Assignment 3 — Classification

Haoran Duan[1]

MSc Data Science, School of Computing, Newcastle University

Abstract. Classification technology assignment 3.

1 Question (a)

```
#Check the response variable
> table(plasma$ESR)
[1]  0  1
   26  6
> plasma_data = data.frame(plasma[,1:2], ESR = as.integer(plasma$ESR) -
    1)
> xtable(head(plasma_data))
```

```
> apply(plasma_data[-3], 2, var)
fibrinogen  globulin
 0.4058565 21.0070565
> apply(plasma_data[-3], 2, mean)
fibrinogen  globulin
   2.78875  35.65625
> cor(plasma_data[-3])
           fibrinogen  globulin
fibrinogen 1.00000000 0.08071681
globulin  0.08071681 1.00000000
```

– There are two predictors varibles named fibrinogen and globulin.And the response variables ESR has two categories, and I use 0 or 1 to represent them.

– A little positive relationship by correlation matrix for two predictors.

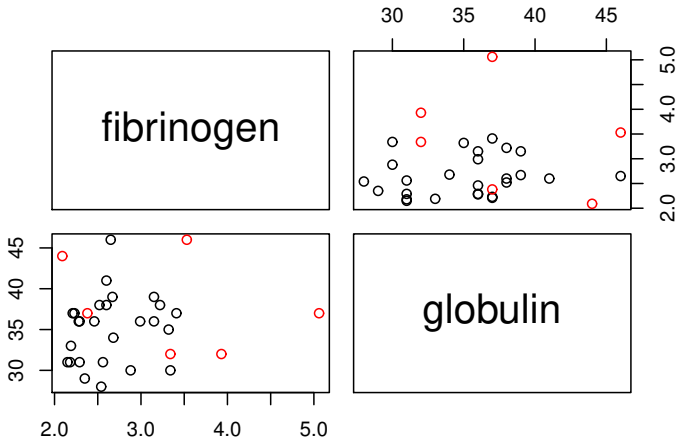

Fig. 1. Scatterplot for the data

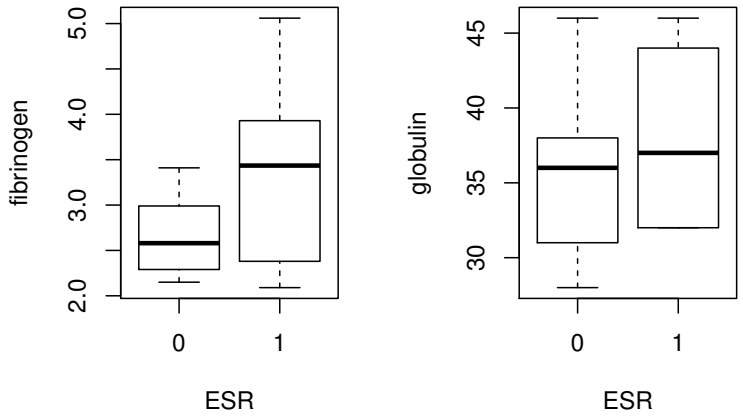

Fig. 2. Boxplot for the two predictors variables

– On the basis of the scatterplot in Fig 1, there is no obvious pattern.So I draw the boxplot next.

– We can check the media(black lines), they are different so i think the fibringe will influence the classification of ESR much more than gobulin. And the relationship between gobulin and ESR are not obvious.

2 Question (b)

```
> full_Model = glm(ESR ~ . ,data =plasma_data, family = 'binomial')
> summary(full_Model)
[1] Call:
glm(formula = ESR ~ ., family = "binomial", data = plasma_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9683  -0.6122  -0.3458  -0.2116   2.2636

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.7921     5.7963  -2.207   0.0273 *
fibrinogen    1.9104     0.9710   1.967   0.0491 *
globulin      0.1558     0.1195   1.303   0.1925
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1
    1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885 on 31 degrees of freedom
Residual deviance: 22.971 on 29 degrees of freedom
AIC: 28.971

Number of Fisher Scoring iterations: 5
```

– On the basis of the summaries of the full model, the only fibrinogen predictor will be included in my 'final model'. We only have two predictor variables, and the globulin has large p-value, this means that, individually, it contributes little to a model which contains both of the predictors.

– Also, we can use the the summary for the final model,the final model is better than a model with no predictors. Then, if we use the analysis of deviance,we can perform the hypothesis as(1), and the large p-value in the final column(Table 2) tells us we can not reject the final model and so we prefer the final model than full model.

Table 1.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	30	24.84			
2	29	22.97	1	1.87	0.1716

$$\begin{aligned} H_0 &: Final\_Model \\ H_1 &: Full\_Model \end{aligned} \tag{1}$$

3 Question (c)

```
> final_Model = glm(ESR ~ . ,data =plasma_data[,-2], family
                    ='binomial')
> summary(final_Model)
[1] Call:
glm(formula = ESR ~ ., family = "binomial", data = plasma_data[, -2])

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-0.9298 -0.5399 -0.4382 -0.3356  2.4794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.8451    2.7703 -2.471  0.0135 *
fibrinogen   1.8271    0.9009  2.028  0.0425 *
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1
    1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 30.885 on 31 degrees of freedom
Residual deviance: 24.840 on 30 degrees of freedom
AIC: 28.84

Number of Fisher Scoring iterations: 5
```

– The fibrinogen variable, it is positive, this indicates that the high(greater than 20) erythrocyte sedimentation rate generally more likely to associate with the high fibrinogen.

4 Question (d)

```
> table(plasma_data$ESR)
[1] 0  1
[2]26  6
> phat = predict(final_Model, plasma_data, type = 'response')
> yhat = ifelse(phat > 0.5, 1, 0)
> confusion_1 = table(Observed =plasma_data$ESR, Predicted = yhat)
              Predicted
[1]              0    1

          0     26    0
Observed
          1      4    2
> 1 - mean(plasma_data$ESR == yhat)
[1]0.125
```

– The confusion matrix, when the ESR greater than 20, the correct prediction 2 out of 2+4 = 6 occations.When the ESR smaller than 20, we get this correctly all the time, on 26 out of 26.So i think in training stage, our final model can classify the '0'(ESR less than 20) situations better than the '1'(ESR greater than 20).

5 Question (e)

```
> train_plasma = rbind(plasma_data[1:13,], plasma_data[27:29,])
> test_plasma = rbind(plasma_data[14:26,], plasma_data[30:32,])
> final_train = glm(ESR ~ fibrinogen, data = train_plasma, family   =
    'binomial')
> phat_e = predict(final_train, test_plasma, type = 'response')
> yhat_e = ifelse(phat_e > 0.5, 1, 0)
> confusion_2 = table(Observed =test_plasma$ESR, Predicted = yhat_e)
> 1 - mean(test_plasma$ESR == yhat_e)
[1] 0.125
```

– 0.125

6 Question (d)

```
> train_plasma = rbind(plasma_data[1:13,], plasma_data[27:29,])
> test_plasma = rbind(plasma_data[14:26,], plasma_data[30:32,])
> lda_train = lda(ESR ~ fibrinogen, data = train_plasma)
> lda_test = predict(lda_train, test_plasma, type = 'response')
> yhat_lda_test = lda_test$class
> 1 - mean(test_plasma$ESR == yhat_lda_test)
```

```
[1] 0.125
> train_plasma = rbind(plasma_data[1:13,], plasma_data[27:29,])
> test_plasma = rbind(plasma_data[14:26,], plasma_data[30:32,])
> qda_train = qda(ESR ~ fibrinogen, data = train_plasma)
> qda_test = predict(qda_train, test_plasma, type = 'response')
> yhat_qda_test = qda_test$class
> 1 - mean(test_plasma$ESR == yhat_qda_test)
[1] 0.125
```

– In my opinion, firstly, the test error is same, I think QDA is a bit complex .The dataset in this project are small and extra flexibility offered by QDA seems unnecessary,also low variance can be obtained by LDA in small dataset. So move on to LDA and Logistic Regression.Logistic Regression and LDA are a bit same, the estimate methods can be different, this project use a small dataset, if I consider to minimum the variance for the model, I think I prefer LDA, also the histogram for scaled fibrinogen Fig 3 ,it seems like a normal distribution, also the Fig 1 shows that there is no obviouse relationship between fibrinogen and ESR, for example, even can not assume it is like linear relationships.So I will choose LDA.

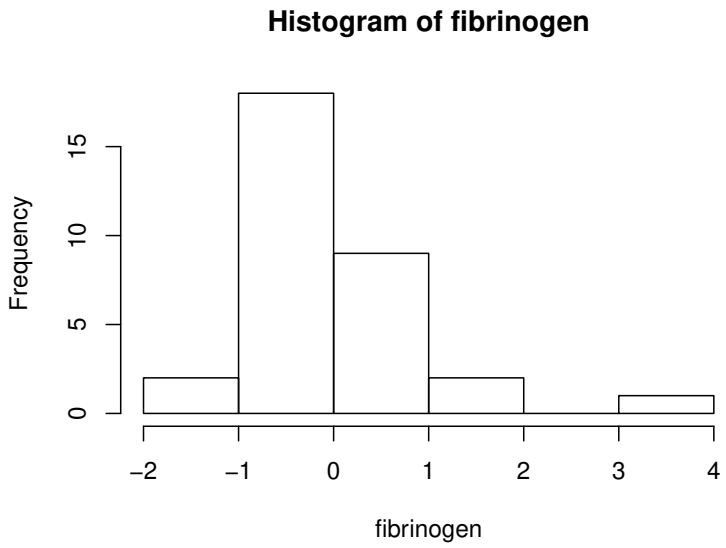

Fig. 3. histogram of fibrinogen