MSCR 509: High Dimensional Analysis

Homework 6

Anish Shah

February 24, 2020

Description

Data were collected as part of a larger study at Baystate Medical Center in Springfield, MA. This data set contains information on 189 births to women seen in the obstetrics clinic. Fifty-nine of these births were low birth weight. The goal of the current study was to determine whether the variables included in the data set were risk factors for having low birth weight in the clinic population being served by the Baystate Medical Center. Actual observed variable values have been modified to protect subject confidentiality.

Variables are below. Description, and then SAS variable (short name):

- Low Birth Weight ('no' = Birth Weight >= 2500g, 'yes' = Birth Weight < 2500g) ... *LOW_BIRTH_WEIGHT (LOW)
- Age of the Mother in Years ... AGE
- Weight in Pounds at the Last Menstrual Period ... WEIGHT (LWT)
- Race ('white', 'black', 'other') ... *RACE
- Smoking Status During Pregnancy ('yes', 'no') ... *SMOKE
- History of Premature Labor ('yes', 'no') ... *PREMATURE_LABOR (PTD)
- History of Hypertension ('yes', 'no') ... *HYPERTENSION (HT)
- Presence of Uterine Irritability ('yes', 'no') ... *UTERINE_IRRITABILITY (UI)
- These variables are coded in SAS as 'character variables,' and therefore require use of the CLASS statement to model properly. Be sure to specify 'reference cell coding' and the chosen reference group. For example, CLASS RACE (param=ref ref='white').

Question 1

Fit a logistic regression model for predicting the risk of low birth weight in terms of age, race, and smoking status. Use reference cell coding, with reference groups as "not-smoking" and "white" for smoking status and race, respectively.

- Write down the predicted logistic regression model.
- Interpret the odds ratios corresponding to smoking and race.

$$log\left(\frac{P(LBW)}{1-P(LBW)}\right) = \beta_0 + \beta_1 age + \beta_2 race + \beta_3 smoke$$

Table 1: Prediction of LBW Babies

	Dependent variable:
	low_birth_weight
age	0.970 (0.907, 1.034)
raceother	2.507** (1.125, 5.772)
raceblack	4.329*** (1.717, 11.310)
smokeyes	2.637** (1.277, 5.646)
Constant	0.345 (0.060, 1.922)
Observations	189
Log Likelihood	−109.100
Akaike Inf. Crit.	228.300
Note:	*p<0.1; **p<0.05; ***p<0.01

The OR for having a LBW baby if the patient is black (compared to white) is **4.33**, with adjustment for age and smoking status. The OR for having a LBW if a smoker (versus nonsmoker) is **2.64**.

Question 2

When evaluating a predictive rule, it is recommended to use a validation set. Split the data into training (70%) and validation (30%) sets, using the seed 19850604 in SAS.

The data has been split into 70/30 training and test data.

Using your model selection procedure of choice, develop a logistic regression model using only the training set. Write down the final estimated model, and provide justification. (Use a forward/backward/stepwise selection /hybrid procedure to decide on final main effects model, use backward selection (w/ a 0.05 significance threshold) to test for potential pairwise interactions.)

Table 2: Best Model from Training Data

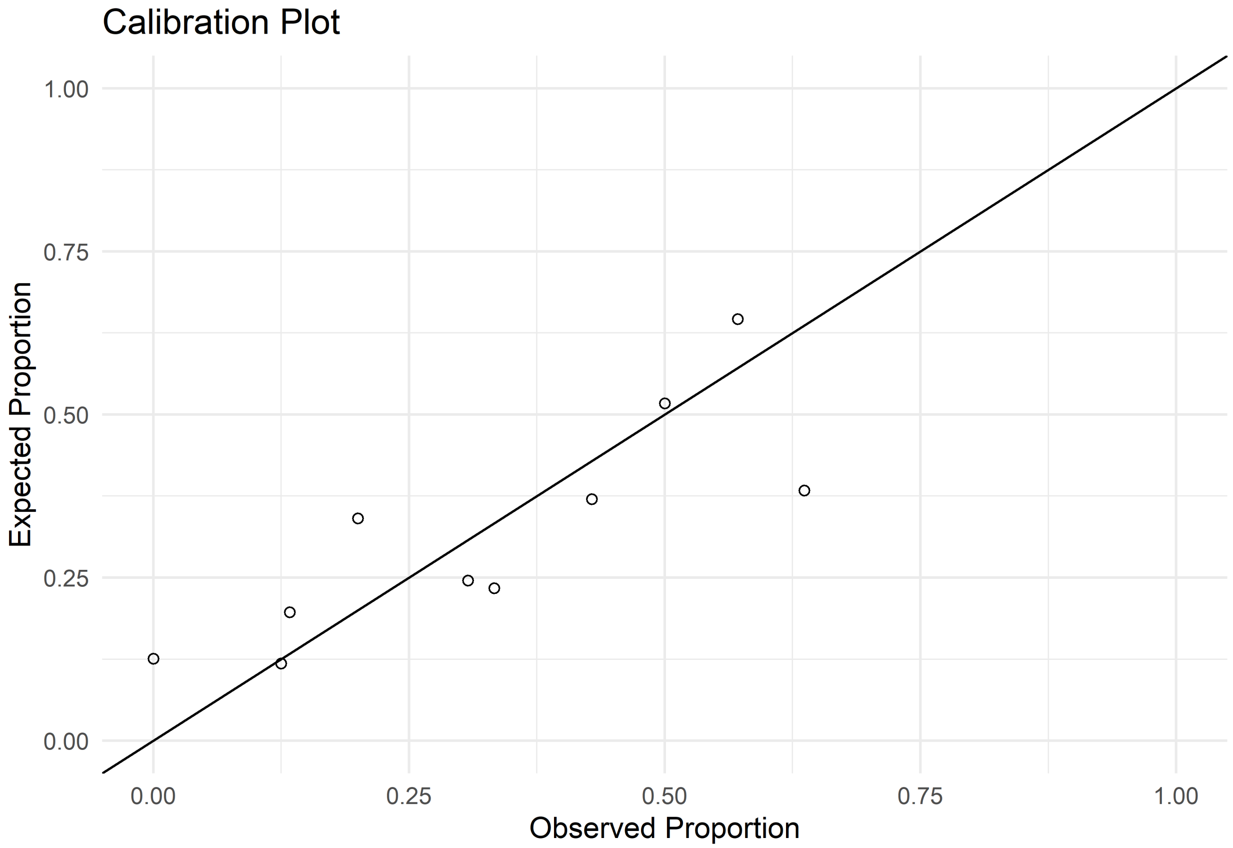
	Dependent variable:
	low_birth_weight
raceother	9.783***
	(2.440, 66.120)
raceblack	24.000***
	(4.551, 195.400)
smokeyes	6.111**
	(1.472, 41.840)
raceother:smokeyes	0.167
	(0.011, 1.845)
raceblack:smokeyes	0.136
	(0.010, 1.532)
Constant	0.067***
	(0.011, 0.221)
Observations	133
Log Likelihood	−72.640
Akaike Inf. Crit.	157.300
Note:	*p<0.1; **p<0.05; ***p<0.01

The final model, as seen above, was selected using an exhaustive model selection method, with pairwise interactions tested bidirectionally between all terms. The final model incorporates race and smoking status, and adjusts for the interaction between smoking and race.

Question 3

Perform the Hosmer-Lemeshow test on the training set, using the 'lackfit' option.

- Report the test statistic and p-value.
- Interpret the results.
- Construct a calibration plot, and interpret.



This is for the full training data, as the question does not specify to use the final model that has interaction terms.

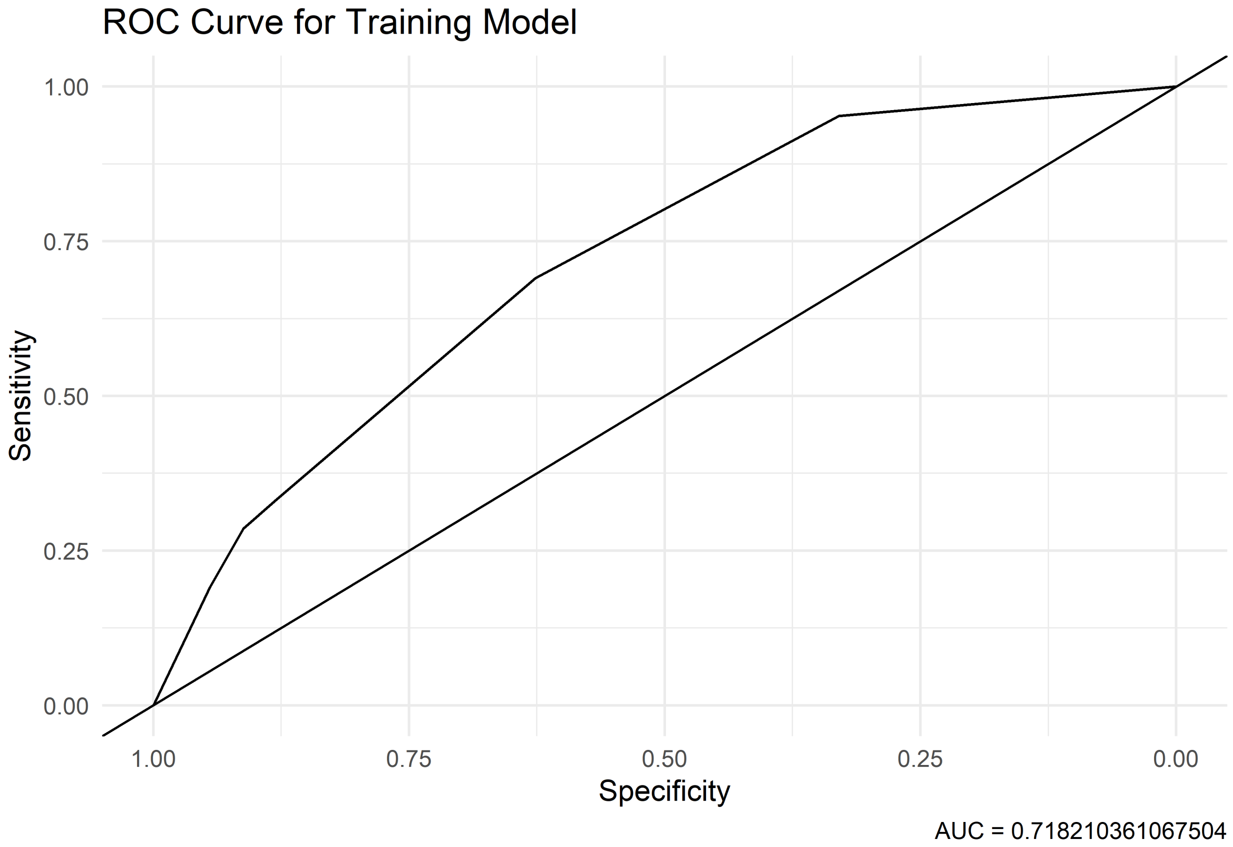
The Hosmer-Lemeshow test on the training data had a χ^2 of 7.7652 for 8 degrees of freedom, with 0.4567. This suggsts that, for an $\alpha = 0.05$, we have insufficient evidence to reject the null hypothesis, and thus can accept the model fits the data well.

The calibration plot shows that there is a linear pattern to the 10 groups from the Hosmer-Lemeshow test that roughly fit the 1:1 correlation line. This helps show that although there is a pattern, with the training set, there are potentially outliers or random noise that may need to be addressed (e.g. bootstrapping or multiple iterations).

Question 4

Derive the ROC curve and AUC for your final model, using the training data. Comment on the predictability of this model?

This ROC curve is from the final model, after evaluating 1st and 2nd level interactions of parameters. This appears to be a somewhat decent predictability of the data, wiht an $AUC > 0.7$.



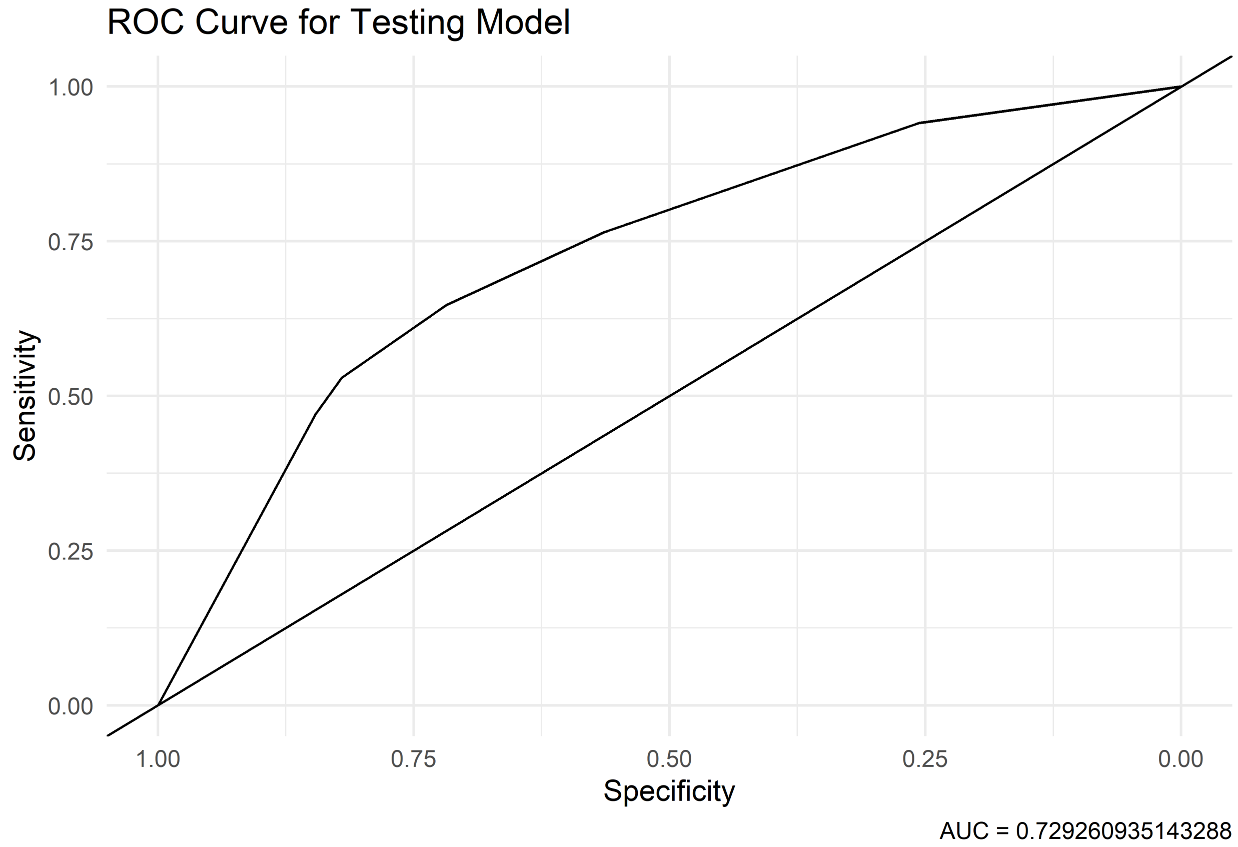
Question 5

Use the validation set to evaluate your final model.

- Report the AUC.

- Show a ROC plot for validation dataset.
- Comment on the overall performance of your final model.

Below is the predictivity of our model via ROC curve. It similarly shows an $AUC > 0.7$, suggesting reasonable predictability. It is likely that the test data is underpowered to fully evaluate this model.



Question 6

State your final estimated model and explain how you would interpret for future patients. Specifically, describe patient characteristics that are associated with the outcome.

Best Model from Training Data

```
   Dependent variable:
---------------------------
    low_birth_weight
```

raceother	2.500 (0.270, 55.050)
raceblack	3.333 (0.263, 81.370)
smokeyes	13.330** (1.805, 281.100)
raceother:smokeyes	0.150 (0.004, 2.924)
raceblack:smokeyes	0.225 (0.003, 13.420)
Constant	0.100** (0.005, 0.522)

Observations 56
Log Likelihood -30.120
Akaike Inf. Crit. 72.240

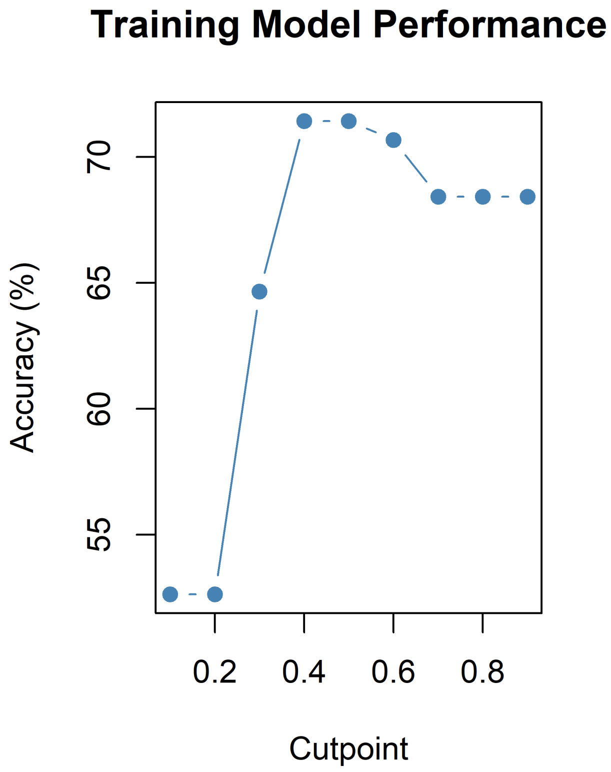
== Note: *p<0.1;* ***p<0.05;*** p<0.01

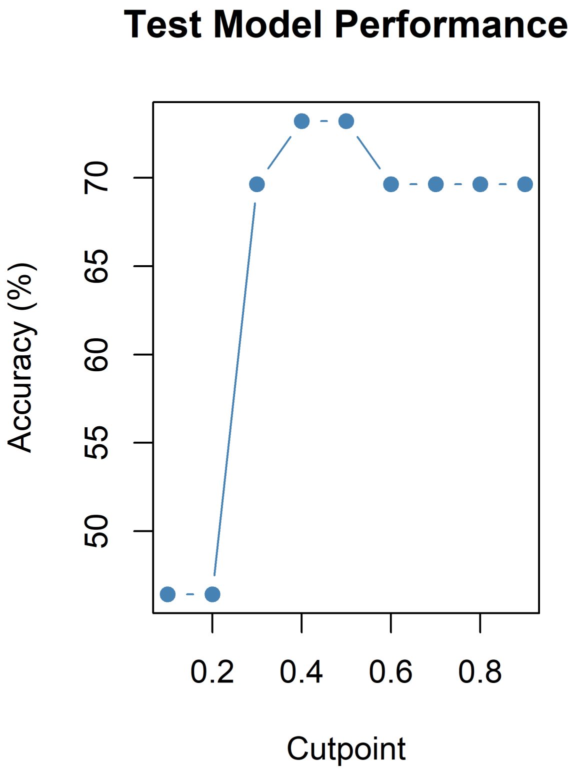
In this validation dataset, the parameter statistics have a wider confidence interval, suggesting we are underpowered. However, even with this, smoking remains a strong predictor of LBW babies. Compared to the training model, the trend of black race and smoking appear to be high risk factors, and thus would be patients of special interest or of a higher risk category.

Question 7

Choose a cut-point that you think is reasonable for deriving a classification rule from your final model.

- Specify your classification rule.
- Report the sensitivity and specificity in the training set.
- Report the sensitivity and specificity in the validation set.
- Comment on the performance of your classification rule.


Training Model Performance


Test Model Performance

Based on assessing the cutpoint at multiple levels, the value of **0.5** is reasonable to use for an accuracy of approximately 70%. The graphs here help to show my confidence in the performance of my classification rule.

For the training data set, with the specified cutpoint, the sensitivity of the model is 0.9121, and the specificity is 0.2857.

For the test/validation data set, with the specified cutpoint, the sensitivity of the model is 0.8205, and the specificity is 0.5294.

Question 8

Given that the dataset is not very large, the Principal Investigator suggests that the whole data set should be used to develop the model. Build your final model, using the entire data set.

- State your estimated model.
- Report the AUC under this model. Do you think this AUC is an accurate estimate of your model's true predictive abilities? Is it an under- or over-estimate?
- Using cross-validation, estimate the AUC.
- Discuss the results. How well does your model predict? Would you consider using this model in the clinical setting?

Table 4: Model using Entire Data Set

	Dependent variable:
	low_birth_weight
raceother	2.671^{**}
	(1.600, 13.090)
raceblack	4.587^{***}
	(2.734, 28.450)
smokeyes	2.736^{***}
	(0.906, 6.333)
Constant	0.162^{***}
	(0.011, 0.956)
Observations	189
Log Likelihood	−109.600
Akaike Inf. Crit.	227.200
Note:	*p<0.1; **p<0.05; ***p<0.01

The model stated above in table format. Using hte full dataset, the "best" model, using exhaustive model selection processes, yields the same model with the training data set.

The AUC is 0.6647 for the full data set. We see that it is lower than in the training or data set. I think this is a more reflective of the true predictive value of our model, but potentially influenced by noise or systemic bias in the data set initial creation. With 10-fold validation, the AUC is 0.6237. It appears that this full model is less powerful/predictive than we would safely like for assess patients at risk for LBW. I would not use this for clinical testing at this point, however I would still point out that the risk is likely greater in both black race and smoking.