Modern Data Mining, HW 3

Jose Cervantez

Bethany Hsaio

Rob Kuan

$11{:}59~\mathrm{pm},\,03/17,\,2024$

Contents

1	PartI: Mo	del Building							2
2	Part II: L	ogistic Regression							3
	2.1 Frami	ngham heart disease stu	dy			 	 	 	3
	2.1.1	Identify risk factors .				 	 	 	4
		2.1.1.1 Understand t	he likelihood func	tion		 	 	 	4
		2.1.1.2 Identify impo	rtant risk factors	for Heart.Di	sease.	 	 	 	6
		2.1.1.3 Model building	ng			 	 	 	8
	2.1.2	Classification analysis				 	 	 	10
		2.1.2.1 ROC/FDR.				 	 	 	10
		2.1.2.2 Cost function	/ Bayes Rule			 	 	 	13

1 PartI: Model Building

Multiple regression is one of the most popular methods used in statistics as well as in machine learning. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we could to determine the form of the response as well as the function format for the factors. Then, when we have many possible features to be included in the working model it is inevitable that we need to choose a best possible model with a sensible criterion. Regularizations such as LASSO are introduced. Be aware that if a model selection is done formally or informally, the inferences obtained with the final lm() fit may not be valid. Some adjustment will be needed. This last step is beyond the scope of this class. Check the research line that Linda and collaborators have been working on.

The main job in this part is a rather involved case study about devastating covid19 pandemic. Please read through the case study first. This project is for sure a great one listed in your CV.

For covid case study, the major time and effort would be needed in EDA portion.

Answer:

See file covid_case_study_2024.pdf for our work and answers to Part I.

2 Part II: Logistic Regression

Logistic regression is used for modeling categorical response variables. The simplest scenario is how to identify risk factors of heart disease? In this case the response takes a possible value of YES or NO. Logit link function is used to connect the probability of one being a heart disease with other potential risk factors such as blood pressure, cholestrol level, weight. Maximum likelihood function is used to estimate unknown parameters. Inference is made based on the properties of MLE. We use AIC to help nailing down a useful final model. Predictions in categorical response case is also termed as Classification problems. One immediately application of logistic regression is to provide a simple yet powerful classification boundaries. Various metrics/criteria are proposed to evaluate the quality of a classification rule such as False Positive, FDR or Mis-Classification Errors.

LASSO with logistic regression is a powerful tool to get dimension reduction. We will not use it here in this work.

2.1 Framingham heart disease study

We will continue to use the Framingham Data (Framingham.dat) so that you are already familiar with the data and the variables. All the results are obtained through training data.

Liz is a patient with the following readings: AGE=50, GENDER=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0. We would be interested to predict Liz's outcome in heart disease.

To keep our answers consistent, use a subset of the data, and exclude anyone with a missing entry. For your convenience, we've loaded it here together with a brief summary about the data.

We note that this dataset contains 311 people diagnosed with heart disease and 1095 without heart disease.

0 1 1095 311

After a quick cleaning up here is a summary about the data:

```
DBP
HD
               AGE
                               SEX
                                              SBP
0:1086
          Min.
                 :45.0
                          FEMALE:730
                                         Min.
                                                 : 90
                                                                : 50.0
                                                        Min.
1: 307
          1st Qu.:48.0
                          MALE
                                :663
                                         1st Qu.:130
                                                        1st Qu.: 80.0
          Median:52.0
                                         Median:142
                                                        Median: 90.0
          Mean
                  :52.4
                                         Mean
                                                 :148
                                                        Mean
                                                                : 90.2
          3rd Qu.:56.0
                                                        3rd Qu.: 98.0
                                         3rd Qu.:160
          Max.
                  :62.0
                                         Max.
                                                 :300
                                                        Max.
                                                                :160.0
                     FRW
     CHOL
                                    CIG
Min.
        : 96
                       : 52
                               Min.
                                       : 0
               Min.
1st Qu.:200
               1st Qu.: 94
                               1st Qu.: 0
Median:230
               Median:103
                               Median: 0
                                       : 8
Mean
        :235
               Mean
                       :105
                               Mean
3rd Qu.:264
               3rd Qu.:114
                               3rd Qu.:20
Max.
        :430
                       :222
                               Max.
                                       :60
               Max.
```

Lastly we would like to show five observations randomly chosen.

```
HD AGE
               SEX SBP DBP CHOL FRW CIG
        61
                              248 104
                                        20
643
     1
              MALE 140
                         68
11
     0
        45
              MALE 110
                         88
                              183
                                   90
                                         0
        58
                              296 100
576
     1
              MALE 150
                         95
                                        15
        59
              MALE 260 130
                              246 111
                                        20
560
     1
702
     0
        45 FEMALE 122
                        74
                              178
                                         5
```

2.1.1 Identify risk factors

- **2.1.1.1** Understand the likelihood function Conceptual questions to understand the building blocks of logistic regression. All the codes in this part should be hidden. We will use a small subset to run a logistic regression of HD vs. SBP.
 - i. Take a random subsample of size 5 from hd_data_f which only includes HD and SBP. Also set set.seed(471). List the five observations neatly below. No code should be shown here.

Answer:

```
HD SBP
643 1 140
11 0 110
576 1 150
560 1 260
702 0 122
```

ii. Write down the likelihood function using the five observations above.

Answer: The likelihood function using the five observations is:

$$\mathcal{L}\left(\beta_{0},\beta_{1}\mid \text{Data}\right) = \frac{e^{\beta_{0}+156\beta_{1}}}{1+e^{\beta_{0}+156\beta_{1}}} \cdot \frac{e^{\beta_{0}+164\beta_{1}}}{1+e^{\beta_{0}+164\beta_{1}}} \cdot \frac{1}{1+e^{\beta_{0}+156\beta_{1}}} \cdot \frac{1}{1+e^{\beta_{0}+138\beta_{1}}} \cdot \frac{1}{1+e^{\beta_{0}+155\beta_{1}}}$$

iii. Find the MLE based on this subset using glm(). Report the estimated logit function of SBP and the probability of HD=1. Briefly explain how the MLE are obtained based on ii. above.

Answer:

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
## Call:
  glm(formula = HD ~ SBP, family = binomial, data = hd data.f[index,
##
##
##
##
  Coefficients:
##
                Estimate Std. Error z value Pr(>|z|)
                          622609.03
                                           0
##
                 -334.96
  (Intercept)
                                           0
## SBP
                    2.56
                            4730.05
                                                    1
##
##
  (Dispersion parameter for binomial family taken to be 1)
##
##
       Null deviance: 6.7301e+00 on 4 degrees of freedom
## Residual deviance: 4.0651e-10 on 3 degrees of freedom
  AIC: 4
##
## Number of Fisher Scoring iterations: 25
```

Based on maximum likelihood estimation, the equation for the estimated model is as follows.

$$Logit = -334.96 + 2.56 \times SBP$$

The probability of HD=1 is equal to:

$$\hat{P}(HD = 1 \mid SBP) = \frac{e^{-334.96 + 2.56 \times SBP}}{1 + e^{-334.96 + 2.56 \times SBP}}$$

NOTE: the five samples that are randomly chosen, when classified, have **perfect separation**. This is why the parameters values are so extreme.

The maximum likelihood was obtained using the Fisher Scoring algorithm (which uses the Fisher Information Matrix) to estimate the parameters that are most likely to have generated the observed data. The algorithm iteratively updates the parameter estimates until the log-likelihood converges to a maximum value.

iv. Evaluate the probability of Liz having heart disease.

Answer:

The probability of Liz having heart disease is 2.22e-16 (which is an extremely small probability close to zero). Again, this is because the five data observations selected using seed(471) have **perfect separation** and can be classified with 100% accuracy.

[1] "Liz's probability of heart disease is 2.22044604925031e-16"

- 2.1.1.2 Identify important risk factors for Heart.Disease. We focus on understanding the elements of basic inference method in this part. Let us start a fit with just one factor, SBP, and call it fit1. We then add one variable to this at a time from among the rest of the variables. For example
 - i. Which single variable would be the most important to add? Add it to your model, and call the new fit fit2.

We will pick up the variable either with highest |z| value, or smallest p value. Report the summary of your fit2 Note: One way to keep your output neat, we will suggest you using **xtable**. And here is the summary report looks like.

Answer:

SEX is the most important factor, as it is has the highest Z-value and results in the lowest AIC of all the other models. A summary of the table is below:

std.error $_{\rm term}$ estimate statistic p.value (Intercept) -4.570.3897-11.739.289e-32SBP 0.018720.0023248.053 8.071e-16SEXMALE 0.90340.13986.4641.02e-10

Table 1: Summary of fit2 model

ii. Is the residual deviance of fit2 always smaller than that of fit1? Why or why not?

Answer:

Yes, the residual deviance of fit2 (no matter which predictor variables is used) will always be smaller than that of fit1. This is because adding an additional predictor will always add additional information to the model, which will reduce the residual deviance.

However, the AIC may not improve, because that takes into account the number of parameters in the model.

iii. Perform both the Wald test and the Likelihood ratio tests (Chi-Squared) to see if the added variable is significant at the .01 level. What are the p-values from each test? Are they the same?

Answer:

The p-values of both tests are displayed in the tables below. They are not the same, because the Wald test, the likelihood ratio test, and the score test (not mentioned in class) are all different ways of estimating the significance of the added variable. Even those these three tests are asymptotically equivalent, in finite samples they may have different estimates.

Generally speaking, the likelihood ratio test is superior than the Wald Test because it does not require estimating an additional parameter (the standard deviation of the coefficient in fit 2).

Table 2: Likelihood Ratio Test (Chisquared)

#Df	LogLik	Df	Chisq	Pr(>Chisq)
3	-686.9	NA	NA	NA
2	-708.7	-1	43.7	3.828e-11

Table 3: Wald Test (Chisquared)

Res.Df	Df	Chisq	Pr(>Chisq)
1391	NA	NA	NA
1390	1	41.78	1.02e-10

- **2.1.1.3** Model building Start with all variables. Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).
 - i. Use backward selection method. Only keep variables whose coefficients are significantly different from 0 at .05 level. Kick out the variable with the largest p-value first, and then re-fit the model to see if there are other variables you want to kick out.

Answer:

Table 4: Summary of Best Model (Backwards Selection Method)

term	estimate	std.error	statistic	p.value
(Intercept)	-8.409	0.9086	-9.255	2.151e-20
AGE	0.05664	0.0145	3.906	9.377e-05
SEXMALE	0.9899	0.1451	6.824	8.842e-12
SBP	0.01696	0.002362	7.179	7.021e-13
CHOL	0.00448	0.001495	2.996	0.002737

ii. Use AIC as the criterion for model selection. Find a model with small AIC through exhaustive search. Does exhaustive search guarantee that the p-values for all the remaining variables are less than .05? Is our final model here the same as the model from backwards elimination?

Answer:

Exhaustive search does not gaurantee that all the remaining variables are less than 0.05. It only guarantees that the model with the smallest AIC is selected.

This model is different than the model selected from backwards elimination.

Table 5: Summary of Best Model (Exhaustive Search Method)

term	estimate	std.error	statistic	p.value
(Intercept)	-9.228	0.9962	-9.263	1.979e-20
AGE	0.06153	0.01478	4.164	3.122e-05
SEXMALE	0.9113	0.1571	5.8	6.633e-09
SBP	0.01597	0.002487	6.42	1.366e-10
CHOL	0.004493	0.001503	2.99	0.002794
FRW	0.006039	0.004004	1.508	0.1315
CIG	0.01228	0.006088	2.017	0.04369

iii. Use the model chosen from part ii. as the final model. Write a brief summary to describe important factors relating to Heart Diseases (i.e. the relationships between those variables in the model and heart disease). Give a definition of "important factors".

Answer:

Table 6: Increase in odds for each one unit increase in predictor

predictor	odds_increase
AGE	6.3%
SEXMALE	148.7%

predictor	odds_increase
SBP	1.6%
CHOL	0.5%
FRW	0.6%
CIG	1.2%

The important factors are not just coefficients that are statistically significant in the model. They must also have large betas that influence the probability of somebody having heart disease.

The final model shows that the following factors are significantly different than 0, and are correlated with increased odds of being diagnosed with heart disease: SBP, SEX, AGE, CHOL, and CIG.

- SEX: Being male vs. female increases your odds of heart disease by 148%
- AGE: For every year increase in age, the odds of having heart disease increase by 6.3%
- SBP: For every unit increase in systolic blood pressure, the odds of having heart disease increase by 1.6%
- CHOL: For every unit increase in cholestrol, the odds of having heart disease increase by 0.5%
- CIG: For every weekly self-reported cigarette smoked, the odds of having heart disease increase by 1.2%

Summary

When taking into account the beta coefficients, and the potential range of values for the individuals in the dataset, it seems that SEX, AGE, and SBP are the most important factors in predicting heart disease. AGE has a large absolute unit level impact on odds, and SEX and AGE also have a large impact when taking into account that they are continuous variables and there is a wide range of values in the dataset. CIG and CHOL may also be an important factor if individuals have extreme values for these variables.

Controlling for these other variables, we do not have enough evidence to reject the null hypothesis that the coefficient for FRW = 0, hence we cannot say if FRW is an important factor.

iv. What is the probability that Liz will have heart disease, according to our final model?

Answer:

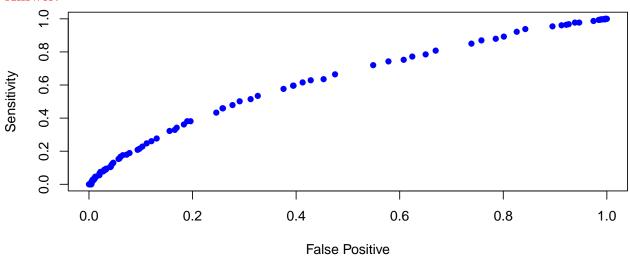
Liz's probability of heart disease is 3.46%.

2.1.2 Classification analysis

2.1.2.1 ROC/FDR

i. Display the ROC curve using fit1. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

Answer:



Optimal threshold: 0.298

True Positive Rate at this threshold: 0.215 False Positive Rate at this threshold: 0.0976

The ROC curve is a graphical representation of the true positive rate (TPR) and the false positive rate (FPR) for every possible threshold value. The ROC curve is also helpful for calculating metrics such as AUC (Area Under Curve) which can summarize the discrimination ability of the model.

The classifer where the false positive rate is less than 0.1 and the true positive rate is as high as possible is when the threshold is set at 0.298%:

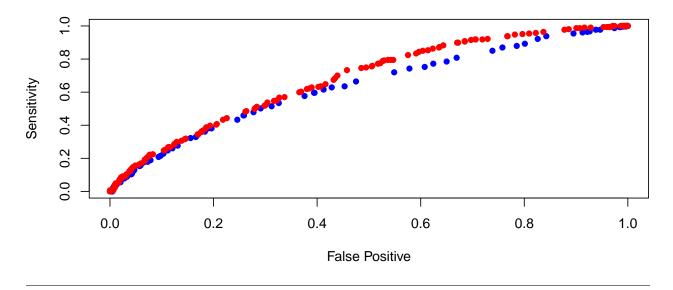
$$\widehat{HD} = 1$$
 if $\widehat{P}(HD = 1 \mid SBP) > 0.298$

ii. Overlay two ROC curves: one from fit1, the other from fit2. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

Answer:

The ROC curve from fit1 is always contained within the ROC curve from fit2. This is because fit2 has all the predictor variables in fit1, and so the ROC curve from fit2 will always have a higher true positive rate and a lower false positive rate than fit1 because of the additional incremental information that can help improve the classifier.

See the chart below.



iii. Estimate the Positive Prediction Values and Negative Prediction Values for fit1 and fit2 using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values?

Answer:

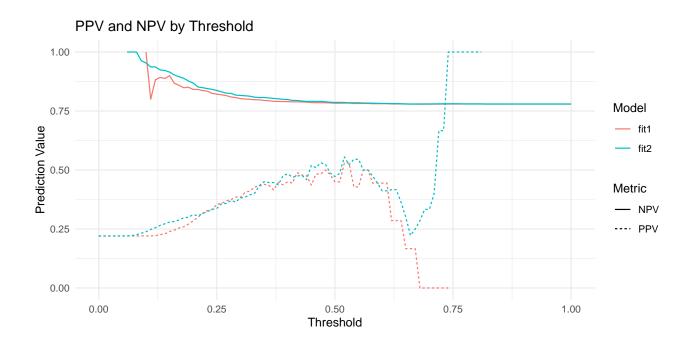
fit2 is more desirable if we prioritize the Positive Prediction values. See table below:

Model	Positive_Prediction_Value	Negative_Prediction_Value		
fit1	0.02932	0.9899		
$\operatorname{fit} 2$	0.05537	0.9825		

Answer:

If positive and negative prediction values are the concerns, I would choose fit2 as it generally has higher positive prediction values (specifically for thresholds 0.1 to 0.5) and higher negative prediction values (on average) than fit1.

iv. For fit1: overlay two curves, but put the threshold over the probability function as the x-axis and positive prediction values and the negative prediction values as the y-axis. Overlay the same plot for fit2. Which model would you choose if the set of positive and negative prediction values are the concerns? If you can find an R package to do so, you may use it directly.



2.1.2.2 Cost function/ Bayes Rule Bayes rules with risk ratio $\frac{a_{10}}{a_{01}} = 10$ or $\frac{a_{10}}{a_{01}} = 1$. Use your final model obtained from Part 1 to build a class of linear classifiers.

- i. Write down the linear boundary for the Bayes classifier if the risk ratio of $a_{10}/a_{01} = 10$.
- ii. What is your estimated weighted misclassification error for this given risk ratio?
- iii. How would you classify Liz under this classifier?
- iv. Bayes rule gives us the best rule if we can estimate the probability of HD-1 accurately. In practice we use logistic regression as our working model. How well does the Bayes rule work in practice? We hope to show in this example it works pretty well.

Now, draw two estimated curves where x = threshold, and y = misclassification errors, corresponding to the thresholding rule given in x-axis.

- v. Use weighted misclassification error, and set $a_{10}/a_{01} = 10$. How well does the Bayes rule classifier perform?
- vi. Use weighted misclassification error, and set $a_{10}/a_{01} = 1$. How well does the Bayes rule classifier perform?