Practical 6: Logistic Regression Model Diagnostics (Binary Data)

The data to be used in this practical are from a survey of female labor force participation. The data is stored in a STATA file called "mroz.dta", which is available on Canvas.

```
library(foreign)
mroz.df<-read.dta("C:\\Temp folder\\mroz.dta")
attach(mroz.df)
```

View the names of the variables in the data-frame:

```
names(mroz.df)
```

1. Fit a logistic regression model on labor force participation with **age**, number of kids under age 6 (**kidslt6**) and the number of kids aged 6 or over (**kidsge6**).

```
glm1 <- glm(inlf ~ age + kidslt6 + kidsge6,
family = binomial (link = logit))

summary(glm1)
```

2. Calculate the residuals and other diagnostic statistics for the model as follows:

```
h<-lm.influence(glm1)$hat

rpear <-residuals(glm1, "pearson")/sqrt(1-h)

rdev <-residuals(glm1, "deviance")/sqrt(1-h)

phat<- glm1$fitted.values
```

```
D<-rpear*rpear*h/(4*(1-h))
```

3. Plot the Pearson residuals versus the linear predictor. What should you expect?

```
plot(glm1$linear.predictors, rpear,main="Plot of
Pearson Residuals v Linear Predictor")
```

From notes:
Of the methods studied for Binomial data, the following are applicable to binary data:

- **Index Plot: plot of residuals versus case number.**
- **Leverage Values**
- **Influence Values**

4. Obtain an index plot of the residuals. What should you expect? What do you conclude?

```
plot(rpear)
```

```
identify(rpear, n = 2)
```

5. Obtain an index plot of the leverage values. What do you conclude?

```
plot(h, main="Index Plot of Leverage Values")
abline(h=0.0106, lty=1)
```

```
identify(h, n = 2)
```

6. Obtain an index plot of the influence (D) values. What do you conclude?

```
plot(D, main="Index Plot of Cook's Distance
Values")
```

```
identify(D, n = 2)
```

7. Investigate the cases of concern.

Outliers/High influence:

```
mroz.df[c(74,400), c(1,3,4,5)]
```

High leverage:

```
mroz.df[c(53, 720), c(1,3,4,5)]
```