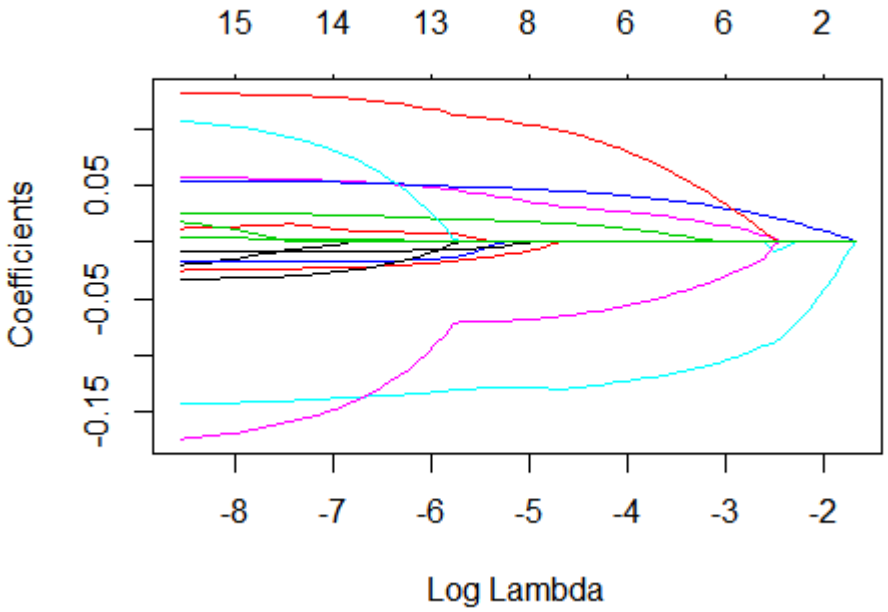
The codes & results

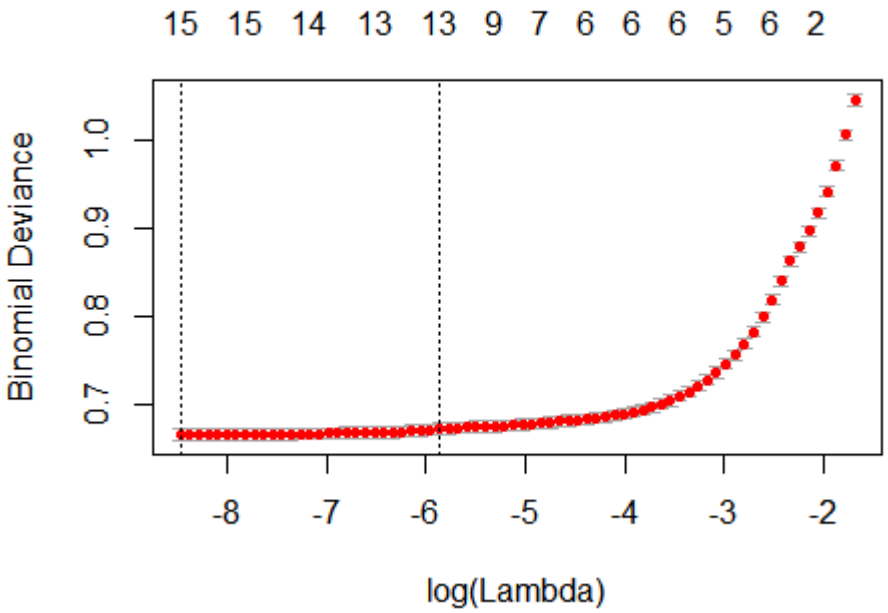
季莹莹，15220162202134

```
> ## Prepare the data
> library(ISLR)
> library(glmnet)
> library(caret)
> rm(list = ls())
> set.seed(123)
>
> weather <- read.csv("weather.csv")
> weather$RainTomorrow <- factor(weather$RainTomorrow,levels=c(0,1),labels=c("FALSE","TRUE"))
> summary(weather)
    MinTemp          MaxTemp          Rainfall        Evaporation        Sunshine
 Min.   :-6.70   Min.   : 4.10   Min.   :  0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 8.40   1st Qu.:18.60   1st Qu.:  0.000   1st Qu.: 2.800   1st Qu.: 5.000
 Median :13.10   Median :23.80   Median :  0.000   Median : 4.800   Median : 8.600
 Mean   :13.34   Mean   :24.13   Mean   :  2.117   Mean   : 5.445   Mean   : 7.699
 3rd Qu.:18.30   3rd Qu.:29.60   3rd Qu.:  0.600   3rd Qu.: 7.400   3rd Qu.:10.700
 Max.   :31.40   Max.   :48.10   Max.   :206.200   Max.   :81.200   Max.   :14.500
 WindGustSpeed     WindSpeed9am    WindSpeed3pm    Humidity9am      Humidity3pm
 Min.   :  9.00   Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   :  0.0
 1st Qu.: 31.00   1st Qu.: 9.00   1st Qu.:13.00   1st Qu.: 55.00   1st Qu.: 36.0
 Median : 39.00   Median :15.00   Median :19.00   Median : 67.00   Median : 51.0
 Mean   : 40.56   Mean   :15.24   Mean   :19.58   Mean   : 66.22   Mean   : 49.7
 3rd Qu.: 48.00   3rd Qu.:20.00   3rd Qu.:24.00   3rd Qu.: 80.00   3rd Qu.: 63.0
 Max.   :124.00   Max.   :67.00   Max.   :76.00   Max.   :100.00   Max.   :100.0
  Pressure9am      Pressure3pm       Cloud9am        Cloud3pm        Temp9am          Temp3pm
 Min.   : 980.5   Min.   : 977.1   Min.   :0.000   Min.   :0.000   Min.   :-0.90   Min.   : 3.70
 1st Qu.:1012.7   1st Qu.:1010.1   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:12.90   1st Qu.:17.30
 Median :1017.3   Median :1014.8   Median :5.000   Median :5.000   Median :17.70   Median :22.30
 Mean   :1017.3   Mean   :1014.9   Mean   :4.247   Mean   :4.328   Mean   :18.09   Mean   :22.63
 3rd Qu.:1022.0   3rd Qu.:1019.5   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:23.20   3rd Qu.:27.80
 Max.   :1040.4   Max.   :1038.9   Max.   :8.000   Max.   :9.000   Max.   :39.40   Max.   :46.10
 RainTomorrow
 FALSE:45361
 TRUE :12729
```

```
> ## Split data into training and test data sets
> n <- nrow(weather)
> train <- sample(n,n/2)
> weather.tr <- weather[train,]
> weather.te <- weather[-train,]
>
> ## Logistic regression on the full model
>
> # Fit on training data
> logitfit <- glm(RainTomorrow ~.,weather.tr,family='binomial')
> summary(logitfit)

Call:
glm(formula = RainTomorrow ~ ., family = "binomial", data = weather.tr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2016  -0.5106  -0.2802  -0.1239   3.1784

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   60.458777   3.414477  17.707  < 2e-16 ***
MinTemp       -0.028145   0.010499  -2.681  0.00735 **
MaxTemp        0.014853   0.016522   0.899  0.36867
Rainfall       0.026064   0.002773   9.400  < 2e-16 ***
Evaporation   -0.016305   0.008183  -1.993  0.04631 *
Sunshine      -0.143970   0.008470 -16.997  < 2e-16 ***
WindGustSpeed  0.057950   0.002178  26.606  < 2e-16 ***
WindSpeed9am  -0.007068   0.002865  -2.467  0.01362 *
WindSpeed3pm  -0.026008   0.003056  -8.510  < 2e-16 ***
Humidity9am    0.005659   0.002182   2.594  0.00949 **
Humidity3pm    0.053407   0.002310  23.117  < 2e-16 ***
Pressure9am    0.117159   0.011110  10.545  < 2e-16 ***
Pressure3pm   -0.183281   0.011214 -16.345  < 2e-16 ***
Cloud9am      -0.033416   0.010427  -3.205  0.00135 **
Cloud3pm       0.131259   0.011434  11.480  < 2e-16 ***
Temp9am        0.031831   0.015460   2.059  0.03950 *
Temp3pm       -0.009444   0.018714  -0.505  0.61379
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30394  on 29044  degrees of freedom
Residual deviance: 19337  on 29028  degrees of freedom
AIC: 19371

Number of Fisher Scoring iterations: 6

>
> # Predict on test data
> p = predict(logitfit,weather.te,type="response")
> logitpred = as.factor(p > 0.5)
> table(logitpred,weather.te$RainTomorrow,dnn=c("predicted","true"))
         true
predicted FALSE  TRUE
    FALSE 21438  3020
    TRUE   1184  3403
> logiterr <- 1-mean(logitpred==weather.te$RainTomorrow) #misclassification error rate
> logiterr
[1] 0.1447409


> ## Lasso regression (variable selection)
>
> # Fit on training data
> x <- model.matrix(RainTomorrow ~.,weather.tr)[,-1] #no intercept
> y <- weather.tr$RainTomorrow
> lassofit.all <- glmnet(x,y,alpha=1,family="binomial")
> plot(lassofit.all,xvar="lambda")
```




```
> ## Cross validation (select a good value of lambda)
> cv.lasso <- cv.glmnet(x,y,alpha=1,family="binomial")
> plot(cv.lasso)
```



```
> # Refit the model using optimal lambda
> lambda.star <- cv.lasso$lambda.min
> lassofit.star <- glmnet(x,y,alpha=1,lambda=lambda.star,family="binomial")
> coef(lassofit.star)
17 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)   60.888882472
MinTemp       -0.021500267
MaxTemp        0.011799566
Rainfall       0.025529100
Evaporation   -0.016679747
Sunshine      -0.142546045
WindGustSpeed  0.057006871
WindSpeed9am  -0.007453108
WindSpeed3pm  -0.024969331
Humidity9am    0.004271247
Humidity3pm    0.054003484
Pressure9am    0.105753795
Pressure3pm   -0.172193570
Cloud9am      -0.032027381
Cloud3pm       0.130605021
Temp9am        0.019225615
Temp3pm        .
> # Predict on test data
> newx <- model.matrix(RainTomorrow ~.,weather.te)[,-1]
> lassopred <- predict(lassofit.star,newx,type="class")
> table(lassopred,weather.te$RainTomorrow,dnn=c("predicted","true"))
         true
predicted FALSE  TRUE
    FALSE 21446  3022
    TRUE   1176  3401
> lassoerr <- 1-mean(lassopred==weather.te$RainTomorrow)
> lassoerr
[1] 0.1445343
```