Breast Cancer Inference and Prediction using Logistic Regression

Xiaoliu Wu, University of California, Davis, Department of Statistics

Dataset

- The data set contains 116 patients. 64 patients suffers breast cancer and 52 are healthy.
- 9 Continuous measurements from blood tests:
  - age, BMI, glucose (mg/dL), insulin (μU/mL),HOMA, leptin (ng/mL), adiponectin (μg/mL), resistin (ng/mL), MCP-1 (pg/dL), and BMI (kg/m2).

Objectives

- Build a logistic regression model that fits the data well and construct confidence intervals for slopes.
- Construct a predictive model using regularized logistic regression models.

Methods

- For inference, we fit a logistic regression model:

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

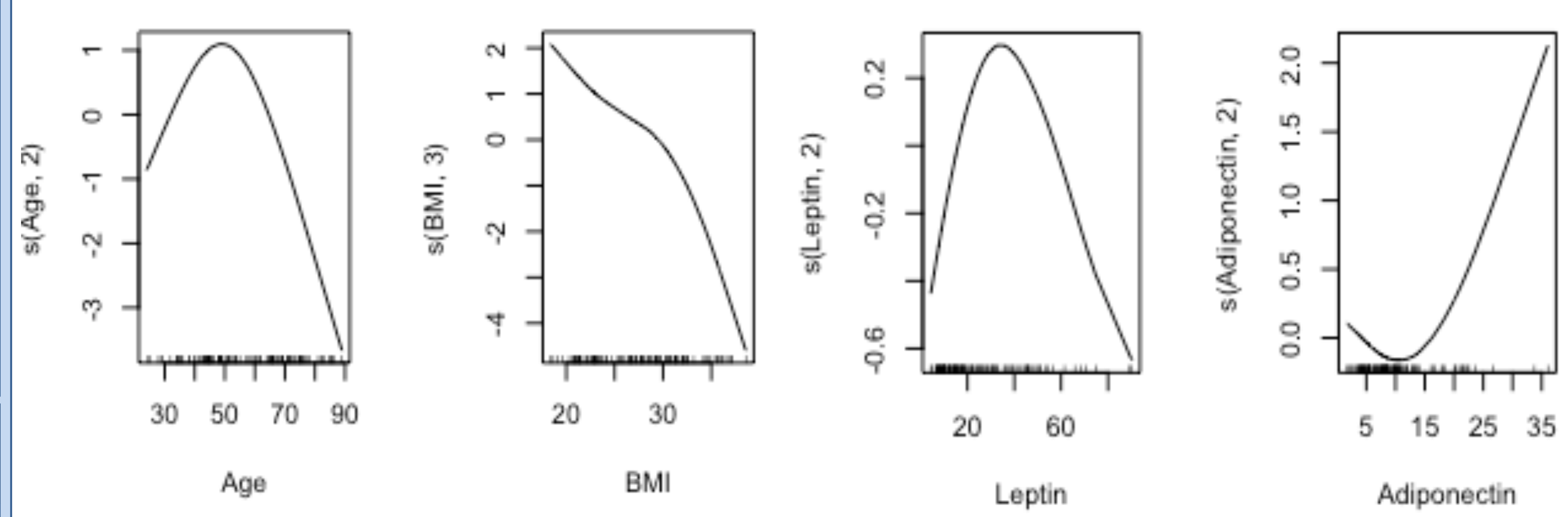
- Use GAM to discover non-linear relationship between the response and predictors.

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

- Delete outliers based on the “full” logistic model.
- Select the inference model by considering BIC, AIC and cross validation.
- Construct the confidence intervals using likelihood and bootstrap.
- For prediction, we consider logistic regression model with different predictors and regularization terms and use leave-one-out cross validation for hyperparameter tuning.

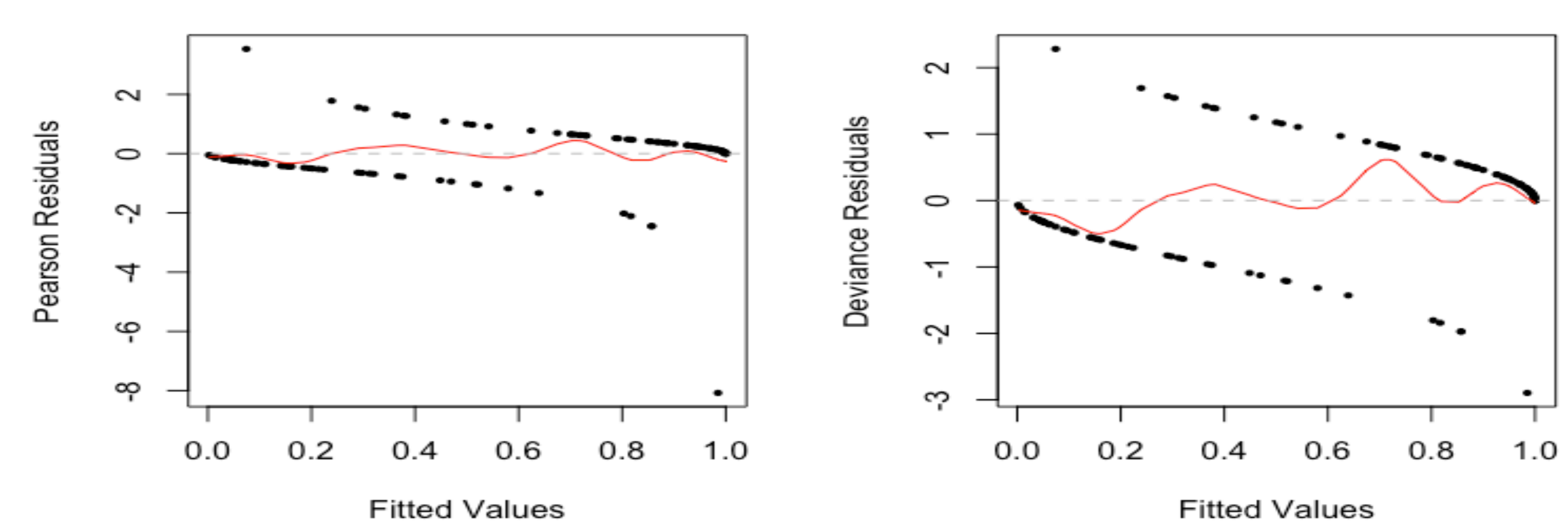
Inference Results

- After deleting outliers, GAM still found clear non-linear relationship between the response and predictors.



- We use the model selected by BIC criteria for inference. The residual plot shows no sign of lack of fit.

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



- 99% bootstrap confidence intervals are reported here, as they are more conservative. Bonferroni correction is used.

	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

- The transformation of slopes $-\frac{Age}{2\ Age\text{^}2}$ tells us the age where the risk of breast cancer peaks. The point estimation for the ratio is 54 and the 99% confidence interval is (49.0145,58.5212).

Classification

	Train Error	CV Error	Test Error	Precision	Recall
Baseline	0.2195	0.2778	0.2174	0.7647	0.9286
Baseline 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***

- The baseline model uses only the linear terms.
- Regularization is crucial for prediction as it avoids overfitting.
- The small sample size may cause the inconsistency between the CV error and test error.
- The non-linear relationship found by GAM helps prediction.
- Future study may consider other predictive models with non-linear decision boundaries.

References

- 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).
- 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.
- Other reference are listed in the report.