Logistic Regression Excercises

```
library(dplyr)
library(caret)
library(caTools)
library(pROC)
```

Problem 1: Predicting the Popularity of Music Records

Predict whether a song will rank in top ten. The data used are from 2009 and before 2009. The threshold probability for the logistic regression model is 0.45.

```
songs = read.csv("songs.csv")
songs = songs[songs$Year <= 2009,][,c(-2, -3, -4, -5)]
```

Use stratified sampling to split the data into a training set and a test set. The test size is 0.2.

```
set.seed(144)
split = createDataPartition(songs$Top10, p = 0.8, list = FALSE)
songs_train = songs[split,]
songs_test = songs[-split,]
```

Build the logistic regression model.

```
logreg = glm(Top10 ~ ., data=songs_train, family="binomial")
summary(logreg)
```

```
##
## Call:
## glm(formula = Top10 ~ ., family = "binomial", data = songs_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0402  -0.5339  -0.3359  -0.1775   3.2188
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.314e+02  1.651e+01   7.963 1.68e-15 ***
## Year                      -5.809e-02  8.118e-03  -7.156 8.31e-13 ***
## timesignature              1.137e-01  9.746e-02   1.167 0.243177
## timesignature_confidence   8.195e-01  2.211e-01   3.707 0.000210 ***
## loudness                   3.265e-01  3.309e-02   9.865  < 2e-16 ***
## tempo                      5.985e-04  1.896e-03   0.316 0.752315
## tempo_confidence           3.713e-01  1.610e-01   2.307 0.021059 *
## key                        1.806e-02  1.175e-02   1.537 0.124244
## key_confidence             2.800e-01  1.588e-01   1.764 0.077786 .
## energy                    -1.651e+00  3.485e-01  -4.738 2.16e-06 ***
## pitch                     -4.922e+01  7.871e+00  -6.253 4.02e-10 ***
## timbre_0_min               2.738e-02  4.823e-03   5.678 1.36e-08 ***
## timbre_0_max              -3.356e-01  2.884e-02 -11.637  < 2e-16 ***
```

```
## timbre_1_min              5.254e-03  8.846e-04   5.939 2.86e-09 ***
## timbre_1_max             -5.756e-04  8.258e-04  -0.697 0.485753
## timbre_2_min             -1.003e-03  1.285e-03  -0.781 0.434950
## timbre_2_max              1.278e-03  1.038e-03   1.231 0.218379
## timbre_3_min              6.194e-04  6.776e-04   0.914 0.360703
## timbre_3_max             -2.383e-03  6.530e-04  -3.649 0.000264 ***
## timbre_4_min              1.207e-02  2.240e-03   5.388 7.14e-08 ***
## timbre_4_max              5.886e-03  1.754e-03   3.355 0.000794 ***
## timbre_5_min             -4.999e-03  1.427e-03  -3.502 0.000462 ***
## timbre_5_max             -3.081e-04  9.053e-04  -0.340 0.733589
## timbre_6_min             -1.673e-02  2.567e-03  -6.516 7.24e-11 ***
## timbre_6_max              3.640e-03  2.450e-03   1.485 0.137432
## timbre_7_min             -5.260e-03  2.015e-03  -2.610 0.009047 **
## timbre_7_max             -4.171e-03  2.091e-03  -1.995 0.046052 *
## timbre_8_min              5.126e-03  3.201e-03   1.601 0.109269
## timbre_8_max              3.435e-03  3.379e-03   1.017 0.309350
## timbre_9_min             -2.308e-03  3.357e-03  -0.687 0.491809
## timbre_9_max              4.865e-03  2.740e-03   1.775 0.075842 .
## timbre_10_min             3.251e-03  2.074e-03   1.568 0.116996
## timbre_10_max             3.197e-03  2.045e-03   1.563 0.118023
## timbre_11_min            -2.674e-02  4.136e-03  -6.467 9.98e-11 ***
## timbre_11_max             1.936e-02  3.861e-03   5.013 5.37e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4785.9  on 5760  degrees of freedom
## Residual deviance: 3750.4  on 5726  degrees of freedom
## AIC: 3820.4
##
## Number of Fisher Scoring iterations: 6
```

The confusion matrix of the prediction is

```
pred_prop = predict(logreg, songs_test, type="response")
predicted_values = ifelse(pred_prop > 0.45, 1, 0)
actual_values = songs_test$Top10
table(predicted_values, actual_values) # confusion matrix
```

```
##                 actual_values
## predicted_values    0    1
##                0 1171  158
##                1   49   62
```

Therefore,

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{62+1171}{62+49+1171+158} = 0.8563\ ,$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN} = \frac{62}{62+158} = 0.2818\ ,$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN} = \frac{49}{49+1171} = 0.0402\ .$$

Problem 2: Framingham Heart Study

Read the data and split them randomly into a training set and a test set. The test size is 0.25.

```
data = read.csv("framingham.csv")
data$TenYearCHD = factor(data$TenYearCHD)
data$male = factor(data$male)
data$currentSmoker = factor(data$currentSmoker)
data$BPMeds = factor(data$BPMeds)
data$prevalentStroke = factor(data$prevalentStroke)
data$prevalentHyp = factor(data$prevalentHyp)
data$diabetes = factor(data$diabetes)
data[data$education=='Some college/vocational school',]$education =
  'College/Vocational' # display the model results in a neat way
set.seed(38)
N = nrow(data)
idx = sample.split(data$TenYearCHD, 0.75)
train = data[idx,]
test = data[!idx,]
```

a. Calculate Expected Costs and the Value of p

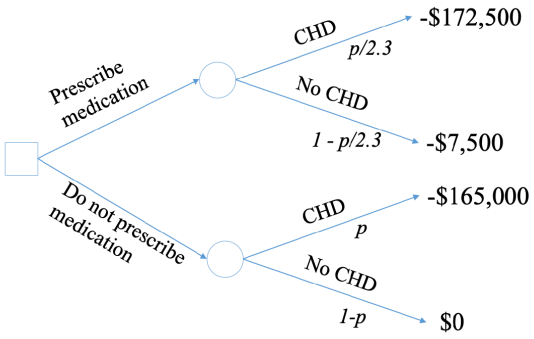

Figure 1: Decision tree for prescribing the approved medication to prevent CHD.

The expected cost borne by a patient who does not take the preventive medication:

$$\mathbb{E}\left[\text{cost without medication}\right] = p \times 165000 + (1-p) \times 0 = 165000p$$

The expected cost borne by a patient who takes the preventive medication:

$$\mathbb{E}\left[\text{cost with medication}\right] = p/2.3 \times 172500 + (1 - p/2.3) \times 7500 = 7500 + 71739p$$

To solve the value of p when we would recommend the medication, we write

$$\mathbb{E}\left[\text{cost without medication}\right] = \mathbb{E}\left[\text{cost with medication}\right]$$

```
p = 7500 / (165000 - (172500/2.3 - 7500/2.3))
print(paste("The value of p is", round(p, 4)), quote=F)
```

```
## [1] The value of p is 0.0804
```

This equation gives $p = 0.08$. Therefore, when $p \geq 0.08$, we would recommend the medication.

b. Logistic Regression: First Fit

We use all the independent variables in the dataset to construct a logistic regression model, which predicts the probability that a patient will experience CHD within the next 10 years.

```
fit1 = glm(TenYearCHD ~ ., data=train, family="binomial")
summary(fit1)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8784  -0.5960  -0.4270  -0.2837   2.8314
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -8.200424   0.812233 -10.096  < 2e-16 ***
## male1                         0.509274   0.126005   4.042 5.31e-05 ***
## age                           0.062535   0.007742   8.077 6.61e-16 ***
## educationCollege/Vocational -0.087740   0.228471  -0.384  0.70095
## educationHigh school/GED    -0.020900   0.204758  -0.102  0.91870
## educationSome high school    0.081260   0.190941   0.426  0.67042
## currentSmoker1               0.028271   0.179325   0.158  0.87473
## cigsPerDay                   0.015663   0.007192   2.178  0.02942 *
## BPMeds1                      0.050114   0.279362   0.179  0.85763
## prevalentStroke1             0.836722   0.589452   1.419  0.15576
## prevalentHyp1                0.289766   0.157951   1.835  0.06658 .
## diabetes1                    0.229506   0.348199   0.659  0.50982
## totChol                      0.003259   0.001281   2.544  0.01095 *
## sysBP                        0.014240   0.004496   3.168  0.00154 **
## diaBP                       -0.004888   0.007547  -0.648  0.51721
## BMI                          0.006547   0.014829   0.442  0.65883
## heartRate                   -0.004047   0.004905  -0.825  0.40932
## glucose                      0.006202   0.002496   2.484  0.01298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2341.9  on 2743  degrees of freedom
## Residual deviance: 2074.6  on 2726  degrees of freedom
## AIC: 2110.6
##
## Number of Fisher Scoring iterations: 5
```

According to the model `fit1`, the most important risk factors for 10-year CHD are the statistically significant variables at the 95% level. They are `male`, `age`, `cigsPerDay`, `totChol`, `sysBP`, and `glucose`. These make intuitive clinical sense as male, elder people, people who smokes more cigarettes, and people with high cholesterol, high blood pressure and high blood sugar are more likely to experience CHD.

c. Predict Whether a Certain Patient Will Experience CHD

The patient is a 55-year old college-educated male, smokes 10 cigarettes per day, is not on blood pressure medication, has not had a stroke, has hypertension, has not been diagnosed with diabetes, has a Cholesterol

of 220, as a systolic blood pressure of 140 and a diastolic blood pressure of 100, has a BMI of 30, has a heart rate of 60, and has a glucose level 80.

We predict the probability that this patient will experience CHD in the next ten years

```
new_patient = data.frame(
  male = as.factor(1),
  age = 55,
  education = 'College',
  currentSmoker = as.factor(1),
  cigsPerDay = 10,
  BPMeds = as.factor(0),
  prevalentStroke = as.factor(0),
  prevalentHyp = as.factor(1),
  diabetes = as.factor(0),
  totChol = 220,
  sysBP = 140,
  diaBP = 100,
  BMI = 30,
  heartRate = 60,
  glucose = 80
)
pred_prop_new_patient = predict(fit1, new_patient, type="response")
print(paste("The predicted probability is",round(pred_prop_new_patient,4)), quote=F)
```

```
## [1] The predicted probability is 0.2487
```

Since $0.2487 > p = 0.08$, the physician should prescribe the medication.

d. Other Interventions for the Patient

From Question b., we know that variables `male`, `age`, `cigsPerDay`, `totChol`, `sysBP`, and `glucose` are important risk factors for 10-year CHD. Other variables are not easy to change or cannot be changed, but this patient can smoke less to decrease the risk of experiencing CHD in the next ten years. Therefore, we try three levels of the variable `cigsPerDay`, and then we predict the probability that the patient will experience CHD in the next ten years. The three levels are `cigsPerDay`=5, `cigsPerDay`=2, and {`currentSmoker`=0, `cigsPerDay`=0} . Note `cigsPerDay`=0 makes `currentSmoker`=0.

```
# level1: cigsPerDay = 5
new_patient_cig1 = new_patient
new_patient_cig1$cigsPerDay = 5
# level1: cigsPerDay = 2
new_patient_cig2 = new_patient
new_patient_cig2$cigsPerDay = 2
# level3: currentSmoker = 0, cigsPerDay = 0
new_patient_cig3 = new_patient
new_patient_cig3$currentSmoker = as.factor(0)
new_patient_cig3$cigsPerDay = 0
# predict the probability that the patient will experience CHD
prob1 = predict(fit1, new_patient_cig1, type="response")
prob2 = predict(fit1, new_patient_cig2, type="response")
prob3 = predict(fit1, new_patient_cig3, type="response")
data.frame(cigsPerDay10=pred_prop_new_patient, cigsPerDay5=prob1,
           cigsPerDay2=prob2, cigsPerDay0=prob3, row.names=c("Probability"))
```

```
##             cigsPerDay10 cigsPerDay5 cigsPerDay2 cigsPerDay0
## Probability     0.248699   0.2343559   0.2260302   0.2157748
```

If this patient stops smoking, his excessive `totChol` and `sysBP` levels will decrease to normal (his `glucose` level is normal). Let us assume `totChol`=180 and `sysBP`=120. Then the predicted probability that this patient will experience CHD in the next ten years is

```
new_patient_cig3_improve = new_patient_cig3
new_patient_cig3_improve$totChol = 180
new_patient_cig3_improve$sysBP = 120
prob4 = predict(fit1, new_patient_cig3_improve, type="response")
print(paste("The predicted probability is", round(prob4, 4)), quote=F)
```

```
## [1] The predicted probability is 0.1537
```

We can see that the predicted probability decreases from 0.2487 to 0.1537, nearly 10%! Therefore, in addition to prescribing the medication, the physician should suggest this patient stop smoking, which can greatly reduce the patient's risk of experiencing CHD.

e. Prediction on the Test Set & Confusion Matrix

We use the threshold determined in Question a., which is $p = 0.08$, to build prediction on the test set. Then we compute its confusion matrix, accuracy, True Positive Rate, and False Positive Rate.

```
CHD_pred_prop = predict(fit1, test, type="response")
CHD_predicted_values = ifelse(CHD_pred_prop > p, 1, 0)
CHD_actual_values = test$TenYearCHD
table(CHD_predicted_values, CHD_actual_values) # confusion matrix
```

```
##                      CHD_actual_values
## CHD_predicted_values   0   1
##                    0 287  15
##                    1 488 124
```

Therefore,

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{124 + 287}{124 + 488 + 287 + 15} = 0.4497 \ ,$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} = \frac{124}{124 + 15} = 0.8921 \ ,$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} = \frac{488}{488 + 287} = 0.6297 \ .$$

Finally, we calculate the expected economic cost for all patients in the test set if patients are prescribed the medication using the strategy implied by the model.

$$\begin{aligned} \mathbb{E}\left[\text{total cost}\right] &= 172500 \times TP/2.3 + 7500 \times FP + 165000 \times FN \\ &= 172500 \times 124/2.3 + 7500 \times 488 + 165000 \times 15 \\ &= 15435000 \end{aligned}$$

The expected economic cost for all patients in the test set is $15,435,000.

f. The Expected Economic Costs in the Baseline and Ideal Models

We consider a "baseline" model and an "ideal" model. The baseline model reflects the current practice, where the medication is not prescribed to any patient. In the ideal model, medication is only prescribed to patients that would otherwise develop CHD, assuming perfect *ex post* information on the test set. For each of these models, we compute the expected economic cost for all patients in the test set.

$$\mathbb{E}[\text{baseline cost}] = 165000 \times (TP + FN) = 165000 \times (124 + 15) = 22935000 \ ,$$
$$\mathbb{E}[\text{ideal cost}] = 172500 \times (TP + FN)/2.3 = 172500 \times (124 + 15)/2.3 = 10425000 \ .$$

The baseline cost is \$22,935,000, and the ideal cost is \$10,425,000. We can see that if we use logistic regression to predict whether a person will experience CHD in the next ten years and prescribe the medication according to the predictions, we can reduce the expected economic cost for all patients in the test set from \$22,935,000 to \$15,435,000, which is 32.7%. These results indicate that it is necessary and significant to do the predictions and the prescriptions according to the predictions.

g. ROC Curve and AUC Score

We construct the ROC curve and compute the AUC score for the logistic regression model (`fit1`) on the test set. Sensitivity = TPR, Specificity = 1 - FPR.

```
par(pty="s")
CHD_roc = roc(response=CHD_actual_values, predictor=CHD_pred_prop)
plot.roc(CHD_roc, print.auc=T, auc.polygon=T, auc.polygon.col='linen', print.thres=T)
plot.roc(smooth(CHD_roc), add=TRUE, col='firebrick')
```

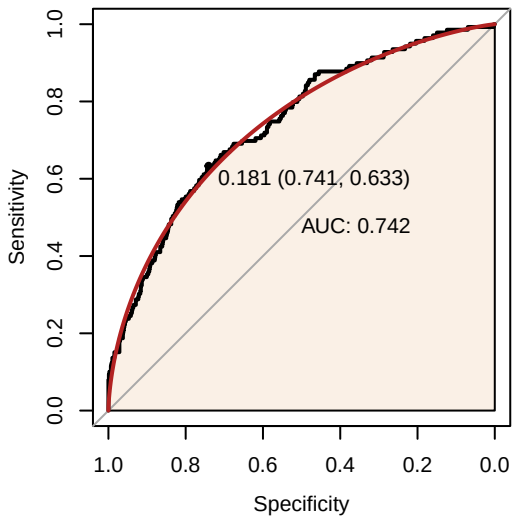

Figure 2: ROC curve and AUC score.

If decision-makers look to further study possible medications for preventing CHD, they can use the ROC curve to choose the prediction probability that achieves high TPR (high sensitivity) and low FPR (high specificity). They can use the prediction probability to decide whether a medication should be prescribed to patients. Since the condition that a patient predicted to not experience CHD actually experiences CHD has serious consequences, decision-makers should put emphasis on increasing the TPR.

h. Rebuild the Logistic Regression Model

We choose three risk factors to fit a logistic regression model on the training data. They are `male`, `age`, and `cigsPerDay`.

```
train_subset = train[,c(1, 2, 5, 16)]
test_subset = test[,c(1, 2, 5, 16)]
```

```r
fit2 = glm(TenYearCHD ~ ., data=train_subset, family="binomial")
summary(fit2)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = "binomial", data = train_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1731  -0.6083  -0.4508  -0.3285   2.5987
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.378551   0.383420 -16.636  < 2e-16 ***
## male1        0.406601   0.117465   3.461 0.000537 ***
## age          0.083941   0.006833  12.285  < 2e-16 ***
## cigsPerDay   0.014740   0.004766   3.093 0.001982 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2341.9  on 2743  degrees of freedom
## Residual deviance: 2156.4  on 2740  degrees of freedom
## AIC: 2164.4
##
## Number of Fisher Scoring iterations: 5
```

We evaluate the model performance on the test data.

```r
CHD_pred_prop2 = predict(fit2, test_subset, type="response")
CHD_predicted_values2 = ifelse(CHD_pred_prop2 > p, 1, 0)
CHD_predicted_values2 = as.factor(CHD_predicted_values2)
confusionMatrix(data=CHD_predicted_values2, reference=CHD_actual_values, positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 225  14
##          1 550 125
##
##                Accuracy : 0.3829
##                  95% CI : (0.3513, 0.4153)
##     No Information Rate : 0.8479
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0734
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8993
##             Specificity : 0.2903
##          Pos Pred Value : 0.1852
##          Neg Pred Value : 0.9414
```

```
##              Prevalence : 0.1521
##          Detection Rate : 0.1368
##    Detection Prevalence : 0.7385
##       Balanced Accuracy : 0.5948
##
##        'Positive' Class : 1
##
```

The results show that accuracy = 0.3829, TPR = sensitivity = 0.8993, and FPR = 1 - specificity = 0.7097. From the comparison between `fit1` and `fit2`, we see that the simplified logistic regression model does not perform much worse than the full model.

	Accuracy	True Positive Rate	False Positive Rate
`fit1`	0.4497	0.8921	0.6297
`fit2`	0.3829	0.8993	0.7097

i. Some Ethical Concerns of the Analysis

Finally, we discuss some ethical concerns of the analysis. Issues of privacy and informed consent may be the ethical concerns of the analysis. The researchers should protect the participants' privacy when using their demographic information and physical data in the study. The researchers should inform the participants how they will use the data in the study and whether they will publish the data. They should also get consent from the participants to let them to use the data.