Data 603:Statistical Modelling with Data

Logistic Regression

Part II :Introduction to the Multiple Logistic Regression Model

Before discussing about the Multiple Logistic Regression, let's discover a Logistic Regression with a qualitative variable (more than 2 levels).

Example: The German Credit Data contains data on 6 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles. The independent variables are listed below

Creditability= (1 if good credit, 0 if bad credit)

Balance=Account Balance (Categorical variable with 4 levels)

$$Balance = \begin{cases} 1 \text{ if balance is more than } 5000 \\ 2 \text{ if balance is } 3001 - 5000 \\ 3 \text{ if balance is } 1001 - 3000 \\ 4 \text{ if balance is less than } 1000 \end{cases}$$

Duration= Duration of credit in months (months)

Employment=Length of current employment (years)

Amount=Credit amount (dollars)

Age=Age (year)

Is there any sufficient evidence to indicate that Balance is an important predictor for predicting the Creditability?

```
library("readxl")
creditdata <- read_excel("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/creditbi
lity.xlsx")
mylogit<-glm(Creditability~factor(Balance),data=creditdata,family="binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = Creditability ~ factor(Balance), family = "binomial",
##     data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0725  -1.1898   0.4983   0.9948   1.1650
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.0292     0.1208   0.242 0.809059
## factor(Balance)2   0.4167     0.1739   2.397 0.016531 *
## factor(Balance)3   1.2236     0.3262   3.750 0.000177 ***
## factor(Balance)4   1.9944     0.1980  10.071  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1090.4  on 996  degrees of freedom
## AIC: 1098.4
##
## Number of Fisher Scoring iterations: 4
```

The dummy coding with 4 levels for Balance is provided

Dummy coding with 4 levels

	Balance 2	Balance 3	Balance 4
Balance1	0	0	0
Balance2	1	0	0
Balance3	0	1	0
Balance4	0	0	1

From the output, the logit model is

$$\hat{logit} = 0.0292 + 0.4167X_1 + 1.2236X_2 + 1.9944X_3$$

The logistic Regression model is

$$E(y) = P(y = 1|X) = \frac{e^{0.0292+0.4167X_1+1.2236X_2+1.9944X_3}}{1 + e^{0.0292+0.4167X_1+1.2236X_2+1.9944X_3}}$$

We can predict the Creditability for each Balance group as following

$$\hat{\pi} = \begin{cases} = \frac{e^{0.0292}}{1+e^{0.0292}} \text{ if balance is more than } 5000 \\ = \frac{e^{0.0292+0.4167}}{1+e^{0.0292+0.4167}} \text{ if balance is } 3001 - 5000 \\ = \frac{e^{0.0292+1.2236}}{1+e^{0.0292+1.2236}} \text{ if balance is } 1001 - 3000 \\ = \frac{e^{0.0292+1.9944}}{1+e^{0.0292+1.9944}} \text{ if balance is less than } 1000 \end{cases}$$

FAQ!

If one of the parts of the dummies is not significant, what should I do?

You have to use all dummies. If you exclude one dummy it will change the interpretation of your effects.

Multiple Logistic Regression for a Binary Dependent Variable

We now consider the problem of predicting a binary response using multiple predictors. By analogy with the extension from simple to multiple linear regression, we can also extend a simple logistic regression to a multiple logistic regression as follows:

Binary Logistic Regression Model

$$E(y) = P(y = 1|X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}$$

$$\begin{aligned} where & \\ y &= 1 \text{ if category A occurs} \\ &= 0 \text{ if category B occurs} \\ E(y) &= P(\text{Category A occurs}) = \pi \\ X &= (X_1, X_2, \ldots X_p) \text{ are p predictors} \end{aligned}$$

For example; for the Default data, the logistic regresssion model that uses balance and income to predict probability of default can be written as

```
library(ISLR) # for Default data Set
summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

```
mylogit <- glm(default ~ balance+income, data = Default, family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = default ~ balance + income, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

```
#comparing with the simple logistic model
mysimplelogit <- glm(default ~ balance, data = Default, family = "binomial")
summary(mysimplelogit)
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

The output uses the maximum likelihood method to estimate $\beta_0, \beta_1, \beta_2$, shows the coefficient estimates for a logistic regression model that uses balance and income (in thousands of dollars) to predict probability of default. From the output the logistic regression model is

$$\hat{\pi} = P(y = 1|X) = \frac{e^{-10.154+0.005647X_1+0.00002081X_2}}{1 + e^{-10.154+0.005647X_1+0.00002081X_2}}$$

$$\hat{\text{logit}} = -10.154 + 0.005647X_1 + 0.00002081X_2$$

Interpretations of Multiple Logistic Regression Coefficients

In general, the coeffcient $\hat{\beta}_1$ and $\hat{\beta}_2$ in the logistic model estimates the change in the log-odds (same concept as linear regression). For example, from the default output,

$\hat{\beta}_2$ = 0.00002081. This indicates that an increase in income is associated with an increase in the probability of default when balance is held constant.

To be precise,

a one-unit **increase** in income is associated with an **increase** in the log odds of default by 0.00002081 units when balance is held constant. OR

For every $1 increase in income, we estimate the log odds of a default increase by 0.00002081 when balance is held constant. What does it really mean??

By computing $e^{\hat{\beta}_2}$ (antilog of the coefficient) , so we interpret in terms of the odds ratio.

$$\hat{\beta}_1 = 0.00002081$$
$$e^{\hat{\beta}_1} = e^{0.00002081} = 1.00002$$

For every 1 dollar increases in balance, we estimate **the odds** of a default increases by about 0.002%.

Note! R function to calculate the antilog for β_i

```
library(ISLR) # for Default data Set
mylogit <- glm(default ~ balance+income, data = Default, family = "binomial")
sum.coef<-summary(mylogit)$coef
est<-exp(sum.coef[,1])
print(est)
```

```
##  (Intercept)      balance       income
## 9.728329e-06 1.005663e+00 1.000021e+00
```

For Example, a survey that asked students in their final year of a high school near Dayton Ohio, whether they had ever used marijuana. The explanatory variables are gender and race.

		Marijuana Use	
Race	**Gender**	**Yes**	**No**
White	Female	420	620
	Male	483	579
Other	Female	25	55
	Male	32	62

Source: Alan Agresti (2019),An Introduction to Categorical data Analysis, 3rd edition, New York: Wiley.

Use of marijuana by gender and race

```
marijuana=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/Marijuana.tx
t",header = TRUE,sep="\t")
marijuana
```

race	gender	yes	no
<fctr>	<fctr>	<int>	<int>

race	gender	yes	no
<fctr>	<fctr>	<int>	<int>
white	female	420	620
white	male	483	579
other	female	25	55
other	male	32	62

4 rows

```
# Option 1 to fit logitic using a contingency table
model1<-glm(yes/(yes+no)~gender+race,weights=yes+no,family = "binomial",data=marijuana)
summary(model1)  # for Wald Z test
```

```
##
## Call:
## glm(formula = yes/(yes + no) ~ gender + race, family = "binomial",
##     data = marijuana, weights = yes + no)
##
## Deviance Residuals:
##        1         2         3         4
## -0.04513   0.04402   0.17321  -0.15493
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.83035    0.16854  -4.927 8.37e-07 ***
## gendermale   0.20261    0.08519   2.378  0.01739 *
## racewhite    0.44374    0.16766   2.647  0.00813 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12.752784  on 3  degrees of freedom
## Residual deviance:  0.057982  on 1  degrees of freedom
## AIC: 30.414
##
## Number of Fisher Scoring iterations: 3
```

```
# Option 2 to fit logitic using a contingency table
model2<-glm(as.matrix(marijuana[,3:4])~gender+race,family = "binomial",data=marijuana)
# The matrix response sould not be a data frame in R
summary(model2)  # for Wald Z test
```

```
##
## Call:
## glm(formula = as.matrix(marijuana[, 3:4]) ~ gender + race, family = "binomial",
##     data = marijuana)
##
## Deviance Residuals:
##        1         2         3         4
## -0.04513   0.04402   0.17321  -0.15493
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.83035    0.16854  -4.927 8.37e-07 ***
## gendermale   0.20261    0.08519   2.378  0.01739 *
## racewhite    0.44374    0.16766   2.647  0.00813 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12.752784  on 3  degrees of freedom
## Residual deviance:  0.057982  on 1  degrees of freedom
## AIC: 30.414
##
## Number of Fisher Scoring iterations: 3
```

```
#
```

Testing a relationship Between the Response and predictors

Full Model Test

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \text{at least one } \beta_i \neq 0$$

The formula for Wald's test is complicated (in terms of matrix), so we will use R software to compute the p-value

For example; for the Default data, the logistic regresssion model that uses balance and income to predict probability of default. Test the full logistic regression model with 2 predictors

```
library(aod) # for Wald test
library(ISLR) # for Default data Set
mylogit <- glm(default ~ balance+income, data = Default, family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = default ~ balance + income, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

```
#The Wald chi square test for the full model
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 2:3)  #Terms tells R which terms in the model are to be tested.
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 617.9, df = 2, P(> X2) = 0.0
```

```
#p-value for the overall test using the likelihood-ratio test
with(mylogit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
```

```
## [1] 4.540906e-292
```

```
## Or the p-value is...
 pchisq(2920.6-1579.0, 9999-9997, lower.tail = FALSE)
```

```
## [1] 4.734363e-292
```

From the output, the Wald χ^2=617.9 with the pvalue =0.0 <0.05. Moreover, the likelihood-ratio test (see below for more details) confirms the the result with p-value =4.540906e-292 from the optional R funtion. This indicates that the probability of default depends on at least one predictor at $\alpha = 0.05$

Testing For The Significance of The Coefficients (Individual Test)

For maximum likelihood estimates, the Wald Z test or the Wald χ^2 test is the test of significance for individual regression coefficients in logistic regression .

$$H_0 : \beta_i = 0$$

The probability of Y does not depend on X

$$H_1 : \beta_i \neq 0$$

The probability of Y depends on X

$$Zcal = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim Normal(0, 1)$$

$$or$$

$$\chi^2 = (\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)})^2 \sim \chi^2_1$$

we reject the null hypothesis when $|Z_{cal}| > Z_{\alpha/2}$ or p-value$<\alpha$

Likelihood Ratio Test

The test statistic for testing the overall/subset adequacy of the logistic model is called **the Likelihood Ratio Test** Suppose two alternative models are under consideration, one model is simpler or more parsimonious than the other. This is performed using the likelihood ratio test, which compares the likelihood of the data under the full model against the likelihood of the data under a model with fewer predictors. **Removing predictor variables from a model will almost always make the model fit less well** (i.e. a model will have a lower log likelihood), but it is necessary to test whether the observed difference in model fit is statistically significant.

$$H_0 : \beta_{r+1} = \beta_{r+2} \ldots = \beta_p = 0$$
reduced model is true

$$H_1 : \text{at least one } \beta_i \neq 0$$
larger model is true

The likelihood ratio statistic is

$$\triangle G^2 = -2logL \text{ from the reduced model} - (-2logL \text{ from larger model})$$

The p-value is $p(\chi^2 > \triangle G^2)$

Larger values of $\triangle G^2$ ("-2 Log L") lead to small p-values, which provide evidence against the reduced model in favor of the larger model;

In this class we will not calculate the likelihood ratio statistic, we use R function for testing

```
library(ISLR) # for Default data Set
#logit<-glm(default~1, data = Default, family = "binomial")
simplelogit <- glm(default ~ balance, data = Default, family = "binomial")
multiplelogit <- glm(default ~ balance+income, data = Default, family = "binomial")
#anova(logit,multiplelogit,test="Chisq")
anova(simplelogit,multiplelogit,test="Chisq")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	9998	1596.452	*NA*	*NA*	*NA*
2	9997	1578.966	1	17.48541	2.895205e-05

2 rows

```
library(lmtest)# for lrtest() function
#Another option to perform the Likelihood Ratio Test
lrtest(simplelogit,multiplelogit)
```

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2	-798.2258	*NA*	*NA*	*NA*
2	3	-789.4831	1	17.48541	2.895205e-05

2 rows

R functions anova(model1,model2,test="Chisq"): performs the likelihood ratio test

lrtest(model1,model2):performs the likelihood ratio test

From the output, it can be seen that $\triangle G^2$ =17.485 with p-value =2.895e-05 < 0.05, so we reject Ho, therefore the larger model performs better than the reduced model at $\alpha = 0.05$. The probability of default depends on both both income and balance predictors.

Model fit in Multiple Logistic Regression

Similarly to Simple Logistic Regression, we look for Deviance, AIC, and ROC with AUC (area under ROC curve) to tell us how well the model predicts the probability of Y. We can compare the performance of logistic models by using AIC and AUC values.

For example, we would like to compare 2 models for predicting the probability of Default by using only balance, and with both balance and income.

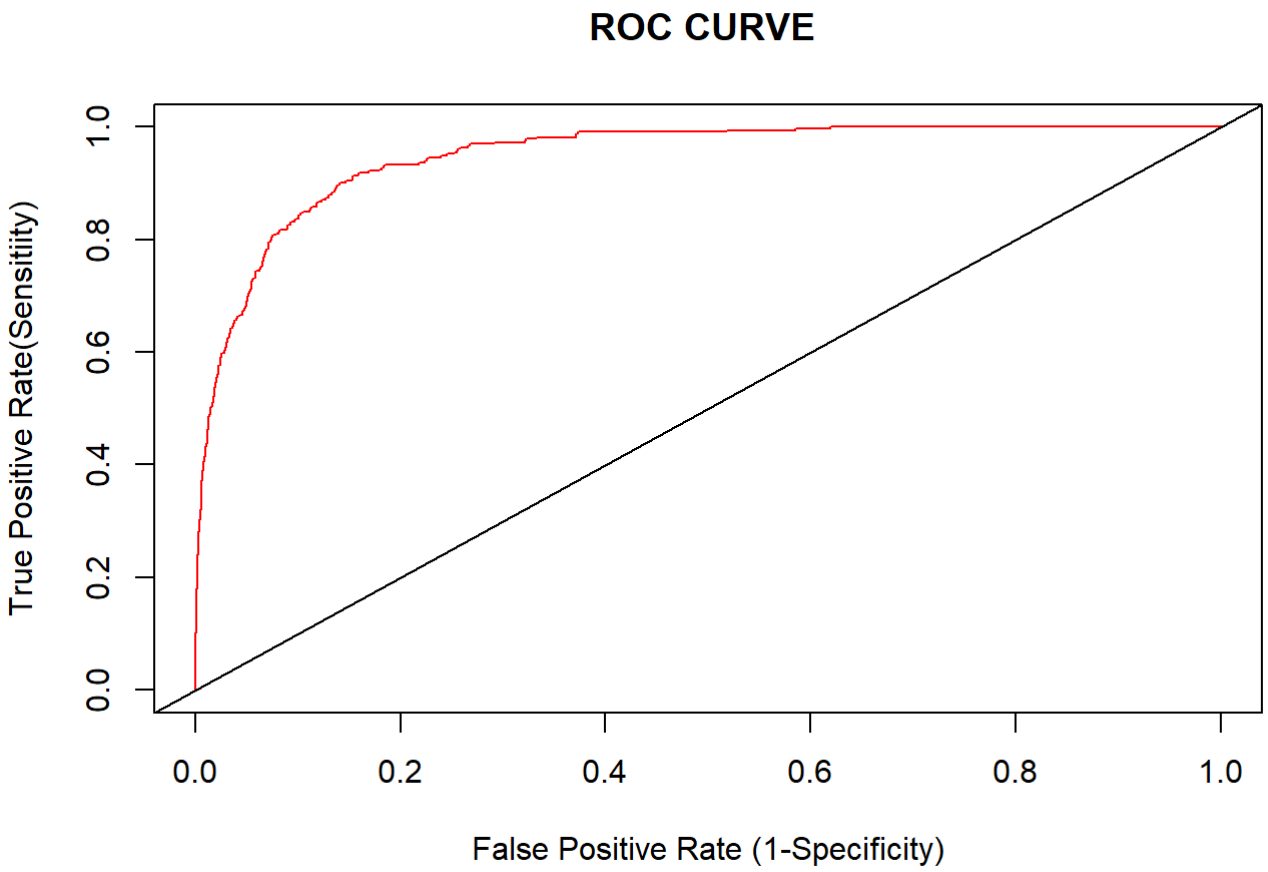
```
library(ISLR) # for Default data Set
library(ROCR) # for ROC
library(pROC) # for AUC
mymullogit <- glm(default ~ balance+income, data = Default, family = "binomial")
summary(mymullogit)
```

```
##
## Call:
## glm(formula = default ~ balance + income, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4725  -0.1444  -0.0574  -0.0211   3.7245
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

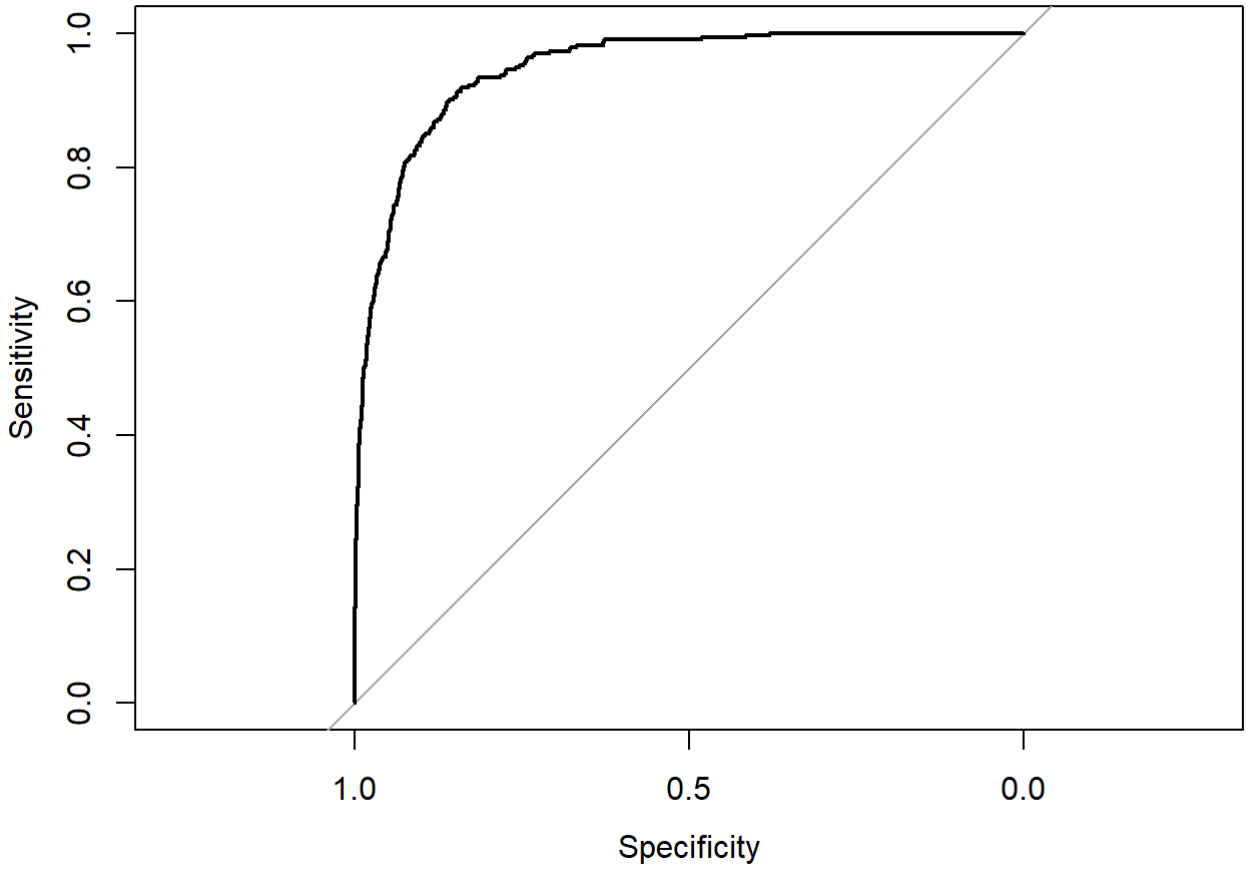
```
mylogit <- glm(default ~ balance, data = Default, family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

```
# ROC&AUC for default = balance
#--------ROC Curve-----------
prob1=predict(mylogit,type=c("response"))
# Option 1
pred<-prediction(prob1,Default$default)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```



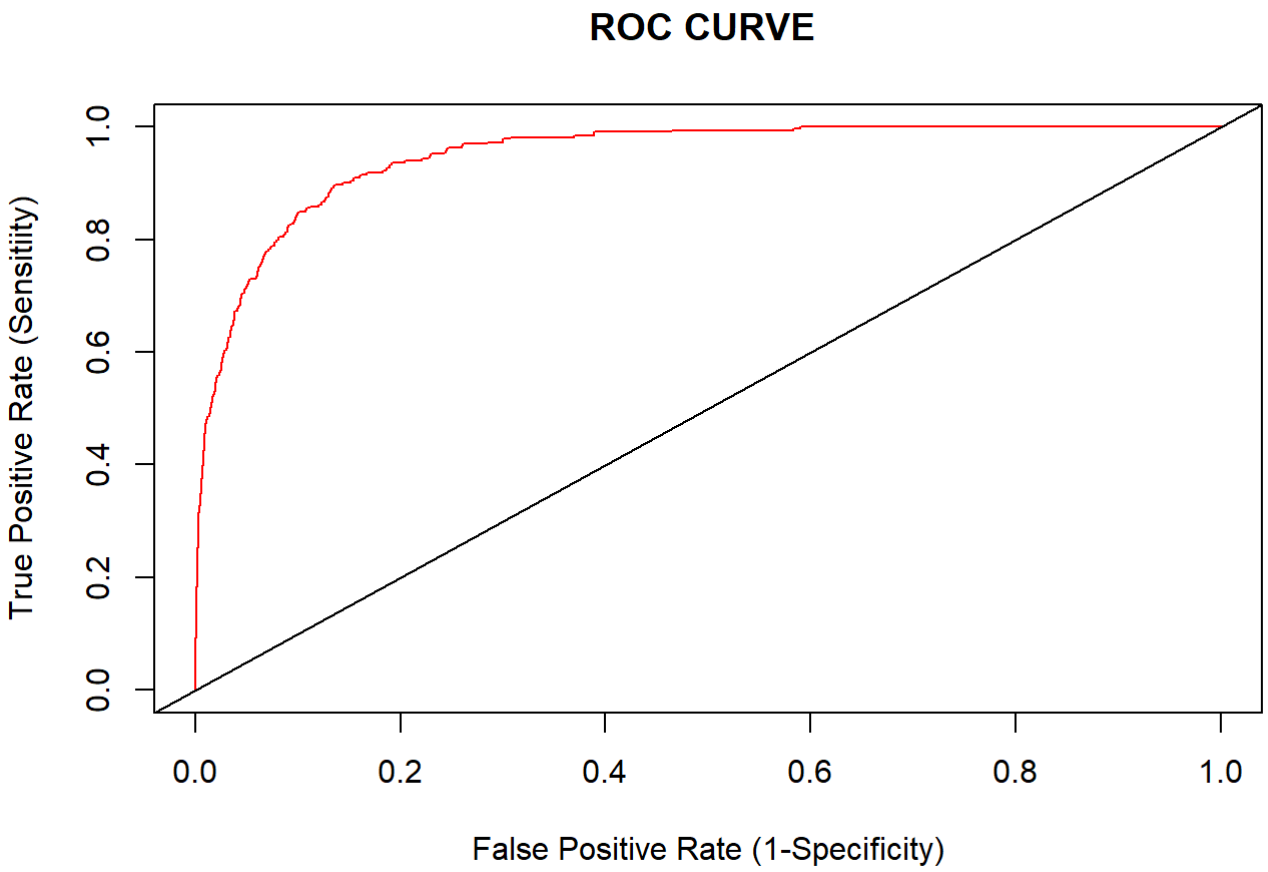

```
#  Option 2
roc1<-roc(Default$default,prob1)
plot(roc1) #also provides the ROC curve
```



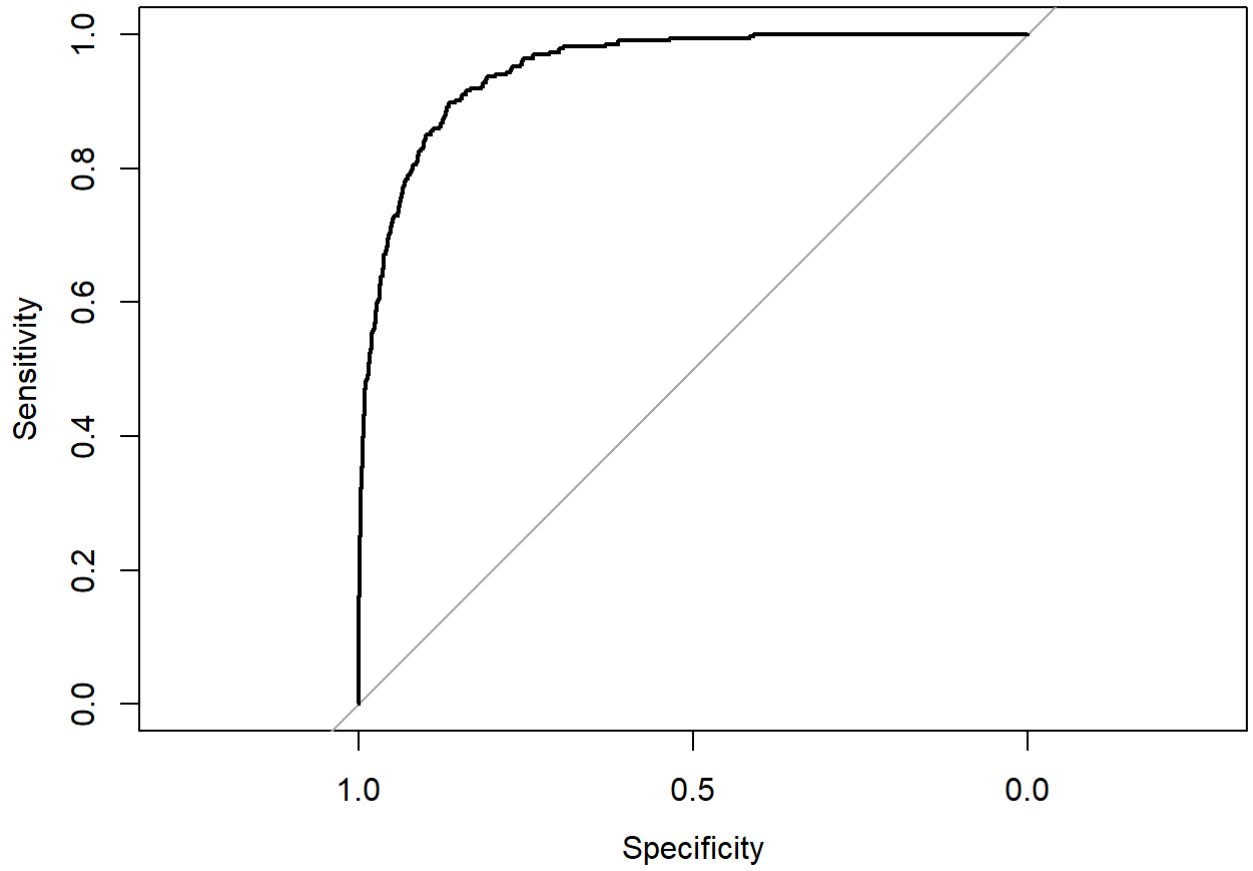

```
#-------AUC------------------
auc(roc1)
```

```
## Area under the curve: 0.948
```

```
# ROC&AUC for default = balance+income
#--------ROC Curve-----------
prob2=predict(mymullogit,type=c("response"))
pred<-prediction(prob2,Default$default)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate (S
ensitiity)")
abline(0,1)
```



```
#---------AUC------------
roc2<-roc(Default$default,prob2)
plot(roc2) #also provides the ROC curve
```




```
auc(roc2)
```

```
## Area under the curve: 0.9491
```

From the output, you can see that the AIC and AUC for predicting the probability of default by using only balance predictor was 1600.5 and 0.9482 respectively, while using both balance and income reduced the AIC to 1585 and increased AUC to 0.9491. We can summarise that having both variables in the model increase the precision of predicting the probability of default.

Inclass Practice Problem

Consider the problem of collusive bidding among road construction contractors. Contractors sometimes scheme to set bid prices higher than the fair market (or competitive) price. Suppose an investigator has obtained information on the bid status (1 if fixed bid or 0 if competitive bid) for a sample of 31 contracts. In addition, two variables thought to be related to bid status are also recorded for each contract: number of bidders x1 and the difference between the winning (lowest) bid and the estimated competitive bid (called the engineer's estimate) x2, measured as a percentage of the estimate. The data are provided in **ROADBIDS.csv** file

a. Use the Wald Z test to check if the probability of status depends on NUMBIDS and/or DOTEST at $alpha = 0.05$

b. From the result in a) find the Wald χ^2 statistic for the full logistic regression model with the p-value?

c. Use the Likelihood Ratio Test to check if the multiple logistic model (with NUMBIDS and DOTEST) performs better than the reduced model with NUMBIDS variable at $\alpha = 0.05$

d. Write both the logistic regression model of STATUS on NUMBIDS and DOTEST and the logit transformation of this logistic regression model.

e. Compare the model fit for models in part c) to confirm that the multiple logistic model performs better.

f. Predict the probability of STATUS when there were 4 bidders and the difference between the winning (lowest) bid and the estimated competitive bid (called the engineer's estimate) was 27.3% higher.

```
library(aod)  # for Wald test
library(ROCR) # for ROC
library(pROC) # for AUC
bid=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/ROADBIDS.csv", heade
r = TRUE)
summary(bid)
```

```
##      STATUS          NUMBIDS           DOTEST
##  Min.   :0.0000   Min.   : 2.000   Min.   :-11.70
##  1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.:  1.75
##  Median :0.0000   Median : 4.000   Median : 10.60
##  Mean   :0.3871   Mean   : 5.161   Mean   : 13.98
##  3rd Qu.:1.0000   3rd Qu.: 6.500   3rd Qu.: 23.10
##  Max.   :1.0000   Max.   :13.000   Max.   : 72.80
```

```
mymullogit <- glm(STATUS ~ NUMBIDS+DOTEST, data = bid, family = "binomial")
mylogit <- glm(STATUS ~ NUMBIDS, data = bid, family = "binomial")
#part a)
summary(mymullogit)
```

```
##
## Call:
## glm(formula = STATUS ~ NUMBIDS + DOTEST, family = "binomial",
##     data = bid)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5898  -0.5514  -0.1119   0.3973   2.3260
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.42120    1.28677   1.104   0.2694
## NUMBIDS     -0.75534    0.33880  -2.229   0.0258 *
## DOTEST       0.11220    0.05139   2.183   0.0290 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.381  on 30  degrees of freedom
## Residual deviance: 22.843  on 28  degrees of freedom
## AIC: 28.843
##
## Number of Fisher Scoring iterations: 6
```

```
summary(mylogit)
```

```
##
## Call:
## glm(formula = STATUS ~ NUMBIDS, family = "binomial", data = bid)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5481  -0.9030  -0.2786   1.0406   1.7086
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8565     1.0276   1.807   0.0708 .
## NUMBIDS      -0.5086     0.2320  -2.193   0.0283 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.381  on 30  degrees of freedom
## Residual deviance: 33.296  on 29  degrees of freedom
## AIC: 37.296
##
## Number of Fisher Scoring iterations: 5
```

```
#part b)

#The Wald chi square test for the full model
wald.test(b = coef(mymullogit), Sigma = vcov(mymullogit), Terms = 2:3)  #Terms tells R which terms in the
model are to be tested.
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 6.4, df = 2, P(> X2) = 0.04
```

```
#part c
anova(mylogit,mymullogit,test="Chisq")
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	29	33.29638	*NA*	*NA*	*NA*
2	28	22.84302	1	10.45335	0.001224271

2 rows

```
library(lmtest)
lrtest(mylogit,mymullogit)
```

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2	-16.64819	*NA*	*NA*	*NA*
2	3	-11.42151	1	10.45335	0.001224271

2 rows

```
#part e
# ROC&AUC for  status~numbid
#--------ROC Curve-----------
#option1
prob=predict(mylogit,type=c("response"))
roc<-roc(bid$STATUS,prob)
```

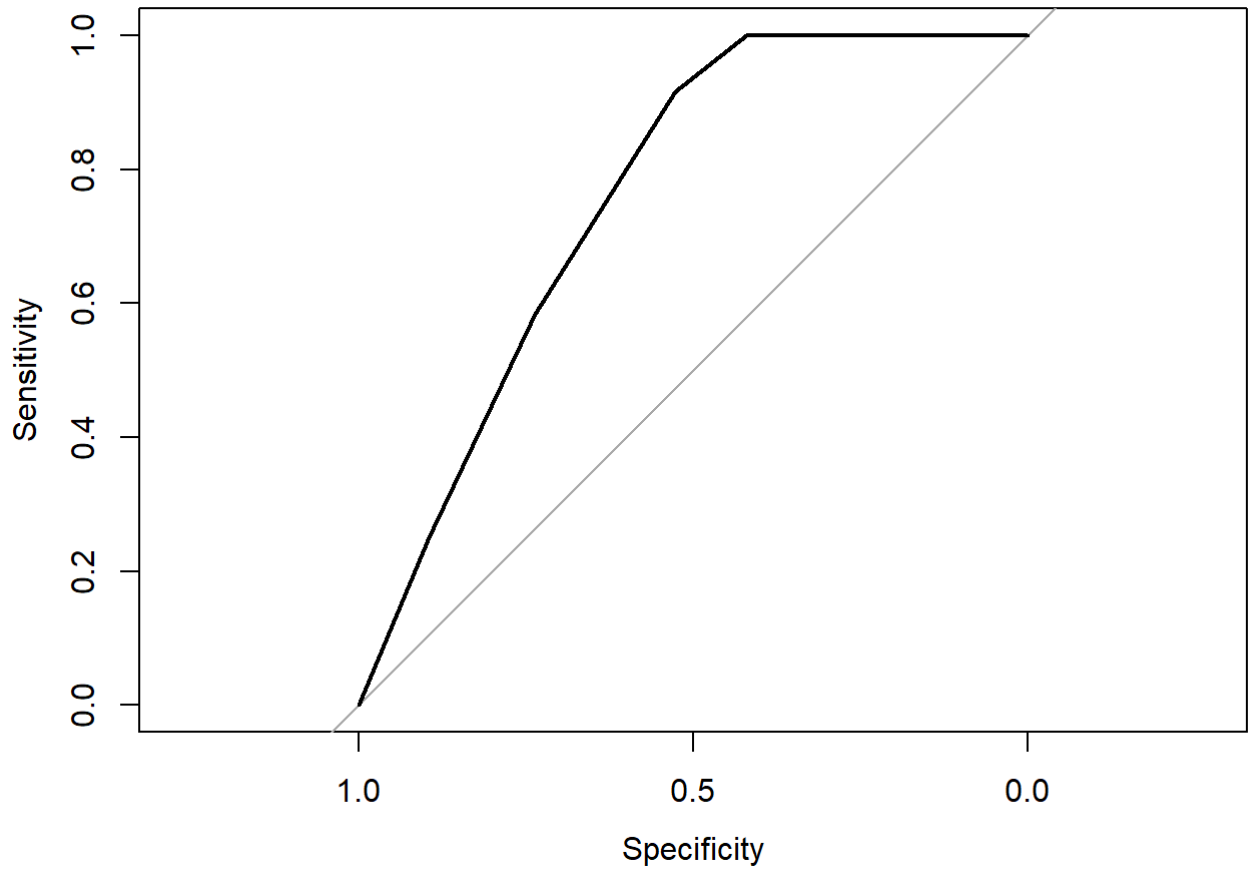
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

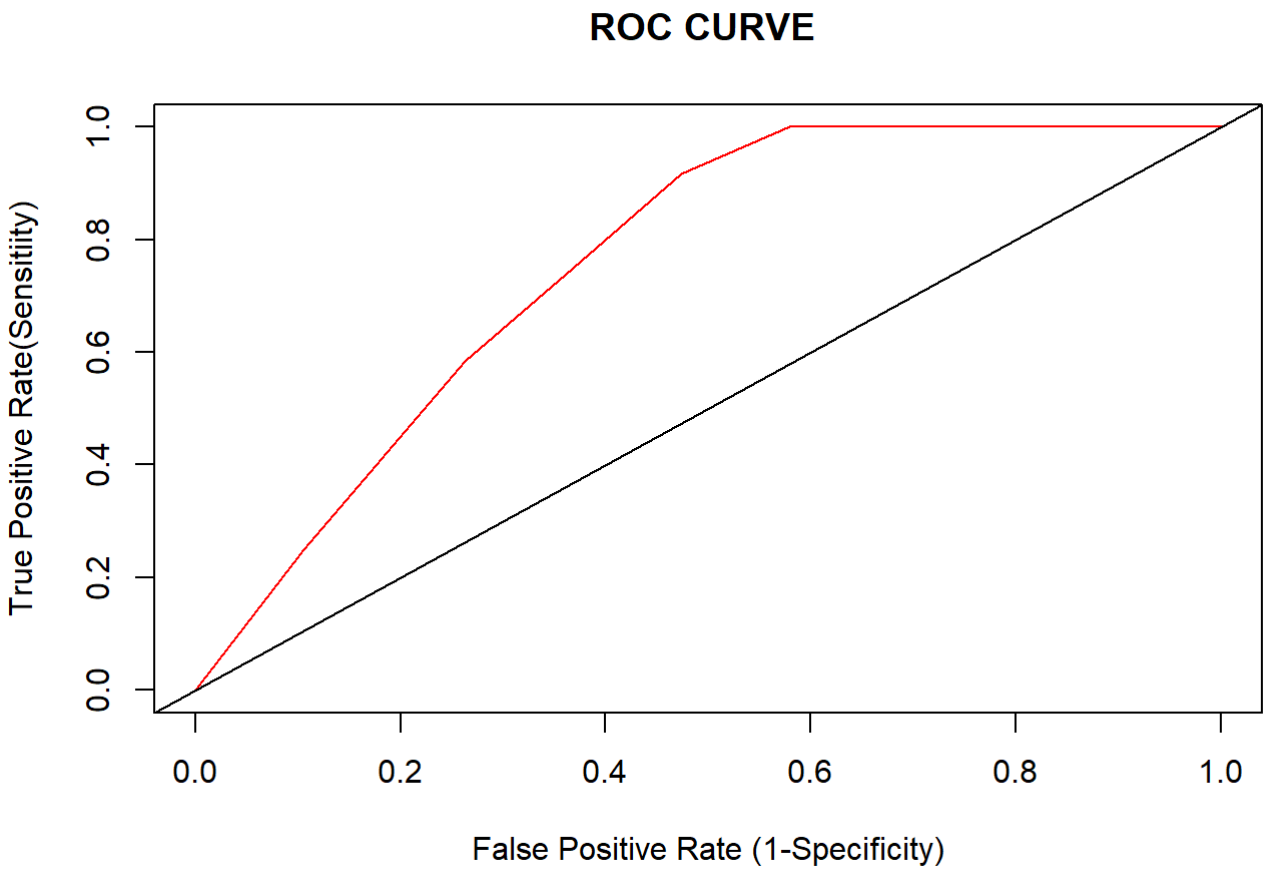
```
auc(roc)
```

```
## Area under the curve: 0.7588
```

```
plot(roc)
```



```
#option2
pred<-prediction(prob,bid$STATUS)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```




```
#---------AUC------------
roc<-roc(bid$STATUS,prob)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

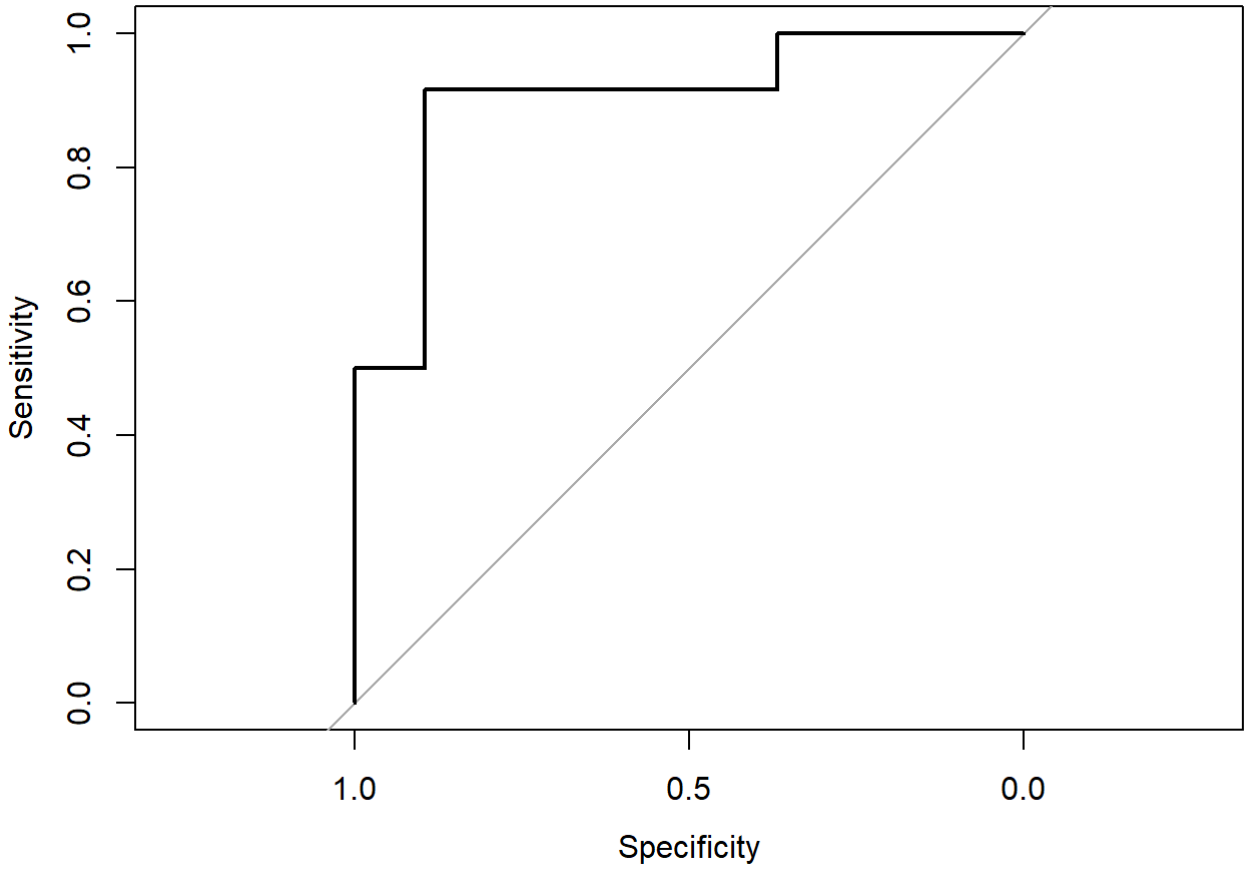
```
auc(roc)
```

```
## Area under the curve: 0.7588
```

```
# ROC&AUC for STATUS ~ NUMBIDS+DOTEST
#--------ROC Curve-----------
prob=predict(mymullogit,type=c("response"))
roc<-roc(bid$STATUS,prob)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

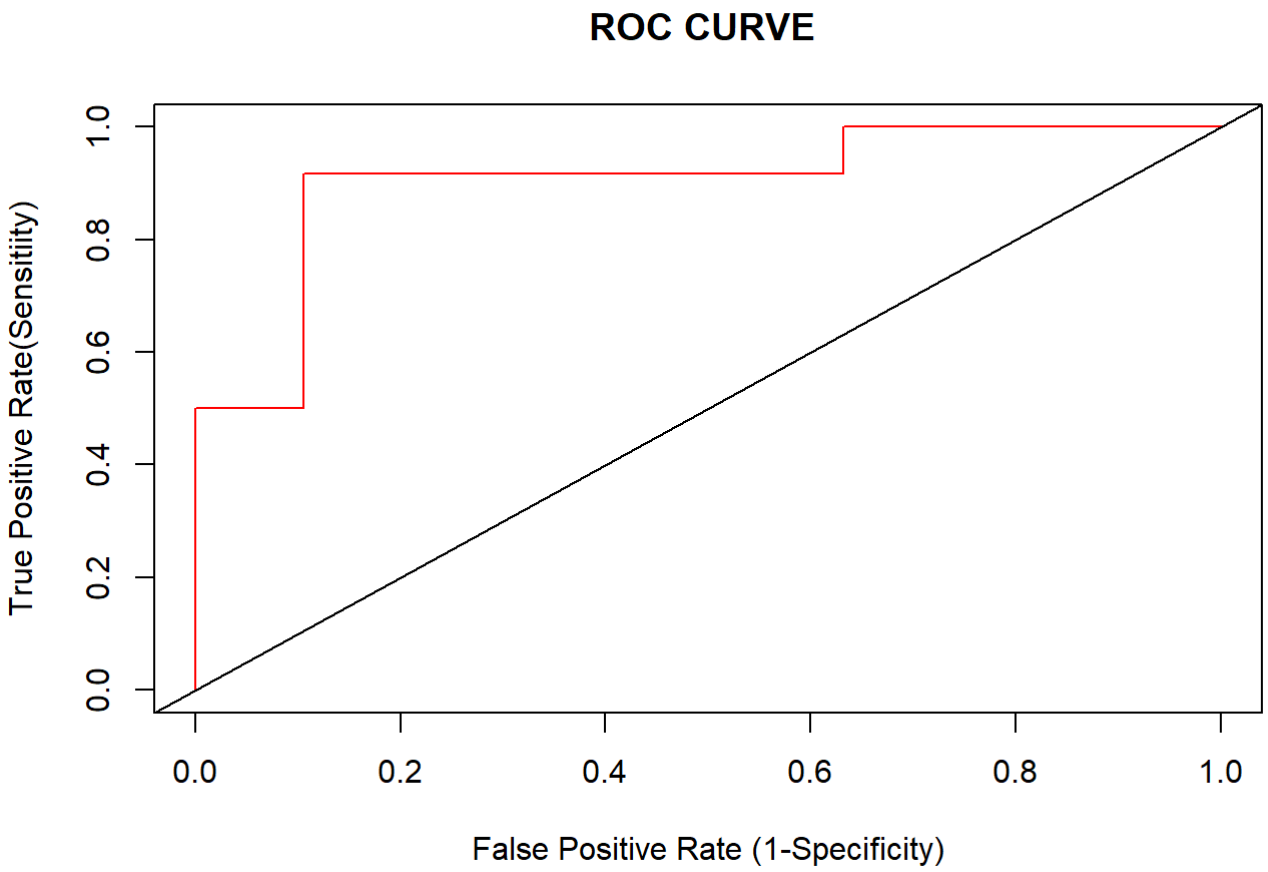
```
plot(roc)
```




```
auc(roc)
```

```
## Area under the curve: 0.9035
```

```
#option2
pred<-prediction(prob,bid$STATUS)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```




```
#---------AUC------------
roc<-roc(bid$STATUS,prob)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
auc(roc)
```

```
## Area under the curve: 0.9035
```

```
#part f)
newdata = data.frame(NUMBIDS=4, DOTEST=27.3)
predict(mymullogit, newdata, type="response")
```

```
##         1
## 0.8119964
```

Multiple Logistic Regression for a Binary Dependent Variable with both Quantitative and Qualitative variables

For Example: The desire data, showing the distribution of 24 currently married and fecund women interviewed in the Fiji Fertility Survey, according to age, education, desire for more children. the data is provided in disire.xlsx file

X1= age (year)

X2= education (0=none, 1=some),

Y= desire for more children (0=no more, 1=more),

```
library("readxl")
desire <- read_excel("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/desire.xlsx"
)
mymullogit <- glm(desire ~ age+factor(education), data = desire, family = "binomial")
summary(mymullogit)
```

```
##
## Call:
## glm(formula = desire ~ age + factor(education), family = "binomial",
##     data = desire)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.83453  -0.37036   0.04505   0.52734   2.16236
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         11.1323     4.4513   2.501  0.01239 *
## age                 -0.3714     0.1440  -2.579  0.00992 **
## factor(education)1   2.3006     1.4534   1.583  0.11344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.271  on 23  degrees of freedom
## Residual deviance: 15.987  on 21  degrees of freedom
## AIC: 21.987
##
## Number of Fisher Scoring iterations: 6
```

```
mylogit1 <- glm(desire ~ age, data = desire, family = "binomial")
summary(mylogit1)
```

```
##
## Call:
## glm(formula = desire ~ age, family = "binomial", data = desire)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1495  -0.5071   0.1165   0.4776   1.7500
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.2882     4.0169   2.810  0.00495 **
## age          -0.3493     0.1236  -2.826  0.00471 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33.271  on 23  degrees of freedom
## Residual deviance: 19.090  on 22  degrees of freedom
## AIC: 23.09
##
## Number of Fisher Scoring iterations: 5
```

```
anova(mylogit1,mymullogit,test='Chisq')
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	22	19.08977	*NA*	*NA*	*NA*
2	21	15.98675	1	3.103026	0.0781469

2 rows

Inclass Practice Problem

Suppose you are investigating allegations of experience, level of education and gender in the hiring practices of a particular firm. Data were collected on 60 former applicants to be used to fit the logit model where

Y= 1 if hired, 0 if not

X1= Years of education

X2= Years of experience

X3=1 if male applicant, 0 if female applicant

From **the DISCRIM.csv** data,

a. Test if The probability of hiring depends on education, experience, and/or gender at $alpha = 0.05$

Should we drop Education predictor?

b. Use the log likelihood ratio test to check if Education predictor should be added into the model.

c. Use the model fit methods (AIC, ROC curve with AUC) to confirm that adding Education into the model will improve the model performance.

d. write both the logistic regression model of HIRE on and the logit transformation of this logistic regression model.

e. Interpret the logistic regression coefficient $e^{\hat{\beta}_1}$ in logistic model

f. Predict the probability of hiring when a man who has worked for 6 years, has 4 years of education. Would he be hired for this firm.

```
library(ROCR)# for ROC
library(lmtest)# for lrtest() function
library(pROC)

discrim=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/DISCRIM.csv", he
ader = TRUE)
summary(discrim)
```

```
##       HIRE           EDUC            EXP              GENDER
##  Min.   :0.0   Min.   :4.00   Min.   : 0.000   Min.   :0.0
##  1st Qu.:0.0   1st Qu.:4.00   1st Qu.: 2.000   1st Qu.:0.0
##  Median :0.0   Median :6.00   Median : 3.500   Median :0.5
##  Mean   :0.4   Mean   :5.65   Mean   : 4.383   Mean   :0.5
##  3rd Qu.:1.0   3rd Qu.:6.00   3rd Qu.: 7.000   3rd Qu.:1.0
##  Max.   :1.0   Max.   :8.00   Max.   :12.000   Max.   :1.0
```

```
#part a),d),e)
mylogit1 <- glm(HIRE ~ factor(GENDER)+EXP+EDUC, data = discrim, family = "binomial")
summary(mylogit1)
```

```
##
## Call:
## glm(formula = HIRE ~ factor(GENDER) + EXP + EDUC, family = "binomial",
##     data = discrim)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5224  -0.4214  -0.1321   0.2853   3.2570
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.2627     2.5427  -3.250 0.001156 **
## factor(GENDER)1   2.1482     0.9319   2.305 0.021160 *
## EXP               0.7602     0.2100   3.620 0.000295 ***
## EDUC              0.5509     0.2974   1.852 0.063984 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80.761  on 59  degrees of freedom
## Residual deviance: 36.207  on 56  degrees of freedom
## AIC: 44.207
##
## Number of Fisher Scoring iterations: 6
```

```
mylogit2 <- glm(HIRE ~ factor(GENDER)+EXP, data = discrim, family = "binomial")
summary(mylogit2)
```

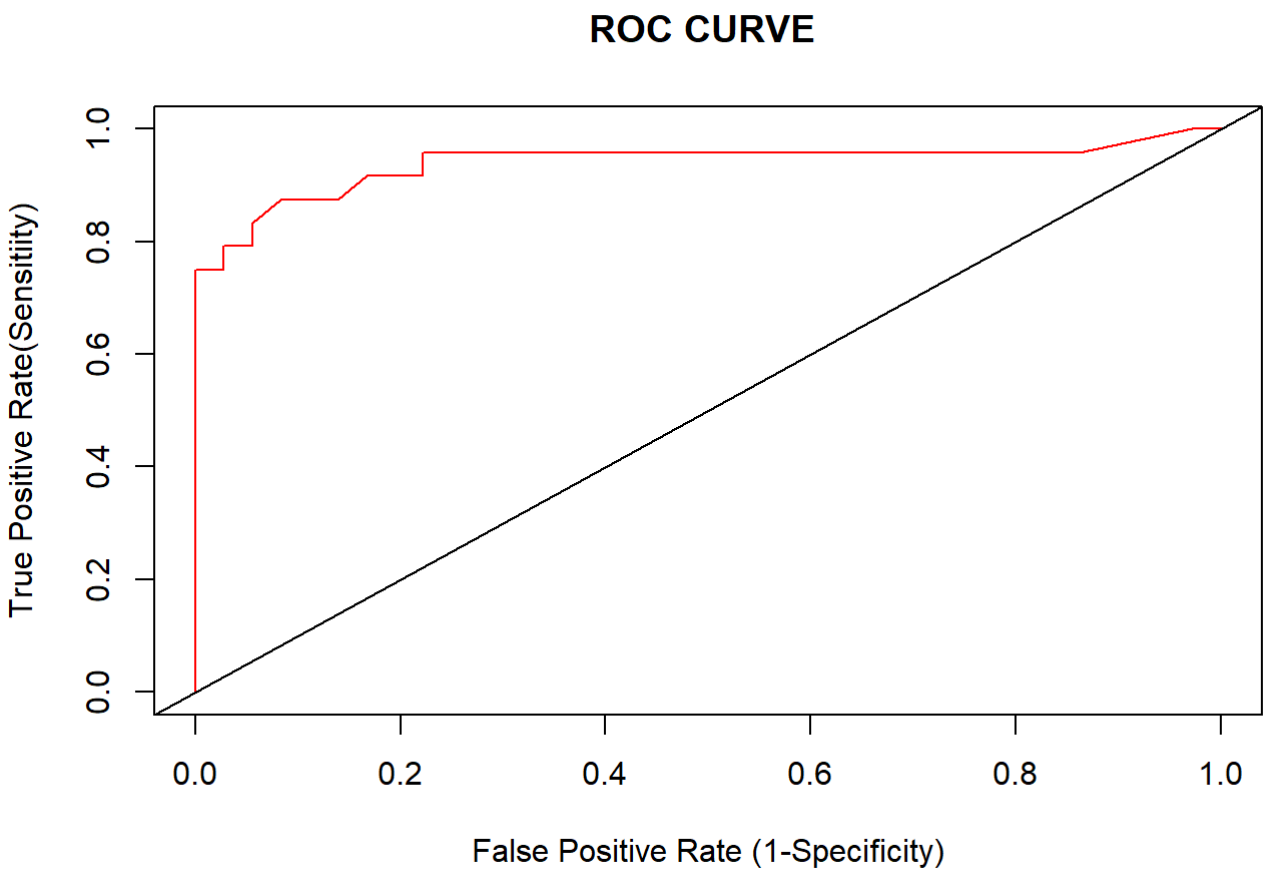
```
##
## Call:
## glm(formula = HIRE ~ factor(GENDER) + EXP, family = "binomial",
##     data = discrim)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8985  -0.5211  -0.1901   0.3437   2.8363
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.7073     1.1744  -4.008 6.12e-05 ***
## factor(GENDER)1   2.0647     0.8770   2.354 0.018555 *
## EXP               0.7032     0.1816   3.872 0.000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80.761  on 59  degrees of freedom
## Residual deviance: 40.128  on 57  degrees of freedom
## AIC: 46.128
##
## Number of Fisher Scoring iterations: 6
```

```
#part b)
#Likelihood Ratio Test
lrtest(mylogit2,mylogit1)
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	3	-20.06417	*NA*	*NA*	*NA*
2	4	-18.10364	1	3.921053	0.04768499

2 rows

```
# ROC&AUC for HIRE ~ factor(GENDER)+EXP+EDUC
#--------ROC Curve-----------
prob=predict(mylogit1,type=c("response"))
pred<-prediction(prob,discrim$HIRE)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```

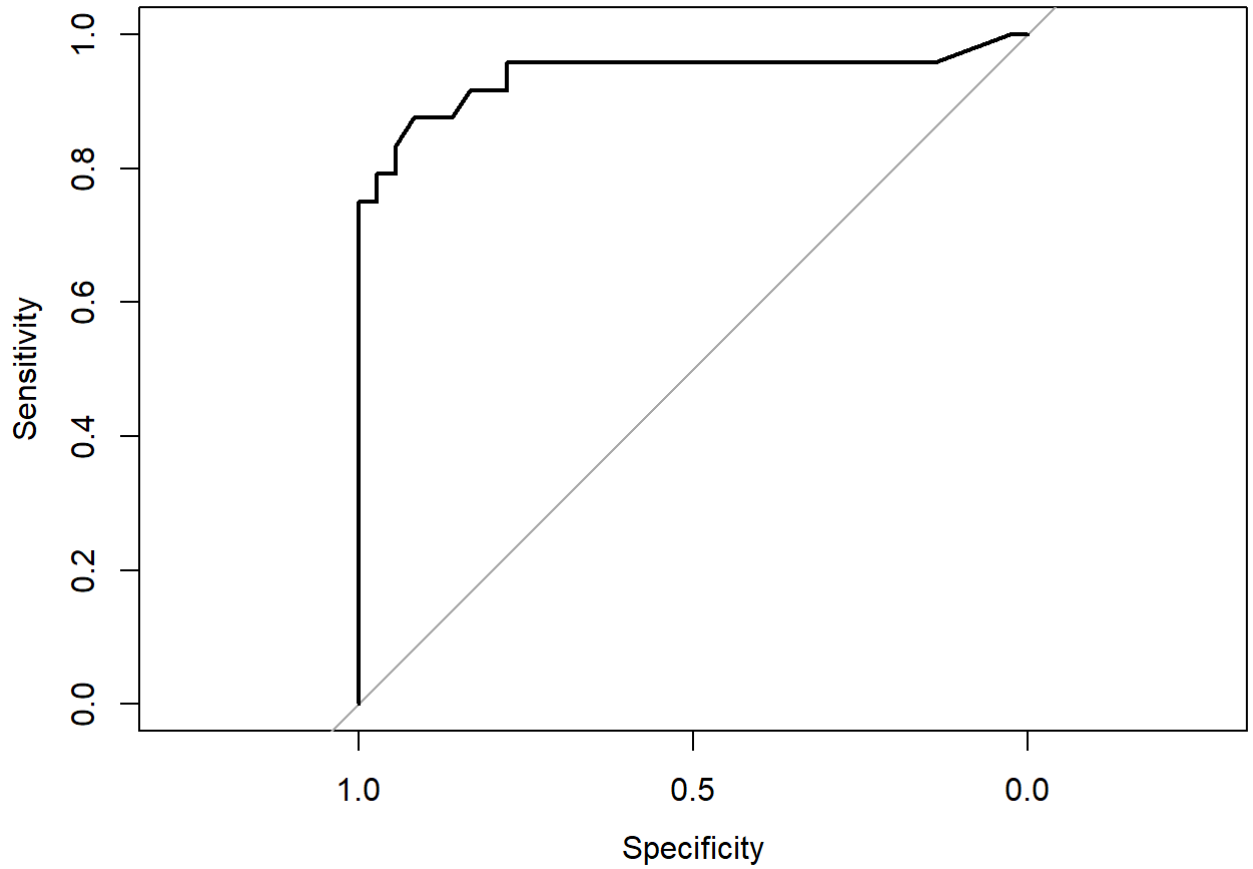


```
#---------AUC------------
roc<-roc(discrim$HIRE,prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

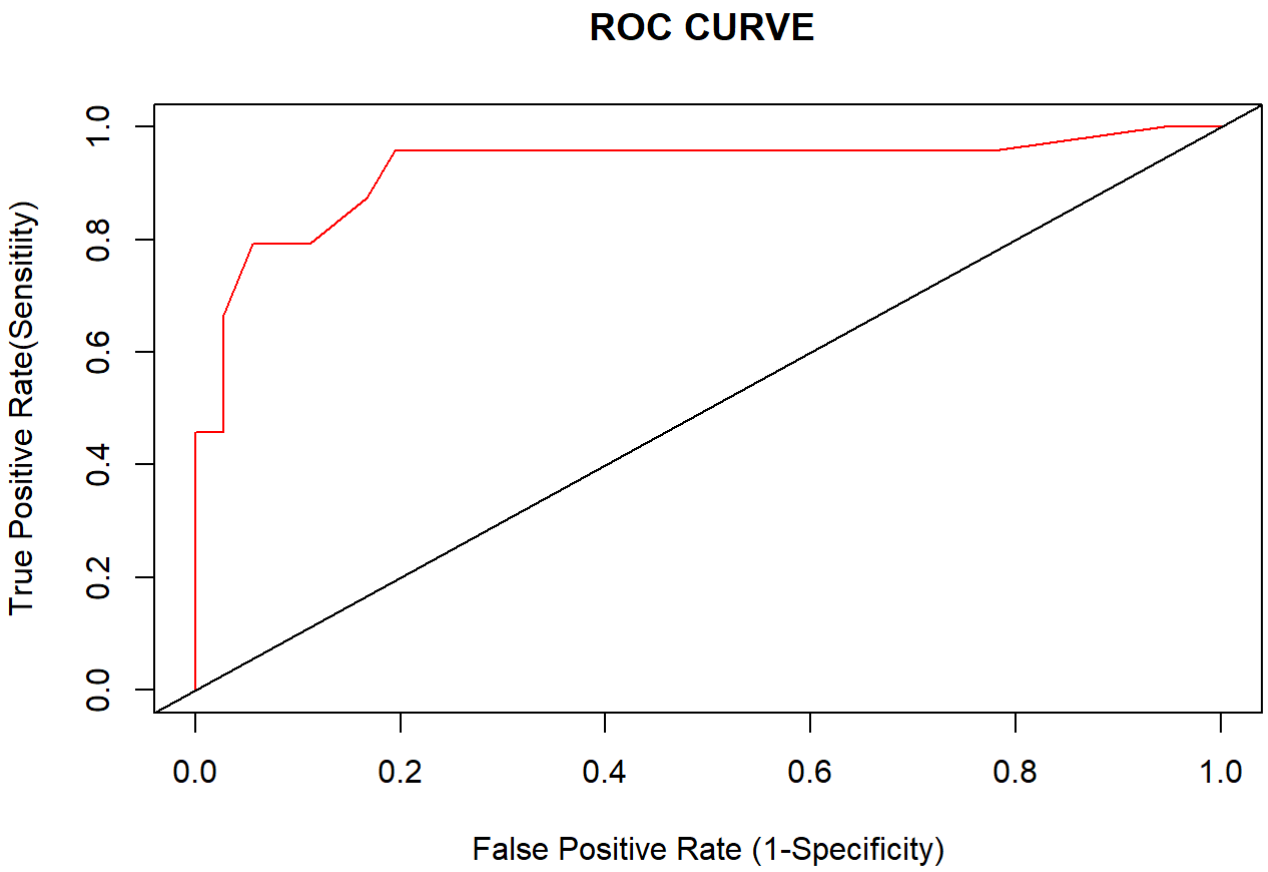
```
plot(roc)
```



```
auc(roc)
```

```
## Area under the curve: 0.9398
```

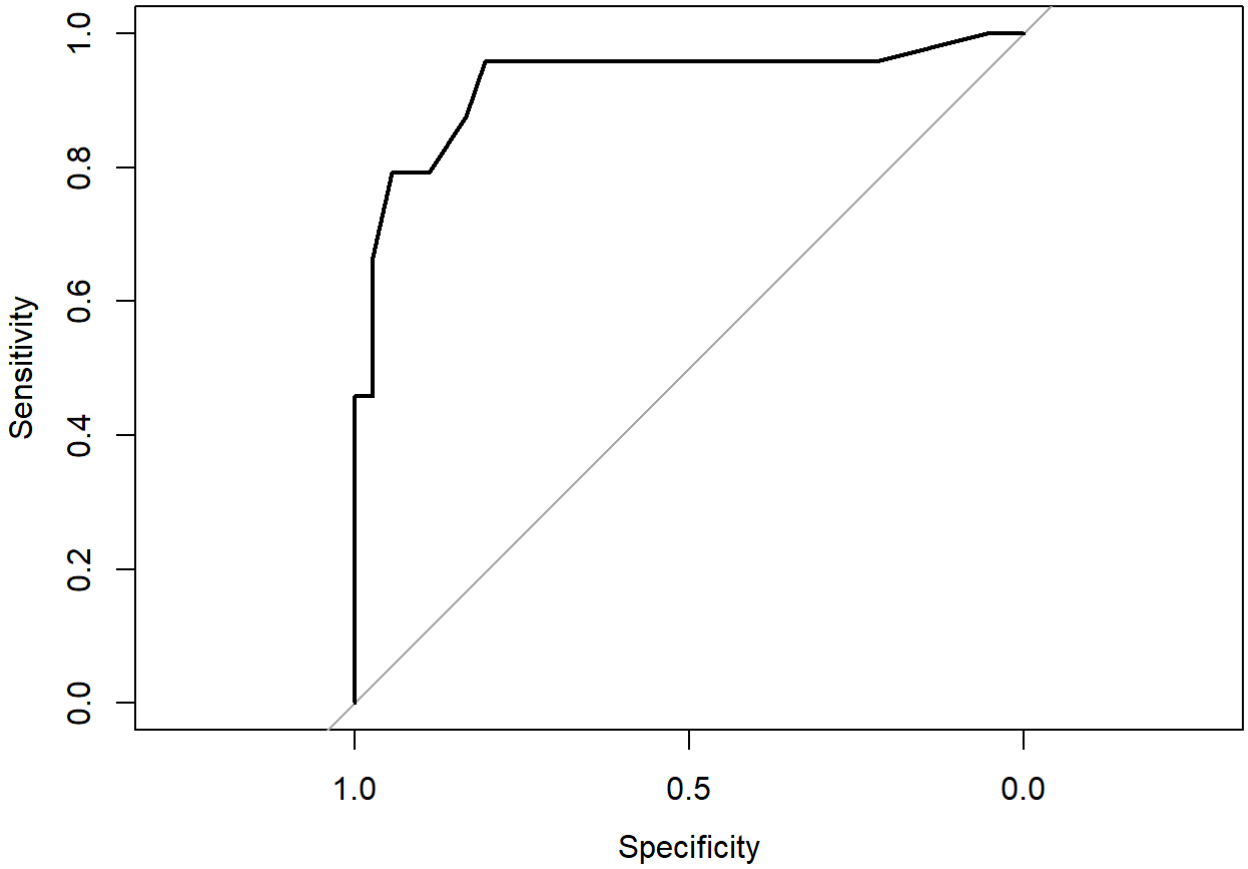
```
# ROC&AUC for HIRE ~ factor(GENDER)+EXP
#--------ROC Curve-----------
prob=predict(mylogit2,type=c("response"))
pred<-prediction(prob,discrim$HIRE)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```




```
#---------AUC------------
roc<-roc(discrim$HIRE,prob)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot(roc)
```




```
auc(roc)
```

```
## Area under the curve: 0.9265
```

```
#part f)
newdata = data.frame(GENDER=1, EXP=6, EDUC=4)
predict(mylogit1, newdata, type="response")
```

```
##         1
## 0.6570043
```

Inclass Practice Problem

The German Credit Data contains data on 6 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. A predictive model developed on this data provided in *creditbility.xlsx* file is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.The independent variables are listed below

Creditability= (1 if good credit, 0 if bad credit)

Balance=Account Balance (Categorical variable with 4 levels)

$$Balance = \begin{cases} 1 \text{ if balance is more than } 5000 \\ 2 \text{ if balance is } 3001 - 5000 \\ 3 \text{ if balance is } 1001 - 3000 \\ 4 \text{ if balance is less than } 1000 \end{cases}$$

Duration= Duration of credit in months (months)

Employment=Length of current employment (years)

Amount=Credit amount (dollars)

Age=Age (year)

```
library("readxl")
library(lmtest)# for lrtest() function
creditdata <- read_excel("c:/Users/thuntida.ngamkham/OneDrive - University of Calgary/dataset603/creditbi
lity.xlsx")
myfulllogit<-glm(Creditability~Age+Amount+employment+Duration+factor(Balance),data=creditdata,family="bin
omial")
summary(myfulllogit)
```

```
##
## Call:
## glm(formula = Creditability ~ Age + Amount + employment + Duration +
##     factor(Balance), family = "binomial", data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4298  -0.9897   0.4731   0.8329   1.7524
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.062e-01  3.320e-01  -0.320  0.74910
## Age               1.173e-02  7.094e-03   1.653  0.09829 .
## Amount           -2.820e-05  3.267e-05  -0.863  0.38807
## employment        1.653e-01  6.553e-02   2.523  0.01165 *
## Duration         -3.459e-02  7.835e-03  -4.414 1.01e-05 ***
## factor(Balance)2  5.441e-01  1.820e-01   2.990  0.00279 **
## factor(Balance)3  1.073e+00  3.329e-01   3.222  0.00127 **
## factor(Balance)4  1.992e+00  2.035e-01   9.787  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1039.1  on 992  degrees of freedom
## AIC: 1055.1
##
## Number of Fisher Scoring iterations: 4
```

```
mylogit1<-glm(Creditability~Age+employment+Duration+factor(Balance),data=creditdata,family="binomial")
mylogit2<-glm(Creditability~employment+Duration+factor(Balance),data=creditdata,family="binomial")
summary(mylogit1)
```

```
##
## Call:
## glm(formula = Creditability ~ Age + employment + Duration + factor(Balance),
##     family = "binomial", data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4221  -0.9881   0.4769   0.8372   1.7071
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.097657   0.331700  -0.294  0.76844
## Age               0.011108   0.007051   1.575  0.11517
## employment        0.169384   0.065376   2.591  0.00957 **
## Duration         -0.038762   0.006200  -6.252 4.06e-10 ***
## factor(Balance)2  0.529841   0.181103   2.926  0.00344 **
## factor(Balance)3  1.085209   0.332655   3.262  0.00111 **
## factor(Balance)4  1.983625   0.203106   9.766  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1039.9  on 993  degrees of freedom
## AIC: 1053.9
##
## Number of Fisher Scoring iterations: 4
```

```
summary(mylogit2)
```

```
##
## Call:
## glm(formula = Creditability ~ employment + Duration + factor(Balance),
##     family = "binomial", data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3519  -1.0079   0.4854   0.8433   1.7237
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.216968   0.264923   0.819 0.412794
## employment        0.193602   0.063280   3.059 0.002217 **
## Duration         -0.039002   0.006197  -6.294  3.1e-10 ***
## factor(Balance)2  0.521296   0.180830   2.883 0.003942 **
## factor(Balance)3  1.098172   0.331798   3.310 0.000934 ***
## factor(Balance)4  1.983720   0.202793   9.782  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1042.4  on 994  degrees of freedom
## AIC: 1054.4
##
## Number of Fisher Scoring iterations: 4
```

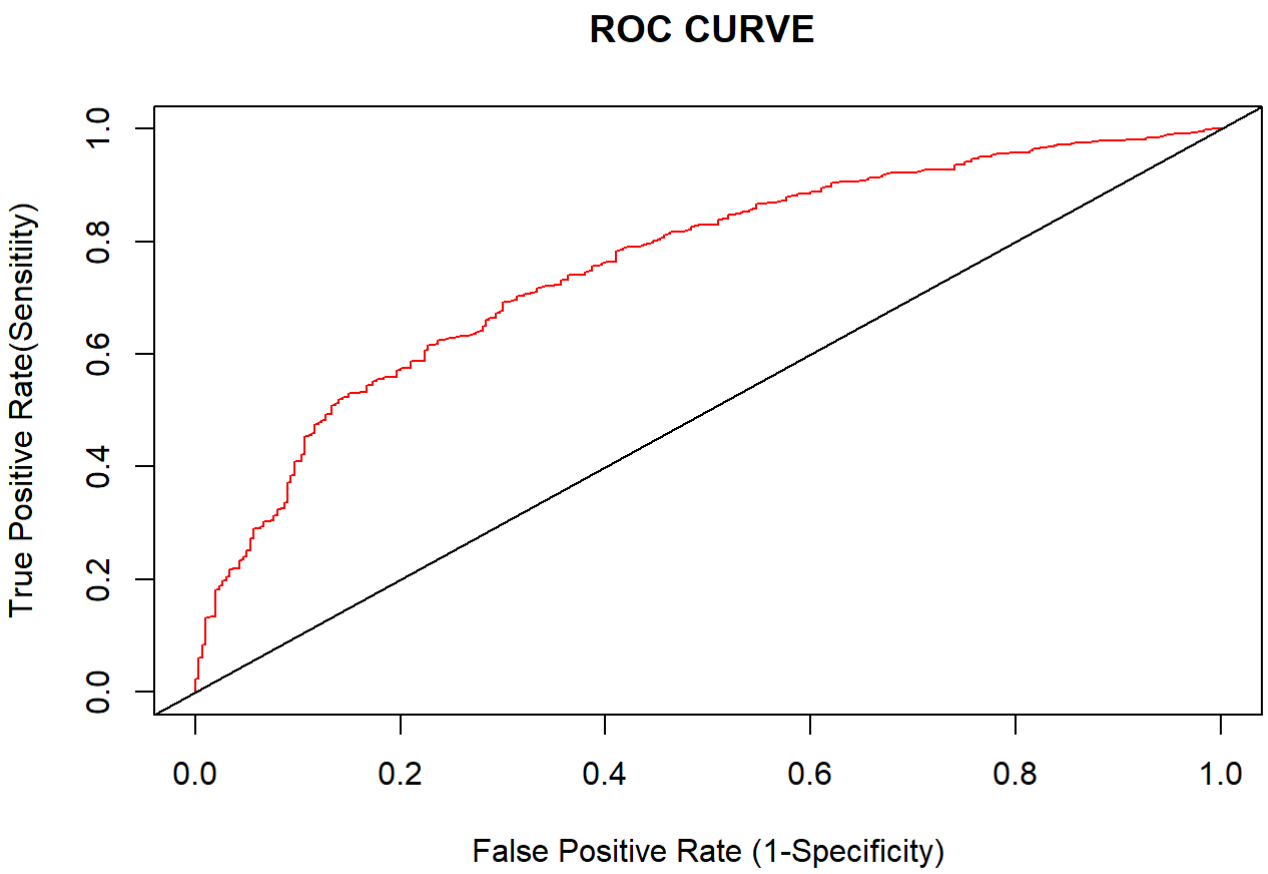
```
#part a)
#Likelihood Ratio Test
lrtest(myfulllogit,mylogit2)
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	8	-519.5724	*NA*	*NA*	*NA*
2	6	-521.2059	-2	3.266944	0.1952504

2 rows

```
#part b)
# ROC&AUC for Creditability~Age+employment+Duration+factor(Balance)
#--------ROC Curve-----------
prob=predict(mylogit1,type=c("response"))
pred<-prediction(prob,creditdata$Creditability)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```




```
#---------AUC------------
roc<-roc(creditdata$Creditability,prob)
```

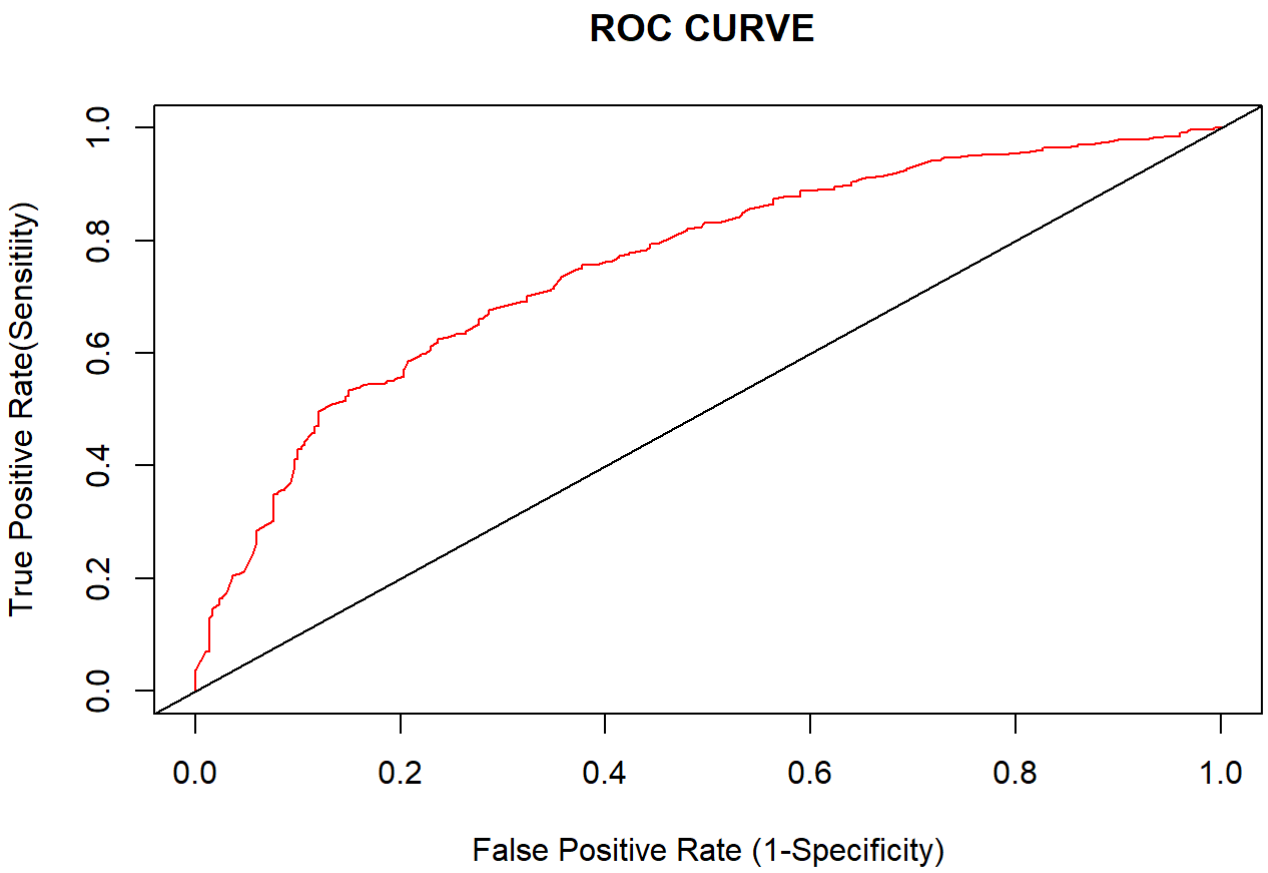
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc(roc)
```

```
## Area under the curve: 0.7582
```

```
# ROC&AUC for Creditability~employment+Duration+factor(Balance)
#--------ROC Curve-----------
prob=predict(mylogit2,type=c("response"))
pred<-prediction(prob,creditdata$Creditability)
perf<-performance(pred,measure = "tpr",x.measure="fpr")
plot(perf,col=2,main="ROC CURVE ", xlab="False Positive Rate (1-Specificity)",ylab="True Positive Rate(Se
nsitiity)")
abline(0,1)
```



```
#---------AUC------------
roc<-roc(creditdata$Creditability,prob)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
auc(roc)
```

```
## Area under the curve: 0.7568
```

```
#part f)
newdata = data.frame(employment=2,Duration=20,Balance=3)
predict(mylogit2, newdata, type="response")
```

```
##         1
## 0.7155111
```

a. Test if The probability of Creditbility depends on all predictors at $alpha = 0.05$

Should we drop both AGE and AMOUNT predictors? (if not sure, use log likelihood ratio to test between reduced model and full model)

b. Use the model fit methods(AIC, ROC curve with AUC) to check the model performance.

c. Write a logistic model for a person who has 2500 dollars account balance.

d. From part c) would you consider this applicant as a bad credit risk if the duration of credit in months = 16 months with the Length of current employment = 2 years.