Validação cruzada

Os dados foram coletados em 845 áreas metropolitanas (AM) dos EUA no ano de 1992. A variável resposta é a venda *per capita* no varejo (em US$ 1000). As variáveis explicativas são X_1: número de pontos de venda *per capita*, X_2: renda *per capita* (em US$ 1000), X_3: despesa federal *per capita* (em US$ 1000) e X_4: número de homens para cada 100 mulheres. A descrição das variáveis e os dados podem ser obtidos nas páginas http://www.stat.ufl.edu/~winner/data/retail92.txt e http://www.stat.ufl.edu/~winner/data/retail92.dat.

```
## Dados
mydata <- ... completar ...
colnames(mydata) <- c("AM", "Y", paste("X", 1:4, sep =""))

cat("\n n =", n <- nrow(mydata))
```

```
     n = 845
```

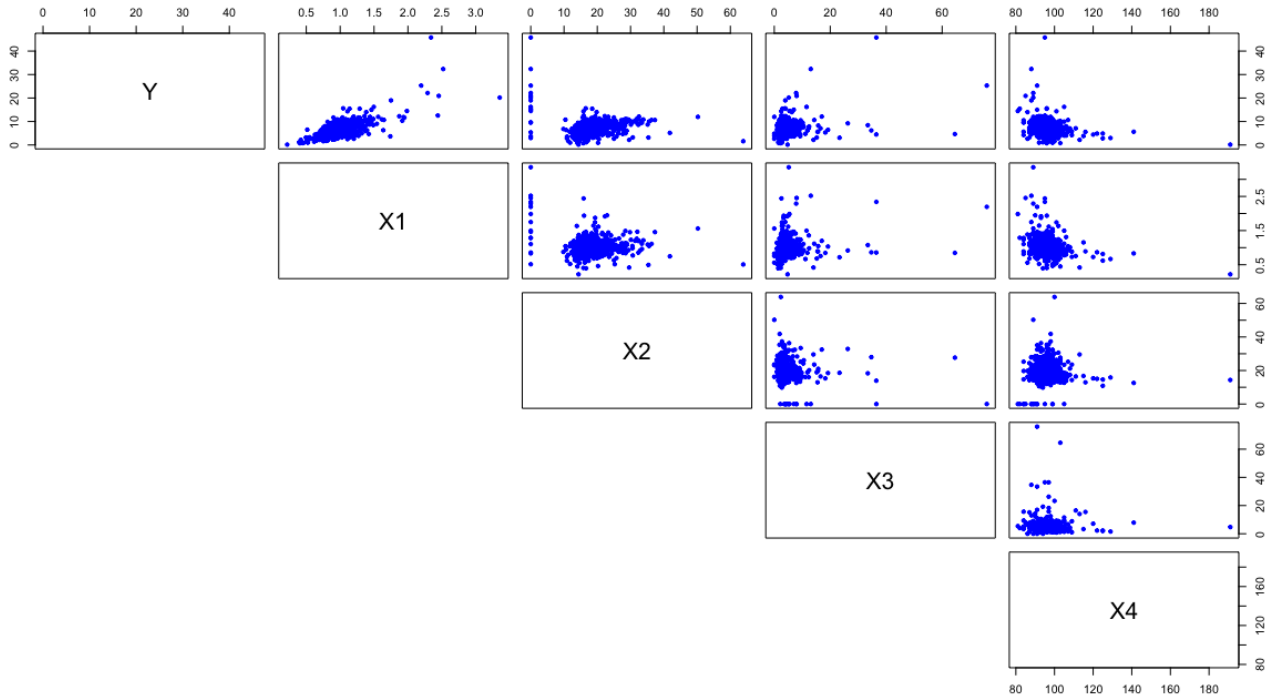
```
summary(mydata[, -1])
```

```
       Y                X1               X2        
 Min.   : 0.152   Min.   :0.219   Min.   : 0.00  
 1st Qu.: 5.419   1st Qu.:0.905   1st Qu.:15.88  
 Median : 7.139   Median :1.014   Median :17.68  
 Mean   : 7.190   Mean   :1.031   Mean   :18.38  
 3rd Qu.: 8.677   3rd Qu.:1.126   3rd Qu.:20.20  
 Max.   :45.759   Max.   :3.354   Max.   :63.87  
       X3               X4        
 Min.   : 0.000   Min.   : 81.00  
 1st Qu.: 2.584   1st Qu.: 93.00  
 Median : 3.253   Median : 96.00  
 Mean   : 4.154   Mean   : 96.03  
 3rd Qu.: 4.366   3rd Qu.: 98.00  
 Max.   :75.952   Max.   :191.00  
```

Existem observações com valores nulos das variáveis X_2 e X_3. O exemplo será desenvolvido incluindo estas observações.

```
pairs(mydata[, -1], pch = 20, lower.panel = NULL, col = "blue")
```

Nota 1. Comente o gráfico abaixo.



A validação cruzada é baseada em $k = 5$ pastas (ou seja, *five-fold crossvalidation*). Desta forma, os conjuntos de teste e treinamento são formados por 169 (= n/k) e 676 (= $n - n/k$) observações, respectivamente. Os índices das observações são armazenados em uma matriz (`indC`).

```
## Validação cruzada
set.seed(14901)
k <- 5  # Número de pastas
indC <- matrix(sample(n), nrow = k)
```

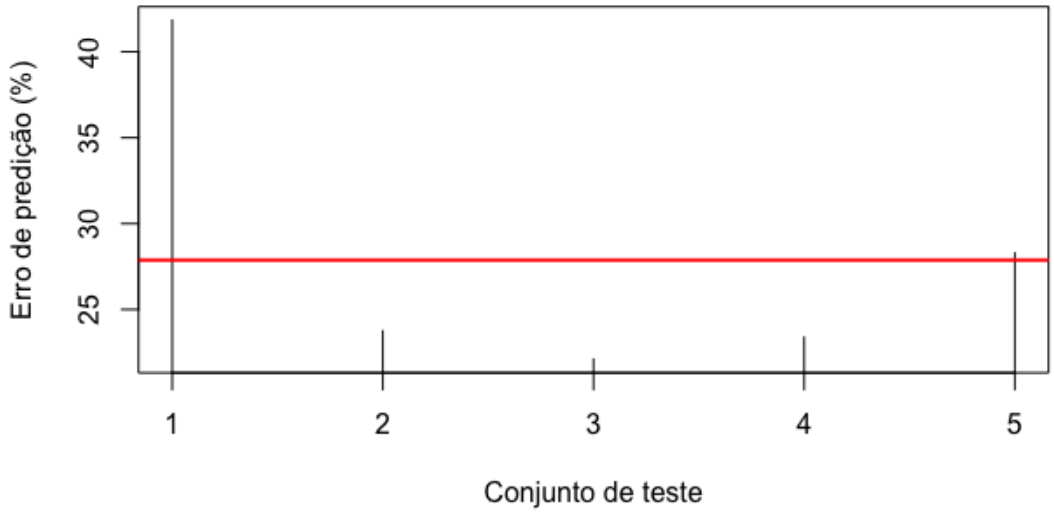
Na j–ésima etapa de treinamento, o argumento `subset` seleciona os índices das observações utilizadas no ajuste do modelo. Em seguida, a função `predict` calcula as predições para as observações do j–ésimo conjunto de teste (`newdata`). Finalmente, o erro de predição percentual (baseado no módulo do erro de predição relativo) na j–ésima é calculado, $j = 1,...,k$.

```
erropred <- c()
for (j in 1:k) {
   # Treinamento
   modc <- lm(Y ~ X1 + X2 + X3 + X4, data = mydata, subset =
         indC[-j,])
   # Teste
   ypred <- predict(modc, newdata = mydata[indC[j,],
      c("X1",   "X2", "X3", "X4")])
   # Erro percentual
    erropred[j]  <-  mean(abs((mydata$Y[indC[j,]]  -  ypred)/
     mydata$Y[indC[j,]])) * 100
}
erropredm <- mean(erropred)
```

```
cat("\n Erro de predição percentual:", round(erropredm, 2))
```

```
     Erro de predição percentual: 27.87
```

```
plot(erropred, xlab = "Conjunto de teste", ylab = "Erro de predição",
     type = "h")
abline(h = erropredm, col = "red", lwd = 2)
```



Nota 2. Verifique em um dos conjuntos de treinamento se o modelo faz um bom ajuste.

Nota 3. Refaça o exemplo eliminando as observações com valores nulos das variáveis X_2 e X_3.

Nota 4. Refaça o exemplo utilizando 10 pastas (10-*fold crossvalidation*).