WEEK 6 ASSIGNMENT

Concepts of Statistics 2 – DATA-51200 | Spring 2 2020

Christina Morgenstern

---

1. On your own words, summarize (in less than one page) the validation of results in Logistic Regression

Building a Logistic Regression model for prediction requires a final analysis to assess the generalizability of the model. This validation step ensures that the results are most descriptive of the data and can be generalized to the total population. Most often this procedure involves the use of a holdout or validation sample which comprises of data separate from the analysis sample used to develop the model. In using this holdout sample the logistic model is used to assess the predictive accuracy of this unseen data. Since this data has not been used for the model building, the result should demonstrate the generalizability of the logistic model.

Cross-validation is an alternative approach in model validation where the data is partitioned into equally sized fragments and each subsample used to compute the logistic model. The "jackknife method" is widely used in this approach and makes use of the "leave-one-out" principle. Thereby the analysis is performed on k-1 subsamples with one observation eliminated at a time. Applying the logistic model and estimating the group membership of the eliminated observation. Once all subsamples have been analyzed, the confusion matrix and calculated hit ratios are used to validate the results.

References:

[1] Multivariate Data Analysis by Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolphe E. Anderson, Pearson, 8th edition, 2019

2. For the data set associated with this homework (HBAT) (you may use any software and programming language you feel comfortable dealing with. Make sure to include your codes, diagrams and results). Using X4 as the non-metric response variable and (X6 up to X15) as the metric variables:

a. Apply forward selection binary logistic regression (1 is the level of interest with single non-cross effects) and report what variable is entered into the model after each step. (Use 0.05 significance level). Report the final summary of the regression model and the ROC curve and the area under the ROC curve after each step.

The objective of applying Logistic Regression to the HBAT data set is to identify differences in the perception of customers based on their geographic location.

The HBAT data set was loaded into SAS Studio and the model Linear Logistic Regression applied using X4 as the dependent categorical variable and variables X5 to X15 as independent metric variables.

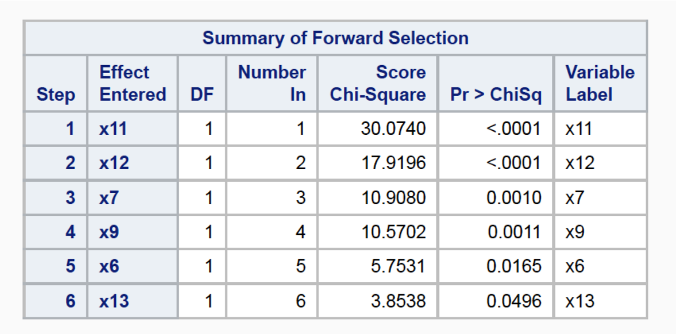
```
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.IMPORT plots=(roc);
      model x4(event='1')=x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 / link=logit lackfit
            rsquare selection=forward slentry=0.05 hierarchy=single details
            technique=fisher;
run;
```

Using the forward selection procedure and a significance level of 0.05, the following variables are entered into the model after each step:

Step 1: X11
Step 2: X12
Step 3: X7
Step 4: X9
Step 5: X6
Step 6: X13

No additional effects meet the 0.05 significance criterion and thus no further variables are entered to the model after step 6. The model is summarized in table 1.



Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	x11	1	1	30.0740	<.0001	x11
2	x12	1	2	17.9196	<.0001	x12
3	x7	1	3	10.9080	0.0010	x7
4	x9	1	4	10.5702	0.0011	x9
5	x6	1	5	5.7531	0.0165	x6
6	x13	1	6	3.8538	0.0496	x13

Table 1. Summary of forward selection.

The ROC (Receiver Operating Characteristic) curve plots sensitivity (the true positive rate) on the y-axis and 1-specificity (the true negative rate) and is used for interpreting model fit. The result for the ROC curve of the logistic regression model applied to the HBAT data set with forward selection is shown in Figure 1. The performance of the model is depicted as the blue line which lies above the diagonal and thus represents good classification results (better than random). The area under curve (AUC) is 0.9676 and provides an overall test of predictive accuracy. Since this value is close to 1 (> 0.90) the predictive accuracy of this model is considered excellent.

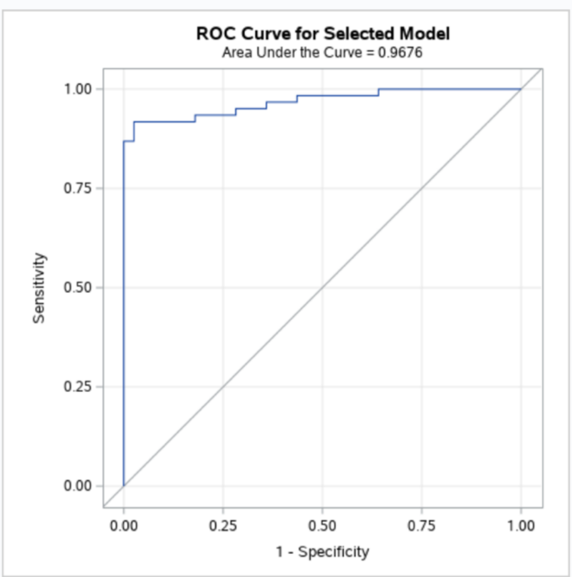


Figure 1. ROC curve for linear regression model with forward selection.

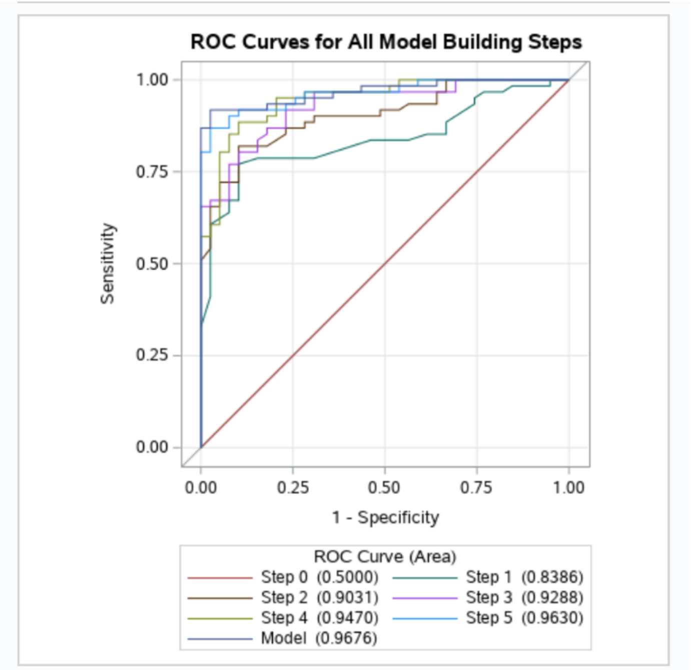


Figure 2. ROC curves for all model building steps of linear regression model with forward selection.

Upon adding the variables after each step of forward selection, the area under the curve increases from 0.500 at step 0 and the base model with random explanation to 0.8386 after adding variable X11 at step 1. The model performance as judged by the area under the curve increases with each step until the maximum value of 0.9676 is reached.

b. Apply backward selection binary logistic regression (1 is the level of interest with single noncross effects) and report what variable is eliminated from the model after each step. (Use 0.05 significance level). Report the final summary of the regression model and the ROC curve and the area under the ROC curve after each step.

The following SAS code was generated, when applying binary logistic regression with backward selection using a 0.05 significance level.

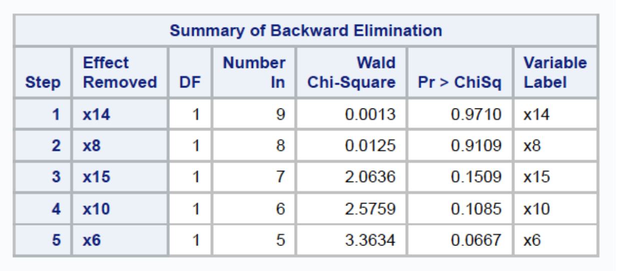
```
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.IMPORT plots=(roc);
      model x4(event='1')=x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 / link=logit lackfit
             rsquare selection=backward slstay=0.05 hierarchy=single details
             technique=fisher;
run;
```

Using the backward selection process of generating the logistic regression model, the following variables were removed after each step:

Step 1: X14
Step 2: X8
Step 3: X15
Step 4: X10
Step 5: X6

No additional effects meet the 0.05 significance criterion and thus no further variables were removed from the model after step 5. The summary of the model building process can be seen in table 2.



Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	x14	1	9	0.0013	0.9710	x14
2	x8	1	8	0.0125	0.9109	x8
3	x15	1	7	2.0636	0.1509	x15
4	x10	1	6	2.5759	0.1085	x10
5	x6	1	5	3.3634	0.0667	x6

Table 2. Summary of backward selection.

The performance of the logistic regression model estimated using backward selection is visualized using the ROC curve (Figure 2). Since the blue curve lies above the diagonal, the predictive ability of this model is considered better than random. The value for the area under the curve is 0.9651 demonstrating, a perfect prediction.

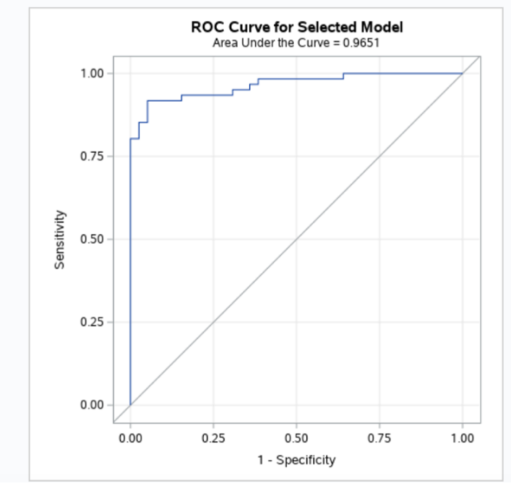


Figure 3. ROC curve for linear regression model with backward selection.

Comparing the ROC curves and area under the curves for all separate model building steps from the backward selection process (see Figure 3), shows that the performance of the model was decreasing after each step with removal of individual variables. The initial model containing all variables had an AUC value of 0.9731. After step 2 (and removal of variables X14 and X8), the performance also slightly increased to 0.9739 suggesting that this model would have been a perfect fit.

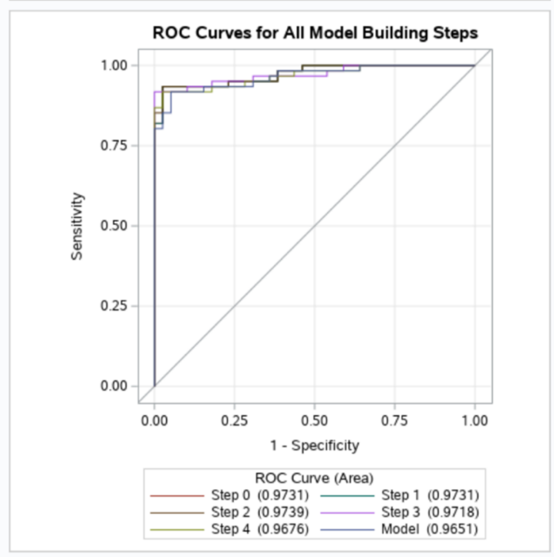


Figure 4. ROC curves for all model building steps from backward selection linear regression model.

c. Which selection method from (a) or (b) provides better model? Explain.

Comparing the two final logistic regression models after forward and backward selection shows that the forward selection model does slightly better. When comparing the AUC values for the final models, the forward selection model achieves a value of 0.9676 whereas the backward selection model reaches a value of 0.9651.