Assignment-3.R

axelj

2022-11-17

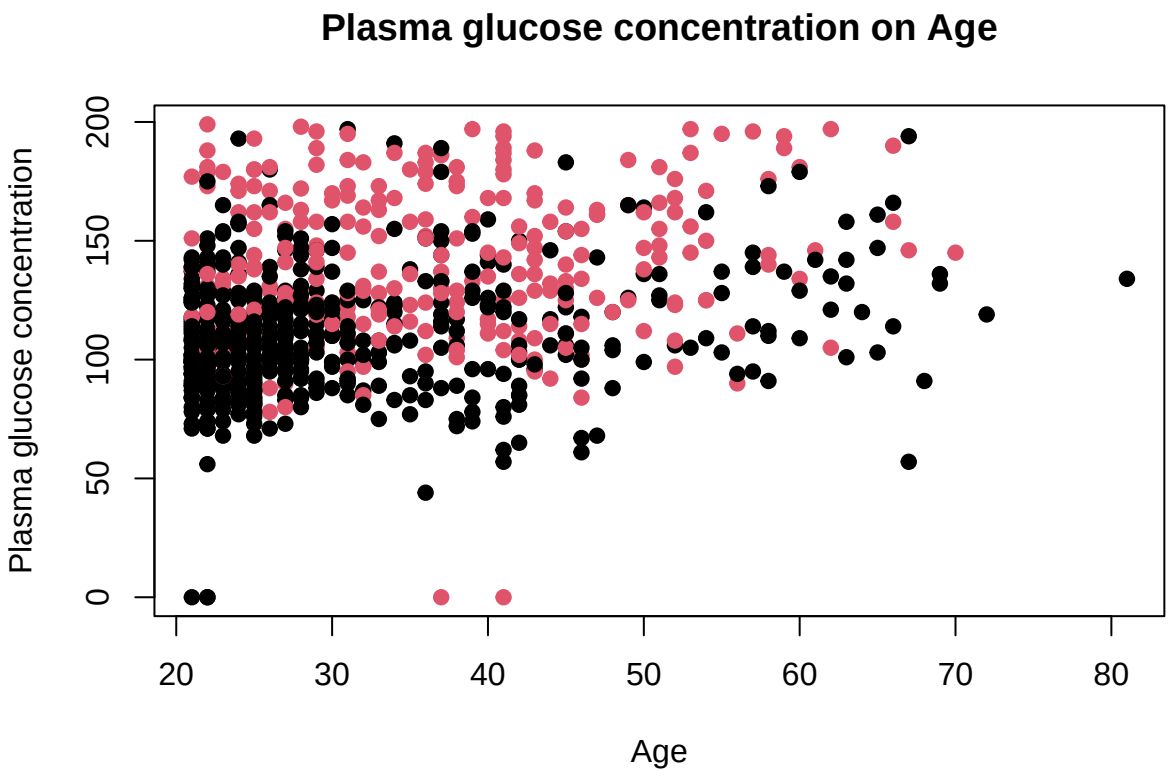
```
#Assignment 3 lab1
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
#Part 1
diabetes = read.csv('pima-indians-diabetes.csv')
diabetes1 = as.data.frame(diabetes)
x = diabetes1[[8]] #Vector of plasma glucose concentration levels
y = diabetes1[[2]] #Vector of ages
plot(x,y, col=as.factor(diabetes1$X1), pch=19,
     main="Plasma glucose concentration on Age",
     xlab="Age", ylab="Plasma glucose concentration")
```




```
#Part 2
set.seed(12345)
train=diabetes%>%select(X1, X148, X50)
m1=glm(as.factor(diabetes1$X1)~., train, family="binomial")
coef(m1)
```

```
## (Intercept)        X148         X50
## -5.89785793  0.03558250  0.02450157
```

```
Prob=predict(m1, type="response")
Pred=ifelse(Prob>0.5, "1", "0")
table(train$X1, Pred)
```

```
##    Pred
##       0   1
##   0 436  64
##   1 140 127
```

```
summary(m1)
```

```
##
## Call:
## glm(formula = as.factor(diabetes1$X1) ~ ., family = "binomial",
##     data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3303  -0.7775  -0.5095   0.8370   3.1617
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.897858   0.462450  -12.75  < 2e-16 ***
## X148         0.035582   0.003288   10.82  < 2e-16 ***
## X50          0.024502   0.007379    3.32 0.000899 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 796.49  on 764  degrees of freedom
## AIC: 802.49
##
## Number of Fisher Scoring iterations: 4
```

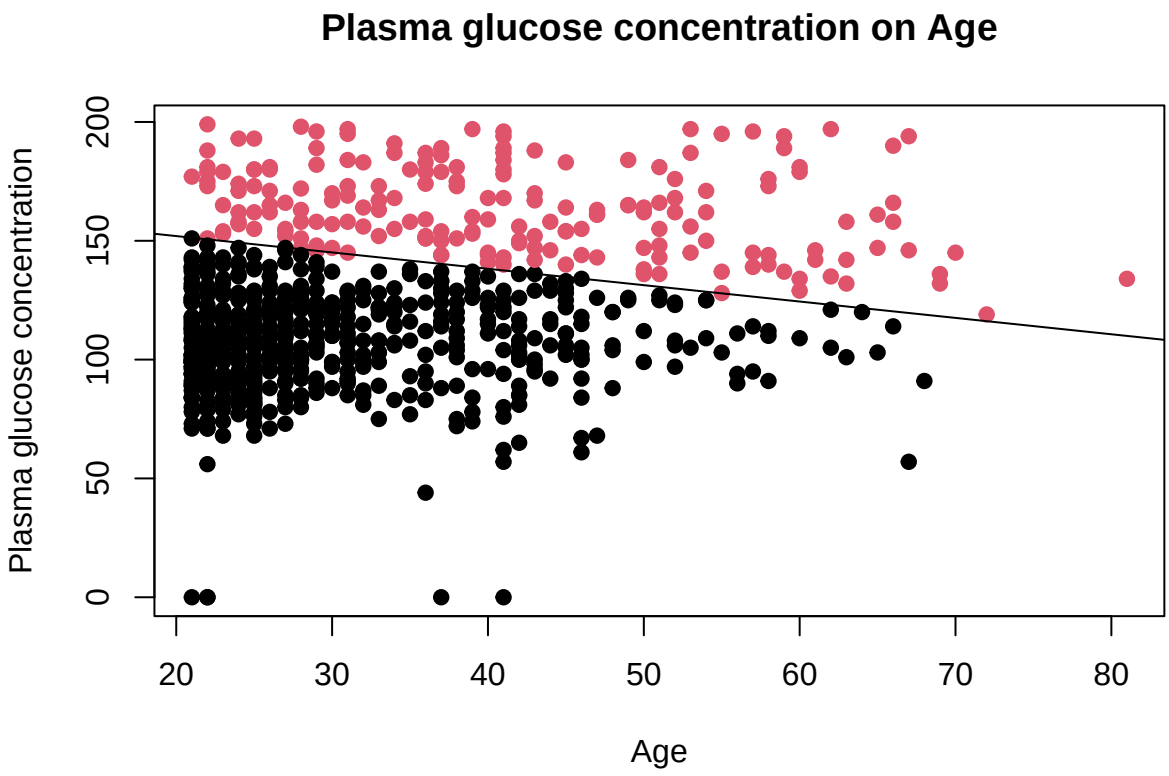
```
missclass=function(X,X1) {
  n=length(X)
  return(1-sum(diag(table(X,X1)))/n)
}

missclass(as.factor(diabetes1$X1), Pred)
```

```
## [1] 0.2659713
```

```
#Part 3


plot(x, y, col=as.factor(Pred), pch=19,
     main="Plasma glucose concentration on Age",
     xlab="Age", ylab="Plasma glucose concentration")
#The values in abline below are calculated by hand. Can be seen in the report
abline(165.7539767, -0.6886066)
```




```
#Part 4
# r = 0.2
Pred=ifelse(Prob>0.2, "1", "0")
table(train$X1, Pred)
```

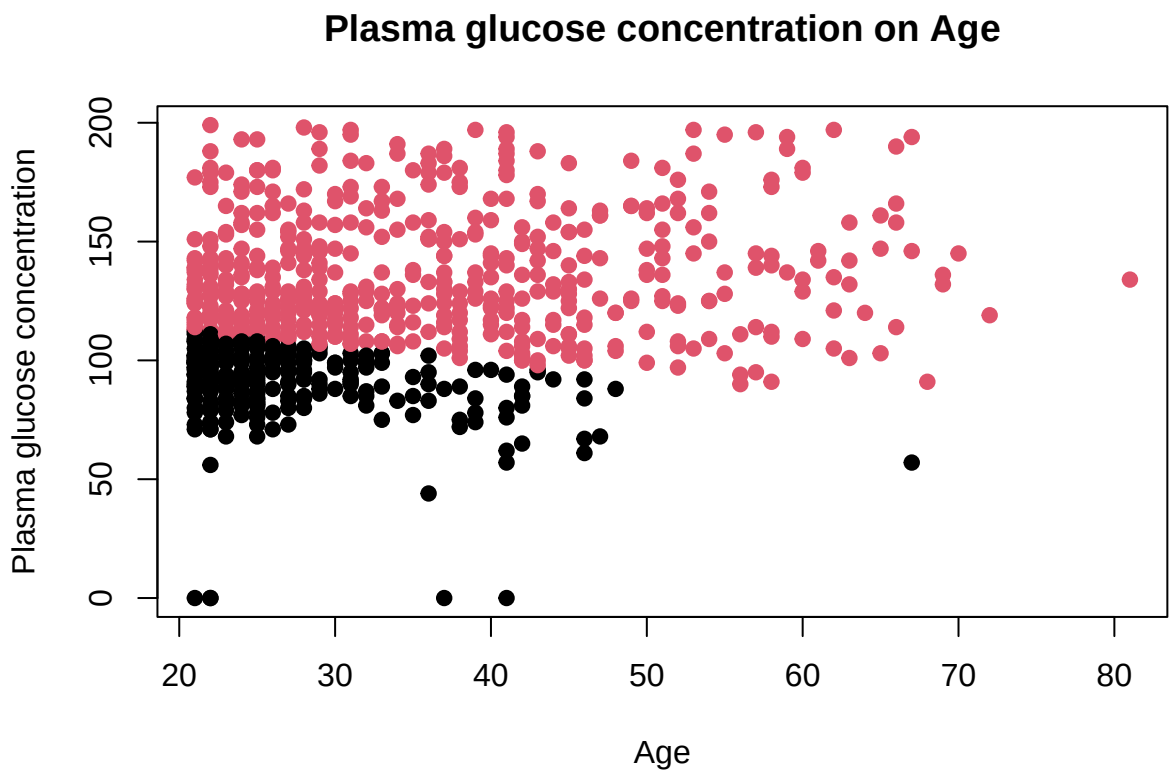
```
##    Pred
##       0   1
##   0 238 262
##   1  25 242
```

```
summary(m1)
```

```
##
## Call:
## glm(formula = as.factor(diabetes1$X1) ~ ., family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3303  -0.7775  -0.5095   0.8370   3.1617
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.897858   0.462450  -12.75  < 2e-16 ***
## X148         0.035582   0.003288   10.82  < 2e-16 ***
```

```
## X50          0.024502   0.007379    3.32 0.000899 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 796.49  on 764  degrees of freedom
## AIC: 802.49
##
## Number of Fisher Scoring iterations: 4
```

```
plot(x, y, col=as.factor(Pred), pch=19,
     main="Plasma glucose concentration on Age",
     xlab="Age", ylab="Plasma glucose concentration")
```



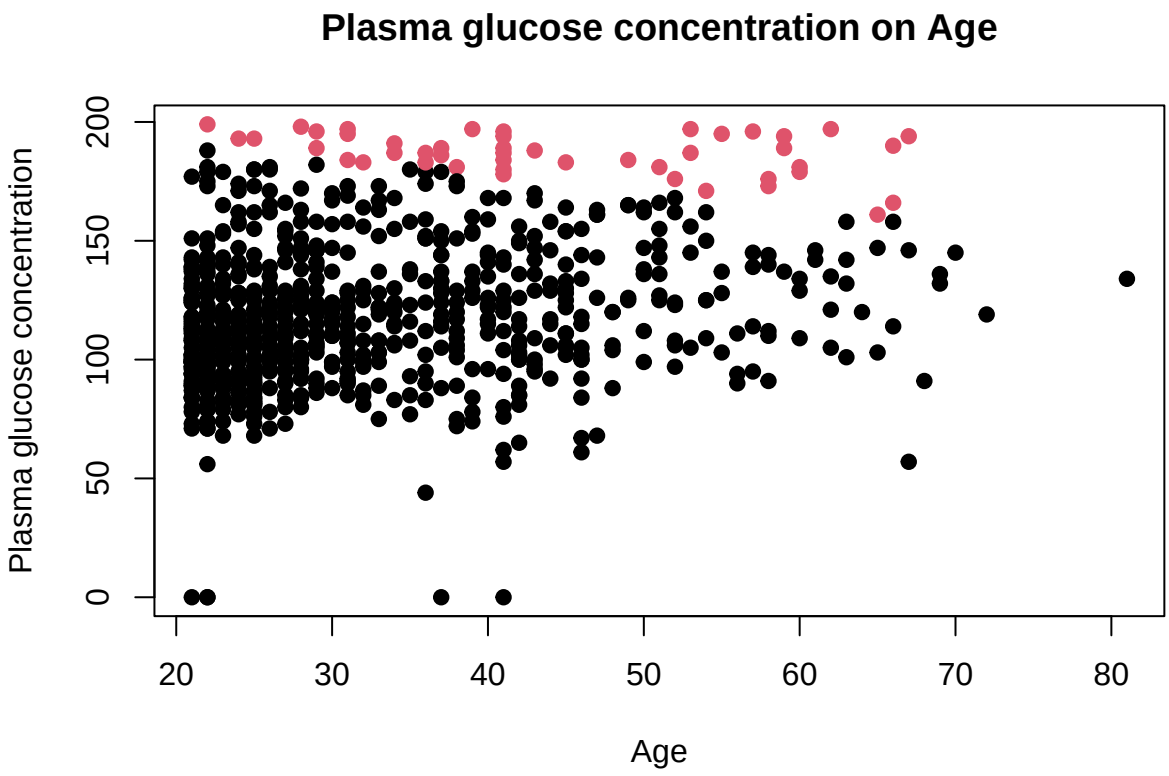

```
# r = 0.8
Pred=ifelse(Prob>0.8, "1", "0")
table(train$X1, Pred)
```

```
##    Pred
##       0   1
##   0 490  10
##   1 231  36
```

```r
summary(m1)
```

```
##
## Call:
## glm(formula = as.factor(diabetes1$X1) ~ ., family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3303  -0.7775  -0.5095   0.8370   3.1617
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.897858   0.462450  -12.75  < 2e-16 ***
## X148         0.035582   0.003288   10.82  < 2e-16 ***
## X50          0.024502   0.007379    3.32 0.000899 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 796.49  on 764  degrees of freedom
## AIC: 802.49
##
## Number of Fisher Scoring iterations: 4
```

```r
plot(x, y, col=as.factor(Pred), pch=19,
     main="Plasma glucose concentration on Age",
     xlab="Age", ylab="Plasma glucose concentration")
```




```
#Part 5

head(df)
```

```
##
## 1 function (x, df1, df2, ncp, log = FALSE)
## 2 {
## 3     if (missing(ncp))
## 4         .Call(C_df, x, df1, df2, log)
## 5     else .Call(C_dnf, x, df1, df2, ncp, log)
## 6 }
```

```
z1 = c((diabetes1$X50)^4)
z2 = c(((diabetes1$X50)^3)*diabetes1$X148)
z3 = c(((diabetes1$X50)^2)*((diabetes1$X148)^2))
z4 = c(diabetes1$X50*((diabetes1$X148)^3))
z5 = c(diabetes$X148^4)
X1 = diabetes1$X1
df = data.frame(z1, z2, z3, z4, z5, X1)
head(df)
```

```
##        z1      z2       z3        z4         z5 X1
## 1  923521 2532235  6943225  19037875   52200625  0
## 2 1048576 5996544 34292736 196111584 1121513121  1
## 3  194481  824229  3493161  14804349   62742241  0
```

```
## 4 1185921 4923369 20439441  84854649  352275361  1
## 5  810000 3132000 12110400  46826880  181063936  0
## 6  456976 1370928  4112784  12338352   37015056  1
```

```
train=df%>%select(X1, z1, )
m1=glm(as.factor(diabetes1$X1)~., train, family="binomial")
coef(m1)
```

```
##   (Intercept)            z1
## -7.479194e-01  4.771323e-08
```

```
Prob=predict(m1, type="response")
Pred=ifelse(Prob>0.5, "1", "0")
table(train$X1, Pred)
```

```
##    Pred
##       0   1
##   0 483  17
##   1 263   4
```

```
summary(m1)
```

```
##
## Call:
## glm(formula = as.factor(diabetes1$X1) ~ ., family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7582  -0.8928  -0.8845   1.4337   1.5028
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.479e-01  8.890e-02  -8.413  < 2e-16 ***
## z1           4.771e-08  1.789e-08   2.668  0.00764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 984.09  on 765  degrees of freedom
## AIC: 988.09
##
## Number of Fisher Scoring iterations: 4
```