Homework #4

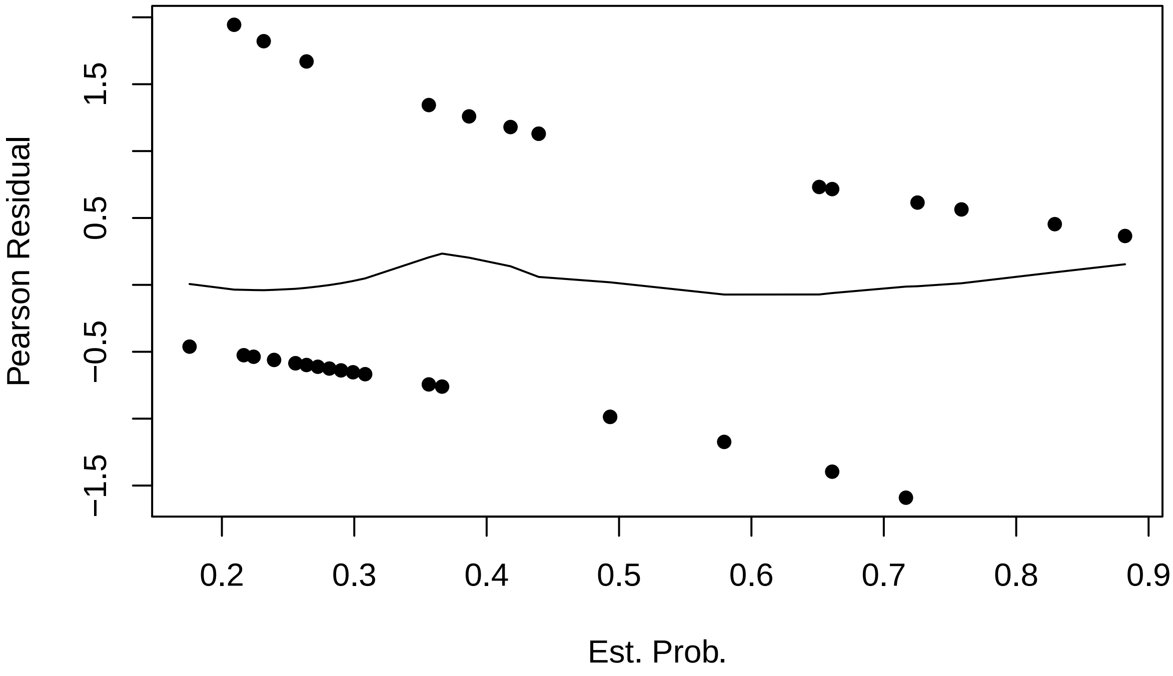
Zhijian Liu

14.27

b. Obtain the studentized Pearson residuals and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

- Use a first order model with 1 predictors X1. (This is the result from Problem 14.21 (d).)
- In addition to the plot and your comments, report the values of the studentized Pearson residuals (some books may call them "standardized Pearson residual") of the first 3 cases.

```
reg <- glm(Y ~ X1, family = binomial(link = logit), data = df, x=T)
p.hat <- reg$fitted
# Get Hat matrix for Logistic Regression
W <- diag(p.hat*(1-p.hat))
X<- reg$x
H<- sqrt(W)%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrt(W)
h <- diag(H)
# Get studentized Pearson residuals
e <- df$Y - p.hat #original residual
rp<- e/sqrt(p.hat*(1-p.hat))  # Pearson Residual
rsp <- rp/sqrt(1-h)  # Studentized/Standardized Pearson
# plot
index <- sort(p.hat,index.return=T)$ix
loe<-loess(rp ~ p.hat, degree=1)
plot(p.hat, rp, type="p", pch=16, xlab="Est. Prob.", ylab="Pearson Residual")
lines(p.hat[index], loe$fitted[index], type="l")
```



```
# first 3 cases
rsp[1:3]
```

```
##          1          2          3
## -0.7583796 -1.0080273  0.7621641
```

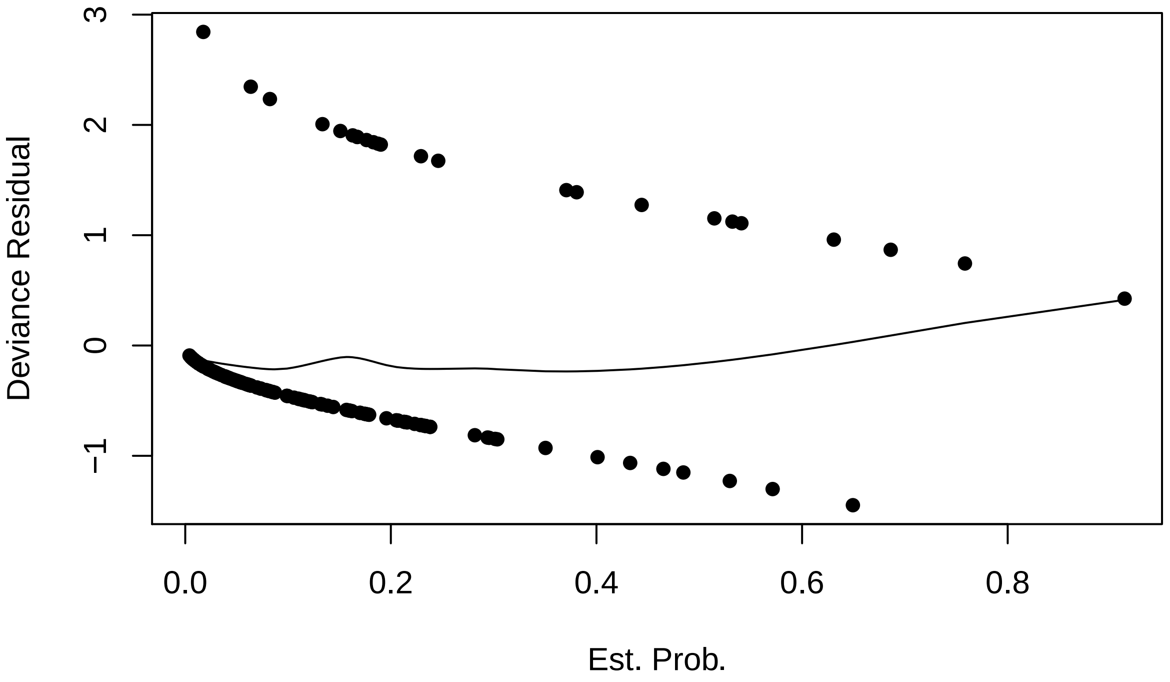
The plot of studentized Pearson residuals against the estimated model probabilities and the first 3 cases of the studentized Pearson residuals are shown as above. The roughly flat lowess line in the plot suggests a good fit the model.

14.28

c. Obtain the deviance residuals and plot them against the estimated model probabilities with a lowess smooth superimposed. What does the plot suggest about the adequacy of the fit of the logistic regression model?

- Use a first order model with 2 predictors X1 and X2. (This is the result from Problem 14.22 (d).)
- In addition to the plot and your comments, report the values of the deviance residuals of the first 3 cases.

```
reg <- glm(Y ~ X1 + X2, family = binomial(link = logit), data = df, x=T)
p.hat <- reg$fitted
# Get Hat matrix for Logistic Regression
W <- diag(p.hat*(1-p.hat))
X<- reg$x
H<- sqrt(W)%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrt(W)
h <- diag(H)
# Get sDeviance residual
dev <- residuals(reg)  # Deviance residual
sdev <- dev/sqrt(1-h)  # Standardized Deviance residual
# plot
index <- sort(p.hat,index.return=T)$ix
loe<-loess(dev ~ p.hat, degree=1)
plot(p.hat, dev, type="p", pch=16, xlab="Est. Prob.", ylab="Deviance Residual")
lines(p.hat[index], loe$fitted[index], type="l")
```



```
# first 3 cases
dev[1:3]
```

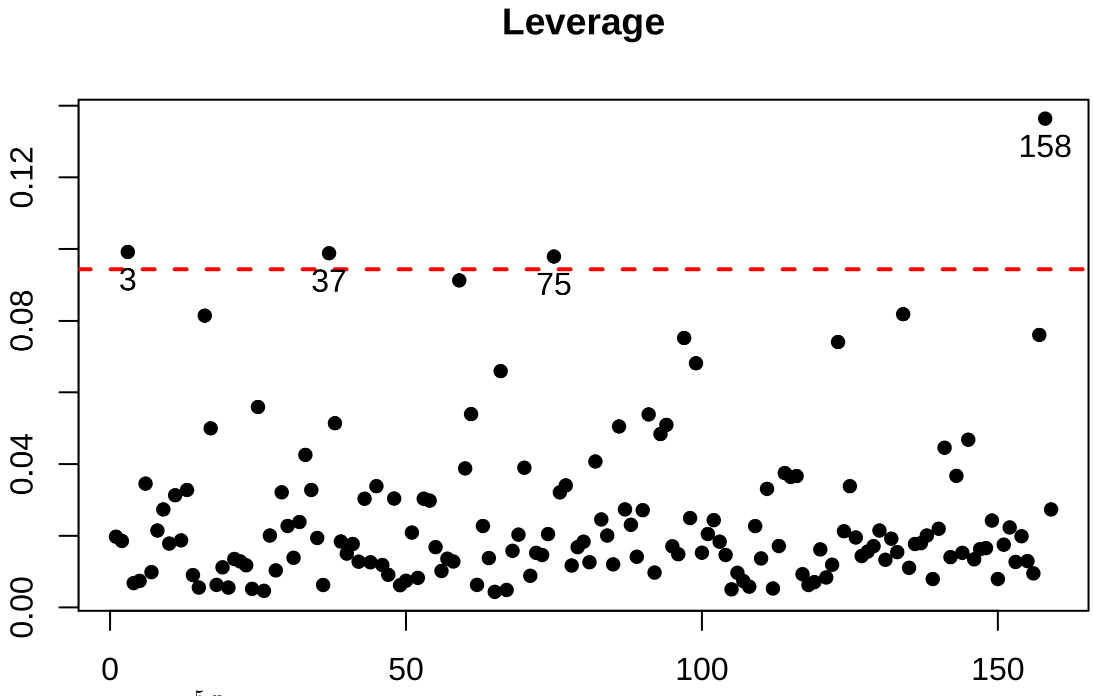
```
##          1          2          3
## -0.5460231 -0.5137326  1.1526024
```

The plot of deviance residuals against the estimated model probabilities and the first 3 cases of the deviance residuals are shown as above. The lowess line in the plot appears to have an upward trend, and this signal of having a pattern indicates some drawback of the model

14.32

a. For the logistic regression fit in Problem 14.14a, prepare an index plot of the diagonal elements of the estimated hat matrix (14.80). Use the plot to identify any outlying X observations.

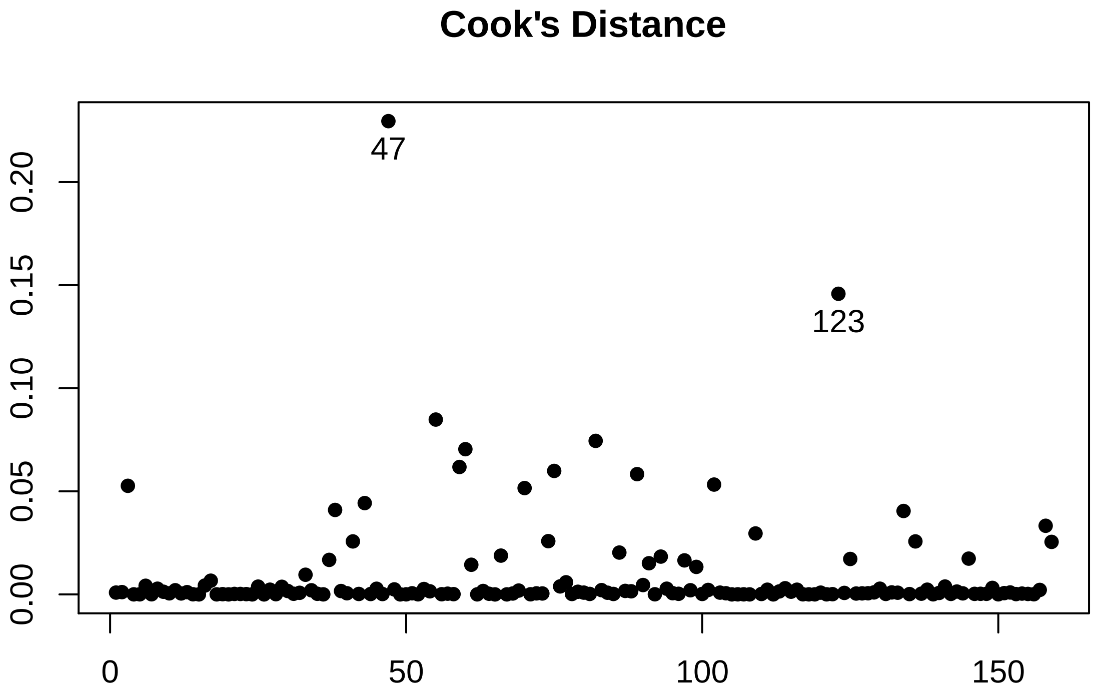
```
reg <- glm(Y ~ X1 + X2 + X3, family = binomial(link = logit), data = df, x=T)
p.hat <- reg$fitted
# Get Hat matrix for Logistic Regression
W <- diag(p.hat*(1-p.hat))
X<- reg$x
H<- sqrt(W)%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrt(W)
h <- diag(H)
# plot
plot(h, main="Leverage", pch=16, ylab="", xlab="")
p <- 3
I <- nrow(df)
lv.line <- 5*p/I # 6*sum(h)/I
lv<- h > 5*p/I
abline(lv.line, 0, lty=2,lwd=2, col=2)
text(c(1:length(h))[lv], h[lv], labels=c(1:length(h))[lv], pos=1)
```



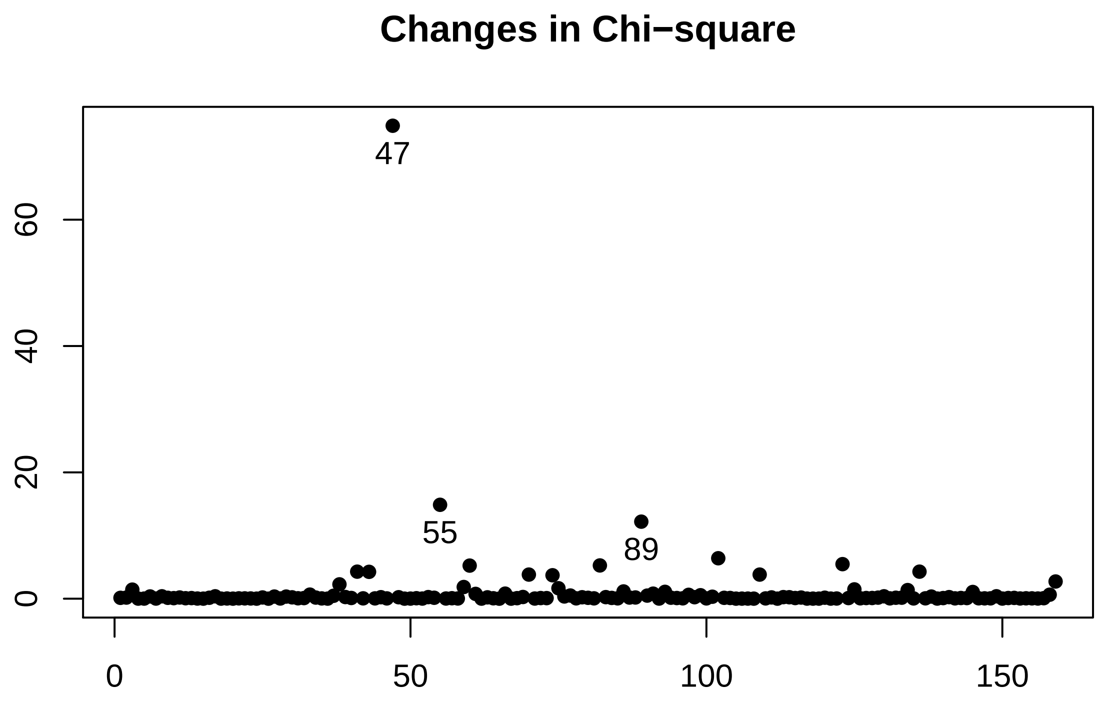
I set the level $\frac{5 \cdot p}{I} = 0.09433962$ to identify the outliers if the leverage of an observation is higher than this level. The index plot pinpoints the 3^{rd}, 37^{th}, 75^{th} and 158^{th} observations are the most extrem outliers in the X space.

b. To assess the influence of individual observations, obtain the delta chi-square statistic (14.85), the delta deviance statistic (14.86), and Cook's distance (14.87) for each observation. Plot each of these in seperate index plots and identify any influential observations. Summarize your findings.

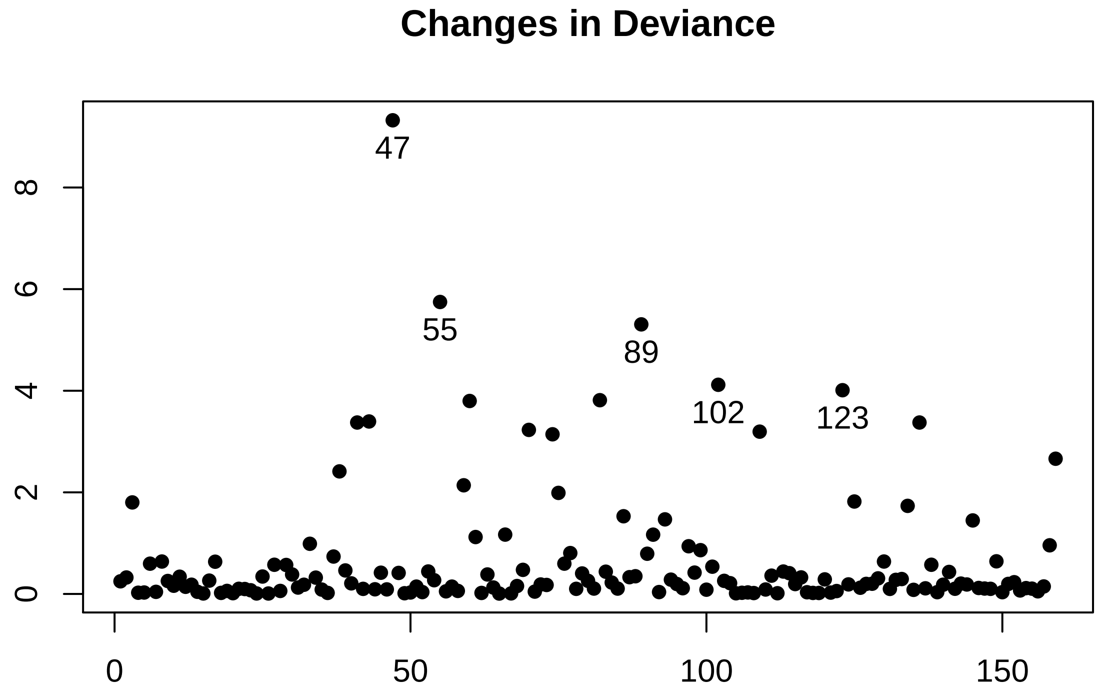
```
# Cook's Distance
rp <- (df$Y-p.hat)/sqrt(p.hat*(1-p.hat))  # Pearson Residual
D <- rp^2*(h)/(p*(1-h)^2)
# Change in Pearson Chi-square
rsp <- rp/sqrt(1-h)  # Studentized Pearson residual
dChi<-rsp^2
# Change in Deviance
dev<-residuals(reg)
ddev<-h*rsp^2+dev^2
# plots
# Cook's Distance
plot(D, main="Cook's Distance", ylab="", xlab="", pch=16)
lv <- D > 9*mean(D)
text(c(1:length(D))[lv], D[lv], labels=c(1:length(h))[lv], pos=1)
```



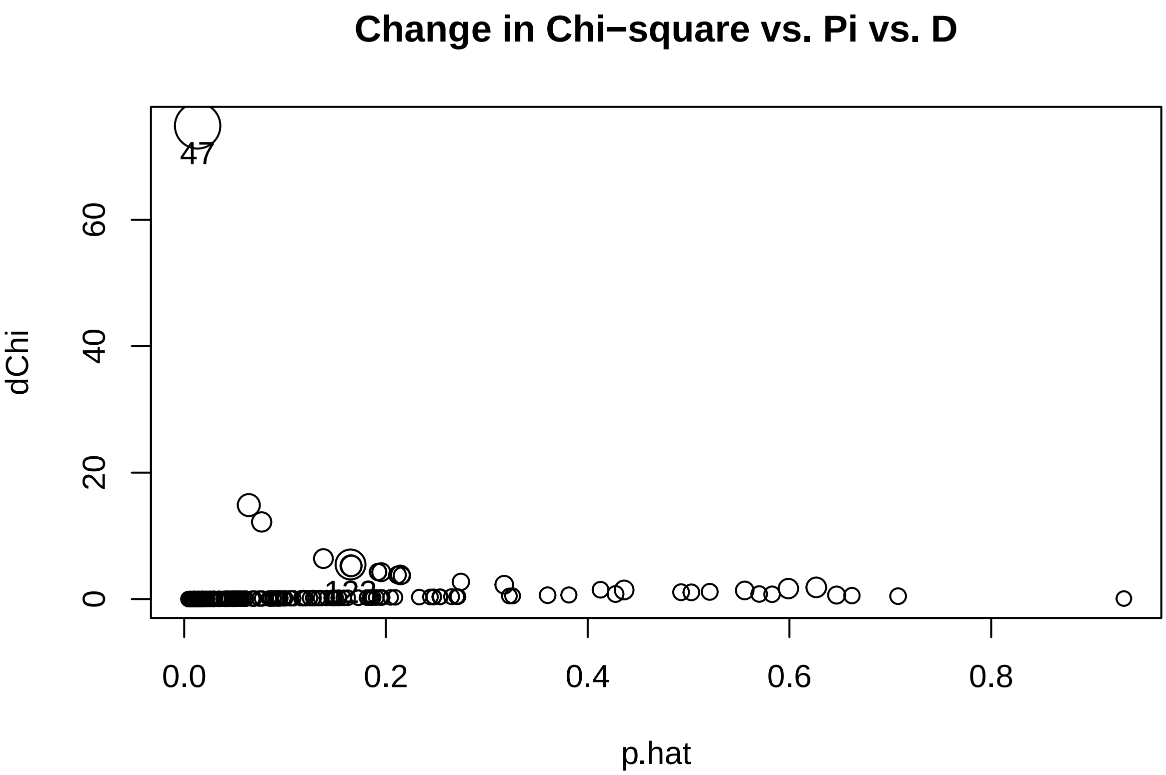
```
# Change in Chi-square
plot(dChi, main="Changes in Chi-square", pch=16, ylab="", xlab="")
lv <- dChi > 7
text(c(1:length(dChi))[lv], dChi[lv], labels=c(1:length(dChi))[lv], pos=1)
```



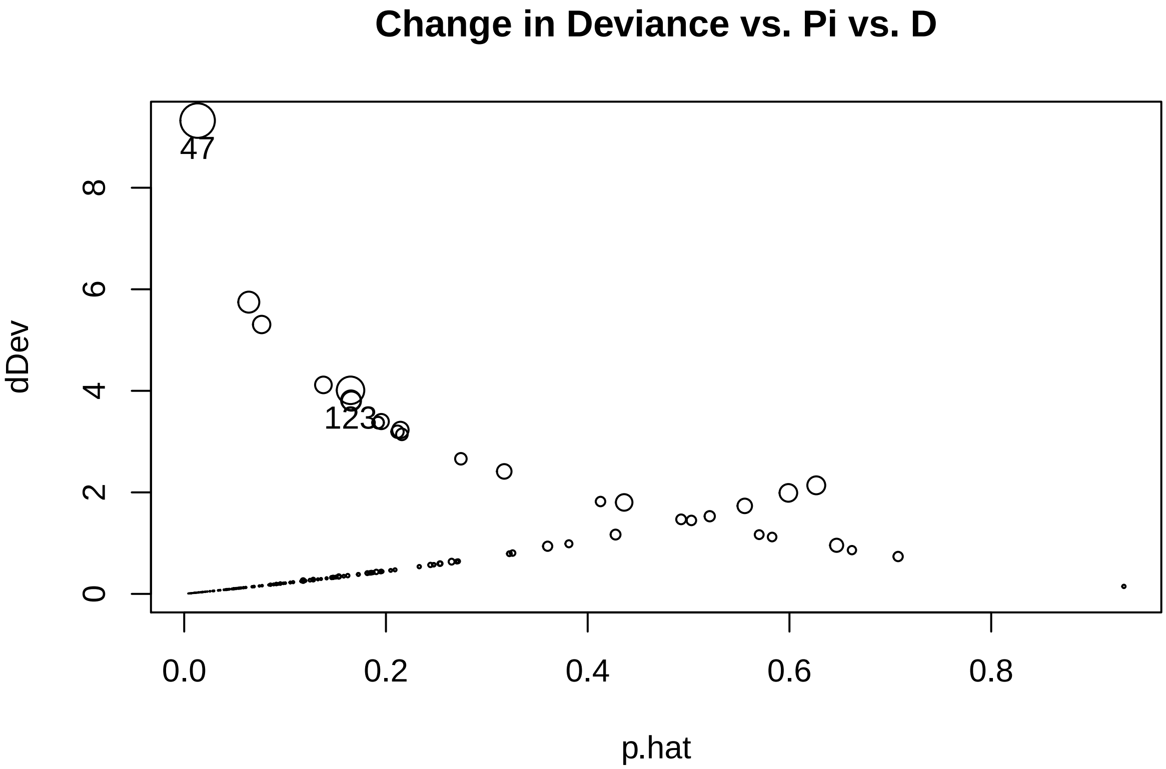
```
# Change in Deviance
plot(ddev, main="Changes in Deviance", pch=16, ylab="", xlab="")
lv<- ddev > 4
text(c(1:length(ddev))[lv], ddev[lv], labels=c(1:length(ddev))[lv], pos=1)
```



```
# bubble plot
lv<- D > 9*mean(D)
plot(p.hat, dChi, cex=exp(D*5), main="Change in Chi-square vs. Pi vs. D", ylab="dChi")
text(p.hat[lv], dChi[lv], labels=c(1:length(p.hat))[lv], pos=1)
```



```
plot(p.hat, ddev, cex=sqrt(D)*5, main="Change in Deviance vs. Pi vs. D", ylab="dDev")
text(p.hat[lv], ddev[lv], labels=c(1:length(p.hat))[lv], pos=1)
```



The plots of different statistics share the same suggestion that the 47^{th} observation is the most influential to the model. Meanwhile, the 123^{th} observation is also highly suspected to be influential, since it has a very high Cook's distance, Chi-square change and deviance change.