Logistic Regression

Consider the diabetes dataset. We will take Outcome as the response, the other variables as predictors. We will build a reasonably good" logistic regression model for these data.

a) **Figure** 1 shows the boxplots of `Outcome` as a function of other predictor variables. Based on the plots it can be observe that all the predictors will be helpfull when separating patients with diabities and without diabities as there is a difference between distributions of patients with diabities and other for every predictor. Among them `Pregancy`, `Glucose`, `Insulin` and `Age` will be really helpful.

```
library(car)
library(lmtest)
library(ggplot2)
library(ISLR)
library(MASS)
```

```
diab<-read.csv("diabetes.csv")
head(diab)
```

```
  Pregnancies.. Glucose.. BloodPressure.. SkinThickness.. Insulin.. BMI..
1             2       138              62              35         0  33.6
2             0        84              82              31       125  38.2
3             0       145               0               0         0  44.2
4             0       135              68              42       250  42.3
5             1       139              62              41       480  40.7
6             0       173              78              32       265  46.5
  DiabetesPedigreeFunction.. Age.. Outcome
1                      0.127    47       1
2                      0.233    23       0
3                      0.630    31       1
4                      0.365    24       1
5                      0.536    21       0
6                      1.159    58       0
```

```
diab$Outcome<-as.factor(diab$Outcome)
names(diab)<-c("Pregnancies","Glucose","BP","Thickness","Insulin","BMI","DPB","Age","Outcome")
str(diab)
```

```
'data.frame':   2000 obs. of  9 variables:
 $ Pregnancies: int  2 0 0 0 1 0 4 8 2 2 ...
 $ Glucose    : int  138 84 145 135 139 173 99 194 83 89 ...
 $ BP         : int  62 82 0 68 62 78 72 80 65 90 ...
 $ Thickness  : int  35 31 0 42 41 32 17 0 28 30 ...
 $ Insulin    : int  0 125 0 250 480 265 0 0 66 0 ...
 $ BMI        : num  33.6 38.2 44.2 42.3 40.7 46.5 25.6 26.1 36.8 33.5 ...
 $ DPB        : num  0.127 0.233 0.63 0.365 0.536 ...
 $ Age        : int  47 23 31 24 21 58 28 67 24 42 ...
 $ Outcome    : Factor w/ 2 levels "0","1": 2 1 2 2 1 1 1 1 1 1 ...
```

```
contrasts(diab$Outcome)
```

```
  1
0 0
1 1
```

```
plot3a<-ggplot(diab, aes(x=Outcome, y=Pregnancies,fill=Outcome)) +
    geom_boxplot() + theme(legend.position = "none",axis.title.x = element_text(size=10), axis.title.y = element
plot3b<-ggplot(diab, aes(x=Outcome, y=Glucose,fill=Outcome)) +
    geom_boxplot() + theme(legend.position = "none",axis.title.x = element_text(size=10), axis.title.y = element

plot3c<-ggplot(diab, aes(x=Outcome, y=BP,fill=Outcome)) +
    geom_boxplot() + ylab("Blood Pressure")+ theme(legend.position = "none",axis.title.x = element_text(size=10)

plot3d<-ggplot(diab, aes(x=Outcome, y=Thickness,fill=Outcome)) +
    geom_boxplot() +  ylab("Skin Thickness") + theme(legend.position = "none",axis.title.x = element_text(size=1

plot3e<-ggplot(diab, aes(x=Outcome, y=Insulin,fill=Outcome)) +
    geom_boxplot() + theme(legend.position = "none",axis.title.x = element_text(size=10), axis.title.y = element

plot3f<-ggplot(diab, aes(x=Outcome, y=BMI,fill=Outcome)) +
    geom_boxplot() + theme(legend.position = "none",axis.title.x = element_text(size=10), axis.title.y = element

plot3g<-ggplot(diab, aes(x=Outcome, y=DPB,fill=Outcome)) +
    geom_boxplot() + ylab("Diabetes Pedigree Function") + theme(legend.position = "none",axis.title.x = element_

plot3h<-ggplot(diab, aes(x=Outcome, y=Age,fill=Outcome)) +
    geom_boxplot() + theme(legend.position = "none",axis.title.x = element_text(size=10), axis.title.y = element

require(gridExtra)
grid.arrange(plot3a, plot3b, plot3c, plot3d, plot3e, plot3f, plot3g, plot3h, ncol=8)
```

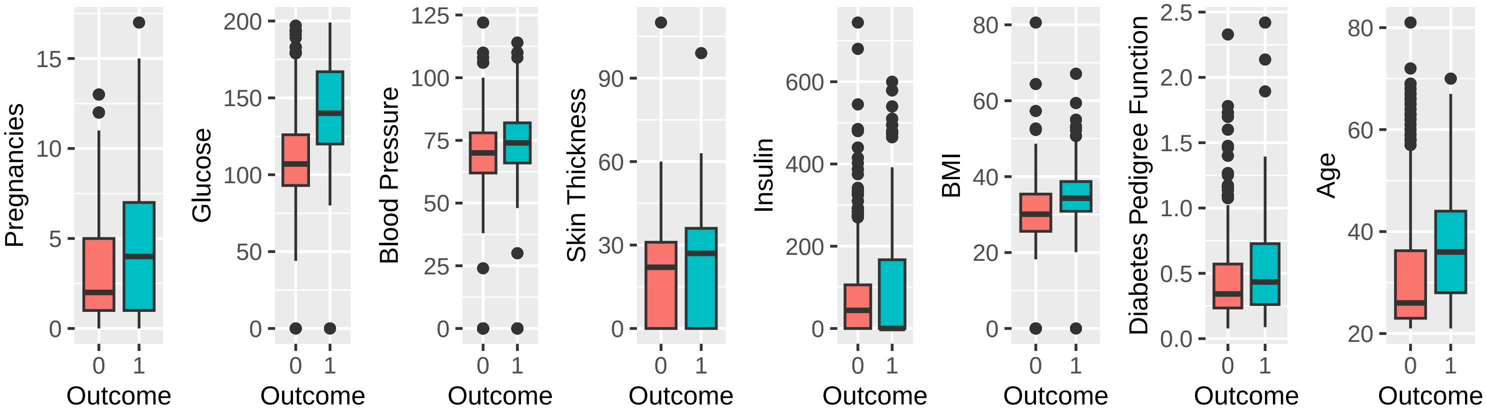

Figure 1: *Class conditional distributions for the diabities data.*

```
#part b)
full.model <- glm(Outcome ~ Pregnancies + Glucose + BP + Thickness + Insulin + BMI + DPB + Age, family = binomia
summary(full.model)
```

```

Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BP + Thickness +
    Insulin + BMI + DPB + Age, family = binomial, data = diab)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.0264511  0.4306345 -18.639  < 2e-16 ***
Pregnancies  0.1263845  0.0199997   6.319 2.63e-10 ***
Glucose      0.0337202  0.0022258  15.150  < 2e-16 ***
BP          -0.0096446  0.0032441  -2.973  0.00295 **
Thickness    0.0005185  0.0042301   0.123  0.90244
Insulin     -0.0012426  0.0005786  -2.148  0.03175 *
BMI          0.0775549  0.0088819   8.732  < 2e-16 ***
DPB          0.8877583  0.1860275   4.772 1.82e-06 ***
```

```
Age          0.0129414  0.0057020   2.270  0.02323 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2569.4  on 1999  degrees of freedom
Residual deviance: 1914.3  on 1991  degrees of freedom
AIC: 1932.3

Number of Fisher Scoring iterations: 5
```

```r
red.model <- glm(Outcome ~ Pregnancies + Glucose + BP + Insulin + BMI + DPB + Age, family = binomial, data = dia
summary(red.model)
```

```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BP + Insulin + 
    BMI + DPB + Age, family = binomial, data = diab)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -8.0273146  0.4306244 -18.641  < 2e-16 ***
Pregnancies  0.1263707  0.0199944   6.320 2.61e-10 ***
Glucose      0.0336810  0.0022020  15.296  < 2e-16 ***
BP          -0.0095806  0.0032013  -2.993  0.00276 ** 
Insulin     -0.0012123  0.0005228  -2.319  0.02042 *  
BMI          0.0778743  0.0084946   9.167  < 2e-16 ***
DPB          0.8894946  0.1855205   4.795 1.63e-06 ***
Age          0.0128944  0.0056879   2.267  0.02339 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2569.4  on 1999  degrees of freedom
Residual deviance: 1914.3  on 1992  degrees of freedom
AIC: 1930.3

Number of Fisher Scoring iterations: 5
```

```r
anova(red.model, full.model, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Outcome ~ Pregnancies + Glucose + BP + Insulin + BMI + DPB + 
    Age
Model 2: Outcome ~ Pregnancies + Glucose + BP + Thickness + Insulin + 
    BMI + DPB + Age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1992     1914.3                     
2      1991     1914.3  1 0.015033   0.9024
```

```r
null.model <- glm(Outcome ~ 1, family = binomial, data = diab)
anova(null.model, red.model, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Outcome ~ 1
Model 2: Outcome ~ Pregnancies + Glucose + BP + Insulin + BMI + DPB + 
    Age
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
```

```
1      1999     2569.4
2      1992     1914.3  7   655.06 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

- First, built a **model 1** by taking `Outcome` as response variable and all other variable as predictors. When testing the significance of jth coefficient: i.e $H_0 : \beta_j = 0$ vs $H_0 : \beta_j \neq 0$, we do not reject the null hypothesis for predictors `Thickness` as its p value is $0.90244 > 0.05$. Thus we can conclude that predictor `Thickness` is not associated with response `Outcome` after adjusting for the other predictors.
- Then built a **model 2** by removing `Thickness` from the **model 1**. To compare two nested models - full(**model 1**) and reduced(**model 2**), change in deviance statistic is calculated and p value for chisquare test is $0.9024 > 0.05$. Therefore, we do not reject null hypothesis (H_0 : full model = reduced model) and conclude that **model 2** is pretty much good as the **model 1**.
- Finally test of model significance for **model 2** was carried by taking null model(model which has a common intercept and no predictors) as the reduced model and **model 2** as the full model. Change in deviance statistic is calculated and p value for chisquare test is $2.2 \times 10^{-16} < 0.05$. Therefore, we can reject null hypothesis (H_0 : full model = reduce model) and conclude that **model 2** is significant.

c) Let p be probability of success (probability of getting diabetes). Then the final model:

$$\begin{aligned} logit(p) = &-8.0273 + 0.1264 Pregnancies + 0.0337 Glucose - 0.0096 BP - 0.0012 Insulin \\ &+ 0.0779 BMI + 0.8895 BPB + 0.0129 Age \end{aligned}$$

- exp(0.1264) = 1.13470. Therefore we expect to see about 13.5% increase in the odds of having diabities, for a one-unit increase in Pregnancies given that other variables held constant..
- exp(0.0337) = 1.0342. Therefore we expect to see about 3% increase in the odds of having diabities, for a one-unit increase in Glucose given that other variables held constant.
- Training error rate for the model is 0.216

```r
#part c)
CI<-confint(red.model)
rownames(CI)<-NULL
std.coef<-coef(summary(red.model))[, "Std. Error"]
names(std.coef)<-NULL
est<-cbind(coef(red.model),std.coef,CI)
colnames(est)<-c("Estimate","Std.Error","2.5%","97.5%")
```

```r
library(xtable)
xtab<-xtable(est,caption="summary of estimates of the regression
coefficients")
digits(xtab)<-c(0,4,4,4,4)
print(xtab,table.placement="H")
```

% latex table generated in R 4.3.0 by xtable 1.8-4 package % Sat Nov 18 00:17:21 2023

	Estimate	Std.Error	2.5%	97.5%
(Intercept)	-8.0273	0.4306	-8.8896	-7.2009
Pregnancies	0.1264	0.0200	0.0874	0.1659
Glucose	0.0337	0.0022	0.0294	0.0381
BP	-0.0096	0.0032	-0.0159	-0.0033
Insulin	-0.0012	0.0005	-0.0022	-0.0002
BMI	0.0779	0.0085	0.0615	0.0948
DPB	0.8895	0.1855	0.5275	1.2549
Age	0.0129	0.0057	0.0017	0.0240

Table 1: summary of estimates of the regression coefficients

```r
exp(coef(red.model))
```

```
 (Intercept)  Pregnancies      Glucose           BP      Insulin          BMI
0.0003264236 1.1347027661 1.0342546108 0.9904651313 0.9987884571 1.0809867255
         DPB          Age
2.4338993683 1.0129778520
```

```r
pred.prob.diab <- predict(red.model,diab, type = "response")
pred.diab <- ifelse(pred.prob.diab >= 0.5, "1", "0")
err.rate = 1 - mean(pred.diab == diab[, "Outcome"])
```