## Preparing data

In Breast Cancer Wisconsin (Prognostic) Data Set Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis.

These are attributes that sample-code-number is just an ID and we don't count it a feature, and the class attribute is our output for regression.

```
#  Attribute                     Domain
-- -----------------------------------------
1. Sample code number            id number
2. Clump Thickness               1 - 10
3. Uniformity of Cell Size       1 - 10
4. Uniformity of Cell Shape      1 - 10
5. Marginal Adhesion             1 - 10
6. Single Epithelial Cell Size   1 - 10
7. Bare Nuclei                   1 - 10
8. Bland Chromatin               1 - 10
9. Normal Nucleoli               1 - 10
10. Mitoses                      1 - 10
11. Class:                       (2 for benign, 4 for malignant)
```

First of all we need to remove the Data that have missing values (16 row).

# Analysis

## Linear regression

We have 9 Features, we calculated the regression for this features, and save the result of them in Results Data

In this Model we have coefficient from 0.23, all p-values have the same values, so they have same importance. There is some thing that surprise us and it was the small coefficients for all features, but we understood that because of our X's that are in {1,2,3,4,5,6,7,8,,9,10} and our y that are in {2,4} the coefficients always are near 0.2

- R-squared: R-squared is a statistical measure of how close the data are to the fitted regression line
- p-value : When we perform a hypothesis test in statistics, a p-value helps us determine the significance of your results.
- Adjusted R-squared : The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance

- F-statistics : The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. The value of Prob(F) is the probability that the null hypothesis for the full model is true ( that all of the regression coefficients are zero).

1. bareNuclei

```
                        OLS Regression Results                            
=============================================================================
=
Dep. Variable:                  class   R-squared:                       
0.677
Model:                            OLS   Adj. R-squared:                  
0.676
Method:                 Least Squares   F-statistic:                     
1426.
Date:                Sun, 14 Oct 2018   Prob (F-statistic):          
3.40e-169
Time:                        18:21:25   Log-Likelihood:                
-551.15
No. Observations:                 683   AIC:                             
1106.
Df Residuals:                     681   BIC:                             
1115.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      
0.975]
-----------------------------------------------------------------------------
-
const          1.9359      0.029     66.754      0.000       1.879       
1.993
bareNuclei     0.2155      0.006     37.766      0.000       0.204       
0.227
=============================================================================
=
Omnibus:                      218.770   Durbin-Watson:                   
1.791
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              
796.545
Skew:                           1.479   Prob(JB):                    
1.08e-173
Kurtosis:                       7.386   Cond. No.                         
7.23
=============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.677 do the data almost have a good fit with regression line Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

2. clump thickness

```
                        OLS Regression Results
=============================================================================
=
Dep. Variable:                  class   R-squared:
0.511
Model:                            OLS   Adj. R-squared:
0.510
Method:                 Least Squares   F-statistic:
711.4
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
7.29e-108
Time:                        18:21:25   Log-Likelihood:
-692.64
No. Observations:                 683   AIC:
1389.
Df Residuals:                     681   BIC:
1398.
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
=====
                     coef    std err          t      P>|t|      [0.025
0.975]
-----------------------------------------------------------------------------
-----
const              1.6253      0.048     34.065      0.000       1.532
1.719
clumpThickness     0.2419      0.009     26.673      0.000       0.224
0.260
=============================================================================
=
Omnibus:                       46.918   Durbin-Watson:
1.742
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
54.786
Skew:                           0.676   Prob(JB):
1.27e-12
```

# Kurtosis: 3.314 Cond. No. 10.1

p-value is 0.000 so this feature is significant R-squared is 0.511 do the data almost have a good fit with regression line Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

3. uniformity of cell size

```
                            OLS Regression Results
=============================================================================
=
Dep. Variable:                  class   R-squared:
0.674
Model:                            OLS   Adj. R-squared:
0.673
Method:                 Least Squares   F-statistic:
1406.
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
8.92e-168
Time:                        18:21:25   Log-Likelihood:
-554.42
No. Observations:                 683   AIC:
1113.
Df Residuals:                     681   BIC:
1122.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===========
                           coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
const                    1.8944      0.030     63.242      0.000       1.836
1.953
uniformityOfCellSize     0.2556      0.007     37.498      0.000       0.242
0.269
==============================================================================
=
Omnibus:                      141.261   Durbin-Watson:
1.769
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
268.178
Skew:                           1.191   Prob(JB):
5.83e-59
Kurtosis:                       4.937   Cond. No.
6.48
==============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.674 do the data almost have a good fit with regression line Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

4. uniformity of cell shape

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  class   R-squared:
0.676
Model:                            OLS   Adj. R-squared:
0.675
Method:                 Least Squares   F-statistic:
1418.
```

```
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
1.37e-168
Time:                        18:21:25   Log-Likelihood:
-552.54
No. Observations:                 683   AIC:
1109.
Df Residuals:                     681   BIC:
1118.
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
============
                            coef    std err          t      P>|t|      [0.025
0.975]
-----------------------------------------------------------------------------
------------
const                     1.8558      0.031     60.654      0.000       1.796
1.916
uniformityOfCellShape     0.2625      0.007     37.652      0.000       0.249
0.276
=============================================================================
=
Omnibus:                      111.909   Durbin-Watson:
1.849
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
189.327
Skew:                           1.013   Prob(JB):
7.73e-42
Kurtosis:                       4.595   Cond. No.
6.63
=============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.675 do the data almost have a good fit with regression line Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

5. marginal adhesion

```
                            OLS Regression Results
=============================================================================
=
Dep. Variable:                  class   R-squared:
0.499
Model:                            OLS   Adj. R-squared:
0.498
Method:                 Least Squares   F-statistic:
677.9
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
2.98e-104
Time:                        18:21:25   Log-Likelihood:
-700.97
No. Observations:                 683   AIC:
1406.
```

```
Df Residuals:                     681   BIC:
1415.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=======
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-------
const                2.0337      0.036     55.888      0.000       1.962
2.105
marginalAdhesion     0.2354      0.009     26.036      0.000       0.218
0.253
==============================================================================
=
Omnibus:                      126.961   Durbin-Watson:
1.629
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
199.005
Skew:                           1.240   Prob(JB):
6.12e-44
Kurtosis:                       3.915   Cond. No.
5.84
==============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.498 do the data fitted with regression line not -bad Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

6. single epithelial cell size

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  class   R-squared:
0.477
Model:                            OLS   Adj. R-squared:
0.477
Method:                 Least Squares   F-statistic:
622.2
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
4.73e-98
Time:                        18:21:25   Log-Likelihood:
-715.27
No. Observations:                 683   AIC:
1435.
Df Residuals:                     681   BIC:
1444.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===============
                               coef    std err          t      P>|t|
[0.025      0.975]
-----------------------------------------------------------------------------
---------------
const                        1.7403      0.047     37.287      0.000
1.649       1.832
singleEpithelialCellSize     0.2967      0.012     24.943      0.000
0.273       0.320
=============================================================================
=
Omnibus:                       68.911   Durbin-Watson:
1.769
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
87.713
Skew:                           0.831   Prob(JB):
8.98e-20
Kurtosis:                       3.564   Cond. No.
7.24
=============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.477 do the data fitted with regression line not -bad Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

7. bland chromatin

```
                            OLS Regression Results
=============================================================================
=
Dep. Variable:                  class   R-squared:
0.575
Model:                            OLS   Adj. R-squared:
0.574
Method:                 Least Squares   F-statistic:
921.0
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
1.27e-128
Time:                        18:21:25   Log-Likelihood:
-644.76
No. Observations:                 683   AIC:
1294.
Df Residuals:                     681   BIC:
1303.
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
=====
                     coef    std err          t      P>|t|      [0.025
0.975]
-----------------------------------------------------------------------------
-----
const              1.6820      0.041     40.878      0.000       1.601
1.763
```

```
blandChromatin     0.2955      0.010     30.348      0.000       0.276
0.315
==============================================================================
=
Omnibus:                       84.860   Durbin-Watson:
1.821
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
118.645
Skew:                           0.899   Prob(JB):
1.72e-26
Kurtosis:                       3.968   Cond. No.
7.57
==============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.575 do the data almost have a good fit with regression line Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

8. normal nucleoli

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  class   R-squared:
0.516
Model:                            OLS   Adj. R-squared:
0.516
Method:                 Least Squares   F-statistic:
727.5
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
1.47e-109
Time:                        18:21:25   Log-Likelihood:
-688.73
No. Observations:                 683   AIC:
1381.
Df Residuals:                     681   BIC:
1391.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=====
                     coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-----
const              2.0549      0.035     58.886      0.000       1.986
2.123
normalNucleoli     0.2247      0.008     26.972      0.000       0.208
0.241
==============================================================================
=
Omnibus:                      147.506   Durbin-Watson:
1.848
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
254.504
Skew:                           1.328   Prob(JB):
5.43e-56
Kurtosis:                       4.374   Cond. No.
5.91
==============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.677 do the data almost have a good fit with regression line Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

9. mitoses

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  class   R-squared:
0.179
Model:                            OLS   Adj. R-squared:
0.178
Method:                 Least Squares   F-statistic:
148.8
Date:                Sun, 14 Oct 2018   Prob (F-statistic):
4.30e-31
Time:                        18:21:25   Log-Likelihood:
-869.41
No. Observations:                 683   AIC:
1743.
Df Residuals:                     681   BIC:
1752.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-
const          2.3258      0.045     51.536      0.000       2.237
2.414
mitoses        0.2333      0.019     12.198      0.000       0.196
0.271
==============================================================================
=
Omnibus:                      226.302   Durbin-Watson:
1.665
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
110.185
Skew:                           0.839   Prob(JB):
1.18e-24
Kurtosis:                       1.972   Cond. No.
3.51
```

```
==============================================================================
=
```

p-value is 0.000 so this feature is significant R-squared is 0.179 do the data fitted with regression line so bad Prob(f-statistics) is so small so the null hypothesis for features in this model is not true

### plots

As you see our data you understand that all of our data X are 1 or 2 or 3, ...or 10 , and all y are 2 or 4 so our coefficient will be so small because y to x ratio is small. and all figures for our features are like this , because all plot is just 20 point, and it's because of our data set

## Multiple Linear Regression

As we see in Linear Regression all features were significant, but if we want to know that if the features are truly significant and the effects are not from other features, we calculate the multiple regression with these features and we face with this

All features multiple regression result

```
Dep. Variable:                  class   R-squared:                       0.843
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     402.5
Date:                Wed, 10 Oct 2018   Prob (F-statistic):         4.46e-264
Time:                        20:24:42   Log-Likelihood:                -303.90
No. Observations:                 683   AIC:                             627.8
Df Residuals:                     673   BIC:                             673.1
Df Model:                           9
Covariance Type:            nonrobust

                             coef          std err       t         P>|t|       [0.025
0.975]
const                        1.5047        0.033     45.807      0.000        1.440
1.569
clumpThickness               0.0634      0.007      8.898      0.000       0.049
0.077
uniformityOfCellSize         0.0437      0.013      3.428      0.001       0.019
0.069
uniformityOfCellShape        0.0313      0.012      2.508      0.012       0.007
0.056
marginalAdhesion             0.0165      0.008      2.065      0.039       0.001
0.032
singleEpithelialCellSize     0.0202      0.010      1.924      0.055      -0.000
0.041
bareNuclei                   0.0908      0.006     14.091      0.000       0.078
0.103
blandChromatin               0.0384      0.010      3.802      0.000       0.019
0.058
normalNucleoli               0.0371      0.007      4.981      0.000       0.022
0.052
```

```
mitoses                       0.0020      0.010      0.197      0.844      -0.018
0.021
```

So the mitoses and singleEpithelialCellSize have p-value > 0.05 then these are insignificant features and we can remove them and again calculate the multiple linear regression with other 7 features and get these. and prob(F-statistic) is significant, this means at least one feature has relationship with response.

All significant features multiple regression result

```
Dep. Variable:                  class   R-squared:                       0.842
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     515.4
Date:                Thu, 11 Oct 2018   Prob (F-statistic):         6.47e-266
Time:                        21:22:05   Log-Likelihood:                -305.95
No. Observations:                 683   AIC:                             627.9
Df Residuals:                     675   BIC:                             664.1
Df Model:                           7
Covariance Type:            nonrobust

                          coef        std err     t          P>|t|      [0.025
0.975]
const                     1.5318      0.030     51.224      0.000       1.473
1.591
clumpThickness            0.0638      0.007      8.960      0.000       0.050
0.078
uniformityOfCellSize      0.0504      0.012      4.096      0.000       0.026
0.075
bareNuclei                0.0913      0.006     14.187      0.000       0.079
0.104
blandChromatin            0.0386      0.010      3.833      0.000       0.019
0.058
normalNucleoli            0.0393      0.007      5.359      0.000       0.025
0.054
uniformityOfCellShape     0.0331      0.012      2.654      0.008       0.009
0.058
marginalAdhesion          0.0177      0.008      2.237      0.026       0.002
0.033
```

As we see in second multiple regression, R-squared decreases but this is natural because when number of features decrease then R-squared decreases too, so the best way is to compare Adjusted R-squared in two models that in these tho models are equal (0.841), so it tells us removing 2 feature doesnt decrease R-squared too much so they are not important features. In other hand we can compare Prob(F-statistics) too, that when removing 2 Features the prob(F-statistics) decreases so we can say that this deleting features give us better result.

# Regularization

When Features or samples are too much , the over fitting problem may happen so we must regularization the features and remove or some features that dont't have significant p-values or shrink samples data, we use 3 ways to regularize our data in this assignment

For use the best aphpha in formula we check from 0.01 to 100 in a loop to find best R-squared then choose that alpha

## Ridge

Ridge regression Result

```
Alpha  = 0.09
Alpha  = 0.1

R-squared  = 0.8433241288874997

                    Feature  Coefficients  t values  Standard Errors  Probabilites
0                 constants        1.5047    45.807            0.033         0.000
1            clumpThickness        0.0634     8.898            0.007         0.000
2      uniformityOfCellSize        0.0437     3.428            0.013         0.001
3     uniformityOfCellShape        0.0313     2.508            0.012         0.012
4          marginalAdhesion        0.0165     2.065            0.008         0.039
5  singleEpithelialCellSize        0.0202     1.924            0.010         0.055
6                bareNuclei        0.0908    14.091            0.006         0.000
7            blandChromatin        0.0384     3.801            0.010         0.000
8            normalNucleoli        0.0371     4.981            0.007         0.000
9                   mitoses        0.0020     0.197            0.010         0.844
```

The Ridge method is based on the restriction of βi , and the value of α is small, So we can conclude that compression is low. the p-value for mitoses (0.844) and singleEpithelialCellSize(0.055) are not small enough so in this model these features are not significant and can be remove We also observe that the p-value of other parameters in the ridge method and the multiple linear regression almost is the same, so we find that the parameters in the two methods are equally important.

## Lasso

In Lasso method, we see that the mitosis coefficient is zero, so we find that this parameter has been omitted. By comparing the two methods of Lasso and Multiple Linear Regression, we observe that the p-value of marginalAdhesion , singleEpithelialCellSize , blandChromatin , the coefficient of mitoses in the lasso method is zero so we know in this model which means that the significance of these parameters is zero in this method, singleEpithelialCellSize p-value increase to ( 0.876) so in this model singleEpithelialCellSize is not significant, also the p-value in uniformityOfCellSize has been reduced, so it has been increased in the lasso method, and in other cases p-value is the same, we find out that the parameters have the same importance. Lasso regression Result

```
Alpha  = 0.1

R-squared  = 0.8402016223438626
```

```
                    Feature  Coefficients  t values  Standard Errors  Probabilites
0                 constants        1.6064    48.423            0.033         0.000
1            clumpThickness        0.0558     7.757            0.007         0.000
2       uniformityOfCellSize        0.0553     4.293            0.013         0.000
3     uniformityOfCellShape        0.0315     2.502            0.013         0.013
4          marginalAdhesion        0.0116     1.436            0.008         0.152
5  singleEpithelialCellSize        0.0017     0.156            0.011         0.876
6                bareNuclei        0.0952    14.640            0.007         0.000
7            blandChromatin        0.0256     2.514            0.010         0.012
8            normalNucleoli        0.0370     4.918            0.008         0.000
9                   mitoses        0.0000     0.000            0.010         1.000
```

## Elastic net

Elastic net regression Result The elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. like Lasso results we can see that also in this model mitoses coefficient is zero and has been removed.

```
Alpha  = 0.1

R-squared  = 0.8425236919081593

                    Feature  Coefficients  t values  Standard Errors  Probabilites
0                 constants        1.5567    47.270            0.033         0.000
1            clumpThickness        0.0594     8.312            0.007         0.000
2      uniformityOfCellSize        0.0488     3.819            0.013         0.000
3     uniformityOfCellShape        0.0320     2.559            0.013         0.011
4          marginalAdhesion        0.0145     1.811            0.008         0.071
5  singleEpithelialCellSize        0.0115     1.095            0.011         0.274
6                bareNuclei        0.0924    14.308            0.006         0.000
7            blandChromatin        0.0320     3.164            0.010         0.002
8            normalNucleoli        0.0370     4.961            0.007         0.000
9                   mitoses        0.0000     0.000            0.010         1.000
```

## Conclusion

As we see in results of above models we see that the R-squared in Lasso, Ridge and Elastic net doesnt increase , so for our data-set the least square regression and regularization method almos have same result

## References

- Simple and Multiple Linear Regression in Python

- How to Interpret the F-test of Overall Significance in Regression Analysis