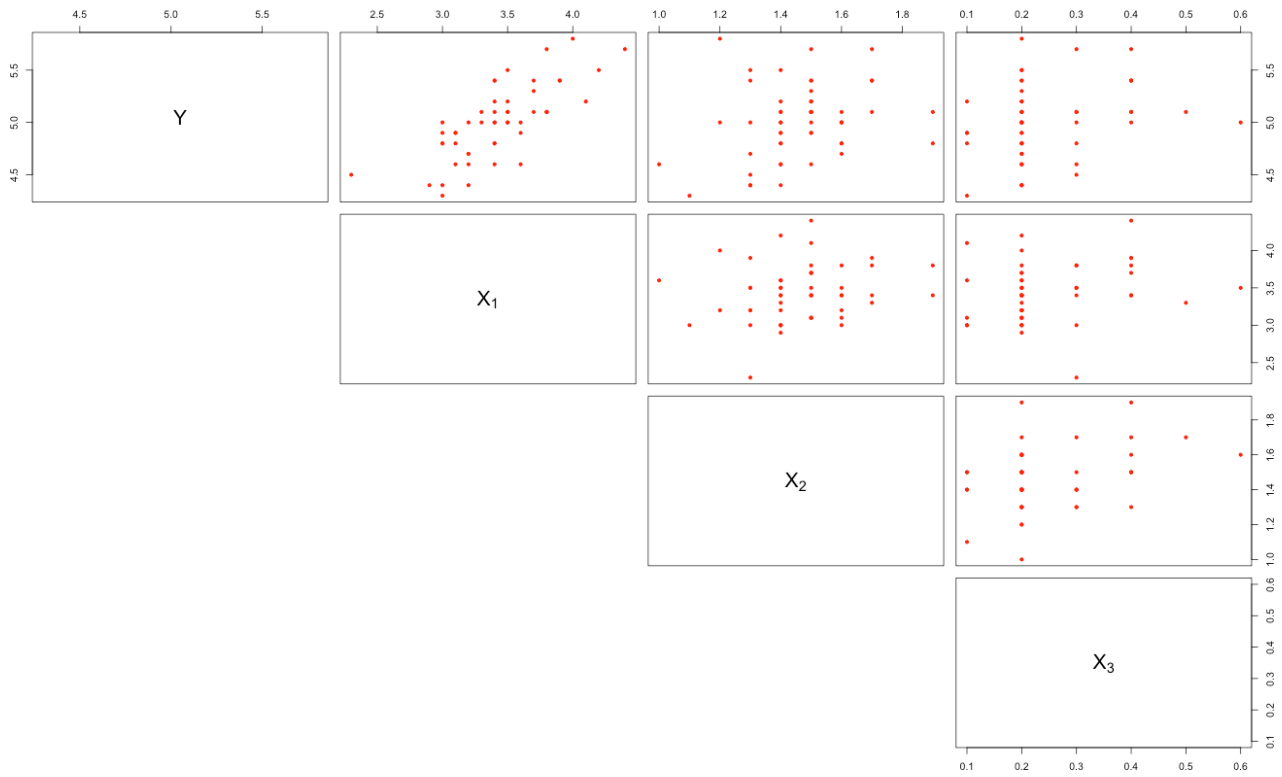
Modelo linear

É apresentado um exemplo de ajuste em **R** com algumas ferramentas de diagnóstico aplicadas a um modelo linear normal. Os dados estão descritos e listados no exercício 12 no arquivo http://wiki.icmc.usp.br/images/e/e5/Lista1SME0812-2014.pdf.

```
dados <- . . . # completar
n <- nrow(dados)
p <- ncol(dados) # Neste exemplo

summary(dados)
```

```
       Y               X1              X2              X3
 Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
 1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
 Median :5.000   Median :3.400   Median :1.500   Median :0.200
 Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
 3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
 Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
```

```
pairs(dados, labels = c("Y", expression(X[1]), expression(X[2]),
   expression(X[3])), pch = 20, lower.panel = NULL, col = "red")
```



Nota 1. Comente os gráficos acima.

Um modelo linear é ajustado com a função `lm`.

```
m1 <- lm(Y ~ X1 + X2 + X3, data = dados)
names(m1)
```

```
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
```

Nota 2. Explique cada um dos componentes do objeto `m1` acima.

Nota 3. Refaça o ajuste utilizando a função `glm`.

```
summary(m1)
```

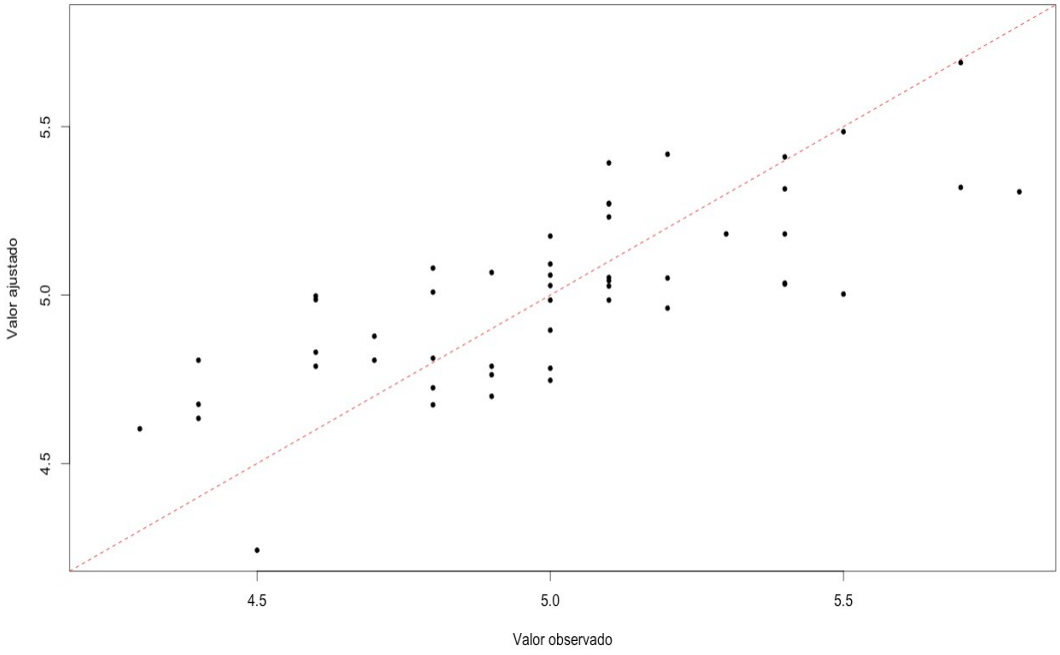
```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40662 -0.17721  0.01222  0.13388  0.49693

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.35189    0.39287   5.986 3.03e-07 ***
X1           0.65483    0.09245   7.083 6.83e-09 ***
X2           0.23756    0.20802   1.142    0.259
X3           0.25213    0.34686   0.727    0.471
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2371 on 46 degrees of freedom
Multiple R-squared: 0.5751,     Adjusted R-squared: 0.5474
F-statistic: 20.76 on 3 and 46 DF,  p-value: 1.192e-08
```

Nota 4. Comente os resultados gerados pela função `summary`.

Os valores observados e ajustados pelo modelo são apresentados abaixo.

```
rm1 <- range(dados$Y, m1$fitted.values)
plot(dados$Y, m1$fitted.values, pch = 20, xlab = "Valor observado",
     ylab = "Valor ajustado", xlim = rm1, ylim = rm1)
abline(0, 1, lty = 2, col = "red")
```

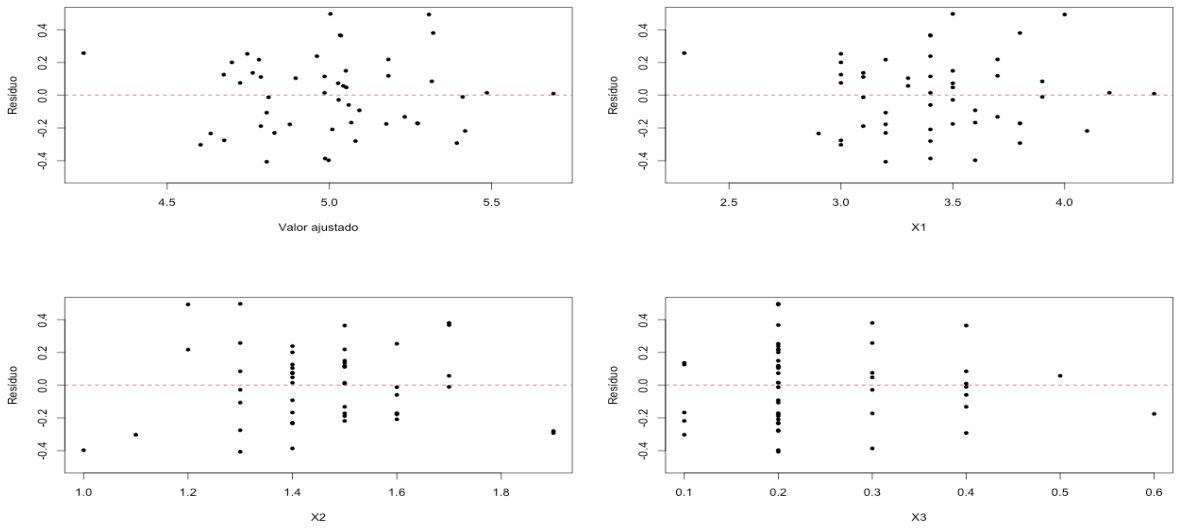
Nota 5. Justifique o uso da função `range` no trecho acima.



Nota 6. Comente o resultado do ajuste com base no gráfico acima.

Os gráficos dos resíduos ordinários *versus* valores preditos e também variáveis explicativas são mostrados em seguida.

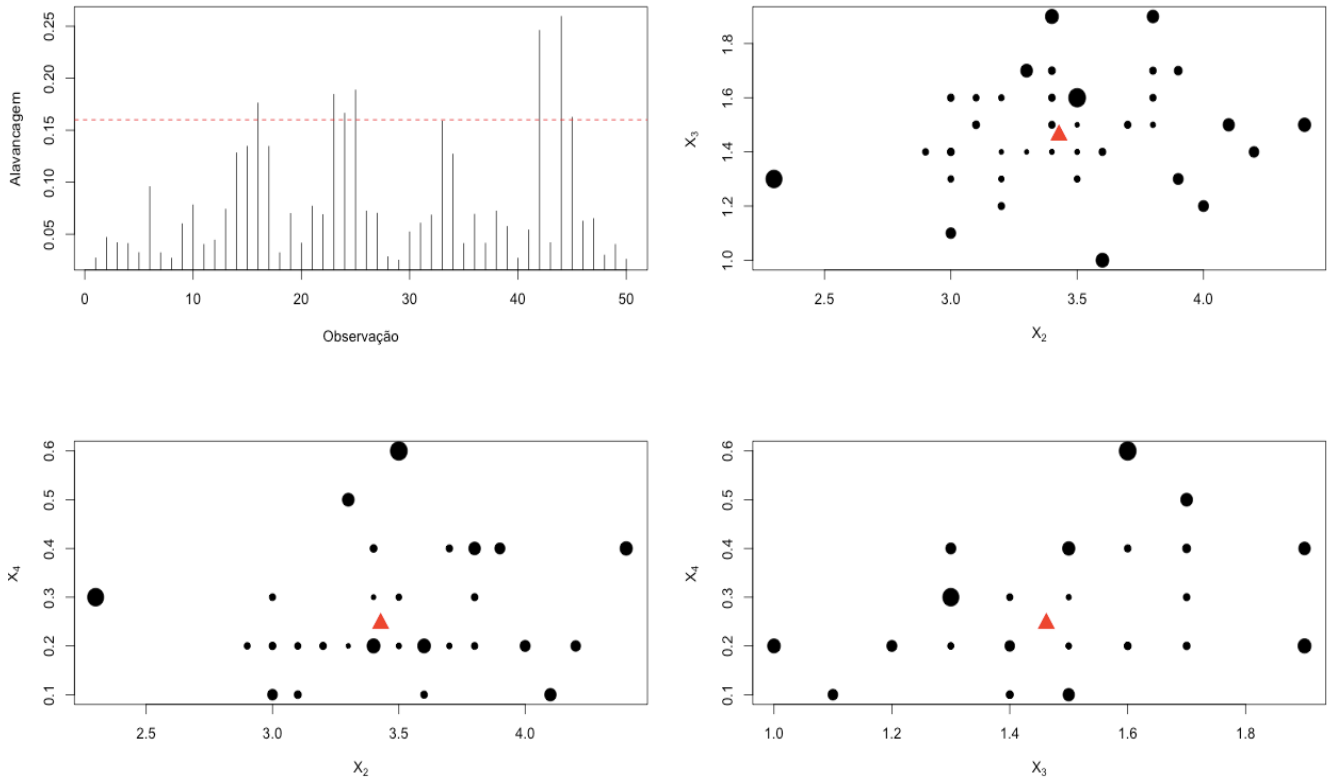
```
par(mfrow = c(2, 2))
maxe <- max(abs(m1$resid))
plot(m1$fitted.values, m1$residuals, pch = 20, ylab = "Resíduo",
     xlab = "Valor ajustado", ylim = c(-maxe, maxe))
abline(h = 0, lty = 2, col = "red")
for (j in 2:4) {
   plot(dados[, j], m1$residuals, pch = 20, ylab = "Resíduo",
        xlab = names(dados)[j], ylim = c(-maxe, maxe))
   abline(h = 0, lty = 2, col = "red")
}
```



Nota 7. Comente os gráficos acima relacionando-os com alguma suposição do modelo.

Alguns gráficos representando a medida de alavancagem são mostrados abaixo.

```
par(mfrow = c(2, 2))
minh <- min(lm.influence(m1)$h)
cexh <- 3 * (lm.influence(m1)$h - minh) / (max(lm.influence(m1)$h) - minh)
+ 1
plot(lm.influence(m1)$h, type = "h", xlab = "Observação",
     ylab = "Alavancagem")
abline(h = 2 * p / n, lty = 2, col = "red")
plot(X[, 2], X[, 3], pch = 20, cex = cexh, xlab = expression(X[2]), ylab =
expression(X[3]))
points(mean(X[, 2]), mean(X[, 3]), pch = 17, col = "red", cex = 2)
plot(X[, 2], X[, 4], pch = 20, cex = cexh, xlab = expression(X[2]), ylab =
expression(X[4]))
points(mean(X[, 2]), mean(X[, 4]), pch = 17, col = "red", cex = 2)
plot(X[, 3], X[, 4], pch = 20, cex = cexh, xlab = expression(X[3]), ylab =
expression(X[4]))
points(mean(X[, 3]), mean(X[, 4]), pch = 17, col = "red", cex = 2)
```



Nota 8. Explique os resultados gerados pela função `lm.influence`.

Nota 9. Nos gráficos acima identifique as observações com alavancagem mais alta.

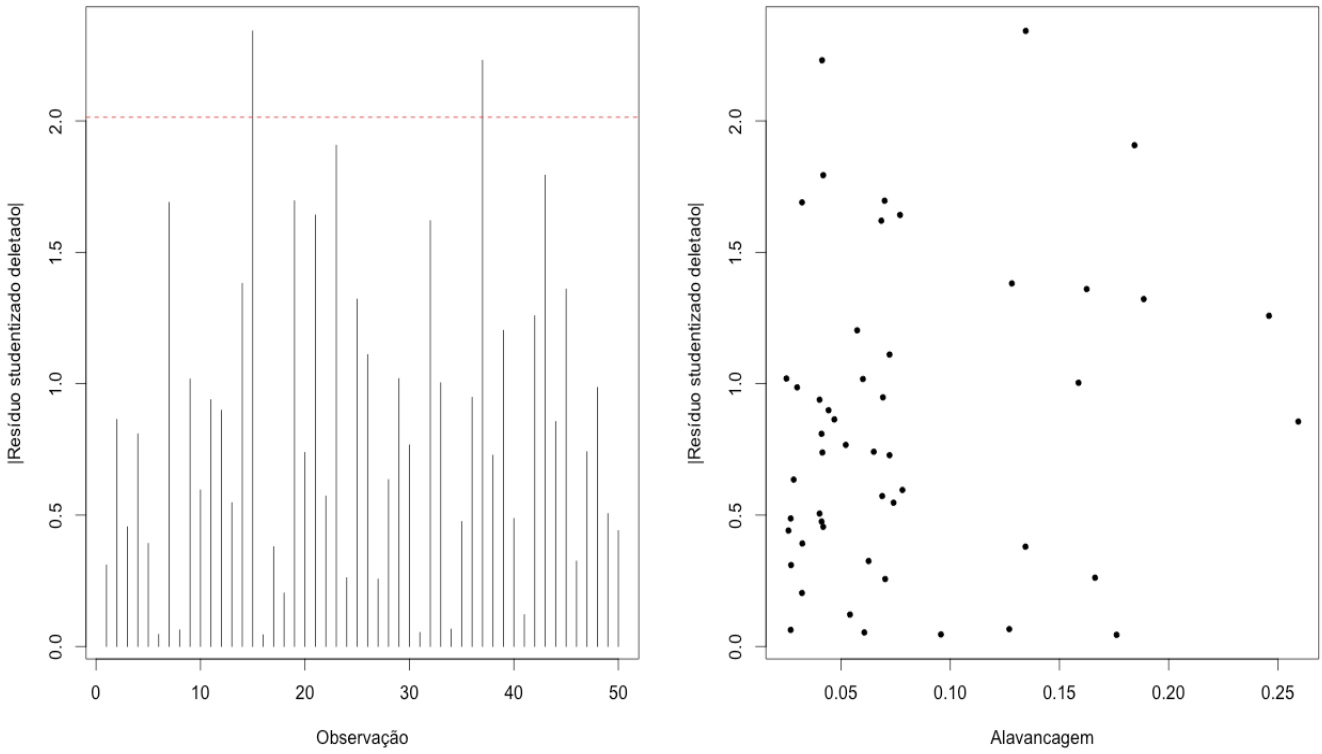
Nota 10. Calcule a matriz chapéu (***H***).

Dois gráficos com os resíduos studentizados deletados são apresentados a seguir.

```
# Resíduo studentizado deletado
ti <- m1$resid * sqrt((n - p - 1) /
   (summary(m1)$s^2 * m1$df.residual * (1 - lm.influence(m1)$h) -
   m1$resid^2))

summary(ti)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.908000 -0.798800  0.053680 -0.001755  0.583500  2.343000
```

```
plot(abs(ti), type = "h", xlab = "Observação",
     ylab = "|Resíduo studentizado deletado|")
abline(h = qt(0.975, df = n - p - 1), lty = 2, col = "red")

plot(lm.influence(m1)$h, abs(ti), pch = 20, xlab = "Alavancagem",
     ylab = "|Resíduo studentizado deletado|")
```

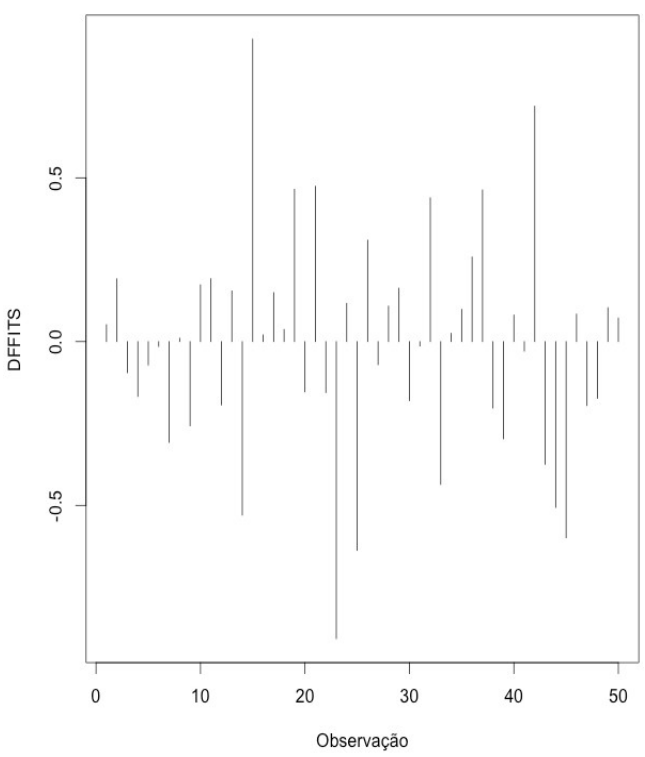


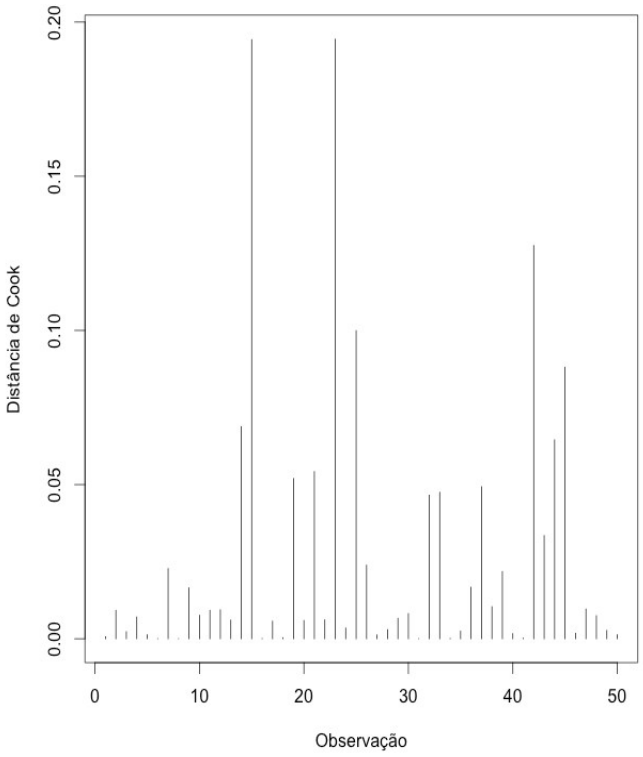
Nota 11. Comente os gráficos acima relacionando-os com alguma suposição do modelo.

Nota 12. Realize o teste da hipótese de que não existem observações afastadas na variável resposta.

Para ajudar a identificar observações influentes, são apresentados gráficos de índices das medidas DFFITS e distância de Cook. Além destas, para cada coeficiente do modelo, é apresentada a mudança na estimativa decorrente da exclusão de uma observação de cada vez.

```
# Observações influentes
dffits <- ti * sqrt(lm.influence(m1)$h / (1 - lm.influence(m1)$h))
dcook <- m1$resid^2 * lm.influence(m1)$h /
   (p * summary(m1)$s^2 * (1 - lm.influence(m1)$h)^2)

par(mfrow = c(1, 2))
plot(dffits, type = "h", xlab = "Observação", ylab = "DFFITS")
plot(dcook, type = "h", xlab = "Observação", ylab = "Distância de Cook")
```



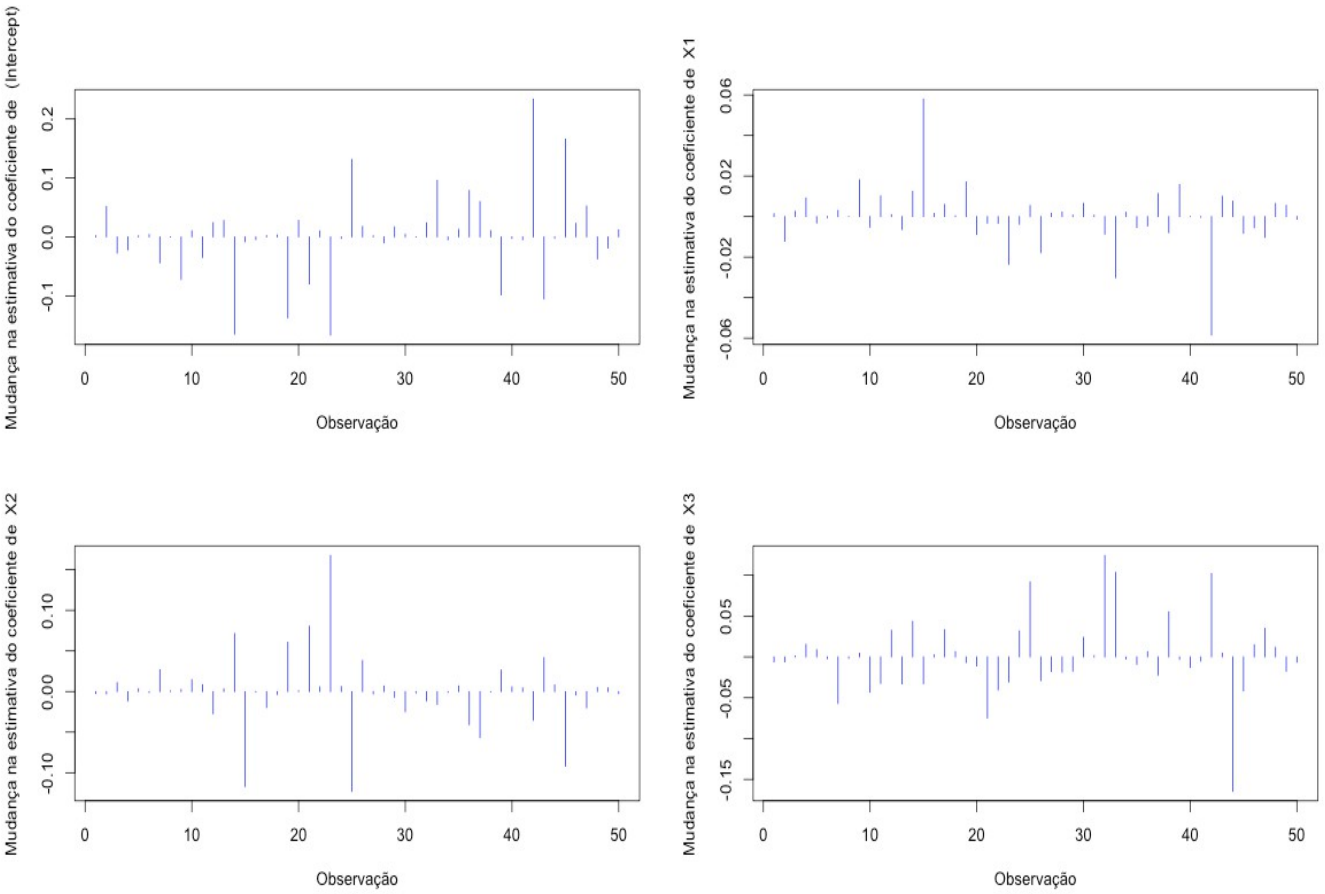


Nota 13. Identifique as observações mais influentes nos gráficos acima. Tente apontar alguma diferença marcante delas em relação às demais.

```
par(mfrow = c(2, 2))
for (j in 1:4) {
   plot(lm.influence(m1)$coef[, j], type = "h", xlab = "Observação",
      ylab = paste("Mudança na estimativa do coeficiente de ",
      names(coef(m1))[j]), col = "blue")
}
```

Nota 14. Identifique as observações mais influentes nos gráficos abaixo. Tente apontar alguma diferença marcante delas em relação às demais.

Nota 15. Calcule DFBETAS e represente graficamente.



Nota 16. Efetue o teste da hipótese H: $\beta_{X2} = \beta_{X3} = 0$. Caso seja possível simplificar o modelo, refaça o exemplo com o modelo mais simples.

Nota 17. Procure reproduzir todos os resultados utilizando outros pacotes estatísticos (SAS, SPSS, Minitab e Statistika, por exemplo).