Data Modeling for Breast Cancer Cells Research

by
Vanban L. Wu, Ph.D., PMP, PSM


Abstract:

The goal of this paper is to use a Data Modeling approach segregating malignant breast cancer cells from the benign ones. The data science technique used in this research is called Logistic Regression categorization. Due to the transparency of the operations, we are able to identify the statistical significance and multicollinearity of the variables under study and contrast the performance of the final model against a perfect model scenario.

Introduction:

Per reference described in Kaggle.com [1], cells variables are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The characteristics of the cell nuclei present the image in the 3-dimensional space [2]. In this paper, we are using this background as a testbed for the data science application.

In the data science domain, many data modeling techniques can be used to categorize or predict outcomes of an event, and Logistic Regression is one of such a technique. In this process, we utilize heavily on statistical evaluations to eliminate insignificant variables and identify multicollinearity among them for a higher performance and robust model.

Environment:

The data was downloaded from Kaggle.com on 10/02/2018 [1], an original source from the University of Wisconsin Hospitals, Madison. The size of the CSV file contains a sample set of 569 specimens with 32 variables per entry.

To support the data processing, model creation and performance evaluation, two commonly used tools are being employed: Excel and GNU Regression, Econometrics and Time-series Library (Gretl).

Data Processing:

Per original reference descriptions, the file contains the following variables:

1. ID number
2. Diagnosis (M = malignant, B = benign)

Ten real-valued variables are computed for each cell nucleus:

1. Radius (mean of distances from the center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter$^2$ / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these variables was computed for each image, resulting to 30 independent variables. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All variable values are re-coded with four significant digits. Class distribution divides into 357 benign, and 212 malignant cases.

In the processing phase, we perform the following data conversion and integrity checks:

- Select and convert fields into formats needed for the subsequent phase of analysis
- Identify and wrangle missing values, for better representation of the variables

To fit into the Logistic Regression model, the “Diagnosis” variable is being converted into two dummy variables (data science terminologies): “Malignant” variable and “Benign” variable using Gretl tool. Both dummy variables are dependent variables to the equations (1), (2) and (3) stated in the next session.

For the remaining variables (3 – 32), integrity checks are being conducted on missing values; any zero (0) values will be replaced by the variable mean of that classification, malignant or benign.

**Limitation due to File Size:**

As a primary step in the Logistic Regression modeling, the algorithm tries to fit all variables into a Sigmoid function by optimizing all coefficients, $a_0, a_1, ..., a_n$, where $p$ is the probability of occurrence (predicted outcome), and $X's$ are the independent variables under study:

$$p = \frac{1}{1+e^{-y}} \text{ where } y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n \quad (1)$$

$$\Leftrightarrow \ln\left(\frac{p}{1-p}\right) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n \quad (2)$$

$$\Leftrightarrow \frac{p}{1-p} = e^{a_0} \cdot e^{a_1X_1} \cdot e^{a_2X_2} \cdots \cdot e^{a_nX_n} \quad (3)$$

Due to the size of data source provided, which only contains 569 records, the algorithm used in the Logistic Regression runs into the “Perfect Prediction” paradox [3], with an 80:20 ratio distribution; 80% training set, 20% testing test. To overcome this shortfall, we have to allocate 99% of the records for the algorithm to fit all independent variables into the formula.
This drawback makes the 1% for the testing set meaningless as we may over-fit the model with the size of the training set. Thus, we need to adopt another method to evaluate the model performance to be discussed in the last phase of the process.

**Logistic Regression Processes:**

The regression addressed in this paper includes the following phases:

- Building the model by backward elimination process from the regressors list
- Transforming variables to optimize the predictive power
- Verifying the multicollinearity for robustness.
- Assessing the final model by the Cumulative Accuracy Profile (CAP).

**Building the Logistic Regression Models:**

As an initial step, a guideline Significant Level (SL) is selected for the process. To follow the norm, we choose SL = 0.05, i.e. any variable of highest calculated p-values above 0.05 will be treated as insignificant, thus a candidate to be removed in subsequent iterations. Initially, all thirty independent variables are included in the computation (see Figure 1). The “Malignant” dummy variable is assigned as the binary dependent variable.

In the first iteration, the calculated result shows “Perimeter Worst” having the highest p-value, 0.9541 > 0.05 (see Figure 2), which reflects its insignificance to the topic, thus the first candidate to be excluded from the regressors’ list.

In the second iteration, the Adjusted R-squared improves from 0.873871 (1st iteration) to 0.876560 (2nd iteration) proving the action of eliminating the “Perimeter Worst” variable a valid one. The result operation also shows “Smoothness Se” (0.9511 > 0.05) to be the next candidate for removal (see Figure 3).
Table 1 below summarizes the elimination process with all pertinent information generated per iteration:

<table>
<thead>
<tr>
<th>Iteration No.</th>
<th>Variable to be removed next</th>
<th>p-value of Variable to be removed</th>
<th>Adjusted R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Perimeter Worst</td>
<td>0.9541</td>
<td>0.873871</td>
</tr>
<tr>
<td>2</td>
<td>Smoothness Se</td>
<td>0.9511</td>
<td>0.876560</td>
</tr>
<tr>
<td>3</td>
<td>Fractal Dimension Mean</td>
<td>0.9029</td>
<td>0.879248</td>
</tr>
<tr>
<td>4</td>
<td>Texture Mean</td>
<td>0.8271</td>
<td>0.881921</td>
</tr>
<tr>
<td>5</td>
<td>Radius Se</td>
<td>0.7123</td>
<td>0.884550</td>
</tr>
<tr>
<td>6</td>
<td>Concavity Mean</td>
<td>0.5146</td>
<td>0.887052</td>
</tr>
<tr>
<td>7</td>
<td>Compactness Worst</td>
<td>0.2363</td>
<td>0.889147</td>
</tr>
<tr>
<td>8</td>
<td>Smoothness Worst</td>
<td>0.4781</td>
<td>0.889662</td>
</tr>
<tr>
<td>9</td>
<td>Symmetry Mean</td>
<td>0.4273</td>
<td>0.891616</td>
</tr>
<tr>
<td>10</td>
<td>Area Worst</td>
<td>0.2549</td>
<td>0.893398</td>
</tr>
<tr>
<td>11</td>
<td>Area Mean</td>
<td>0.2134</td>
<td>0.894413</td>
</tr>
<tr>
<td>12</td>
<td>Symmetry Se</td>
<td>0.2125</td>
<td>0.895858</td>
</tr>
<tr>
<td>13</td>
<td>Symmetry Worst</td>
<td>0.2322</td>
<td>0.896146</td>
</tr>
<tr>
<td>14</td>
<td>Perimeter Mean</td>
<td>0.1208</td>
<td>0.896509</td>
</tr>
<tr>
<td>15</td>
<td>Perimeter Se</td>
<td>0.2326</td>
<td>0.895547</td>
</tr>
<tr>
<td>16</td>
<td>Compactness Se</td>
<td>0.2559</td>
<td>0.896202</td>
</tr>
<tr>
<td>17</td>
<td>Radius Mean</td>
<td>0.0875</td>
<td>0.897057</td>
</tr>
<tr>
<td>18</td>
<td>Concave Points Mean</td>
<td>0.1173</td>
<td>0.895191</td>
</tr>
<tr>
<td>19</td>
<td>Concave Points Worst</td>
<td>0.4086</td>
<td>0.894148</td>
</tr>
<tr>
<td>20</td>
<td>Convexity Se</td>
<td>0.1026</td>
<td>0.895894</td>
</tr>
<tr>
<td>21</td>
<td>N/A</td>
<td>N/A</td>
<td>0.894526</td>
</tr>
</tbody>
</table>

By the 17th iteration (see Figure 4), the model has reached its stability with most p-values below 0.05, except three at the borderline. For the sake of discussion, we continue the elimination process; the three subsequent models show deterioration in R-squared values, thus an unwarranted effort. As a consequence, the remaining fourteen variables, from Radius Mean to Fractal Dimension Worst (see Figure 4), constitute a significant list to the Malignancy study and will be used in the optimization phase that follows.
Transforming Independent Variables:

In this phase, there are ten common transformations described in past literatures (B. Lund [4] and Royston, Sauerbrei [5]):

- 8 monotonic: The “fractional polynomials” \( X^i \) where \( i \) belongs to \( S = \{-2, -1, -0.5, 0, 1, 0.5, 2, 3\} \), where \( X^0 \) explicitly denotes log(X) – not a value 1 in math notation, \( X^1 \) is \( X \) itself (i.e. no transformation needed), \( X^{-0.5} \) and \( X^{0.5} \) represent \( 1/\sqrt{X} \) and \( \sqrt{X} \) respectively.
- 3 quadratic: \( (X – \text{median})^2 \), \( (X – p_{25})^2 \), and \( (X – p_{75})^2 \), where median, \( p_{25} \), \( p_{75} \) are the 50th, 25th, and 75th percentiles for \( X \) respectively.

In this paper, we won’t perform a brute-force transformation on all 14 variables, as it takes 14^{10} iterations which is impossible to accomplish. Thus, we decide the following selections to illustrate the transformation process:

- Apply \( \sqrt{X} \) on Radius Mean
- Apply \( \log(X) \) on Area Se
- Apply \( X^2 \) on Concavity Se

To apply \( \sqrt{X} \) on Radius Mean, we replace Radius Mean from the regressors list with the newly defined variable Square_Root_RadiusMean. The generated model 22 shows an improved Adjusted R-Squared and better p-values result (see Figure 5).

Next applying \( \log(X) \) on Area Se using the same process, the generated model 23 shows no improvement with a slight decrement in the Adjusted R-Squared (see Figure 6). Thus, we decide to reverse the replacement and move on to the next transformation.

Applying \( X^2 \) on Concavity Se generates a better model 24 than model 22 (see Figure 7), thus a preferred transformation for this operation.
Checking for multicollinearity:

By calculating the collinearity, we first check the Variance Inflation Factors (VIF) of model 24. The four variables of high collinearity are: Radius Worst (VIF = 29.149), Square Root Radius Mean (VIF = 21.128), Concave Point Worst (VIF = 18.981), and Concave Points Mean (VIF = 16.968) (see Figure 8).
From the correlation matrix of model 24 (see Figure 9.1 & 9.2), we decide to build two separated models evolved from the model 24 for best selection: Model 25.1 by removing the variable Radius Worst from the aggressors list, and model 25.2 by removing the variable Squared Root Radius Mean (see Figure 10.1 & 10.2).

**Figure 9.1**
Out of the two options, we decide to adopt model 25.2 over 25.1 due to higher Adjusted R-squared.
From the model 25.2, we eliminate variable Concave Points Worst, the resulting model 26 is depicted in Figure 11.

Table 2 below depicts the iteration process to eliminate variables of high p-values. The iteration stops at model 27 due to Adjusted R-squared deteriorates in model 28:

Table 2

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Variable to be removed next</th>
<th>p-value of Variable to be removed</th>
<th>Adjusted R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Square Root Radius Mean</td>
<td>0.0745</td>
<td>0.897508</td>
</tr>
<tr>
<td>23</td>
<td>Concavity Se</td>
<td>0.0610</td>
<td>0.892968</td>
</tr>
<tr>
<td>24</td>
<td>N/A</td>
<td>N/A</td>
<td>0.899584</td>
</tr>
<tr>
<td>25.1</td>
<td>Concave Points Mean</td>
<td>0.1227</td>
<td>0.887506</td>
</tr>
<tr>
<td>25.2</td>
<td>Concave Points Worst</td>
<td>0.1179</td>
<td>0.896185</td>
</tr>
<tr>
<td>26</td>
<td>Concave Points Mean</td>
<td>0.3417</td>
<td>0.895038</td>
</tr>
<tr>
<td>27</td>
<td>Square Concavity Se</td>
<td>0.0980</td>
<td>0.896758</td>
</tr>
<tr>
<td>28</td>
<td>Texture Se</td>
<td>0.0542</td>
<td>0.894526</td>
</tr>
</tbody>
</table>

The following charts (see Figure 12) depict the p-values of all remaining independent variables, their VIF calculations, the corresponding correlation matrix and heat map of the model 27, no flaw indications in the final model.
Assessing the Model:

The Cumulative Accuracy Profile (CAP) curve is a method used in the data science domain to assess the performance of a model being built, a standard to contrast between models, and detect deterioration of a model over time [6]. In this illustration, the Model CAP
closely converges to the Perfect CAP curve performance (see Figure 13); at 38% Percentile, the Perfect CAP curve reaches 100%, while the Model CAP curve is at 98.5% mark. At 48.5% the Model CAP curve reaches 100%.

**Figure 13**

**A Side Note on Coefficients of Logistic Regression:**

Based on the equation (3) stated at the beginning of the paper, any increase of an independent variable $x_i$ by 1 unit will trigger the malignancy odds ($\frac{P}{1-P}$) by a multiplicative factor of $e^{a_i}$. By this mathematical correlation, any independent variable of large coefficient, will definitely influence the outcome of malignancy; if the sign of a coefficient is positive, the variable will impact positively to the malignancy outcome, otherwise the benign nature of the study. This observation still needs to be consulted with medical pathology research, and AMA literatures as correlation does not always imply causation.

**Concluding Remark:**

In conclusion, we successfully illustrate the major Logistic Regression methodology in the analysis of breast cancer cells on their malignancy characteristics. In this process, we identify the significant variables which are worth noticing in cancer medical analysis; we start with thirty variables and scale the list down to just eleven variables of statistical significance and low correlation. This result in conjunction with the coefficients are worth to be cross-checked with pathologist for medical interpretation and confirmation.
References


