# Indian Institute of Technology Kanpur

SESSION:-2014-15

# Regression Analysis Project

**Predicting the chance of recurrence of Breast Cancer and the number of days it will recur in**

Submitted to:
Dr. Sharmishtha Mitra

Submitted by:
Samadipa Saha
Sandeep Begad
Sheallika Singh
Vibhuti Mahajan

**DATA SET:**
Title: Wisconsin Prognostic Breast Cancer (WPBC)
Source: Source Information

a) Creators:

Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792
wolberg@eagle.surgery.wisc.edu

W. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street@cs.wisc.edu 608-262-6619

Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi@cs.wisc.edu

b) Donor: Nick Street

c) Date: December 1995

Number of instances: 198

Number of attributes: 34 (ID, outcome, 32 real-valued input features)

Attribute information

1) ID number
2) Outcome (R = recur, N = nonrecur)
3) Time (recurrence time if field 2 = R, disease-free time if field 2= N)
The Recurrence Surface Approximation (RSA) method is a linear programming model which predicts Time To Recur using both recurrent and nonrecurrent cases.

4-33) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Several of the papers are listed on the site http://www.google.com/url?q=http%3A%2F%2Farchive.ics.uci.edu%2Fml%2Fdatasets%2FBreast%2BCancer%2BWisconsin%2B%2528Prognostic%2529&sa=D&sntz=1&usg=AFQjCNHrBM4U7mSnJME2VjrRiN1IVUbgqQ which contain detailed description of how these features are computed.

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 4 is Mean Radius, field 14 is Radius SE, field 24 is Worst Radius.

Values for features 4-33 are recorded with four significant digits.

34) Tumor size - diameter of the excised tumor in centimeters
35) Lymph node status - number of positive axillary lymph nodes
observed at time of surgery

Missing attribute values:
Lymph node status is missing in 4 cases.

Class distribution: 151 non-recur, 47 recur

# Acknowledgements

We would like to thank our instructor of the course Dr. Sharmishtha Mitra for providing constant guidance and motivation for this project, without which it would have been an impossible task to accomplish.

Next we would like to thank the effort of the UCI machine learning repository team to give access to the innumerable datasets to students like us. A well structured data have been an important facet in the project.

Last but not the least we appreciate each others' contribution and hard work for the fulfillment of the project.

## ABSTRACT

In this project we have tried to predict the number of patients who are at a chance of having the disease again. For this, we have used a logistic regression model to classify the patients into those who shall have breast cancer again in the future and those that shall not. From among the patients in our observation set who show recurrent breast cancer, we have used multiple linear regression model to predict the time period in which the disease shall recur.

The columns in the data set have been coded in the following way:

- rad=radius
- tex=texture
- peri=perimeter
- smooth=smoothness
- compac=compactness
- conc=concavity
- cpts=concave points
- sym=symmetry
- fd=facial dimension
- tum_size=tumor size
- lymph=lymph node status
- m=mean
- se=standard error
- w=”worst” or largest(mean of the three largest values)

So rad_m implies mean radius of the nuclei, rad_se implies standard error from the nuclei radii values, etc.

## Logistic Regression model

We fit a logistic regression model on our training dataset with 156 observations and 33 regressors and obtain the prediction obtained from our training data set[80% of the data]

```
fit=glm(formula = outcome ~ time + rad_m + tex_m + peri_m + area_m +
    smooth_m + compac_m + conc_m + cpts_m + sym_m + fd_m + rad_se +
    tex_se + peri_se + area_se + smooth_se + compac_se + conc_se +
    cpts_se + sym_se + fd_se + rad_w + tex_w + peri_w + area_w +
    smooth_w + compac_w + conc_w + cpts_w + sym_w + fd_w + tum_size +
    lymph, family = "binomial", data = train)
```

The number of observations misclassified =12.
The misclassification percentage=7.69%

The test set consists of 38 data points.The prediction obtained is shown below where 1 denotes that the disease is likely to happen again and 0 specifies otherwise.
[R code:
predict(fit, testdata, type=response) ]

```
157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   1   0
175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192
  0   0   1   0   0   0   0   0   0   0   0   0   0   0   1   0   0   1
193 194
  0   0
```

And the number of observations misclassified=14
The misclassification percentage=36.84%

However the VIF's of some regressors are quite high, indicating high multicollinearity, hence we need to remove multicollinearity from the model by variance decomposition method.
[R code:
>library(faraway)
>vif(fit)
]

```
   time            rad_m          tex_m      peri_m          area_m         smooth_m
  102.38364 63796.38497   217.29630 67398.32418 11982.17207   479.09867
   compac_m     conc_m      cpts_m         sym_m       fd_m        rad_se
 1401.49209  1107.55100  1022.00796   237.12490   441.04466  3762.40524
    tex_se       peri_se        area_se     smooth_se   compac_se     conc_se
  207.63171  3822.36763  1278.34208   249.07073  1169.82759  1509.19851
   cpts_se     sym_se        fd_se        rad_w          tex_w         peri_w
  244.12018   465.43306   799.22344 19129.92818   352.19082  7207.77312
    area_w      smooth_w       compac_w     conc_w      cpts_w       sym_w
10248.26733   264.43651  1132.72258  1047.78158   279.53559   597.17596
      fd_w        tum_size       lymph
 1054.13322    30.96778    40.46549
```

**The model after removing successively removing collinearity using variance decomposition method.**

outcome ~ time + tex_m + compac_m + tex_se + peri_se +
    compac_se + conc_se + cpts_se + sym_se + fd_se + area_w +
    conc_w + tum_size + lymph

The VIF's of the regressors:

```
    time       tex_m     compac_m  tex_se     peri_se    compac_se   conc_se
 17.75276  10.28624  45.40091  17.02112  35.26033  64.50147  60.46200
  cpts_se    sym_se     fd_se        area_w    conc_w     tum_size    lymph
 35.40312  16.86470  40.21718  18.62755  37.72544  10.15669  10.31293
```

The number of observations misclassified in the training data set of the model after removing collinearity=27
The misclassification percentage=17.31%

The prediction on the test data is as below.

```
157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174
  1   0   1   0    0   0    0   0    0   0   1    0  0    0   0    0    0   1
175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192
  1   0   1    1   0    0   0   0    0   1   1   1   1   0   0   1    1   0
193 194
  1   1
```

The number of observations misclassified in the test data set=11
The misclassification percentage=28.94%
We further apply variable selection to reduce the number of regressors.

## **Model after removing collinearity and variable selection**

After successively removing collinearity and variable selection, we obtain the model as:
outcome ~ time + compac_m + peri_se + conc_se + fd_se + conc_w

And the VIF's of the regressors obtained

```
   time      compac_m peri_se   conc_se    fd_se      conc_w
15.28565 35.46256 13.19941 32.81442 23.31598 29.28844
```

The number of observations misclassified in the training data set of the model =29
The misclassification percentage=18.56%

The prediction on the test data is as below.

```
157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174
  1    0    1    0    0    0    0    0    1    1    1    1    0    0    0    0    0    1
175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192
  1    0   1    0    1    1    0    0    0    1    1    1    1    1   1    0    1    0
193 194
  0   1
```

The number of observations misclassified in the test data set=15
The misclassification percentage=39.47%

## MLR Model

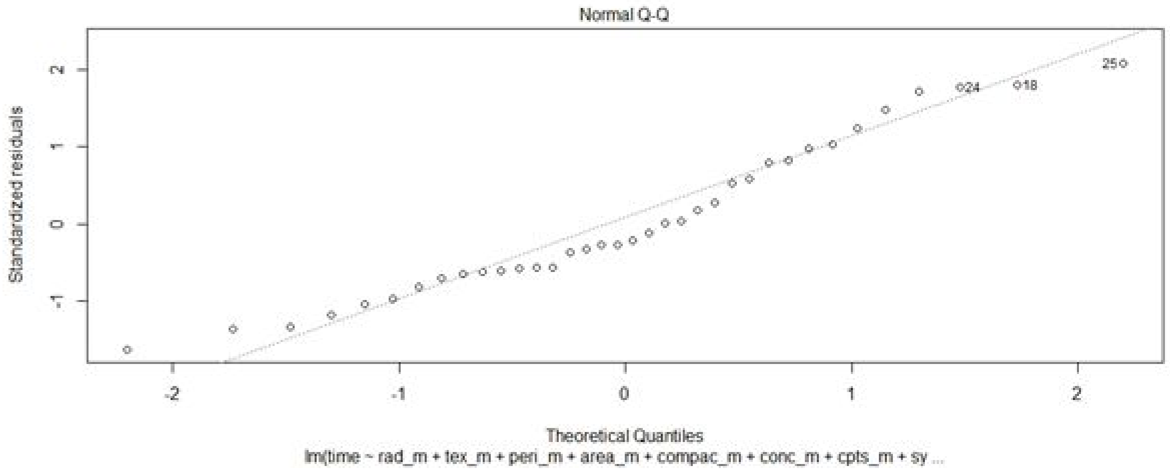
For MLR model, the actual dataset included only 46 observations with the initially 33 regressors. 80% of the data (36 observations) were used for training and rest for testing.
Issues associated:

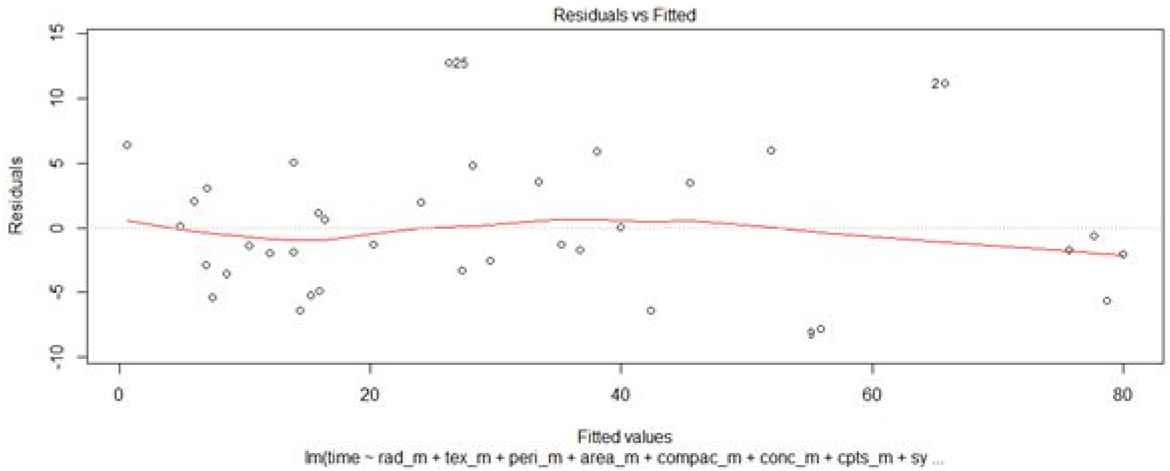
a. High number of regressors i.e n~p
b. Multicollinearity
c. Non-linear model

Step 1: To handle the data more efficiently an initial variable selection was done so that adjusted r-squared was maximum. This was performed using regsubsets() command from “car” library in R.
As a result, 26 variables were selected accordingly and the summary of the fit is as follows:

```
lm(formula = time ~ rad_m + tex_m + peri_m + area_m + compac_m +
   conc_m + cpts_m + sym_m + fd_m + rad_se + tex_se + peri_se +
   smooth_se + compac_se + conc_se + cpts_se + sym_se + tex_w +
   peri_w + area_w + smooth_w + compac_w + conc_w + cpts_w +
   sym_w + fd_w + tum_size, data = train)

Residuals:
   Min   1Q Median     3Q     Max
-7.873 -3.005 -1.278  3.149 12.749

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.299e+02  1.153e+02   1.127 0.292548
rad_m        1.448e+02  4.045e+01   3.579 0.007197 **
tex_m       -4.974e+00  2.573e+00  -1.933 0.089321 .
peri_m      -2.728e+01  6.451e+00  -4.228 0.002884 **
area_m       5.172e-01  1.176e-01   4.399 0.002289 **
compac_m     1.940e+03  4.635e+02   4.185 0.003058 **
conc_m       2.249e+03  3.849e+02   5.844 0.000386 ***
cpts_m      -5.072e+03  7.581e+02  -6.691 0.000154 ***
sym_m        8.418e+02  3.717e+02   2.265 0.053323 .
fd_m        -5.343e+03  2.222e+03  -2.405 0.042872 *
rad_se      -4.875e+02  1.078e+02  -4.523 0.001942 **
tex_se      -4.005e+01  1.767e+01  -2.267 0.053176 .
peri_se      7.601e+01  1.818e+01   4.180 0.003078 **
smooth_se   -1.455e+04  4.182e+03  -3.480 0.008314 **
```

```
compac_se   -8.842e+02  6.444e+02  -1.372 0.207246
conc_se         -7.235e+03  1.228e+03  -5.891 0.000365 ***
cpts_se          2.503e+04  5.042e+03   4.965 0.001100 **
sym_se           2.847e+03  1.255e+03   2.268 0.053031 .
tex_w            5.667e+00  2.175e+00   2.606 0.031343 *
peri_w          -2.251e-01  1.666e+00  -0.135 0.895827
area_w          -4.229e-02  5.642e-02  -0.750 0.475014
smooth_w         2.259e+03  4.994e+02   4.523 0.001942 **
compac_w        -2.693e+02  9.563e+01  -2.816 0.022630 *
conc_w           3.180e+02  1.242e+02   2.561 0.033590 *
cpts_w          -1.582e+03  3.189e+02  -4.960 0.001107 **
sym_w           -5.330e+02  2.164e+02  -2.463 0.039144 *
fd_w             1.012e+03  7.664e+02   1.320 0.223309
tum_size         8.857e+00  1.821e+00   4.864 0.001250 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.07 on 8 degrees of freedom
Multiple R-squared:  0.9588,   Adjusted R-squared:  0.8196
F-statistic:  6.89 on 27 and 8 DF,  p-value: 0.003978
```





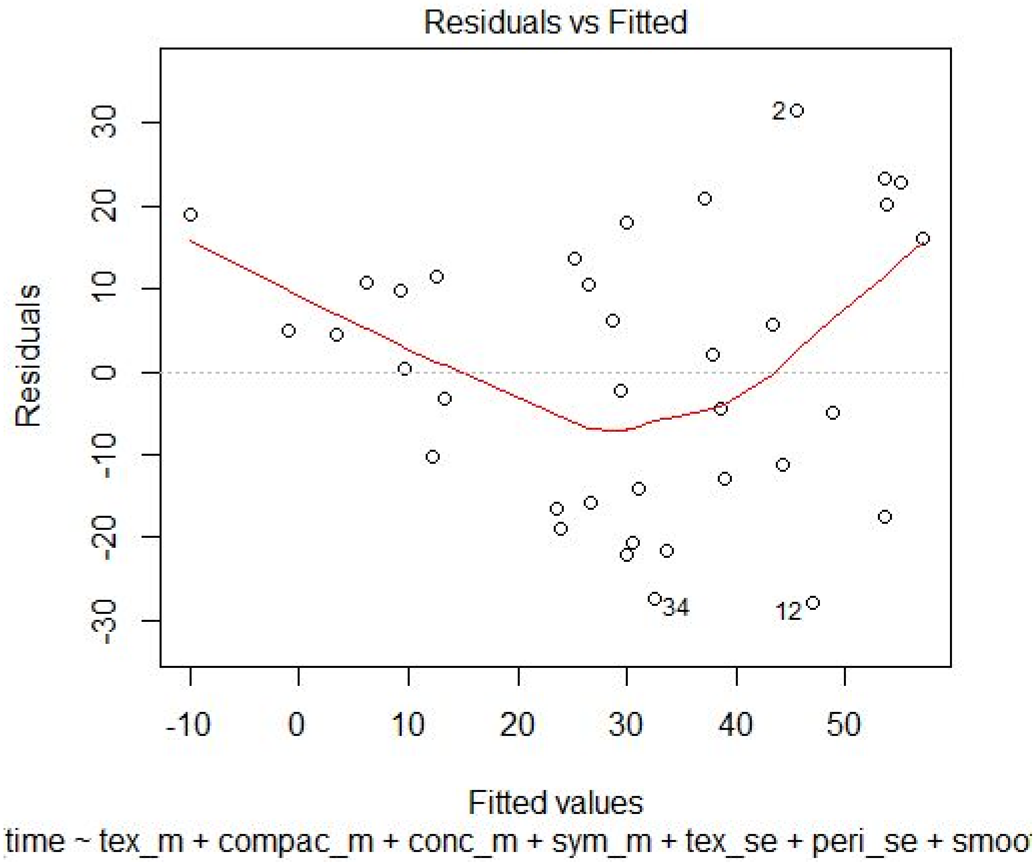
The plots show a relatively better fitted model but the predicted values for the testing data did not give proper results.

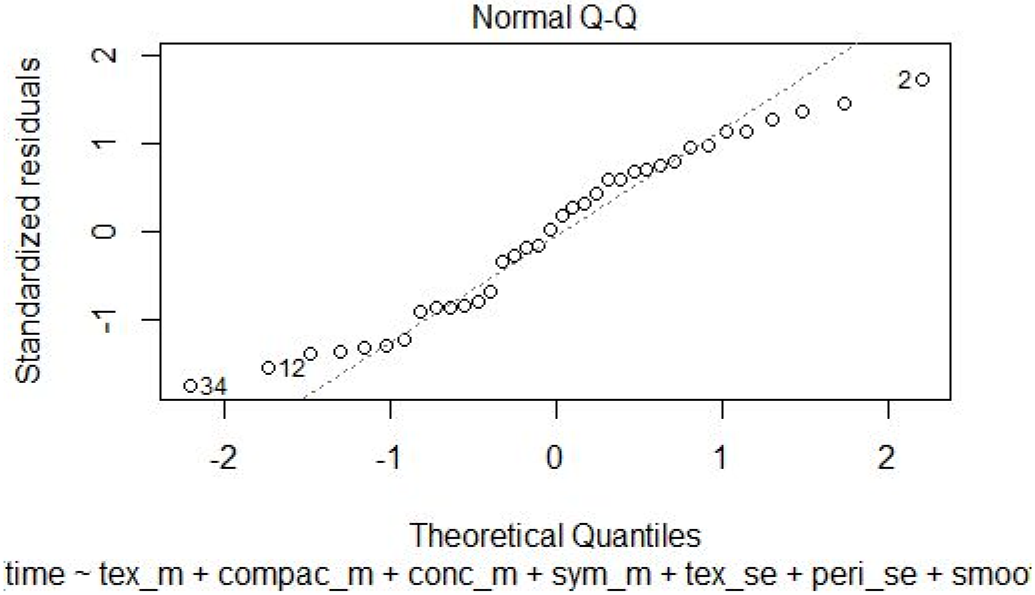
```
pred_init
         37         38         39         40         41         42         43         44         45         46
 -0.695916   5.113816  35.780140 -23.043968  -5.652990  39.342355  99.413660 192.139541
62.405499   6.137621
```

Step2: Next we checked for multicollinearity using variance decomposition method, the same way we did in Logistic model.
Removal of variables was stopped when VIF's reached a value around 20  ( max VIF=21.03), but variance decomposition gave non significant variance decomposition proportions for high VIF regressors.
The model obtained after removing multicollinearity is as follows:

```
Call:
lm(formula = time ~ tex_m + compac_m + conc_m + sym_m + tex_se + 
         peri_se + smooth_se + compac_se + sym_se + area_w + smooth_w + 
         compac_w + conc_w + cpts_w + fd_w + tum_size, data = train)

Residuals:
         Min      1Q  Median       3Q     Max 
-28.004 -14.492   1.234  12.045  31.476 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.496e+01  7.020e+01   0.640   0.5295
tex_m    -1.788e+00  1.530e+00  -1.169   0.2568
compac_m       -3.795e+02  3.842e+02  -0.988   0.3356
conc_m           2.692e+02  2.501e+02   1.076   0.2953
sym_m            3.575e+02  2.654e+02   1.347   0.1939
```

```
tex_se          2.713e+01  1.495e+01   1.815   0.0854 .
peri_se         6.189e+00  4.920e+00   1.258   0.2236
smooth_se   -2.825e+03  4.175e+03  -0.677   0.5068
compac_se   -5.289e+02  7.322e+02  -0.722   0.4789
sym_se          -1.047e+03  7.111e+02  -1.472   0.1574
area_w          -2.456e-02  1.681e-02  -1.461   0.1603
smooth_w        2.571e+02  5.271e+02   0.488   0.6313
compac_w        1.048e+02  1.279e+02   0.820   0.4227
conc_w          -4.978e+01  8.454e+01  -0.589   0.5629
cpts_w  -2.805e+02  2.382e+02  -1.177   0.2536
fd_w     -3.423e+01  6.591e+02  -0.052   0.9591
tum_size        4.182e-01  2.595e+00   0.161   0.8737
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.21 on 19 degrees of freedom
Multiple R-squared:  0.5232,   Adjusted R-squared:  0.1216
F-statistic: 1.303 on 16 and 19 DF,  p-value: 0.2882
```

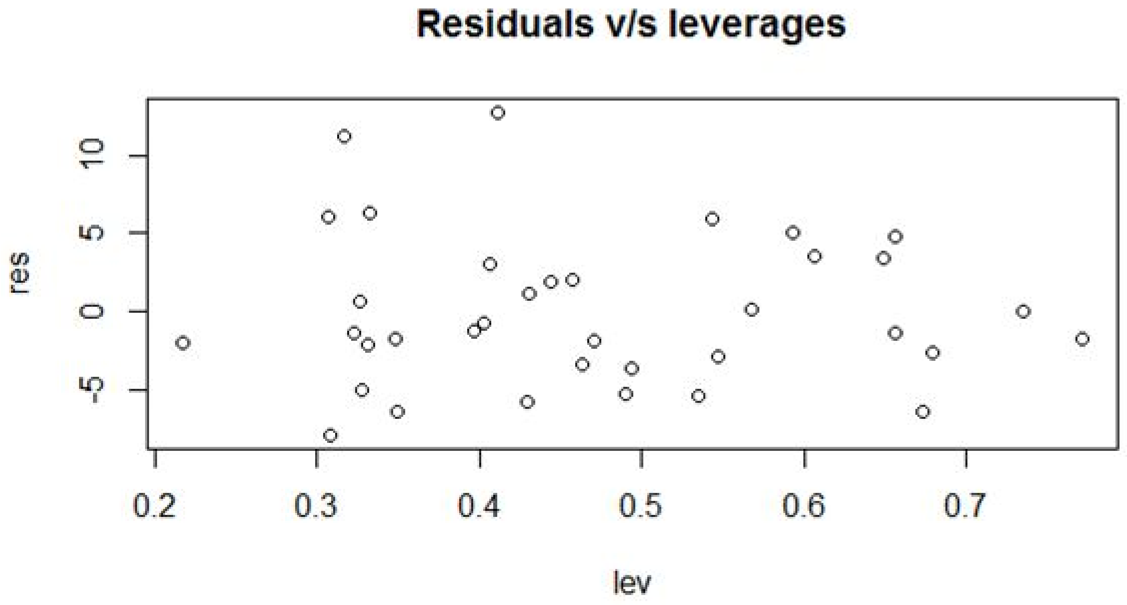


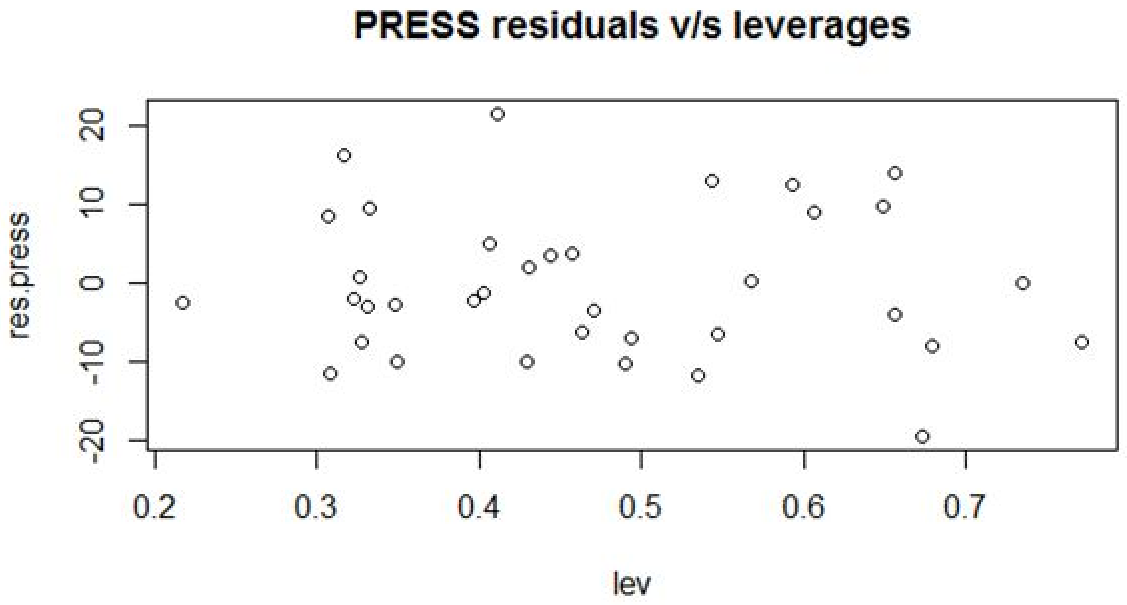


```
   37        38        39        40        41        42        43        44        45 
 1.694452 39.198678 12.546668 22.692806 42.760911 34.724093 29.558598 14.852857 
42.213418 
       46 
 9.471197 
```

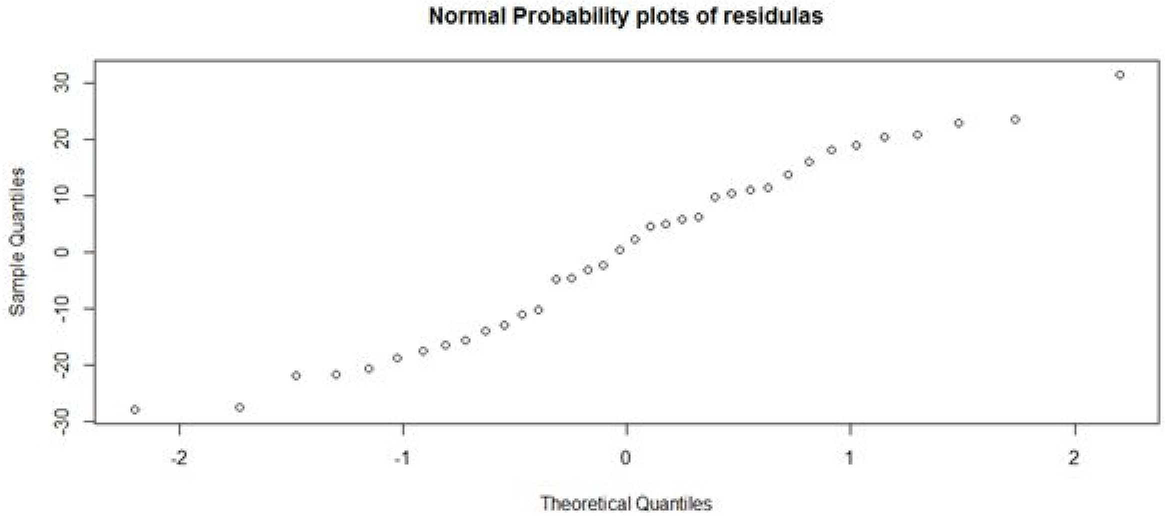
The model did not fit as well as the previous model but  the prediction was better in this case. The residual plots and the q-q plots gave a clear idea of deviation from our assumptions of normality and presence of a linear model.

Step3: Residual analysis was done to check model inadequacies. PRESS residuals were calculat0ed. Plots of leverage values (hii) and residuals were as follows. Also no leverage value came significantly different from others and were close to the mean value of 0.477





No influential observation was recorded as ordinary residuals and Press residuals had similar values.

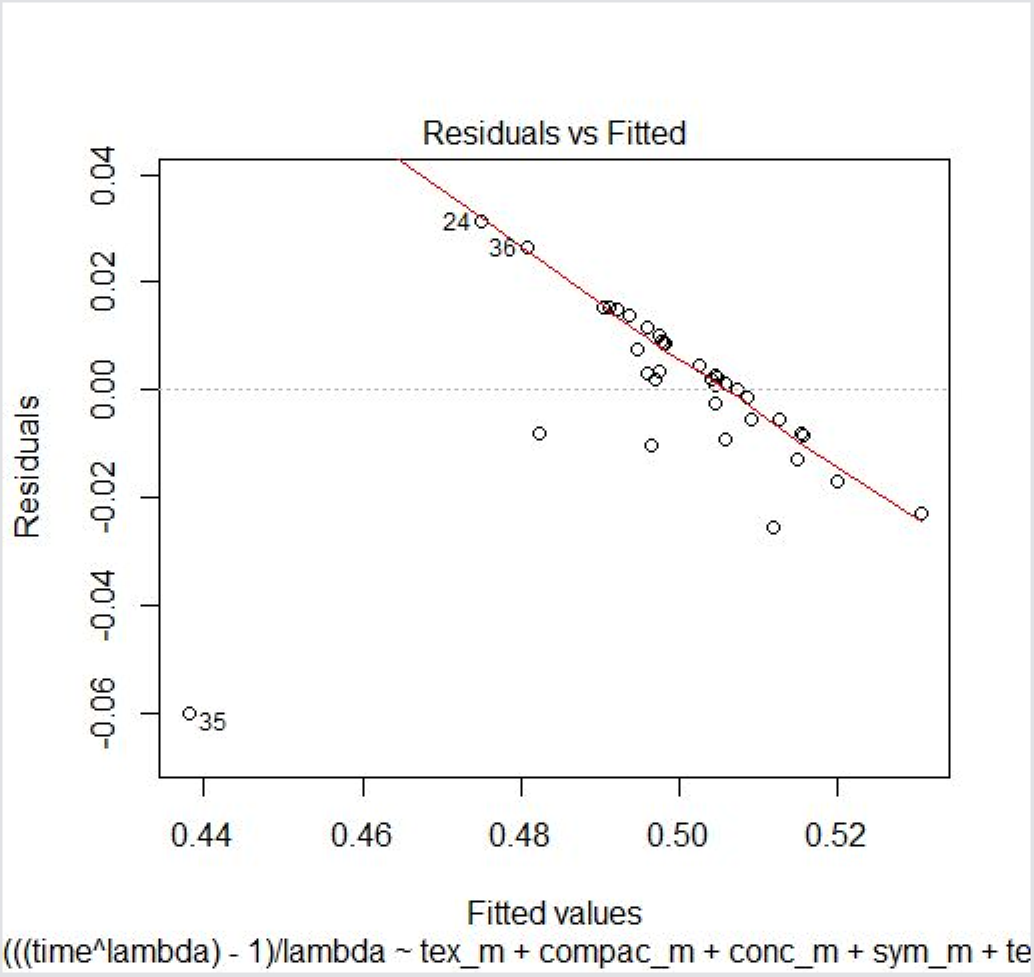


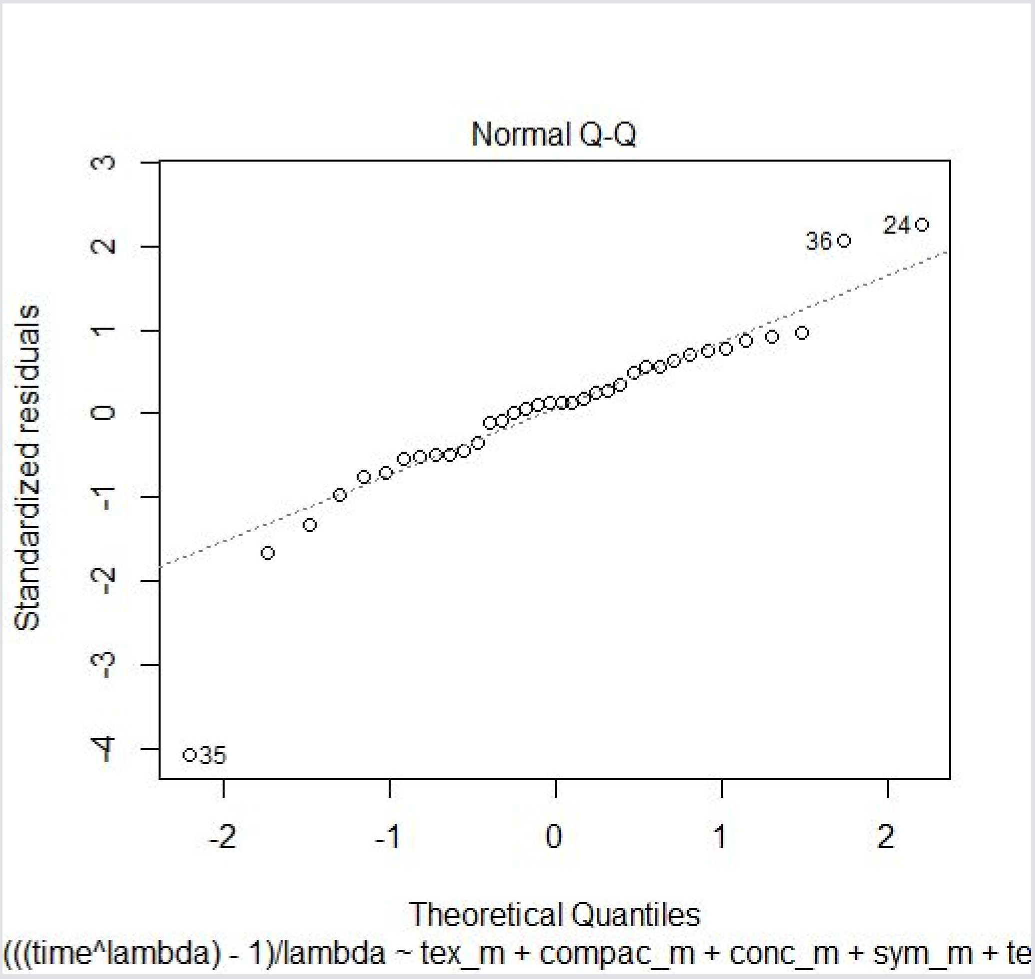
Step4: The residual plot was assumed to be an opening funnel type and box-cox was applied to check for variance dependence. The lambda chosen was -1.9 by subsequent checking of SSRes for each iteration.

The final output showed that the assumption was wrong

```
Call:
lm(formula = ((time^lambda) - 1)/lambda ~ tex_m + compac_m +
        conc_m + sym_m + tex_se + peri_se + smooth_se + compac_se +
        sym_se + area_w + smooth_w + compac_w + conc_w + cpts_w +
        fd_w + tum_size, data = train)

Residuals:
        Min       1Q      Median          3Q       Max
-0.060048 -0.007831  0.002109  0.008929  0.031181

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.569e-01  6.816e-02   6.704 2.08e-06 ***
tex_m   -2.214e-03  1.485e-03  -1.491   0.1524
compac_m        -7.574e-02  3.730e-01  -0.203   0.8412
conc_m          -2.598e-01  2.428e-01  -1.070   0.2981
sym_m           9.306e-02  2.577e-01  0.361   0.7220
tex_se          2.665e-02  1.451e-02  1.837   0.0820 .
peri_se -4.819e-03  4.776e-03  -1.009   0.3257
smooth_se       2.089e+00  4.053e+00   0.515   0.6122
compac_se       4.048e-01  7.109e-01  0.569   0.5758
sym_se          7.445e-02  6.905e-01  0.108   0.9153
area_w          3.070e-05  1.632e-05  1.881   0.0753 .
```

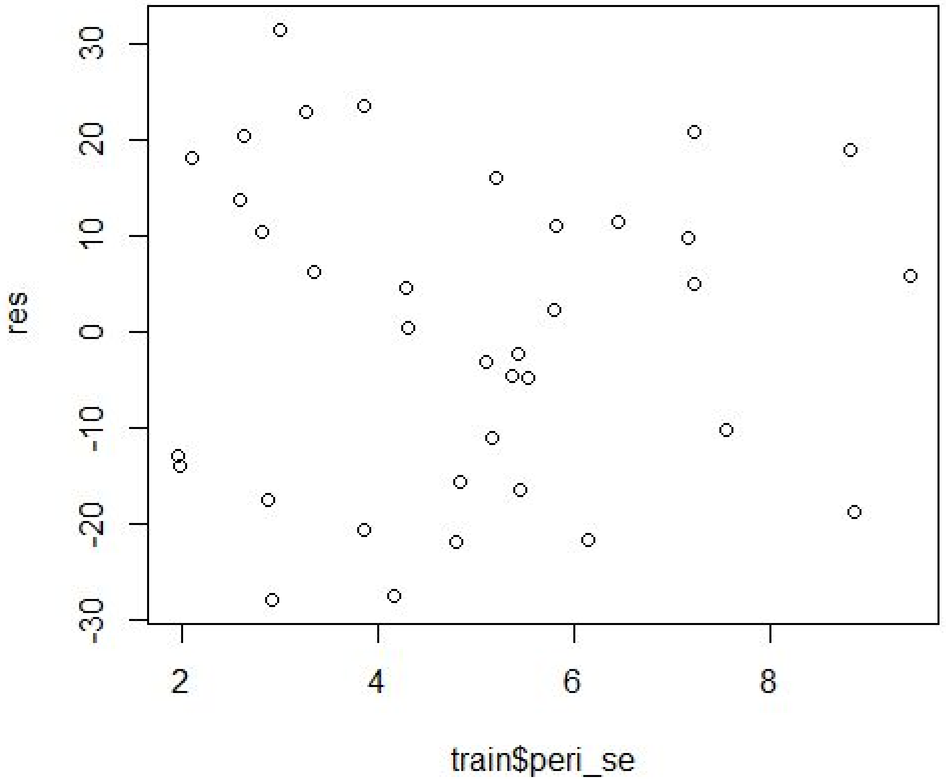
```
smooth_w          2.798e-01  5.118e-01   0.547   0.5909
compac_w          -5.431e-03  1.242e-01  -0.044   0.9656
conc_w            7.446e-02  8.208e-02   0.907   0.3757
cpts_w  -2.730e-01  2.313e-01  -1.180   0.2525
fd_w     1.297e-01  6.399e-01   0.203   0.8415
tum_size          1.793e-03  2.520e-03   0.712   0.4853
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02157 on 19 degrees of freedom
Multiple R-squared:  0.4827,   Adjusted R-squared:  0.04711
F-statistic: 1.108 on 16 and 19 DF,  p-value: 0.411
```





Step5: Non-linearity was assumed finally. The fit of residuals v/s regressor for each regressor( in the model or not) was compared with residuals v/s fitted.

One such example:



Since the number of observations were less, we could not make a proper estimation of the regressor that could be creating the non-linearity.

Results and conclusions.
The final model that was considered was the one after removing multicollinearity as it gave the best predictions among all the models and our basic aim was predictio only. But still there were significant diffrences from the actual values.

```
model: model2<-lm(formula = time ~   tex_m   + compac_m +
             conc_m + sym_m  +  tex_se + peri_se +
             smooth_se + compac_se +  sym_se +
              area_w + smooth_w + compac_w + conc_w + cpts_w +
              fd_w + tum_size, data = train)
```

predictions:

| predicted | actual |
|---|---|
| 1.69 | 1 |
| 39.19 | 9 |
| 12.54 | 16 |
| 22.69 | 9 |
| 42.760 | 14 |
| 34.724 | 12 |
| 29.55 | 11 |
| 14.85 | 7 |
| 42.21 | 14 |
| 9.47 | 1 |

The data available is not a good dataset for prediction as it gives just 10% accuracy, but it is a good dataset to predict whether the cancer will recur in a particular patient in due time.
Some other parameters would have been needed for making a better prediction.