Prediction of Concavity of breast cancer cell using Multiple Linear Regression

Harshita Jain (2020CH70166)
Lalwala Taukir Mohamad (2020CH70174)
Meeta Rajput (2020CH70179)
Spandan Dutta (2018CH10248)

November 11, 2021

Supervisor: Prof. Jayati Sarkar

Department of Chemical Engineering
Indian Institute of Technology Delhi

Abstract:

Concavity of breast cancer cell is predicted by various models developed in this term paper. Majorly two models are developed, Multiple Linear Regression and Multiple Polynomial Regression. Multiple Linear Regression is modified and Errors are reduced. Multiple Linear Regression performs better when modifications are done and manually some features are squared by observing how the data is scattered.

Key Findings:

Concavity can be predicted with acceptable accuracy and error. Multiple Linear Regression can predict concavity with Root Mean Squared Error (RMSE) **0.03034** and Coefficient of Determination (R^2) **88.68 %.** Multiple Linear regression model can be modified and model with RMSE **0.02388** and R^2 **92.79 %** can be obtained.

Nomenclature and abbreviations:

1. MLR – Multiple Linear Regression
2. PLR – Polynomial Linear Regression
3. MAE – Mean Absolute Error
4. MSE – Mean Squared Error
5. RMSE – Root Mean Squared Error
6. R^2 – Coefficient of Determination
7. AIC – Akaike Information Criterion
8. SS_{res} – Sum of Residuals
9. SS_{tol} – Total Sum of Squares

Content **Page No.**

Introduction

Breast cancer is a one among the numerous cancers, a lethal disease with several variations. Breast cancer develops when a subset of breast cells occurs when there is an uncontrolled growth of breast cells. These cells multiply and grow faster than the healthy cells and clump together to create a mass. Malignant Cells in the breast may spread to the areas with nodes of lymph or other body parts. The group of people who are prone to breast cancer are females, someone having personal ancestral history of breast cancer, specific genetic makeup susceptible to cancer, someone working in the radiations, having excess fat, having their first child at an older age, having menarche at earlier or having delayed menopause, someone drinking alcohol, consuming tobacco.

Developed countries are at higher risk than in developing countries in terms of breast cancer as a result certain high-standard practices such as obesity due to unhealthy diet menopausal therapy, declining fertility rates, use oral contraceptive pills, lesser breastfeeding time, etc.
Preliminary analysis and prediction of malignant progression have become critical concerns in tumor cell research because they may assist patients in obtaining better therapy.

Women are more to Breast cancer than men. Experts believe that DNA abnormalities cause 5% to 10% of breast cancers passed down through generations of a family. In 2020, 2.3 million new cases of breast cancer will get added and 685,000 deaths due to breast cancer globally, with the number of cases, predicted to rise to 4.4 million by 2070. Breast cancer accounted for 24.5% of total cases and 15.5% of it resulting in deaths among women in 2020, ranking first in terms of incidence and mortality in much of the globe. Its incidence and mortality vary from one nation to another, with the incidence of age-standard being the greatest. China had the highest number of breast cancer fatalities, accounting for around 17.1% of all cancer deaths while United States was at the second with for 6.2% of deaths due to breast cancer across the world. In contrast to incidence rates, high-income countries had lower mortality rates, for example, South Korea, United States, Japan, China while developing nations, for example, Samoa, Fiji, Nigeria, Cameroon, and Jamaica had lower death rates.

In the base research paper, MLR and PLR of 2^{nd} order are implemented. However, there can be many modifications (some of which are implemented in the term paper). Assuming linear dependence for all features in MLR ignores any feature's higher order dependence and in PLR considering 2^{nd} degree dependence for all features also squares features which have linear dependence. Our prime objective is to overcome these issues and propose a better model.

Problem formulation

The concavity of a breast cancer cell tells a lot about how much the cells can expand. In today's world, mammography is the most widely used procedure for detecting breast cancer. Many numerical approaches, Machine Learning and Deep Learning models, such as Logistic regression, have previously been created for predicting breast cancer and determining whether it is a malignant (grows at an unusually rapid pace) or benign (grows at a slower rate) tumour, as well as its stage. Predicting concavity of the cancer cells helps in estimating pressure on the surrounding cells qualitatively as concavity tells how much the cell can expand and at what rate. We have attempted to represent concavity as linear combination of 8 other features of cell, these results can be further used for studying cancel cells and their effects.

Numerical Analysis

Multiple Linear Regression (MLR):

It uses several predictor values to predict the dependent value. The link between independent variables and dependent variable is established using MLR.

MLR Equation with **m** independent variable can be written as:

$$y_1 = a_0 + a_1x_1 + a_2x_2 + \ldots + a_mx_m + e \quad \text{..... equation 1}$$

If we have **n** data points,

$$\widehat{y_1} = a_0 + a_1x_{11} + a_2x_{12} + \ldots + a_mx_{1m}$$

$$\widehat{y_2} = a_0 + a_1x_{21} + a_2x_{22} + \ldots + a_mx_{2m}$$

$$\vdots$$

$$\widehat{y_n} = a_0 + a_1x_{n1} + a_2x_{n2} + \ldots + a_mx_{nm}$$

where,

$\widehat{y_n}$ = dependent variable (concavity_mean for this model)
a_i = regression coefficients
x_i = dependent variables
e = residual error

For the model (n=140 for Training set and 60 for Testing set, m=8 as there are 8 features.)

Dependent variable:

1. *concavity_mean* – measure of severity of concave portions.

Independent variables in the model:

1. *radius_mean*: The average of the distances between the cancer cell's centre and the points around its perimeter.
2. *texture_mean*: It is the standard deviation (S.D) of greyscale values given in the data set.
3. *area_mean:* It is the mean of areas from the dataset.
4. *compactness_mean*: It is one less than the squared perimeter per unit area. i.e., compactness_mean =perimeter^2 / area - 1.0
5. *smoothness_mean*: It displays the local variance in radius length.
6. *perimeter_mean*: It is the mean of perimeters from the dataset.
7. *concave points_mean*: It's the number of segments contour graph which is concave
8. *fractal_dimension_mean*

Assumptions in MLR – no correlation between independent variables, linear relationship between dependent and independent variable, normal distribution of residuals.

Matrix Least square method: *(for calculating regression coefficients)*

$$y_1 = a_0 + a_1x_1 + a_2x_2 + \ldots + a_mx_m + e$$

$$\widehat{y_1} = a_0 + a_1x_{11} + a_2x_{12} + \ldots + a_mx_{1m}$$

$$\widehat{y_2} = a_0 + a_1x_{21} + a_2x_{22} + \ldots + a_mx_{2m}$$

.

.

$$\widehat{y_n} = a_0 + a_1x_{n1} + a_2x_{n2} + \ldots + a_mx_{nm}$$

$$\begin{bmatrix}\widehat{y_1}\\ \widehat{y_2}\\ .\\ .\\ .\\ \widehat{y_n}\end{bmatrix} = \begin{bmatrix}1 & \hat{x}_{11} & \hat{x}_{12} & . & . & . & \hat{x}_{1m}\\ 1 & \hat{x}_{21} & \hat{x}_{22} & . & . & . & \hat{x}_{2m}\\ . & . & . & . & . & . & .\\ . & . & . & . & . & . & .\\ . & . & . & . & . & . & .\\ 1 & \hat{x}_{n1} & \hat{x}_{n2} & . & . & . & \hat{x}_{nm}\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ a_2\\ .\\ .\\ a_m\end{bmatrix}$$

$$\begin{bmatrix}\widehat{y_1}\\ \widehat{y_2}\\ .\\ .\\ .\\ \widehat{y_n}\end{bmatrix} = \hat{y} \begin{bmatrix}1 & \hat{x}_{11} & \hat{x}_{12} & . & . & . & \hat{x}_{1m}\\ 1 & \hat{x}_{21} & \hat{x}_{22} & . & . & . & \hat{x}_{2m}\\ . & . & . & . & . & . & .\\ . & . & . & . & . & . & .\\ . & . & . & . & . & . & .\\ 1 & \hat{x}_{n1} & \hat{x}_{n2} & . & . & . & \hat{x}_{nm}\end{bmatrix} = X \begin{bmatrix}a_0\\ a_1\\ a_2\\ .\\ .\\ a_m\end{bmatrix} = \emptyset$$

Least Square Method minimizes $\sum e_i^2$

$$\underline{Y} = X\underline{\emptyset} + \underline{e}$$

$$\underline{e} = \underline{Y} - X\underline{\emptyset}$$

$$S(\underline{\emptyset}) = \sum e_i^2 = (\underline{Y} - X\underline{\emptyset})^T(\underline{Y} - X\underline{\emptyset})$$

$$= \underline{YY^T} - \underline{Y^T}X\underline{\emptyset} - \left(X\underline{\emptyset}\right)^T\underline{Y} + \left(X\underline{\emptyset}\right)^T X\underline{\emptyset}$$

$$= \underline{YY^T} - \underline{Y^T}X\underline{\emptyset} - \underline{\emptyset^T}X^T\underline{Y} + \underline{\emptyset^T}X^TX\underline{\emptyset}$$

$$\frac{dS}{d\emptyset} = -2X^T\underline{Y} + 2X^TX\underline{\emptyset} = 0$$

$$\Rightarrow X^TX\underline{\emptyset} = X^T\underline{Y}$$

$$\underline{\emptyset} = (X^TX)^{-1}X^T\underline{Y}$$

$\emptyset$ can be now used for predicting concavity of cancer cell.

$$\mathrm{y} = a_0 + a_1x_1 + a_2x_2 + \ldots. + a_mx_m$$

where, y is the predicted concavity_mean.

Equation from MLR:

concavity_mean = 0.052794 - 0.667856*radius_mean + 0.016929*texture_mean + 0.281758*perimeter_mean + 0.309720*area_mean - 0.061102*smoothness_mean + 0.175801*compactness_mean + 0.327907*concave_points_mean - 0.020731*fractal_dimension_mean

Polynomial Linear Regression (PLR):

Dependence on independent variable is not always linear, many a time there is a curvature possible. PLR can be used in these case for fitting the model well and predicting value of dependent variable. The dependence can be quadratic, cubic, or biquadratic or any other degree.

According to base journal article the 2nd order polynomial implemented here fits a wide range of curvatures. In this PLR, quadratic dependence is considered on all the independent variables and analyzed the performance in comparison to the MLR.

Note: Here quadratic dependence is considered, however there can be different dependence like cubic or exponential. It is a matter of data analysis and predicting dependence. Many relations are proposed in the later part of the term paper.

$$y = a_0 + a_1x_1^2 + a_2x_2^2 + \cdots + a_mx_m^2 + e$$

Rather than studying each feature individually, considering any specific dependence (quadratic here) for all independent variables may lead to overfitting although accuracy can be increased.

For calculating regression coefficients, we can implement same method as used in MLR but while calculating coefficients by Matrix least square method all the x_i are squared.

$$y_1 = a_0 + a_1x_1^2 + a_2x_2^2 + \ldots + a_mx_m^2 + e$$

$$\widehat{y_1} = a_0 + a_1x_{11}^2 + a_2x_{12}^2 + \ldots + a_mx_{1m}^2$$

$$\widehat{y_2} = a_0 + a_1x_{21}^2 + a_2x_{22}^2 + \ldots + a_mx_{2m}^2$$

.

.

$$\widehat{y_n} = a_0 + a_1 x_{n1}^2 + a_2 x_{n2}^2 + \ldots + a_m x_{nm}^2$$

$$\begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \\ . \\ . \\ . \\ \widehat{y_n} \end{bmatrix} = \begin{bmatrix} 1 & \hat{x}_{11}^2 & \hat{x}_{12}^2 & . & . & . & \hat{x}_{1m}^2 \\ 1 & \hat{x}_{21}^2 & \hat{x}_{22}^2 & . & . & . & \hat{x}_{2m}^2 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & \hat{x}_{n1}^2 & \hat{x}_{n2}^2 & . & . & . & \hat{x}_{nm}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ . \\ a_m \end{bmatrix}$$

$$\begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \\ . \\ . \\ . \\ \widehat{y_n} \end{bmatrix} = \hat{y} \begin{bmatrix} 1 & \hat{x}_{11}^2 & \hat{x}_{12}^2 & . & . & . & \hat{x}_{1m}^2 \\ 1 & \hat{x}_{21}^2 & \hat{x}_{22}^2 & . & . & . & \hat{x}_{2m}^2 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & \hat{x}_{n1}^2 & \hat{x}_{n2}^2 & . & . & . & \hat{x}_{nm}^2 \end{bmatrix} = X \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ . \\ a_m \end{bmatrix} = \emptyset$$

Now using Matrix Least square method, $\emptyset$ can be calculated and $\emptyset$ can be now used for predicting concavity of cancer cell.

$$y = a_0 + a_1 x_1^2 + a_2 x_2^2 + \ldots. + a_m x_m^2$$

where, y is the predicted concavity_mean.

Equation from PLR :

concavity_mean = 0.012475 + 0.478038*radius_mean^2 + 0.019002*texture_mean^2 - 0.098446*perimeter_mean^2 - 0.389872*area_mean^2 - 0.008414*smoothness_mean^2 + 0.125206*compactness_mean^2 + 0.316206*concave_points_mean^2 + 0.109285*fractal_dimension_mean^2

Error Analysis

Each model is analysed on the basis on Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2).

MAE – Mean of absolute value of residuals

$$MAE = \frac{\sum_{i=1}^{M}|y_i - x_i|}{M}$$

MSE – also known as variance, measure of how the predicted values are spread out from mean value.

$$MSE = \frac{\sum_{i=1}^{M}(y_i - x_i)^2}{M}$$

RMSE – also known as standard deviation

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{M}(y_i - x_i)^2}{M}\right)}$$

Where,
y_i is predicted value
x_i is true value
M is total number of testing point (M=60 in this model)

R^2 – It is the measure of variation in dependent variable (concavity mean) that can be predictable from independent variables.

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

Where,
SS_{res} is sum of residuals and SS_{tol} is total sum of squares, it is sum of squares of difference in predicted value and mean value of the observed values.

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

$\bar{y}$ is the mean value of all observed (predicted values)

$$SS_{res} = \sum_{i=1}^{M}(y_i - x_i)^2$$

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

Where,
y_i is predicted value
x_i is true value

Data set

Data set is obtained from *kaggle.com*.
Information in data set.

1) ID number
2) Diagnosis (Malignant(M), Benign(B))

Real valued features of each cell nucleus:

1) Concavity
2) Radius
3) Texture
4) Perimeter
5) Area
6) Smoothness
7) Compactness
8) Concave points
9) Symmetry
10) Fractal dimension

Corresponding to each feature, Data set contains mean, standard error and worst of that feature. There is no missing data in the data set.

Algorithm:

1. Creating training set and testing set:

   In this model 200 data set are used in total and they are divided in 70 : 30 ratio. 140 data points are used to train the model in other words to determine the weights of all independent variables (regression coefficients) and 60 data points for prediction of concavity and error analysis.

   Training data is stored in *Training_Set.csv* and Testing data in *Testing_Set.csv* in Data folder.

2. Feature scaling:

   In this model min-max normalization is used to normalize the date. The data can have varying value and magnitudes. Its an important step as without it the training model can while calculating coefficients can weigh higher to greater values and lower to smaller values without considering

units. Eg. If any model uses 100m and 0.2 km as data then without feature scaling it can consider 100m to be greater. So feature scaling is done. Min-max normalization scales the data between 0 and 1.

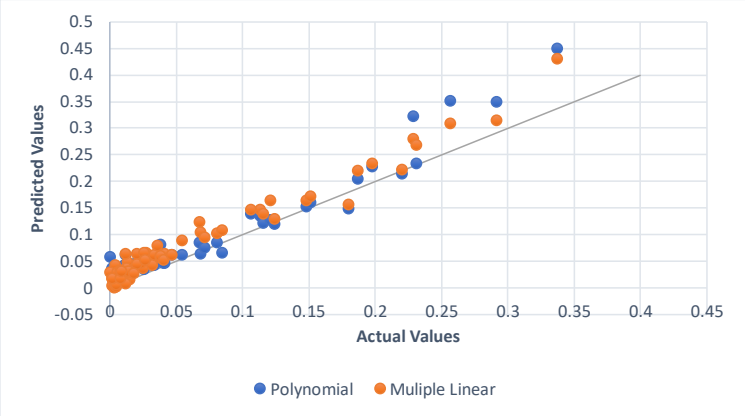
$$x_{new} = \frac{(x_i - \min(X))}{\max(X) - \min(X)}$$

Where, x_{new} is scaled value, x_i is original value, X is the set of all the values of the feature which is scaled.

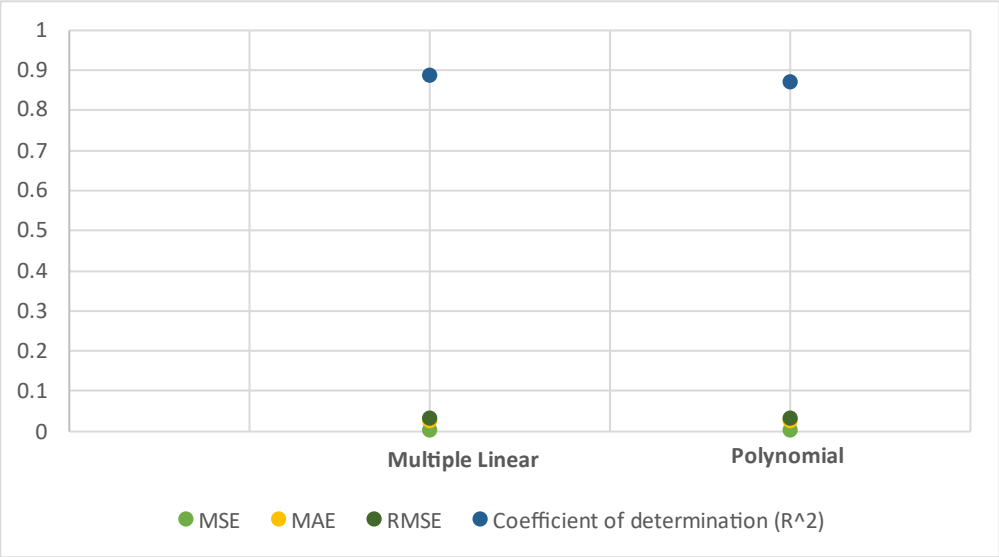
3. Calculate regression coefficients by matrix least square method in training model.
4. Use regression coefficients in the testing model, calculate errors and coefficient of determination.
5. Use testing model for feeding in new input and predict concavity.

Results

Regression Technique	MSE	MAE	RMSE	R^2
MLR	0.000920881	0.0254193	0.030346	88.6767 %
PLR (2nd order)	0.00108499	0.0250093	0.0329392	87.0639 %



Inference from Graph – predicted values of concavity mean in MLR are closer to y=x (i.e. Actual concavity mean) line than PLR and residual errors are lesser in MLR.



Inference from Graph – MLR clearly has larger coefficient of determination. Error values are close so not distinguishable from the graph.

Error and performance analysis

MSE and RMSE is lesser in MLR and MAE can be considered equivalent. Coefficient of Determination (R^2) is more in MLR. From this error analysis MLR can be considered as a better model. Performance of MLR > PLR of 2^{nd} order.

Reason for above error analysis and model behavior

In PLR implemented as directed in base journal, 2^{nd} order is taken and all features are squared. There can be possible overfitting in the model and may be some features were not in squared dependence but got squared, so it performed less effectively in comparison to MLR.

Level 2

Motivation: It concluded in *Reason for above error analysis and model behavior* of Level 1, that possible reason of MLR performing better than PLR of 2nd order can be squaring all the features which would have led to overfitting.

For improving the model rather than squaring all the features and making the equation 1 2nd order, Data visualization can be done for predicting degrees of features and optimizing the performance.

Proposal 1: It is obvious that area is directly proportional to [length]^2 while radius and perimeter are proportional to [length]^1, so if we consider degree of area to be 1 in the equation 1 then using 2nd degree for perimeter_mean and radius_mean should predict concavity_mean with lesser error and more R^2 value.

Equation from MLR with proposal 1 (First improved model):

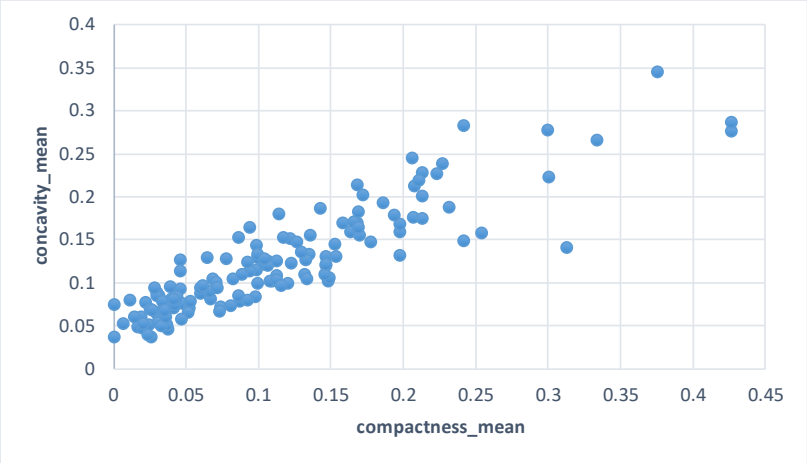
concavity_mean = 0.018540 + 0.053110*radius_mean^2 + 0.016036*texture_mean + 0.174703*perimeter_mean^2 - 0.248679*area_mean - 0.058336*smoothness_mean + 0.149443*compactness_mean + 0.313350*concave_points_mean - 0.020258*fractal_dimension_mean

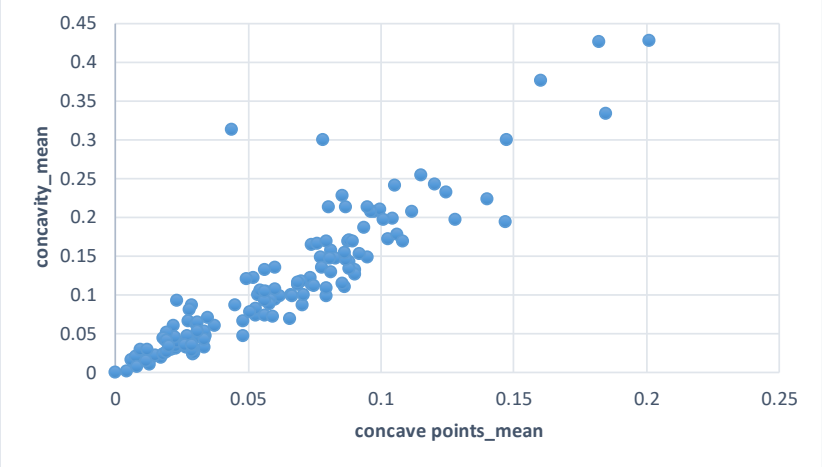
Result

Regression Technique	MSE	MAE	RMSE	R^2
MLR with proposal 1 (First Improved Model)	0.000581925	0.0189347	0.0241231	92.8611 %

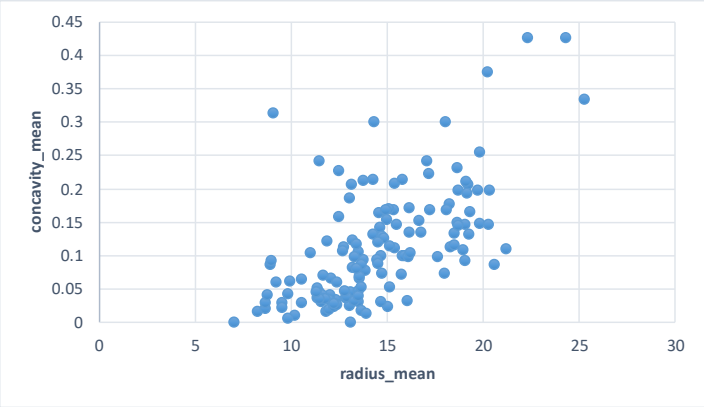
Proposal 2: Some modification can be made in equation 1 by carefully observing training set data.

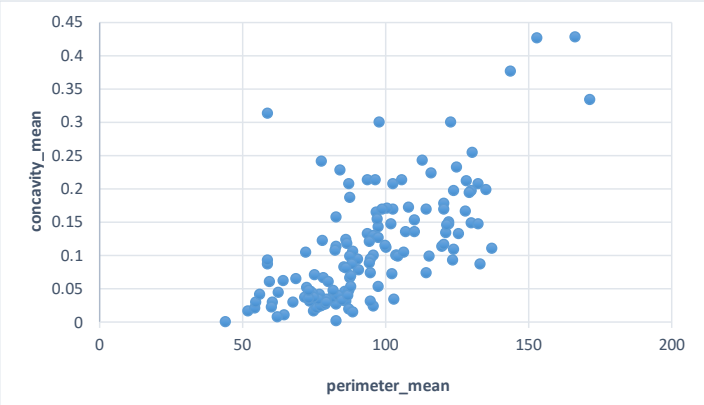
Proposal-2a - Below compactness_mean and concave points_mean are not much scattered so, considering linear dependence for them.

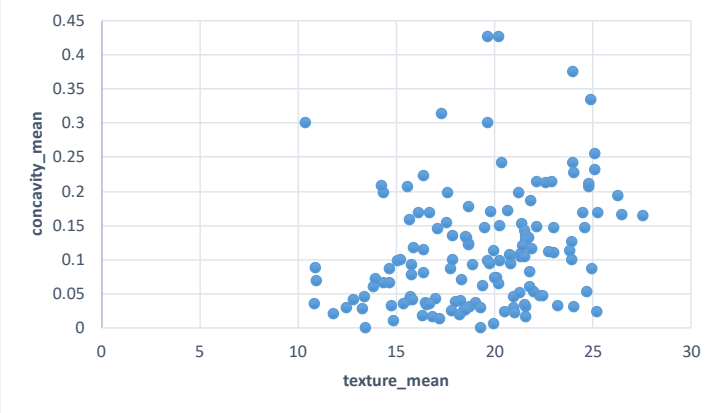


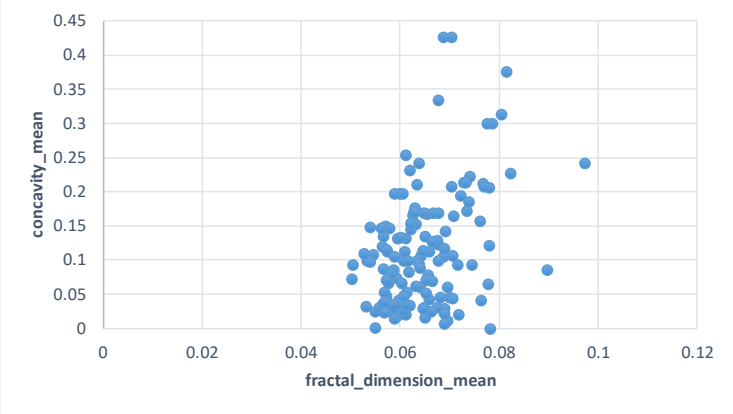


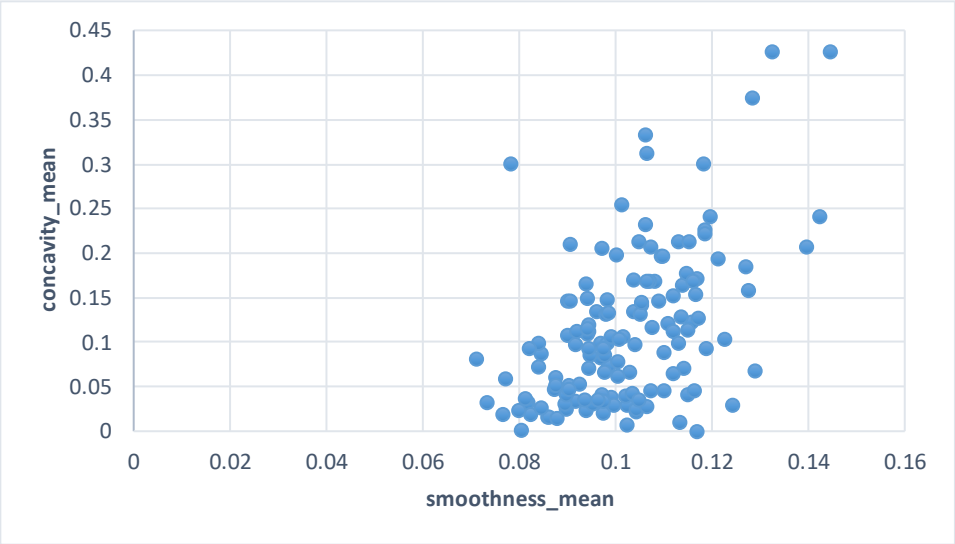
Proposal-2b - Below radius_mean, perimeter_mean, texture_mean, smoothness_mean and fractal_dimension_mean, they seem to be similarly scattered so according to proposal 1, as we are considering second order dependence for radius_mean and perimeter_mean, we propose second order dependence on texture_mean, smoothness_mean and fractal_dimension_mean also.


concavity_mean
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
radius_mean


concavity_mean
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150
200
perimeter_mean


concavity_mean
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
texture_mean


concavity_mean
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.02
0.04
0.06
0.08
0.1
0.12
fractal_dimension_mean



Equation from MLR with proposal 2 (Second improved model):

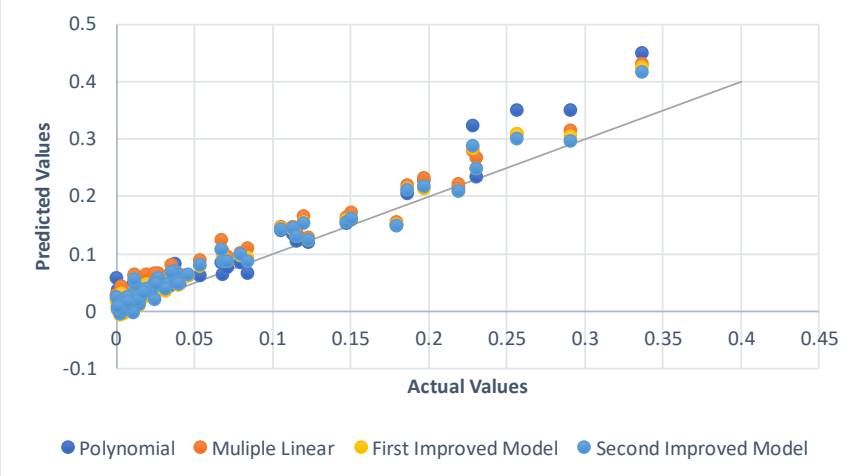
concavity_mean = 0.008312 - 0.088533*radius_mean^2 + 0.010066*texture_mean^2 + 0.278948*perimeter_mean^2 - 0.181611*area_mean - 0.041188*smoothness_mean^2 + 0.131236*compactness_mean + 0.292271*concave_points_mean - 0.041467*fractal_dimension_mean^2

Result

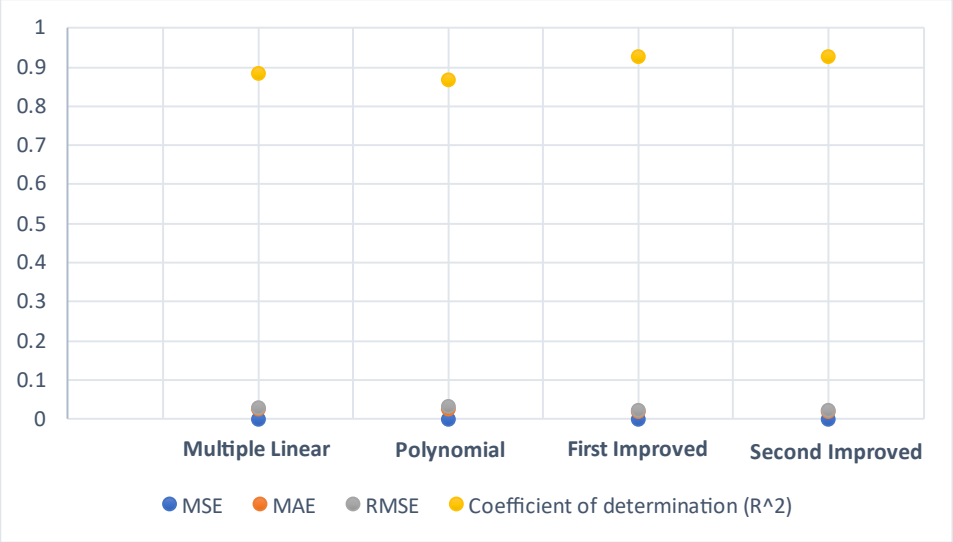
Regression Technique	**MSE**	**MAE**	**RMSE**	**R^2**
MLR with proposal 1 and proposal 2 (Second improved model)	0.000570292	0.018915	0.0238808	92.7929 %

Combined Results of all models:

Regression Technique	MSE	MAE	RMSE	R^2
MLR	0.000920881	0.0254193	0.030346	88.6767 %
PLR (2nd order)	0.00108499	0.0250093	0.0329392	87.0639 %
MLR with proposal 1(First improved model)	0.000581925	0.0189347	0.0241231	92.8611 %
MLR with proposal 1 and proposal 2 (Second improved model)	0.000570292	0.018915	0.0238808	92.7929 %



Inference from Graph – predicted values of concavity mean in Second improved Model are closer to y=x (i.e Actual Concavity mean) line then First improved model then MLR and PLR. Residual error has order Second Improved Model < First Improved Model < MLR < PLR.



Inference from graph – R^2 is larger of First improved and Second improved method than MLR and PLR. MAE, MSE, RMSE are quite close for all so not distinguishable from this graph.

Analysis:

In level 1 it was concluded that MLR performs better than PLR. After proposal 1, MAE, MSE, RMSE of the model are reduced and R^2 is also increases. After proposal 2 R^2 is almost same as proposal 1 with slight decrease in MSE, MAE, RMSE.
So it can be concluded that performance of second improved model ~ First improved model > MLR > PLR.

Conclusion:

In this term paper concavity of the cancer cell is predicted by 4 models, MLR, PLR(2^{nd} order), and two improved models which are developed by modifying MLR. They provide good approxiamation of concavity of the cancer cell. MLR with RMSE **0.03034** R^2 value **88.68 %** perform better than 2^{nd} order PLR with RMSE **0.03294** and R^2 **87.06%.** After 2 proposal and modification in MLR yeild model with RMSE **0.02388** and R^2 value **92.79 %.** Our results can be used for further research in this area and models can be further modified to decrease the errors and possible faults in the improved models.

Additional Analysis:

After analyzing the data, the following are the summary of the dependent and all the independent variables.

```
> #descriptive statistics
> summary(Y)
 concavity_mean
 Min.   :0.00000
 1st Qu.:0.02956
 Median :0.06154
 Mean   :0.08880
 3rd Qu.:0.13070
 Max.   :0.42680
> summary(X)
  radius_mean      texture_mean   perimeter_mean     area_mean      smoothness_mean
 Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5   Min.   :0.05263
 1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637
 Median :13.370   Median :18.84   Median : 86.24   Median : 551.1   Median :0.09587
 Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9   Mean   :0.09636
 3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530
 Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0   Max.   :0.16340
 compactness_mean  concave_points_mean fractal_dimension_mean
 Min.   :0.01938   Min.   :0.00000     Min.   :0.04996
 1st Qu.:0.06492   1st Qu.:0.02031     1st Qu.:0.05770
 Median :0.09263   Median :0.03350     Median :0.06154
 Mean   :0.10434   Mean   :0.04892     Mean   :0.06280
 3rd Qu.:0.13040   3rd Qu.:0.07400     3rd Qu.:0.06612
 Max.   :0.34540   Max.   :0.20120     Max.   :0.09744
```

Using the Multiple Linear Regression (MLR) model, the data was fitted, and a model was approximated for predicting the concavity values. The summary of the model is shown below:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.079788 -0.010586 -0.000376  0.008937  0.223573

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.936e-02  3.434e-02   2.311  0.02118 *
Xradius_mean            -5.323e-02  1.095e-02  -4.861 1.52e-06 ***
Xtexture_mean            4.272e-04  2.487e-04   1.718  0.08633 .
Xperimeter_mean          6.928e-03  1.763e-03   3.931 9.53e-05 ***
Xarea_mean               5.462e-05  2.084e-05   2.621  0.00900 **
Xsmoothness_mean        -8.741e-01  1.134e-01  -7.708 5.84e-14 ***
Xcompactness_mean        2.842e-01  8.644e-02   3.288  0.00107 **
Xconcave_points_mean     1.582e+00  9.680e-02  16.338  < 2e-16 ***
Xfractal_dimension_mean  9.142e-01  3.522e-01   2.596  0.00969 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02346 on 560 degrees of freedom
Multiple R-squared:  0.9146,    Adjusted R-squared:  0.9134
F-statistic:   750 on 8 and 560 DF,  p-value: < 2.2e-16
```

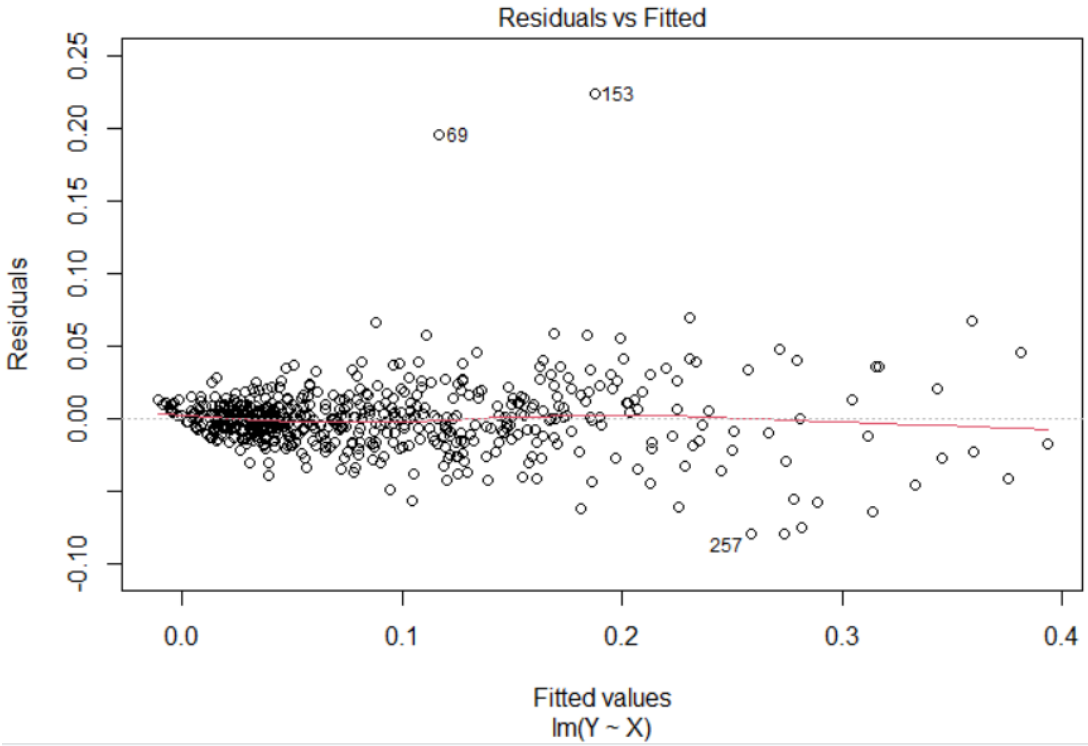
So, the proposed model can be:

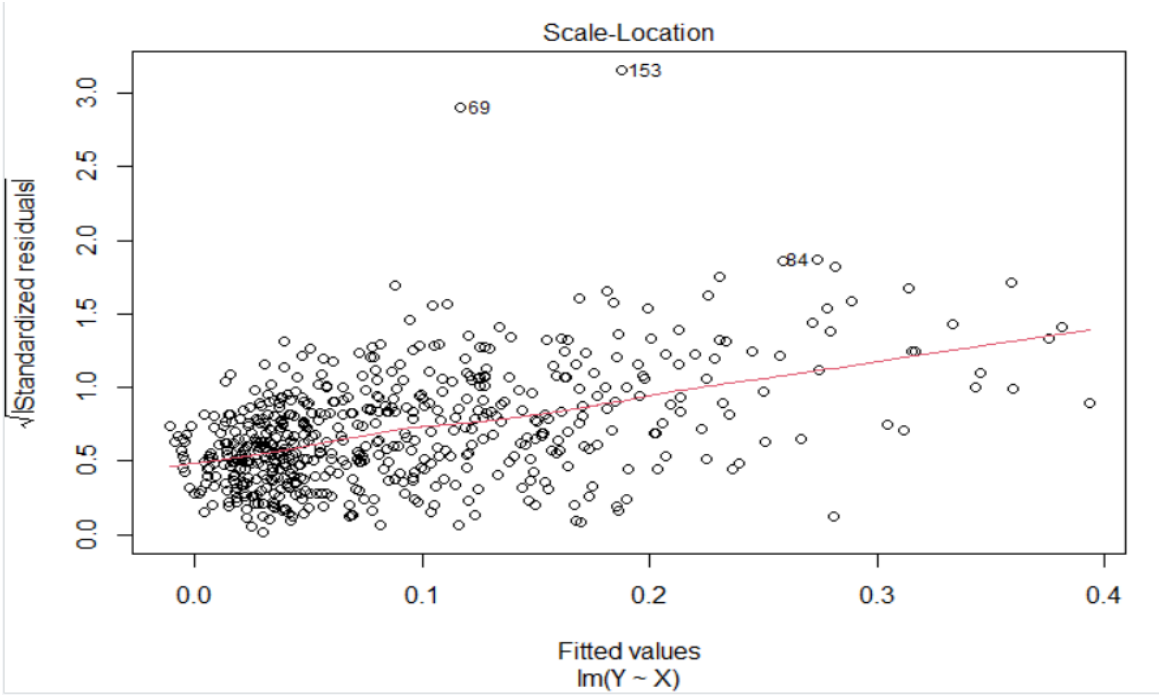
Concavity = $0.07936 - 0.05323$(radius_mean) + $4.272*10^{-4}$(texture_mean) + $6.928*10^{-3}$(perimeter_mean) + $5.462*10^{-5}$(area_mean) – $8.741*10^{-1}$(smoothness_mean) + $2.842*10^{-1}$(compactness_mean) + 1.582(concave points_mean) + $9.142*10^{-1}$(fractal_dimension_mean)

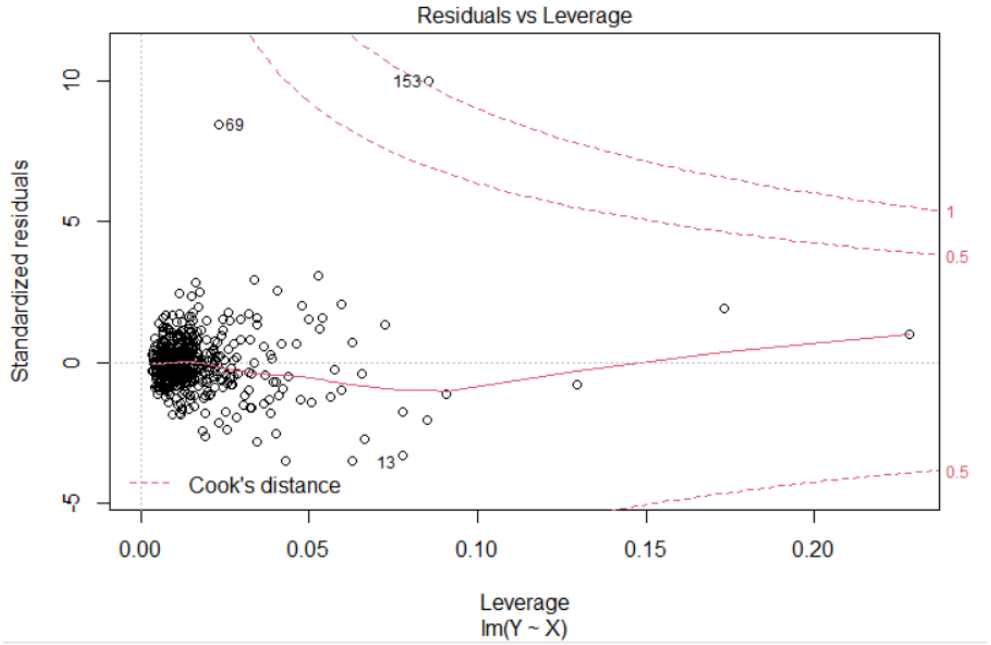
Due to multiple variables the model becomes a bit complex. In order reduce the complexity of the model, only the significant variables can be included without much affecting the R-squared value.

After the analysis, we find that only texture_mean is not significant whereas the slope and all the other independent variables are significant ($p < 0.05$). The variable (to optimize) that are not significant enough can be removed if by omitting that variable, the value of Residual Deviation should not increase and value of AIC should decrease.

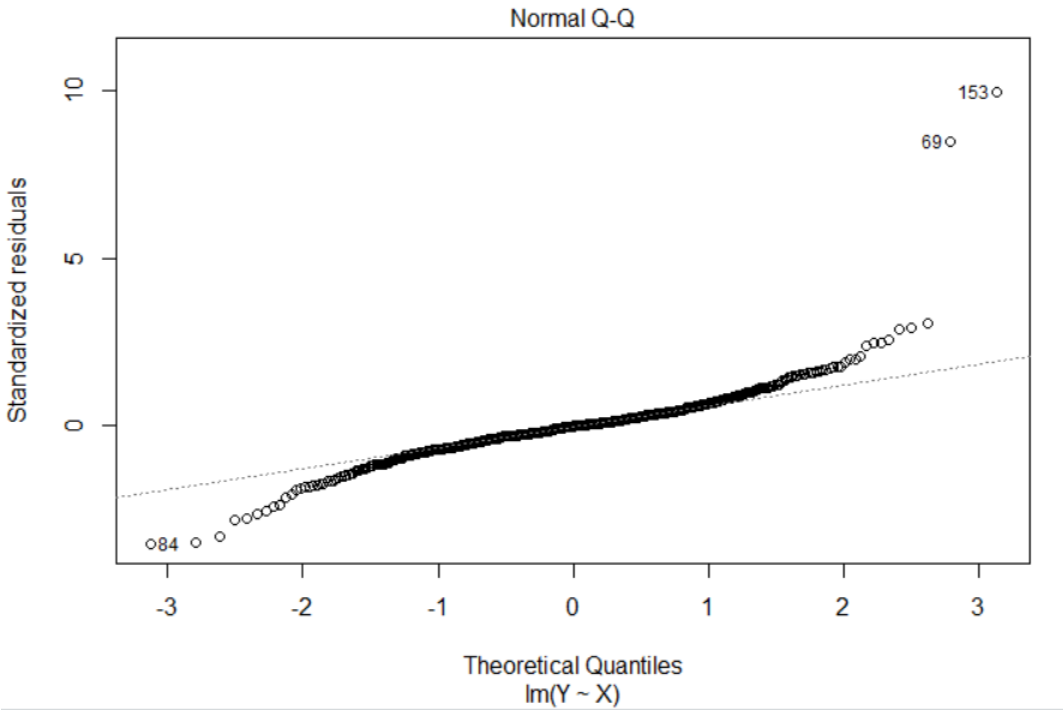
We did further analysis in order to find if our assumption that the residual is normally distributed is correct or not. The graphs obtained in the process are shown below:


Residuals vs Fitted
Residuals
Fitted values
lm(Y ~ X)
153
69
257
0.25
0.20
0.15
0.10
0.05
0.00
-0.10
0.0
0.1
0.2
0.3
0.4


Scale-Location
√|Standardized residuals|
Fitted values
lm(Y ~ X)
153
69
84
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.1
0.2
0.3
0.4



If the residuals (for given X, Y value) are normally distributed, the Q-Q plot should be a straight line and the residual values should fit in a straight line. But, the quantile-quantile (Q-Q) plot shown below, shows that the residuals values do not form a straight line which implies that the residuals are not normally distributed which contradicts our assumption made earlier.



Self-Assessment:

According to us, this term paper can be called **Level-2**. We developed code (C++ language used) for models (MLR and PLR) proposed in base research paper and our results also correspond with their results. We obtained Coefficient of Determination (R^2) for MLR and PLR as 88.6767 % and 87.0639 % respectively in comparison to 93.7 % and 90.3 % in original paper. Our results do not exactly match with original paper's result as there can be possible variation in data set for instance in base research paper training set data points: testing set data points ratio is 75:25 whereas we considered 70:30 ratio just for testing on more data points and total number of data points taken was not mentioned in base paper so we considered 200 total data points which were then split into 70:30 ratio. Our results follow the same trend as mentioned in the paper. (MLR performs better than PLR with more R^2 value and less RMSE, MAE turn out to be almost similar in both, slightly more in MLR).

We extended models to produce more accurate results with lesser errors and more R^2 value by proposing modifications in MLR on the basis of observation of data. R^2 value is increased to 92.7929% and RMSE decreased to 0.0238808.

Thus, we were able to successfully generate our own **original results**, which produces **more accurate** results with **lesser errors**.

Moreover, the report goes beyond reproducing results and we have done further analysis in order to check the validity of the assumptions made in the paper for generating the results. We have proved that the assumptions (such as assuming the residuals to be normally distributed) are actually not true and needs to be taken in account for developing a better (reliable) model which can provide more accuracy and less error. This becomes a part of our way forward and we aim to achieve this in future that will help to develop a better model for predicting cancer cells which may have some significant impact in the healthcare industry.

Way Forward:

These were the assumptions we made earlier:

- There is no collinearity between the independent variables.
- The residuals (for given X, Y) are normally distributed.

In reality, these assumptions may not be true. Further analysis needs to be done by including the collinearity between the independent variables which is likely to be the case. In case of collinearity found Ridge regression method can be used to regularize the model.

Also, from our analysis (included in the Level-2) we observed and concluded that the residuals are not normally distributed which contradicts our assumption. Due to which, the model needs to be further optimized by including these factors which may yield error if not taken into account. Some predicted values (7 predicted values out of 60*4 values predicted collectively by 4 models) of our MLR and modified models comes out to be negative when predicted which indicates that these models need further modifications for removing these faults.

References:

1. *This is the original Research paper we have referred https://iopscience.iop.org/article/10.1088/1757-899X/1166/1/012029/pdf*
2. *Global Stats and risk factors were taken https://onlinelibrary.wiley.com/doi/full/10.1002/cac2.12207*
3. *Briefings of Breast Cancer https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470*
4. *CLL113 lecture notes 16 Prof. Jayati Sarkar - for multiple linear regression and matrix least square*
5. *For the study of Mean Absolute error https://en.wikipedia.org/wiki/Mean_absolute_error*
6. *For the study of Mean squared error https://en.wikipedia.org/wiki/Mean_squared_error*
7. *For referring to Root mean square deviation https://en.wikipedia.org/wiki/Root-mean-square_deviation*
8. *Theory and definitions related to Coefficient of determination https://en.wikipedia.org/wiki/Coefficient_of_determination*
9. *Source for the data set related to breast cancer https://www.kaggle.com/uciml/breast-cancer-wisconsin-data*
10. *Feature scaling formula and reference https://www.geeksforgeeks.org/ml-feature-scaling-part-2/*