 2.1.10

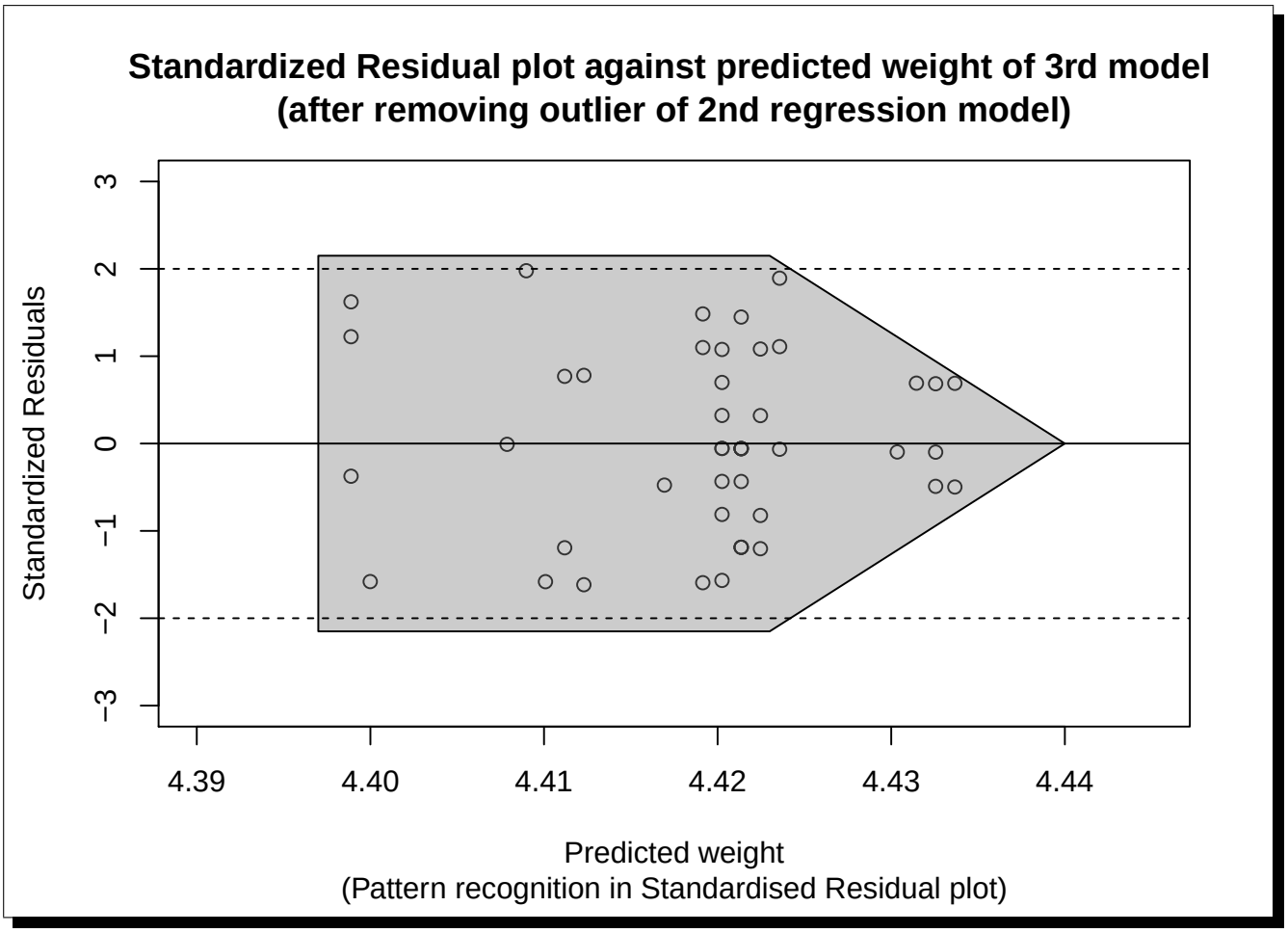

Fig:-2.1.4

Here we observe a horizontal cylindrical pattern at the left and cone shaped pattern at the right. As residuals are difference of predicted value from observed value, thus error is larger at left side and relatively smaller at the right. So the residuals are relatively much more spreaded at the left than the right. Thus the spread is not constant throughout the region. This is called **heteroscedasticity** or **non-constant variance**.

2.2 INFERENTIAL DATA ANALYSIS :

So far we have done a lot of exploratory data analysis. Now it's time to infer and verify our claims or findings based on the idea acquired from exploratory data analysis. Here we have used "lm()" function to fit a multiple linear regression model. At first we fitted our first regression model. And after analysing the **residuals** (i.e. *obseved value−predicted value*) and **standard residual plots** against predicted values of the model, we could identify outliers with respect to our fitted model and removing outliers again fitted new second regression model. And this process of removing

outlier and modifing regression model is continued until all the outliers of corresponding fitted models are removed completely. Thus aagain we got outlier from second model and fitted our third model. We see there is no outlier in this finally fitted third model. And for each modification of model we can see some changes or increase in **multiple R-Squared** value, which is coefficient of determining how much our fitted model can explain the given data. This R-Squared value lie between 0 and 1. Now we would like to mention our three fitted models and corresponding multiple R-Squared value of each model.

1. Thus our **estimated multiple regression equation of 1st model** is -

$$\hat{y} = 5.5749207 + 0.0083687x_1 + 0.0002617x_2 - 0.0240877x_3$$

   And **multiple R-Squared value** is = 0.08982 i.e. the given data is only 8.982% explainable by the 1st multiple regression model. And there is some outlier with respect to this model.

2. Removing outlier of 1st model **estimated multiple regression equation of 2nd model** is -

$$\hat{y} = 5.2905645 + 0.0106772x_1 + 0.0007567x_2 - 0.0224924x_3$$

   And **multiple R-Squared value** is = 0.09728 i.e. the given data is only 9.728% explainable by the 2nd multiple regression model. And here is also some outlier with respect to this model.

3. Removing outlier of 2nd model **estimated multiple regression equation of 3rd model** is -

$$\hat{y} = 5.090054 + 0.012305x_1 + 0.001106x_2 - 0.021367x_3$$

   And **multiple R-Squared value** is = 0.1024 i.e. the given data is only 10.24% explainable by the 3rd multiple regression model. This is the final model free from outliers.

So, we will use our final model i.e. 3rd regression model for inferencing.

Interpretation of coefficients :

- 0.012305 unit of increase in weight of the product can be expected when x_1 i.e. extruder speed is increased 1 unit and other two variables x_2(current flown) and x_3(conveyer speed) are remained constant.
- 0.001106 unit of increase in weight of the product can be expected when x_2 i.e. extruder speed is increased 1 unit and other two variables x_1(extruder speed) and x_3(conveyer speed) are remained constant.
- 0.0224924 unit of increase in weight of the product can be expected when x_3 i.e. extruder speed is increased 1 unit and other two variables x_1(extruder speed) and x_2(current flown) are remained constant.

Now, using our final model i.e. 3rd model we will find the 5% confidence interval of each regression parameters (i.e. a, b, c, d) of this model, where a, b, c and d are respectively **intercept**, **slope of extruder speed (EXTRUDER.RPM)**, **slope of current flown (CURRENT)** and **slope of conveyer speed (ConveyerSpeed)**.

R Chunk - 2.1.11

```
                     2.5 %       97.5 %
(Intercept)    3.564096011  7.020752289
EXTRUDER.RPM  -0.008850536  0.030914771
CURRENT       -0.006413255  0.006531401
ConveyerSpeed -0.042936444 -0.001296527
```

So, from the above inferential analysis we can notice that except intercept all the multiple regression parameters' 95 % confidence interval includes 0 at approximately middle of each confidence interval and each confidence interval has very very short length. Thus there is no reason to reject the clam that all the **regression parameters except intercept are not significantly different from** 0 **at the** 5 % **level of significance in the light of given data**. So, there is **no such statistically significant linear relationship between the variables of interest**.

3 CONCLUSION AND DECISION

From all of our above exploratory and inferential data analysis, we can conclude some significant report based on the given data as following -

1. **The correlation between weight of product and extruder speed (in RPM) is not statistically significant (i.e. very close to** 0**).**
2. **The correlation between weight of product and current flown is not statistically significant.**
3. **All the regression parameters except intercept are not significantly different from at the** 5 % **level of significance.**
4. **There is no such statistically significant linear relationship between the variables of interest.**

From this entire analysis we can realize that the concept of multiple linear regression model is a very poor model to explain this given data set. So, one can think of fitting another model may be some quadratic or exponential model (say) instead of using multiple regression model to improve the research study in future. Because the multiple linear regression model can't explain a huge portion of irregularity in the data set.

APPENDIX :-

Mathematical Formulas :

- $Correlation(r_{xy}\, or,\ r_{yx}) = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{Var(x)}\sqrt{Var(y)}}$
- Let y_i's are the observed values of a variable under study and $\hat{y}_i$'s are the corresponding predicted value for all $i = 1(1)n$.

  Define the **residuals** as

$$\epsilon_i = y_i - \hat{y}_i, \ for\ all\ i = 1(1)n$$

If $\bar{y}$ is the mean observed data such that

$$\bar{y} = \sum_{i=1}^{n} y_i,$$

then the **variability** of the data set can be measured using three **sums of squares** formulas:

1. The total sum of squares (proportional to the variance of the data) is -

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

2. The **regression sum of squares**, also called the **explained sum of squares** is -

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

3. The **sum of squares of residuals**, also called the **residual sum of squares** is -

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The most general definition of the **coefficient of determination (R-Squared)** is -

$$R^2 = 1 - Fraction\ of\ variance\ unexplained(FVU)$$

$$i.e.\ R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation}$$

$$i.e.\ R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R Codes :

- **For Exploratory Data Analysis -**

R Chunk - 2.1.1

```
data=read.table(file="C:/Users/Ankan Chakraborty/Extruder.csv",sep=",",header=T)
attach(data)
data[1:2,] #Extracting 1st two rows of the data set
```

R Chunk - 2.1.2

```
par(mfrow=c(1,3)) #Plotting 3 graphs in one window
    plot(EXTRUDER.RPM,WEIGHT,main="Scatterplot of weight \nand extruder speed")
    plot(CURRENT,WEIGHT,main="Scatterplot of weight \nand current")
    plot(ConveyerSpeed,WEIGHT,main="Scatterplot of weight \nand conveyer speed")
```

R Chunk - 2.1.3

```
cat("The correlation between weight and extruder speed is -")
cor(WEIGHT,EXTRUDER.RPM)
cat("The correlation between weight and current is -")
cor(WEIGHT,CURRENT)
cat("The correlation between weight and conveyer speed is -")
cor(WEIGHT,ConveyerSpeed)
```

R Chunk - 2.1.4

```
model=lm(formula = WEIGHT~.,data = data) #Fitting multiple regression model
model
```

R Chunk - 2.1.5

```
summary(model)        #For knowing the whole information about the fitted model
```

R Chunk - 2.1.6

```
predicted=5.5749207+0.0083687*EXTRUDER.RPM+0.0002617*CURRENT-0.0240877*ConveyerSpeed
residual=WEIGHT-predicted
m=max(residual)
par(mfrow=c(1,2))
  plot(predicted,residual,main="Residual plot against \npredicted weight of 1st model",
  ylim=c(-m,m),xlab="Predicted weight", ylab="Residuals")
  abline(0,0)         #Plotting X-axis
  std.residual=rstandard(model)
  plot(predicted,std.residual,main="Standardized residual plot against \npredicted weight
  of 1st model",ylim = c(-3,3),xlab="Predicted weight", ylab="Standardized Residuals")
  abline(0,0)         #Plotting X-axis
  abline(-2,0,lty=2) #Plotting x=2 line
  abline(2,0,lty=2)  #Plotting x=-2 line
  text(x=4.410,y=2.85,labels="Outlier")  #For writing text in the graphing plot
  arrows(4.414,2.8,4.424,2.8,angle=20)   #For drawing arrow in the graphing plot
```

R Chunk - 2.1.7

```
outlier_1=which(abs(std.residual)>2)#To know the index position of outlying observation
data_1=data[-outlier_1,]            #Removing outlying observation
model_1=lm(formula = WEIGHT~.,data = data_1)#Fitting new model after removing outlier
summary(model_1)
```

R Chunk - 2.1.8

```
predicted_1=5.2905645+0.0106772*data_1$EXTRUDER.RPM+0.0007567*data_1$CURRENT
-0.0224924*data_1$ConveyerSpeed
residual_1=data_1$WEIGHT-predicted_1
std.residual_1=rstandard(model_1)
plot(predicted_1,std.residual_1,main="Standardized Residual plot against predicted weight
of 2nd model \n(after removing outlier of 1st regression model)",xlim=c(4.39,4.445),
ylim=c(-3,3),xlab="Predicted weight", ylab="Standardized Residuals")
abline(0,0)
abline(-2,0,lty=2)
abline(2,0,lty=2)
text(x=4.410,y=2.85,labels="Outlier")
arrows(4.413,2.8,4.4215,2.13,angle=20)
```

R Chunk - 2.1.9

```
outlier_2=as.numeric(which(abs(std.residual_1)>2))
data_2=data_1[-c(outlier_1,outlier_2),]
model_2=lm(formula = WEIGHT~.,data = data_2)
summary(model_2)
```

R Chunk - 2.1.10

```
outlier_2=as.numeric(which(abs(std.residual_1)>2))
data_2=data_1[-c(outlier_1,outlier_2),]
model_2=lm(formula = WEIGHT~.,data = data_2)
predicted_2=5.090054+0.012305*data_2$EXTRUDER.RPM+0.001106*data_2$CURRENT
-0.021367*data_2$ConveyerSpeed
residual_2=data_2$WEIGHT-predicted_2
std.residual_2=rstandard(model_2)
plot(predicted_2,std.residual_2,main="Standardized Residual plot against predicted weight
of 3rd model \n(after removing outlier of 2nd regression model)",xlim=c(4.39,4.445),
ylim=c(-3,3),xlab="Predicted weight", ylab="Standardized Residuals",sub="(Pattern
recognition in Standardised Residual plot)")
x=c(4.397,4.423,4.44)
y=c(2.15,2.15,0,0,-2.15,-2.15)
polygon(c(x,rev(x)),y,border="black",col=rgb(0.5,0.5,0.5,0.4))
abline(0,0)
abline(-2,0,lty=2)
abline(2,0,lty=2)
```

- **For Inferential Data Analysis -**

R Chunk - 2.1.11

```
confint(model_2)      #By default it gives 95% confidence interval
detach(data)
```

BIBLIOGRAPHY :-

Some useful sources used to get plenty of informations are -

- Linear Regression Residual Analysis - https://youtu.be/gLENW2AdJWg
- Linear Regression, Outliers and Influential Observations - https://youtu.be/fJSXS4oVf88
- Multiple Regression Residual Plots - https://youtu.be/wtT4zkxNy-A
- Multiple Linear Regression, The Very Basics - https://youtu.be/dQNpSa-bq4M
- https://en.wikipedia.org/wiki/Coefficient_of_determination
- https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R
- https://www.statisticshowto.datasciencecentral.com/what-is-a-standardized-residuals/
- http://www.r-tutor.com/elementary-statistics/simple-linear-regression/standardized-residual
- https://stats.stackexchange.com/questions/155586/confidence-intervals-of-coefficients-of-multiple-regression
- https://rpubs.com/aaronsc32/regression-confidence-prediction-intervals
- https://rstudio-pubs-static.s3.amazonaws.com/71339_d0b8346f41314979bc394448c5d60d86.html
- https://www.investopedia.com/terms/r/r-squared.asp

Acknowledgement :-

I want to thank Prof. Atanu Kumar Ghosh for providing me with this project and help me to learn some interesting result and facts from the data set. Most significantly what I learnt is to have an insight of a real life oriented data set and problems. I shall remain ever grateful to him for pushing my limit into real life working field.