Reflective Essay

STAT448

Tianqi Wu

12/15/2018

1.Introduction

After taking this course, I learnt a lot of techniques involved in advanced data analysis that I never knew before. Based on the types of the predictors and response variables, I now can apply different kinds of methods to analyze the data sets accordingly. I am also aware of the importance of model diagnosis that we should check relevant test results to ensure if assumptions are satisfied. By interpreting the parameter estimates and model results, I can understand the behavior of the model better. From simple inference of basic statistics to analysis of complicated model, I surely developed better understanding of advanced data analysis overall. In what follows, I will discuss three unique statistical methods that I believe are quite useful in my future career.

2.Data Analysis

2.1 Horseshoe Crab Data — Logistic Regression

This dataset first appears in chapter 8's programming exercises and it captures factors that affect whether the female crab had any other males, called satellites, residing nearby her. First, I learned that we should understand the variables. Color and Spine are the categorical predictors. Width and weight are continuous predictors. Since the response variable y is binary, we may use logistic regression model to analyze the dataset.

Before taking this course, I may simply put all the variables in the model. Now, I am aware that we could use automatic procedure(forward, backward and stepwise) or information criterion measure(AIC, BIC and Mallow's Cp) to select the predictors. Among the three automatic procedures, forward and stepwise are believed to have better results. Hence, I will fit a logistic regression model using stepwise automatic selection for the response y with the 4 predictors.

Here is the code that produces the logistic regression:

```
proc logistic data=crab;
 class color(param=ref ref='4') spine(param=ref ref='3');
 model y(event='1') = width weight color spine/
selection=stepwise rsquare lackfit;
 output out=crab_cbar CBAR = CBAR;
run;

proc print data = crab_cbar;
where CBAR>0.5;
run;
```

First, we may check if there is any extreme influential point. This step is important since influential point could impact the result of model largely, which I never knew before. Here, I check the Cbar and see if there is any point with Cbar > 0.5. With this cut-off value suggested from the lecture, there is no extreme influential point and we do not need to remove any of the observations. We see that only width is kept using the stepwise selection since it has p-value <0.0001, which meets the 0.05 significance level.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Width		1	1	27.8752		<.0001

Then, we may also interpret the parameter estimates using odds ratios. I learnt that we should specify the response level or category of interest for response variable and use reference coding for categorical predictors. Here, I specify the category of interest for response to be 1(the crab has at least one satellite) and I used reference coding 4 and 3 for predictors color and spine so that I may easily interpret them. However, continuous predictor width is the only significant variable in this case. The odds ratio may be understood as following: the odds of crab having at

least one satellite is 1.644 times odds for a 1-unit increase in width. Hence, crabs with larger width are more likely to have satellites.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Width	1.644	1.347	2.007

Finally, we can interpret the model results. The max re-scaled R square of the model is 0.2271. Hence, the predictor width captures 22.71 percent of variation in the response. It means that the model does not predict well. Also, the area under the curve c value of 0.742 indicates relatively weak predicting power. However, the Hosmer and Lemeshow Goodness-of-Fit Test significance (p-value > 0.7309) does not reject the null and we conclude that the model predicts fine.

R-Square	0.1655	Max-rescaled R-Square	0.2271

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	73.5	Somers' D	0.485
Percent Discordant	25.0	Gamma	0.492
Percent Tied	1.5	Tau-a	0.224
Pairs	6882	c	0.742

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.2465	8	0.7309

2.2 Horseshoe Crab Data — Discriminant Function Analysis

At the end the course, I learnt discriminant function analysis, which I has never used before. It is really useful to classify new observations into one of the known classes assuming we have a data set containing a classification variable and some numeric variables. I will use the Horseshoe Crab Data for illustration. The binary classification variable is y and numeric variables are width and weight. We may also select the predictors using function stepdisc in this case.

Here is the code for stepdisc procedure:

```
proc stepdisc data=crab sls=0.05 sle=0.05;
class y;
var width weight;
run;
```

Again, only width is kept in the model. The selection is the same as the logistic regression. Then, we may perform a discriminant analysis:

```
proc discrim data=crab manova pool=test crossvalidate;
class y;
var width;
run;
```

After taking the course, I am aware that there are two types of parametric methods: Linear Discriminant Analysis(LDA) and Quadratic Discriminant Analysis(QDA). To test which one is more appropriate, we check Test of Homogeneity of Within Covariance Matrices. The Chi-square test is significant at the 0.1 level since it has p-value > 0.0677. Hence, we need to account for differences of covariance across groups and QDA is more appropriate.

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
3.337917	1	0.0677

Then, we may look at the MANOVA result. All tests are all highly significant with p-value < .0001. Therefore, we reject the null and conclude that group means are different. Hence, it is very likely to discriminate between response Y based on width. The misclassification rate for re-substitution is 32.91%. It indicates that the model does not work very well. 22.58% of Y=0 is misclassified as Y=1 and 43.24% of Y=1 is misclassified as Y=0. I now know that re-substitution may create bias in the classification rates. It is better to use cross-validation to address the concern. In this case, the misclassification rate is the same for re-substitution and cross-validation.

The DISCRIM Procedure

Multivariate Statistics and Exact F Statistics					
S=1 M=-0.5 N=84.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.83887157	32.85	1	171	<.0001
Pillai's Trace	0.16112843	32.85	1	171	<.0001
Hotelling-Lawley Trace	0.19207759	32.85	1	171	<.0001
Roy's Greatest Root	0.19207759	32.85	1	171	<.0001

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CRAB
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Y			
From Y	0	1	Total
0	48 77.42	14 22.58	62 100.00
1	48 43.24	63 56.76	111 100.00
Total	96 55.49	77 44.51	173 100.00
Priors	0.5	0.5	

Error Count Estimates for Y			
	0	1	Total
Rate	0.2258	0.4324	0.3291
Priors	0.5000	0.5000	

2.3 Baseball Data — Principal Component Analysis

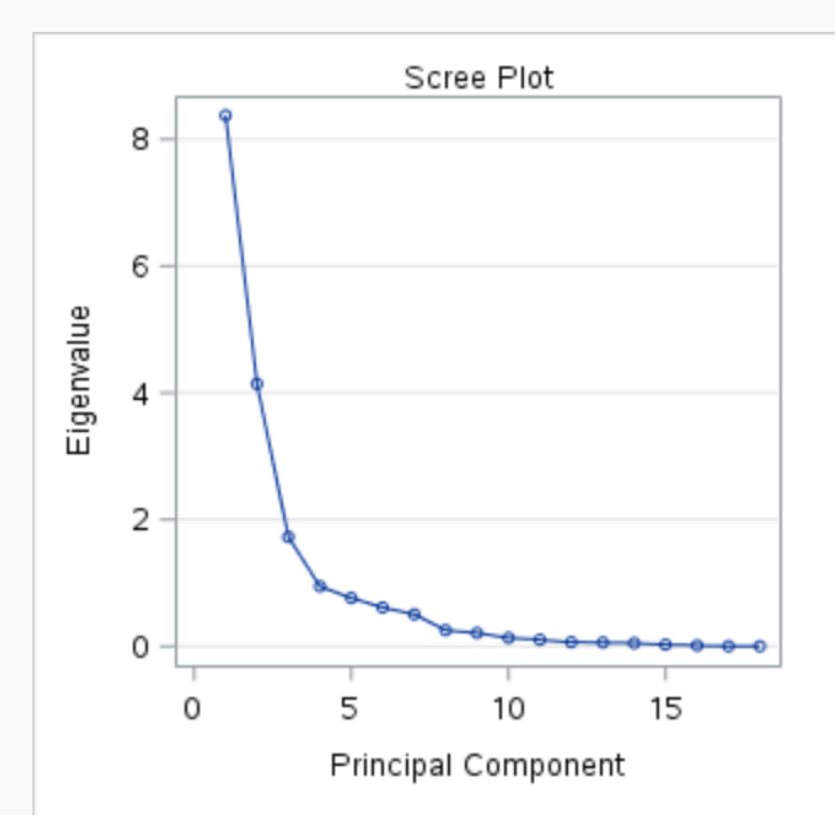
This dataset appears in the chapter 16's programming exercise. Looking the dataset, it has a lot of variables and I may not know how to deal with it before taking this class. Now, I learnt that Principal Component Analysis(PCA) may be applied to this kind of 'big data' to achieve dimension reduction. Here is the code for PCA:

```
proc princomp data=sashelp.baseball;
run;
```

In order to decide how many principal components to keep, there are several ways. First, we may retain the number of components with particular cumulative proportion of variations. We decide to keep first three principal components since they retain reasonably high cumulative proportion of variations (79.15%). We can also keep components with corresponding eigenvalues greater than the average eigenvalue. Since we use the correlation matrix in this case, the average eigenvalue is just 1. There are three components having eigenvalues greater than 1. Finally, we could also draw a scree plot and select the number of eigenvalues above the elbow. From the graph, we can see that the elbow occurs at 4 and we just keep the first three. Hence, all three methods of choosing principal components agree.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	8.37937437	4.23953681	0.4655	0.4655
2	4.13983756	2.41283055	0.2300	0.6955
3	1.72700701	0.78150352	0.0959	0.7915
4	0.94550349	0.18233653	0.0525	0.8440



With PCA, we could describe large amount of original variation in much fewer new ‘variables’. Also, this process removes correlation underlying. However, there are also some trade-offs. For example, we do not really know what the principal components represent and it makes hard for us to interpret. Moreover, it may throw away some percentage of the original information.

3. Conclusion

To conclude, with different statistical methods that I learnt in the course, I feel more confident about the general procedure of data analysis. I am able to decide which model to use depending on the variables of the data sets and goal of investigation. I also developed better understanding of the models discussed in the course. Learning the theory behind the models, I can now do model diagnosis and interpret the results better. Most of the techniques taught in the course are

so popular and useful that I will definitely benefit from them in my future career. I am glad that I took this course of advanced data analysis.