MSCR 509: High Dimensional Analysis

Homework 5

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

Run a model with [age, race, Hx Hypertension] and report AUC.

Table 1: Logistic regression of LBW

	Dependent variable:
	low_birth_weight
age	0.963
	(0.901, 1.025)
raceother	1.642
	(0.802, 3.373)
raceblack	3.117^{**}
	(1.303, 7.535)
hypertensionyes	2.214
	(0.575, 8.506)
Constant	0.709
	(0.144, 3.523)
Observations	189
Log Likelihood	−111.900
Akaike Inf. Crit.	233.800
Note:	*p<0.1; **p<0.05; ***p<0.01

The AUROC of *Model 1* is 0.6492.

Question 2

Run a model with [age, race, Hx Hypertension, smoking during pregnancy, premature labor] and report AUC. Compare the AUC to the model derived in Question 1 and comment. Also report sensitivity, specificity, false positive rate and false negative rate for prob level = 0.500 for the better model.

Table 2: Logistic regression of LBW with Further Covariates

	Dependent variable:
	low_birth_weight
age	0.951
	(0.883, 1.020)
raceother	2.400**
	(1.032, 5.765)
raceblack	4.016***
	(1.546, 10.790)
hypertensionyes	2.565
	(0.655, 10.090)
smokeyes	2.207**
	(1.022, 4.892)
premature_laboryes	4.376***
	(1.848, 10.780)
Constant	0.425
	(0.066, 2.649)
Observations	189
Log Likelihood	−102.700
Akaike Inf. Crit.	219.500
Note:	*p<0.1; **p<0.05; ***p<0.01

The AUC of *Model 2* is 0.7265. The AUC of *Model 1* is 0.6492, suggesting that *Model 2* is more effective or a better predictor. When we set the probability level to be 0.5, we can determine other statistics, as seen below, for this better model.

	x
Sensitivity	0.9077
Specificity	0.3559
Pos Pred Value	0.7564
Neg Pred Value	0.6364
Precision	0.7564
Recall	0.9077
F1	0.8252
Prevalence	0.6878
Detection Rate	0.6243
Detection Prevalence	0.8254
Balanced Accuracy	0.6318

Question 3

Write down estimated final multivariate model you obtained in homework 4. What is predictability of this model? Plot ROC curve and find a cut-point with reasonable sensitivity and specificity and report results and justify your answer.

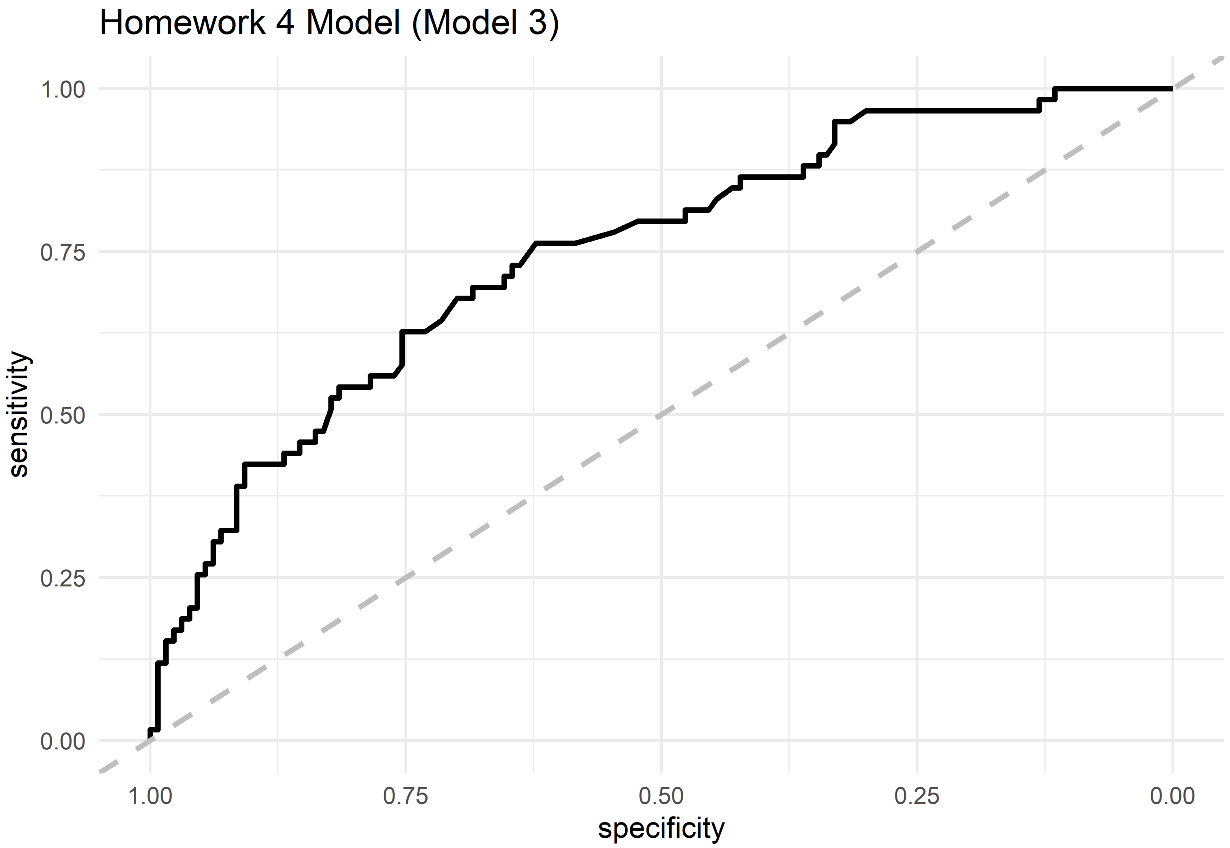
Table 3: Homework 4 Stepwise Regression Model

	Dependent variable:
	low_birth_weight
weight	0.984**
	(0.970, 0.996)
raceother	2.231*
	(0.949, 5.428)
raceblack	5.331***
	(1.991, 14.990)
smokeyes	2.206*
	(1.009, 4.969)
premature_laboryes	3.583***
	(1.517, 8.739)
hypertensionyes	4.633**
	(1.085, 21.000)
Constant	1.052
	(0.175, 7.017)
Observations	189
Log Likelihood	−100.400
Akaike Inf. Crit.	214.700
Note:	*p<0.1; **p<0.05; ***p<0.01

The final model from Homework 4 is…

$$log\left(\frac{P(LBW)}{1-P(LBW)}\right) = \beta_0 + \beta_1 weight + \beta_2 race + \beta_3 smoke + \beta_4 premature_labor + \beta_5 hypertension$$

We can look at the coefficients in the following table, for *Model 3.* We can also examine a ROC curve of this model.



The overall curve has an $AUC = 0.7477$. In choosing a cutpoint, we should consider first two concepts: 1) clinical utility of predicting LBW babies, and 2) effectiveness of the model.

1. Clinically, a LBW baby is a high-risk baby, and knowing if this will be the case is important. We would rather overdiagnosis the potential of a LBW then be incorrect. We want a high sensitivity, and can tolerate a low specificity.
2. Statistically, which we will not do here, we could use the maximum distance from the diagonal line as our choice for cutpoint selection. We could use bootstrapping methods to optimize this cutpoint, but it may not be as clinically relevant.

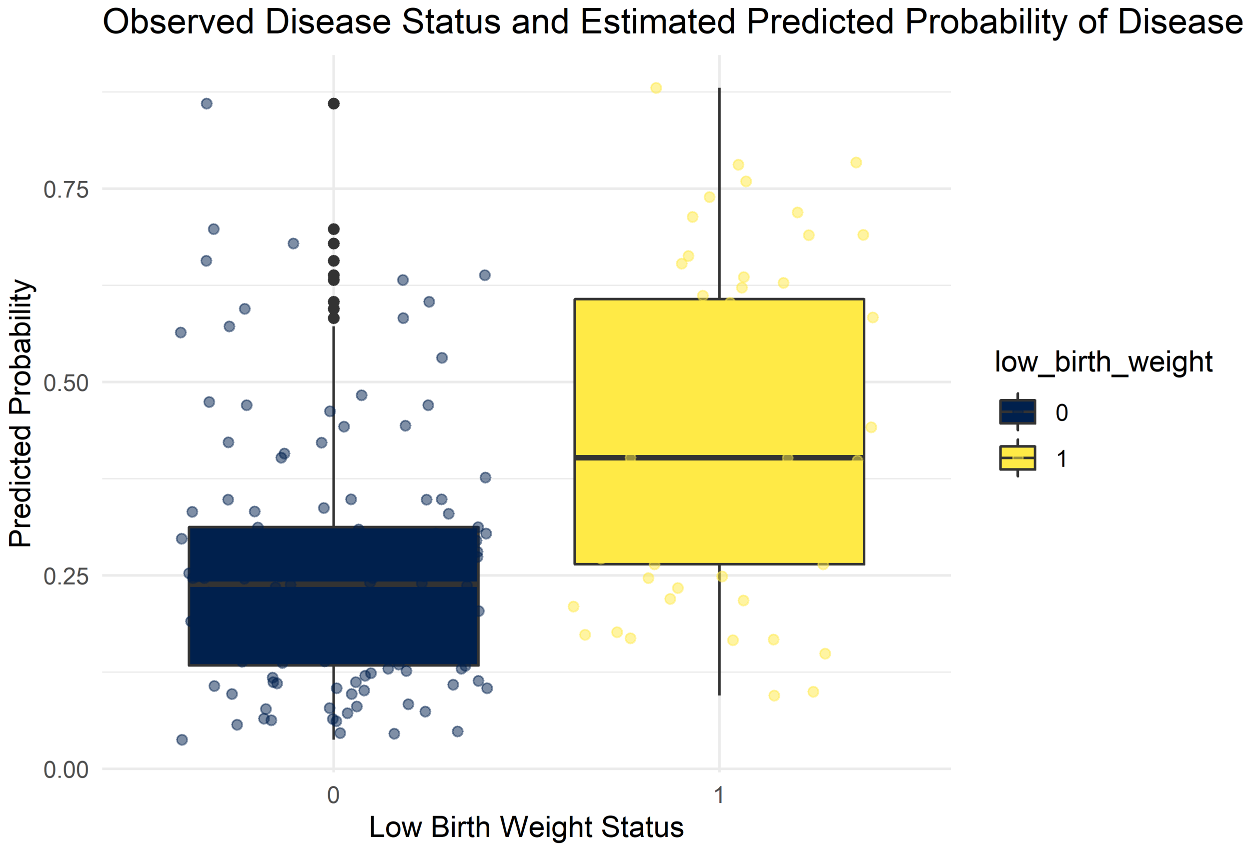
We will change our cutpoint such that a probabiliy greater than 2/3 is considered a positive prediction, and less is a negative prediction. With that level of cutoff, we have the following attributes of our algorithm's predictivity.

	x
Sensitivity	0.9769
Specificity	0.1525
Pos Pred Value	0.7175
Neg Pred Value	0.7500
Precision	0.7175
Recall	0.9769
F1	0.8274
Prevalence	0.6878
Detection Rate	0.6720
Detection Prevalence	0.9365
Balanced Accuracy	0.5647

Question 4

Using better model, do a box plot for observed disease status and estimated predicted probability for disease.

The best model that we have generated so far is *Model 3*, developed from stepwise regression from *Homework 4*.



Question 5

Report Hosmer-Lemeshow test and likelihood ratio test for better model.

For the Hosmer-Lemeshow test for *Model 3*, seen in *Question 3*, the $\chi^2 = 10.094$, with the associated $p = 0.2585$. This suggests that there is not enough evidence to reject the null, so we can accept that this model is a good fit.

We will compare this *Model 3* with *Model 1*, using the Likelihood Ratio Test. The $\chi^2 = 23.0991$, which has an associated $p = 0$. This suggests that *Model 3* is a better fit than *Model 1* as well.