STA 223 Final Project

Xiaoliu Wu

March 2020

1 Data Description

The breast cancer data [1] was collected by researchers at University of Coimbra and Hospitais da Universidade de Coimbra and is available on the UCI Machine Learning Repository. The data set contains 116 patients' information including age (years), 8 medical measurements which can be easily obtained from a routine bloody analysis[1] and an indicator variable of the presence of breast. 8 medical measurements are glucose (mg/dL), insulin ($\mu U/mL$), HOMA, leptin (ng/mL), adiponectin ($\mu g/mL$), resistin (ng/mL), MCP-1 (pg/dL), and Body Mass Index (kg/m^2). Note that

$$\text{HOMA} = \frac{\text{Insulin} * \text{Glucose}}{405},$$

therefore, we should consider it as the interaction term rather than an independent predictor. 64 patients in the study suffering breast cancer and 52 patients who do not. Thus, there is no severe class imbalance problem.

2 Objectives

We strike to build a logistic regression model that fits the data well and construct confidence intervals for slopes. Next, we construct a predictive model using regularized logistic regression models.

3 Methods

Since all 9 predictors are continuous, we first use the generalized additive model (GAM) to discover potential non-linear relationships between predictors and the response. Namely, we fit the following model

$$logit(P(Y_i = 1|X = x_i)) = \beta_0 + \sum_{j=1}^{9} g_j(x_{ij}).$$

Next, we fit a logistic regression model

$$Y_i|X_i \sim \text{Bernoulli}(P_i)$$
$$logit(P_i) = \eta = X_i^T\beta$$

where X_i^T is a column vector of predictors. We will construct confidence intervals for slopes after outlier detection, model selection and model diagnostic.

4 Inference

4.1 Exploratory Analysis & Outlier Detection

GAM (Figure 4 in the Appendix) indicates clear nonlinear relationship between age, BMI, resistin, MCP-1, leptin, Adiponectin and the $logit(P(Y_i = 1|X = x_i))$. We fit a full logistic regression model with all 9 predictors together with quadratic terms of age, BMI, and resistin as well as cubic terms for MCP-1, leptin, and Adiponectin.

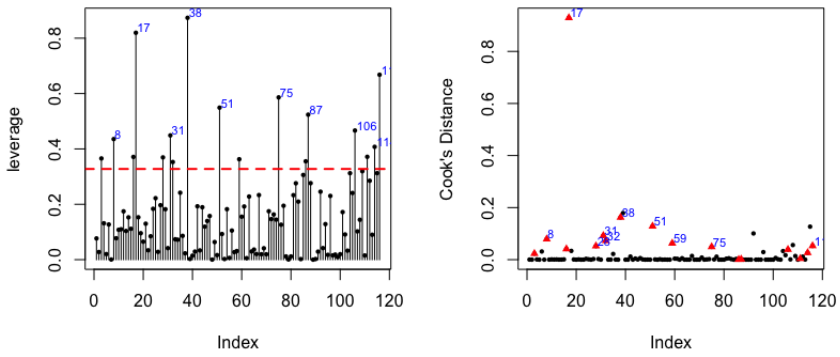

Figure 1: Plots of Leverage Points and Cook's Distance of the Initial Logistic Model

Based on Figure 1, observations 17, 38 and 51 seem to have large Cook's distance and happen to be high leverage points. We visualize these three dots using the first two principal components. The first two principal components account for about 50% variability of the data.

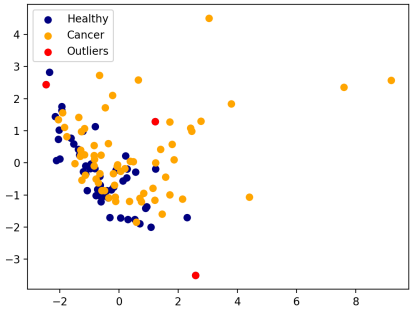

Figure 2: Plots of the First Two Principal Components.

All three patients don't suffer breast cancer, and they seem to be far away from the cloud of other healthy patients in Figure 2. Therefore, we deem them as outliers and delete them from the analysis.

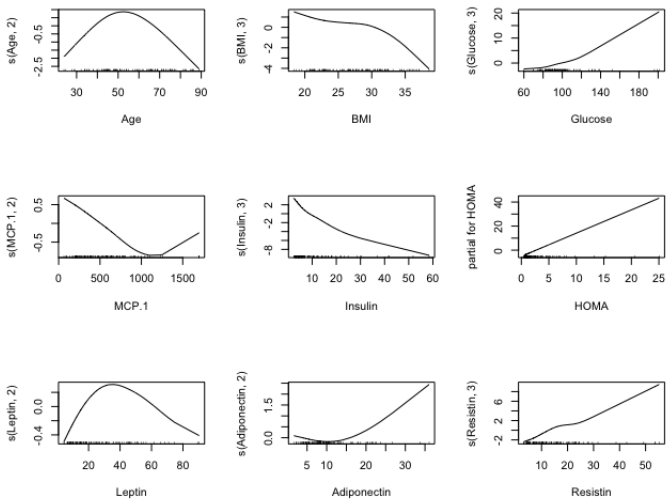

Figure 3: Plots of Smoothing Functions from the GAM Model after Deleting Outliers.

We refit the GAM and Figure 3 suggests that we may drop all the cubic terms and the quadratic term for resistin from the full logistic model.

4.2 Model Selection

We apply stepwise AIC and BIC for model selection. AIC arrives at the model

$$\eta = \beta_0 + \beta_1 * \text{age} + \beta_2 * \text{age}^2 + \beta_3 * \text{BMI} + \beta_4 * \text{BMI}^2 + \beta_5 * \text{Glucose} + \beta_6 * \text{HOMA} + \beta_7 * \text{Resistin} + \beta_7 * \text{Adiponectin}^2,$$

whereas, BIC arrives at the model

$$\eta = \beta_0 + \beta_1 * \text{age} + \beta_2 * \text{age}^2 + \beta_3 * \text{BMI}^2 + \beta_4 * \text{Glucose} + \beta_5 * \text{Resistin}.$$

We chose the BIC model for three reasons. Because of the consistency of the BIC criteria, BIC is usually used for inference purposes. Since BIC selected a more parsimonious model than AIC did, we will have a smaller confidence region. Finally, we used cross-validation to check if either model is a reasonable fit for unseen data. The results demonstrate that both models have similar generalizability based on mean CV-errors. Details of CV-errors can be found in Section 5.

The residual plots (Figure 5 in the Appendix) confirm the goodness of fit of the BIC model and the leverage and Cook's distance plot (Figure 6 in the Appendix) shows no obvious outliers. Therefore, we used the BIC model for inference.

4.3 Confidence Intervals for Slopes

We report 99.8% confidence intervals produced by likelihood method and non-parametric bootstrap. The confidence level 99.8% is a result of Bonferroni correction, namely $1 - \frac{0.01}{5} =$

	Point Estimate	Lower Bound	Upper Bound	Width
Intercept	-23.2066	-46.5858	-8.5596	38.0261
Age	0.5194	0.1112	1.1031	0.9919
Age^2	-0.0048	-0.0098	-0.0012	0.0086
BMI^2	-0.0039	-0.0079	-0.0007	0.0072
Glucose	0.1224	0.0498	0.2260	0.1762
Resistin	0.1735	0.0388	0.3648	0.3259

Table 1: 99.8% Confidence Intervals of Slopes based on the Likelihood.

	Point Estimate	Lower Bound	Upper Bound	Width
Intercept	-23.2066	-76.3744	-12.0062	64.3682
Age	0.5194	0.1776	1.7701	1.5925
Age^2	-0.0048	-0.0160	-0.0018	0.0141
BMI^2	-0.0039	-0.0117	-0.0007	0.0111
Glucose	0.1224	0.0536	0.3957	0.3421
Resistin	0.1735	0.0364	0.6821	0.6457

Table 2: 99.8% Confidence Intervals of Slopes based on Bootstrap Samples.

0.998. We use the method introduced in Example 2 on page 8 of the lecture notes to construct likelihood confidence intervals. Bootstrap confidence intervals are based on 10000 bootstrap samples and use the 0.001th and 0.999th percentile of the empirical distribution of slope estimates. The confidence intervals are conservative compared to the confidence region constructed by the ellipsoid, but they are easier to report. If the researcher is interested in the exact 99% confidence regions of all slopes, he or she may use the formula given on page 10 of the lecture notes to see if a set of slopes is in the confidence region or not. The relevant distribution is the Chi-squared distribution with 6 degrees of freedom.

From Table 1 and 2, we see that all slopes are significant regardless of the methods of constructing confidence intervals. The bootstrap confidence intervals are almost twice wider than those constructed by the likelihood approach. There are significant portions of likelihood CIs overlapping the bootstrap CIs. We favor the bootstrap confidence intervals because we want to be conservative with inference after performing outlier detection and model selections.

4.4 Interpretation

We can interpret the slope for BMI^2, glucose and resistin in terms of the odds ratio. For example, when a patient's glucose level increases 1 mg/dL, the odds ratio of breast cancer is $\exp 0.1224 = 1.13$. In other words, the risk of breast cancer increases when the patient's glucose increases. Since the slope of resistin is positive, it is also positively correlated with the risk of breast cancer. So far the results have been consistent with previous studies [2] [3]. The negative slope of the quadratic term of BMI is counter-intuitive because obesity is usually a contributing factor for diseases. However, there is a study [4] also observed the negative correlation between BMI and the risk of breast cancer. Further investigation of the

correlation between BMI and breast cancer is required.

The interpretation of coefficients of age is tricky. Both linear and quadratic terms of the it are present, and their slopes have different sign. We can rewrite $\beta_1 \text{age} + \beta_2 \text{age}^2$ as

$$\hat{\beta}_2(\text{age} + \frac{\hat{\beta}_1}{2\hat{\beta}_2})^2 + c.$$

Since $\hat{\beta}_2$ is negative, assuming other predictors being constant, the odds of getting breast cancer first increases with age and is peaked at age of $-\frac{\hat{\beta}_1}{2\hat{\beta}_2} = 54.1042 \approx 54$. After 54, the risk of breast cancer declines with age. The 99% bootstrap confidence interval for $-\frac{\hat{\beta}_1}{2*\hat{\beta}_2}$ is $(49.0145, 58.5212)$ and the likelihood confidence interval is $(48.4378, 60.2280)$. The construction of the likelihood confidence interval uses the delta method and confidence ellipsoid. Details are discussed in the appendix. The result may not apply to all population as the study didn't control factors such as age. Future study should control more factors in order to study the age where the breast cancer risk peaks.

5 Prediction

	Mean Training Error	Mean CV Error	Test Error	Precision	Recall
Baseline	0.2195	0.2778	0.2174	0.7647	0.9286
Bseline $L2$	0.2223	0.2667	0.1739	0.8125	0.9286
AIC	0.1077	0.1444	0.2174	0.7647	0.9286
AIC $L2$	0.1101	**0.1333**	0.1739	0.7778	**1**
BIC	0.1477	0.1889	0.2174	0.8	0.8571
BIC $L2$	0.1256	0.1556	**0.1304**	**0.8235**	**1**
Full Model $L1$	0.0858	0.1556	0.2174	0.7647	0.9286
Full Model $L2$	0.1034	0.1667	0.1739	0.7778	**1**

Table 3: 99.8% Predictive Performance of Various Models. Error the misclassification rate.

We consider 4 logistic regression models and their regularized version. The baseline model only contains all the predictors with first order terms only. The full model contains all predictors and higher order terms suggested by GAM. The AIC and BIC models are introduced in section 4.2. We randomly divide the data set without observation 17,38 and 51 into a training set and a test set. The training set contains 40 healthy cases and 50 cancer cases and the numbers are 9 and 14 for the test set. We use the leave-one-out cross-validation for the hyperparameter tuning.

The BIC model with $L2$ regularization has the best out-of-sample predictive ability in terms of misclassification rate, precision, and recall. The regularized model outperforms the unregularized logistic regression for all three cases. The regularization also enables full model to have comparable predictive performance despite the small sample size. This study once again shows that regularization is crucial for prediction.

The cross validation error is a poor predictor for the test error for this data set. The small sample size of the data set may be the reason of the inconsistency between two errors.

6 Conclusions

We conclude that higher glucose, as well as resistin level, are associated with a higher risk of breast cancer. On the other hand, squared BMI has a negative association with breast cancer. The chance of getting breast cancer initially increases with age; the risk hit the maximum in one's 50s then decreases. For future studies, researchers should control factors such as races to better understand the association between breast cancer and measurements present in this study. Ideally, the study can fit the model based on the BIC model used in this study without a model selection procedure so that the inference can indeed reach the nominal level.

The logistic regression model combined with thoughtful exploratory analysis (GAM) and regularization achieves a reasonably well performance. The experiment shows that the linear method can still be useful and regularization is important for prediction. Future study may compare the logistic regression model with other predictive models which have non-linear decision boundaries. Ideally, we would like to have more data to compare the performance of different models.

References

[1] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 18, 29 (2018). https://doi.org/10.1186/s12885-017-3877-1.

[2] Hou Y., Zhou M., Xie J., Chao P., Feng Q., Wu J. High glucose levels promote the proliferation of breast cancer cells through GTPases. Breast Cancer. 2017;9:429–436. doi: 10.2147/BCTT.S135665.

[3] Zeidan, B., Manousopoulou, A., Garay-Baquero, D.J. et al. Increased circulating resistin levels in early-onset breast cancer patients of normal body mass index correlate with lymph node negative involvement and longer disease free survival: a multi-center POSH cohort serum proteomics study. Breast Cancer Res 20, 19 (2018). https://doi.org/10.1186/s13058-018-0938-6.

[4] David Bai, PharmD. Higher BMI Decreases Risk of Breast Cancer in Premenopausal Women. (2018) AJMC.

7 Appendix

7.1 Code

The code to produce the analysis is publicly available on the GitHub. Here is the link of the repository `https://github.com/charles19920528/Predict-Breast-Cancer-.git`.

7.2 Graphs Omitted from the Main Report

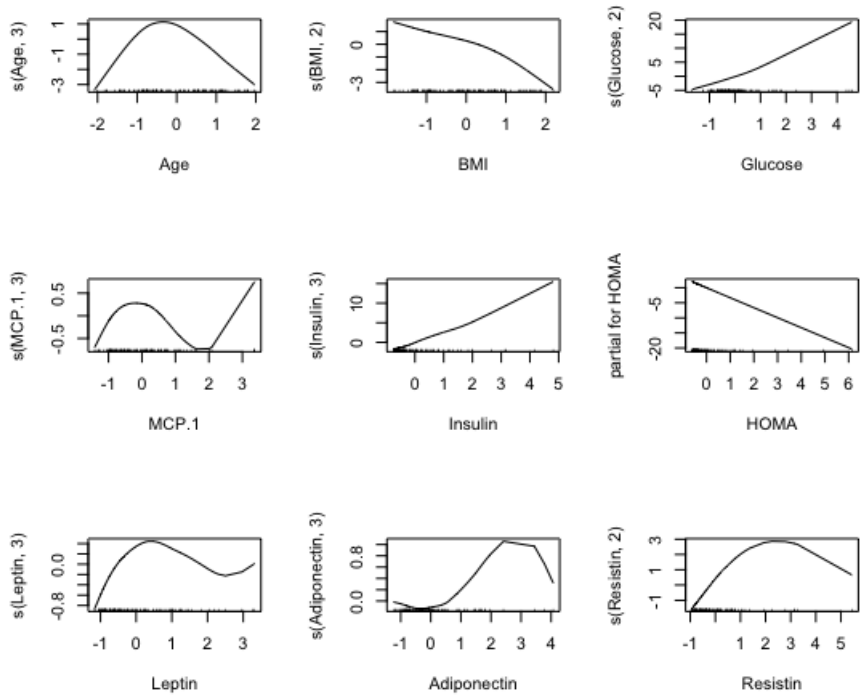

Figure 4: Plots of Smoothing Functions from the Inital GAM Model

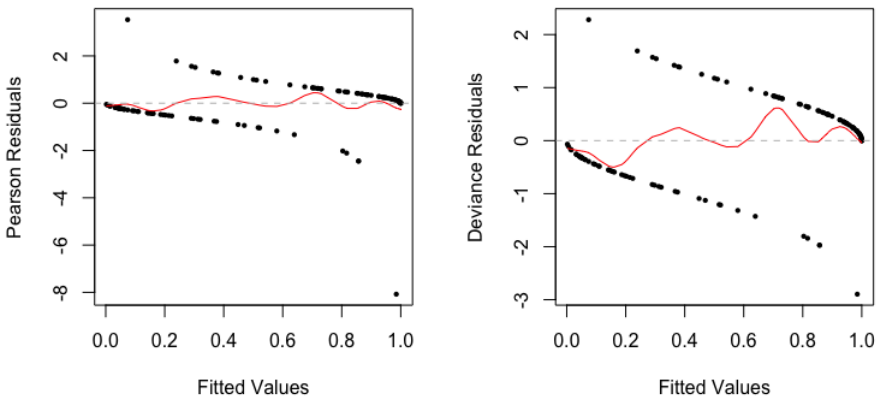

Figure 5: Residual Plots of the BIC Model

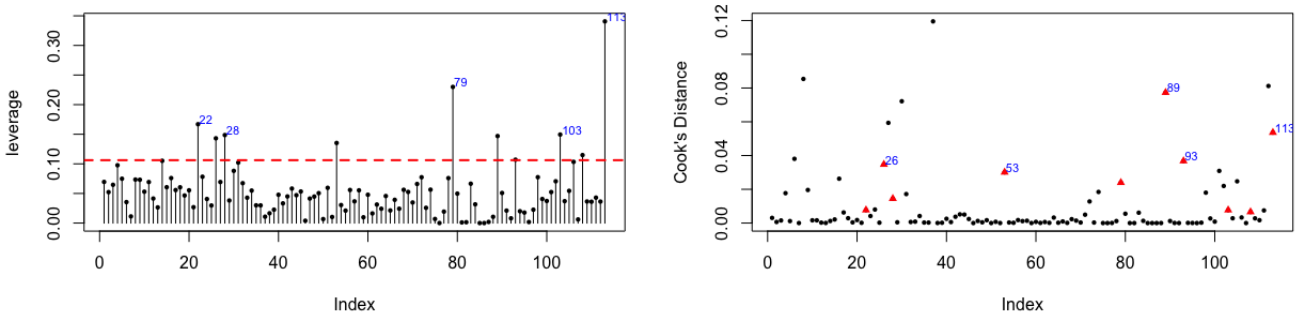

Figure 6: Plots of Leverage Points and Cook's Distance of the BIC Model

7.3 Delta Method

In Section 4.4, we want to construct a confidence interval for $-\frac{\beta_1}{2\beta_2}$. Let $\beta = [\beta_1, \beta_2]^T$ and $h(\beta) = -\frac{\beta_1}{2\beta_2}$. Then the gradient $\nabla h(\beta) = [-\frac{1}{2\beta_2}, \frac{\beta_1}{2\beta_2^2}]$. From lecture notes, we know that

$$(\hat{\beta} - \beta) \sim N(0, I^{-1}(\hat{\beta}))$$

. By the delta method,

$$(h(\hat{\beta}) - h(\beta))^T \sim N(0, \nabla h(\hat{\beta})^T I^{-1}(\hat{\beta}) \nabla h(\hat{\beta}))$$

Since both $h(\hat{\beta})$ and $\nabla h(\hat{\beta})^T I^{-1}(\hat{\beta}) \nabla h(\hat{\beta})$ is a scalar, the 99% confidence interval of () is given by

$$(h(\hat{\beta}) + Z_{0.005} \nabla h(\hat{\beta})^T I^{-1}(\hat{\beta}) \nabla h(\hat{\beta}),\ h(\hat{\beta}) - Z_{0.005} \nabla h(\hat{\beta})^T I^{-1}(\hat{\beta}) \nabla h(\hat{\beta}))$$

where $Z_{0.005}$ is the 0.005th percentile of a standard normal distribution.