> rm(list = ls())

> setwd("E:/Data of R")

>

> #Question 1

> #(a)

>

> y=c(0,0,0,0,1,1,1,1)

> x1=c(1,2,3,3,5,6,10,11)

>

> data11=data.frame(cbind(y,x1))

>

> model11=glm(y~x1,data=data11,family = binomial(link = logit))

Warning message:

glm.fit:拟合機率算出来是数值零或一

> summary(model11)

Call:

glm(formula = y ~ x1, family = binomial(link = logit), data = data11)

Deviance Residuals:

Min 1Q Median 3Q Max

-8.605e-06 -2.167e-06 0.000e+00 2.110e-08 1.288e-05

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -94.87 202572.35 0 1

x1 23.62 48491.51 0 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.109e+01 on 7 degrees of freedom

Residual deviance: 3.139e-10 on 6 degrees of freedom

AIC: 4

Number of Fisher Scoring iterations: 25

>

> #Coefficients and standard errors

> coef(summary(model11))

Estimate Std. Error z value Pr(>|z|)

(Intercept) -94.86875 202572.35 -0.0004683203 0.9996263

x1 23.61643 48491.51 0.0004870218 0.9996114

>

> exp(model11$coefficients)

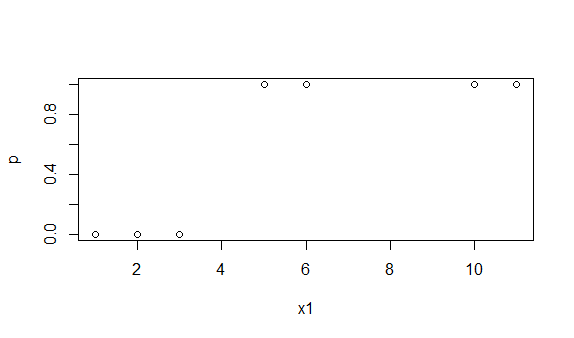
(Intercept) x1

6.295460e-42 1.805026e+10

>

> p=predict(model11,type = 'response')

> qqplot(x1,p)

> 

> #As can be seen from the plot, warning message means that

> #When y=1,pi\_hat=1. When y=0, pi\_hat=0

> #Which is also an indication of complete seperation

> #Another signal is the unnaturally large standard errors

>

> #(b)

> y=c(0,0,0,0,1,1,1,1)

> x2=c(1,2,3,3,3,6,10,11)

>

> data12=data.frame(cbind(y,x2))

>

> model12=glm(y~x2,data=data12,family = binomial(link = logit))

Warning message:

glm.fit:拟合機率算出来是数值零或一

> summary(model12)

Call:

glm(formula = y ~ x2, family = binomial(link = logit), data = data12)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.9005 -0.2252 0.0000 0.0000 1.4823

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -54.08 18834.18 -0.003 0.998

x2 17.80 6278.06 0.003 0.998

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11.0904 on 7 degrees of freedom

Residual deviance: 3.8191 on 6 degrees of freedom

AIC: 7.8191

Number of Fisher Scoring iterations: 21

>

> #Coefficients and standard errors

> coef(summary(model12))

Estimate Std. Error z value Pr(>|z|)

(Intercept) -54.08260 18834.176 -0.002871514 0.9977089

x2 17.79649 6278.059 0.002834711 0.9977382

>

> exp(model12$coefficients)

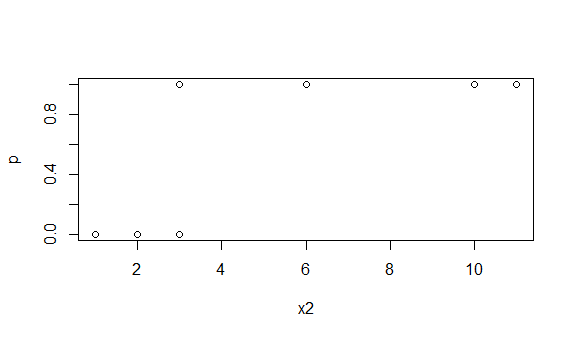
(Intercept) x2

3.252551e-24 5.356922e+07

>

> p=predict(model11,type = 'response')

> qqplot(x2,p)

> 

> #As can be seen from the plot, warning message means that

> #Some observations have pi\_hat=1 or 0, there is not perfect discrimination

> #Which is an indication of quasi-complete separation

> #Another signal is also the unnaturally large standard errors

> #Question 3

>

> data3=read.table("donner.txt",header = T)

>

> #(a)

> model3a=glm(survival~age,data=data3, family = binomial(link = logit))

> summary(model3a)

Call:

glm(formula = survival ~ age, family = binomial(link = logit),

data = data3)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5946 -1.2017 0.8436 0.9882 1.5765

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.97917 0.37460 2.614 0.00895 \*\*

age -0.03689 0.01493 -2.471 0.01346 \*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 120.86 on 87 degrees of freedom

Residual deviance: 114.02 on 86 degrees of freedom

AIC: 118.02

Number of Fisher Scoring iterations: 4

>

> #(b)

> model3a$coefficients[2]

age

-0.03688823

> #Interpretation for

> #For every one year increase in age, the log odds of survial decreased by 0.03689

>

> 1/exp(model3a$coefficients[2])

age

1.037577

> #Interpretation for

> #For every one year increase in age, a person is 1.0375770 times less likely to survive

>

> #(c)

> #H0:

> #Ha:

>

> #Z= =0.03689/(0.01493)=2.471>1.96

>

> #P-value is 0.01346

> #Reject H0

>

> #(d)

> coef(summary(model3a))

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.97917294 0.37459933 2.613921 0.008950982

age -0.03688823 0.01492559 -2.471476 0.013455674

>

> #Confidence Interval for

> lb=coef(summary(model3a))[2,1]-qnorm(0.975)\*coef(summary(model3a))[2,2]

> ub=coef(summary(model3a))[2,1]+qnorm(0.975)\*coef(summary(model3a))[2,2]

> c(lb,ub)

[1] -0.066141852 -0.007634613

>

> #Confidence Interval for

> exp(c(lb,ub))

[1] 0.9359981 0.9923945

>

> #Effect of age on survial is statistically significant

> #because the CI for does not contain 0 and the CI for does not contain 1

>

> #(e)

> age2=data3$age^2

> model3b=glm(survival~age+age2+sex+status, data = data3, family = binomial(link = logit))

> summary(model3b)

Call:

glm(formula = survival ~ age + age2 + sex + status, family = binomial(link = logit),

data = data3)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0431 -1.0391 0.5120 0.8664 2.0797

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.986e-01 6.172e-01 0.322 0.7476

age 1.675e-01 7.107e-02 2.357 0.0184 \*

age2 -3.889e-03 1.525e-03 -2.550 0.0108 \*

sex -6.637e-01 5.588e-01 -1.188 0.2349

statusHired -1.625e+00 7.481e-01 -2.173 0.0298 \*

statusSingle -1.852e+01 1.760e+03 -0.011 0.9916

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 120.855 on 87 degrees of freedom

Residual deviance: 92.363 on 82 degrees of freedom

AIC: 104.36

Number of Fisher Scoring iterations: 16

>

> #(f)

> #H0:reduced model

> #Ha:full model

>

> #

>

> #Reject if

>

> diff.dev=deviance(model3a)-deviance(model3b)

> diff.dev

[1] 21.65525

> qchisq(0.95,4)

[1] 9.487729

>

> #21.65525 > 9.487729

> #Reject H0

>

> #(g)

> model3b$coefficients[4]

sex

-0.663728

> #Interpretation for

> #A male decreases the log odds of survival by 0.663728050

>

> 1/exp(model3b$coefficients[4])

sex

1.942019

> #Interpretation for

> #A male is 1.942019 times less likely to survive

>

> #Question 4

> #(a)

> H=rep(0,88)

> for (i in 1:88) {

+ if(data3$status[i]=='Hired'){

+ H[i]=1

+ }

+ }

> S=rep(0,88)

> for (i in 1:88) {

+ if(data3$status[i]=='Single'){

+ S[i]=1

+ }

+ }

> X=cbind(1,data3$age,age2,data3$sex,H,S)

> fv=X%\*%model3b$coefficients

> summary(fv)

V1

Min. :-17.8917

1st Qu.: -0.3345

Median : 0.3622

Mean : -0.7447

3rd Qu.: 1.1821

Max. : 2.0004

> #(c)

> pi\_hat=exp(fv)/(1+exp(fv))

> summary(pi\_hat)

V1

Min. :0.0000

1st Qu.:0.4171

Median :0.5896

Mean :0.5568

3rd Qu.:0.7653

Max. :0.8808

>

> #(d)

> #(1)

> y\_hat1=rep(0,88)

> for(i in 1:88){

+ if(pi\_hat[i]>0.5){

+ y\_hat1[i]=1

+ } else{

+ y\_hat1[i]=0

+ }

+ }

>

> T1=table(data3$survival,y\_hat1)

> T1

y\_hat1

0 1

0 23 16

1 7 42

>

> #(2)

> y\_hat2=rep(0,88)

> for(i in 1:88){

+ if(pi\_hat[i]>(sum(data3$survival)/88)){

+ y\_hat2[i]=1

+ } else{

+ y\_hat2[i]=0

+ }

+ }

>

> T2=table(data3$survival,y\_hat2)

> T2

y\_hat2

0 1

0 27 12

1 10 39

>

> #(3)

> port1=T1[1,1]/88

> port1

[1] 0.2613636

> port2=(T2[1,2]+T2[2,1])/88

> port2

[1] 0.25

> #0.2613636 > 0.25

> #Second cutoff is better

>

> #(4)

> FPR1=T1[1,2]/(sum(T1[1,]))

> FNR1=T1[2,1]/(sum(T1[2,]))

> FPR2=T2[1,2]/(sum(T2[1,]))

> FNR2=T2[2,1]/(sum(T2[2,]))

>

> #(5)

> c(FPR1,FPR2)

[1] 0.4102564 0.3076923

> #0.4102564 > 0.3076923

> c(FNR1,FNR2)

[1] 0.1428571 0.2040816

> #0.1428571 < 0.2040816

> #It is hard to decide

>

> #(6)

> library(pROC)

> data4=data.frame(cbind(data3$survival,pi\_hat,y\_hat1,y\_hat2))

> roc(data4$V1,data4$V2,plot = TRUE)

Call:

roc.default(response = data4$V1, predictor = data4$V2, plot = TRUE)

Data: data4$V2 in 39 controls (data4$V1 0) < 49 cases (data4$V1 1).

Area under the curve: 0.8004

> points(1-FPR1,1-FNR1,col='red', cex=2, pch=21)

> points(1-FPR2,1-FNR2,col='blue', cex=2, pch=24)

