Predicting Breast-Cancer using Logistic, Probit and Similar Regression Techniques

End-semester project
Statistical Methods 3

Indian Statistical Institute, Kolkata

Fall, 2022

Srijan Chattopadhyay (BS2126)
Alaap Kumar Mukhopadhyay(BS2125)

Table of Contents

1. Aim of the Project
2. Theoretical Background
3. Data Selection and Splitting
4. Fitted Models
5. Analysis and Inference

Aim of the Project

Our goal in this project is to find a suitable regression model we can use to predict the presence of breast cancer in a person given certain predictors. We fit our data into four different regression models, namely, logit, probit, t-distribution and double exponential, and try to find out which model provides us with the best results in terms of (?) metric. We will see and define some criteria and use some result to reach at a conclusion in terms of fitting the best model(?) that we can also use for our future purpose. Whenever needed, we will do some estimation and hypothesis testing.

Theoretical Background of the regression models

Logistic Regression: Here the model assumption is $y_i \sim$ Bern(p_i), where $E(y_i|x_{i1},..,x_{i(k-1)}) = p_i$. So, $log(\frac{p_i}{1-p_i}) = x_i^T\beta$. We perform the mle estimation of β. Our log likelihood is $l(\beta|data) = \sum_{i=1}^{n} y_i lnp_i + (1-y_i)ln(1-p_i) =$ $\sum_{i=1}^{n} y_i(x_i^T\beta) - ln(1+e^{x_i^T\beta}$ Then we have to find the roots of $\frac{dl(\beta)}{d\beta} = 0$.For that different iteration method like Fisher-Scoring or IRLS etc. are used.

Probit Regression: It is the same as that of logistic, except that here we use $p_i = \Phi(x_i^T\beta)$. And then we perform the same process as of the previous.

Theoretical Background of the regression models

Something Similar: Inspired by the idea of probit regression, it is very clear that in spite of normal, if we take any symmetric distribution, then that will also work. So, we also tried this just by replacing the normal assumption by another distributional assumption, i.e. replacing Φ by F in probit, where F is the corresponding CDF.So $p_i = F({x_i}^T \beta)$ We tried with t-distribution with different degrees of freedom and double exponential. We didn't consider cauchy becuase it doesn't have any finite moments and hence the results won't be consistent any more.

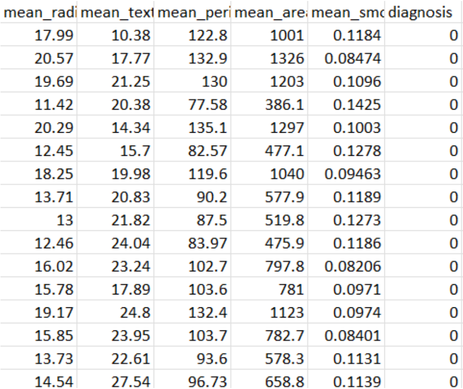
Data Selection and Splitting

Our data consists of 569 rows and 6 columns namely "mean radius","mean texture","mean perimeter","mean area","mean smoothness","diagnosis". The dataset is available at https://www.kaggle.com/datasets/merishnasuwal/breast-cancer-prediction-dataset.

Before fitting the model, we have to split the dataset first into two pieces- Training data and Testing data. We have done that in the following way. For cross validation (to avoid selection bias), we first took a random permutation of the rows and then took the first 469 rows as training data and rest 100 rows for testing. This gives us the splitting of the whole dataset. Now we move on to the fitting of the models.

Data Selection and Splitting

The data looks like this



mean_radi	mean_text	mean_peri	mean_are	mean_smc	diagnosis
17.99	10.38	122.8	1001	0.1184	0
20.57	17.77	132.9	1326	0.08474	0
19.69	21.25	130	1203	0.1096	0
11.42	20.38	77.58	386.1	0.1425	0
20.29	14.34	135.1	1297	0.1003	0
12.45	15.7	82.57	477.1	0.1278	0
18.25	19.98	119.6	1040	0.09463	0
13.71	20.83	90.2	577.9	0.1189	0
13	21.82	87.5	519.8	0.1273	0
12.46	24.04	83.97	475.9	0.1186	0
16.02	23.24	102.7	797.8	0.08206	0
15.78	17.89	103.6	781	0.0971	0
19.17	24.8	132.4	1123	0.0974	0
15.85	23.95	103.7	782.7	0.08401	0
13.73	22.61	93.6	578.3	0.1131	0
14.54	27.54	96.73	658.8	0.1139	0

Figure: Data

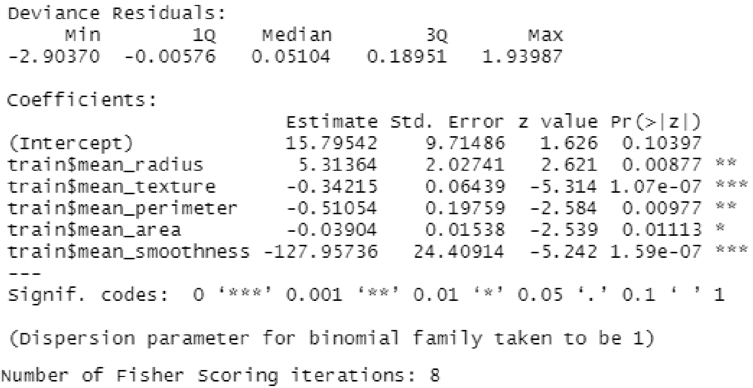
Data Description

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

- **Mean Radius**: Mean of distances from center to points on the perimeter
- **Mean Texture**: Standard deviation of gray-scale values
- **Mean Perimeter**
- **Mean Area**
- **Mean Smoothness**: Local variation in radius lengths
- **Diagnosis**

Fitted Models

Logistic Regression Model:

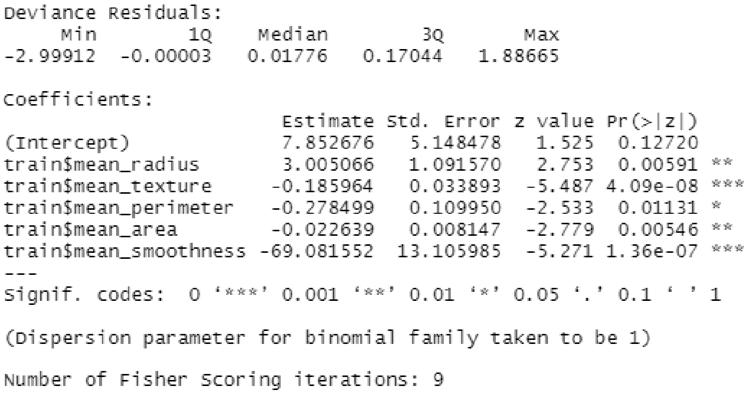


```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.90370  -0.00576   0.05104   0.18951   1.93987

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             15.79542    9.71486   1.626  0.10397
train$mean_radius        5.31364    2.02741   2.621  0.00877 **
train$mean_texture      -0.34215    0.06439  -5.314 1.07e-07 ***
train$mean_perimeter    -0.51054    0.19759  -2.584  0.00977 **
train$mean_area         -0.03904    0.01538  -2.539  0.01113 *
train$mean_smoothness -127.95736   24.40914  -5.242 1.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Number of Fisher Scoring iterations: 8
```

Figure: Summary of Logit Regression Model

Fitted Models

Probit Regression Model:

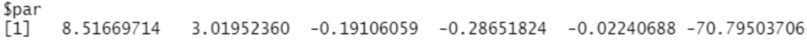


```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.99912  -0.00003   0.01776   0.17044   1.88665

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             7.852676   5.148478   1.525  0.12720
train$mean_radius       3.005066   1.091570   2.753  0.00591 **
train$mean_texture     -0.185964   0.033893  -5.487 4.09e-08 ***
train$mean_perimeter   -0.278499   0.109950  -2.533  0.01131 *
train$mean_area        -0.022639   0.008147  -2.779  0.00546 **
train$mean_smoothness -69.081552  13.105985  -5.271 1.36e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Number of Fisher Scoring iterations: 9
```

Figure: Summary of Probit Regression Model

Fitted Models

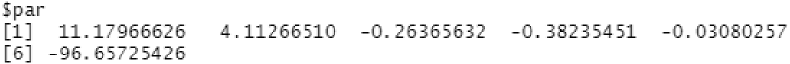
T-Distribution Regression Model:



```
$par
[1]   8.51669714   3.01952360  -0.19106059  -0.28651824  -0.02240688 -70.79503706
```

Figure: Coefficients of the T-Distribution Regression Model

Double Exponential Regression Model:

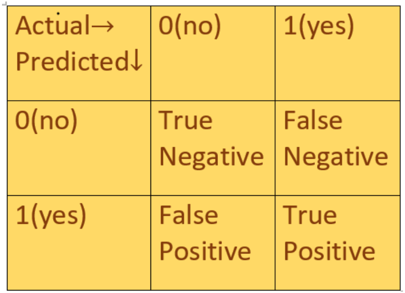


```
$par
[1]  11.17966626   4.11266510  -0.26365632  -0.38235451  -0.03080257
[6] -96.65725426
```

Figure: Coefficients of the Double Exponential Regression Model

Analysis and Inference

For judging how much effective the model is we can try to see the confusion matrix whose elements are true negative(TN), false positive(FP), false negative(FN), true positive(TP). Here is how it looks like:



Actual→ Predicted↓	0(no)	1(yes)
0(no)	True Negative	False Negative
1(yes)	False Positive	True Positive

Figure: Confusion Matrix

Analysis and Inference

Here are different mostly used measures w.r.t. which we can judge a model from the confusion matrix:

1. **Accuracy:**$=\frac{TP+TN}{TP+FP+TN+FN}$
2. **Sensitivity:**$=\frac{TP}{TP+FN}$
3. **Specificity:**$=\frac{TN}{TN+FP}$

But, if we choose model considering a certain metric of the matrix, then we may be misguided, i.e. the decision may vary from person to person due to different psychology. For example, one may think that after admitting a non cancer person into a hospital, he/she may die out of tension. Also another may think from a different perspective. So, clearly there is no universal conclusion if we use these type of measures.

Analysis and Inference

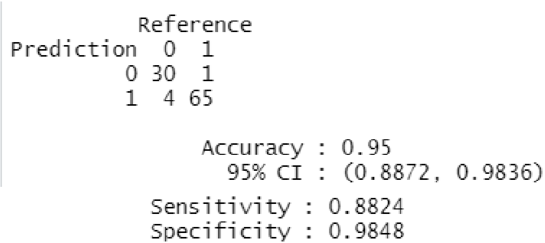
To get rid of this, we can move into two paths. The **first one** is to use Matthews correlation coefficient (MCC). As proved in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312/, "**The Matthews correlation coefficient (MCC)**, instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset". Hence we can judge based on the MCC, which is $\frac{TP.TN-FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. Intuitively, we can see that, here four balance ratios are considered, where as it is not there in the previous metrics, due to that, it is more informative about the matrix.

Analysis and Inference

The **second pathway** is to maximize total (economical) revenue in terms of cost, penalty, advantage etc. It depends upon experience, correct information about the field etc. But for our data, as we don't have correct information about the price of treatment etc., so we won't consider this pathway. Rather, we will try to select the **best model(possibly)** based on the first pathway.
So, we will first take the model(i.e., cutoff) which has the highest MCC for each regression and then compare among them again by MCC to reach our conclusion. Cutoff means the probability above which we tell it 1.

Analysis and Inference

Logistic Regression Model:



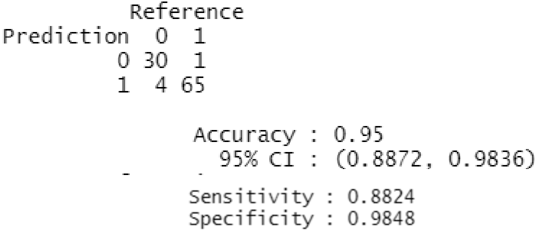
```
          Reference
Prediction  0  1
         0 30  1
         1  4 65

               Accuracy : 0.95
                 95% CI : (0.8872, 0.9836)

            Sensitivity : 0.8824
            Specificity : 0.9848
```

Figure: Confusion Matrix and related metrics of Logit Regression Model at 0.18 cutoff

The MCC obtained here is maximum(0.8882312) here among all the cutoffs. So, this is the best model for logistic regression(according to MCC metric)

Analysis and Inference

Probit Regression Model:



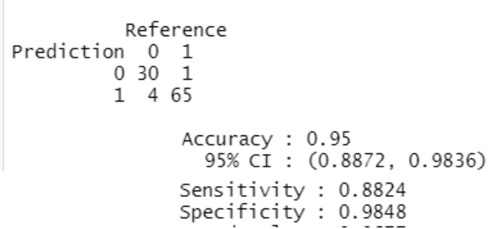
```
          Reference
Prediction  0  1
         0 30  1
         1  4 65

               Accuracy : 0.95
                 95% CI : (0.8872, 0.9836)

            Sensitivity : 0.8824
            Specificity : 0.9848
```

Figure: Confusion Matrix and related metrics of Probit Regression Model at 0.20 cutoff

The MCC obtained here is maximum(0.8882312) here among all the cutoffs. So, this is the best model for Probit regression(according to MCC metric)

Analysis and Inference

T-Distribution Regression Model:



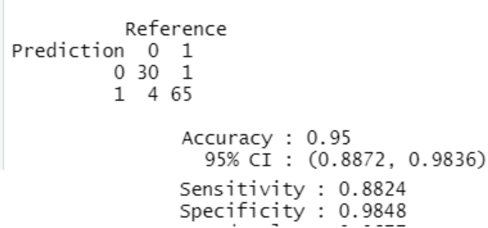
```
           Reference
Prediction  0  1
         0 30  1
         1  4 65

               Accuracy : 0.95
                 95% CI : (0.8872, 0.9836)

            Sensitivity : 0.8824
            Specificity : 0.9848
```

Figure: Confusion Matrix and related metrics of T-Distribution(df=43) Regression Model at 0.20 cutoff

The MCC obtained here is maximum(0.8882312) here among all the cutoffs. So, this is the best model for t-distribution regression(according to MCC metric)

Analysis and Inference

Double Exponential Regression Model:



```
          Reference
Prediction  0  1
         0 30  1
         1  4 65

               Accuracy : 0.95
                 95% CI : (0.8872, 0.9836)

            Sensitivity : 0.8824
            Specificity : 0.9848
```

Figure: Confusion Matrix and related metrics of Double Exponential Regression Model at 0.17 cutoff

The MCC obtained here is maximum(0.8882312) here among all the cutoffs. So, this is the best model for double exponential regression(according to MCC metric)

Analysis and Inference

As four of them are giving same MCC, so we will be taking the one which is less unbiased in terms of the second important error,i.e., the maximum cutoff, here 0.20. Here are such two, one is T-Distribution(df=43) Regression and another is Probit regression. So, now let's compare these two by random testing datasets.

So, what we did is the following: We did a random sampling from our dataset only, i.e. for i = 100 to 469, we randomly selected i many rows of the dataset and called that to be our testing data, and tested both the models and for each i, we got a cutoff which gives the maximum MCC. So, we can think it in this way that we did a random sampling from the distribution of the optimum cutoffs. So, we have 370 points from the distribution of optimum cutoff for both the models.

Analysis and Inference

Similarly, we now have enough datapoints fro the distribution of the best mcc for both the models. So, we now do the following hypothesis testing: We assume that the random variable 'best Mcc for probit model(X)' follows $N(\mu_1, \sigma^2)$ and that for t-distrbution(df=43)(Y) follows $N(\mu_2, \sigma^2)$

$$H_0 : \mu_2 = \mu_1 \quad H_1 : \mu_2 \neq \mu_1$$

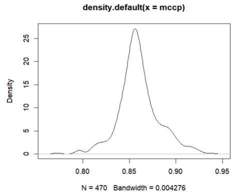


Figure: MCC distn Probit

Analysis and Inference

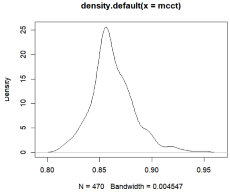


Figure: MCC distrn T

We assume σ to be unknown. So, a size α test will be $\mathbb{1}[\frac{|\bar{Y}-\bar{X}|}{S_p\sqrt{\frac{1}{n}+\frac{1}{n}}} \geq t_{2n-2,\frac{\alpha}{2}}]$ Corresponding p value will be $P(X, Y) = 2(1 - F(\frac{|\bar{Y}-\bar{X}|}{S_p\sqrt{\frac{2}{n}}}))$, F is the cdf of t_{2n-2}.

Analysis and Inference

In our case, $\bar{Y} = 0.8624161$, $\bar{X} = 0.8600604$, $S_p = 0.4642769$ and the value of corresponding test-statistic is 0.07778159. But $t_{938,0.025}$ is 1.962496. So, we accept(fail to reject) our null hypothesis at 0.05 level(size) of significance and corresponding p value is 0.9380184.

So, both model are similarly good in terms of MCC. The average cutoff for probit regression is 0.3389574, and that of t-distribution regression(df=43) is 0.3356596. So, for future prediction, we can use the cutoff as 0.33 and can use any of the two models. So, our final model is these two, because as we have shown that, on an average they are same effective.